Unimodularity in Random Networks: Applications to the Null recurrent Doeblin Graph and Hierarchical Clustering

Sayeh Khaniha

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure de Paris

Unimodularity in Random Networks: Applications to the Null recurrent Doeblin Graph and Hierarchical Clustering

Soutenue par
Sayeh Khaniha
Le 26 September 2023

École doctorale n°386
Sciences Mathématiques de Paris Centre

Spécialité
Mathematics

Composition du jury :

Nicolas Curien Université Paris-Sud Orsay	*Rapporteur*
Ádám TIMAR University of Iceland	*Rapporteur*
Kasra Alishahi Sharif University of Thechnology	*Examinateur*
Javad Ebrahimi Broojeni Sharif University of Thechnology	*Examinateur*
Alexandre Gaudilliere Institut de Mathematiques de Marseille	*Examinateur*
Eva Locherbach Universite de Paris 1 Pantheon Sorbonne	*présidente du jury*
Hermann Thorisson University of Iceland	*Examinateur*
Mir-Omid Haji-Mirsadeghi Sharif University of Thechnology	*Co-Directeur de thèse*
François Baccelli École Normale Supérieure	*Directeur de thèse*

1794 ENS | PSL

Résumé

Cette thèse repose sur la notion d'unimodularité dans le contexte des réseaux aléatoires et explore deux domaines d'application : le couplage par le passé des chaînes de Markov dans le cas de la récurrence nulle, basé sur les graphes de Doeblin associés, et la classification non supervisée basée sur le regroupement hiérarchique des points d'un processus ponctuel.
La première partie de cette thèse se concentre sur les propriétés d'un graphe aléatoire spécifique appelé le graphe de Doeblin, qui est associé à l'algorithme du couplage par le passé utilisé pour l'échantillonnage parfait de la distribution stationnaire d'une chaîne de Markov. Cette thèse étudie le cas récurrent nul, où il est montré que le graphe des ponts, un sous-graphe du Doeblin Graph, est soit un arbre infini, soit une forêt composée d'une collection dénombrable d'arbres infinis. Dans le premier cas, l'arbre infini possède une unique extrémité, n'est généralement pas unimodularisable, mais présente une unimodularité locale. Ces propriétés sont exploitées pour étudier le régime stationnaire des dynamiques aléatoires de processus à valeurs mesures sur l'arbre des ponts de Doeblin, en particulier les dynamiques aléatoires tabou et potentielle.
La deuxième partie de cette thèse présente un nouveau modèle de regroupement hiérarchique adapté à la classification non supervisée d'ensembles de données qui sont dénombrablement infinis. L'algorithme proposé utilise plusieurs niveaux de regroupement, construisant des clusters à chaque niveau en utilisant des chaînes de voisins les plus proches de points ou de clusters. Cet algorithme est appliqué au processus ponctuel de Poisson pour lequel il est démontré que l'algorithme de regroupement définit une forêt phylogénétique, qui est un facteur du processus ponctuel et est donc unimodulaire. Diverses propriétés de cette forêt aléatoire, telles que les tailles moyennes des clusters à chaque niveau ou la taille moyenne du cluster d'un nœud typique, sont examinées.

Graphes Unimodulaires aléatoires ⋆ Processus ponctuels stationnaires ⋆ classification des feuillets ⋆ Échantillonnage parfait⋆ Classification non supervisée hiérarchique⋆ Processus ponctuels⋆ Arbres Aléatoires

Abstract

This thesis is based on the notion of unimodularity in the context of random networks and explores two domains of application: Coupling from the Past for Markov Chains in the null recurrent case based on the associated Doeblin Graphs, and unsupervised classification based on hierarchical clustering on point processes.
The first part of this thesis focuses on the properties of a specific random graph called the Doeblin Graph, which is associated with the Coupling from the Past algorithm used for the perfect sampling of the stationary distribution of a Markov Chain. This thesis studies the null recurrent case, where it is shown that the Bridge Doeblin Graph, a subgraph of the Doeblin Graph, is either an infinite tree or a forest composed of a countable collection of infinite trees. In the former case, the infinite tree possesses a single end, is not generally unimodularizable, but exhibits local unimodularity. These properties are leveraged to investigate the stationary regime of measure-valued random dynamics on the Bridge Doeblin Tree, particularly the taboo and potential random dynamics.
The second part of this thesis introduces a novel hierarchical clustering model tailored for unsupervised classifications of datasets that are countably infinite. The proposed algorithm employs multiple levels of clustering, constructing clusters at each level using nearest-neighbor chains of points or clusters. This algorithm is applied to the Poisson point process. It is proven that the clustering algorithm defines a phylogenetic forest on the Poisson point process, which is a factor of the point process and is therefore unimodular. Various properties of this random forest, such as the mean sizes of clusters at each level or the mean size of the cluster of a typical node, are examined.

Unimodular Random graphs⋆Stationary Point Processes⋆Foil Classification⋆Perfect Sampling⋆Unsupervised Classification⋆Point Processes⋆Random Trees

Acknowledgments

Before starting this section of the thesis, I thought writing it would be an easy task. However, as I attempted to put everything in my mind onto paper, I realized it's much more challenging than I anticipated. Now, as I write these sentences, it's the final days before submitting it to the university library. I'm not certain whether everyone I express gratitude to in this section will see it or not. Still, I write it so that in the future, when I look back at this thesis, I remember the atmosphere of these days.

The beginning of my work on my doctoral project coincided with my first entry into France. Before entering any society, the image one has of that community is based on hearsay. Upon entering that society, observations either present a new image or confirm those preconceived notions. These observations can vary from person to person since individuals each interact with different people. Most of the people I mention here are those who shaped my observations in this society, creating an image of Paris in my mind that goes beyond its beautiful architecture and charming streets. They are the individuals who made life in this city more beautiful. Perhaps I cannot express with my pen all that they were, but I will forever hold them in my heart and mind.

I believe the first person I should write about is my advisor, François. Besides being a knowledgeable and skilled professor in his field, he is a compassionate humanist. Throughout the four years of my doctoral program, I felt his presence by my side at every moment. He wasn't just a mentor teaching me mathematics and research; with him, I felt like a child again, this time in the world of mathematics, with a caring and knowledgeable father. Despite all shortcomings and mistakes, his dedicated efforts are aimed at helping me bring out a better version of myself. What more can define being an advisor than this? I am grateful to him for everything he was.

The second person I express my gratitude to is my second advisor, Mir-Omid, to whom I owe my experience in this academic environment. Mir-Omid always instilled in me a sense of confidence, enabling me to pursue my work with self-assurance and greater strength. His interesting ideas and unique perspectives on issues were often the key to unraveling the knots in my work.

Following him, it is time to express my gratitude to all members of our group: Bharath, Michel, and Pierre that the small group of "Baccelli Boys and Girls" start with them and

later continued with the joining of Ke. They were not just colleagues for me but also my friends. Bharath, the encyclopedic mind of our group, who could find the answer to any question either instantly or in a fraction of a second. Michel, who always generously shared his knowledge and made the passage of time more enjoyable with his humor and witty remarks. Bharath, both himself and his lovely wife, Sayanika, were steadfast friends and pillars of support for me during this period. They were always ready to listen to my concerns and issues, making the hard situation more bearable. Ke, who could always be relied upon for weekend getaways. She dedicated a significant amount of time to teach me climbing, and whenever I asked for assistance in this regard, I was met with a kind "yes" in response. I am confident that every time I look back on this period in the future, I will fondly remember this small group, and a smile of the sweetness of collaboration and friendship with them will play on my lips.

Next in line is my dear friend, Mahsa. Our acquaintance dates back to a brief meeting during our master's program in Iran. I wonder if on that particular Wednesday, she had not randomly come across my name in a shared group in France or had decided not to message me to find out about my presence in Paris, how different my life would have been. The day we went to the Sen River together marked the division of my life in France into before and after that. Her presence was a source of comfort during moments of distress because she always made time to listen to my words with kindness. She was a motivation for my endeavors, as she is a consistently energetic girl who always puts her best effort into tasks without considering the outcome. She is patient in tackling new challenges. Her passion for Western classical music introduced me to this captivating genre, and she was always ready to spend time on weekend outings and leisure activities. She was my constant companion for two-person chats over chocolate and tea at any hour of any day. I am very happy to have had her as a friend, and I am grateful for her constant companionship.

Contents

Chapter

1 Introduction

This thesis is comprised of two pieces of work on random networks. The first part is based on a preprint available on [1].
A random rooted network is considered unimodular when, heuristically, any vertex has an equal probability of being the root. To rigorously define this concept for networks with an infinite number of vertices, the assumption is made that the random rooted network satisfies the mass transport principle. The literature has studied the properties of unimodular networks and of covariant dynamics on them ([2] and [3]).
This thesis focuses on two specific random graphs that are associated with stochastic algorithms, and it investigates the properties of these graphs using the notion of unimodularity. The first algorithm under consideration is the Coupling from the Past algorithm, which is utilized for the perfect sampling of the stationary distribution of a Markov Chain. The second random graph pertains to a model used for hierarchical clustering of data. The thesis is structured into two distinct parts. A preliminary is given in Chapter 3 provides a common foundation for both parts. The subsequent material and results in each part are independent and can be read separately. This brief introduction provides an overview of the main themes explored in each part.

I. Doeblin Graph

Let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov Chain with countable state space, $\mathscr{S}$. It is well known that when $\{X_t\}_{t\in\mathbb{N}}$ is irreducible, aperiodic, and positive recurrent, it has a unique stationary distribution. On the other hand, when it is null recurrent, it admits no stationary probability distribution, but a unique stationary measure σ, i.e., the measure σ satisfies $\sigma = \sigma P$ with $\sigma(s^*) = 1$, where P is the transition probability matrix of the Markov Chain, and s^* is an arbitrary fixed point in $\mathscr{S}$.

One can consider a Markov Chain as a dynamics on $\mathscr{S}$-valued random variables. This dynamics can be written as the following equation [1]

$$X_{t+1} = \sum_{x\in\mathscr{S}} \mathbb{1}_{\{X_t=x\}} h(x, \xi_t^x), \tag{1.0.1}$$

[1] Another representation for Equation (1.0.1) is the stochastic recurrence equation $X_{t+1} = h(X_t, \xi_t)$.

where $\{\xi_t^x, x \in \mathscr{S}\}_{t\in\mathbb{Z}}$ is the source of randomness, which is assumed to be i.i.d. for different $t \in \mathbb{Z}$ (see Remark 5.1.1) and such that $\mathbb{P}[h(x, \xi_t^x) = y] = p_{xy}$. Here h is an update rule which allows one to construct the random variable at time $t + 1$ from that at time t. When $\{X_t\}_{t\in\mathbb{Z}}$ is positive recurrent, Equation (1.0.1) has a stationary solution. This means there is a random variable X, with distribution σ, and such that $X \overset{d}{=} h(X, \xi_t^X)$, with $\overset{d}{=}$ meaning equality in distribution. In the null recurrent case, this dynamic does not admit such a stationary solution.

The Doeblin Graph of $\{X_t\}$ is a random graph with vertices in $\mathbb{Z} \times \mathscr{S}$. In this graph, the horizontal axis ($\mathbb{Z}$-axis) represents time and is referred to as the *time axis* and the vertical axis ($\mathscr{S}$-axis), which is referred to as the *state axis*, represents the state space. The edges of the Doeblin Graph are defined using the transition probabilities of the Markov Chain: there is an edge from each vertex (t, x) to vertex $(t + 1, h(x, \xi_t^x))$, with $\{\xi_t^x\}_{x,t}$ as defined above. The Bridge Graph with respect to s^* is the union of all paths of the Doeblin Graph starting from $\mathbb{Z} \times \{s^*\}$, where s^* is an arbitrary fixed point in $\mathscr{S}$.
It is shown in [4] that for an irreducible, aperiodic, and positive recurrent Markov Chain with a totally independent source of randomness (see Remark 5.1.1), the Bridge Graph is almost surely a tree. This tree is both locally finite and unimodularizable, as defined in Section 3.3 and discussed in Subsection 4.3. Using the unimodular property, it is shown that, in this case, there is a unique bi-infinite path in the this graph and that the points of this bi-infinite path form a stationary sequence with marginal distribution the stationary distribution.

The first aim of the Part I of this thesis is the study of the properties of the Bridge Graph constructed from an irreducible, aperiodic and *null* recurrent Markov Chain. In this case, one can show that the Bridge Graph is not unimodularizable in general, that it can be both connected (it is then a tree) or not connected (it is then a forest made of an infinite number of trees), and that it contains no bi-infinite path when it is connected or when it satisfies some additional condition given in Section 5.1. It is also shown that the Bridge Graph is nevertheless locally unimodular. More precisely, it contains a unimodular and one ended random tree, the *Recurrence Time EFF*. This allows one to show that the Bridge Graph is one ended as well (See Section 5.3 for more details).

In the recurrent case (both positive and null), two measure-value dynamics are defined on the Bridge Graph, the Taboo and the Potential dynamics. These dynamics are measure-valued and of the form

$$M_{t+1} = H(M_t, \xi_t), \tag{1.0.2}$$

where for each t, M_t is a random measure on $\mathscr{S}$ and $\xi_t = \{\xi_t^x\}_{x\in\mathscr{S}}$ is the same as in Equation (1.0.1). The Taboo dynamics and the Potential dynamics are each defined by a specific update rule H defined in Section 6.1. It is shown that the Taboo dynamics has a stationary state on the space of random measures on $\mathscr{S}$, called the Taboo Point Process (TPP). A key point is that the mean measure of the TPP is equal to the invariant measure of the Markov Chain. The Potential Dynamics is also studied. In the null recurrent case, this dynamics also has a stationary state on the space of random measures on $\mathscr{S}$. This random measure is called the Potential Point Process (Potential PP). This point process is locally finite, but its mean measure is infinite. These two point processes can also be defined in the positive recurrent case as well, and their properties are also discussed in this case. They are hence fundamental objects in that they can be defined for all recurrent discrete time discrete space Markov Chains, they are left invariant by the Markov dynamics, and they provide, as it will be shown, important informations on the CFTP algorithm as well as complementary informations on the two most important deterministic measures of Markov Chain theory, namely the invariant measure and the potential measure.

After studying the existence of stationary regimes for these dynamics, their uniqueness is also discussed. For this purpose, these dynamics are considered as Markov Chains on the space of locally finite counting measures on $\mathscr{S}$, $\mathscr{N}(\mathscr{S})$. It is shown that the dynamical systems introduced above may have other stationary solutions than the one built on the Bridge Graph of the MC.

II. Clustering on Point Processes

Clustering is a popular method in unsupervised learning that plays an important role in data science as an alternative to supervised learning, which requires no training mechanisms. Given a representation of the objects, clustering aims to identify groups, or clusters, based on some similarity measure between them. Usually, the objects are represented as a finite number of points in $\mathbb{R}^d$, or equivalently, as a point process.
Clustering algorithms can be organized into two broad categories: point assignment clustering and hierarchical clustering. In point assignment algorithms, some clusters are predefined, points are considered in some order, and each one is assigned to the cluster into which it best fits. An important parameter here is the number of clusters. The best-known point assignment algorithm is the k-means algorithm (see [5] for definition and details). Hierarchical algorithms start with each point forming its own cluster. Clusters are then connected/merged based on their similarities and form new clusters. The process of connecting and merging clusters is repeated until reaching a single cluster to which all the points belong. A key component of hierarchical clustering is the function used to measure distances between clusters. This

includes minimum distance (single linkage), Hausdorff distance (complete linkage), and centroid distances (see [6], [7], and Subsection 3.4 for more details).

The aim of part II is to study the properties of a natural clustering algorithm that essentially belongs to the hierarchical class but also shares some features of point assignment. An important property of this algorithm, which is called *Hierarchical Thinning Nearest Neighbor Clustering (HTNNC)*, is that it allows one to handle countably infinite data sets, specifically point processes of the Euclidean space.

Here, a concise overview of the HTNNC algorithm is presented. In the algorithm, each cluster connects to its nearest neighbor cluster at each hierarchical step. The distance considered between clusters is known as centroid (clustroid) [2] distance. In each cluster, one or more representative points, referred to as centroids or clustroids, are selected to collectively represent the entire cluster. The distance between clusters is calculated based on the distances between their respective centroids (clustroids). The key idea of the HTNNC algorithm lies in the selection of clustroids. The clustroid of a cluster is chosen as the pair that achieves the minimum distance in the cluster.

Let us first describe the hierarchical nature of the algorithm. This is illustrated in the case where the pseudo distance between two clustroids is the minimal distance between the points of the two clustroids.

At level zero, each point connects to its nearest neighbor point. Under natural assumptions, this defines a collection of finite connected components called *clusters of order 0* on the data set. Each cluster is made up of directed trees, called *sub-cluster trees of order 0*, that are connected towards a cycle containing a pair of mutually nearest neighbor points (MNN). Call each such pair a *cycle of order 0* and the points belongs to a cycle of order 0 *cluster heads of order 0*.
Consider now all such order 0-cycles. Connect each 0-cycle to the nearest 0-cycle (nearest for the above pseudo distance). This defines a new collection of directed trees of 0-cycles, where again, each tree is connected to a cycle containing a pair of 0-cycles which are MNN for this pseudo distance. Define the exit point of a 0-cycle to be the point of the cycle which is the closest to the nearest 0-cycle. For each such order 0 exit point, replace the initial directed connection to the other point of its own cycle with a directed connection to the closest point of the nearest 0-cycle. This defines a directed graph on cluster heads of order 0, with again finite connected components. Each connected component consists of two directed trees that are connected towards a cycle containing a pair of 0-cluster heads, which are mutual nearest neighbors (MNN) for the pseudo distance between 0-cycle. The union of this directed

[2] When the representative point of a cluster is not precisely located at the center of the cluster's data, the term *clustroid* is employed in place of *centroid*

graph and of the sub-cluster trees of order 0 defines the clusters of order 1 on the data set. Each such cluster contains a collection of directed trees called sub-cluster trees of order 1, where each tree is connected towards a cycle of length 2 connecting two cluster heads of order 0. The last pair is called a cycle of order 1. The above union consists of hanging each order 0 sub-cluster tree to an appropriate order 0 cluster-head, which will be called *gateway point.*

By iterating this procedure, one hierarchically defines cluster-heads of order 1 or more, as well as clusters of order 2 or more. The end result is a tree (or a forest) where each cluster-head of order k is the gateway point of a descendant sub-cluster tree of order k, made of cluster heads of order lower than k and of regular points (a point is regular if it is not a cluster-head of any order).

Let us illustrate the basic idea by a species classification pseudo-example. Assume that the classification algorithm bears on genetic data of species and first finds the cluster of horse-related species that includes plough horses, race horses, ponneys, donkeys, etc., and an archetype (cluster heads) for this horses cluster defined as the MNN pair in this order-0 cluster. Other order 0-clusters could e.g. that of sheep, with again lots of variants and a 0-cluster head pair which is an archetype for sheep, or that of cows with yet another MNN pair, etc. Rather than looking at the proximity of say the horse cluster as a whole and the sheep cluster as a whole in terms of the smallest distance from the two sets of species, which is what certain algorithms would do, HTNN looks at the distances between the archetype of horses, that of sheep, that of cows, etc, which are the corresponding cycles of order 0. Based on these 0-cycles, it then determines a cluster head pair of order 1, which could be, for example, mammals. Then it iterates to find, for instance, vertebrates, and so on.

Here are now a few observations on HTNN. It is defined as a hierarchical algorithm. However, one can also say that it is point assignment based on the fact that cluster-heads of a given order are in a sense predefined through a mechanism that does not require the whole data set but a sufficient data set which is linked to the data set in the tree for which it is a gateway.

For all k, the cluster heads (or archetypes) of level k are a deterministic functional of the data set. Their definition involves no extra randomness, as in, for example, the techniques based on the thinning of point processes described in more detail in Section 8.3.

It is applicable to infinite point processes. In addition, using the pseudo-distance between pairs leads to a limited computational effort, well adapted to large data sets. This feature is distinctive of other known distance based hierarchical clustering algorithms, e.g., the nearest-neighbor chain described in Section 3.4.

In addition, it supports *parallelization*, in that one can start building the trees in two

separated regions in parallel, enabling faster processing and scalability.

As mentioned before, the aim of this part goes beyond the introduction of a new hierarchical classification algorithm. It is also to analyze this algorithm acting on point processes. Here is a summary of the analytical results:

This algorithm is first considered on the Poisson point process (PPP). It is shown that the algorithm can be defined as a sequence of point shifts on the PPP, each point shift representing a step of the algorithm (see [8] and Section 8.1 for more details). As a result, at each step, n, the point shift gives a random directed graph on the PPP, which is a forest. The resulting random forests are unimodular graphs (see [2] and Section 3.3). Each step contains infinitely many finite clusters (trees).

It is shown that when n goes toward infinity, the weak limit point shift, which will be called *Recursive Thinning Nearest Neighbor Point Shift (RTNN PS)*, exists. The limiting (weak limit) graph (the limiting point shift graph) is unimodular and is called the *Recursive Thinning Nearest Neighbor Forest /Tree (RTNNF/T)*. The RNNF/T gives an infinite spanning forest or tree on the PPP. It will be shown that the infinite random tree containing the origin is one-ended on the PPP. This result holds for any sub point process of the PPP or any point process that contains no second-order descending chain a.s. (see Sections 8.2 and 9.1 for more details).

While these analytical results provide valuable insights that will be discussed, several open questions and conjectures remain regarding the RNNPS and RTNNF/T. One such question revolves around whether the RTNNF/T is a tree when constructed on the PPP. The conjecture is that, in all dimensions, this point shift graph is a tree on the PPP. Another question is whether the point map probability of this point shift exists on the PPP, and how to compute the point map probability if it exists. Consider a point shift f on a PP and the action of the semigroup of translations by $-f$ on probability distributions of counting measures that have a point at the origin. The f-probabilities of the point process are then defined as the limits of the orbit of this semigroup action on the Palm distribution of the PP. Answering these questions and validating the conjecture would greatly contribute to our understanding of the algorithm's behavior.

Furthermore, in the RTNNF/T, one can define the following notion of distance for two given points (species). It is defined as the smallest k such that they both belong to the same k cluster. Additionally, evaluating the mean number of points in the descendant (phylogenetic) tree of a typical point within the RNNF/T on the PPP provides valuable insights into the clustering structure and characteristics of the algorithm.

Chapter

2 Introduction en Français

Cette thèse est composée de deux travaux sur les réseaux aléatoires. La première partie est basée sur une prépublication disponible sur [1].
Un réseau aléatoire enraciné est considéré comme unimodulaire lorsque, d'un point de vue heuristique, tout sommet a une probabilité égale d'être la racine. Pour définir rigoureusement ce concept pour les réseaux ayant un nombre infini de sommets, on suppose que le réseau aléatoires enraciné satisfait le principe de transport de masse. La littérature a étudié les propriétés des réseaux unimodulaires et des dynamiques covariantes sur ces réseaux ([2] et [3]).
Cette thèse se concentre sur deux graphes aléatoires spécifiques qui sont associés à des algorithmes stochastiques et elle étudie les propriétés de ces graphes en utilisant la notion d'unimodularité. Le premier algorithme étudié est l'algorithme de couplage par le passé, qui est utilisé pour l'échantillonnage parfait de la distribution stationnaire d'une chaîne de Markov. Le second graphe aléatoire se rapporte à un modèle utilisé pour la classification hiérarchique des données.
La thèse est structurée en deux parties distinctes. Un préliminaire, présenté dans le chapitre 3, fournit une base commune aux deux parties. Les résultats de chaque partie sont indépendants et peuvent être lus séparément. Cette brève introduction donne un aperçu des principaux thèmes explorés dans chaque partie.

I. Graphe de Doeblin

Soit $\{X_t\}_{t\in\mathbb{N}}$ une chaîne de Markov avec un espace d'état dénombrable, $\mathscr{S}$. Il est bien connu que lorsque $\{X_t\}_{t\in\mathbb{N}}$ est irréductible, apériodique et récurrente positive, elle possède une distribution stationnaire unique. En revanche, lorsqu'elle est récurrente nulle, elle n'admet pas de distribution de probabilité stationnaire, mais elle a une mesure stationnaire unique σ qui satisfait $\sigma = \sigma P$ avec $\sigma(s^*) = 1$, où P est la matrice de probabilité de transition de la chaîne de Markov, et s^* est un point fixe arbitraire dans $\mathscr{S}$.

On peut considérer une chaîne de Markov comme une dynamique sur des variables aléatoires à valeurs $\mathscr{S}$. Cette dynamique peut être écrite sous la forme de l'équation

suivante [1]

$$X_{t+1} = \sum_{x \in \mathscr{S}} \mathbb{1}_{\{X_t = x\}} h(x, \xi_t^x), \tag{2.0.1}$$

où $\{\xi_t^x, x \in \mathscr{S}\}_{t\in\mathbb{Z}}$ est la source d'aléatoire, qui est supposée être i.i.d. pour différents $t \in \mathbb{Z}$ (voir la remarque 5.1.1) et telle que $\mathbb{P}[h(x, \xi_t^x) = y] = p_{xy}$. Ici h est une règle de mise à jour qui permet de construire la variable aléatoire au temps $t+1$ à partir de celle au temps t. Lorsque $\{X_t\}_{t\in\mathbb{Z}}$ est récurrente positive, l'équation (2.0.1) a une solution stationnaire. Cela signifie qu'il existe une variable aléatoire X, avec une distribution σ, et telle que $X \overset{d}{=} h(X, \xi_t^X)$, avec $\overset{d}{=}$ signifiant l'égalité en distribution. Dans le cas de la récurrence nulle, cette dynamique n'admet pas de solution stationnaire.

Le graphe de Doeblin de $\{X_t\}$ est un graphe aléatoire dont les sommets se trouvent dans $\mathbb{Z} \times \mathscr{S}$. Dans ce graphe, l'axe horizontal (axe $\mathbb{Z}$) représente le temps et est appelé *axe du temps* et l'axe vertical (axe $\mathscr{S}$), qui est appelé *axe d'état*, représente l'espace d'état. Les arêtes du graphe de Doeblin sont définies à l'aide des probabilités de transition de la chaîne de Markov : il existe une arête entre chaque sommet (t, x) et le sommet $(t+1, h(x, \xi_t^x))$, avec $\{\xi_t^x\}_{x,t}$ comme défini ci-dessus. Le graphe des ponts par rapport à s^* est l'union de tous les chemins du graphe de Doeblin à partir de $\mathbb{Z} \times \{s^*\}$, où s^* est un point fixe arbitraire dans $\mathscr{S}$.
Il est démontré dans [4] que, pour une chaîne de Markov irréductible, apériodique et récurrente positive avec une source d'aléatoire complètement indépendante (voir la remarque 5.1.1), le graphe des ponts est presque sûrement un arbre. Cet arbre est à la fois localement fini et unimodularisable, comme défini dans la Section 3.3 et discuté dans la Sous-section 4.3. En utilisant la propriété unimodulaire, on montre que, dans ce cas, il existe un unique chemin bi-infini dans ce graphe et que les points de ce chemin bi-infini forment une séquence stationnaire dont la distribution marginale est la distribution stationnaire.

Le premier objectif de la partie I de cette thèse est l'étude des propriétés du graphe des ponts de Doeblin construit à partir d'une chaîne de Markov irréductible, apériodique et récurrente *nulle*. Dans ce cas, on peut montrer que le graphe des ponts n'est pas unimodularisable en général, qu'il peut être à la fois connecté (c'est alors un arbre) ou non connecté (il s'agit alors d'une forêt composée d'un nombre infini d'arbres), et qu'il ne contient aucun chemin bi-infini lorsqu'il est connecté ou lorsqu'il satisfait à une condition supplémentaire donnée dans la Section 5.1. On montre également que le graphe des ponts est localement unimodulaire. Plus précisément, il contient un arbre aléatoire unimodulaire et une seule extrémité, le *Recurrence Time EFF*. Cela

[1] Une autre représentation de l'équation (2.0.1) est l'équation de récurrence stochastique $X_{t+1} = h(X_t, \xi_t)$.

permet de montrer que le graphe des ponts contient également une seule extrémité (voir la Section 5.3 pour plus de détails).

Dans le cas récurrent (à la fois positif et nul), deux dynamiques à valeurs mesures sont définies sur le graphe des ponts, la dynamique tabou et la dynamique potentielle. Ces dynamiques à valeurs mesures sont de la forme

$$M_{t+1} = H(M_t, \xi_t), \tag{2.0.2}$$

où pour chaque t, M_t est une mesure aléatoire sur $\mathscr{S}$ et $\xi_t = \{\xi_t^x\}_{x\in\mathscr{S}}$ est identique à ce qui est définie pour l'équation (2.0.1). La dynamique tabou et la dynamique potentielle sont chacune définies par une règle de mise à jour H donnée dans la Section 6.1. On montre que la dynamique tabou a un état stationnaire, dans l'espace des mesures aléatoires sur $\mathscr{S}$, appelé le processus ponctuel tabou. Un point clé est que la mesure moyenne du processus du point tabou est égale à la mesure invariante de la chaîne de Markov. La dynamique potentielle est également étudiée. Dans le cas récurrent nul, cette dynamique possède aussi un état stationnaire dans l'espace des mesures aléatoires sur $\mathscr{S}$. Cette mesure aléatoire est appelée processus ponctuel potentiel. Ce processus ponctuel est localement fini, mais sa mesure moyenne est infinie. Ces deux processus ponctuels peuvent également être définis dans le cas récurrent positif et leurs propriétés sont également discutées dans ce cas. Ce sont donc des objets fondamentaux en ce sens qu'ils peuvent être définis pour toutes les chaînes de Markov récurrentes à temps discret et à espace discret et invariant par la dynamique de Markov. Comme on le verra, ils apportent des informations importantes sur l'algorithme CFTP ainsi que des informations complémentaires sur les deux mesures déterministes les plus importantes de la théorie des chaînes de Markov, à savoir la mesure invariante et la mesure potentielle.

Après avoir étudié l'existence de régimes stationnaires pour ces dynamiques, leur unicité est également discutée. Pour cela, ces dynamiques sont considérées comme des chaînes de Markov sur l'espace des mesures de comptage localement finies sur $\mathscr{S}$, $\mathscr{N}(\mathscr{S})$. On montre que les systèmes dynamiques introduits ci-dessus peuvent avoir d'autres solutions stationnaires que celle construite sur le graphe des ponts de la chaîne de Markov.

II. Regroupement sur des Processus Ponctuels

Le regroupement est une méthode populaire d'apprentissage non supervisé qui joue un rôle important dans la science des données en tant qu'alternative à l'apprentissage supervisé et qui ne nécessite pas de mécanismes de formation. Étant donné une représentation des objets, le regroupement vise à identifier des groupes, ou grappes,

sur la base d'une certaine mesure de similarité entre eux. En général, les objets sont représentés par un nombre fini de points dans $\mathbb{R}^d$.
Les algorithmes de regroupement peuvent être classés en deux grandes catégories : le regroupement par affectation de points et le regroupement hiérarchique. Dans les algorithmes d'affectation de points, certaines grappes sont prédéfinies, les points sont considérés dans un certain ordre, et chacun d'entre eux est affecté à la grappe qui lui convient le mieux. Le nombre de grappes est un paramètre important. L'algorithme d'affectation de points le plus connu est l'algorithme des k-moyennes (voir [5] pour la définition et les détails). Dans les algorithmes hiérarchiques, chaque point forme d'abord sa propre grappe. Les grappes sont ensuite connectées/fusionnées sur la base de leurs similitudes et forment de nouvelles grappes. Le processus de connexion et de fusion des grappes est répété jusqu'à l'obtention d'une grappe unique à laquelle tous les points appartiennent. La fonction utilisée pour mesurer les distances entre les grappes est un élément clé de la classification hiérarchique. Il s'agit notamment de la distance minimale (lien simple), de la distance de Hausdorff (lien complet), et les distances entre centroïdes (voir [6], [7] et la sous-section 3.4 pour plus de détails).

L'objectif de la partie II est d'étudier les propriétés d'un algorithme de regroupement naturel qui appartient essentiellement à la classe hiérarchique mais partage également certaines caractéristiques de l'affectation de points. Une propriété importante de cet algorithme, appelé *Hierarchical Thinning Nearest Neighbor Clustering (HTNNC)*, est qu'il permet de traiter des ensembles de données dénombrables, en particulier des processus ponctuels de l'espace euclidien.

Nous présentons ici un aperçu concis de l'algorithme HTNNC. Dans l'algorithme HTNNC, chaque grappe se connecte à sa grappe voisine la plus proche à chaque étape hiérarchique. La distance considérée entre les grappes est appelée centroïde (clustroïde)[2]. Au sein de chaque grappe, un ou plusieurs points représentatifs appelés centroïdes (clustroïdes) sont sélectionnés pour représenter l'ensemble de la grappe. La distance entre les grappes est calculée sur la base des distances entre leurs centroïdes (clustroïdes) respectifs. L'idée clé de l'algorithme HTNNC réside dans la sélection des clustroïdes. Le clustroïde d'une grappe est choisi comme la paire qui realise la distance minimale dans la grappe.

