Statistical Learning and Model Criticism for Networks and Point Processes

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Jiasen Yang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF DISSERTATION APPROVAL

Prof. Jennifer Neville, Chair

Departments of Computer Science and Statistics

Prof. Vinayak Rao

Department of Statistics

Prof. Petros Drineas

Department of Computer Science

Prof. David Gleich

Department of Computer Science

Prof. Hao Zhang

Department of Statistics

Approved by:

Prof. Jun Xie

Head of the Statistics Graduate Program

Essentially, all models are wrong, but some are useful.

— George E. P. Box (1919–2013)

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

SYMBOLS

$\mathbb{R}$	Real numbers
$\mathbb{R}_+$	Non-negative real numbers
$\mathcal{X}$	Discrete or continuous domain (Chapter 1–5); sampled configuration from a point process (Chapter 6)
$\mathbb{X}$	Ground space of a point process (typically $\mathbb{X} \subseteq \mathbb{R}^d$)
$\mathcal{N}_{\mathbb{X}}$	Space of finite point configurations on $\mathbb{X}$
$N(t)$	Counting process
$\lambda(t)$	Intensity (rate) function of a point process
$\lambda(t\|\mathcal{H}_t)$	Conditional intensity function given history $\mathcal{H}_t$
ϕ, ψ	Point configurations (counting measures) (Section 2.2.2 and Chapter 6)
Φ, Ψ	Point processes (random counting measures)
δ_x	Dirac measure at x (Chapter 6); Dirac evaluation functional at x (Section 3.1)
$\rho(x\|\phi)$	Papangelou conditional intensity at point x given configuration ϕ
$\mathscr{F}$	Function space
$\mathscr{H}$	Reproducing kernel Hilbert space (RKHS) with inner-product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ and norm $\| \cdot \|_{\mathscr{H}}$
$k(\cdot, \cdot)$	Positive-definite kernel
MMD	Maximum mean discrepancy
$\mathcal{A}$	Stein operator
$\mathcal{L}, \mathcal{T}$	Linear operators
$\mathbb{D}(\cdot \,\|\, \cdot)$	Stein discrepancy
$\mathbb{S}(\cdot \,\|\, \cdot)$	Square of Stein discrepancy
$\mathcal{N}(\cdot, \cdot)$	Normal distribution

$\xrightarrow{\mathcal{D}}$	Convergence in distribution
$\neg, \neg$	Cyclic permutation and inverse permutation
Δ, Δ^*	Difference operators with respect to $\neg$ and $\neg$
$\mathcal{D}^+, \mathcal{D}^-$	Inclusion and exclusion functionals
$\mathbb{I}\{\cdot\}$	Indicator function

ABBREVIATIONS

i.i.d.	Independent and identically distributed
ERGM	Exponential random graph model
HP	Hawkes process model
PLS	Poisson-rate latent space model
BLS	Hawkes base-rate latent space model
RLS	Hawkes reciprocal latent space model
DLS	Hawkes dual latent space model
RKHS	Reproducing kernel Hilbert space
p.d.	Positive-definite
IPM	Integral probability metric
MMD	Maximum mean discrepancy
KSD	Kernelized Stein discrepancy
KDSD	Kernelized discrete Stein discrepancy

ABSTRACT

Yang, Jiasen Ph.D., Purdue University, August 2019. Statistical Learning and Model Criticism for Networks and Point Processes. Major Professor: Jennifer Neville.

Networks and point processes provide flexible tools for representing and modeling complex dependencies in data arising from various social and physical domains. Graphs, or networks, encode *relational* dependencies between entities, while point processes characterize *temporal* or *spatial* interactions among events.

In the first part of this dissertation, we consider dynamic network data (such as communication networks) in which links connecting pairs of nodes appear continuously over time. We propose latent space point process models to capture two different aspects of the data: (*i*) communication occurs at a higher rate between individuals with similar latent attributes (*i.e.*, homophily); and (*ii*) individuals tend to reciprocate communications from others, but in a varied manner. Our framework marries ideas from point process models, including Poisson and Hawkes processes, with ideas from latent space models of static networks. We evaluate our models on several real-world datasets and show that a *dual* latent space model, which accounts for heterogeneity in both homophily and reciprocity, significantly improves performance in various link prediction and network embedding tasks.

In the second part of this dissertation, we develop nonparametric goodness-of-fit tests for discrete distributions and point processes that contain intractable normalization constants, providing the first generally applicable and computationally feasible approaches under those circumstances. Specifically, we propose and characterize Stein operators for discrete distributions, and construct a general Stein operator for point processes using the Papangelou conditional intensity function. Based on the proposed Stein operators, we establish *kernelized Stein discrepancy* measures for discrete distributions

and point processes, which enable us to develop nonparametric goodness-of-fit tests for unnormalized density/intensity functions. We apply the kernelized Stein discrepancy tests to discrete distributions (including network models) as well as temporal and spatial point processes. Our experiments demonstrate that the proposed tests typically outperform two-sample tests based on the maximum mean discrepancy, which, unlike our goodness-of-fit tests, assume the availability of exact samples from the null model.

1. INTRODUCTION

Machine learning has proven very successful in the analysis of independent and identically distributed (*i.i.d.*) data. Much of the recent research in machine learning and statistics focuses on developing models and methods that recognize and exploit dependencies and heterogeneities in various aspects of the data. Two classes of models will play an import role in this dissertation—network models for describing relational dependencies, and point processes for characterizing temporal/spatial dependencies and heterogeneities. The ability to capture sophisticated dependencies in data comes at a cost of increased model complexity, rendering it difficult to assess the statistical quality of the model fit. The contributions of this dissertation are twofold: we propose latent space point process models for dynamic network data, and we develop statistical *goodness-of-fit* tests for complex models involving intractable normalization constants, with examples including network models and point processes.

1.1 Modeling Dependencies with Networks and Point Processes

Graphs, or networks, represent relational data that arise naturally in various social, physical, and biological domains. Examples include social networks consisting of users and their friendships, citation networks comprising articles with their co-citations, protein interaction networks characterizing the physical contacts among protein molecules, communication networks recording messages sent between individuals, and the World Wide Web containing web pages and the hyperlinks between them.

Point pattern data, consisting of the locations of objects in some ambient space, occur widely in the natural and social sciences. Point process models have been applied to describe stars and galaxies (Babu and Feigelson, 1996), trees in a forest (Diggle, 2003),

earthquakes and aftershocks (Ogata, 1988), neurons in the brain (Linderman et al., 2014), and the dynamics of crime (Linderman and Adams, 2014).

In many applications, we encounter *dynamic* networks whose structures evolve over time. For example, in a communication network, links encoding messages sent among individuals (nodes) are recorded over a continuous period with their timestamps. Such temporal information offers valuable insight into the evolution of the network structure and the underlying relationships among individuals. In Chapter 4, we investigate two main characteristics governing the appearance of links in a dynamic network: (*i*) communication occurs at a higher rate between individuals with similar latent attributes—an observation referred to as *homophily* in the social science literature; and (*ii*) individuals tend to *reciprocate* communications from others, but in a manner that varies across different individuals (see Figure 1.1 for an illustration). To capture both characteristics, we employ latent space models to learn hidden node attributes underlying the network, and point processes including Poisson and Hawkes processes to model the temporal dynamics of link generation. Specifically, we model the communications between each pair of nodes as realizations from a point process whose intensity function contains two components: (*i*) a baseline rate depending on the distance between the nodes in a *homophily latent space*; and (*ii*) a reciprocation rate depending on the positions of the nodes in a different *reciprocal latent space*. Through careful ablation experiments on several real-world networks, we show that such a *dual* latent space model (depicted in Figure 1.2) achieves superior performance over models with a single latent space in a variety of link prediction and network embedding tasks, and allows one to decouple the influences of homophily and reciprocity in temporal interactions.