Décrivons d'abord la nature hiérarchique de l'algorithme. Ceci est illustré dans le cas où la pseudo-distance entre deux clustroïdes est la distance minimale entre les points des deux clustroïdes.

Au niveau zéro, chaque point est relié au point voisin le plus proche. Sous des hy-

[2]Lorsque le point représentatif d'une grappe n'est pas précisément situé au centre des données de la grappe, le terme *clustroïde* est employé à la place de la distance *centroïde.*

pothèses naturelles, ceci définit une collection de composantes connectées finies appelées *grappes d'ordre 0* sur l'ensemble des données. Chaque grappe est constituée d'arbres dirigés, appelés *arbres de sous-grappes d'ordre 0*, qui sont connectés à un cycle contenant une paire de points mutuellement les plus proches (MNN). Chacune de ces paires est appelée *cycle d'ordre 0* et les points appartiennent à un cycle d'ordre 0 de *têtes de grappes d'ordre 0.*

Considérez maintenant tous ces cycles d'ordre 0. Reliez chaque cycle d'ordre 0 au cycle d'ordre 0 le plus proche (le plus proche pour la pseudo-distance ci-dessus). Ceci définit une nouvelle collection d'arbres dirigés de cycles d'ordre 0, où chaque arbre est connecté à un cycle contenant une paire de cycles d'ordre 0 qui sont MNN pour cette pseudo-distance. Définissez le point de sortie d'un cycle d'ordre 0 comme étant le point du cycle qui est le plus proche du cycle d'ordre 0 le plus proche. Pour chaque point de sortie d'ordre 0, remplacez la connexion dirigée initiale vers l'autre point de son propre cycle par une connexion dirigée vers le point le plus proche du cycle d'ordre 0 le plus proche. Cesi définit un graphe orienté sur les têtes de grappe d'ordre 0, avec à nouveau des composantes connectées finies. Chaque composante connectée consiste en deux arbres dirigés qui sont connectés à un cycle contenant une paire de têtes de grappe d'ordre 0, qui sont les voisins mutuels les plus proches pour la pseudo-distance entre les cycles d'ordre 0. L'union de ce graphe orienté et des arbres de sous-groupes d'ordre 0 définit les grappes d'ordre 1 sur l'ensemble des données. Chacun de ces grappes contient une collection d'arbres dirigés appelés arbres de sous-groupes d'ordre 1, où chaque arbre est connecté à un cycle de longueur 2 reliant deux têtes de grappe d'ordre 0. La dernière paire est appelée un cycle d'ordre 1. L'union ci-dessus consiste à accrocher chaque arbre de sous-grappe d'ordre 0 à une tête de grappe d'ordre 0 appropriée, qui sera appelée *point de passage*.

En itérant cette procédure, on définit hiérarchiquement des têtes de grappes d'ordre 1 ou plus, ainsi que des grappes d'ordre 2 ou plus. Le résultat final est un arbre (ou une forêt) où chaque tête de grappe d'ordre k est le point d'entrée d'un sous-arbre de grappe descendant d'ordre k, composé de têtes de grappe d'ordre inférieur à k et de points réguliers (un point est régulier s'il n'est pas une tête de grappe d'un ordre quelconque).

Illustrons l'idée de base par un pseudo-exemple de classification des espèces. Supposons que l'algorithme de classification s'appuie sur les données génétiques des espèces et trouve d'abord le groupe d'espèces liées aux chevaux, qui comprend les chevaux de labour, les chevaux de course, les poneys, les ânes, etc. et un archétype (têtes de grappe) pour ce groupe de chevaux défini comme la paire les voisins mutuels les plus proches dans ce groupe d'ordre 0. D'autres grappes d'ordre 0 pour-

raient par exemple être celles des moutons, avec là encore de nombreuses variantes et une paire de têtes de grappe 0 qui est un archétype pour les moutons, ou celle des vaches avec encore une autre paire, les voisins mutuels les plus proches, etc. Plutôt que d'examiner la proximité de l'ensemble des chevaux et de l'ensemble des moutons en fonction de la plus petite distance entre les deux ensembles d'espèces, comme le feraient certains algorithmes, HTNNC examine les distances entre l'archétype des chevaux, celui des moutons, celui des vaches, etc. sur la base de ces cycles d'ordre 0, il détermine ensuite une paire de têtes de groupe d'ordre 1, qui pourrait être celle des mammifères, par exemple. Il faut ensuite itérer pour trouver, par exemple, les vertébrés, etc.

Voici maintenant quelques observations sur le HTNNC. Il est défini comme un algorithme hiérarchique. Toutefois, on peut également dire qu'il s'agit d'une affectation de points basée sur le fait que les têtes de grappe d'un ordre donné sont en quelque sorte prédéfinies par un mécanisme qui ne nécessite pas l'ensemble des données, mais un ensemble de données suffisant qui est lié à l'ensemble des données de l'arbre pour lequel il constitue un point de passage.

Pour tout k, les têtes de grappe (ou archétypes) de niveau k sont une fonction déterministe de l'ensemble de données. Leur définition n'implique aucun élément aléatoire supplémentaire, comme c'est le cas, par exemple, pour les techniques basées sur l'amincissement (thinning) des processus ponctuels décrites plus en détail dans la Section 8.3.

Elle est applicable aux processus ponctuels infinis. En outre, l'utilisation de la pseudo-distance entre les paires demande un effort de calcul limité, bien adapté aux grands ensembles de données. Cette caractéristique distingue HTNNC d'autres algorithmes connus de regroupement hiérarchique basés sur la distance, par exemple la chaîne du plus proche voisin (nearest-neighbor chain) décrite dans la Section 3.4.

En outre, cet algorithme peut être parallélisé, en ce sens que l'on peut commencer à construire les arbres dans deux régions séparées en parallèle, ce qui permet un traitement plus rapide et une plus grande évolutivité. Il est *adaptatif*, en raison de la nature de sa définition du clustroïde au sein de chaque grappe. Dans les algorithmes de regroupement, l'adaptativité fait référence à la capacité d'ajuster dynamiquement les paramètres ou le comportement en fonction des caractéristiques des données : contrairement aux approches traditionnelles qui reposent sur des hypothèses concernant le nombre de grappes ou les paramètres de densité, il construit des grappes de manière autonome sans avoir besoin d'une initialisation explicite ou d'un ajustement des paramètres. Le centroïde, défini comme la paire atteignant la distance minimale au sein d'un grappe, s'adapte naturellement à la structure inhérente des données.

Comme mentionné précédemment, l'objectif de cette partie dépasse l'introduction d'un nouvel algorithme de classification hiérarchique. Il s'agit surtout d'analyser cet algorithme agissant sur des processus ponctuels. Voici un résumé des résultats de l'analyse.

Tout d'abord, cet algorithme est considéré sur le processus ponctuel de Poisson (PPP). On montre que l'algorithme peut être défini comme une séquence de déplacements de points (point shifts) sur le PPP, chaque déplacement de point représentant une étape de l'algorithme (voir [8] et la Section 8.1 pour plus de détails). À chaque étape, n, le déplacement de points donne un graphe dirigé aléatoire sur PPP, qui est une forêt. Les forêts aléatoires résultantes sont des graphes unimodulaires (voir [2] et la Section 3.3). On montrera que chaque étape contient une infinité de grappes finies (arbres).

Lorsque n tend vers l'infini, on montrera qu'il existe un déplacement de point limite (pour le convergence faible), qui sera appelé *Recursive Thinning Nearest Neighbor Point Shift (RTNN PS)*. Le graphe limite Pour la convergence locale faible (le graphe de déplacement du point limite) est unimodulaire et est appelé *Recursive Thinning Nearest Neighbor Forest /Tree (RTNNF/T)*. RNNF/T donne une forêt ou un arbre infini sur le PPP. Il sera démontré que l'arbre aléatoire infini contenant l'origine est à une extrémité sur le PPP. Ce résultat est valable pour tout sous-processus ponctuel du PPP ou tout processus ponctuel qui ne contient pas de chaîne descendante du second ordre a.s. (voir les Sections 8.2 et 9.1 pour plus de détails).

Bien que ces résultats analytiques fournissent des informations précieuses qui seront discutées, plusieurs questions et conjectures restent ouvertes concernant le RNNPS et le RTNNF/T. L'une de ces questions consiste à savoir si le RTNNF/T est un arbre lorsqu'il est construit sur le PPP. La conjecture est que, dans toutes les dimensions, ce graphe de déplacement de points est un arbre sur le PPP. L'autre question est de savoir si la probabilité de déplacement au point (point map probability) de ce déplacement de point existe sur le PPP, et de calculer cette probabilité de déplacement au point si elle existe. Considérons un déplacement de point f sur un PP et l'action du semigroupe des translations par $-f$ sur les distributions de probabilité des mesures de comptage qui ont un point à l'origine. Les f probabilités du processus ponctuel sont alors définies comme les limites de l'orbite de l'action de ce semigroupe sur la distribution de Palm du PP. Répondre à ces questions et valider la conjecture contribuerait grandement à notre compréhension du comportement de l'algorithme.

En outre, dans RTNNF/T, on peut définir la notion suivante de distance pour deux points (espèces) donnés. Elle est définie comme le plus petit k tel qu'ils appartiennent à la même k grappe. En outre, l'évaluation du nombre moyen de points dans l'arbre descendant (phylogénétique) d'un point typique dans RNNF/T sur le PPP

fournit des indications précieuses sur la structure de regroupement et les caractéristiques de l'algorithme.

Chapter

3 Preliminaries

Chapter content

3.1 Point Process

3.1.1 Introduction

In this section, we introduce the concept of a point process, which will be used throughout the thesis. Let M be a locally compact, second countable, Hausdorff topological space, and let $\mathscr{B}$ be the Borel σ-algebra on M. Let N be the set of all locally finite counting measures on $(M, \mathscr{B})$, i.e., any $\phi \in N$ can be written as a finite or countably infinite sum $\sum_i \delta_{x_i}$ of Dirac measures located at some points $x_i \in M$. Let $\mathscr{N}$ be the σ-algebra on N generated by the mapping $\phi \to \phi(B)$, where $B \in \mathscr{B}$. This σ-algebra is the Borel σ-algebra with respect to the vague topology on the space of counting measures.

A *point process* Φ is a measurable mapping from some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to the measurable space $(N, \mathscr{N})$. The distribution of Φ is the probability measure P_Φ on $(N, \mathscr{N})$ obtained by taking the image of $\mathbb{P}$ under Φ. In this thesis, we only consider the case where $M = \mathbb{R}^d$.

The mean measure (or intensity measure) of Φ is a measure defined on $(M, \mathscr{B})$ as $\lambda(B) = \mathbb{E}(\Phi(B))$.

Two concepts from the theory of point processes that will be used in this thesis are the notion of stationarity and the Palm probability. The point process Φ (on $\mathbb{R}^d$) is called *stationary* if its finite-dimensional distributions are invariant under translations of $\mathbb{R}^d$.

The Palm probability is defined on the point process Φ as follows: Fix a Borel subset B of $\mathbb{R}^d$ with positive and finite Lebesgue measure $||B||$. For all $F \in \mathscr{N}$, define

$$P_0(F) = \frac{1}{\lambda(B)} \int \sum_{t \in \Phi \cap B} 1\{\theta_t \Phi \in F\} P(d\Phi), \tag{3.1.1}$$

where θ_t is a measurable flow in Ω. We assume the measurable space $(\Omega, \mathscr{F})$ is equipped with a measurable flow $\theta_t : \Omega \to \Omega$ for $t \in \mathbb{R}^d$, i.e., a family of mappings such that $(\omega, t) \mapsto \theta_t \omega$ is measurable. The flow satisfies θ_0 being the identity on Ω, and for any $s, t \in \mathbb{R}^d$, $\theta_s \circ \theta_t = \theta_{s+t}$. The interpretation of P_0 (or P_x) is the conditional distribution of point process given that fact that there is a point at the origin (at point $x \in \mathbb{R}^d$). The Palm probability needed in this thesis is contained in the refined Campbell theorem (see [9]).

Theorem 3.1.1 (Refined Campbell Theorem)**.** *Let Φ be a point process on $\mathbb{R}^d$ with mean measure λ and Palm distribution $P_x, x \in \mathbb{R}^d$. For all non-negative measurable function f on $\mathbb{R}^d \times N$ we have,*

$$\mathbb{E}\left[\int_{\mathbb{R}^d} f(x, \Phi)\Phi(dx)\right] = \int_{\mathbb{R}^d} \int_N f(x_0, \mu) P_x(d\mu)\lambda(dx). \tag{3.1.2}$$

A point process Φ is said to be a Poisson Point Process (PPP) of intensity Λ if $p(\phi(B) = n) = \frac{e^{\Lambda(B)}(\Lambda(B))^n}{n!}$ for all bounded $B \in \mathscr{B}$, and for disjoint $\{B_i\}_{i=1}^n$ (Borel, bounded sets of $\mathbb{R}^d$) random variables $(\Phi(B_1, ..., \Phi(B_n))$ are independent. About the Palm distribution of PPP, *Slyvniak's Theorem* says that a point process Φ with finite mean measure is Poisson iff for λ-almost $x \in \mathbb{R}^d$, $P_x = P_{\Phi + \delta_x}$.

3.1.2 Second order intensity measure of a Point Process

This subsection provides a review of the definition of the second order intensity measure of a point process and its relation to inhibition and clustering. The local version of the intensity measure is used to study the presence of local clustering or inhibition in a point process, based on perfect samples obtained from it. This will be discussed further in Section 6.2. For more details and result about the second order intensity measure, see [10].

Similar to the notion of covariance in a random variable, the second order intensity measure measures the dependence between pairs of points in a point process. For a point process Φ, the second order intensity measure can be written as follows:

$$\lambda_2(x, y) = \frac{\mathbb{E}[\Phi(dx)\Phi(dy)]}{|dx||dy|} \quad \text{as } |dx|, |dy| \to 0. \tag{3.1.3}$$

Here, $\lambda_2(x, y)$ measures the expected number of points in the infinitesimal areas dx and dy. Thus, the second order intensity measure provides information about the interaction between points in a point process, such as clustering or inhibition.

In the stationary and isotropic point processes, Ψ, Ripley's K-function is commonly used to study the dependence between pairs of points. The K-function is defined as follows:

$$K(r) = \lambda^{-1}\mathbb{E}[\text{Number of extra points within distance } r \text{ of a randomly chosen point}], \tag{3.1.4}$$

where The parameter r represents the distance between the two points, and λ is the intensity measure of the point process. Given ψ, a perfect sample of Ψ, with in a predefined study area, A, if the number of observed data points is N, one can estimate $K(r)$ with the following function,

$$K(r) = \frac{|A|}{N}\frac{1}{N-1}\sum_i \sum_{j\neq i} 1\{d(S_i, S_j) \leq r\} \quad S_i, S_j \in \psi. \tag{3.1.5}$$

Here, $d(S_i, S_j)$ is the distance between points S_i and S_j, and the sum is taken over all pairs of points in the sample ϕ. In the case of a Poisson point process in $\mathbb{R}^2$, the K-function simplifies to $K(r) = \pi r^2$, which is known as *complete spatial randomness.* A value of $K(r) > \pi r^2$ indicates *clustering*, while a value of $K(r) < \pi r^2$ indicates *inhibition* in the point process. (See also [10] and [11] for more details.)

In the case of a non-homogeneous point process, the local K-function $K_x(r)$ can be used to measure the dependence between points. The local K-function is defined as:

$$K_x(r) = \frac{\mathbb{E}\left[\Phi(x-r, x+r)|\Phi(x) = 1\right]}{\mathbb{E}\left[\Phi(x-r, x+r)\right]}. \tag{3.1.6}$$

Here, $K_x(r)$ measures the expected number of points in the annulus of radius r centered at point x, given that there is a point at x. The denominator is the expected number of points in the annulus without conditioning on the existence of a point at x. In the case of a Poisson point process, the local K-function is constant and equal to 1 for all values of r. If $K_x(r) > 1$ for some $x \in \mathbb{R}^d$, then the point x is considered to be part of a cluster in the point process. If $K_x(r) < 1$ for some, $x \in \mathbb{R}^d$ the point is considered as an inhibition point in the point process.

Without the assumption of stationarity and isotropicity on point process Φ the local K function can be estimated naturally if there exists a large enough number of sampling of Φ.

3.1.3 Descending Chain in Point Processes

This section provides a preliminary overview of Part II of the thesis. This section reviews the definition and the results that exist about the existence of a descending chain in station-

ary point processes. For seeing the details and more results, see [12], [13], and [14].

A descending chain in a point process is an infinite sequence of points in the process for which the distances between consequensive points form a descending sequence. More fomally, let Φ be a point process on $(\mathbb{R}^d, \rho)$, where ρ is a metric on $\mathbb{R}^d$, and let $\{x_n\}_{n\in\mathbb{N}}$ be an infinite sequence of points in Φ such that $\rho(x_{i-1}, x_i) > \rho(x_i, x_{i+1})$ for all $i \geq 1$. Then $\{x_n\}_{n\in\mathbb{N}}$ is a descending chain in Φ.

The existence of a descending chain is an important concept in point process theory. It is related to other problems like percolation (see [14] and [12]), and existence of perfect matching (see [13]) in stationary point processes.

In [12] it is shown that there exists no descending chain in Poisson Point Process in $\mathbb{R}^d$, for all d. In paper [14], Daley and Last show that for a much broader class of point processes the result holds. This class includes Poisson cluster processes, Cox processes, and Gibbs processes satisfying some exponential moment conditions. On the other hand, the non-existence of a descending chain does not hold for all stationary point processes (see example 3.4 of [14]).

Section 9.1 introduces another concept similar to the descending chain, called the "Second Order Descending Chain." Additionally, it will be demonstrated that the Poisson point process does not admit a second order descending chain. The proof of this fact is inspired by the proof of non-existence of the descending chain in a class of point processes that satisfy some factorial moment condition. Here the sketch of this proof is provided, borrowed from [14] .

For any $\phi \in N$ (defined in Subsection 3.1.1) and $n \in \mathbb{N}$, let $\phi^{(n)}$ denote the set of all $(x_1, \ldots, x_n) \in \phi^n$ with pairwise different entries. Identify $\phi^{(n)}$ with the measure

$$\phi^{(n)}(B) := \sum_{(x_1,\ldots,x_n)\in\phi^n} 1((x_1, \ldots, x_n) \in B), \quad (B \in \mathscr{B}((\mathbb{R}^d)^{(n)})).$$

Define the nth factorial moment measure of a point process Φ, on such a set B, by

$$\begin{aligned}\alpha^{(n)}(B) &:= \mathbb{E}\left[\sum_{(x_1,\ldots,x_n)\in\Phi^n} 1((x_1, \ldots, x_n) \in B)\right] \\ &= \mathbb{E}\left[\Phi(B_1)(\Phi(B)_1 - 1)\ldots(\Phi(B)_1 - n + 1)\right] \quad \text{if } B = B_1^{(n)} \text{ with } B_1 \in \mathscr{B}(\mathbb{R}^d).\end{aligned}$$

Assume Φ is a stationary point process satisfying

$$\alpha^{(n)}(dx_1, \ldots, dx_n) \leq c^n dx_1 \ldots dx_n \quad (n \in \mathbb{N} \text{ and } c > 0) \tag{3.1.7}$$

As an example, one can see that the PPP satisfies this condition with $\alpha^{(n)}(dx_1, \ldots, dx_n) = \lambda^n dx_1 \ldots dx_n$.

Using inequality (3.1.7), one can bound the probability of a stationary point process Φ being in the set $C_n(b)$, which consists of all measures containing a finite descending chain of points whose first point has norm at most b. Specifically, one can write

$$\begin{aligned} P(\Phi \in C_n(b)) &\leq \alpha^n \int \cdots \int 1(b \geq |x_1| \geq |x_2 - x_1| \geq \cdots \geq |x_n - x_{n-1}|) dx_1 \cdots dx_n \\ &= \frac{c^n}{n!}(b^d v_d)^n \\ &\to 0 \quad \text{as } n \to \infty, \end{aligned}$$

where v_d is the volume of the unit ball in $\mathbb{R}^d$. Consequently, we have $P(\Phi \in C(b)) = 0$, which implies that Φ almost surely contains no descending chain of points whose first point has norm at most b. Importantly, this argument does not rely on any independence properties of Φ.

3.2 Coupling

The concept of coupling of probability measures plays a central role in Part I of this thesis. In this section, we provide a brief overview of the definition and key results related to coupling. For a more detailed treatment of the topic, the interested reader is referred to [15], [16], and [17].

Consider two probability distributions, $\mathscr{P}_1$ and $\mathscr{P}_2$, on some spaces S_1 and S_2, respectively. A joint distribution, denoted by $\mathscr{P}_{1,2}$, on the product space $S_1 \times S_2$ is said to be a coupling of $\mathscr{P}_1$ and $\mathscr{P}_2$ if its marginal distributions are $\mathscr{P}_1$ and $\mathscr{P}_2$. In other words, a coupling of two probability distributions preserves the individual distributions of each distribution.

One common way to construct a coupling is by defining two random variables X_1 and X_2, distributed according to $\mathscr{P}_1$ and $\mathscr{P}_2$, respectively, on a common probability space. This allows us to define a joint distribution, denoted by $\mathscr{P}(X_1, X_2)$, that is a coupling of $\mathscr{P}_1$ and $\mathscr{P}_2$.

One of the couplings used in the thesis is the *monotone coupling* of probability measures. Let S be a set with partial order $\preceq$, and let ν and μ be two probability measures on S. We say that ν stochastically dominates μ, denoted by $\mu \preceq \nu$, if $\mathbb{E}\mu(f) \leq \mathbb{E}\nu(f)$ for all increasing $f : S \to \mathbb{R}$. When S is a subset of $\mathbb{R}$, this definition is equivalent to $\mu(t, \infty) \leq \nu(t, \infty)$ for all $t \in \mathbb{R}$. The *monotone coupling theorem*, also known as Strassen's theorem (1965), provides an important result about the coupling of two probability measures.

Theorem 3.2.1 (Strassen 1965)**.** *Suppose that S is a Polish space equipped with a partial order. Two probability measures μ and ν on S satisfy $\mu \leq \nu$ if and only if there exists an*

$S \times S$-valued random element (X, Y) such that X has distribution μ, Y has distribution ν, and $P(X \leq Y) = 1$.

Let S be a set with a countable partial order $\preceq$, and consider a Markov chain $\{Z_t\}_{t\geq 0}$ on S with transition probabilities P. We call this Markov chain a *monotone chain* if Pf is an increasing function whenever f is increasing. Here $Pf(x) = \mathbb{E}[f(X)]$, where X is a random variable with distribution $P(x, .)$. The definition of a monotone Markov chain is equivalent to the existence of a coupling (X, Y) of $P(x, .)$ and $P(y, .)$ satisfying $X \leq Y$ for every comparable $x, y \in S$.

In this thesis, the state spaces considered are $\mathbb{Z}$ or $\mathbb{N}$. In this case where the Markov chain is monotone, there exists a *complete monotone coupling* for the Markov chain, which is a coupling $(X_i)_{i\in\mathbb{Z}}$ between all the states of the state space of the Markov chain, i.e., $\mathbb{Z}$, such that for all $j < k$, $X_j \leq X_k$ where $X_i \sim P(i, .)$ for all $i \in \mathbb{Z}$. The coupling $(X_i)_{i\in\mathbb{Z}}$ can be defined as follows:

$$\phi_i(u) := \inf_t\{F_i(t) \geq u\}, \tag{3.2.1}$$

where $F_i(t) = P(i, (-\infty, t))$. Let $X_i = \phi_i(U)$, where U is uniform on $(0, 1)$. Then $(X_i)_{i\in\mathbb{Z}}$ is a monotone coupling.

3.3 Unimodular Networks

The theory of unimodular networks is central to both parts I and II of the thesis. This chapter presents the primary definitions and theorems that are relevant to multiple chapters. For a more detailed explanation of random networks and unimodularity, see [2] and [3].

3.3.1 Introduction

A *network* is a graph $G = (V, E)$ equipped with a complete separable metric space Ξ called the *mark space*, and two maps from v, and $\{(v, e), v \in V, e \in E, v \sim e\}$ to Ξ, where $\sim$ denotes the adjacent vertices or edges. The image of v (resp. (v, e)) in Ξ is called its *mark*. Note that the graphs and directed graphs are special cases of networks. Unless otherwise stated, networks are considered to be connected and locally finite, indicating that the degree of each vertex is finite.

An *isomorphism* between two networks is a graph isomorphism that preserves the marks. A rooted network is represented by a pair (G, o) where G is a network and o is a distinguished vertex of G, called *root*. Similarly, a doubly rooted network (G, o, v) is a network with a pair of distinguished vertices. Let $\mathscr{G}$ denote the set of isomorphism class of connected and locally finite networks, and $\mathscr{G}^*$ (resp. $\mathscr{G}^{**}$) be the set of isomorphism class of (resp. doubly) rooted networks. The isomorphism class of a network G (rep. (G, o) or

(G, o, v)) is denoted by $[G]$ (resp. $[G, o]$ or $[G, o, v]$). The sets $\mathscr{G}^*$, and $\mathscr{G}^{**}$ can be equipped with a metric and its Borel sigma field (see [2]). The distance between (G, o) and (G', o') is $2^{-\alpha}$, where α is the supremum of those r such that there is a rooted isomorphism between $N_r(G, o)$ and $N_r(G', o')$ such that the distance of the marks of the corresponding elements is at most $\frac{1}{r}$. The distance on $\mathscr{G}^{**}$ is defined similarly. So $\mathscr{G}^*$ and $\mathscr{G}^{**}$ are complete separable non compact metric spaces. Borel measurable subsets of $\mathscr{G}^*$ and $\mathscr{G}^{**}$ and Borel measurable functions on $\mathscr{G}^*$ and $\mathscr{G}^{**}$ are roughly speaking those which can be determined by looking only at the finite neighborhood of the root or roots. For example, the degree of the root or roots are Borel measurable functions. A *random network* is a random element in $\mathscr{G}^*$, that is, a measurable function from some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $\mathscr{G}^*$. This function will be denoted by $[\mathbf{G}, \mathbf{o}]$ with bold symbols.

3.3.2 Unimodular Networks

A random network $[\mathbf{G}, \mathbf{o}]$ is unimodular if for all measurable functions $g : \mathscr{G}^{**} \to \mathbb{R}^{\geq 0}$, the following mass transport principle holds,

$$\mathbb{E}[\sum_{v \in V(\mathbf{G})} g[\mathbf{G}, o, v]] = \mathbb{E}[\sum_{v \in V(\mathbf{G})} g[\mathbf{G}, v, o]]. \tag{3.3.1}$$

A probability measure on $\mathscr{G}^*$ is referred to as unimodular if, when considered as a random network, it results in a unimodular network.

A simple example of a unimodular network is when the graph G is deterministic, connected, and finite. In this case, $[G, o]$ is unimodular if and only if o is chosen uniformly from $V(G)$ (see [2]). Another example of a unimodular graph is $(\mathbb{Z}, 0)$, where $\mathbb{Z}$ is the integer graph with vertices corresponding to integers and edges connecting adjacent vertices. If a unimodular graph is enriched with i.i.d. marks on its vertices or edges, the resulting network is still unimodular (Lemma 4 in [2]). Therefore, the i.i.d. mark of $\mathbb{Z}$ rooted at 0 is unimodular. This example will be used later.

A lemma that will be used in the later chapters is the so called *No Infinite/Finite Inclusion lemma* from [3] paper. Before discussing this lemma, we need to introduce some definitions.

A *covariant subset* of the set of vertices is a map S that associates to each network G a set $S(G) \subseteq V(G)$, which is covariant under network isomorphisms, and such that the function $[G, o] \to 1_{o \in S_G}$ is measurable. A map Π, which associates a partition Π_G of $V(G)$ to each network G, is a *covariant vertex partition* if the partition Π_G is covariant under network isomorphisms, and the subset $\{[G, o, s] : s \in \Pi_G(o)\}$ is well-defined and measurable, where $\Pi_G(o)$ is the element of Π_G that contains o. The lemma is as follows:

Lemma 3.3.1 (No Infinite/Finite Inclusion, Lemma 2.11 in [3])**.** *Let* $[\mathbf{G}, \mathbf{o}]$ *be a unimodular*

network, Π *a covariant partition, and* S *a covariant subset. Almost surely, there is no infinite element* E *of* Π_G *such that* $E \cap S_G$ *is finite and non-empty.*

The important theorem that will be used in both part of the thesis is the *Foil Classification in Unimodular Networks* from [3] paper. To state it the following definitions are needed. A *covariant vertex-shift* (vertex shift) is a map f which associates to each network G a function $f_G : V(G) \to V(G)$ which is covariant under isomorphisms and the well-defined function $[G, o, s] \to 1_{\{f_G(o)=s\}}$ on $\mathscr{G}^{**}$ is measurable.

Given a network G and an arbitrary vertex-shift f, let $\sim_f$ be the equivalence relation on $V(G)$ such that $\sim_f$ if and only if there exists an integer n such that $f^n(x) = f^n(y)$. Each equivalence class of $\sim_f$ is called a *foil*, and the partition of $V(G)$ generated by the foils is called the f-foliation of G. The f-graph of G, denoted by G^f, is a directed graph with vertices $V(G)$ and directed edges $(x, f(x))$ for each vertex x in $V(G)$. Directed paths and directed cycles in G^f are called f-paths and f-cycles, respectively. Considering an acyclic connected component in G^f, the foils can be given a total order such that each foil L is older than $f^{-n}(L)$ for any $n > 0$. A connected component of G^f may have a youngest foil, which is a foil L such that $f^{-1}(L) = \emptyset$. For $x \in V(G)$ and $n \geq 0$, let $D_n(x) := f^{-n}(x) = \{y \in V(G) : f^n(y) = x\}$, and let $d_n(x) := \#D_n(x)$. Similarly, define $D(x) := \cup_{n=1}^{\infty} D_n(x)$. The sequence of sets $f^n(V(G))$ is decreasing in n. Its limit is denoted by $f^\infty(V(G))$. For a connected component C of G^f, define $f^\infty(C)$ as the set of vertices $x \in C$ such that $D_n(x)$ is non-empty for all $n \geq 0$. The following is the classification theorem from [3]. Foil Classification in Unimodular Networks, Theorem 3.10 in [3]] Let $[\mathbf{G}, \mathbf{o}]$ be a unimodular network and f be a vertex-shift. Almost surely, every vertex has finite degree in the graph G^f . In addition, each component C of G^f has at most two ends, and it belongs to one of the following three classes:

- Class $\mathscr{F}/\mathscr{F}$: C and all its foils are finite. If $n = n(C)$ is the number of foils in C, then
  - C has a unique f-cycle and its length is n;
  - $f_G^\infty(C)$ is the set of vertices of the cycle;
  - Each foil of C contains exactly one vertex of the cycle.
- Class $\mathscr{I}/\mathscr{F}$: C is infinite but all its foils are finite. In this case,
  - The (undirected) f-graph on C is a tree;
  - There is a unique bi-infinite f-path in C, each foil in C contain exactly one vertex of the path, and $f_G^\infty(C)$ coincides with the set of vertices of the path;
  - The order of the foils of C is of type $\mathbb{Z}$; that is, there is no youngest foil in C.

- Class $\mathscr{I}/\mathscr{I}$: C and all foils of C are infinite. In this case,
  - The (undirected) f-graph on C is a tree;
  - C has one end, there is no bi-infinite f-path in C, $D(v)$ is finite for every vertex $v \in C$, and $f_G^{\infty}(C) = \emptyset$;
  - The order of the foils of C is of type $\mathbb{N}$ or $\mathbb{Z}$, that is, there may or may not be a youngest foil in C.

Another definition that will be used in multiple chapters is the notion of *Eternal Family Tree/Forest*. A Family Tree/Forest is a specific type of directed tree/forest, where each vertex has at most one outgoing edge. The parent of a vertex v is the unique vertex w that it points to, denoted as $F(v) = w$. Note that some vertices may not have a parent. When every vertex has exactly one outgoing edge, the tree is referred to as an Eternal Family Tree/Forest (EFT/EFF). The Family Trees are in one-to-one correspondence with undirected trees with a selected vertex or end. A rooted Family Tree define in a same way as unimoduar graph, which root may or may not have a parent. In the case of an Eternal Family Tree where F is defined for all vertices, F is a covariant vertex-shift and the graph T is identical to its F-graph.