1.2 Statistical Model Criticism and Goodness-of-Fit Testing

Statistical techniques for model checking and diagnostics have lagged behind the development of increasingly sophisticated machine learning models. Such techniques are essential for us to gauge the utility and defects of opaque models, to enhance

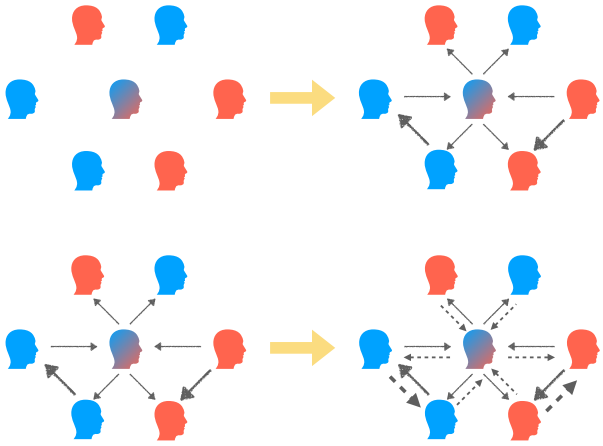


Figure 1.1.: Influences of homophily (*top*) and reciprocity (*bottom*) in social interactions.

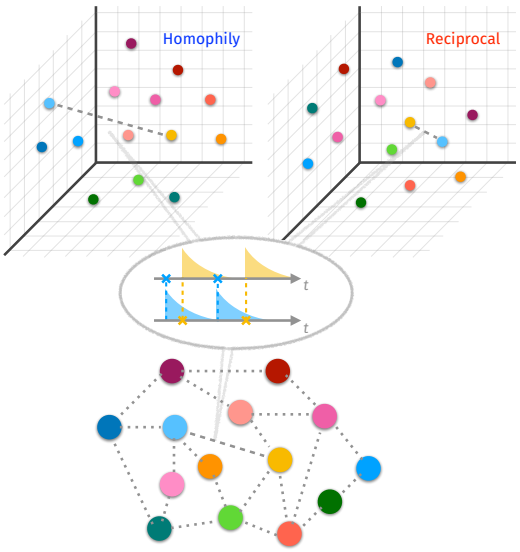


Figure 1.2.: Dual latent space model.

their interpretability, and to identify aspects for their improvement. Statistical *model criticism*, a term attributed to George Box, refers to the process of assessing how well a model fits the observed data, without explicit reference to alternative models or assumptions (O'Hagan, 2003). As Box (1976); Box and Draper (1986) famously noted, "all models are wrong, but some are useful"; one could never validate whether a model is true, but one could attempt to measure the degree to which it falsely describes the data (Gelman and Shalizi, 2013). By identifying deficiencies in the current model, criticism leads to a *revised* model, which could again be subject to criticism, and the process continues until a satisfactory model is found (Seth et al., 2018). This iterative process constitutes part of the data-analysis cycle, as illustrated in Figure 1.3, which Blei (2014) termed *Box's loop* based on the ideas in Box (1976, 1980).

A fundamental and well-studied model criticism technique is the *goodness-of-fit test*. Classical goodness-of-fit tests, such as the χ^2 test (Pearson, 1900); the Kolmogorov–Smirnov test (Kolmogorov, 1933; Smirnov, 1948); and the Anderson–Darling test (Anderson and Darling, 1954), have become essential tools in the practitioner's toolbox. These existing tests typically require the model distribution to be fully specified. In modern applications, however, the distribution is often known only up to an intractable

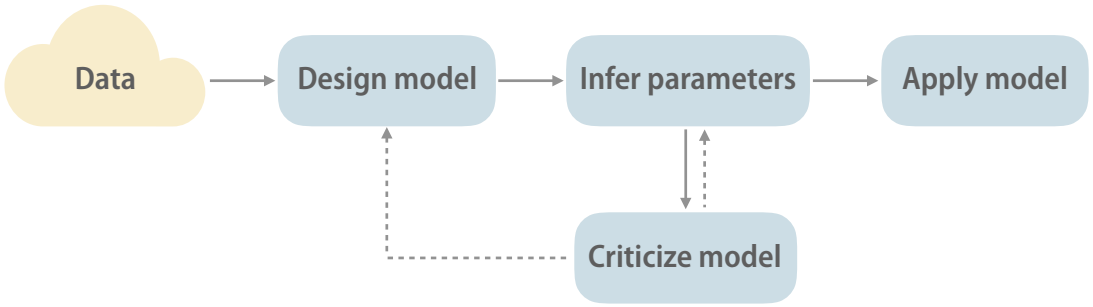


Figure 1.3.: The iterative process of data analysis; adapted from Box's loop (Blei, 2014).

normalization constant; examples include large-scale graphical models, deep generative models, and statistical models for network data. While a variety of approximate inference techniques such as Markov chain Monte Carlo (MCMC) and variational methods have been studied to allow learning and inference in these models, it is usually hard to quantify the approximation errors involved, rendering it difficult to establish statistical tests with calibrated uncertainty estimates.

In the case of point processes, well-established goodness-of-fit tests are only available under the simplest scenarios—such as when the null model is a Poisson process. For more general point processes, the construction of such tests typically rely on pseudo-likelihood approximations (Strauss and Ikeda, 1990) which introduce biases and errors that are hard to quantify, or heuristic summary statistics (such as Ripley's K-function) which only capture certain aspects of the observed data and may lead to a considerable loss of statistical power. For widely used models that capture pairwise or higher-order dependencies between points (*e.g.*, Gibbs processes), their density/intensity functions can often be evaluated only up to a normalization constant, because summing over all possible configurations leads to an intractable infinite-dimensional integral.

Recently, a new line of research (Gorham and Mackey, 2015; Oates et al., 2017; Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017; Jitkrittum et al., 2017) has developed goodness-of-fit tests which work directly with unnormalized model distributions. Central to these tests is the notion of a *Stein operator*, originating from

Stein's method (Stein, 1972, 1986) for characterizing convergence in distribution. Given a distribution $p(\boldsymbol{x})$ on $\mathcal{X}^d$ and a class of test functions $f \in \mathscr{F}$ on $\mathcal{X}^d$, a Stein operator $\mathcal{A}_p$ satisfies $\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p f(\boldsymbol{x})\right] = 0$, so that when $\mathcal{A}_p$ is applied to any test function f, the resulting function $\mathcal{A}_p f$ has zero-expectation under p. Additionally, the expectation under any other distribution $q \neq p$ should be non-zero for at least some function f in $\mathscr{F}$. When $\mathscr{F}$ is sufficiently rich, the maximum value $\sup_{f\in\mathscr{F}} \mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p f(\boldsymbol{x})\right]$ serves as a discrepancy measure, called *Stein discrepancy*, between distributions p and q.

The properties of the Stein discrepancy measure depends on two objects: the Stein operator $\mathcal{A}_p$, and the set $\mathscr{F}$. Different authors have studied different choices of $\mathscr{F}$: Gorham and Mackey (2015) considered test functions in the $\mathcal{W}^{2,\infty}$ Sobolev space, and the resulting test statistic requires solving a linear program under certain smoothness constraints. On the other hand, Oates et al. (2017); Chwialkowski et al. (2016); Liu et al. (2016) proposed taking $\mathscr{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS), which leads to test statistics that can be computed in closed form and with time quadratic in n, the number of samples.