$\mathscr{T}$, $\mathscr{T}^*$, and $\mathscr{T}^{**}$ represent sets of isomorphism classes of Family Trees, and they form closed subspaces of $\mathscr{G}^*$ and $\mathscr{G}^{**}$, respectively, in a suitable mark space. *A random rooted Family Tree (FT)* is a random network that almost surely belongs to $\mathscr{T}^*$. *Unimodular FT* is defined as in the definition of unimodular networks.

For example, a tree with one end can be considered an EFT by directing the edge from a vertex v toward the neighbor in the unique semi-infinite path that starts at v and passes through the end.

3.3.3 Local Unimodularity

Tom Hutchcroft introduced the concept of local unimodularity in random networks in his paper [18]. Let $\mathscr{G}_L^*$ be the set of isomorphism classes of triples (G, A, o), where (G, o) is a rooted network, and A is a distinguished subset of vertices, $V(G)$. We denote the isomorphism class of the triple (G, A, o) by $[G, A, o]$. The local topology on $\mathscr{G}_L^*$ is defined similarly to that of $\mathscr{G}^*$. Specifically, two classes of triple networks, $[G, A, o]$ and $[G', A', o']$, are considered to be close in the local topology if there exists a large r and a rooted isomorphism from $Nr(G, o)$ to $N_r(G', o')$ such that the intersection of A' with the $N_r(G', o')$ is equal to the image, under the isomorphism, of the restriction of A to the r-ball around o. Similarly, the set of isomorphism classes $\mathscr{G}_L^{**}$ of quadruples (G, A, o, v), where (G, o, v) is a doubly rooted network and A is a distinguished subset of vertices, is defined in the same way.

A random triple network $[\mathbf{G}, \mathbf{A}, \mathbf{o}]$ is said to be *locally unimodular* if, for all measurable functions $g : \mathscr{G}_L^{**} \to \mathbb{R}^{\geq 0}$, $o \in A$ almost surely, and

$$\mathbb{E}[\sum_{v \in A} g[\mathbf{G}, A, o, v]] = \mathbb{E}[\sum_{v \in A} g[\mathbf{G}, A, v, o]]. \tag{3.3.2}$$

To see an example, consider an arbitrary connected, locally finite graph G with a finite set of vertices A. If o is a random vertex that is uniformly chosen from A, then the resulting triple (G, A, o) is locally unimodular.

In Section 5.2, the Renewal EFT (EFF) graph will be introduced. This graph will be constructed from another random graph called the Renewal Bridge Graph. Using the definition of local unimodularity in random networks, it will be shown that the null recurrent Bridge graph is unimodular on the vertices that are contained in the Renewal EFT (EFF).

3.4 Clustering Algorithms

Part II of this thesis introduces a clustering model for points called the "Hierarchical Mutual Nearest Neighbor Chain (HMNNC)." This section includes a discussion of preliminary clustering methods for (data) points. For further details on this topic, please refer to [5], [6], and [7].

Clustering is the process of grouping a collection of (data) points into "clusters" based on a distance measure, such that the points in the same cluster have a small distance from one another. Clustering algorithms typically assume that the data set, i.e., the set of points, is a finite set, and that the points are located in Euclidean space. In this section, unless otherwise specified, we consider the data set to be a set of points $X = \{x_1, x_2, ..., x_n\}$ in $\mathbb{R}^d$.

A key function in clustering algorithms is the distance function used to measure the distance between clusters. Popular metrics include:

- Minimum Distance (Single Linkage): Given two clusters A and B, the minimum distance function between them, denoted d_1, is defined as $d_1(A, B) = \min_{a \in A, b \in B} d(a, b)$, where $d(a, b)$ is the Euclidean distance between points a and b.

- Maximum Distance (Complete Linkage): In the complete linkage function, the distance function between clusters A and B, denoted d_2, is defined as $d_2(A, B) = \max_{a \in A, b \in B} d(a, b)$, where $d(a, b)$ is the Euclidean distance between points a and b.

- Centroid (Clustroid) distance: The centroid distance between two clusters A and B is defined as the distance between the center point of A and B. The term "centroid" typically refers to the average of the points within the cluster, or the point in the cluster that is nearest to the average. However, there are cases where it is necessary

to represent a cluster with a point that is not related to the center of the cluster. In such cases, people use the term "clustroid" instead of centroid and consider the distance between clustroids as the distance between clusters.

Clustering algorithms can be broadly categorized into two fundamentally different strategies: Point Assignment and Hierarchical Clustering.

In *Point Assignment* algorithms, the algorithm starts with a short phase in which initial clusters are estimated. Then the points are considered in some order, and each one is assigned to the cluster that it best fits. The best-known point assignment algorithm is the K-mean algorithm.

Example 3.4.1 (K-mean algorithm)**.** Let the dataset $X = x_1, x_2, ..., x_n$ be given. The K-mean algorithm assumes that the number of clusters is known and is equal to K. This K can be optimized by some trial and error approach. Next, using some approach that is not discussed here, K points in X are initialized as the representatives or centroids of the K clusters. In the main step of the algorithm, each point in X other than the K selected points is considered and assigned to the closest cluster, where "closest" means by the centroid distance. When all the points are assigned to the clusters, the centroid of each cluster is updated by the points that are in the cluster, and the points are reassigned to the new centroids. This procedure continues until updating the centroids of the clusters does not change them.

On the other hand, *Hierarchical Clustering* algorithms build a hierarchy of clusters from the bottom up or from the top down. There are two main types of hierarchical clustering: Agglomerative and Divisive. In Agglomerative clustering, each data point starts as its own cluster, and then pairs of clusters are iteratively merged into larger clusters until a stopping criterion is met or until there is only one cluster left. In Divisive clustering, all data points start in one cluster, and then the algorithm recursively splits the clusters into smaller ones until a stopping criterion is met. The HMNNC algorithm proposed in this thesis is an example of agglomerative hierarchical clustering. Here the Mutual Nearest Neighbor Chain algorithm, which is a known agglomerative hierarchical clustering and is related to our algorithm is reviewed.

Example 3.4.2 (Nearest Neighbor Chain Algorithm)**.** The nearest neighbor chain (NNC) algorithm was proposed in the 1980s ([19], [20]) as a way to speed up certain hierarchical clustering algorithms. NNC is defined for a finite number of points, let $X = x_0, x_1, ..., x_n \in \mathbb{R}^d$ be a set of n points. Consider each point in X as a singleton cluster. The algorithm starts with an arbitrary cluster, say x_{j_0}. From x_{j_0}, the chain $C = x_{j_0}, x_{j_1}, x_{j_2}, ...$ is constructed, where x_{j_k} is the nearest neighbor of $x_{j_{k-1}}$ with respect to a distance previously chosen between the clusters. Consequently, the distances between the consecutive clusters in the

chain keep decreasing until the chain reaches mutual nearest neighbor (MNN) pairs. The MNN pairs merge to form a new cluster, which is then removed from the chain C. The procedure continues from the last cluster that is in chain C. When all the clusters in chain C are removed, a new random cluster is chosen, and the process repeats until only one single cluster remains (or a stopping criterion is met).

Part I

Doeblin Graph

Summary

Here is a summary of Part I of the thesis. Chapter 4 reviews the definition of the Deoblin Graph, Bridge Graph, and its properties in the case where it is constructed from an irreducible, aperiodic, and positive recurrent MC. The definition and the results in this chapter are from paper [4]. Sections 5, 6, and 7 are our new results that are also available in [1].

Section 5.1 contains the main definitions and results, whereas the other subsections gather the proofs. All the results are first established in a simple example, called the *Renewal Markov Chain*. They are then extended to the general null recurrent MC. Subsection 5.2 introduces this simple example of null recurrent Markov Chain and the *Renewal Bridge Graph* constructed from this MC. Then a random graph on $\mathbb{Z}$, called the *Renewal Eternal Family Tree (Forest)* is introduced. It will be shown that this random graph can be connected and form a tree, or disconnected and form a forest. For the definition of Eternal Family Trees, see [3]. Then other properties of the Renewal EFT, such as its unimodularity, are proved.

In Subsection 5.2.3, a coupling between the Renewal Bridge Graph and the Renewal EFT is defined. This coupling helps one study the properties of the Renewal Bridge Graph using the properties of the Renewal EFT. One of the important properties of the Bridge Graph in the positive recurrent case is the existence of a unique bi-recurrent path as mentioned above. The intersection of this path with time zero is the perfect sample in this case. We will see that this type of path does not exist in the null recurrent Renewal Bridge Graph.

Section 5.3 considers the general null recurrent Markov Chain. It is shown that the structural properties established for the simple example hold for the general Bridge Graphs.

Chapter 6 defines dynamics on the Bridge Graph and studies the properties of these random dynamics. Section 6.1 considers the Taboo and Potential Dynamics on the Bridge Graph and studies their properties. The Taboo PP is introduced in this section and strongly relates to the unique invariant measure of the null recurrent Markov Chain (Theorem 5.2.3). The constructibility of the Taboo and Potential PPs is also discussed. These random measures have infinite support, and due to their definition on the Bridge Graph, it becomes apparent that the entire measure relies on infinite information from the Bridge Graph. Consequently, these measures are not globally constructible. However, it will be shown that they are locally finitely constructible, meaning that the mass of the measure at each point depends only

on an almost surely finite subgraph of the Bridge Graph. Nonetheless, this does not imply universal algorithmic constructibility or the practical determination of this finite subgraph in all cases. It signifies that there are scenarios where finding a stopping time in the Bridge Graph connected to the mass of a point in these measures may not always be possible. This algorithmic constructibility, however, holds in cases where $\{X_t\}_{t\in\mathbb{N}}$ is *monotone.*, allowing for perfect sampling from the Taboo and Potential PPs. To provide further intuition, in cases where paths that construct the Bridge Graph are coupled monotonically, the monotonicity property ensures that the points in the supports of the Taboo PP (or Potential PP) appear in the Bridge Graph in a specific order (from small to large). When a new point of the measure is observed in the Bridge Graph, it indicates that the masses of the smaller points have been completely constructed. A concrete example pertaining to queuing theory is also discussed in detail in Section 6.2. The $GI/GI/1$ queue allows one to illustrate the meaning and the practical interest of these two point processes.

Section 6.3, considers the properties of the two point processes when the MC is positive recurrent. For the Taboo PP, the connection between perfect sampling in the CFTP sense, and the one obtained using the definition of the TPP is discussed. The properties of the Potential PP in the positive recurrent case are also considered. This section also gives some results about the properties of these two dynamics in more general state spaces. Instead of considering these dynamics on the Bridge Graph, they are regarded as Markov Chains on the space of the random measures on S. Chapter 7 gives some results about the properties of these dynamics in more general state spaces. Instead of considering these dynamics on the Bridge Graph, they are regarded as Markov Chains on the space of the random measures on S.

Chapter

4 Doeblin Graph and Bridge Graph

Résumé: Ce chapitre traite de l'algorithme Couplage du Passé (CFTP), qui génère un échantillon parfait à partir de la distribution stationnaire d'une chaîne de Markov. Il exploite le couplage de Doeblin, créant un graphe dirigé aléatoire appelé le Graphe de Doeblin. La première partie de la thèse explore les propriétés de ce graphe pour un espace d'états dénombrable, des chaînes de Markov apériodiques, irréductibles et récurrentes nulles.

Ce chapitre se concentre sur les chaînes de Markov récurrentes positives, mettant en lumière le Graphe de Doeblin unimodularisable et un chemin bi-récurrent unique, dont les sommets servent d'échantillons parfaits de la distribution stationnaire. Le chapitre fournit des définitions précises du Graphe de Pont et du Graphe de Doeblin, examinant leurs propriétés dans les chaînes de Markov récurrentes positives.

Chapter content

4.1 Introduction

The coupling from the past (CFTP) algorithm is a method designed by Propp and Wilson in [21] used to generate a perfect sample from the stationary distribution π of an irreducible and aperiodic Markov Chain on a finite state space. It accomplishes this by utilizing a family of copies of Markov Chains started in all possible states at all possible times and merging them when they meet. This coupling of the chains is known as Doeblin coupling. This procedure results a random directed graph called the Doeblin Graph of the Markov Chain. The study of this graph is a by-product of research on perfect simulation.
In the framework of stochastically recursive sequences (SRS), introduced by Borokov and Foss in [22] and [23], Markov chains are a special case. The existence of a stationary ver-

sion of an SRS to which non-stationary versions converge in a certain sense was established in their work. The CFTP algorithm, introduced by Propp and Wilson in [21], was developed to obtain samples from the stationary distribution of a Markov Chain in finite state space, and can be viewed as a specialization of the general idea of [22] for SRS to the Markov Chain case. Foss and Tweedie in [24] gave a necessary and sufficient condition for the CFTP algorithm to converge.
The Doeblin Graph is used extensively in the first part of this thesis. It is noteworthy that the definition of the Doeblin Graph requires no initial assumption on the Markov Chain that it is constructed from, and therefore can be defined for any Markov Chain. The objective of the first part of this thesis is to investigate its properties in the case where the state space of the Markov Chain is countable and the MC is aperiodic, irreducible, and null recurrent.

In [4], the properties of the Doeblin Graph were studied in the case where the state space of the Markov Chain is countable and the MC is aperiodic, irreducible, and positive recurrent. The most important properties are the fact that this graph is unimodularizable and the existence of a unique bi-recurrent path $\{\beta_t\}_{t\in\mathbb{Z}}$. The bi-recurrent path is a path such that the number of times that it meets each state x in the state space, in both positive and negative times, is infinite a.s. Based on this definition, when a random path is bi-recurrent, it is bi-infinite. The existence of a bi-recurrent path is established from the unimodularizability of the positive recurrent Bridge Graph in the sense of [2]. The importance of the bi-recurrent path is that the vertices belonging to this path are distributed as the stationary distribution of the MC from which the Bridge Graph was constructed. So each vertex in this path can be considered as a perfect sample of the stationary distribution of the MC.
In this chapter, Section 4.2 provides the precise definition of the Bridge Graph and the Doeblin Graph, while Section 4.3 reviews the results regarding the properties of the Doeblin Graph and the Bridge Graph in the case where it is constructed by a positive recurrent Markov Chain.

4.2 Definition

Let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov chain on a state space $\mathscr{S}$. Here $\mathscr{S}$ is considered to be a countable set. As already mentioned in the introduction, the Doeblin Graph is a random graph constructed from the Markov Chain and the SRS framework will be used. The vertices of the Doeblin Graph are $\Sigma = \mathbb{Z} \times \mathscr{S}$. The first component of the vertices is considered as time, and hence the horizontal axis will be referred to as the **time axis**. The second component corresponds to the state of the vertices, and hence the vertical axis will be referred to as the **state axis**. There is an identically distributed and independent source of randomness $\{\xi_t^x, x \in \mathscr{S}\}_{t\in\mathbb{Z}}$, with $\xi_t^x \in \Xi = [0,1]$ such that $\{\xi_t^x\}$ is independent of the initial distribu-

tion of $\{X_t\}_{t\in\mathbb{N}}$. The function $h : \mathscr{S} \times \Xi \to \mathscr{S}$ defines the transitions between the states of $\mathscr{S}$. In addition, h satisfies $P(h(x, \xi_t^x) = y) = p_{x,y}$ for all $x, y \in \mathscr{S}$. The edges of the Doeblin Graph, D_X, are directed edges which are defined from a vertex (x, t) at time t, to a vertex at time $t + 1$, through the random map

$$(t, x) \mapsto (t + 1, h(x, \xi_t^x)). \tag{4.2.1}$$

Consider the subgraph of the Doeblin Graph of X that contains those vertices which are in the union of the trajectories starting from all (t, s^*), $t \in \mathbb{Z}$. This gives a subgraph of the Doeblin Graph called the **(Doeblin) Bridge Graph,** B_X. Here s^* is a fixed arbitrary state in $\mathscr{S}$.

Example 4.2.1. Consider the *lazy random walk*, W, defined on the state space $\mathscr{S} = \mathbb{Z}$, with the following transition probabilities: the walk stays at the current state, i, with probability $1/3$, and moves to each neighbor of i, i.e., $i + 1$ or $i - 1$, at random with probability $1/3$. Then, one can consider the Doeblin Graph of W, constructed from the driving sequence $\{\xi_t^y, y \in \mathbb{Z}\}_{t\in\mathbb{Z}}$, and the transition function h, where $\{\xi_t^y, y \in \mathbb{Z}\}$ are maximally coupled for a given t, i.e., at each time t, for all i, $\xi_t^i = \xi_t^0$, and for all i the transition $h(i, \xi_t^i) = h(i, \xi_t^0) = i + h(0, \xi_t^0)$. In this example, one can show that the Bridge Graph of this lazy random walk coincides with its Doeblin Graph (see Figure 4.1).

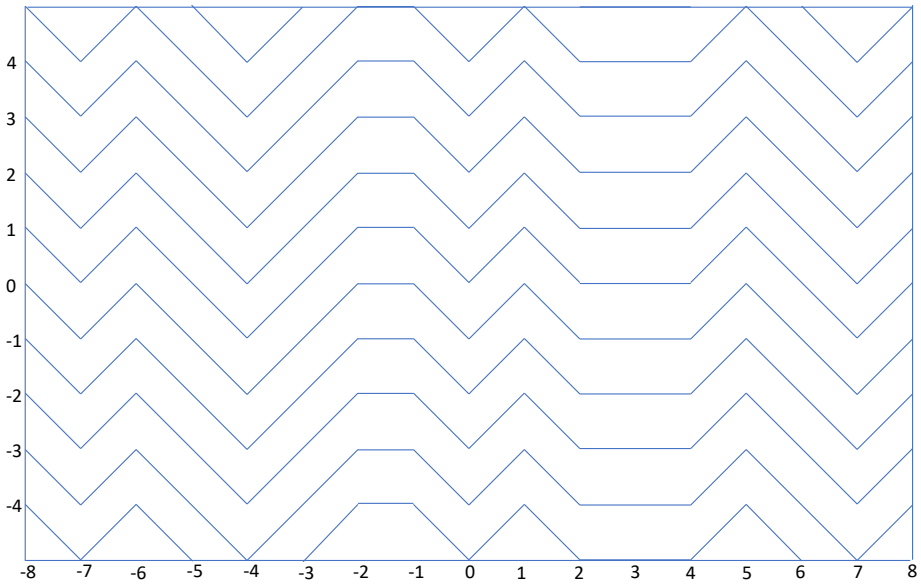


Figure 4.1: The maximally coupled Doeblin Graph and Bridge Graph of lazy random walk

4.3 Properties in the Positive Recurrent Case

This section reviews the main properties of the Doeblin Graph and Bridge Graph constructed from an irreducible, aperiodic, positive recurrent Markov chain (MC). The case under study is discussed in the paper by Baccelli et al. [4]. We will now review the main properties presented in that paper. In the following chapters, we will compare and contrast this case with the case where the MC is irreducible, aperiodic, and null recurrent.
First, it is important to note that the Doeblin Graph D is acyclic as an undirected graph. This is because all outgoing edges point forward by one unit in time, and each vertex has only one outgoing edge. Therefore, the Bridge Graph, being a subgraph of the Doeblin Graph, is also acyclic. Both of these graphs can be represented as trees or forests. The following results pertain to the connected components of these graphs.

Proposition 4.3.1 (Proposition 2.3 in [4])**.** *Let X be a MC with transition probabilities $\mathbf{P}$, where $\mathbf{P}$ is irreducible and positive recurrent with period d. Suppose D_X is the Doeblin graph constructed from, X which has fully independent transitions. Then, almost surely, D_X has d connected components. In particular, if $\mathbf{P}$ is irreducible, aperiodic, and positive recurrent, then D_X is a tree.*

Lemma 4.3.2 (Lemma 2.11 in [4])**.** *Let X be a MC with transition probabilities $\mathbf{P}$, where $\mathbf{P}$ is irreducible, aperiodic, and positive recurrent. If $u, v \in V(D)$ are in the same connected component of D_X, they are also in the same connected component of B_X.*

These results indicate that in the case of an irreducible, aperiodic, and positive recurrent MC X, both the Bridge Graph and the Doeblin Graph constructed from X are trees. The condition that D_X is a tree is generally equivalent to the strong coupling convergence [[22], [23]] of the path starting from the state $(0, x^*) \in \mathbb{Z} \times \mathscr{S}$ to a stationary version of the stochastic recursive sequence (SRS).
From this point onward, we will focus exclusively on the case where the MC is irreducible, aperiodic, and positive recurrent.
Another important property is the unimodularizability of the Bridge Graph. A random graph G is called *unimodularizable* when it is possible to choose a root o from the set of vertices $V(G)$ such that (G, o) is unimodular, as defined in [2](see also Section 3.3).

Theorem 4.3.3 (Theorem 4.1 in [4])**.** *The Bridge Graph constructed from an irreducible, aperiodic, and positive recurrent Markov chain is unimodularizable.*

One key property that helps prove unimodularizability is that the intersection of the Bridge Graph with the time axis is almost surely finite with finite expectation. Roughly speaking, if we consider the intersection of the Bridge Graph with time zero, denoted as

B_X^0, and choose a vertex o uniformly from B_X^0, then (B_X, o) becomes unimodular.
One can naturally define the vertex shift f^0 on B_X. For each $(x,t) \in V(B_X)$, consider the unit directed edge that starts from (x, t) and goes to (y, t + 1). Then define $f^0(x,t) = (y, t+1)$. With this definition, the f^0-graph of B_X coincides with B_X . It is known that there is only one connected component in B_X. Using the foil classification theorem in unimodular networks (see Theorem 3.3.2), B_X belongs to one of the classes $\mathscr{F}/\mathscr{F}$, $\mathscr{I}/\mathscr{F}$, or $\mathscr{I}/\mathscr{I}$.

Lemma 4.3.4 (See Section 4.2 in [4]). *Let the Markov chain $X = \{X_t\}_{t\in\mathbb{N}}$ be irreducible, aperiodic, positive recurrent and consider the vertex shift f^0 on B_X, the Bridge Graph of the vertex x^*. The Bridge Graph B_X belongs to the $\mathscr{I}/\mathscr{F}$ class of the Foil Classification Theorem in unimodular networks.*

As explained in Section 3.3, according to Theorem 3.3.2, the connected component of the $\mathscr{I}/\mathscr{F}$ class, in unimodular networks, has the property of containing a unique bi-infinite path. Therefore, there exists a unique bi-infinite path in B_X. The following theorem states the properties of this unique bi-infinite path.

Theorem 4.3.5 (See Section 4.2 in [4]). *Suppose D_X is the Doeblin Graph constructed from an irreducible, aperiodic, positive recurrent Markov Chain $\{X_n\}_{n\in\mathbb{N}}$, and B_X is the Bridge Graph of the vertex x^*. The Bridge Graph B_X contains a unique bi-infinite path $(\beta_t)_{t\in\mathbb{Z}}$. This path is uniquely bi-recurrent for the state x^* (and every $x \in \mathscr{S}$) in D_X, meaning that the set $t \in \mathbb{Z} : x_t = x^*$ is unbounded both above and below. Moreover, the path $(\beta_t)_{t\in\mathbb{Z}}$ is stationary.*

Theorem 4.3.5 shows that similar to the standard CFTP setup, there exists a β_0 at time 0 in D_0 that is a perfect sample from the stationary distribution of the Markov chain. However, unlike in the standard CFTP setup, it is not known whether there is an algorithm that can find β_0 in finite time.

Chapter

5 Null recurrent Bridge Graph

Résumé: La Section 5.1 contient les principales définitions et résultats, tandis que les autres sous-sections regroupent les démonstrations. Tous les résultats sont d'abord établis dans un exemple simple, appelé la *Chaîne de Markov de Renouvellement*. Ils sont ensuite étendus à la chaîne de Markov générale à récurrence nulle. La sous-section 5.2 introduit cet exemple simple de chaîne de Markov à récurrence nulle, ainsi que le *Graphe de Pont de Renouvellement* construit à partir de cette chaîne. Ensuite, un graphe aléatoire sur $\mathbb{Z}$, appelé l'*Arbre (Forêt) Familiale Éternelle de Renouvellement*, est présenté. Il sera démontré que ce graphe aléatoire peut être connecté pour former un arbre, ou déconnecté pour former une forêt. Ensuite, d'autres propriétés de l'arbre (forêt) familiale éternelle de renouvellement, telles que son unimodularité, sont prouvées.

Dans la sous-section 5.2.3, un couplage entre le graphe de pont de renouvellement et l'arbre (forêt) familiale éternelle de Renouvellement est défini. Ce couplage aide à étudier les propriétés du graphe de pont de renouvellement en utilisant les propriétés de l'arbre (forêt) familiale éternelle de renouvellement. Une des propriétés importantes du graphe de pont dans le cas récurrent positif est l'existence d'un unique chemin bi-récurrent, comme mentionné ci-dessus. L'intersection de ce chemin au temps zéro est l'échantillon parfait dans ce cas. Nous verrons que ce type de chemin n'existe pas dans le graphe de pont de renouvellement à récurrence nulle.

La Section 5.3 concerne la chaîne de Markov générale à récurrence nulle. On montre que les propriétés structurelles établies pour l'exemple simple s'appliquent aux graphe de ponts généraux.

Chapter content

5.1 Main Definition and Results

Consider a Markov Chain $X = \{X_t\}_{t\in\mathbb{N}}$ defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with a countable state space $\mathscr{S}$ and transition probabilities $P = (p_{x,y})_{x,y\in\mathscr{S}}$. As mentioned in the introduction, two different dynamics are considered on the random counting measures or point processes with multiplicity on $\mathscr{S}$, satisfying (1.0.2).
The first dynamics is the **Taboo Dynamics**, denoted by H^T which is defined with respect to a reference point $s^* \in \mathscr{S}$. It is defined by $M^T_{t+1} = H^T(M_t, \xi_t)$, with, for each $x, y \in \mathscr{S}$,

$$M^T_{t+1}(y) = \begin{cases} \sum_{x\in\mathscr{S}} M^T_t(x)\mathbb{1}_{\{h(x,\xi^x_t)=y\}}, & y \neq s^* \\ 1, & y = s^*. \end{cases} \tag{5.1.1}$$

This dynamics constructs the random measure at time $t+1$ from the random measure at time t. It sends some mass from each state x to state y with rule h while adding up the masses sent to the same state y. It ignores all the masses that enter s^* at time $t+1$ and puts mass 1 at this point.
The second dynamics is the **Potential Dynamics**. It is denoted by H^P. One has $M^P_{t+1} = H^P(M_t, \xi_t)$, with, for each $x, y \in \mathscr{S}$,

$$M^P_{t+1}(y) = \begin{cases} \sum_{x\in\mathscr{S}} M^P_t(x)\mathbb{1}_{\{h(x,\xi^x_t)=y\}}, & y \neq s^* \\ \sum_{x\in\mathscr{S}} M^P_t(x)\mathbb{1}_{\{h(x,\xi^x_t)=y\}} + 1 & y = s^*. \end{cases} \tag{5.1.2}$$

As the Taboo Dynamics, there is a reference point $s^* \in \mathscr{S}$. For constructing the random measure at time $t+1$ from the random measure at time t, the Potential Dynamics sends some mass from each state x to state y with rule h while adding up the masses sent to the same state y; in addition it adds mass one at point s^*.
The difference between the Taboo Dynamics and the Potential Dynamics is that the former always puts mass one at s^* and does so by deleting the masses arriving at this point and adds mass one at s^*. In contrast, the Potential Dynamics just adds mass one at s^*.
The update rules of these two dynamics are related to the Doeblin coupling of the MC. Therefore, the main tool that will be leveraged to study these dynamics is the Doeblin Graph and its subgraph, the Bridge Graph.

Remark 5.1.1. In the definition of the Doeblin Graph and the dynamics defined in (1.0.2), the same source of randomness $\{\xi^x_t, x \in \mathscr{S}\}_{t\in\mathbb{Z}}$ is used. This sequence is always considered to be independent in t but not necessarily in x. i.e., at each time t, the random variables

$\{\xi_t^x\}$ can be coupled. When the random variables $\{\xi_t^x\}$ are independent in both t and x, this sequence will be called *totally independent*.

It will be shown in Section 5.3 that the (irreducible, aperiodic) null recurrent Bridge Graph has the following properties:

Proposition 5.1.2. *The null recurrent Bridge Graph is either connected, in which case it is a tree, or disconnected, in which case it is a forest made of an infinite number of trees, and both cases can happen.*

When this null recurrent Bridge Graph is connected, the following property will be proved in Section 5.3:

Proposition 5.1.3. *If the null recurrent Bridge Graph is a tree, it has no bi-infinite path.*

It will also be proved in Section 5.3 (see also Remark 5.3.6) that a similar result is also valid under the following condition:

Proposition 5.1.4. *Consider a null recurrent Bridge Graph, B_X, which may not be connected. Suppose the Bridge Graph satisfies the condition that for all (t_1, s_1) and $(t_2, s_2) \in \mathbb{Z} \times \mathscr{S}$, the paths passing through (t_1, s_1) and (t_2, s_2) meet each other with positive probability in finite time. Then, the Bridge Graph has no bi-infinite path.*

The condition of the last proposition does not imply that the graph B_X is connected. The Renewal EFF provides an example where the graph may satisfy this property and be disconnected. The following result is proved in Section 5.3:

Proposition 5.1.5. *When the null recurrent Bridge Graph is a tree, it is not unimodularizable, in general.*

Concerning unimodularizability, defined in Section 3.3, a second and instrumental result is that this random graph is always locally unimodular. The unimodular graph H in this case is that with the set of vertices $V = \{x^*, t\}_{t\in\mathbb{Z}}$ and the set of edges $F = \{(t, x^*), (\mathscr{T}_t, x^*)\}_{t\in\mathbb{Z}}$, where $\mathscr{T}_t$ is the time of first return to x^* after time t. This graph, which is called the **Recurrence Time EFF**, is unimodular indeed (Proposition 5.3.3).

Consider all the vertices in the intersection of the Bridge Graph and the zero timeline, i.e., those on the state axis. This random set will be referred to as the **S-set** (Support Set). The properties of the S-set in the null recurrent Bridge Graph are studied in Section 5.3.
Two multiplicities for a point in the S-set are now defined. One can look at each vertex in the Bridge Graph (or any directed graph) as an individual. Moreover, by following the outgoing edge, go from each vertex to its parent vertex. In the Bridge Graph, one can consider the descendants of a vertex that lie on the time axis, i.e., belonging to $\mathbb{Z} \times \{s^*\}$. Descendants of this type are referred to as ***-descendants**.

Definition 5.1.6. The **Taboo multiplicity** of a point in the S-set of a Bridge Graph is the number of its *-descendants such that the path from this descendant to the S-set does not visit state s^* before time zero. See Figure 5.1. Note that by definition, the Taboo multiplicity is positive at all points of the S-set.

Remark 5.1.7. Note that the order of the generations is not consistent with the time direction, as ancestors live in the future, and these notions should not be mixed up.

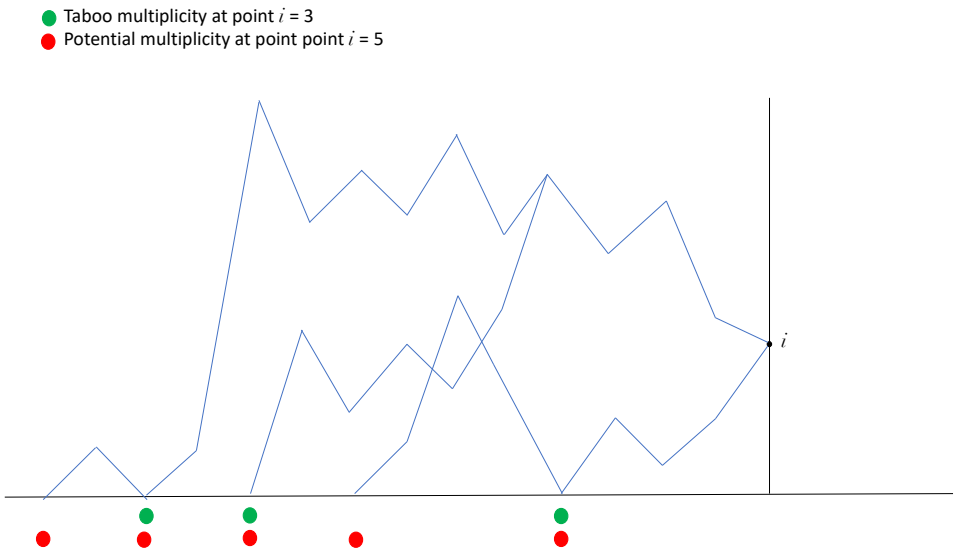


Figure 5.1: The Taboo multiplicity and the Potential multiplicity

It will be shown in Proposition 5.2.19 that the Taboo multiplicity of any vertex in the Bridge Graph is a.s. finite, so the definition of Taboo multiplicity gives a locally finite random measure whose support is the S-set. This random measure is called the **Taboo Point Process (Taboo PP)** and is denoted by τ. Below, $\tau_t(j)$ denotes the random mass (multiplicity) that the Taboo PP puts on j at time t. The following result is proved in Section 6.1:

Theorem 5.1.8. *The Taboo PP, τ, is a steady state of the Taboo Dynamics,*

$$\tau_{t+1} = H^T(\tau_t, \xi_t). \tag{5.1.3}$$

The following theorem, also proved in Section 6.1, shows that there is a relation between the Taboo PP in the null recurrent Bridge Graph and the stationary measure of the null recurrent Markov Chain:

Theorem 5.1.9. *Let $\{X_t\}_{t\in\mathbb{Z}}$ be an aperiodic and recurrent MC, and B_X be its associated Bridge Graph with the driving sequence $\{\xi_t^x\}_{t\in\mathbb{Z},x\in S}$. Then $\mathbb{E}[\tau_t(i)]$, the mean measure of*

the Taboo PP at points, does not depend on t, nor on the coupling of $\{\xi_t^x\}_{t\in\mathbb{Z},x\in S}$ *in* x*, and it is equal to the stationary measure of that point in the Markov Chain* $\{X_n\}$*. That is,*

$$\mathbb{E}[\tau_t(i)] = \sigma(i), \quad \forall i \in \mathscr{S}, \tag{5.1.4}$$

where σ *is the invariant measure of the Markov Chain* $\{X_n\}$ *and* $\sigma(s^*) = 1$*.*

Note that the last theorem holds regardless of the fact that the associated null recurrent Bridge Graph is connected or not.

The second multiplicity that will be considered for a point in the S-set is the "Potential multiplicity":

Definition 5.1.10. The **Potential multiplicity** of a point in the S-set is the number of all its *-descendants in the Bridge Graph. See Figure 5.1.

The Potential multiplicity on the S-set gives a random measure with support the S-set itself. This random measure is called the **Potential Point Process (Potential PP)** and denoted by π. Again, $\pi_t(j)$ denotes the random mass that the Potential PP puts on j at time t. Proposition 5.3.10 shows if the null recurrent Bridge Graph is a tree, then the Potential multiplicity of the vertices in the null recurrent Bridge Graph is a.s. finite. In addition the following result holds:

Theorem 5.1.11. *In the null recurrent case, if the null recurrent Bridge Graph is a tree or, under the assumption that there is no bi-infinite path in its associated recurrent time EFF (as stated in Proposition 5.1.4), the Potential PP,* π*, is a locally finite steady state of the Potential Dynamics*

$$\pi_{t+1} = H^P(\pi_t, \xi_t). \tag{5.1.5}$$

The relation between the Potential PP of a null recurrent Bridge Graph and the associated MC is summarized in the following theorem.

Theorem 5.1.12. *Consider a null recurrent Markov Chain* $\{X_n\}$ *and its associated Bridge Graph* B_X*. The mean measure of the Potential PP is equal to the entries of a row in the*

potential matrix [1] *of the Markov Chain* $\{X_n\}$*. So,*

$$\mathbb{E}[\pi_t(i)] = \infty, \quad \forall i \in \mathscr{S}. \tag{5.1.6}$$

Remark 5.1.13. Equation (5.1.6) remains valid in the positive recurrent case. Also, in the transient Bridge Graph, it can be shown that the potential multiplicities are such that their means are equal to the entries of the classical potential matrix for MCs [25]. This is why the multiplicity, the associated point process and the dynamics are called "potential". In the positive recurrent case, since the value of the potential multiplicity, in one state is infinite, it is not a locally finite measure. So Theorem 5.1.11 does not hold in the positive recurrent case.

5.2 Renewal Bridge Graph

This section first introduces the Renewal Markov Chain, a simple example of recurrent Markov Chain, which may be positive or null recurrent. After that, the Renewal Bridge Graph, which is the Bridge Graph constructed from the Renewal Markov Chain, is introduced. Before going through the proof of the properties of the general null recurrent Bridge Graph, the proofs are first established in this particular example.
Consider random variable η on $\mathbb{N}^*$ [2] with distribution

$$\Lambda = \{p_k, k \in \mathbb{N}^*\}, \tag{5.2.1}$$

where $p_k = \mathbb{P}(\eta = k)$. Below, it is assumed that η is such that the set $A = \{k \in \mathbb{N}^*; p_k > 0\}$ has infinite cardinality and the greatest common divisor of A is equal to 1. From this distribution, one can define the following Markov Chain:

Definition 5.2.1 (Renewal Markov Chain)**.** Consider the following transition probabilities on the non-negative integers: for $i \neq 0$

$$p_{ij} = \begin{cases} 1 & \text{if } j = i - 1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.2.2}$$

[1] Suppose you have a Markov Chain X with a countably infinite state space $\mathscr{S}$ and transition probabilities $P = [p_{xy}]$ for $x, y \in \mathscr{S}$. The potential matrix $U = [u(x, y)]$ of X is an infinite matrix, and its entries are defined as:

$$u(x, y) = \sum_{k=0}^{\infty} P^k(x, y)$$

Here, $P^k(x, y)$ represents the probability that, starting from state x, the Markov Chain reaches state y after exactly k steps.

[2] In this Thesis, $\mathbb{N}^*$ denotes the natural numbers without zero and $\mathbb{N}$ natural numbers with zero.

and for $i = 0$

$$p_{0j} = p_{j+1}, \tag{5.2.3}$$

where the p_js are the probabilities of the random variable η defined in (5.2.1). This Markov Chain is called the **Renewal Markov Chain**. The assumptions that A is infinite and $gcd(A) = 1$ make the Renewal MC irreducible and aperiodic. These assumptions are assumed to hold throughout this section. Starting from 0, it a.s. returns to this point. So point 0 is recurrent, and thus the Markov Chain is recurrent. Let T_0^+ be the first return time to 0 starting from 0. Then

$$\mathbb{E}[T_0^+] = \sum_i \mathbb{E}[T_0^+ | \text{first jump is i}] p(\text{first jump is } i) = \sum_i (i+1).p_{0(i+1)} = \mathbb{E}[\eta]. \tag{5.2.4}$$

So if $\mathbb{E}[\eta] = \infty$, this Markov Chain is null recurrent, and hence in this case it is called the **null recurrent Renewal Markov Chain**.

The Doeblin Graph of the Renewal Markov Chain is as follows: the set of vertices of this random graph is $\Sigma = \mathbb{Z} \times \mathbb{N}$, and the driving sequence is $\xi_t^n, n \in \mathbb{N}$. For $i \neq 0$, vertex (t, i) has a single outgoing edge which goes to the vertex $(t+1, h(i, \xi_t^i))$ with $h(i, \xi_t^i) = i - 1$ a.s. For $i = 0$, the outgoing edge from $(t, 0)$ goes to vertex $(t+1, h(0, \xi_t^0))$ with $\mathbb{P}(h(i, \xi_t^0) = j) = p_{0j}$ in (5.2.3).The union of the trajectories starting from $(t, 0)$, $t \in \mathbb{Z}$, forms the Bridge graph with respect to $s^* = 0$. This graph is called the *Renewal Bridge Graph*.

5.2.1 Renewal Eternal Family Forest and Tree

The Renewal Eternal Family Forest (EFF) is a random graph defined from Renewal Bridge Graph.

Definition 5.2.2. Consider the directed random graph $G^\eta = (V, E)$ on $\mathbb{Z}$, with vertices $V = \mathbb{Z}$. The set of edges, E, is as follows: at each vertex i, there is an edge to vertex $i + \eta_i$, where the random variables $\{\eta_i\}_{i \in \mathbb{Z}}$ are i.i.d. with $\eta_i \sim \eta$ defined in (5.2.1). One can verify that this graph has no loops, and hence it is either a tree or a forest. If this graph is a tree, it has all the properties of an Eternal Family Forest as defined in [4]. So G^η is called the **Renewal Eternal Family Forest (Renewal EFF)**. In the connected case, it is referred to as the **Renewal Eternal Family Tree (Renewal EFT)**.

Below, it is assumed that $\mathbb{E}(\eta) = \infty$. Proposition 5.2.3 shows that both the connected case (tree) and the disconnected case (forest made of more than one tree) can arise. This proposition considers a specific distribution for η, satisfying the infinite mean property.

Proposition 5.2.3. *Let $0 < \alpha \leq 1$ be fixed. Suppose that η has the following probability distribution,*

$$P(\eta = j) = q_j = \frac{c_1}{j^{\alpha+1}}, \quad , j \geq 1 \tag{5.2.5}$$

which gives the following tail distribution :

$$P(\eta > j) = \frac{c_2}{j^{\alpha}}, \quad j \geq 1, \tag{5.2.6}$$

where c_1 and c_2 are normalizing constants. Then the random graph constructed in Definition 5.2.2 is a.s. a Renewal EFT when $\alpha \geq \frac{1}{2}$ and a.s. a forest when $\alpha < \frac{1}{2}$.

The proof of Proposition 5.2.3 is provided in Subsection 11.0.1 of the appendix. For the remainder of the document, it is assumed that η satisfies $\mathbb{E}(\eta) = \infty$, unless mentioned otherwise.

Remark 5.2.4. The Renewal EFF is not limited to distribution (5.2.5), which is considered only for showing that both the connected and disconnected cases exist when η has an infinite mean.

5.2.2 Properties of the Renewal EFF

Here are some properties of Renewal EFF to be used later. Proposition 5.2.5 studies the unimodular property of the Renewal EFF. For the definition and some examples of unimodular random networks, see [2].

Proposition 5.2.5. *onsider the Renewal EFF with additional edges connecting vertex n to $n+1$ for all $n \in \mathbb{Z}$. This graph is referred to as the* Renewal EFF with ghost edges. *The Renewal EFF with ghost edges is a unimodular graph.*

Proof. Let (G, o) be the deterministic graph with vertices $V = \mathbb{Z}$ and edge set $E = \{(n, n+1)|\forall n \in \mathbb{Z}\}$ rooted at 0. This is a unimodular graph. For all unimodular networks, it is possible to enrich vertices and edges with i.i.d. marks and preserve unimodularity (see [2]). Since the Renewal EFF with ghost edges is a random graph obtained by adding i.i.d. marks to (G, o), it is a unimodular network. □

Remark 5.2.6. Note that the inclusion of ghost edges in the Renewal EFF is motivated by the requirement in the definition of unimodularity, which is specified for connected, locally finite graphs. Without the ghost edges, the Renewal EFF is not connected in general. However, by assigning a mark of zero to these ghost edges and a mark of one to the edges of the Renewal EFF, all properties of unimodularity, such as the Mass Transport Principle and local finiteness, are satisfied for the Renewal EFF.

The same holds true for the Recurrence Time EFF, as defined in Section 5.3. Therefore, throughout the rest of Part I, when referring to the unimodularity of the Renewal EFF or Recurrence Time EFF, we are specifically addressing the random graph with the inclusion of ghost edges, with mark zero assigned to these ghost edges.

Proposition 5.2.7. *Let $G^\eta = G^\eta(V, E)$ be a Renewal EFF. Then each connected component of G^η is $\mathscr{I}/\mathscr{I}$ in the sense of the foil classification theorem of unimodular networks.*

Proof. Consider the vertex shift f on G^η, which maps each vertex to its right adjacent vertex. Each connected component of the Renewal EFF is an infinite tree, so it is either in the $\mathscr{I}/\mathscr{I}$ class or the $\mathscr{I}/\mathscr{F}$ class of the foil classification theorem. First, suppose that G^η is a.s. connected and that it is $\mathscr{I}/\mathscr{F}$. It follows that there is a unique bi-infinite f-path, $\mathscr{P}$, in this component (see Theorem 3.3.2). Since $\mathscr{P}$ is unique, it is distinguishable in the whole graph. So it is a covariant subgraph of G^η. Using Lemma 2.8 of [3], $\mathscr{P}$ has a positive density in $\mathbb{Z}$. In the sense that $\mathbb{P}(0 \in V_{\mathscr{P}}) > 0$, where $V_{\mathscr{P}}$ is the set of vertices of $\mathscr{P}$. Consider the measurable function g defined as follows: $g[G^\eta, x, y] \equiv 0$, if there is no bi-infinite path in G^η. When the bi -infinite path $\mathscr{P}$ exists:

- $g[G^\eta, x, y] = 1$, if x, y are two consecutive vertices in $\mathscr{P}$ such that $x < y$,
- $g[G^\eta, x, y] = 1$, if $x \notin V_{\mathscr{P}}$, and y is the nearest vertex to the left of x that belongs to $V_{\mathscr{P}}$, and
- $g[G^\eta, x, y] = 0$, otherwise.

Using the mass transport principle, one can write:

$$\mathbb{E}\left[\sum_{x\in\mathbb{Z}} g[G^\eta, 0, x]\right] = \mathbb{E}\left[\sum_{x\in\mathbb{Z}} g[G^\eta, x, 0]\right]. \tag{5.2.7}$$

The left-hand side of equation (5.2.7) is equal to the probability of the existence of $\mathscr{P}$ in G^η, and the right-hand side of the equation is equal to

$$\mathbb{P}(0 \in V_{\mathscr{P}}).\mathbb{E}[\text{ number of the vertices between 0 and its right neighbor in } \mathscr{P}|0 \in V_{\mathscr{P}}]. \tag{5.2.8}$$

Since $\mathbb{P}(0 \in V_{\mathscr{P}}) > 0$, the equality of (5.2.7) and (5.2.8) gives that the expectation of the number of the vertices between 0 and its right vertex in $V_{\mathscr{P}}$ given that $0 \in V_{\mathscr{P}}$ is finite.
On the other hand, note that the existence of a bi-infinite path is a property of the left-hand side of the graph. In the sense that if there exists a path that comes from $-\infty$ and reaches zero, it is bi-infinite. The path on the right-hand side of 0 is "fresh", and the distribution of the length of $\mathscr{P}$ edges on this side is the same as η. So the distance between 0 and its right neighbor in $\mathscr{P}$ has an infinite mean, while it is shown above that this expectation is finite.

Thus this path does not exist. Hence, the tree belongs to the $\mathscr{I}/\mathscr{I}$ class.
Consider now the case where G^η is not connected. Suppose only one bi-infinite path exists in the graph. Then this path is again a covariant subgraph of G^η, i.e., with positive probability, zero belongs to this path, and the same argument as in the EFT case shows that it is impossible. So either there is no bi-infinite path in the graph, or there is more than one bi-infinite path. Suppose that the latter case happens. The variables $\eta_i, i < 0$ determine the number of bi-infinite paths in the EFF. Let $\mathscr{P}_1$ and $\mathscr{P}_2$ be two bi-infinite paths that come from $-\infty$ and reach time zero. Since after 0, these two paths do not depend on the past, and since the gcd of A (this set is defined at the beginning of the section) is equal to one, these two paths meet each other with positive probability. So there is more than one bi-infinite path in one connected component of the EFF with positive probability, which is impossible due to the foil classification theorem of unimodular networks. □

Rephrased in terms of the classification of unimodular EFTs, the last result complements the known fact that the renewal EFT is $\mathscr{I}/\mathscr{F}$ in the case where the renewal distribution has finite mean, by showing that it is $\mathscr{I}/\mathscr{I}$ when this mean is infinite.

Remark 5.2.8. The Renewal EFF is obviously linked to *renewal theory*. It is also linked to the so-called *subtractive random forest.* The latter were recently introduced for the modeling of online recommendation systems in [26]. In fact, such a random forest is a subgraph of some Renewal EFF. The analysis of Renewal EFF proposed in the present part of this thesis can hence shed light on the properties of subtractive random forests. In particular, the Recurrence Time EFF introduced in Section 5.3 which generalizes the Renewal EFF allows one to consider stationary instead of i.i.d. random variables that can be interesting for studying the properties of subtractive random forests in more general cases.

5.2.3 Properties of the Renewal Bridge Graph

Basic Properties

In the Renewal MC, the transition probabilities from zero are the same as the jumps distribution in the Renewal EFT (EFF). Using this, one can define a coupling between the Renewal EFT (EFF) and the Renewal Bridge Graph.

Definition 5.2.9. In the Renewal Bridge Graph, consider e_t, the outgoing edge [3] at vertex (t, s^*). Let $(t+1, s')$ be the end of edge e_t. Then the jump at time t is defined by $|s' - s^*|$.

Let $\{\eta_i\}_{i\in\mathbb{Z}}$ be the length of the outgoing edge at vertex i in the Renewal EFT, and

[3] In the Bridge Graph (and in the EFF), there is a natural direction for the edges from time t to time $t+1$. With this direction, each edge has a beginning vertex and an end vertex.

$\{\eta_i'\}_{i\in\mathbb{Z}}$ be the jump at time i in the Renewal Bridge Graph. Then

$$\eta_i \sim \eta_i' + 1 \quad \forall i.$$

The coupling between (η_i, η_i') is defined by taking

$$\eta_i = \eta'_i + 1 \quad \forall i \quad \text{a.s.} \tag{5.2.9}$$

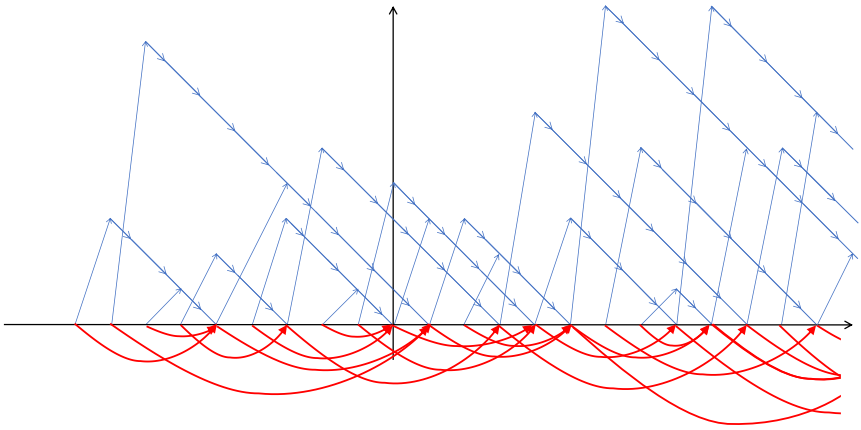
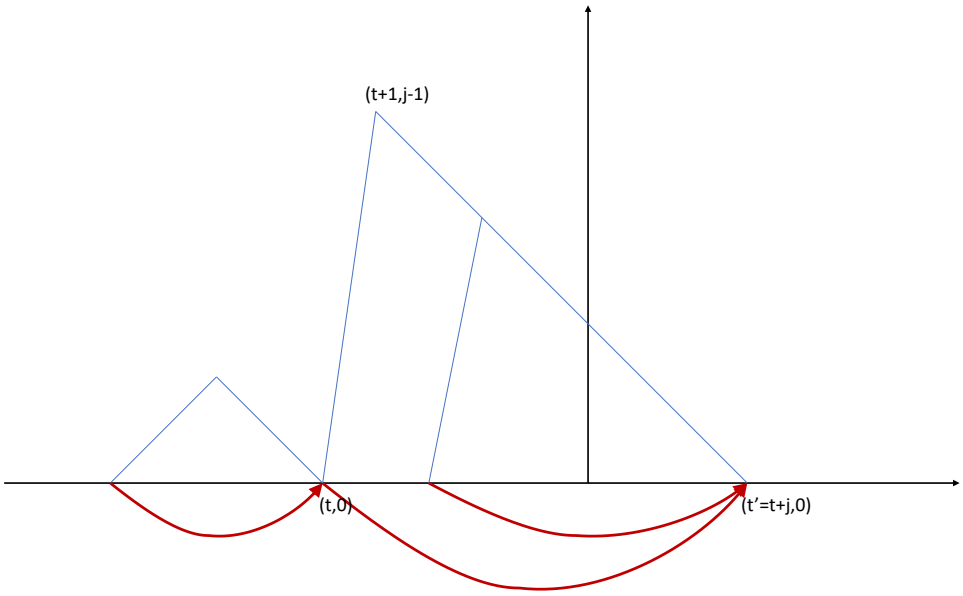


Figure 5.2: Coupling between Renewal EFT and Renewal Bridge Graph

Proposition 5.2.10. *The irreducible aperiodic null recurrent Renewal Bridge Graph is either connected or disconnected. Both cases can happen, i.e., there are examples where it is a tree and examples where it is a forest made of infinitely many trees.*

Proof. In the Bridge Graph, there is only one outgoing edge from each vertex. Also, the edges are just going forward in time. So there is no cycle. Hence the Bridge Graph is either a tree or a forest. It remains to show both cases are possible. This is because, given the coupling (5.2.9), by the following argument, the connectedness of the Renewal Bridge Graph and the Renewal EFT(EFF) are equivalent.
Suppose that in the Renewal EFT, there is an edge from vertex t to vertex $t' = t + j$, where j is the value of η_t. Correspondingly, using the coupling defined in (5.2.9), in the Renewal Bridge Graph, there is an edge $\eta_i' = \eta_i - 1 = j - 1$ between the vertices $(t, 0)$ and $(t + 1, j - 1)$. Due to the construction of the Renewal Bridge Graph, it has a decreasing path from vertex $(t+1, j-1)$ to vertex $(t+j, 0)$. It means that there is a path in the Bridge Graph starting from vertex $(t, 0)$ in the time axis and back to the time axis for the first time

Figure 5.3: Equivalence of connectedness of the Renewal Bridge Graph and Renewal EFT



again at vertex $(t+j, 0)$.
So if two paths in the Renewal EFT (EFF) starting from two different vertices in the Renewal EFT (EFF) meet each other at a given time, the paths starting from the corresponding vertices in the Renewal Bridge Graph meet each other and vice versa. See Figure 5.3. Thus a Renewal Bridge graph is a tree if and only if its corresponding Renewal EFT is a tree. Then the result follows from Proposition 5.2.3. □

Proposition 5.2.11. *Every bi-infinite path, $\{\beta_t\}_{t\in\mathbb{Z}}$, in the Renewal Bridge Graph, B_X, corresponds to a bi-infinite path in its associated Renewal EFF.*

Proof. The proof consists in proving that, every bi-infinite path $\{\beta_t\}_{t\in\mathbb{Z}}$, in the Renewal Bridge Graph, is bi-recurrent, i.e., it meets the time axis in both the positive and negative parts a.s., infinitely many times. Since the MC X is recurrent, every bi-infinite path meets the positive part of the time axis a.s. infinitely many times. So it is enough to show that it meets time axis infinitely many times in the negative part.
let T be an arbitrary element of the time axis. Consider the following set in the Renewal Bridge Graph:

$D =\{T-t$; the path that starts from $(T-t, s^*)$ goes back to the time axis for the first

time at time $T\}$.

For a fix t, the probability that $T - t \in D$ equals the probability that at time $T - t$, the jump is equal to $t - 1$, i.e. p_{t-1}. Since $\sum_{t=1}^{\infty} p_t = 1$, one can conclude, using the Borel-Cantelli lemma, that the cardinality of D is a.s. finite. On the other hand, suppose there exists an infinite path, $\{\beta_t\}_{t\leq T}$, in the Renewal Bridge Graph that comes from $(-\infty, +\infty)$ and reaches the time axis for the first time at T. Hence

$$\{\beta_t\}_{t\leq T} = \{(t, T - t)\}_{t\leq T}. \tag{5.2.10}$$

Since $\{\beta_t\}_{t\leq T}$ is a path in the Renewal Bridge Graph, every vertex in this path has a backtrack to the time axis. It means that there exist infinitely many edges starting from the time axis and ending up at $\{\beta_t\}_{t\leq T}$. Note that (5.2.10) gives that the probability that this happen is equal to the probability that D be infinite, which is equal to zero. So a.s., in the Renewal Bridge Graph, there is no bi-infinite path that comes from $(-\infty, +\infty)$ and reaches the time axis for the first time at some T. So every bi-infinite path in the Bridge graph is bi-recurrent. □

Proposition 5.2.12. *The Renewal Bridge Graph has no bi-infinite path.*

Proof. Proposition 5.2.11 states that every bi-infinite path in the Renewal Bridge Graph is bi-recurrent (for the definition, see Theorem 4.3.5). So, if there is a bi-infinite path in the Renewal Bridge Graph, some vertices in this bi-infinite path have infinitely many descendants in the time axis. It means that correspondingly some vertices in the Renewal EFF have infinitely many descendants, which contradicts the fact that every connected component of EFF is $\mathscr{I}/\mathscr{I}$, as shown by Proposition 5.2.7. □

Remark 5.2.13. It is easy to check that, under the assumption that the gcd of A is 1, the Renewal Bridge Graph satisfies the conditions of Proposition 5.1.4, and this regardless of the infinite mean distribution chosen. Two observations are then in order. Firstly, Proposition 5.2.12 hence also follows from this more general result. Secondly, since there are distributions for which the Renewal Bridge Graph is not connected, this shows that the conditions of Proposition 5.1.4 are weaker then the connectedness assumption.

The following proposition provides a result regarding the unimodularizability of the Renewal Bridge Graph.

Proposition 5.2.14. *The null recurrent Renewal Bridge Graph is not unimodularizable in general.*

Proof. The proof is similar to that of Proposition 5.1.5, which will be presented in Section 5.3. □

In the Renewal Bridge Graph, the function that maps every vertex to its right adjacent vertex, in the next time, is a vertex shift in the sense of [3]. As defined in Section 5.2.1, considering this vertex shift, one can considers its foils in the Bridge Graph.

Proposition 5.2.15. *The foils of a connected components of the Renewal Bridge Graph are its intersections with vertical timelines. There are infinitely many foils in each connected component of the Bridge Graph, and the order of the foils is of type* $\mathbb{Z}$[4].

Proof. It will be shown in Proposition 5.3.7 that the same property holds in the Bridge Graphs constructed by a general null recurrent Markov Chain. So the result is valid for the Renewal Bridge Graph as well. □

Properties of the S-set in the Renewal Bridge Graph

Definition 5.2.16. Consider the intersection of the Bridge Graph with the zero timeline. This set is a random subset of the state space $\mathscr{S}$, referred to as the S**-set**.

In the Renewal Bridge Graph, suppose that vertex $(0, y)$ belongs to the S-set. Then, since the vertices of the Renewal Bridge Graph have a backtrack to a vertex in the time axis, if one goes backward in time, from vertex $(0, y)$, one eventually reaches a vertex in the time axis for the first time. This vertex is denoted by $(t_y^-, 0)$. Also, by continuing the path that passes through the vertex $(0, y)$ forward in time, it will also reach a vertex on the time axis. Denote this vertex by $(t_y^+, 0)$. Note that by the definition of the Renewal Markov Chain, $(t_y^+, 0) = (y, 0)$. So for each vertex $y \neq 0$ in the S-set of the Renewal Bridge Graph, there is a path in the graph that starts from a vertex on the time axis before time zero and returns to the time axis, for the first time, after time zero. Correspondingly, under the coupling (5.2.9), there is an edge in the Renewal EFT that starts from vertex t_y^- before time zero and ends at vertex t_y^+ after time zero. See Figure 5.4.

Definition 5.2.17. In the Renewal EFT (EFF), an edge that starts before zero and ends after zero is called **flying over zero**.

The following propositions give results about the S-set of the Renewal Bridge Graph when it is assumed that the Renewal Bridge Graph is connected.

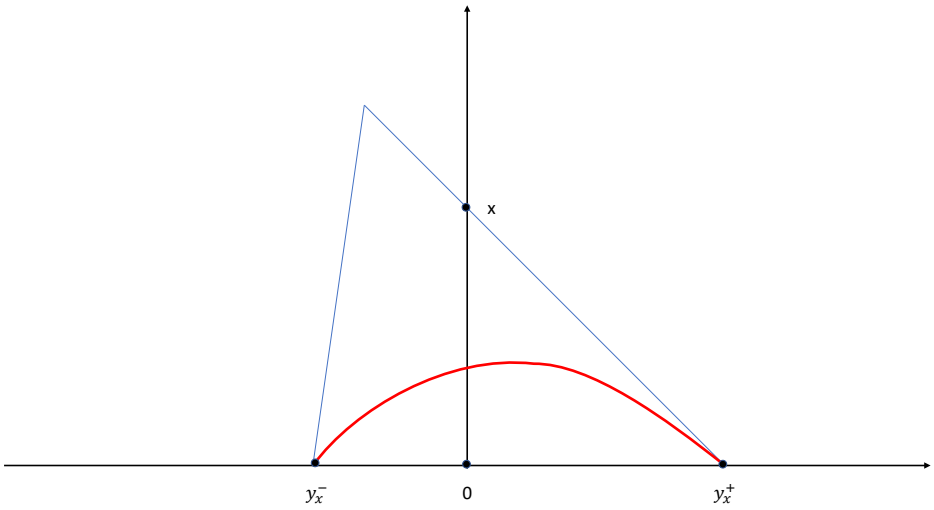
Proposition 5.2.18. *The S-set has an a.s. infinite cardinality.*

Proof. Assume the cardinality of the S-set is a.s finite. Then, in the corresponding Renewal EFT, there are finitely many edges flying over zero. That is, the set

$$F := \{t \in \mathbb{N};\ \text{there is a flying edge over zero in the Renewal EFT with the end vertex } t\}, \tag{5.2.11}$$

[4] See [3] for a precise definition.

Figure 5.4: Correspondence between an edge that flying over zero in the Renewal EFT and a path that started before time zero and ended after time zero in the Renewal Bridge Graph.



is a.s. finite. So in the Renewal EFT, all the vertices before vertex 0 are the descendants of one of the vertices in the F set. So at least one of the vertices in the F set has infinitely many descendants, which is impossible because all trees of the Renewal EFF are $\mathscr{I}/\mathscr{I}$. □

Proposition 5.2.19. *Almost surely, every vertex in the S-set has finitely many *-descendants in the Renewal Bridge Graph.*

Proof. Let $(0, y)$ be a vertex in the S-set. The number of *-descendants of $(0, y)$ is equal to the number of children (the first generation of descendants) of vertex t_y^+ in the Renewal EFF. Each connected component of a fixed vertex in the Renewal EFF is unimodular (as established in Proposition 5.2.5), which implies that the number of children of any vertex within the component containing t_y^+ is almost surely finite (see Proposition 3.11 in [3]). Consequently, the number of *-descendants for any vertex in the S-Set of the Renewal Bridge Graph is also almost surely finite. □

Remark 5.2.20. In the Renewal Bridge Graph, it is known that the connected components of the corresponding Renewal EFF are in the $\mathscr{I}/\mathscr{I}$ class of the Foil Classification Theorem (see Proposition 5.2.7). Using the same arguments as those presented in the proof of Proposition 5.2.19, it can be established that not only the number of *-descendants of the vertices in the S-set is finite, but the total number of their descendants is also finite (see Theorem

3.10 in [3]). This *finite number of descendant property* is not limited to the vertices in the S-set; it holds for all vertices of the Bridge Graph.

5.3 Properties of the Bridge Graph of a General Null Recurrent Markov Chain

This section extends most of the results on the Renewal Bridge Graph to the Bridge Graph of a general irreducible, aperiodic, and null recurrent Markov Chain. Many properties of the Renewal Bridge Graph are concluded from the coupling defined in (5.2.9). The following object, which generalizes the Renewal EFT, will be used for the general Bridge Graph:

Definition 5.3.1. Consider an irreducible, aperiodic and null recurrent Markov Chain $\{X_n\}$, $n \in \mathbb{N}$, and its associated Bridge Graph B_X, with reference vertex, s^*. In this setup the time axis is $\{(t, s^*); t \in \mathbb{Z})$. In B_X, consider the paths starting from a point in the time axis for example (t, s^*), then look at the times when these paths get back to this axis again. Let the random variable $\mathscr{T}_t$ denote the time that it takes for the path starting from vertex (t, s^*) in the Bridge Graph to return to the time axis for the first time. Define $G^{\mathscr{T}} = G^{\mathscr{T}}(V, E)$ be the random graph, where V, the set of the vertices, is the whole $\mathbb{Z}$, and where E, the set of the edges, is defined as follows: each point t has a single outgoing edge, e_t, with length $\mathscr{T}_t$. This random graph is called the **Recurrence Time EFF** of B_X. In the connected case, it is referred to as the Recurrence Time EFT.

Remark 5.3.2. Note that in the Recurrence Time EFF:

- Since the Markov Chain $\{X_n\}_{n\in\mathbb{N}}$ is null recurrent for all t, $\mathbb{E}(\mathscr{T}_t) = \infty$.
- The random variables $\{\mathscr{T}_t\}$ are identically distributed. However, since the paths that start from two different vertices (t_1, s^*) and (t_2, s^*) in the Bridge Graph may meet each other before returning to s^*, the random variables $\{\mathscr{T}_t\}$ are not independent in general, even in the totally independent case defined in Remark 5.1.1.