Regarding the choice of the Stein operator $\mathcal{A}_p$, all the aforementioned works consider the case when $\mathcal{X} \subseteq \mathbb{R}$ is a continuous domain, $p(\boldsymbol{x})$ is a smooth density on $\mathcal{X}^d$, and the Stein operator is defined in terms of the *score function* of p, $\boldsymbol{s}_p(\boldsymbol{x}) = \nabla \log p(\boldsymbol{x}) = {}^{\nabla p(\boldsymbol{x})}/_{p(\boldsymbol{x})}$, where ∇ is the gradient operator. Observe that any normalization constant in p cancels out in the score function, so that if the Stein operator $\mathcal{A}_p$ depends on p only through $\boldsymbol{s}_p$, then the Stein discrepancy $\sup_{f\in\mathscr{F}} \mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p f(\boldsymbol{x})\right]$ can still be computed when p is unnormalized. However, constructing the Stein operator using the gradient becomes restrictive when one moves beyond distributions with smooth densities.

For discrete distributions, even in the simple case of Bernoulli random variables, none of the aforementioned tests apply, since the probability mass function is no longer differentiable. This motivates more general constructions of tests based on Stein's method that would also be applicable to discrete domains. In Chapter 5, we extend the notion of Stein discrepancy to discrete distributions by defining an appropriate Stein operator based on partial differences. Then, adopting a similar strategy as Chwialkowski

et al. (2016); Liu et al. (2016), we develop a nonparametric goodness-of-fit test for unnormalized discrete distributions. Furthermore, we propose a general characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators. For any Stein operator constructed as such, we could then define a kernelized Stein discrepancy measure to establish a valid goodness-of-fit test. Finally, we apply our proposed goodness-of-fit test to the Ising model, the Bernoulli *restricted Boltzmann machine* (Hinton, 2002), and the *exponential random graph model* (Wasserman and Pattison, 1996). Our experiments show that the proposed test typically outperforms a two-sample test based on the *maximum mean discrepancy* (Gretton et al., 2012) in terms of power while maintaining control on false-positive rate.

Unlike distributions over fixed-length vectors, point processes are inherently *infinite-dimensional* distributions over *sets* containing an arbitrary number of points in some underlying space. This fundamental difference precludes the construction of Stein operators using existing techniques, and requires a new set of tools. In Chapter 6, we construct a suitable Stein operator for general point processes. While such constructions have been well-studied for Poisson process approximations in the probability literature (Barbour, 1988; Barbour and Brown, 1992), constructions for general point processes have been largely unexplored. Our key technical tool in constructing a general Stein operator is the *Papangelou conditional intensity* of a point process (Papangelou, 1974). Importantly, any (intractable) normalization constant in the density or intensity function of the point process cancels out when evaluating the Papangelou conditional intensity. Next, we define a positive-definite kernel on the space of point configurations using the maximum mean discrepancy, which captures both extrinsic and intrinsic characteristics of the point configurations. Using our proposed Stein operator and kernel function, we then define a kernelized Stein discrepancy measure between point processes. This allows us to develop a computationally feasible, nonparametric goodness-of-fit test for general point processes, including those whose density/intensity functions contain intractable normalization constants, such as Gibbs processes. We apply our proposed goodness-of-fit test to

the Poisson process as well as two processes with inter-point interactions: the Hawkes process (Hawkes, 1971) exhibiting self-excitation, and the Strauss process (Strauss, 1975) featuring repulsion. Our experiments demonstrate that the proposed test outperforms a two-sample test based on the maximum mean discrepancy in terms of power while maintaining control on false-positive rate.

1.3 Contributions

This dissertation develops statistical models and model-criticism techniques for learning from data exhibiting relational, temporal, and/or spatial dependencies. At a high-level, the contributions of this dissertation fall into two flavors. On the one hand, we propose latent space point process models to study dynamic interactions in communication networks. On the other hand, we develop statistical model criticism techniques (in particular, goodness-of-fit tests) for complex models involving intractable normalization constants, with examples including network models and point processes. More specifically, we make the following contributions:

- In Chapter 4, we study latent space point process models for dynamic networks.
  - We propose a sequence of models, including a Poisson process latent space model, two single-latent space Hawkes process models, and a dual-latent space model, to capture and decouple the influences of homophily and reciprocity in temporal interactions.
  - We develop methodology to evaluate the proposed models, including static and dynamic link prediction tasks, as well as exploration of the learned node embeddings.
  - We evaluate the utility of our models both quantitatively and qualitatively on three real-world datasets, and show that incorporating both homophily and reciprocal latent spaces improves predictive performance and gives rise to interpretable embeddings.

- In Chapter 5, we develop a nonparametric goodness-of-fit test for unnormalized discrete distributions, and propose a general characterization of Stein operators.
  - We propose a difference Stein operator for discrete spaces, which allows us to define a (kernelized) discrete Stein discrepancy measure and establish a goodness-of-fit test for unnormalized discrete distributions.
  - We propose a general characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators.
  - We apply our proposed goodness-of-fit test to the Ising model, the Bernoulli restricted Boltzmann machine, and the exponential random graph model. Our experiments show that the proposed test typically outperforms a two-sample test based on the maximum mean discrepancy in terms of power while maintaining control on false-positive rate.
- In Chapter 6, we propose a Stein–Papangelou operator for general point processes and develop a computationally feasible, nonparametric goodness-of-fit test.
  - We construct a suitable Stein operator for general point processes based on the *Papangelou conditional intensity* function.
  - We propose a positive-definite kernel on the space of point configurations using the maximum mean discrepancy, and define a kernelized Stein discrepancy measure for general point processes.
  - We develop a computationally feasible, nonparametric goodness-of-fit test for general point processes, including those whose density/intensity functions contain intractable normalization constants.
  - We apply our proposed goodness-of-fit test to the Poisson process, the Hawkes process, and the Strauss process. Our experiments show that the proposed test outperforms a two-sample test based on the maximum mean discrepancy in terms of power while maintaining control on false-positive rate.

1.4 Thesis Organization

This dissertation is organized as follows:

- In Chapter 2, we review the existing literature on models for networks and point processes, and introduce notation that shall be used throughout the dissertation. In Chapter 3, we review the fundamentals of reproducing kernel Hilbert spaces, and discuss two classes of nonparametric hypothesis tests—a kernel two-sample test based on the maximum mean discrepancy (Gretton et al., 2012), and a goodness-of-fit test based on the kernelized Stein discrepancy (Chwialkowski et al., 2016; Liu et al., 2016). The latter forms the basis of our developments in Chapters 5 and 6.

- Chapters 4, 5, and 6 constitute the main contributions of this dissertation. In Chapter 4, we propose latent space point process models to capture and decouple the influences of homophily and reciprocity dynamic networks. Chapters 5 and 6 turn to the study of model criticism techniques for complex statistical models. In Chapter 5, we define the notion of kernelized discrete Stein discrepancy, and develop a nonparametric goodness-of-fit test for discrete distributions with intractable normalization constants. In Chapter 6, we propose a Stein–Papangelou operator for general point processes, and establish a computationally feasible kernel-based goodness-of-fit test for general point processes.

- Finally, Chapter 7 concludes with a summary of the contributions in this dissertation, and outlines directions for future research.

Figure 1.4 visualizes the dependency structure of Chapters 2–6. Each of these chapters includes a short summary section reviewing the material covered in that chapter and outlining its connection to other chapters.

The material presented in Chapters 4–6 have appeared in prior publications (Yang et al., 2017, 2018, 2019). All three chapters are based on joint work with Vinayak Rao and Jennifer Neville; Chapter 5 also involves collaboration with Qiang Liu.