5.3.1 Properties of the Recurrence Time EFF and the Null Recurrent Bridge Graph

The Recurrence Time EFF has almost the same properties as the Renewal EFF.

Proposition 5.3.3. *Recurrence Time EFF (with ghost edges defined in Proposition 5.2.5) is a unimodular network. Recurrence Time EFF is a unimodular network. When the Recurrence Time EFF is connected, it is in the $\mathscr{I}/\mathscr{I}$ class of the foil classification theorem for unimodular networks.*

Proof. Since the Recurrence Time EFF with ghost edges (see Remark 5.2.6) is constructed by a stationary marking of $\mathbb{Z}$ so it is a unimodular network. (see [2] for more details).

For the $\mathscr{I}/\mathscr{I}$ structure property in the connected case, the same proof as for Proposition 5.2.7 for the connected case works here. □

Remark 5.3.4. The Recurrence Time EFT, $G^{\mathscr{T}}$, is defined as a function of the Bridge Graph. In the Bridge Graph, B_X, the mass transport principle holds for those vertices that belong to $G^{\mathscr{T}}$. This means that the triple $(B_x, A, 0)$ is locally unimodular in the sense of [18], with $A = \{(x, y); y = s^*\}$.

So far, only the properties of Recurrence Time EFFs were discussed. Their implications on general null recurrent Bridge Graphs are now considered.

Proposition 5.3.5. *To every bi-infinite path, $\{\beta_t\}_{t\in\mathbb{Z}}$, in the null recurrent Bridge Graph B_X, one can associate a bi-infinite path in its Recurrence Time EFF.*

Proof. The proof of this proposition is almost the same as Proposition 5.2.11 in the Renewal case. It is first shown that there is no bi-infinite path in the null recurrent Bridge Graph that intersects the time axis for the first time at time T, where T is an arbitrary time in the time axis. Define the same set D, as in the renewal case:

$D =\{T-t$; the path that starts from $(T-t, s^*)$ in the null recurrent Bridge Graph goes back to the time axis for the first time at time $T\}$.

Note that the definition of D gives that the vertices $(T - t, s^*)$, in B_X, that belong to D, are, in its corresponding recurrence time EFF, the first generation descendants of T. The unimodularity of the Recurrence Time EFF implies that a.s. all vertices have a finite degree. So the cardinality of D is a.s. finite.
If there is a bi-infinite path, $\{\beta_t\}_{t\in\mathbb{Z}}$, in B_X, which is not bi-recurrent, then there is a time T such that $\{\beta_t\}_{t\leq T}$ does not intersect the time axis. Since $\{\beta_t\}_{t\leq T}$ is a path in the Bridge Graph, every vertex in it has a backtrack to the time axis. It means there exist infinitely many paths starting from a vertex $(t, s^*), t < T$, in the time axis and entering $\{\beta_t\}_{t\leq T}$, i.e., the starting vertices of these paths belong to the set D. This contradicts the fact that D is a.s. finite for any arbitrary T. So such a path does not exist. □

Proof of Proposition 5.1.2. Since the Renewal Bridge Graph is an example of a null recurrent Bridge Graph, in the general case also the null recurrent Bridge Graph can be either connected or not connected. □

Proof of Proposition 5.1.3. Proposition 5.3.5 gives that every bi-infinite path in the null recurrent Bridge Graph is bi-recurrent. So if there exists any bi-infinite path in the Bridge

Graph, there is a bi-infinite path in the Recurrence Time EFT. Moreover, since the Recurrence Time EFT is $\mathscr{I}/\mathscr{I}$, there is no bi-infinite path in the null recurrent Bridge Graph. □

Proof of Proposition 5.1.4. Proposition 5.3.5 gives that to every bi-infinite path in B_X, one can associate a bi-infinite path in the associated Recurrence Time EFF.
First, suppose that there is only one bi-infinite path in the Recurrence Time EFF. This bi-infinite path is a covariant subset, and since the Recurrence Time EFF is unimodular, by the same argument as in the proof of Proposition 5.2.7, one can show that this is not possible. So either there is no bi-infinite path in the graph, or there are more than one bi-infinite path. In order to show that the further case is impossible, consider the following set:

$$M = \{(0, s) \in B_X; s \text{ is in a bi-infinite path}\}. \tag{5.3.1}$$

Note that, the set M is determined by the property of the Bridge Graph before time zero. Suppose that $(0, s_1)$ and $(0, s_2)$ are two separate arbitrary elements in M. By assumption, with positive probability, the trajectories starting from these two vertices meet each other in the future, in the Bridge Graph. It means that, correspondingly, in the Recurrence Time EFF, two bi-infinite paths meet each other with positive probability. I.e., with positive probability, there is a connected component in the Recurrence Time EFF that has two bi-infinite paths, which is impossible due to the foil classification theorem of unimodular networks. So there is no bi-infinite path in the null recurrent Bridge Graph. □

Remark 5.3.6. Note that:

1. From the proof of Proposition 5.1.4, one can conclude that if the Recurrence Time EFF has this property that every two vertices in this graph meet each other with positive probability, then every connected component in this graph is $\mathscr{I}/\mathscr{I}$.

2. The assumption of Proposition 5.1.4 does not put any condition on the MC itself, but it puts a condition on the coupling that exists between the driving sequence $\{\xi_t^y\}_{y\in S}$ when t is fixed. In particular, in the case where $\{\xi_t^y\}_{y\in S}$ is totally independent, this condition is satisfied. This condition does not hold is Example 4.2.1, where the random variables $\{\xi_t^y\}_{y\in S}$, in this case, are maximally coupled for a fixed t. In this example, there are infinitely many bi-infinite paths in the Bridge Graph.

3. The Renewal Bridge Graph studied in Section 5.2.3 is an example where the driving sequence is totally independent.

Proposition 5.3.7. *The foils of a connected component of the null recurrent Bridge Graph, B_X, are the intersections of this component of the graph with vertical timelines. Thus the null recurrent Bridge Graph has infinitely many foils, and the order of foils is that of $\mathbb{Z}$.*

Proof. Consider the vertex shift f in the Bridge graph B_X, where f maps each vertex (x, y) to its right adjacent vertex. For each $t \in \mathbb{Z}$, let

$$S_t = \{(x, y) \in B_X; x = t\}.$$

Suppose that for some $t \in \mathbb{Z}$, (t, y_1) and (t, y_2) are in S_t and in the same connected component. Since they are in the same connected component, there is a vertex $(t_0, y_0) \in B_X$ such that the trajectories of these two vertices meet each other. By definition of the Bridge Graph, the number of steps that it takes for vertices (t, y_1) and (t, y_2) to reach (t_0, y_0) is equal to $t_0 - t$, so these two vertices belong to the same foil.
For the converse, note that if two vertices (t_1, y_1) and $t_2, y_2)$ are in the same foil, then the trajectories of these two vertices meet each other after the same number of steps. Since by definition of the Bridge Graph and the vertex shift f, at each step, the trajectories move exactly one unit forward in time, it follows that t_1 and t_2 are equal. □

Proof of Proposition 5.1.5. Consider the Bridge Graph as a network, i.e., a graph with marks on its vertices and edges. Suppose that the Bridge Graph, B_X, is a unimodularizable network. In the general case, as shown in Proposition 5.3.7, in each connected component, the intersections of the Bridge Graph with vertical lines are the foils, and the foils are a.s. infinite. The foils form a covariant vertex partition of the Bridge Graph in the sense of [3]. Moreover, the time axis is a covariant subset of the Bridge Graph if vertices are marked by their coordinates with respect to the time and state axis.
Using the no infinite/finite inclusion lemma in [3] for unimodular networks, the intersection of a foil with the time axis should also be infinite, because each foil is almost surely infinite. However, in the Bridge Graph, this intersection has only one element. So the Bridge Graph is not a unimodularizable network in the sense of Definition **??**. □

Remark 5.3.8. When the Bridge Graph is considered as a graph without any marks with respect to the time axis and the state axis coordinates, it is not unimodularizable, in general as well. Indeed, this is the case when the time axis is distinguishable in the Bridge Graph as a covariant subset and hence the same proof as Proposition 5.1.5 holds. For example, in the Renewal Bridge Graph, the set R is a distinguishable set in the whole graph, so this graph is not unimodularizable. However, considering the Bridge Graph without its marks can also lead to situations where it is unimodularizable. For instance, the maximally coupled Bridge Graph of the lazy random walk in Example 4.2.1 is a unimodular Bridge Graph rooted at $(0, 0)$.

The S-set of the general null recurrent Bridge Graph has the same properties as in the Renewal Bridge Graph.

Proposition 5.3.9. *Under the assumptions of Proposition 5.1.3 (or Proposition 5.1.4), the cardinality of the S-set of a null recurrent Bridge Graph is a.s. infinite.*

Proof. Proposition 5.3.3 states that the Recurrence time EFT of a null recurrent Bridge Graph is $\mathscr{I}/\mathscr{I}$. Also, the proof of Proposition 5.1.4 shows that under its assumptions, the connected components of the Recurrence time EFT is $\mathscr{I}/\mathscr{I}$. So the same proof as in the Renewal Bridge Graph in Proposition 5.2.18 holds here. □

Proposition 5.3.10. *Every vertex in the S-set of a null recurrent Bridge Graph has finitely many *-descendants a.s.*

Proof. The proof of this property in the Renewal Bridge Graph is solely based on the unimodularity of the connected components of the corresponding Renewal EFF. In the case of a general null recurrent Bridge Graph (the Bridge Graph of an aperiodic, irreducible, and null recurrent Markov Chain), Proposition 5.3.3 shows that the connected component of its corresponding recurrence time EFF is unimodular. Consequently, the same argument used in the proof of the Renewal Bridge Graph holds here. □

Chapter

6 Dynamics on the Doeblin Graph

Le Chapitre 6 définit la dynamique sur le graphe de pont et étudie les propriétés de cette dynamique aléatoire. La Section 6.1 examine les dynamiques tabou et Potential sur le graphe de pont et étudie leurs propriétés. La dynamique tabou PP est introduite dans cette section et est fortement liée à la mesure invariante unique de la chaîne de Markov à récurrence nulle (Théorème 5.2.3). La constructibilité des dynamiques tabou et potential PPs est également discutée. Ces mesures aléatoires ont un support infini, et en raison de leur définition sur le graphe de pont, il devient évident que l'ensemble de la mesure repose sur une information infinie du graphe de pont. Par conséquent, ces mesures ne sont pas globalement constructibles. Cependant, il sera démontré qu'elles sont localement finiment constructibles, ce qui signifie que la masse de la mesure en chaque point dépend seulement d'un sous-graphe presque certainement fini du graphe de pont. Néanmoins, cela n'implique pas une constructibilité algorithmique universelle ni la détermination pratique de ce sous-graphe fini dans tous les cas. Cela signifie qu'il existe des scénarios où trouver un temps d'arrêt dans le graphe de pont lié à la masse d'un point dans ces mesures n'est pas toujours possible. Cependant, cette constructibilité algorithmique est vérifiée dans les cas où $\{X_t\}_{t\in\mathbb{N}}$ est *monotone*, permettant un échantillonnage parfait des dynamiques tabou et potential PPs. Pour fournir une intuition supplémentaire, dans les cas où les chemins construisant le graphe de pont sont couplés de manière monotone, la propriété de monotonie assure que les points dans les supports des dynamiques tabou PP (ou potential PP) apparaissent dans le graphe de pont dans un ordre spécifique (du petit au grand). Lorsqu'un nouveau point de la mesure est observé dans le graphe de pont, cela indique que les masses des points plus petits ont été entièrement construites. Un exemple concret lié à la théorie des files d'attente est également discuté en détail dans la Section 6.2. La file d'attente $GI/GI/1$ permet d'illustrer le sens et l'intérêt pratique de ces deux processus ponctuels.

La Section 6.3 examine les propriétés des deux processus ponctuels lorsque la chaîne de Markov est récurrente positive. Pour la dynamique tabou PP, la connexion entre l'échantillonnage parfait dans le sens de la CFTP et celui obtenu en utilisant la définition du TPP est discutée. Les propriétés de la dynamique potential PP dans le cas récurrent positif sont également considérées. Cette section donne également quelques résultats sur les propriétés de ces deux dynamiques dans des espaces d'états plus généraux. Au lieu de considérer ces dynamiques sur le graphe de pont, elles sont considérées comme des chaînes de Markov sur l'espace des mesures aléatoires sur S.

Chapter content

6.1 H^T, and H^P on the Null Recurrent Bridge Graph

This section considers the two dynamics, H^T and H^P, defined in Section 5.1 as dynamics on the measures on the state space of the null recurrent Bridge Graph, and studies their properties.

6.1.1 Taboo Dynamics and Its Relation with the Invariant Measure of Null Recurrent MCs

The first dynamics is the Taboo Dynamics, H^T, defined in (5.1.1). Theorem 5.1.8 states that this dynamics has at least one steady state.

Proof of Theorem 5.1.8. Consider the Taboo PP as the initial state of the Taboo Dynamics. By the definition of this dynamics, at time one, there is mass one at state s^*. Moreover, the

mass of any arbitrary state $y \neq s^*$ is

$$M_1^T(y) = \sum_{x \in \mathscr{S}, x \neq s^*} \tau_0(x) \mathbb{1}_{\{h(x,\xi_0^x)=y\}} + \mathbb{1}_{\{h(s^*,\xi_t^{s^*})=y\}}.$$

Since for all $x \neq s^*$, $\tau_0(x)$ is the number of *-descendant of x which are such that the first return to s^* of the path starting from them takes place after time zero, $M_1^T(y)$ is also the number of *-descendants with the same property, that is $\tau_1(y) = M_1^T(y)$. It is clear from the Bridge Graph construction that $\tau_1 \overset{d}{=} \tau_0$. So the Taboo PP is a stationary state of the Taboo dynamics. □

In the positive recurrent case, there is a known relation between the Bridge Graph and the stationary distribution of the Markov chain. In this last case, the Bridge Graph contains a unique bi-infinite path (see Lemma 4.3.4). So there is a point in the S-set with infinitely many descendants. It is proven in [4] that this point is a perfect sample of the stationary distribution of the Markov Chain. On the other hand, there is no bi-infinite path in the null recurrent Bridge Graph. All the points have finitely many descendants, and the approach of the positive recurrent case does not work. Instead, in the null recurrent case, the finite Taboo multiplicity defined in Definition 5.1.6 can be defined for the vertices on the S-set. Theorem 5.1.9 establishes a connection between this Taboo PP and the stationary measure of the null recurrent Markov Chain.
Before going to the proof of Theorem 5.1.9, here is a classical lemma about computing stationary measure of the MC, using taboo probabilities.

Lemma 6.1.1. *Let X_n be an irreducible and recurrent Markov Chain. For fixed i in the state space, let ζ be defined by*

$$\zeta_j = \sum_{n=0}^{\infty} q_{ij}^n, \quad j \neq i, \tag{6.1.1}$$

where q_{ij}^n is the probability for going from i to j, after n steps, without visiting i, and $\zeta_i = 1$. Then ζ is a positive invariant measure for the chain. This invariant measure is unique up to multiplication by a constant.

Proof of Theorem 5.1.9. Let $G^{\mathscr{T}}$ be the Recurrence Time EFF associated with B_X, and Q_t be the path that starts from (t, s^*) in B_X, forward in time, where s^* is the reference point of the Bridge Graph.
Consider an arbitrary state y in the state space of the MC. Define $g[G^{\mathscr{T}}, t, t'] = 1$, when Q_t passes through vertex (t', y) before returning to s^*. The mass transport principle states

that, for each y in the state space S, the following equation holds:

$$\mathbb{E}[\sum_{t\in\mathbb{Z}} g[G^{\mathscr{T}}, t, 0]] = \mathbb{E}[\sum_{t\in\mathbb{Z}} g[G^{\mathscr{T}}, 0, t]]. \tag{6.1.2}$$

The right-hand side of (6.1.2) is the expectation of the number of times that path Q_0 intersects the state y before returning to s^*, whereas the left-hand side is equal to the expectation of the mass that the Taboo PP puts at point y, at time 0. Since the right-hand side of the equation does not dependent on the coupling of $\{\xi_t^y\}_{y\in S}$ for a given t, the same is true for the left hand side. So the expectation of the Taboo point process at each point in the state space does not dependent on the coupling of $\{\xi_t^y\}_{t\in\mathbb{Z},y\in S}$, for fixed t. Moreover, using Lemma 6.1.1, the right-hand side of (6.1.2) is equal to the invariant measure of the MC, and this shows that Equation (5.1.4) holds for the mean measure of the Taboo PP. □

Remark 6.1.2. Notice that, by Definition 5.1.6, the point s^* always belongs to the S-set. Moreover, the mass that the Taboo PP puts at this point is equal to 1 at all times.

On the Constructibility of the Taboo Point Process

Let vertex $(y, 0)$ belongs to the S-set, the support of the Taboo PP. Then due to Proposition 5.3.10, $\tau_0(y)$, is a.s. finite. In this sense, one can say that the Taboo PP is locally finitely constructible, i.e., the taboo multiplicity at each vertex of the S-set depends on a finite subtree of the Bridge Graph.
The important question here is whether the Taboo multiplicities are algorithmically constructible. A positive answer to this question is equivalent to the possibility of producing a perfect sample of this point process. In general, the answer to this question is unknown. However, there are some results for the special case, where the MC is stochastically monotone. These results are given in the following subsection.

Stochastically Monotone Markov Chain

Here it is shown that in the case where the transition probabilities of the Markov chain are stochastically monotone, the Taboo point process is algorithmically locally constructible. The following definition is borrowed from [27].

Definition 6.1.3. Assume that the state space of the Markov chain $\{X_n\}_{n\in\mathbb{Z}}$, $\mathscr{S}$, is endowed with a partial order denoted by $\leq$. The Markov chain is stochastically monotone if its transition probabilities have the stochastic monotone property. i.e., the probability measures $(P(x, .); x \in \mathscr{S})$, on $\mathscr{S}$ are such that $P(x', .) \preceq P(x, .)$ whenever $x' \leq x$, whereby $\preceq$ means stochastically less than or equal to.

Proposition 6.1.4. *Assume $X = \{X_n\}_{n\in\mathbb{Z}}$ is a stochastically monotone Markov Chain on $\mathscr{S}$, and $\mathscr{S}$ has the minimum element s_0. Moreover, suppose that the Bridge Graph B_X is constructed with the reference point $s^* = s_0$. Then the TPP is algorithmically locally constructible.*

Proof. By a classical coupling argument, there exists a coupling $\xi_t(.)$ for the driving sequence $\{\xi_t^x\}$, at each time t, such that if $y \le z$, then $\xi_t(y) \le \xi_t(z)$ (See [21] and [28]).
Denote the path that starts from $(t_0, s^*), t \in \mathbb{Z}$, by $\mathscr{P}^{t_0} = \{\mathscr{P}_t^{t_0}, t \ge t_0\}$. Note that if $t_0 < t_1$, then

$$\{\mathscr{P}_{t_1}^{t_0}\} \ge s_0 = \{\mathscr{P}_{t_1}^{t_1}\},$$

and hence, with the monotone coupling argument, the path $\mathscr{P}^{t_0}$, at each time, remains larger than or equal to the path $\mathscr{P}^{t_1}$.
For constructing the Taboo PP, first start trajectories from time -1 and step by step add trajectories. Due to the monotone coupling and the later argument, each new trajectory that is added is larger than or coalesces with the pointwise supremum of the trajectories considered before. So, when a new trajectory is added to the Bridge Graph, the image of this trajectory might belong to one of the following scenarios:

Adds no mass to the S-set,

Adds mass to the last point added before,

Creates a new point in the S-set, which is greater than all the points that have been built in the S-set before.

With this observation, once a new point appears in the S-set, the mass of the Taboo PP for the points that are less than or equal to this new point is fully determined. □

Remark 6.1.5. Consider the random walk defined on the state space $\mathbb{N}$ with the following transition probabilities: for $n = 0$, the walk stays at 0, with probability $1/2$, and it moves to state 1 with probability $1/2$. For state $n \in \mathbb{N}, n \neq 0$, the walk moves to each neighbor of n, i.e., $n+1$ and $n-1$ with probability $1/2$. This Markov chain is an example that satisfies the assumption of Proposition 6.1.4.

With this algorithm described above, when a new path is added to the Bridge graph from time t, it might add a new mass to the S-set or not. So computing the expectation of the time until a new mass is added to the S-set gives information about the time it takes to construct the taboo multiplicities locally. This expectation is computed in the following corollary.

Lemma 6.1.6. *Consider the assumptions of Proposition 6.1.4 and the algorithm that is presented in its proof. If $\mathscr{P}^{t_1}$ is a path that adds mass to the S-set, then for all $t < t_1$, $\mathscr{P}^t$ does not intersect the time axis (the state s^* of the state space) between times t_1 and 0.*

Proof. Consider the path $\mathscr{P}^{t_0}$, with $t_0 < t_1$. For all $t > t_1$, $\mathscr{P}^{t_0}_t \geq \mathscr{P}^{t_1}_t$. Since the path $\mathscr{P}^{t_1}$ adds mass to the S-set; it does not intersect the time axis up to time zero, so the same holds for the path $\mathscr{P}^{t_0}$. □

Corollary 6.1.7. *Under the assumptions of Lemma 6.1.6, Suppose that t_1 and $t_1 - T$ are two successive times such that the paths $\mathscr{P}^{t_1}$, and $\mathscr{P}^{t_1-T}$ add mass to the S-set, then*

$$\mathbb{E}[T] = \infty. \tag{6.1.3}$$

Proof. Since $t_1 - T < t_1$, Lemma 6.1.6 gives that for all $t < t_1 - T$, the path $\mathscr{P}^t$ does not intersect the time axis at time t_1. So vertex (t_1, s^*) does not have any descendants before time $t_1 - T$ in B_X. Consequently, in the Recurrence Time EFT, the vertex t_1 does not have descendants before vertex $t_1 - T$. So,

$$\mathbb{E}[T] > \mathbb{E}[\text{Number of descendants of } t_1 \text{ in the Recurrence Time EFT}]. \tag{6.1.4}$$

Note that the backward construction of the Bridge Graph, as mentioned before, starts from time zero and explores the graph step by step in the past. When it is constructed up to time t_1, the left-hand side of t_1, in the Bridge Graph, is not explored yet. So the distribution of the Bridge Graph before time t_1 is the same as the original distribution of the Bridge Graph. The original distribution in the Recurrence Time EFT, is such that the expectation of the number of the $*$-descendants of any vertex is infinite. So the right-hand side of Equation (6.1.4) is infinite, and thus $\mathbb{E}[T] = \infty$. □

6.1.2 Potential Dynamics

The second dynamics defined on the null recurrent Bridge Graph is the Potential Dynamics. Theorem 5.1.11 states that this dynamics also has at least one steady state.

Proof of Theorem 5.1.11. Consider the Potential PP as the initial state of the Potential Dynamics, where the dynamics is associated to a null recurrent MC. Let $y \neq s^*$ be a state in the state space $\mathscr{S}$. The mass that the Potential Dynamics puts at y at time 1, is equal to

$$M_1^P(y) = \sum_{x \in \mathscr{S}} \pi_0(x) \mathbb{1}_{\{h(x,\xi_0^x)=y\}}. \tag{6.1.5}$$

That is, it is obtained by adding all the masses that enter the state y, via the Bridge Graph's edges, from time 0. Since for each $x \in \mathscr{S}$, $\pi_0(x)$ is equal to the number of its $*$-descendants, $M_1^P(y)$ is equal to the number of $*$-descendants of vertex $(1, y)$ in the Bridge Graph, which is the potential multiplicity of this vertex. The same argument shows that the

result holds for $y = s^*$, with this difference that s^* itself should be counted once. So the Potential PP is a steady-state of the Potential Dynamics. □

Theorem 5.1.12 shows the connection between the steady state of the Potential Dynamics and the null recurrent MC associated with it. This connection is related to the potential matrix R of the MC with entries R_{xy}, where R_{xy} is the expected number of visiting of the state y given that the initial state of the MC is the state x. For a recurrent MC, the entries of the potential matrix are all equal to infinity.

Proof of Theorem 5.1.12. Let $G^{\mathscr{T}}$ be the Recurrence Time EFT (EFF) associated with B_X, and Q_t the path that starts from (t, s^*) in B_X, where s^* is the reference point of the Bridge Graph.
Consider an arbitrary state y in the state space of the MC. Define $g[G^{\mathscr{T}}, t, t'] = 1$, when Q_t passes through the vertex (t', y). Using the mass transport principle, for each y in the state space S the following equation holds:

$$\mathbb{E}[\sum_{t\in\mathbb{Z}} g[G^{\mathscr{T}}, t, 0]] = \mathbb{E}[\sum_{t\in\mathbb{Z}} g[G^{\mathscr{T}}, 0, t]]. \tag{6.1.6}$$

The right-hand side of the (6.1.6) is the expectation of the number of times that path Q_0 intersects the state y, which its expectation is equal to R_{s^*y}. On the other hand, the left-hand side of (6.1.6) is equal to the expectation of the mass that the Potential PP puts at point y, at time 0. Since R_{s^*y} is related to a recurrent MC, the mean measure of the Potential PP at each point is infinity. □

Consider the null recurrent Bridge Graph. Proposition 5.3.10 states that in this case, for each $(0, y)$ in the S-set, $\pi_0(y)$ is a.s. finite. So one can consider the algorithmic constructibility of the Potential PP as the Taboo PP. The same result as Proposition 6.1.4 is valid for the Potential PP.

Proposition 6.1.8. *Assume $X = \{X_n\}_{n\in\mathbb{Z}}$ is a stochastically monotone Markov Chain on $\mathscr{S}$, and $\mathscr{S}$ has the minimum element s_0. Moreover, suppose that the Bridge Graph B_X is constructed with the reference point $s^* = s_0$. Then the Potential PP is algorithmically locally constructible.*

Proof. Consider the backward construction algorithm for the Bridge Graph with the monotone coupling introduced in the proof of Proposition 6.1.4. The same argument as in the proof of Proposition 6.1.4 shows that, once a new point appears in the S-set, the mass of the Potential PP for the points that are less than or equal to this new point is fully determined. So the claim is proved. □

Remark 6.1.9. Let X satisfy the assumptions of Proposition 6.1.8, and B_X be its associated Bridge Graph constructed by the monotonically coupled driving sequence. Although the same step-by-step backward construction algorithm for the Bridge Graph gives the potential multiplicities of the point in the S-set, one can consider a faster algorithm for constructing the potential multiplicities.
For showing this, let (t_0, s^*) and (t_1, s^*), where $t_0 < t_1$, be two vertices in the time axis that add mass to the same vertex $(0, y)$ in the S-set when the potential multiplicity is considered. Then all the vertices (t_i, s^*), where $t_0 < t_i < t_1$, add mass to the potential multiplicity of $(0, y)$. So, for finding the Potential multiplicity of $(0, y)$, it is sufficient to find the first time and the last time that add mass to $(0, y)$. So instead of constructing the Bridge Graph step by step, one can use the *exponential search* algorithm (see [29]) to find the last time that adds mass to vertex $(0, y)$. Then use *binary search* for finding the first time. If T is the position of the search time, then exponential search takes $O(logT)$ time to find T. So this new algorithm for finding the Potential multiplicity is faster than the step-by-step construction.

6.2 Perfect Sampling of the Taboo and Potential PPs in the Critical Single Server Queue

6.2.1 Loynes' Theory

This section is focused on the application of the results of the previous sections to the $GI/GI/1$ queue and its associated workload Markov Chain [30]. The service times $\{\varsigma_n; n \in \mathbb{Z}\}$ are assumed to be i.i.d. with finite and nonzero mean $\mathbb{E}[\varsigma]$. The inter-arrival times are i.i.d. and denoted by $\{\upsilon_n; n \in \mathbb{Z}\}$, with finite and nonzero mean $\mathbb{E}[\upsilon]$. That is, $\upsilon_n = T_n - T_{n-1}$, where T_n is the arrival time of the n-th customer. Let $W_n = W(T_n^-)$ denote the workload just before time T_n, which is the amount of service remaining to be done by the server at that time. This workload process satisfies the equation

$$W_{n+1} = (W_n + \varsigma_n - \upsilon_n)^+, \quad \forall n \in \mathbb{Z}, \tag{6.2.1}$$

where $(a)^+ = \max(a, 0)$. Because of the i.i.d. assumptions, (6.2.1) defines an $\mathbb{N}$-valued Markov Chain. To avoid degenerate cases, it is assumed below that the variance of $\varsigma - \upsilon$ is non zero and that $\{\varsigma_n\}_n$ and $\{\upsilon_n\}_n$ are independent, although these assumptions are not essential. The *traffic intensity* is $\rho = \frac{\mathbb{E}[\varsigma]}{\mathbb{E}[\upsilon]}$. It is well known that when $\rho < 1$, this Markov Chain is positive recurrent and when $\rho > 1$, it is transient. In the critical case, when $\rho = 1$, it is null recurrent.
In the case where $\rho < 1$, Loynes' theory allows one to define a perfect sample from the stationary distribution [30]. This section extends this theory to the perfect sampling of the

Taboo PP using the algorithm that is provided in Subsection 6.1.1, which applies since this Markov Chain is stochastically monotone.

The Loynes variable at time $n \geq 0$, L_n, is the value of the workload at time 0 when starting the queue empty at time $-n$, and when coupling the service and inter-arrival times as in the CFTP algorithm, namely

$$L_n = \left(\max_{k=1,\dots,n} \sum_{l=-n}^{-1} (\varsigma_l - \upsilon_l) \right)^+ .$$

The sequence $\{L_n\}_n$ is non-decreasing.

6.2.2 Interpretation and Perfect Sampling of the Taboo PP

It is easy to see that M_0^T, the Taboo PP of this Markov Chain, is simple, and that its support is $\{L_n\}_{n\geq 0}$. Indeed, adding customers $-n$ in the past leads to a new atom in this PP if and only if this addition creates a busy period that starts at the arrival of customer $-n$ and lasts until time 0. The Taboo PP at time 0 is hence equal to

$$M_0^T = \delta_0 + \sum_{n\geq 1} \delta_{L_n} 1\{L_n > L_{n-1}\}. \tag{6.2.2}$$

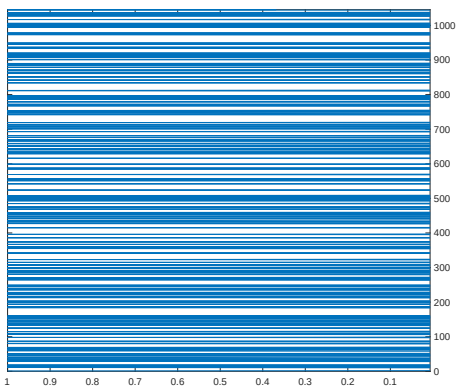
It captures the *joint structure of the workload strict records* in this Loynes type (or CFTP) construction.

In the positive recurrent case, the Loynes sequence converges to an a.s. finite limit L_∞; the Taboo PP is then a.s. finite and the supremum value of its support is the perfect sample L_∞ of the steady state workload. On the other hand, in the null recurrent case, when the condition described in Proposition 5.1.4 holds, the Taboo PP has an infinite support but is locally finite.

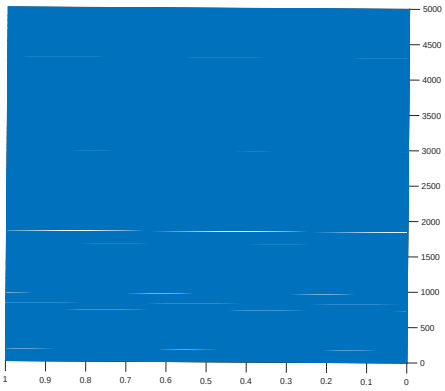
Proposition 6.2.1. *Consider the Bridge Graph of the workload process defined in* (6.2.1)*. If the support of the random variable $\varsigma_n - \upsilon_n$ is unbounded from below, then the condition described in Proposition 5.1.4 holds, that is, for all (t_1, s_1) and (t_2, s_2) in the Bridge Graph, the paths passing through these points intersect with positive probability in finite time.*

Proof. Let $t_1 \leq t_2$, and let $\mathscr{P}_1$ and $\mathscr{P}_2$ be the paths in the Bridge Graph passing through (t_1, s_1) and (t_2, s_2). Suppose that $\mathscr{P}_1$ reach the point $(t_2, s_{1,2})$ at time t_2. Since there is a positive probability that $\varsigma_{t_2} - \upsilon_{t_2} \leq -\max\{s_2, s_{1,2}\}$, it follows that there is a positive probability that at time $t_2 + 1$ the paths $\mathscr{P}_1$ and $\mathscr{P}_2$ meet. □

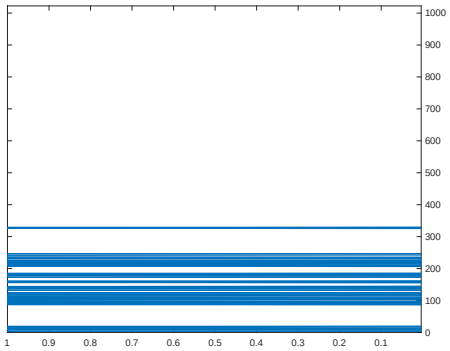
In the simulations provided in this section, it is important to note that the condition of Proposition 6.2.1 holds.