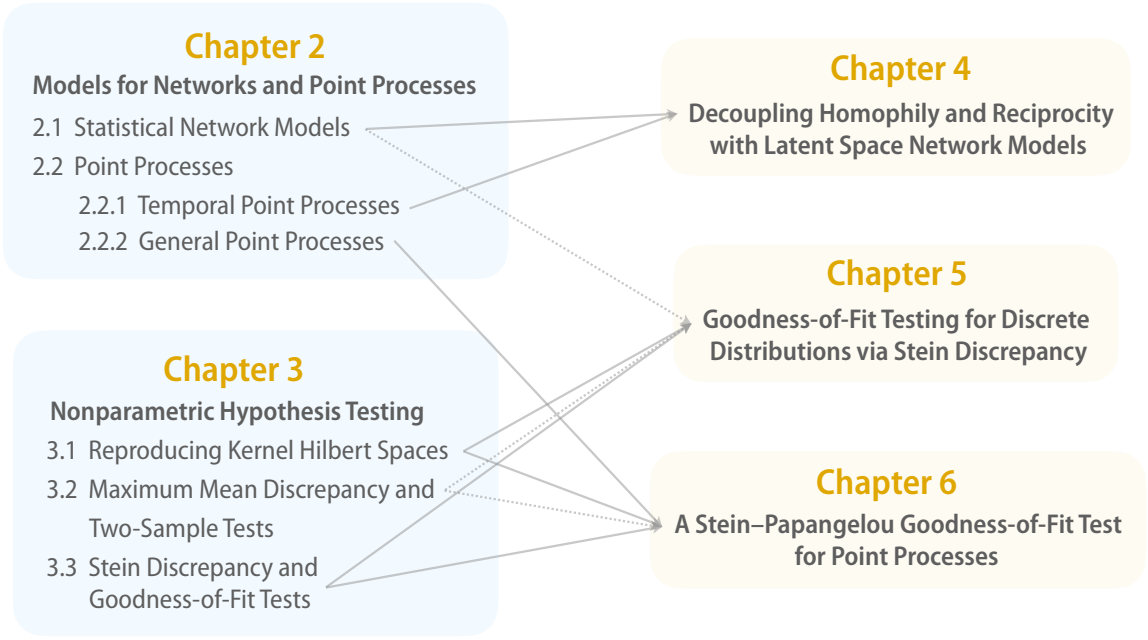


Figure 1.4.: Dependency structure of Chapters 2–6.

2. MODELS FOR NETWORKS AND POINT PROCESSES

Networks and point processes provide flexible tools for representing and modeling complex dependencies in data arising from various physical and social domains. Graphs, or networks, encode *relational* dependencies between entities, while point processes characterize *temporal* or *spatial* interactions among events. In this chapter, we review the relevant literature on network models and point processes, setting the stage for our developments in subsequent chapters. Specifically, in Chapter 4, we propose latent space point process models for decoupling homophily and reciprocity in dynamic networks; in Chapters 5 and 6, we develop kernel-based goodness-of-fit tests for intractable discrete distributions (including network models) and point processes, respectively.

2.1 Statistical Network Models

Graphs or networks, are useful representations of *relational* data natural to various social, physical, and informational domains. Examples of network data include social networks consisting of users and their friendships, citation networks comprising articles with their co-citations, protein interaction networks characterizing the physical contacts among protein molecules, communication networks recording messages sent between individuals, and the World Wide Web containing web pages and the hyperlinks between them. As such, modeling network data has been a topic of interest in fields ranging from mathematics, physics, statistics, computer science, to the social sciences.

Mathematically, a graph (or network) G is written as $G = (V, E)$, where V is the set of vertices (nodes) and $E \subseteq V \times V$ is the set of edges (links). For example, in a network of individuals on a social media platform like Facebook, V represents users and E encodes undirected friendship relations. In a corporate email network, each node $v \in V$ might represent an employee in the corporation, and each edge (u, v), an email message sent

from node u to node v. These two examples represent two different kinds of network data: static and dynamic.

2.1.1 Static Network Models

Static network models have a rich history, with a few representative ones including the *Erdős-Rényi model* (Erdős and Rényi, 1959), the *small-world model* (Watts and Strogatz, 1998), the *preferential attachment model* (Barabasi and Albert, 1999), the *exponential random graph model* (Frank and Strauss, 1986; Wasserman and Pattison, 1996), the *stochastic blockmodel* (Nowicki and Snijders, 2001), the *latent space model* (Hoff et al., 2002), and the *mixed-membership stochastic blockmodel* (Airoldi et al., 2008). A survey of these models can be found in Goldenberg et al. (2010). Two models will be of particular interest to us in this dissertation: the exponential random graph (or p^*) model of Frank and Strauss (1986); Wasserman and Pattison (1996), and the latent space model of Hoff et al. (2002).

Exponential Random Graph Model

The exponential random graph model (ERGM) was developed in parallel within the statistics and social science communities, and have since attracted much attention in both communities. Origins of the ERGM could be traced back to the p_1 model of Holland and Leinhardt (1981) which takes the form of a log-linear model with fixed effects, and the p_2 model of van Duijn et al. (2004) which replaced the fixed effects with random effects. While the most general form of the ERGM (also referred to as the p^* model) was given by Wasserman and Pattison (1996), we shall present a special case known as the *Markov random graph model* (Frank and Strauss, 1986) for ease of interpretation. The reader is referred to the surveys of Robins et al. (2007b); Goldenberg et al. (2010) for details regarding the other models.

Frank and Strauss (1986) showed that by assuming that any two edges in a graph are independent if they do not share common nodes, the probability distribution of such undirected Markov graphs could be characterized as

$$p(\boldsymbol{y}) = \frac{1}{Z(\boldsymbol{\theta}, \tau)} \exp\left\{ \sum_{k=1}^{n-1} \theta_k S_k(\boldsymbol{y}) + \tau T(\boldsymbol{y}) \right\}, \tag{2.1}$$

where $\boldsymbol{y} \in \{0,1\}^{n \times n}$ is a symmetric adjacency matrix, $S_k(\boldsymbol{y})$ counts the number of edges ($k = 1$) or k-stars ($k \geq 2$) in $\boldsymbol{y}$, $T(\boldsymbol{y})$ counts the number of triangles, and $Z(\boldsymbol{\theta}, \tau)$ is a normalization constant. More general forms of ERGMs could include counts of other (higher-order) structures in the formulation.

Notice that the normalization constant $Z(\boldsymbol{\theta}, \tau)$ in Eq. (2.1) involves summing over all possible $2^{\binom{n}{2}}$ configurations of $\boldsymbol{y}$, which is computationally intractable unless n is very small. Thus, to estimate the model parameters $\boldsymbol{\theta}$ and τ, one could not directly maximize the full likelihood function in the presence of $Z(\boldsymbol{\theta}, \tau)$. Common approaches to circumvent this issue include maximizing the *pseudo-likelihood* which calculates the probability of observing each edge *conditional* on all the other dyads in the network, or estimating the normalization constant $Z(\boldsymbol{\theta}, \tau)$ using Monte Carlo methods. In particular, the formulation of the pseudo-likelihood assumes the subgraph counts in Eq. (2.1) to be independent, although this is usually *not* the case in practice (*e.g.*, a triangle consists of three 2-stars). As a result, pseudo-likelihood maximization could lead to unreliable parameter estimates and standard errors (van Duijn et al., 2009). From a model-specification perspective, certain ERGM parameterizations have been shown to exhibit model and inferential *degeneracies*—namely, that the likelihood function places disproportionate probability mass on only a few graph configurations, often those corresponding to the empty or complete graph (Handcock et al., 2003). Such degenerate behavior has been a topic of much theoretical and empirical investigation (Rinaldo et al., 2009; Chatterjee and Diaconis, 2013; Shalizi and Rinaldo, 2013), and various remedies have been proposed in the literature (Hunter and Handcock, 2006; Snijders et al., 2006; Robins et al., 2007a). Notably, Hunter et al. (2008b) has carefully developed a set of routines for the fitting, simulation, and diagnosis of ERGM models in the form of the

`ergm` R package. Figure 2.1 shows five graph samples drawn from an EGRM on $n = 20$ nodes with parameters $\theta_1 = -2$, $\theta_k = 0\,(k \geq 2)$, and $\tau = 0.05$ using the `ergm` package.

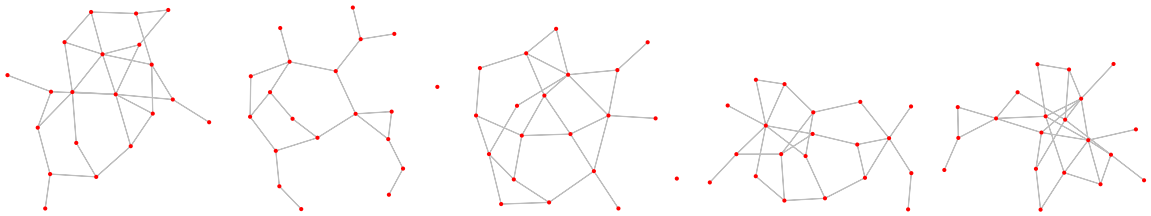


Figure 2.1.: Samples drawn from an EGRM on $n = 20$ nodes with parameters $\theta_1 = -2$, $\theta_k = 0\,(k \geq 2)$, and $\tau = 0.05$ using the `ergm` package (Hunter et al., 2008b).

The various model and inferential degeneracies exhibited by the ERGM raises the question of formal statistical tests for assessing the model fit. Hunter et al. (2008a) presented graphical diagnostics by comparing structural statistics of the observed network with those of networks simulated from the fitted model. In Chapter 5, we develop a kernel-based goodness-of-fit test for discrete distributions with intractable normalization constants, and apply it to the ERGM as one example.

Latent Space Model

Latent space approaches to social network analysis were pioneered by Hoff et al. (2002). In a latent space model, each node $v \in V$ is mapped to a latent representation $\boldsymbol{z}_v$ in some space (*e.g.*, the d-dimensional Euclidean space $\mathbb{R}^d$). The probability p_{uv} of a link between two nodes u and v is modeled as a function of their distance $d(\boldsymbol{z}_u, \boldsymbol{z}_v)$ in the latent space as well as other observed features of the nodes. The latent positions $\boldsymbol{z}_v$ and other model parameters are inferred from the observed network via maximum likelihood estimation (MLE) or Markov chain Monte Carlo (MCMC) methods.

Compared to other models commonly used in social network analysis, the latent space model offers several advantages. First, the latent space model directly reflects the notion of *homophily* that has been observed in many social domains (McPherson

et al., 2001): nodes with similar characteristics are more likely to form a tie as they will be placed closer together in the latent space. Second, by utilizing the properties of the underlying distance metric (specifically, the triangle inequality), the latent space model automatically accounts for the notion of *transitivity* often found in social networks. Third, by estimating the latent representations $\boldsymbol{z}_v$ for each node v, the model effectively *embeds* the nodes of the graph into *e.g.*, the Euclidean space $\mathbb{R}^d$. The resulting embeddings are typically much more amenable to conventional analysis and visualization, and could be directly used as *feature vectors* to perform downstream tasks such as node classification, link prediction, and clustering (community detection).

We note that the popular stochastic blockmodel (Nowicki and Snijders, 2001) and the mixed-membership stochastic blockmodel (Airoldi et al., 2008), both of which cluster network nodes into communities by assigning a latent membership vector for each node, could also be viewed as latent space models where the latent positions reside in a probability simplex rather than an Euclidean space. Hoff (2008) further proposed an *eigenmodel* that encompasses both latent class models (*e.g.*, stochastic blockmodels) and latent distance models as special cases. However, the resulting model is much less interpretable due to the reliance on eigenvectors rather than communities or distances (see *e.g.*, Section 3.9 of Goldenberg et al. (2010) for a more detailed comparison of these models).

Following the seminal work of Hoff et al. (2002), various extensions to the latent space model have appeared in the network analysis literature. Hoff (2005) incorporated random sender and receiver effects in a generalized linear model formulation to model inhomogeneity of the actors, and applied the model to valued (non-binary) networks. Handcock et al. (2007) introduced explicit clustering of the latent positions via a Gaussian mixture model, allowing one to group the network nodes into multiple communities based on their observed interactions. Krivitsky et al. (2009) further combined the two approaches into a latent cluster random effects model that accounts for four common features of social networks: homophily, transitivity, community structure and heterogeneity in node degrees. In addition, Young and Scheinerman (2007) proposed an alternative

to the original latent space model, which they termed the *random dot-product model*, by parameterizing the link probability using the inner-product of the latent node positions instead of their Euclidean distance (see Athreya et al. (2018) for a recent survey on statistical inference for random dot-product graphs).

2.1.2 Dynamic Network Models

In many real-world applications, the network structure evolves over time, and one often has access to fine-grained *temporal* information describing the evolution of the network structure. For instance, new users join social networks like Facebook every day; additionally, existing users may be connected by newly forged friendships. Similarly, in a corporate email network, servers record the precise time-stamps of every message sent between each pair of nodes.

In contrast to the extensive literature on static networks, statistical models for dynamic networks are much less explored. Existing models (Sarkar and Moore, 2006; Miller et al., 2009; Fan and Shelton, 2009; Fu and Xing, 2009; Hanneke and Xing, 2010; Snijders et al., 2010; Durante and Dunson, 2014) typically assume that the available data contain a sequence of graph snapshots captured at discrete time-points, and that the network evolution follows Markov transitions. Such approximations discard important information when one has exact time-stamps available for each link event, and require modelers to choose a particular temporal resolution to study the underlying network dynamics. A much more natural approach is to merge ideas from point process modeling with network models.

Depending on the granularity of the observed temporal information, models for dynamic networks typically fall into two categories: discrete-time or continuous-time. Discrete-time models are used when the data comprise a sequence of graph snapshots taken at regular (often equally spaced) points in time. For example, one could extract the yearly co-authorship network of a group of researchers, or record the friendship network of an undergraduate class across each school year. Continuous-time models typically

assume that precises timestamps are available for the appearance of every node/link in the network, as in the case of email communication networks (*e.g.*, Klimmt and Yang, 2004), where the time-stamps of every email message sent within a corporation is recorded over a time period.

Latent space approaches have been adopted to model dynamic networks in both discrete and continuous time. For the discrete-time setting, Sarkar and Moore (2006) generalized the static model of Hoff et al. (2002) by allowing the latent node positions to transition through a Markov process. Specifically, the node positions in the latent space are allowed to change as time progresses (thus modeling *e.g.*, friendships drifting over time), but large moves are penalized and thus improbable under the model. Sewell and Chen (2015) proposed a similar model that applies to both undirected and directed networks. Also under the discrete-time setup, Heaukulani and Ghahramani (2013) introduced a probabilistic *latent feature propagation* model, which allows the latent features of a node in the current time-point to influence that of other nodes in the next time-point based on the observed network information. Thus, the proposed model is not only able to capture the notion of homophily, but also *social influence* among individuals—*i.e.*, that our current social relationships could influence both our personal interests and our future social interactions.

Under the continuous-time setting, temporal *point processes* provide a natural tool for modeling communications between individuals over time. Perry and Wolfe (2013) used a multivariate point process with a Cox multiplicative intensity model to account for homophily, sender–receiver effects, and multicast interactions (those involving a single sender but multiple receivers) in directed networks. Blundell et al. (2012) applied Hawkes processes (Hawkes, 1971) to model *reciprocity* between groups of individuals in a dynamic network. As a simple example, in a communication network, a message sent from node u to v at one point would increase the likelihood of node v getting back to u with an reply in the near future. Hawkes processes are a class of point processes that are well-suited for modeling such excitation patterns, and we shall define them more formally in the next section. In Chapter 4, we combine the strengths of latent space

models and point processes (including Poisson and Hawkes processes) to capture and decouple the effects of homophily and reciprocity in dynamic networks.