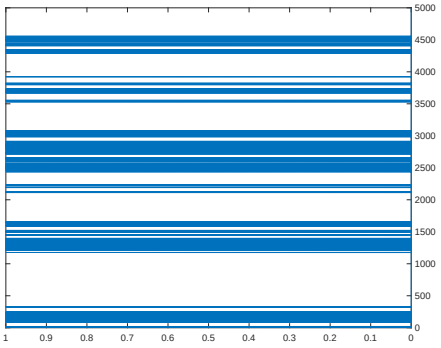
(a) Geo(0.2), $0 \leq i \leq 1000$



(b) Geo(0.2), $0 \leq i \leq 5000$



(c) Zeta(2.5), $0 \leq i \leq 1000$



(d) Zeta(2.5), $0 \leq i \leq 5000$

Figure 6.1: Perfect samples of M^T in the null recurrent case for two different distributions of inter-arrival and service times, for two different intervals of the state space.

Figure 6.1 provides perfect samples of the random measure M_0^T restricted to bounded intervals for different inter-arrival and service time distributions. The samples are perfect because the chain is stochastically monotone.

A corollary of the result on the first moment measure of the Taboo PP is that, in the recurrent case, the invariant measure of this Markov Chain admits the representation

$$\sigma(i) = \mathbb{E}M_0^T(i) = \sum_{n \geq 1} \mathbb{P}[L_n = i, L_n > L_{n-1}], \quad \forall i > 0 \tag{6.2.3}$$

and $\sigma(0) = 1$. The terms in this expression are reminiscent of ladder epochs and heights, but differ from those in that they bear on the backward rather than the forward workload sequence. One can analyze various properties of the Taboo PP using these perfect samples. As Equation (6.2.3) shows, the first moment measure of M_0^T, i.e., its intensity measure, is the invariant measure of the workload Markov Chain. Moment measures of order 2 of M_0^T give some information about the interaction between the points. Using an idea similar to

that of the Ripley K-function in point process theory, one can detect clustering or inhibition in M_0^T by comparing this function to 1, see [10]. For this, consider the following local second-order moment-based function:

$$K_i(r) = \frac{\mathbb{E}[M_0^T(i-r, i+r)|M_0^T(i) = 1]}{\mathbb{E}[M_0^T(i-r, i+r)]}.$$

This function can be estimated using perfect samples of M_0^T. If the points were distributed independently, for all i, we would have $K_i(r) = 1$; this value is used as a benchmark. If $K_i(r) > 1$, there is clustering at point i, whereas if $K_i(r) < 1$, there is inhibition at point i for radius r. Figure 6.2 shows estimates of $K_i(r)$ based on a large collection of perfect samples in some examples of critical queues. As the figures show, there is no general conclusion about clustering or inhibition of M_0^T in this monotone case. The analyzed examples suggest that when inter-arrival and service time variances are finite, there is inhibition for small r, and the value of $K(r)$ tends to 1 for large r (Figures 6.2(a) and 6.2(b)). In contrast, when inter-arrival and service time variances are infinite, there is clustering for small r (Figures 6.2(c) and 6.2(d)).

6.2.3 Interpretation and Perfect Sampling of the Potential PP

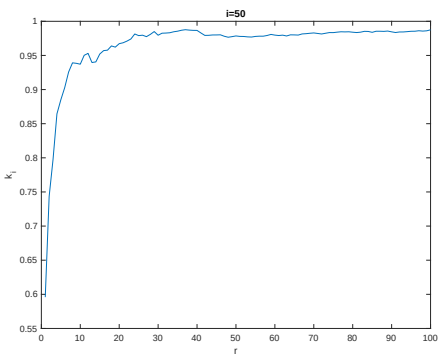
The potential PP has the same support as the Taboo PP, but different multiplicities. It is easy to see that if i is an atom of the Taboo PP, the multiplicity of atom i in the Potential P.P. is the number of epochs that separate in the backward construction the inclusion of atom i in the Taboo PP, due to an increase of the Loynes variable, from its last increase (and atom inclusion). That is

$$M_0^P = \sum_{n\geq 0} \delta_{L_n} 1_{L_n > \max_{0\leq k\leq n-1} L_k} \left(\sum_{k\leq n} 1_{L_n = L_k} \right). \tag{6.2.4}$$

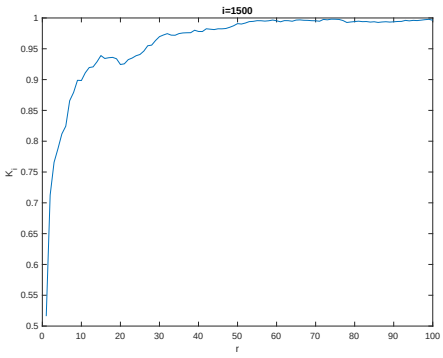
It follows from our general results that, in the null recurrent case, when the Bridge Graph has no bi-infinite paths, which holds under the condition of Proposition 6.2.1, this random measure is a.s. locally finite, though with an infinite first moment measure, whereas it is not locally finite in the positive recurrent case. In other words, this Potential PP gives the *joint time-space structure of the records in Loynes' construction*, with the support of this PP describing the spatial organization of the backward records, as for the Taboo case, and the multiplicities describing their time separation.

It is important to note that, in the simulations provided in this section, the condition of Proposition 6.2.1 holds, which implies that Proposition 5.1.4 also holds. As a result, there are no bi-infinite paths in the Bridge Graph, and the Potential PP is locally finite.

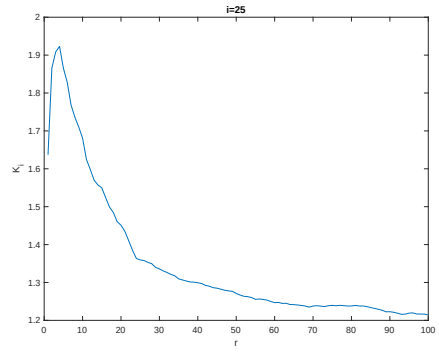
Figure 6.3 gives a perfect sample of an instance of Potential PP at different scales. This



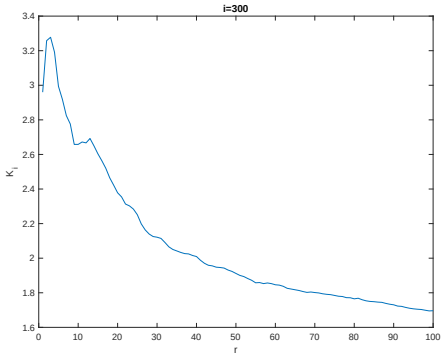
(a) $\varsigma, \upsilon \sim Geo(0.2)$



(b) $\varsigma, \upsilon \sim Poi(25)$



(c) $\varsigma, \upsilon \sim Zeta(2.5)$



(d) $\varsigma, \upsilon \sim Zeta(2.5)$

Figure 6.2: Estimation of $K_i(r)$ with 1000 samples of M_0^T, for two different values of i, and $0 < r \leq 100$.

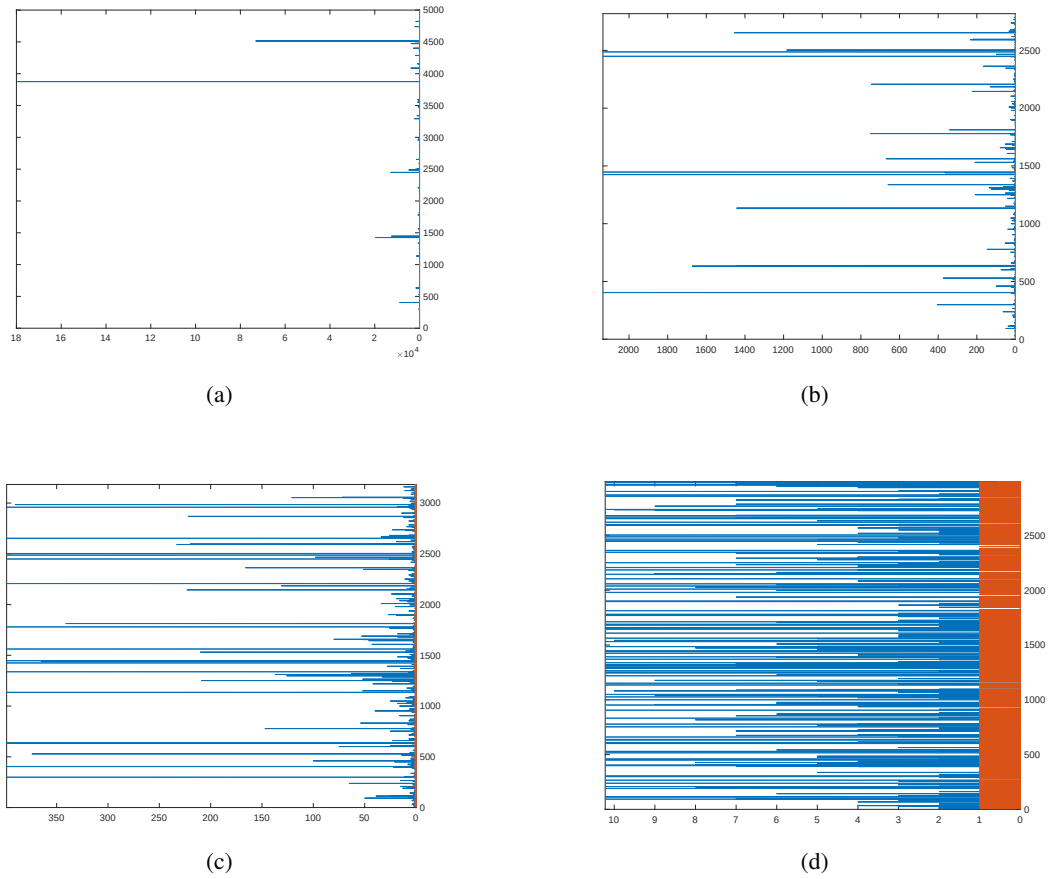


Figure 6.3: Perfect samples of M_0^P for Poisson-distributed inter-arrival and service time, for different scales, the red segments represent the Taboo PP.

point process inherits the complex "correlation" structure of the Taboo PP through their common support. The fact that it has an infinite intensity measure means that, in addition, all its multiplicities are heavy-tailed. These last two properties together with the CFTP space-time interpretation discussed above contribute to making this Potential point process a fascinating object.

The interpretation of the Potential PP survives in the positive recurrent case, with the caveat that it is not locally finite. In fact, in this particular case, all atoms except the largest one have a finite multiplicity, with the same time separation interpretation as above. However, the largest atom, namely L_∞ has an infinite multiplicity as it belongs to the bi-infinite path. It is easy to check that the expressions (6.2.2) and (6.2.4) hold beyond the queuing context, provided the Markov Chain satisfies the stochastic monotonicity assumption.

6.3 Taboo PP on the Positive Recurrent Bridge Graph

6.3.1 Positive Recurrent and Null Recurrent Bridge Graph

The positive recurrent Bridge Graph is studied in [3]. It is shown that, in the irreducible and aperiodic case, when the driving sequence is totally independent, this graph is a.s. connected, it is unimodularizable, and it is I/F in the sense of the foil classification theorem for unimodular networks. Moreover, this I/F property gives that this graph contains a unique bi-recurrent path $\{\beta_t\}_{t\in\mathbb{Z}}$. A positive recurrent MC, X, has a unique stationary distribution, and the intersection of this path with the zero timeline gives a perfect sample of the stationary distribution of X. In contrast, in the null recurrent Bridge Graph, the current work shows that this graph is not connected in general. Moreover, it is not unimodularizable in general. However, it "contains" a unimodular random network (the Recurrence Time EFT/EFF) which allows one to prove several important properties. In contrast with the positive recurrent case, in the null recurrent Bridge Graph, there is no bi-infinite path when it is connected (or under the assumption of Proposition 5.1.4). However, there is an analog of the perfect sample of the positive recurrent case, namely the Taboo PP.

Remark 6.3.1. Of course, there are other ways for constructing a point process that its intensity is equal to the invariant measure at each point of the state space. For example, consider a path of the MC, X, starting from an arbitrary state $s^* \in \mathscr{S}$. For each $s \in \mathscr{S}$, consider the number of times this path meets s before going back to s^*. The expectation of this number for each $s \in \mathscr{S}$ is equal to ζ_s, defined in (6.1.1), which is equal to the stationary measure of s. So in each realization of the Markov Chain starting from s^*, this number can be considered. The advantage of Taboo PP as a technique of sampling is its local constructibility (at least in the monotone case), as mentioned in Subsection 6.1.1.

One can consider the Taboo PP and its properties in the positive recurrent Bridge Graph. A question that arises here is that of the relationship between the Taboo PP of a positive recurrent MC and the classical perfect sample of its stationary distribution.

Proposition 6.3.2. *Consider a positive recurrent MC, X, and its associated Bridge Graph, B_X. Then the S-set of B_X is a.s. finite. Moreover, the Taboo multiplicity of every vertex in the S-set is a.s. finite.*

Proof. In the positive recurrent case, the Bridge Graph is an $\mathscr{I}/\mathscr{F}$ unimodular network [4]. Also, the foils in the Bridge graph are its intersection with the vertical timelines. So the S-set, which is the 0- foil, is a.s. finite.
For proving the second part, note that since B_X is $\mathscr{I}/\mathscr{F}$, every vertex $y \in B_X$, which is not on the bi-infinite path, has a.s. finitely many descendants, specially finitely many

$*$-descendants. To complete the proof, note that, although there are infinitely many $*$-descendants on the bi-infinite path, only one of them does not return to s^* before time zero. Hence the bi-infinite path adds exactly mass one to the S-set. □

6.3.2 Relation between the Taboo PP and Classical Perfect Sampling

The positive recurrent MC, X, has a unique stationary distribution, σ. One can sample from this stationary distribution with the Coupling from the Past algorithm (See [21]). The Taboo PP also gives a samples from the stationary distribution, in the sense that the mean measure of the Taboo PP is the stationary distribution at each point. The relation between these two samplings is discussed in the next proposition:

Proposition 6.3.3. *Let τ be the Taboo PP of the positive recurrent MC, X. Suppose that T is a sample of τ. If one biases T with the number of the points (considering the multiplicities) that belong to it, and chooses a random point from T and denote it by Y, then Y has the stationary distribution of X.*

Proof. Let τ be the set of all possible outcomes of the Taboo PP. For each $T \in \mathscr{T}$, let P_T be the probability that event T occurs, i.e., $\mathbb{P}(\tau = T)$. Then for each $y \in \mathscr{S}$,

$$\begin{aligned}\mathbb{P}(Y = y) &= \sum_{T\in\mathscr{T}} P_T \times \frac{m(T)}{\sum_{T'\in\mathscr{T}} m(T')P_{T'}} \times \frac{T(\{y\})}{m(T)} \\ &= \frac{1}{\sum_{T'\in\mathscr{T}} m(T')P_{T'}} \sum_{T\in\mathscr{T}} P_T \times T(\{y\}) = \frac{\mathbb{E}(\tau(y))}{\mathbb{E}(\tau(\mathscr{S}))},\end{aligned}$$

where $m(T)$ is the sum of the multiplicity of the vertices in T, and $T(\{y\})$ is the multiplicity of y in T. Since this probability is proportional to $\mathbb{E}(\tau(y))$, and the stationary measure in y, $\sigma(y)$, is also proportional to $\mathbb{E}(\tau(y))$, $\mathbb{P}(Y = y) = \sigma(y)$. □

Given a realization of the Taboo PP, a natural question is whether is it possible to get a perfect sample of σ from this realization in the classical sense ?
Here is an algorithm for this, under the extra assumption that M exists such that $M > m(T) \quad \forall T \in \mathscr{T}$.

Algorithm:

1. Generate a sample T from τ.
2. Choose a point Y randomly from T with probability proportional to its multiplicity in T.
3. Accept the point Y with probability $\frac{m(T)}{M}$.

4. If the point Y is rejected, back to step 1.

By using this algorithm one can write the following equations which shows that the algorithm gives a sample of stationary distribution of the MC:

$$\begin{aligned}
\mathbb{P}(Y=y) &= \sum_{T\in\mathscr{T}} P(T)\times\frac{T(\{y\})}{m(T)}\times\frac{m(T)}{M}\times\sum_{n=1}^{\infty}\left(\sum_{T\in\mathscr{T}}P(T)\sum_{y\in T}\frac{T(\{y\})}{m(T)}(1-\frac{m(T)}{M})\right)^{n-1}\\
&= \frac{\mathbb{E}(\tau(y))}{M}\times\sum_{n=1}^{\infty}\left(\sum_{T\in\mathscr{T}}P(T)(1-\frac{m(T)}{M})\right)^{n-1}\\
&= \frac{\mathbb{E}(\tau(y))}{M}\times\sum_{n=1}^{\infty}\left(1-\sum_{T\in\mathscr{T}}P(T)\frac{m(T)}{M}\right)^{n-1}\\
&= \frac{\mathbb{E}(\tau(y))}{M}\times\sum_{n=1}^{\infty}\left(1-\frac{\mathbb{E}(\tau(\mathscr{S}))}{M}\right)^{n-1}\\
&= \frac{\mathbb{E}(\tau(y))}{M}\times\frac{M}{\mathbb{E}(\tau(\mathscr{S}))}=\sigma(y),
\end{aligned}$$

where σ is the stationary distribution of the MC, X.

Chapter

7 Dynamics on the whole Space of the Random measures

Le Chapitre 7 présente quelques résultats sur les propriétés de ces dynamiques dans des espaces d'états plus généraux. Au lieu de considérer ces dynamiques sur le Graphe de Pont, elles sont envisagées comme des chaînes de Markov sur l'espace des mesures aléatoires sur S.

Chapter content

So far, two dynamics H^T and H^P have been considered on the Bridge Graph. In this section, the properties of these dynamics on the general state space, i.e., the space of all integer-valued random measures on $\mathscr{S}$, is studied.
In the Bridge Graph (or Doeblin Graph), at each time t, the Taboo PP and Potential PP are random measures on $\mathscr{S}$. Based on the definitions of H^T and H^P, each of these measures at time t depends only on the measure at time $t-1$ and ξ_{t-1}. So one can consider these dynamics as MCs on the space of $\mathscr{M}(\mathscr{S})$, of all locally finite integer-valued measures on $\mathscr{S}$, and study the properties of these MCs to understand the properties of the dynamics in this more general space.
First, note that $\mathscr{M}(\mathscr{S})$ is a topological (Polish) space [31] that is not countable. So the concepts and notations of MCs on topological state spaces in [32] will be used.

7.1 Kernels

Suppose that Φ^T, and Φ^P are Markov Chains corresponding to the dynamics H^T, and H^P. Let P^T, and P^P be their transition probability kernels on $\mathscr{M}(\mathscr{S})$. Below, it will be shown that these kernels are well-defined by the transition probability kernel definition in [32]. It is enough to show that $P(\mu, B)$, the probabilities of transition from each measure $\mu \in \mathscr{M}(\mathscr{S})$ to each Borel set $B \in \mathscr{M}(\mathscr{S})$ is well-defined.

Proof. Let $\hat{C}_{\mathscr{S}}$ be the class of all continuous and compact support functions $h : \mathscr{S} \to \mathbb{R}^+$, and $\mu h = \int h d\mu$. Since $\mathscr{S}$ is countable the last equation is equivalent to $\mu h = \sum_{x\in\mathscr{S}} h(x)\mu(x)$. The sets of the form

$$K_h^{r,s} = \{\mu | r < \mu h < s \quad , r, s \in \mathbb{R}, h \in \hat{C}_{\mathscr{S}}\} \\ = \{\mu | r < \sum_{x\in\mathscr{S}} h(x)\mu(x) < s, \quad r, s \in \mathbb{R},\ h \in \hat{C}_{\mathscr{S}}\} \tag{7.1.1}$$

is a basis of the vague topology on $\mathscr{M}(\mathscr{S})$.
Let S_h be the support of h. Since h has a compact support, S_h is finite, so $K_h^{r,s}$ belongs to a finite configuration of the masses on S_h, and in the rest of the $\mathscr{S}$ they can take any values. For each h, r, and s one can consider $K_h^{r,s}$ as the following finite intersection of the basis elements:

$$K_h = \cap \mu_{x,i}, \tag{7.1.2}$$

where $\mu_{x,i} \in \mathscr{M}(\mathscr{S})$ put mass i at point $x \in \mathscr{S}$ and, in the rest of the space, it takes any values. For each $\mu \in \mathscr{M}(\mathscr{S})$, $P(\mu, \mu_{x,i})$, the probability of going from μ to $\mu_{x,i}$ using H^T and H^P is computable and well defined. So for each $\mu \in \mathscr{M}(\mathscr{S})$ and $\nu \in K_h^{r,s}$, P^T and P^P is well defined. Since $\{K_h^{r,s}\}$ forms a basis for the vague topology on $\mathscr{M}(\mathscr{S})$, these transition kernels are well-defined. □

Consider the following definitions for these MCs.

Definition 7.1.1. Consider the Taboo/Potential dynamics constructed by a positive (resp. null) recurrent Markov Chain. The MC, Φ^T/Φ^P, corresponding to this dynamics is called positive (resp. null) Taboo/Potential Markov Chain on $\mathscr{M}(\mathscr{S})$.

7.2 Stability Properties of the MCs

In this section the properties of the MCs that are defined in the 7.1 are studied.
The first property is the existence of stationary measures of the MCs. Before going through this, consider the following backward construction for Taboo and Potential PP. Consider the MC $X = \{X_n\}_{n\in\mathbb{N}}$ on $\mathscr{S}$. Let τ be the Taboo Point Process constructed by X, fixing s^* on the Bridge Graph. Consider $\tau'_0 = \delta_{s^*}$. Define τ'_1 by $\tau'_1.\theta = \Phi^T(\tau'_0, \xi)$, where θ is the shift transformation. So τ'_1 is a random measure at time zero when a path started at time -1, from s^* and go forward with the law of Φ^T. For each n, consider

$$\tau'_{n+1}.\theta = \Phi^T(\tau'_n). \tag{7.2.1}$$

Roughly speaking τ'_n is the random measure that there is at time 0 in the Bridge Graph, when the Bridge Graph is constructed from time $-n$. Note that

$$\tau'_n \to \tau, \tag{7.2.2}$$

Where τ is the Taboo Point Process constructed by X. Similarly one can define π'_n for the Potential MC such that

$$\pi'_n \to \pi, \tag{7.2.3}$$

where π is the Potential Point Process. (See [])

Proposition 7.2.1. *The Taboo PP is a stationary distribution of the Positive/Null Taboo MC. More over for each $n \in \mathbb{N}$, the measures that lie in the support of τ'_n, τ and any finite measure is in the domain of attraction of this stationary distribution.*

Proof. Due to the definition of the Taboo PP, it is a stationary distribution of the Taboo MC. And since the equality 7.2.2 holds, starting from eny measure that exists in the support of τ_n and τ, the limit distribution is the Taboo PP.
Suppose that $\mu \in \mathscr{M}(\mathscr{S})$ is a finite measure on $\mathscr{S}$. Every s that is in the support of μ, after finitely many times meet s^* since the initial MC is recurrent. The support of μ is finite. So after finitely many times a.s all the masses that in μ vanishes. So starting from a finite measure, the Markov Chain lies in the Bridge Graph after finitely many times, and the limit is the Taboo PP. □

Proposition 7.2.2. *The Potential PP of a null recurrent MC (in the cases that it is a.s finite at each point), is a stationary distribution of the Null Potential MC. More over for each $n \in \mathbb{N}$, the measures that lie in the support of π'_n, and π is in the domain of attraction of this stationary distribution.*

Proof. The proof is the same as in the Taboo MC case. □

Remark 7.2.3. Note that in the Potential PP is not a stationary distribution of the Positive Potential MC. Since the Potential PP in this case has an infinite mass at one point a.s. So its support does not belongs to $\mathscr{M}(\mathscr{S})$.

So far it is known that the Positive/Null Taboo MC and the Null Potential MC have stationary distributions.The question that arises here is that, is about the uniqueness of this stationary distribution, and consequently the irreducibility of these MCs. Here, since the state space is not countable, the classical definition of irreducibility does not make sense. So the corresponding definition of it for the Markov Chains with general state space, Which is called $\psi - irreducibility$ of the Markov Chain will be used. This definition is with respect to a measure ψ, that is on the state space of the Markov Chain ($\mathscr{M}(\mathscr{S})$).

Definition 7.2.4. A Markov chain Φ defined on $\mathscr{M}$, is called ϕ-irreducible, if there exists a measure ϕ on $B(\mathscr{M})$ such that whenever $\phi(A) > 0$, the probability that Φ reaching A from x is positive, for all x.
The MC is called ψ-irreducible if there exists a maximal measure ψ such that the MC is ψ-irreducible. In the sense that for any other measure ϕ', the chain is ϕ'-irreducible, if and only if $\psi \succ \phi'$.

Proposition 7.2.5. *The positive (null) Taboo MC is not $\psi - irreducible$, in general.*

Before going through the proof, consider the following example.

Example 7.2.6. Consider the Renewal MC in Definition 5.2.1. Consider the Taboo MC constructed by this MC. This measure valued MC is called the **Renewal Taboo MC**, denoted by $\{\Phi_n^{T,R}\}$. In both cases that the Renewal MC is positive recurrent and null recurrent, Taboo PP is a stationary distribution of Renewal Taboo MC. Moreover, any finite measure is in the domain of attraction of this distribution.
First, consider the null recurrent case. Denote the Taboo PP in this case with τ_{Ren}. Consider the following measure on $\mathbb{N}$, which has mass one at each point of the space:

$$\Phi_0^{T,R} = \sum_{k\in\mathbb{N}} 1_k. \tag{7.2.4}$$

And let it be the initial state of the MC $\{\Phi_n^{T,R}\}$. Then

$$lim_{n\to\infty}\Phi_n^{T,R} = \tau_{Ren} + \sum_{k\in\mathbb{N}} 1_k. \tag{7.2.5}$$

The reason why this equality holds is that the mass 1, that there is in the initial state $\Phi_0^{T,R}$, for large k, put an extra mass at every point of the limiting distribution, due to the law of the Renewal MC.
The same equality holds for the positive recurrent case, when the Renewal Taboo MC starts with the same initial state $\Phi_0^{T,R}$.

Proof of Proposition 7.2.5. It will be shown that the Renewal Taboo MC is not $\psi - irreducible$ in both positive and null recurrent case.
Let $B(\mathscr{M})$ be the Borel σ-field on $\mathscr{M}(\mathscr{S})$. Consider the Renewal Taboo MC in 7.2.6. If this MC is $\psi - irreducible$, then, there is a measure φ on $\mathscr{M}(\mathscr{S})$ such that for all $A \in B(\mathscr{M})$ with $\varphi(A) > 0$, and for each $\mu \in \mathscr{M}(\mathscr{S})$, $p_\mu(\tau_A < \infty) > 0$.
Fix $A \in B(\mathscr{M})$ with $\varphi(A) > 0$. Let μ_0 is any finite measure. Starting from μ_0 and $\Phi_0^{T,R} = \sum_{k\in\mathbb{N}} 1_k$ defined in (7.2.4), $\{\Phi_n^{T,R}\}$ has positive probability to enter A in finite time. On the other hand, $lim_{n\to\infty}\Phi_n^{T,R}$, starting from these two starting state is different. So such a φ does not exist.

□

Proposition 7.2.7. *The Null Potential MC is not $\psi - irreducible$, in general.*

Proof. In Null Potential MC, the Potential PP is a stationary distribution on $\mathscr{M}(\mathscr{S})$. With the same proof as in Example 7.2.6, one can show that considering the Null Potential MC constructed by null Renewal MC, And staring from (7.2.4) measure, the limit distribution changes. So with the same argument, in the Renewal example, the Null Potential MC is not $\psi - irreducible$. □

In the Positive Potential MC, it is known up to hear that the Potential PP is not the stationary distribution of this MC. Moreover it will be shown, in the following proposition, that in this case the MC is not tight on the support of π'_n.

Proposition 7.2.8. *The Positive Potential MC is not tight, when the MC starts from a measure that lie in the support of π'_n, for some n. (The Positive Potential MC is not tight on the Bridge Graph.)*

Proof. For the proof it will be shown that the π'_n, defined in (7.2.3), is not tight when X is positive recurrent.
Since $\mathscr{S}$ is countable, a set $K \subset \mathscr{M}(\mathscr{S})$ is compact iff there exists $M_K \in \mathbb{N}$ such that for all $\mu \in K$, and all $s \in \mathscr{S}$, $\mu(s) \leq M_K$. If π'_n is tight, then for every $\varepsilon > 0$ there is a $n_0 \in \mathbb{N}$ and compact set $K \subset \mathscr{M}(\mathscr{S})$ such that $p(\pi'_n \in K) > 1 - \varepsilon$, for all $n > n_0$.
On the other hand the equality (7.2.3) holds, and every measure that is in the support of π'_n has infinite value at one point a.s. It means that π'_n exit any compact after finitely many steps. So π'_n is not tight. □

Part II

Clustering on Point Processes

Summery

Here is a summary of Part II of the thesis. In the Introduction, Chapter 1, the HTNN algorithm is briefly defined. Before delving into the details of this, it is important to mention that there are clustering algorithms in the literature that are related to the HTNN algorithm introduced in this work. Here are a few of them:

The first algorithm is the Nearest-Neighbor Chain (NNC) algorithm, which is a hierarchical algorithm based on the well-known nearest neighbor chain of points. For a detailed definition, see [6] and [7], as well as Example 3.4.2 in Section 3.4. The other clustering algorithm related to HTNN algorithm is Mutual Nearest Neighbor based clustering (MUNEC) algorithm (see [33]), which is a hierarchical algorithm where clusters are connected at each step based on the set distance between them. In contrast, HTNN connects clusters based on the distance between their centroids.

As mentioned before, first, the algorithm is applied to the Poisson point process (PPP). Chapter 8 contains the definition and main results. It will be shown in Subsection 8.1 that the HTNN algorithm can be defined as a sequence of point shifts on the PPP, with each point shift representing a step of the algorithm. Consequently, at each step n, the point shift generates a random directed graph on the PPP, which forms a forest. The resulting random forests are unimodular graphs (see [2] and Section 3.3).

At each step n, the cluster heads of order n form a point process, denoted as Ψ^n. It is evident that the intensity of these point processes are decreasing. Section 8.1 establishes an upper bound, for the intensity of Ψ^n, denoted as ρ_n, in relation to the intensity of the initial Poisson point process, denoted as ρ_n. Specifically, it is shown that ρ_n is bounded above by $\rho/2^n$. This raises interesting questions regarding the statistical properties of Ψ^n. One such question pertains to whether the sequence of scaled versions of Ψ^n by ρ_n converges to a limit or not. The conjecture is that this sequence is tight, implying that at least a subsequence of $\rho_n\Psi^n$ converges to a limit, suggesting that for large n, the random measures $\rho_n\Psi^n$ have nearly identical distributions.

Section 8.2 considers the weak limit of the point shifts, referred to as the Recursive Thinning Nearest Neighbor Point Shift (RTNN PS), and its associated graph, the Recursive Thinning Nearest Neighbor Forest/Tree (RTNNT/F). The limiting structure, RTNNT/F, gives rise to an infinite spanning forest or tree on the PPP. Theorem 8.2.3 demonstrates that, on the PPP,

the point shift graph of each level n of the algorithm contains infinitely many finite clusters (trees), and in the limiting graph when n goes to infinity, the infinite random tree containing the origin is one-ended.

It is important to note that the existing literature contains results regarding the existence and construction of spanning forests (or trees) on point processes or stationary point processes. However, these results have objectives that are distinct from the focus of our work. For instance, the Minimal Spanning Forest/Tree (MSF/T) was introduced as a means to find a natural extension of the minimal spanning tree from finite graphs to infinite graphs and point processes (refer to [34] and [35]). The construction of the MSF (Tree) can be viewed as a hierarchical algorithm where points are gradually connected. However, this procedure requires significant computational effort. In another work, Holroyd and Peres presented a proof of the existence of a factor graph that forms a tree on a Poisson point process in $\mathbb{R}^d$ (see [13]). A factor graph on a point process is a mapping of the point configuration to a graph on it that is measurable and equivariant with respect to the point process, with no additional randomness. With this definition, for example, the RTNNF/T is a factor graph on the PPP. However, their proof involves merging finite connected components in a manner that cannot be algorithmically constructed. In [36], it is shown that the result of Holroyd and Peres can be extended to point processes that are invariant under the isomorphism of Rd, whose groups of symmetries are almost surely trivial. In Subsection 8.3, the connection between RTNNT/F and other spanning forests on the PPP is examined.