2.2 Point Processes

Point pattern data, consisting of the counts and locations of objects in some space (*e.g.*, the d-dimensional Euclidean space $\mathbb{R}^d$), occur widely in the physical and social sciences. When $d = 1$, the underlying space typically indexes time, and the process is called a *temporal* point process. When $d > 1$, the process is often termed a *spatial* point process (one typically considers $d = 2$ or $d = 3$ in applications). A crucial distinction between $d = 1$ and $d > 1$ is the existence of a natural ordering among the elements in $\mathbb{R}$, which is absent in $\mathbb{R}^d$ $(d > 1)$. To establish intuition, we begin by reviewing examples of temporal point processes before presenting the mathematical theory of general point process. We caution that certain notions and properties of temporal point processes relies on the ordering of the real line, and thus fail to generalize to spatial point processes.

2.2.1 Temporal Point Processes

Temporal point processes typically describe events occurring in time, so that the underlying space is the non-negative real line $\mathbb{R}_+$. We write realizations of such events as $\{N(t), t \geq 0\}$, where $N(t)$ is non-negative, integer, and non-decreasing, and gives the number of events occurring in the time interval $[0, t)$. The object $\{N(t), t \geq 0\}$ formally corresponds to a counting measure (*cf.* Section 2.2.2), assigning to any interval $[s, t)$ a measure equal to the number of events in that interval. The special structure of the real line, as well as the causal nature of temporal dynamics have led to a variety of models for point process data on the real line. We describe two of them that will be fundamental to our development in Chapter 4.

Poisson process. The Poisson process is the canonical example of a point process. It is governed by a non-negative intensity (rate) function $\lambda(t)$, and has two properties:

(i) for any $t > s$, the number of events within the time interval $[s, t)$, *i.e.*, $N(t) - N(s)$, follows a Poisson distribution with mean $\int_s^t \lambda(t)\,\mathrm{d}t$; and ($ii$) the number of events in disjoint time intervals are independent random variables. A Poisson process is said to be *homogeneous* if its intensity function is constant ($\lambda(t) \equiv \lambda$), and *inhomogeneous* otherwise. If the intensity $\lambda(t)$ itself is random, then the point process is referred to as a *Cox process* (or a *doubly stochastic* Poisson process).

Hawkes process. Hawkes processes (Hawkes, 1971) have attracted much attention recently (Reinhart, 2018) by capturing deviations from the Poisson process assumptions. Hawkes processes account for causality in the temporal dynamics of point pattern data by modeling self-excitation (when a single point process is involved), and mutual excitation or *reciprocity* (when collections of point processes are under study). The former is relevant to modeling individual activities over time (*e.g.*, hospital visits), while the latter is useful for modeling activities on communication networks (*e.g.*, email communications between members of an organization). In both examples, an initial event is often a trigger for a subsequent burst of activity.

Formally, let $\mathcal{H}_t := \{N(s)\}_{s<t}$ denote the history of the point process prior to time t. Define the *conditional intensity* function

$$\lambda(t|\mathcal{H}_t) := \frac{\mathbb{E}\left[\mathrm{d}N(t)|\mathcal{H}_t\right]}{\mathrm{d}t} \tag{2.2}$$

as the instantaneous arrival rate of the point process given the history $\mathcal{H}_t$. Then, a *Hawkes process* is a point process with conditional intensity function

$$\lambda(t|\mathcal{H}_t) = \gamma + \int_0^t g(t-s)\,\mathrm{d}N(s) = \gamma + \sum_{k:\, t_k<t} g(t-t_k), \tag{2.3}$$

where $\mathcal{H}_t := \{t_k : t_k < t\}$ consists of the event history at time t, γ is the base-rate, and $g(\cdot)$ is a *triggering* kernel that characterizes the excitatory effect that a past event has on the current event rate. For example, $g(\cdot)$ could be the exponential kernel $g(t) = \beta e^{-t/\tau}$, $t \geq 0$, implying that an event has an excitatory boost of magnitude β, which decays exponentially with a time-scale τ. More generally, when we have m

processes $\{N_1(t), N_2(t), \cdots, N_m(t)\}$ that mutually excite one another, a multivariate Hawkes process has the conditional intensity of the j-th process given by

$$\lambda_j(t|\{\mathcal{H}_i(t)\}_{i=1}^m) = \gamma_j + \sum_{i=1}^m \int_0^t g_{ij}(t-s)\,\mathrm{d}N_i(s)\,.$$

Here $\mathcal{H}_i(t)$ denotes the event history associated with the i-th process $N_i(t)$ at time t; this consists of all events up to time t that are seen by process i.

Likelihood function. Given a set of observed events $\{t_i\}_{i=1}^n$ in an interval $[0, T)$, the *density* (*likelihood*) function of a (conditional) intensity $\lambda(t)$ is given by

$$\mathcal{L}\left(\lambda(t) \,|\, \{t_i\}_{i=1}^n\right) = e^{-\Lambda(0,T)} \prod_{i=1}^n \lambda(t_i)\,, \tag{2.4}$$

where $\Lambda(0, T) = \int_0^T \lambda(t)\mathrm{d}t$ is the *cumulative* (conditional) intensity function. Different point process models make different independence assumptions about $\lambda(t)$.

2.2.2 General Point Processes

We now introduce the notation and tools for studying point processes on general spaces. These notions will provide the basis for our development of goodness-of-fit tests for general point processes in Chapter 6. For further details on the rich theory of point processes, we refer the reader to the texts of Kingman (1992); Last and Penrose (2017) for the Poisson process and the comprehensive volumes of Daley and Vere-Jones (2003, 2008) for the theory of general point processes.

Notation. Let $\mathbb{X}$ be a locally compact metric space with $\mathcal{B}_{\mathbb{X}}$ its Borel σ-algebra. We will refer to $\mathbb{X}$ as the *ground space*, and consider point processes with points lying in this space. In practice, $\mathbb{X}$ is usually a compact subset of the d-dimensional Euclidean space $\mathbb{R}^d$.

A *configuration* or *realization* of a point process on $\mathbb{X}$ is a locally finite counting measure on $(\mathbb{X}, \mathcal{B}_{\mathbb{X}})$. We shall be primarily concerned with finite configurations in this

work; these form finite integer-valued measures on $(\mathbb{X}, \mathcal{B}_{\mathbb{X}})$. Let us denote the space of finite configurations on $\mathbb{X}$ by $\mathcal{N}_{\mathbb{X}}$. While a configuration is formally defined as a counting measure, we shall also identify it as a (locally) finite set of points, and describe it using set-theoretic notations. Conversely, any (locally) finite set of points $\phi \subseteq \mathbb{X}$ also defines the configuration with set A having measure

$$\phi(A) := |\phi \cap A|, \quad \forall A \in \mathcal{B}_{\mathbb{X}},$$

where $|\cdot|$ denotes the cardinality of a set. For a point $x \in \mathbb{X}$, let δ_x denote the Dirac measure centered at x. Given a point configuration $\phi \in \mathcal{N}_{\mathbb{X}}$, the configurations $\phi + \delta_x$ and $\phi - \delta_x$ correspond to the point-sets $\phi \cup \{x\}$ and $\phi \backslash \{x\}$, respectively, and we shall use the measure-theoretic and set-theoretic notations interchangeably.

Point process. Formally, a *point process* Φ on $\mathbb{X}$ is a random point configuration on $\mathbb{X}$. Define its *intensity measure* μ as

$$\mu(A) := \mathbb{E}[\Phi(A)], \quad \forall A \in \mathcal{B}_{\mathbb{X}}.$$

When $\mathbb{X} \subseteq \mathbb{R}^d$, the intensity measure is typically given in terms of a positive function $\lambda(\cdot)$ on $\mathbb{X}$, called the *rate* or *intensity function*:

$$\mu(A) = \int_A \lambda(x)\,\mathrm{d}x.$$

Next, we describe a few point processes and introduce some important theoretical tools along the way. We begin by revisiting the Poisson process in its full generality.

Poisson process. A point process Φ with intensity measure μ is called a *Poisson process* if (*i*) the counting measure Φ is *completely random*, *i.e.*, for any disjoint measurable subsets $A_1, A_2, \ldots, A_k \in \mathcal{B}_{\mathbb{X}}$, the point counts $\Phi(A_1), \Phi(A_2), \ldots, \Phi(A_k)$ are independent random variables; and (*ii*) for any set $A \in \mathcal{B}_{\mathbb{X}}$, $\Phi(A)$ follows a Poisson distribution with mean $\mu(A)$.

The following result, known as the *Mecke formula* (see *e.g.*, Last and Penrose, 2017), characterizes the Poisson process through the expectation of integrals (sums) with

respect to a Poisson process, where the integrand depends on both the point process and a location in the ground space.

Theorem 2.2.1 (Mecke formula)**.** *Let μ be an s-finite measure and Φ be a point process on $\mathbb{X}$. Then Φ is a Poisson process with intensity measure μ if and only if*

$$\mathbb{E}\left[\int_{\mathbb{X}} h(x,\Phi)\Phi(\mathrm{d}x)\right] = \int_{\mathbb{X}} \mathbb{E}\left[h(x,\Phi+\delta_x)\right]\mu(\mathrm{d}x).$$

for all measurable functions $h : \mathbb{X} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$.

More complicated point processes relax the assumption of complete randomness. One example is the Hawkes process described in Section 2.2.1. Another example is a general class of point processes, known as *Gibbs processes* (Ripley and Kelly, 1977), that model inter-point interactions in higher-dimensional spaces. These processes originated from statistical physics, and have found a wide range of applications in stochastic geometry, spatial statistics, and image analysis (van Lieshout, 2000).

Gibbs processes. The probability density of a Gibbs process (with respect to the unit-rate Poisson process on $\mathbb{X}$) takes the form:

$$f(\phi) = \frac{1}{Z}\exp\left\{-\sum_{k=1}^{|\phi|}\sum_{\omega\subseteq\phi,\,|\omega|=k} v_k(\omega)\right\},$$

where $v_k : \mathbb{X} \to \mathbb{R}$ is called the k-th order interaction potential, and Z is a normalization constant. Note that this normalization constant involves summing over all possible configurations $\phi \in \mathcal{N}_{\mathbb{X}}$, an *infinite-dimensional* integral which is intractable in all but the simplest situations (*e.g.*, the Poisson process: a Gibbs process with $v_k \equiv 0$, $\forall k > 1$).

Papangelou conditional intensity. The key challenge in generalizing the notion of conditional intensity (Eq. (2.2)) for temporal point processes to spatial point processes is the lack of a natural ordering in $\mathbb{R}^d$ when $d > 1$: the 'history' of the process is not

defined. For a point process Φ with density f, we follow Baddeley et al. (2005) and define its *Papangelou conditional intensity* (Papangelou, 1974) as

$$\rho(x|\phi) = \begin{cases} \dfrac{f(\phi \cup \{x\})}{f(\phi)} & \text{if } x \notin \phi; \\ \dfrac{f(\phi)}{f(\phi \setminus \{x\})} & \text{if } x \in \phi, \end{cases} \tag{2.5}$$

for $x \in \mathbb{X}$ and $\phi \in \mathcal{N}_{\mathbb{X}}$. We set $\rho(x|\phi) = 0$ if $f(\phi) = 0$. Informally, $\rho(x|\phi)\mathrm{d}x$ represents the relative probability of there being a point of Φ lying within an infinitesimal region of area $\mathrm{d}x$ containing x, given that the rest of the point process Φ coincides with ϕ (Baddeley et al., 2005). Thus, the Papangelou conditional intensity provides an intuitive characterization of a point process.[1]

For a Poisson process, its complete randomness ensures that its Papangelou conditional intensity is equivalent to its intensity: $\rho(x|\phi) \equiv \lambda(x)$, $\forall x \in \mathbb{X}$, $\phi \in \mathcal{N}_{\mathbb{X}}$.

For a Hawkes process with density function given by Eq. (2.4), its Papangelou conditional intensity takes the form:

$$\rho(x|\{t_i\}_{i=1}^n) = e^{-\int_0^{T-x} g(s)\,\mathrm{d}s} \cdot \Big[\gamma + \sum_{k:\, t_k < x} g(x - t_k)\Big] \cdot \prod_{i:\, t_i > x} \frac{\gamma + \sum_{k:\, t_k < t_i} g(t_i - t_k) + g(t_i - x)}{\gamma + \sum_{k:\, t_k < t_i} g(t_i - t_k)}.$$

Notice that the Papangelou conditional intensity is different from the conditional intensity function $\lambda(t|\mathcal{H}_t)$ of Eq. (2.2) which conditions only on events prior to t.

For a Gibbs process, although its density f and intensity function λ are both intractable, the normalization constant Z cancels out when evaluating Eq. (2.5), and the Papangelou conditional intensity is fully available:

$$\rho(x|\phi) = \exp\left\{ -\sum_{k=1}^{|\phi|} \sum_{\omega \subseteq \phi,\, |\omega| = k-1} v_k(\{x\} \cup \omega) \right\}. \tag{2.6}$$

An illustrative instance of Gibbs processes is the *Strauss process* (Strauss, 1975), a popular *repulsive* point process.

[1] The Papangelou conditional intensity could also be defined using the notion of *Janossy densities*; see Section 5.3 of Daley and Vere-Jones (2003) and Section 15.5 of Daley and Vere-Jones (2008) for details.

Strauss process. The Strauss process is a spatial point process on $\mathbb{X} \subseteq \mathbb{R}^d$ with conditional intensity

$$\rho(x|\phi) = \beta\gamma^{t_r(x,\phi)}, \tag{2.7}$$

where $\beta > 0$, $\gamma \in [0,1]$, and

$$t_r(x,\phi) := \sum_{y\in\phi} \mathbb{I}\{\|x-y\|_2 \le r\}$$

counts the number of points in ϕ that lie within a distance $r > 0$ of the location x. Notice that Eq. (2.7) can be recovered from Eq. (2.6) by setting

$$v_1(\{x\}) \equiv -\beta, \quad v_2(\{x,y\}) = -(\log\gamma)\cdot\mathbb{I}\{\|x-y\|_2 \le r\}, \quad v_k(\omega) \equiv 0,\ \forall k > 2.$$

While the conditional intensity of the Strauss process takes the simple form of Eq. (2.7), its density and intensity functions are generally computationally intractable for $d \ge 2$.

We conclude this section by reviewing an important identity that generalizes the Mecke formula (Theorem 2.2.1) to any finite point process. This identity, known as the *Georgii–Nguyen–Zessin (GNZ) formula* (see *e.g.*, Daley and Vere-Jones, 2008), will serve as an essential tool in our development of a Stein operator for general point processes in Chapter 6.

Theorem 2.2.2 (Georgii–Nguyen–Zessin (GNZ) formula)**.** *Let Φ be a finite point process on $\mathbb{X}$ with Papangelou conditional intensity ρ. For any measurable function $h : \mathbb{X}\times\mathcal{N}_{\mathbb{X}} \to \mathbb{R}$,*

$$\mathbb{E}\left[\int_{\mathbb{X}} h(x,\Phi\backslash\{x\})\,\Phi(\mathrm{d}x)\right] = \mathbb{E}\left[\int_{\mathbb{X}} h(x,\Phi)\,\rho(x|\Phi)\,\mathrm{d}x\right].$$

2.3 Summary

In this chapter, we have reviewed the existing literature on statistical network models (including in particular, the exponential random graph model and the latent space model), as well as the terminology, concepts, and examples of point processes (including the Poisson, Hawkes, Gibbs, and Strauss processes). In Chapter 4, we shall bring the strengths of latent space models and temporal point processes to model communications

in dynamic networks. In Chapters 5 and 6, we shall develop goodness-of-fit tests for assessing the fit of statistical network models (more generally, unnormalized discrete distributions) and point processes to observed data. The proposed tests rely on the recently introduced notion of *kernelized Stein discrepancy* (Chwialkowski et al., 2016; Liu et al., 2016), which we discuss in the next chapter.