The results of the properties of HTNN clustering on a PPP are as follows. We phrase them in terms of the genetic data of the species example, discussed in the Introduction. All species belong to an infinite phylogenetic tree, which is a connected component (cluster) of the RTNNT/F. Since each component of the RTNNT/F is one-ended, it shows that there exists some sort of universal common ancestor at infinity for all species in a component often called LUCA (Last Universal Common Ancestor) in the literature. The question of whether there are multiple LUCAs or just one arises here for PPPs. This question boils down to finding the number of connected components in the RTNNT/F. There are no general results regarding the connectivity of RTNNT/F. Theorem 8.2.5 shows that the RTNNT/F is a tree in dimension 1. Our conjecture is that this graph is connected in each dimension d when constructed from a PPP, i.e., there is only one LUCA in the RTNNT/F phylogenetic tree of a PPP.

By leveraging the construction, we can calculate the average number of species within descendant trees at different levels. This construction also allows us to define distances between species in the same phylogenetic tree (cluster in RTNNT/F), indicating their level of closeness. Finally, Chapter 9.1 delves into the details of the proof of Theorem 8.2.3.

Chapter

8 Hierarchical Thinning Nearest Neighbor Clustering on Point Process

Ce chapitre examine l'algorithme de regroupement hiérarchique par éclaircissement des voisins les plus proches (HTNNC) sur un processus ponctuel de Poisson (PPP). Comme indiqué précédemment dans l'introduction, nous démontrerons que chaque niveau de l'algorithme peut être caractérisé comme un décalage ponctuel sur le PPP.

Chapter content

This chapter considers the Hierarchical Thinning Nearest Neighbor Clustering (HTNNC) algorithm on Poisson point process (PPP). As previously stated in the introduction, we will demonstrate that each level of the algorithm can be characterized as a point shift on the PPP.

We define a point-shift f on a homogeneous PPP with intensity 1, say

$$\Phi^0 = \sum_i \delta_{x_i}$$

in $\mathbb{R}^d$. The construction described below is scale invariant and there is hence no restriction assuming that the intensity is 1.

8.1 Construction of the pre-limit point shifts

We construct by induction a sequence f^n, $n \geq 0$, of point shifts on Φ^0. For this, we will need the following discrepancy function acting on two pairs of points, say S and T:

$$\delta(S,T) = \min_{x\in S, y\in T} d(x,y), \tag{8.1.1}$$

where d denotes Euclidean distance. The discrepancy between these two sets is hence the shortest distance between the two sets. This function is non-negative and symmetric. However, it does not satisfy the triangle inequality. In addition, different sets have discrepancy 0 when they share one point.

Order 0 The first one, f^0, is the NN (nearest neighbor) point-shift on Φ^0 w.r.t. Euclidean distance d^0 between points. That is for all x, $f^0(x) = \mathrm{NN0}(x)$, where NN0 maps a point to its nearest neighbor in Φ^0.
The graph of the point-shift f^0 has all its connected components of the $\mathscr{F}/\mathscr{F}$ class in the sense of point shift classification in Theorem 3.3.2 i.e., all of its connected components are finite, and have a unique cycle. This is because the PPP does not admit any descending chain (see [14] and Subsection 3.1.3). Each connected component is made of directed trees connected to one cycle which connects two mutual nearest neighbor (MNN) points of Φ^0.
Let Φ^0_c be the sub-PP (point process) of Φ^0 made of the points that belong to f^0-cycles. Let $\Phi^0_a = \Phi^0 \setminus \Phi^0_c$. When deleting the f_0-cycles in the f_0 graph, one gets a collection of directed trees with a gateway point at the points of Φ^0_c which will be referred to as the f^0 sub-clusters of order 0. Each gateway point is a cluster head of order 0 (For the definition of cluster head, see Introduction 1). Hence, Φ^0_c is the PP of cluster heads of order 0 and the tree associated with x, a cluster tree of level 0, is the set of points y of Φ_a such that the f_0-orbit staring from y reaches x before reaching the mutual nearest neighbor of x denoted by $\mathrm{MNN}(x)$.

Order 1 Let $\{S^0_i\}_i$ denote the collection of f_0-cycles. The sets $\{S^0_i\}_i$ form a translation invariant partition of the support of Φ^0_c. Consider the point process of cycles of order 0 on the space of pairs of points of $\mathbb{R}^d$ (closed sets of cardinality 2):

$$\Psi^1 = \sum_i \delta_{S^0_i}.$$

Consider the NN1 graph on these pairs based on δ, namely the graph with a directed edge from S^0_i to S^0_j if $\delta(S^0_j, S^0_i) < \delta(S^0_k, S^0_i)$ for all $k \neq i, j$. The fact that points have densities w.r.t. the Lebesgue measure guarantees that there are no ties a.s. The exit point of the pair S^0_i is defined to be the point of this pair that achieves the minimum in (8.1.1)

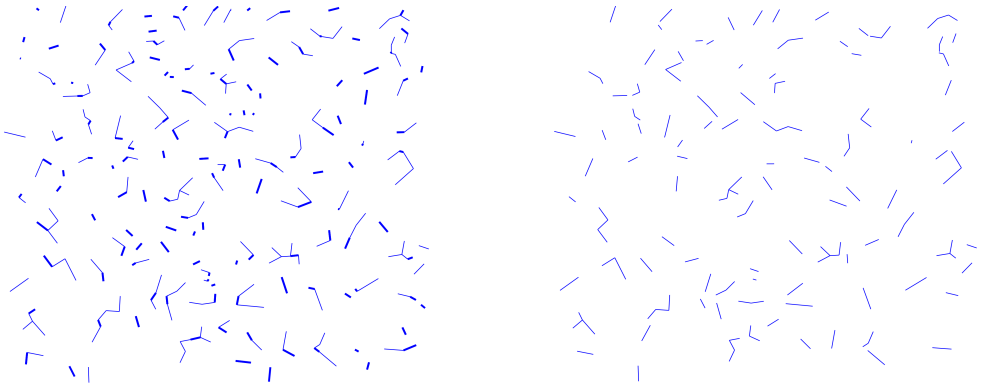


Figure 8.1: Left picture: the graph of f_0 in blue with the f_0 cycles in thick blue; Right: the 0 clusters (the clusters made of a single point are not visible here). These trees are rooted at cluster heads of level 0.

when taking for S_j^0 the set that follows S_i^0 in the NN1 partial order, namely the pair that is the 'nearest' to S_i^0 for δ. It will be shown in Corollary 9.1.6 that the connected components of this graph are finite trees connected to directed cycles of length 2 connecting two exit points, namely two points $x \in S_0^i$ and $y \in S_0^j$ such that S_0^i and S_0^j are MNN1, y is the closest point of S_0^j from S_0^i, and x is the closest point of S_0^i from S_0^j.

This also defines two sub PPs of Ψ^1 which are Ψ_a^1 and Ψ_c^1, where Ψ_c^1 is the subset of points of Ψ^1 which are in a δ cycle and $\Psi_a^1 = \Psi^1 \setminus \Psi_c^1$.

Define f^1 to be the point-shift on Φ^0 that coincides with f^0 everywhere except for exit points of level 0. For each such point, say $x \in S^0$, where S^0 is the pair where x is the exit point, define $f^1(x) = \mathrm{NN1}(x)$, where $\mathrm{NN1}(x)$ denotes the nearest point for d of the pair which is the 'nearest' of S^0 for δ. In words, for each exit point of an f_0-cycle, we obtain f^1 from f^0 by replacing the initial image (which was the nearest point in NN0, namely its MNN for d) by the d-nearest point in the 'δ-nearest' f_0-cycle. The graph of f^1 on Φ^0 is $\mathscr{F}/\mathscr{F}$ (see Corollaly 9.1.6), namely made of finite directed trees connected to directed cycles, the f^1-cycles, which connect two exit points of order 0. These f_1-cycles are disjoint by construction. The points that belong to such f_1-cycles form a sub PP of Φ_c^0 that will be denoted by Φ_c^1. Let $\Phi_a^1 := \Phi^0 \setminus \Phi_c^1$ denote the vertices of the acyclic part of the f^1 graph.

When deleting the f_1 cycles in the f_1 graph, one gets a collection of directed trees with a gateway points at the points of Φ_c^1, which will be referred to as the *sub-cluster trees of order 1*. The gateway points of these trees are called cluster heads of order 1. Equivalently, Φ_c^1 is the point process of cluster heads of order 1. See Figure 8.2 as an example of f^1-graph on a realization of PPP.

Remark 8.1.1. Note that all cluster heads of order 1 are exit points of order 0, but not conversely. Since half of cluster heads of level 0 are exit points, the density of cluster heads of level 1 is less than $\rho/2$, where ρ is the intensity of the cluster heads of order 0.

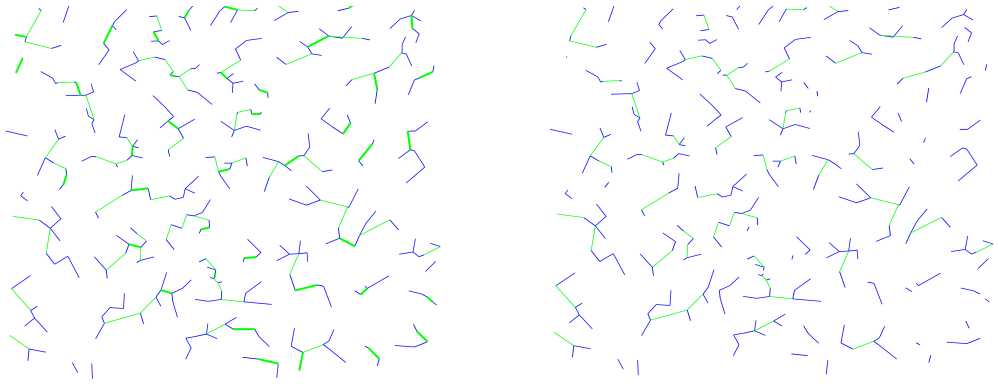


Figure 8.2: On the left, there is a graph of f_1 with the f_1-cycles highlighted in bold. The difference between this graph and the f_0 graph is the addition of new edges from the head points of f_0 cycles to the closest point of the closest f_0 cycle, which are colored green. The right side image shows the 1-sub-cluster trees. The sub-cluster trees made of a single point are not visible here.

Order 2 Let $\{S_i^1\}_i$ be the sets of f^1-cycles and let

$$\Psi^2 = \sum_i \delta_{S_i^1}.$$

Let NN2 be the NN trees on the points of Ψ^2 w.r.t. δ.

By the same arguments as above, this first defines Ψ_c^2, the set of points of Ψ^2 which are in a cycle for NN2 and $\Psi_a^2 = \Psi^2 \setminus \Psi_c^2$.

The exit point (of order 1) of the pair S_i^1 is defined to be the point of this pair that achieves the minimum in δ when taking for S_j^1 the set that follows S_i^1 in the NN2 partial order.

The point-shift f^2 on Φ^0 coincides with f^1 except for exit points of order 1. For such a point, say $x \in S^1$, $f^2(x) = \mathrm{NN2}(x)$, where $\mathrm{NN2}(x)$ denotes the nearest point for d of the set of pairs of order 1 which is the nearest to S^1 for δ. Again f^2 defines a point-shift on Φ^0. Its graph is $\mathscr{F}/\mathscr{F}$, with cycles of length 2 a.s. as there are no ties a.s. The points belonging to these f_2-cycles are exit points of order 1 which form a sub PP of Φ_c^1 that will be called the f_2-cycle PP and denoted Φ_c^2. Let $\Phi_a^2 := \Phi^0 \setminus \Phi_c^2$ denote the vertices of the acyclic part of the f^2 graph.

When deleting the f^2 cycles, i.e. the two directed edges, one gets a collection of finite trees which will be called the sub-cluster trees of order 2. Each tree has a gateway point at a cluster head of level 2. The PP Φ_c^2 is hence also called the set of cluster heads of level 2.

Induction assumption The induction assumption is that

- f^n has been constructed as an $\mathscr{F}/\mathscr{F}$ point-shift on Φ^0 (in the case of PPP this property is shown in Theorem 8.2.3);

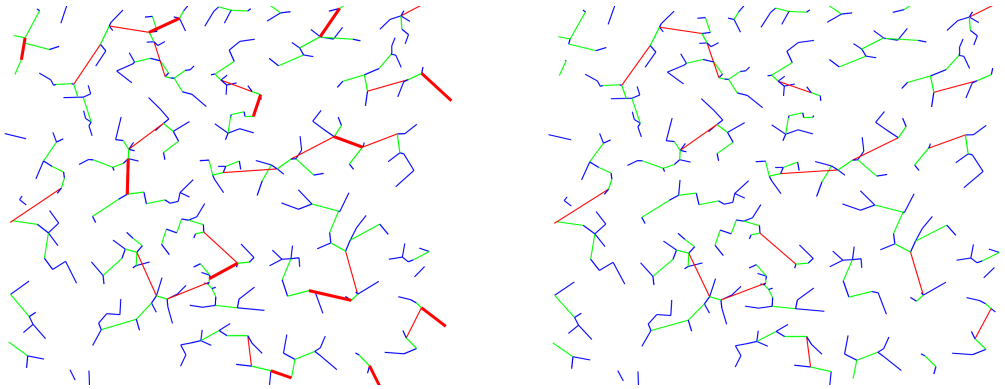


Figure 8.3: Left picture: the graph of f_2 with the f_2-cycles in boldface; The difference with the f_1 graph, which consists in the new edges from head points of f_1 cycles to the closest point of the closest f_1 cycle, are in red. Right: the 2-clusters.

- The f_n-cycles $\{S_n^i\}$ of the connected components are all of length 2;
- The point process Φ_c^n of cyclic points of the f_n-graph has an intensity bounded above by $\rho/2^n$;

Induction step Let Ψ^{n+1} be the set-valued point process

$$\Psi^{n+1} = \sum_i \delta_{S_i^n}.$$

Let NN$(n+1)$ denote the associated NN structure, which we will prove to be $\mathscr{F}/\mathscr{F}$ in Theorem 8.2.3. Let Ψ_a^{n+1} and Ψ_c^{n+1} be the acyclic and the cyclic partition of Ψ^{n+1}. This also defines the set of n-head points. The exit point (of order n) of S_i^n is the point realizing the minimal distance between the points of S_i^n and the set S_j^n which is NN$(n+1)$ next for S_i^n (the nearest neighbor of S_i^n for δ). One defines f^{n+1} from f^n by keeping the images of all points the same except for n-head points, the images of which are defined as follows: the image of such a point say $x \in S_n^k$ by f^{n+1} is NN$(n+1)(x)$, which is the point of S_n^{k+1} which is the closest to x for d. The graph of this point shift $\mathscr{F}/\mathscr{F}$, namely is made of finite directed trees connected to cycles of length 2 connecting certain head points of order n. This defines the f_{n+1} cycles. Since $\Phi_{n+1}^c \subset \Phi_n^c$ and since a f_{n+1}-cycle requires two f_n-cycles (in MNN relation), the assumption that the intensity of Φ_n^c is at most $\rho/2^n$ implies that the intensity of Φ_{n+1}^c is at most $\rho/2^{n+1}$.

By construction, we have $\Phi^0 = \Phi_c^{n+1} \cup \Phi_a^{n+1}$ and $\Phi_c^{n+1} \subset \Phi_c^n$. In addition, since Φ_c^{n+1} is a sub PP of the $(n+1)$-exit points , the intensity of the former is bounded above by $\rho/2^{n+1}$.

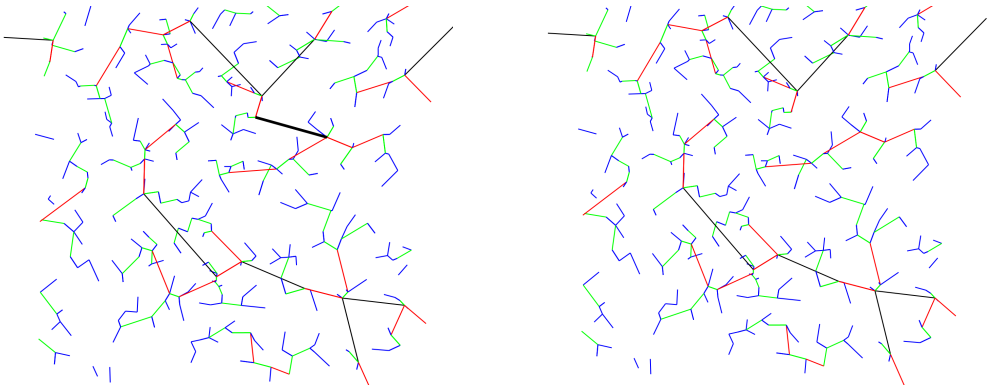


Figure 8.4: Left picture: the graph of f_3 with the f_3-cycles in boldface; The difference with the f_2 graph, which consists in the new edges from head points of f_2 cycles to the closest point of the closest f_2 cycle, are in black. Right: the 3-clusters.

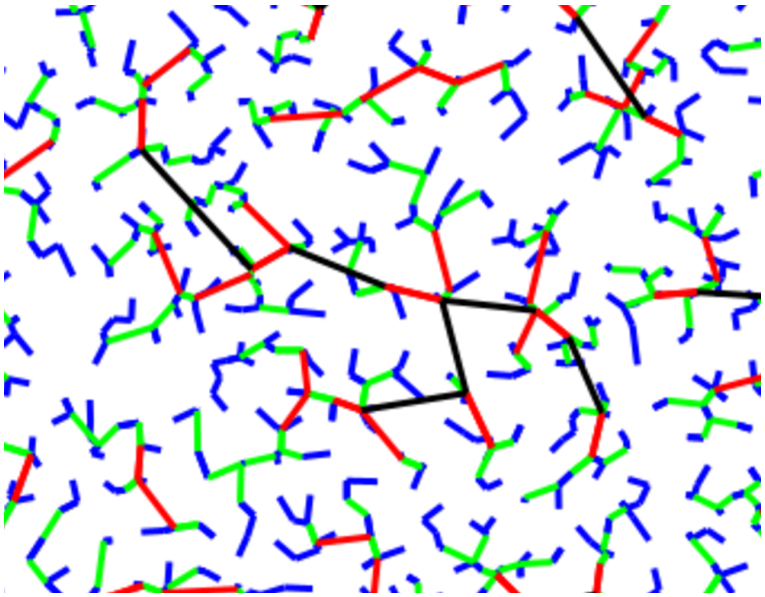


Figure 8.5: A f_3-cluster with the same color code as above

8.2 Construction of the Limiting Point Shift

This section discusses the existence and properties of the limiting point shift and point shift graph of the HMNNC algorithm.

Definition 8.2.1. Consider the sequence f^n, of the point shift constructed in Subsection 8.1 on a PPP. The fact that the intensity of Φ_c^i tends to 0 as i tends to infinity, shown in Subsection 8.1, imply that $\{\Phi_a^{k+1} \setminus \Phi_a^k\}_{k\geq 0}$, where $\Phi_a^0 = \emptyset$, forms a partition of Φ^0. This defines a new point shift f^∞, Its graph, the f^∞-graph, denote by G, is acyclic due to the construction of the f^ns. The f point shift is called the Recursive Thinning Nearest Neighbor Point Shift (RTNN PS) and its graph is called the Recursive Thinning Nearest Neighbor Tree/Forest (RTNNT/F).

Corollary 8.2.2 (Construction of the point shift graph)**.** *The RTNNT/F is a local weak limit for f^n-graphs on Φ^0 as $n \to \infty$. This graph under the palm of Φ^0 and rooted at zero is a unimodular graph.*

Proof. This is assessed under the Palm of Φ^0. For all balls of radius h centered at 0, the restriction of f^n-graphs to this ball converges a.s. to a limit when n tends to infinity. This is because the intensity of the points of Φ_c^i tends to 0 as i tends to infinity. Hence, there is a random n such that the restriction of the graph G^n to this ball is fully determined by f^n (this random integer is the largest integer n such that the ball in question is empty of points of Φ_c^{n+1}). We will denote by $[G, 0]$ the limiting random graph, and by $[G^n, 0]$ the f^n-graph under the palm of Φ^0. Since for each n, $[G^n, 0]$ is an unimodular graph (see [8]) so their limit graph $[G, 0]$ is also unimodular (see [3]). □

Theorem 8.2.3. *For each n, the connected components of the graph $[G^n, 0]$ are of type $\mathscr{F}/\mathscr{F}$. In terms of clustering at each level n, the clusters in the HTNNC algorithm are a.s. finite.*

Proof. In section 9.1, it will be shown that there is no infinite path in $[G^n, 0]$, for all n. This fact demonstrates that for each n, all the clusters of $[G^n, 0]$ are finite. Note that the cycle indicated in the properties of the $\mathscr{F}/\mathscr{F}$ components (see Theorem 3.3.2 in Section 3.3) in each cluster of f^n-graph is the cycle of mutual nearest neighbor points, which is a cycle of length two. □

Corollary 8.2.4. *The connected component of RTNNT/F graph $[G, 0]$ belongs to the $\mathscr{I}/\mathscr{I}$ class of the foil classification theorem in unimodular networks. In other words, on PPP each cluster of RTNNT/F graph is a one-ended tree with all its foils infinite a.s.*

Proof. As mentioned before, for each level n, the equality, $\Phi^0 = \Phi_c^n \cup \Phi_a^n$ holds, where Φ_c^n denote the cluster heads, and Φ_a^n are acyclic points of level n. When for some level

n, a point belongs to Φ_a^n, the number of its descendants is finite and does not change after that due to the construction of RTNNT/F. Since the intensity of Φ_c^n tend to zero as n goes to infinity, $\{\Phi_a^{n+1} \setminus \Phi_a^n\}_{n\geq 0}$, where $\Phi_a^0 =$, forms a partition of Φ^0. So a.s., there is a k such that 0 belongs to $\Phi_a^{k+1} \setminus \Phi_a^k$. So its descendant tree is finite. Therefore, the connected component of zero is in the $\mathscr{I}/\mathscr{I}$ class. In other words, the cluster of the origin in the RTNNT/F graph is a one-ended tree. □

The Existence of a Last Universal Common Ancestor at infinity. Considering the genetic data of the species example discussed in the introduction, we can gain clearer insights by examining the properties of the RTNNT/F graph. Firstly, the $\mathscr{I}/\mathscr{I}$ property of the graph shows that each species (point) belongs to an infinite phylogenetic tree. The $\mathscr{I}/\mathscr{I}$ property shows that there exists a LUCA (Last Universal Common Ancestor) at infinity, from which all species in the cluster are descendants. This raises the question of whether there are different LUCAs or just one in the clustering of the Poisson point process. This is equivalent to determining the number of connected components that exist in the RTNNF/T. Although the number of connected components is generally unknown, in dimension $d = 1$, the following theorem shows that the RTNNT(F) has only one connected component.

Theorem 8.2.5. *The RTNNT(F) in dimension $d = 1$ under the palm of $\Phi^0 \subseteq \mathbb{R}$ is a tree.*

Proof. To prove that $[G, 0]$ contains only one connected component in dimension 1, it suffices to show that for every fixed $r > 0$, almost surely all points of Φ^0 within the ball of radius r centered at 0, denoted as $B_r(0)$, belong to the same component.

Recall that at each level n of the HTNNC algorithm, the point process Φ_c^n consists of the vertices of f^n-cycles S_i^n at level n, (see definition in Section 8.1). Note that the distribution of distances between pairs of S_i^n cycles, for each i, is stationary. Consider the following events:

$$A_{n+1} = \{\text{The first } f^n\text{-cycle after point } 0 \text{ in the positive side of } \mathbb{R} \text{ is connected to an } f^n\text{-cycle in the negative side of } \mathbb{R} \text{ at level } n+1 \text{ of the algorithm}\}.$$

Consider the two points of f^n-cycles, x_i and y_i , belonging to S_i^n. Let's define $M_i^n = \frac{x_i+y_i}{2}$, which represents the middle points of these two points. The sequence of random distances between M_i^ns, $\{d_i\}_{i\in\mathbb{Z}} = \{|M_i^n - M_{i-1}^n|\}_{i\in\mathbb{Z}}$, forms a stationary process. Therefore, the probability of a typical cycle connecting to its left neighbor cycle is equal to the probability of it connecting to its right neighbor cycle. Therefore,

$$\sum_n p(A_n) = \infty.$$

The Borel-Cantelli lemma implies that A_n occurs for infinitely many n. Similarly, one can

show that the same result holds for each n, for the event A'_n,

$$A'_{n+1} = \{\text{The first } f^n\text{-cycle after point } 0 \text{ in the negative side of } \mathbb{R} \text{ is connected to an } f^n\text{-cycle in the positive side of } \mathbb{R} \text{ at level } n+1 \text{ of the algorithm}\}.$$

Furthermore, note that when a cluster from the right-hand side of zero connects to a cluster on the left-hand side of zero and forms a new cluster, all points between them also belong to the same cluster. It is also evident that at step n, the place of the first f^n-cycle from the right-hand side of zero tends to infinity as $n \to \infty$. Thus, for each r, there exists a sufficiently large n such that the first f^n-cycle from the right-hand side of zero connects to the left-hand side, and $B_r(0)$ is contained in the new cluster formed by connecting these f^n-cycles to each other. □

As it is mentioned before in the Summary of this part, our conjecture is that this last theorem is true for all dimensions d.

Remark 8.2.6. All the results shown here for the PPP case are due to its property of not admitting a second-order descending chain (see Definition 9.1.1). Therefore, for any stationary point process that satisfies this property, the results hold.

The Cardinality of Descendant Tree. The construction of the RTNNF/T allows us to calculate bounds on the mean number of species within the descendant tree of a typical cluster at a specific level k. The direction of the edges in the graph signifies the order of descent from children to father. The $\mathscr{I}/\mathscr{I}$ property (of the connected component of zero), ensures a finite number of descendants for each vertex (species), so it enables us to quantify the number of species in the descendant tree.
More precisely, let the intensity of the cluster heads of level k, Φ_c^k, be ρ_k. Then the mean of the cardinality of the descendant tree of a point belongs to Φ_c^k, is $1/\rho_k$. it is known from Remark 8.1.1 that ρ_k tend to zero, as k goes to infinity.

Proposition 8.2.7. *Let G be the RTNNF/T under the Palm probability of $\Phi^0 \subseteq \mathbb{R}^d$. Let N^0 be the cardinality of the descendant tree of 0. Then the mean number of N^0 is infinite. So N^0 is heavy-tailed.*

Proof. The rooted graph $[G, 0]$ is a unimodular graph. Let f be the HTNN PS on $[G, 0]$. Define $g[G, x, y] = \sum_{n=1}^{\infty} \mathbf{1}_{y=f^n(x)}$, which is well-defined and measurable. Therefore, the mass transport principle (3.3.1) holds for g. The left-hand side of (3.3.1) for g is equal to infinity, and its right-hand side is equal to the mean number of descendants of 0. So the mean of N^0 for $[G, 0]$ is infinite. □

The same result remains valid for all the vertices (species) that belong to the foil of zero on the RTNNF/T. This holds true since the following proposition holds.

Proposition 8.2.8. *Let G be the graph of RTNNF/T under the Palm probability of $\Phi^0 \subseteq \mathbb{R}^d$, and consider the RTNN PS on G. Let L_G^0 be the foil of 0. Then the distribution of $[G, 0]$ is identical to the distribution of $[G, v]$, for each $v \in L_G^0$.*

Proof. It is known that there exists a total order on the PPP (Palm Probabilistic Partition) isomorphic to the order on $\mathbb{Z}$, using the depth-first search algorithm (see [13]). Thus, there is a total order on the vertices belonging to the foil of 0, denoted as L_G^0. Consequently, a vertex-shift on G can be obtained by mapping each vertex to its subsequent vertex within the same foil, following this order. This vertex shift is a bijection. By applying Mecke's Point Stationary Theorem (see [3] and Chapter 3), it can be concluded that the distribution of $[G, 0]$ is identical to the distribution of $[G, v]$. □

This means that for each vertex belonging to the foil of 0, the number of points belonging to its descendant tree is a heavy-tailed random variable. If there exists sufficient independence between these heavy-tailed descendant trees, there should be extremely large trees in the foil of zero.

8.3 Connection with Other Spanning Forests on Poisson Point Process

In this subsection, we explore the connection between the hierarchical model for clustering introduced by the HMNNC algorithm and other spanning forests on the Poisson point process (PPP).

The hierarchical model presented in this paper serves as a toy model for hierarchical clustering, with a focus on its properties within the Poisson point process. At each level of the algorithm, we connect a finite number of clusters and construct new finite clusters. Each of these clusters can be viewed as a finite graph with a unique cycle of length two.

It has been demonstrated that the RTNNT/F graph, which is the weak limit graph on the PPP (or any stationary point process that does not admit a second-order descending chain and is non-equidistant), is a spanning forest. Furthermore, each connected component of this graph belongs to the $\mathscr{I}/\mathscr{I}$ class, indicating that it is one-ended.

There exists another model in the literature known as the Minimal Spanning Forest (MSF). Aldous and Steel [] introduced this graph as a natural generalization of the MSF in finite random graphs to infinite graphs. They conjectured that for the PPP, their definition of the MSF consists of a single one-ended tree. However, in this model, it is generally unknown how many components this algorithm produces and how many ends each component has (partial answers can be found in Alexander's work [35] for dimension 2). It is worth noting that the first level of the HTNNC algorithm model, i.e., $[G^0, 0]$, is a subgraph

of the MSF on the PPP. However, in the subsequent levels, due to the greedy nature of the construction, there is a significant difference between them.

In the paper by Holroyed and Peres (see [13]), they demonstrated that for each dimension d, there exists a factor graph of the Poisson point process Φ^0 that is a connected one-ended tree. A factor graph on a point process is a random graph G whose vertex set is the support of the point process, and G is a deterministic function invariant under isomorphisms on the PPP. In this sense, the limiting graph $[G, 0]$ defined on the PPP can be considered a factor graph of the PPP.

In their proof, they construct different levels $n \in \mathbb{N}$ of coarser partitions on the PPP, resulting in infinitely many finite connected components for each n. Each connected component at level n belongs to exactly one connected component at level $n+1$, and eventually, the limiting partition has only one connected component. Using this structure, they define a one-ended graph on the PPP that is a tree. However, their construction of the one-ended tree has a hierarchical nature due to the different levels of coarser partitions, but it is built analytically and does not provide an algorithm for constructing the different levels of connected components.

As an extension, Timar in [36] showed that the existence of the tree factor graph on the PPP can be extended to any stationary non-equidistant point process.

Another natural classification algorithm is based on random thinning. Let us exemplify this on a stationary point process Φ. Let $\pi = \{p_k\}$ be a probability distribution, on the positive integers. Mark independently the points of Φ with an independent random mark sampled according to π. Perform a random and independent thinning of Φ which retains the points of Φ^0 that have mark k. Declare Φ_c^k to be the cluster heads of level k, and declare descendants of a point X of Φ_c^k the points of Φ_c^{k-1} which are in the Voronoi of X w.r.t. Φ_c^k. This algorithm, which was studied in [37], requires extra randomness, whereas HTNNC does not.

Chapter

9 Proof of Theorem 8.2.3

9.1 Proof of $\mathscr{F}/\mathscr{F}$ property of the pre-limits point shift graphs

In Section 8.1, we claimed that for every $n \in \mathbb{N}$, the point shift graph, f^n, contains no infinite path (component). This section shows that the claim is true. For simplicity, it is first demonstrated that the result holds for $n = 1$. Then, it will be shown that the same result holds for all n. Before going through the proof of this, a more general kind of chain called the *Second-order descending chain* is introduced. It is shown, in Proposition 9.1.2, that this kind of chain does not exist in the Poisson point process. Proposition 9.1.5 shows that an infinite path in the f^1-graph is a second-order chain.

Definition 9.1.1. Suppose that N is a point process. Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence of different points in N. $\{x_n\}_{n\in\mathbb{N}}$ is called a *second-order descending chain* if, for all $i \geq 2$, the inequality $d_i < \max(d_{i-1}, d_{i-2})$ holds, where $d_i = \|x_{i+1} - x_i\|$.

Proposition 9.1.2. *Let Φ^0 be a d-dimensional Poisson point process. Then, Φ^0 contains no second-order descending chain.*

Proof. Suppose that there exists a second-order descending chain in Φ^0. Define $g : \mathbb{R}^d \times \mathbb{M}(\mathbb{R}^d) \to \overline{\mathbb{R}}_+$, where $g(x, \Phi)$ is the number of second-order descending chain in Φ starting at x, and $\mathbb{M}(\mathbb{R}^d)$ is the space of locally finite measure on $\mathbb{R}^d$. Campbell's formula and the Slivnyak Theorem for the Poisson point processes gives that,

$$\mathbb{E}[g(x, \Phi^0)] = \lambda \int_{\mathbb{R}^d} p(\text{existence of a second-order descending chain in } \Phi^0 \cup x, \text{ starting at } x)dx.$$

So, for the proof of the proposition, it is enough to show that the probability of having a second-order descending chain in $\Phi^0 \cup x$, starting from x, is zero. Since the distribution of $\Phi^0 \cup x$ is equal to $\Phi^0 \cup 0$, it is enough to show that, this is true for $x = 0$.
Consider the following events ,

$$A = \{ \text{ Existence of a second-order chain,} \{x_n\}_{n\in\mathbb{N}}, \text{ in } \Phi^0 \cup 0, \text{ with } x_0 = 0\},$$

and

$A_R = \{$ Existence of a second-order chain,$\{x_n\}_{n\in\mathbb{N}}$, in $\Phi^0 \cup 0$, with $x_0 = 0$, and $d_1, d_2 < R\}$.

Since $A = \cup_{R=1}^{\infty} A_R$, it is enough to show that $p(A_R) = 0$. Define $A_{R,n}$ the event of existence of a second-order descending chain,$\{x_i\}_{i=0}^{n+1}$, in $\Phi^0 \cup 0$ with length n, $x_0 = 0$, and $d_1, d_2 < R$. Then one can write $A_R = \cap_{n\geq 2} A_{R,n}$. Since for all i, $A_{R,i} \supseteq A_{R,i+1}$, so

$$p(A_R) = \lim_{n\to\infty} p(A_{R,n}). \tag{9.1.1}$$

Thus it is enough to show that $\lim_{n\to\infty} p(A_{R,n}) = 0$. Define $X_{R,n}$, for $n \geq 2$, the number of second order descending chain with length n in $\Phi^0 \cup 0$ with $x_0 = 0$, $d_1, d_2 < R$. Since $p(A_{R,n}) \leq \mathbb{E}[X_{R,n}]$, so it is enough to show that $\lim_{n\to\infty} \mathbb{E}[X_{R,n}] = 0$. Consider the function $g'(x, \Phi^0)$ equal to the number of second order descending chain with length n, $x_0 = 0$, and $x_1 = x$. Then

$$\begin{aligned}\mathbb{E}[X_{R,n}] &= \mathbb{E}[\sum_{x\in\Phi^0\cup\{0\}\cap B_R(0)} g'(x, \Phi^0 \cup \{0\})] \\ &= \lambda \int_{B_R(0)} \mathbb{E}[\, g'(x, \Phi^0 \cup \{0, x\})]\, dx\end{aligned}$$

,where the last equality fallows from Campbell formula. For the next step first note that if there is an infinite second-order descending chain in Φ^0, then there is an infinite second -order descending chain without repeated vertices. So one can consider all the event that defined up to here, to be the chain without repeated vertices. Consider the function $g''(y, \Phi)$ equal to the number of second-order descending chain with length $n(n \geq 2)$, $x_0 = 0$,$x_1 = x$, and ,$x_2 = y$, without repeated vertices. Then it gives

$$\begin{aligned}&= \lambda \int_{B_R(0)} \mathbb{E}[\sum_{y\in B_R(x)\cap(\Phi^0\cup\{0,x\})} g''(y, \Phi^0 \cup \{0, x\})]\, dx \\ &= \lambda \int_{B_R(0)} \lambda \int_{B_R(x)} \mathbb{E}[\, g''(y, \Phi^0 \cup \{0, x, y\})]\, dydx.\end{aligned}$$

The definition of the second-order descending chain, together with considering the chain without repeated vertices implies that the last equation is equal to

$$= \lambda^2 \int_{B_R(0)} \int_{B_R(x)} \mathbb{E}[\, X_{max(\|x\|,\|y-x\|),n-2}]\, dydx.$$

So there is the following recursive equation for the events $X_{R,n}$,

$$\mathbb{E}[X_{R,n}] = \lambda^2 \int_{B_R(0)} \int_{B_R(x)} \mathbb{E}[\, X_{max(\|x\|,\|y-x\|),n-2}]\, dydx. \tag{9.1.2}$$

Lemma 9.1.3. *For a fixed radius $r \in \mathbb{R}$ and for all $n \in \mathbb{N}$, let $X_{R,n}$, be the number of second-order descending chains with length n in $\Phi^0 \cup 0$ with $x_0 = 0$ and $d_1, d_2 < R$. If Φ^0 is a d-dimensional Poisson point process, then for all even value of n the following equation holds.*

$$\mathbb{E}[X_{R,n}] = \frac{(\lambda^2 \omega_d^2 R^{2d})^{n/2}}{n/2!}, \tag{9.1.3}$$

where ω_d is the volume of a unite d-dimensional ball.

Proof of Lemma 9.1.3. The proof is based on the induction. for $n = 0$, $\mathbb{E}[X_{R,0}]$, is the number of second order descending chain with length 0 in $\Phi^0 \cup 0$ with $x_0 = 0$, which is equal to 1, so the base case hold. Assume the (9.1.3) holds for an even number k, it fallows from (9.1.2)

$$\begin{aligned}\mathbb{E}[X_{R,k+2}] &= \lambda^2 \int_{B_R(0)} \int_{B_R(x)} \mathbb{E}[\, X_{max(\|x\|,\|y-x\|),k}]\, dydx \\ &= \frac{\lambda^{k+2}\omega_d^k}{k/2!} \int_{B_R(0)} \int_{B_R(x)} (\max(\|x\|, \|y-x\|))^{kd} dydx \end{aligned} \tag{9.1.4}$$

The inner integral is equal to

$$\int_{B_R(x)} (\max(\|x\|, \|y-x\|))^{kd} dy \tag{9.1.5}$$

$$\begin{aligned} &= \int_{\theta \in s^{n-1}} \int_0^R (\max(\|x\|, \|r\|))^{kd} r^{d-1} dr d\theta \\ &= \int_{\theta \in s^{n-1}} (\int_0^{\|x\|} \|x\|^{kd} r^{d-1} dr + \int_{\|x\|}^R r^{(k+1)d-1} dr)\, d\theta \\ &= \int_{\theta \in s^{n-1}} (\int_0^{\|x\|} \|x\|^{kd} r^{d-1} dr + \int_{\|x\|}^R r^{(k+1)d-1} dr)\, d\theta \\ &= \int_{\theta \in s^{n-1}} (\frac{R^{(k+1)d}}{(k+1)d} + \frac{\|x\|^{(k+1)d} \times k}{(k+1)d})\, d\theta \\ &= \frac{\omega_d}{k+1} (\, R^{(k+1)d} + k\|x\|^{(k+1)d}) \end{aligned} \tag{9.1.6}$$

Equation (9.1.4) together with (9.1.5) gives

$$\begin{aligned}\mathbb{E}[X_{R,k+2}] &= \frac{\lambda^{k+2}\omega_d^{k+1}}{(k+1)(k/2!)} \int_{B_R(0)} (\, R^{(k+1)d} + k\|x\|^{(k+1)d}) dx \\ &= \frac{\lambda^{k+2}\omega_d^{k+1}}{(k+1)(k/2!)} (\int_{B_R(0)} R^{(k+1)d} dx + \int_{B_R(0)} k\|x\|^{(k+1)d} dx) \\ &= \frac{\lambda^{k+2}\omega_d^{k+1}}{(k+1)(k/2!)} (\, \omega_d R^{(k+2)d} + \frac{k}{k+2} \omega_d R^{(k+2)d}) \\ &= \frac{\lambda^{k+2}\omega_d^{k+2}}{\frac{k+2}{2}!} R^{(k+2)d}, \end{aligned}$$

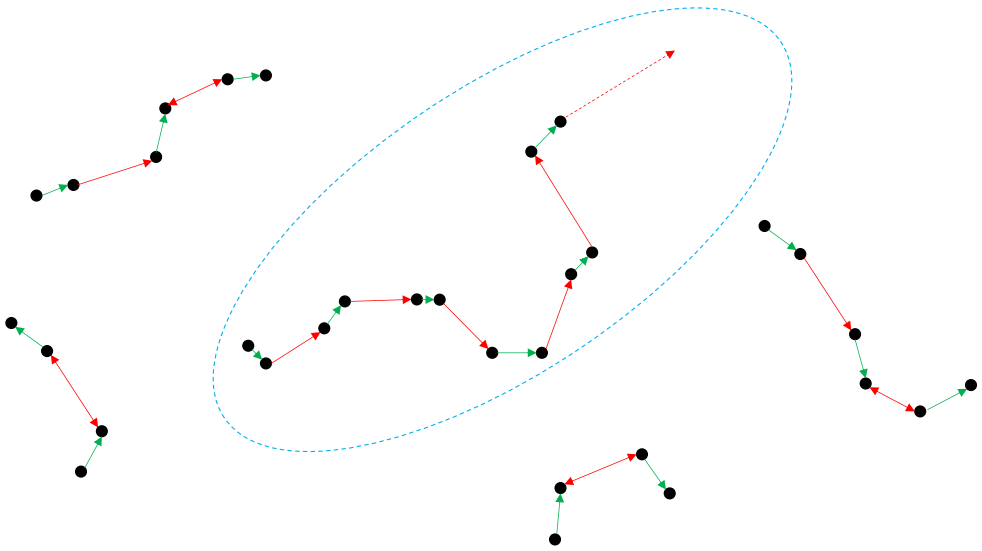


Figure 9.1: The behavior of infinite path in Ψ_c^1 if it exists. The green arrows show $NN0$ links, and the red arrows show NN1 links between the exit points of f^0-cycles and their NN1 points.

which gives the result. □

Con. of the proof of Proposition 9.1.2. Due to the definition of $X_{R,n}$, the relation $X_{R,i} > X_{R,i+1}$, holds for each i. So $\mathbb{E}[X_{R,i}] > \mathbb{E}[X_{R,i+1}]$. This fact, together with Lemma **??**, gives that $\lim_{n\to\infty} \mathbb{E}[X_{R,n}] = 0$, which completes the proof. □

Definition 9.1.4. The definition of the point-shift f^1 implies that the f^1-graph, as a directed graph, includes two distinct types of edges on Φ_c^0. The first type is the edges between the nearest neighbors' points w.r.t. d. These edges are referred to as d-type edges (the green edges in Figure 9.1). The other type is the edges that connect an exit point, x, of level zero to $NN1(x)$. These edges are referred to as δ-type edges of level one (the red edges in Figure 9.1).

Proposition 9.1.5. *Any infinite path in f^1 -graph is a second-order descending chain.*

Proof. Let $\mathscr{P}$ be an infinite path in f^1-graph. Since there is no descending (See Section 3.1.3) in the Poisson point process (see cite), this infinite path must be on Ψ^1. In other words, if there exists an infinite path in the f^1-graph, this path is a descending chain between the pairs of points in Ψ^1. The edges that belong to $\mathscr{P}$ are either of type d or δ, which means they are either between two mutual nearest neighbor points in Φ^0 or between nearest neighbors of the MNN pairs in Ψ^1. (see Figure 9.1).

Let e_i be the i-th edge in $\mathscr{P}$. Then e_i can be d or δ-type of edge. First, suppose that for an arbitrary i, e_i be a d-type edge. Then, there are two possibilities for e_{i-1} and e_{i-2}:

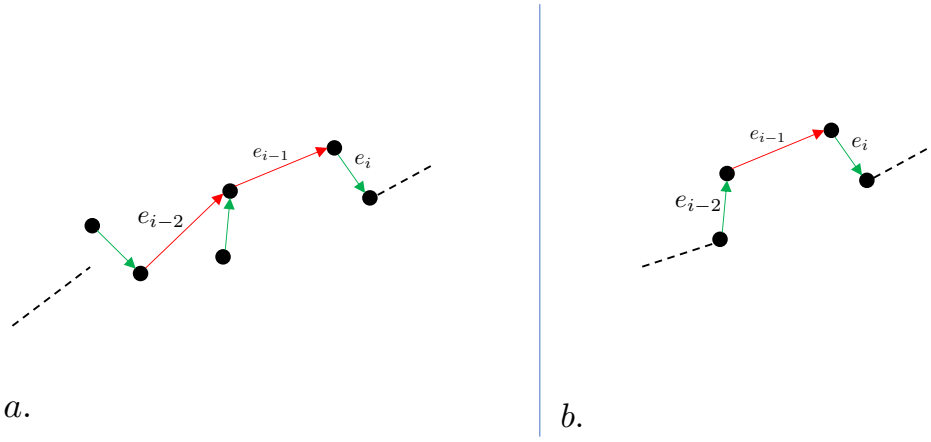


Figure 9.2: Different possibilities for e_{i-1} and e_{i-2} conditioned on e_i is a d^0-type edge in the infinite path of the f^1-graph if it exists.

1. e_{i-1} and e_{i-2} are δ-type edges (see Figure 9.2.a). In this case, since the δ-type edges are consequence edges in the descending chin on Ψ^1, $\|e_{i-1}\| < \|e_{i-2}\|$. And since e_i is the MNN0 edge so $\|e_i\| < \|e_{i-1}\|$. Thus $\|e_i\| < \max(\|e_{i-1}\|, \|e_{i-2}\|)$.

2. e_{i-1} is δ-type edge and e_{i-2} is d-type edge (see Figure 9.2.b).Since e_{i-2}, and e_i both are between MNN0 points it gives

$$\|e_{i-1}\| > \|e_i\|, \|e_{i-2}\| \implies \|e_i\| < \max(\|e_{i-1}\|, \|e_{i-2}\|). \tag{9.1.7}$$

On the other hand, if e_i is a δ-type edge, then there are three possibilities for e_{i-1} and e_{i-2}. The first is that both are δ-type edges (see Figure 9.3.a). In this case, $\|e_i\| < \max(\|e_{i-1}\|, \|e_{i-2}\|)$ since these edges are in descending chain in Ψ^1. The second possibility is that e_{i-1} is a d-type edge, and e_{i-2} is a δ-type edge (see Figure 9.3.b), and it gives

$$\|e_{i-2}\| > \|e_i\|, \quad \& \quad \|e_{i-2}\| > \|e_{i-1}\| \implies \|e_i\| < \max(\|e_{i-1}\|, \|e_{i-2}\|), \tag{9.1.8}$$

where the first inequality is because of the descending chain property, and the second inequality holds since e_{i-1} is the MNN0 edge between points. The last possibility is that e_{i-1} is a δ-type edge, and e_{i-2} is a d-type edge (see Figure 9.3.c), it is easy to show that it this case also the inequality $\|e_i\| < \max(\|e_{i-1}\|, \|e_{i-2}\|)$ holds and this completes the proof. □

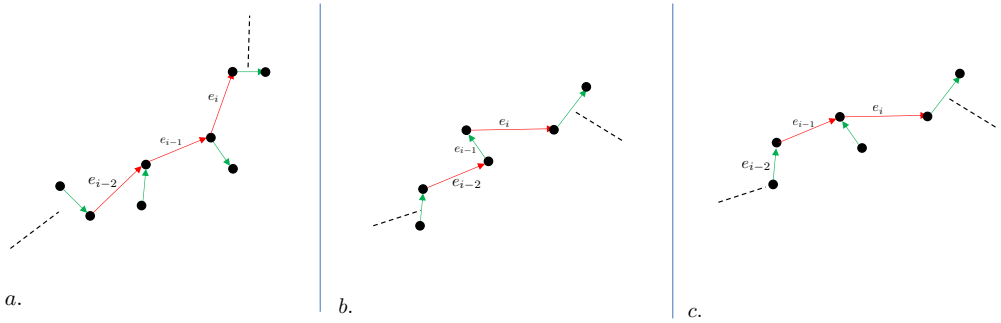


Figure 9.3: Different possibilities for e_{i-1} and e_{i-2} conditioned on e_i is a d^0-type edge in the infinite path of the f^1-graph if it exists.

Corollary 9.1.6. *Proposition 9.1.2, together with proposition 9.1.5, gives that there is no infinite path in the f^1-graph of Φ^0 where Φ^0 is a Poisson Point Process.*

So far, it has been shown that there is no infinite path or cluster in the f^1-graph. The following proposition establishes that this holds true for all f^n-graphs.

Proposition 9.1.7. *For all $n \in \mathbb{N}$, any infinite path in f^n -graph is a second-order descending chain.*

Proof. The proof is by induction. Any infinite path in the f^0-graph is a descending chain in the PPP. Therefore, the base case of the induction is true. For the induction step, suppose there is no infinite path (cluster) in the f^n-graph. Then, if an infinite path P^{n+1} exists in the f^{n+1}-graph, this path lies on Ψ^{n+1}. Thus, there are two types of edges on P^n: the first type is the ones between the $MNN(n)$ points, and the second type is the edges that connect an exit point of level n, x, to $NN(n+1)(x)$. In fact, due to the construction of the f^{n+1} point shift, P^{n+1} forms a descending chain between the f^n-cycles $\{S_n^i\}$. Therefore, one can see the same proof as in Proposition 9.1.5 shows that if P exists, it is a second-order descending chain. □

Chapter

10 Other Examples

10.1 HTNN Algorithm on Cox Point Process

In this section, some examples of clustered point processes are considered. The term "clustered" refers to the fact that the points in these processes can be classified into distinct groups.

Example 10.1.1 (Cox point process)**.** A point process N on a state space $\mathscr{S}$ is called a Cox point process driven by the intensity measure μ, if the conditional distribution of N given μ is a Poisson point process with intensity function μ. Let $\{\xi(x)\}_{x\in\mathscr{S}}$ be a non-negative random field. Then the random-driven measure μ can be defined as

$$\mu(B) = \int_B \xi(x)dx \quad B \subseteq \mathscr{S}. \tag{10.1.1}$$

Consider a deterministic collection of points, $M \in \mathbb{R}^2$, and a fixed radius, $r \in \mathbb{R}$. Let $\widetilde{M}= \{x \in B_r(y); y \in M\}$ and define $\{\xi(y)\}$ as independent Bernoulli random variables with parameter p, and for $x \notin \widetilde{M}, \xi(x) = 0$. Consider the cox point process N' driven by intensity μ defined by (10.1.1), with $\xi(x) = \xi'(x)$. N' is a sub-point process of Poisson point process with intensity p and therefore does not admit any second-order descending chain. The results of Chapter 8 are valid for this point process. According to its definition, N is clustered.

Chapter

11 Appendix

11.0.1 Proof of Proposition 5.2.3

Before going through the proof of this proposition, the first Definition 11.0.1 and Lemma 11.0.2, borrowed from [**?**], are discussed. This last lemma gives the main idea of the proof of Proposition 5.2.3.

Definition 11.0.1. Let μ, ν, and γ be given probability measures on $\mathbb{Z}$. Consider the Markov Chain $\{Y_n\}$ with values in $\mathbb{Z}$ such that $Y_0 = y$ and it has the following transition probabilities:

$$P(Y_{n+1} = k | Y_n = j) = \begin{cases} \mu(k-j) & \text{if } j < 0 \\ \nu(k-j) & \text{if } j > 0 \\ \gamma = \alpha\mu(k) + \beta\nu(k) & \text{if } j = 0\,, \end{cases} \tag{11.0.1}$$

where $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$. This Markov Chain is an ordinary random walk on $\mathbb{Z}$ with jump distribution μ in the positive integers, distribution ν in the negative integers, and γ at 0. This random walk will be referred to as the **oscillating random walk** on $\mathbb{Z}$.
The particular case where $\nu(i) = \mu(-i)$ is called the **anti-symmetric oscillating random walk** . If moreover $\mu(j) = 0$ for $j < 0$, then it is called the **one-sided anti-symmetric** case.

The following lemma from [37] will be used to study the recurrence and transience property of the oscillating random walk.

Lemma 11.0.2. *Consider the one-sided antisymmetric oscillating random walk $\{Y_n\}$ on $\mathbb{Z}$. Then a sufficient condition for zero to be recurrent is that*

$$\sum_{j=n}^{\infty} \mu(j) = O(n^{-\frac{1}{2}}) \quad \text{as } n \to \infty. \tag{11.0.2}$$

A sufficient condition for zero to be transient is

$$\mu(n) \sim cn^{-1-\epsilon} \quad \text{as } n \to \infty, \tag{11.0.3}$$

where c and ϵ denote positive constants, $\epsilon < \frac{1}{2}$.

Proof of Proposition 5.2.3. Consider two independent i.i.d. random sequences

$$\{X_i\}_{i\in\mathbb{N}^*}, \& \quad \{X_i\} \sim \eta$$
$$\{X'_i\}_{i\in\mathbb{N}^*}, \& \quad \{X'_i\} \sim \eta$$

and two random walks on $\mathbb{Z}$ with jumps $\{X_i\}$ and $\{X'_i\}$ respectively, with two arbitrary different starting points X_0 and X'_0, namely,

$$S^l = \sum_{i=0}^{l}\{X_i\} \qquad S'^l = \sum_{i=0}^{l}\{X'_i\}. \tag{11.0.4}$$

The paths created by these two random walks are two paths in G^η starting from the two vertices X_0 and X'_0. For checking the connectedness of G^η, it is needed to check whether these two random walks meet each other in finite time a.s. or not. To this end, the MC $\{Z_n\}_{n\in\mathbb{N}}$ will be considered. For $X_0 < X'_0$ define $Z_0 = X'_0 - X_0$. Moreover fix the S'^l at X'_0 and define Z_i, the difference between X'_0 and X_i up to the time that X_i passes X'_0, i.e.,

$$Z_i = X'_0 - X_i, \quad \text{for } 0 < i \leq t_1,$$

where t_1 is the first t such that $X_t > X'_0$. Then fix S^l at X_{t_1} and look at the next steps of S'^l. For $i > t_1$ define Z_i, the difference between X_{t_1} and X'_i up to the time that X'_i passes X_{t_1}, i.e.,

$$Z_i = X'_i - X_{t_1}, \quad \text{for } t_1 < i \leq t_2,$$

where t_2 is the first time where $X'_t > X_{t_1}$. After that, fix X'_{t_2} and continue this process. With this definition, $\{Z_n\}_{n\in\mathbb{N}}$ is a random walk on $\mathbb{Z}$ which has following transition probability

$$P(Z_{n+1} = k | Z_n = j) = \begin{cases} q_{k-j} & \text{if } j < 0 \text{ and } k > j \\ q_{j-k} & \text{if } j \geq 0 \text{ and } k < j \\ 0 & \text{otherwise ,} \end{cases} \tag{11.0.5}$$

where $\{q_i\}$ is the probability defined in (5.2.5). Our question about the meeting of the two random walks S^l and S'^l reduces to understanding whether the state 0 in $\{Z_n\}_{n\in\mathbb{N}}$ is recurrent or not. But $\{Z_n\}_{n\in\mathbb{N}}$ is a one-sided antisymmetric oscillating random walk where $\mu(j)$ in (11.0.1) is equal to q_j and $\beta = 1$. So

$$\sum_{j=n}^{\infty} q_j = \sum_{j=n}^{\infty} \frac{c_1}{j^{\alpha+1}},$$

which is $O(n^{-\frac{1}{2}})$, as $n \to \infty$, when $\alpha \geq \frac{1}{2}$. So using Lemma 11.0.2, one can conclude

that $\{Z_n\}$ is recurrent when $\alpha \geq \frac{1}{2}$, and it is transient when $\alpha < \frac{1}{2}$. So the two random walks $\{S^l\}$ and $\{S'^l\}$ will meet each other a.s. when $\frac{1}{2} \leq \alpha < 1$ and, in this case, G^η is a Renewal EFT. Correspondingly, when $\alpha < \frac{1}{2}$, G^ξ is a Renewal EFF. $\square$

Bibliography

[1] François Baccelli, Mir-Omid Haji-Mirsadeghi, and Sayeh Khaniha. Coupling from the past for the null recurrent markov chain. 2022. 1, 7, 29

[2] D. Aldous and R. Lyons. Processes on unimodular random networks. *Electronic Journal of Probability*, 12:1454–1508, 2007. 1, 6, 7, 13, 20, 21, 32, 34, 44, 53, 81

[3] F. Baccelli, M.-O. Haji-Mirsadeghi, and A. Khezeli. Eternal family trees and dynamics on unimodular random graphs. *Contemporary Mathematics*, pages 85–127, 01 2018. 1, 7, 20, 21, 22, 29, 45, 50, 51, 52, 55, 70, 89, 92

[4] F. Baccelli, M.-O. Haji-Mirsadeghi, and J.T. Murphy. Doeblin trees. *Electronic Journal of Probability*, 24:1 – 36, 2019. 2, 8, 29, 32, 34, 35, 43, 59, 70

[5] Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning*. Springer, 2009. 3, 10, 24

[6] Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, pages 354–359, 1983. 4, 10, 24, 81

[7] Pedro Contreras Fionn Murtagh. Algorithms for hierarchical clustering: an overview, ii. *WIREs Data Mining and Knowledge Discovery*, 7(6):e1219, 2017. 4, 10, 24, 81

[8] Mohammad-Omar Haji-Mirsadeghi and Francois Baccelli. Point-shift foliation of a point process. *Electronic Journal of Probability*, 23:1–25, 2018. 6, 13, 89

[9] François Baccelli, Bartlomiej Blaszczyszyn, and Mohamed Karray. *Random Measures, Point Processes, and Stochastic Geometry*. INRIA, 2020. 16

[10] D. Stoyan, W. S. Kendal, and J. Mecke. *Stochastic Geometry and Its Applications*. John Wiley & Sons, New York., 1987. 16, 17, 67

[11] Philip M. Dixon. Ripley's k function. *John Wiley & Sons*, 2014. 17

[12] Ronald Meester Olle Haggstrom. Nearest neighbor and hard sphere models in continuum percolation. *Random Structures Algorithms*, 9:295–315, 1996. 18

[13] Alexander Holroyd and Yuval Peres. Trees and matchings from point processes. *Electronic Communications in Probability*, 8:17–27, 2003. 18, 82, 92, 93

[14] Daryl J. Daley and Günter Last. Descending chains, the lilypond model, and mutual-nearest-neighbour matching. *Advances in Applied Probability*, 37(3):604–628, 9 2005. 18, 84

[15] Torgny Lindvall. *Lectures on the Coupling Method*. John Wiley & Sons Inc, Septembre 1992. 19

[16] Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, 2000. 19

[17] Yuval Peres David A. Levin. *Markov Chains and Mixing Times*. American Mathematical Society;, 2017. 19

[18] T. Hutchcroft. Non-intersection of transient branching random walks. *Probability Theory and Related Fields*, 178, 10 2020. 23, 53

[19] J.-P. Benzécri. Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Les cahiers de l'analyse des données*, 7(2):209–218, 1982. 25

[20] J Juan. Programme de classification hiérarchique par l'algorithme de la recherche en chaîne des voisins réciproques. *Les cahiers de l'analyse des données*, 7(2):219–225, 1982. 25

[21] David Bruce Wilson James Gary Propp. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, August-September 1996. 31, 32, 61, 71

[22] S. G. Foss A. A. Borovkov. Stochastically recursive sequences and their generalizations. *Trudy Inst. Mat. SO RAN*, 20:32–103, 1993. 31, 32, 34

[23] Sergey Foss and Alexandr Borovkov. Two ergodicity criteria for stochastically recursive sequences. *Acta Applicandae Mathematicae*, 34(1-2):125–134, 1994. 31, 34

[24] S.G. Fosstand R.L. Tweedi. Perfect simulation and backward coupling. 1998. 32

[25] P. Brémaud. *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, New York, NY, 1999. 42

[26] N. Broutin, L. Devroye, and G. Lugosi. btractive random forests. *arXiv:2210.10544*, 2022. 46

[27] D. Daley. Stochastically monotone Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 10(4):305–317, December 1, 1968. 60

[28] J.A. Fill and M. Machida. Stochastic monotonicity and realizable monotonicity. *The Annals of Probability*, 29(2):938–978, 2001. 61

[29] J. Bentley and A. Chi-Chih Yao. An almost optimal algorithm for unbounded searching. *Information Processing Letters*, 5(3):82–87, 1976. 64

[30] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(3):490–520, 1962. 64

[31] O. Kallenberg. *Random Measures, Theory and Applications*, volume 77. Springer, 2017. 73

[32] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 2005. 73

[33] Frédéric Ros and Serge Guillaume. Munec: A mutual neighbor-based clustering algorithm. *Information Sciences*, 486:148–170, 2019. 81

[34] J. Michael Steele and David Aldous. Asymptotics for euclidean minimal spanning trees on random points. *Probability Theory and Related Fields*, 92:247–258, 1992. 82

[35] Kenneth S. Alexander. Percolation and minimal spanning forests in infinite graphs. *The Annals of Probability*, 23(1):87–104, 1995. 82, 92

[36] Adam Timar. Tree and grid factors of general point processes. *Electronic Communications in Probability*, 9, September 2009. 82, 93

[37] Sergei Zuyev and Konstantin Tchoumatchenko. Aggregate and fractal tessellations. *Probability Theory and Related Fields*, 121:198–218, 2001. 93, 103

RÉSUMÉ

Cette thèse repose sur la notion d'unimodularité dans le contexte des réseaux aléatoires et explore deux domaines d'application : le couplage par le passé des chaînes de Markov dans le cas de la récurrence nulle, basé sur les graphes de Doeblin associés, et la classification non supervisée basée sur le regroupement hiérarchique des points d'un processus ponctuel.

La première partie de cette thèse se concentre sur les propriétés d'un graphe aléatoire spécifique appelé le graphe de Doeblin, qui est associé à l'algorithme du couplage par le passé utilisé pour l'échantillonnage parfait de la distribution stationnaire d'une chaîne de Markov. Cette thèse étudie le cas récurrent nul, où il est montré que le graphe des ponts, un sous-graphe du Doeblin Graph, est soit un arbre infini, soit une forêt composée d'une collection dénombrable d'arbres infinis. Dans le premier cas, l'arbre infini possède une unique extrémité, n'est généralement pas unimodularisable, mais présente une unimodularité locale. Ces propriétés sont exploitées pour étudier le régime stationnaire des dynamiques aléatoires de processus à valeurs mesures sur l'arbre des ponts de Doeblin, en particulier les dynamiques aléatoires tabou et potentielle.

La deuxième partie de cette thèse présente un nouveau modèle de regroupement hiérarchique adapté à la classification non supervisée d'ensembles de données qui sont dénombrablement infinis. L'algorithme proposé utilise plusieurs niveaux de regroupement, construisant des clusters à chaque niveau en utilisant des chaînes de voisins les plus proches de points ou de clusters. Cet algorithme est appliqué au processus ponctuel de Poisson pour lequel il est démontré que l'algorithme de regroupement définit une forêt phylogénétique, qui est un facteur du processus ponctuel et est donc unimodulaire. Diverses propriétés de cette forêt aléatoire, telles que les tailles moyennes des clusters à chaque niveau ou la taille moyenne du cluster d'un nœud typique, sont examinées.

MOTS CLÉS

Graphes Unimodulaires aléatoires ⋆ Processus ponctuels stationnaires ⋆ classification des feuillets ⋆ Échantillonnage parfait⋆ Classification non supervisée hiérarchique⋆ Processus ponctuels⋆ Arbres Aléatoires

ABSTRACT

This thesis is based on the notion of unimodularity in the context of random networks and explores two domains of application: Coupling from the Past for Markov Chains in the null recurrent case based on the associated Doeblin Graphs, and unsupervised classification based on hierarchical clustering on point processes.

The first part of this thesis focuses on the properties of a specific random graph called the Doeblin Graph, which is associated with the Coupling from the Past algorithm used for the perfect sampling of the stationary distribution of a Markov Chain. This thesis studies the null recurrent case, where it is shown that the Bridge Doeblin Graph, a subgraph of the Doeblin Graph, is either an infinite tree or a forest composed of a countable collection of infinite trees. In the former case, the infinite tree possesses a single end, is not generally unimodularizable, but exhibits local unimodularity. These properties are leveraged to investigate the stationary regime of measure-valued random dynamics on the Bridge Doeblin Tree, particularly the taboo and potential random dynamics.

The second part of this thesis introduces a novel hierarchical clustering model tailored for unsupervised classifications of datasets that are countably infinite. The proposed algorithm employs multiple levels of clustering, constructing clusters at each level using nearest-neighbor chains of points or clusters. This algorithm is applied to the Poisson point process. It is proven that the clustering algorithm defines a phylogenetic forest on the Poisson point process, which is a factor of the point process and is therefore unimodular. Various properties of this random forest, such as the mean sizes of clusters at each level or the mean size of the cluster of a typical node, are examined.

KEYWORDS

Unimodular Random graphs⋆Stationary Point Processes⋆Foil Classification⋆Perfect Sampling⋆Unsupervised Classification⋆Point Processes⋆Random Trees