3. NONPARAMETRIC HYPOTHESIS TESTING

Kernel methods provide a powerful toolkit in designing nonparametric hypothesis tests. These kernel-based test statistics typically involve embeddings of probability distributions into *reproducing kernel Hilbert spaces*, a concept we review in Section 3.1. In Section 3.2, we describe a well-known kernel two-sample test statistic, the *maximum mean discrepancy* of Gretton et al. (2012). In Section 3.3, we discuss some recent developments in goodness-of-fit testing using Stein's method (Stein, 1972, 1986; Gorham and Mackey, 2015; Chwialkowski et al., 2016; Liu et al., 2016), which lay the groundwork for our developments in Chapters 5 and 6.

3.1 Reproducing Kernel Hilbert Spaces

Our exposition in this section shall mainly follow Sejdinovic and Gretton (2012); we refer the reader to Rudin (1991) for the relevant background on functional analysis.

Let $\mathcal{H}$ be a Hilbert space with inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\| \cdot \|_{\mathcal{H}}$. The *Riesz representation theorem* (Rudin, 1991) states that for any continuous linear functional $\mathcal{L} : \mathcal{H} \to \mathbb{R}$, there exists a unique $g \in \mathcal{H}$, such that $\mathcal{L}f = \langle f, g \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$.

Definition 3.1.1 (Evaluation functional)**.** *Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. For a fixed $x \in \mathcal{X}$, the linear map $\delta_x : \mathcal{H} \to \mathbb{R}$, $f \mapsto \delta_x(f) = f(x)$ is called the* (Dirac) evaluation functional *at x.*

Definition 3.1.2 (Reproducing kernel Hilbert space)**.** *A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is said to be a* reproducing kernel Hilbert space (RKHS) *if δ_x is continuous (bounded) for all $x \in \mathcal{X}$, i.e., there exists $M > 0$ such that $|\delta_x(f)| = |f(x)| \leq M \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$.*

Definition 3.1.3 (Reproducing kernel)**.** *Let $\mathscr{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. A function $k : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ is called a* reproducing kernel *of $\mathscr{H}$ if*

(i) $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathscr{H}$.

(ii) $\forall x \in \mathcal{X}$, $\forall f \in \mathscr{H}$, $\langle f, k(\cdot, x)\rangle_{\mathscr{H}} = f(x)$.

The second property in Definition 3.1.3 is called the *reproducing property*. In particular, for any $x, y \in \mathcal{X}$, we can write

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathscr{H}}. \tag{3.1}$$

A reproducing kernel, if it exists, is unique. Applying the Riesz representation theorem to the evaluation functional, it can be shown (Sejdinovic and Gretton, 2012) that:

Theorem 3.1.1. *A Hilbert space $\mathscr{H}$ is a reproducing kernel Hilbert space if and only if $\mathscr{H}$ has a reproducing kernel.*

In fact, the function $k(\cdot, x)$ is a *representer of evaluation* at x:

$$\langle f, k(\cdot, x)\rangle_{\mathscr{H}} = f(x) = \delta_x f.$$

Definition 3.1.4 (Kernel)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a* kernel *on $\mathcal{X}$ if there exists a Hilbert space $\mathscr{H}$ and a map $\phi : \mathcal{X} \to \mathscr{H}$, such that for all $x, y \in \mathscr{H}$,*

$$k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathscr{H}}.$$

The map $\phi : \mathcal{X} \to \mathscr{H}$ is called the feature map*, and the space $\mathscr{H}$ the* feature space*.*

Definition 3.1.5 (Positive-definite function)**.** *A symmetric function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is* positive-definite (p.d.) *if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in \mathcal{X}$,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0.$$

The function $h(\cdot, \cdot)$ is strictly positive definite *if for mutually distinct $x_1, \ldots, x_n$, equality holds only when $a_1 = \cdots = a_n = 0$.*

From Definition 3.1.4, it is easy to see that kernel functions are positive-definite. By Eq. (3.1), every reproducing kernel is a kernel. The next theorem shows that every positive-definite function identifies a unique RKHS $\mathscr{H}$, for which k is a reproducing kernel. Therefore, all three notions—reproducing kernel, kernel, and positive-definite function—are effectively equivalent.

Theorem 3.1.2 (Moore–Aronszajn)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite function. There exists a unique RKHS $\mathscr{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ with reproducing kernel k. Specifically, $\mathscr{H}$ is given by the closure of* $\operatorname{span}\{k(\cdot, x) : x \in \mathcal{X}\}$ *endowed with the inner-product*

$$\langle f, g \rangle_{\mathscr{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j)$$

for functions $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j k(\cdot, y_j)$.

The Moore–Aronszajn theorem provides a construction of an RKHS $\mathscr{H}$ from a kernel k without imposing additional assumptions on $\mathcal{X}$ or k. If one assumes that $\mathcal{X}$ is a compact metric space and k a continuous p.d. function, an alternative construction of $\mathscr{H}$ can be obtained using the spectral theory of compact operators.

Mercer representation. Define the *integral operator* associated with kernel k as

$$(\mathcal{T}_k f)(x) = \int k(x, y) f(y) \mathrm{d}y$$

for functions $f \in L_2(\mathcal{X})$. The symmetry of k implies that $\mathcal{T}_k$ is a *self-adjoint* operator, *i.e.*, $\langle f, \mathcal{T}_k g \rangle_{\mathscr{H}} = \langle \mathcal{T}_k f, g \rangle_{\mathscr{H}}$, $\forall f, g \in L_2(\mathcal{X})$; the positive-definiteness of k implies that $\mathcal{T}_k$ is a positive operator, *i.e.*, $\langle f, \mathcal{T}_k f \rangle_{\mathscr{H}} \geq 0$, $\forall f \in L_2(\mathcal{X})$; and the continuity of k implies that $\mathcal{T}_k$ is a compact operator (Sejdinovic and Gretton, 2012). By the spectral theorem (Rudin, 1991), $\mathcal{T}_k$ can be diagonalized in an orthonormal basis comprising its (at most countable set of) eigenfunctions. This leads to the following representation:

Theorem 3.1.3 (Mercer)**.** *Let $k(\cdot, \cdot)$ be a continuous kernel on a compact metric space $\mathcal{X}$. Then for any $x, y \in \mathcal{X}$,*

$$k(x, y) = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(y),$$

where $\{e_j\}$ *and* $\{\lambda_j\}$ *are the orthonormal eigenfunctions and positive eigenvalues of the integral operator* $\mathcal{T}_k$*, and the convergence is absolute and uniform on* $\mathcal{X} \times \mathcal{X}$*.*

Thus, the RKHS $\mathscr{H}$ of k consists of linear combinations of the eigenfunctions of $\mathcal{T}_k$:

$$\mathscr{H} = \left\{ \sum_{j=1}^{\infty} a_j e_j : \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\},$$

equipped with the inner-product

$$\langle f, g \rangle_{\mathscr{H}} = \sum_{j=1}^{\infty} \frac{a_j b_j}{\lambda_j}$$

for functions $f = \sum_j a_j e_j$ and $g(\cdot) = \sum_j b_j e_j$.

We conclude this section by introducing the notion of *vector-valued* reproducing kernel Hilbert spaces which we shall employ in Section 3.3 and in Chapter 5.

Vector-valued RKHS. Assume $\mathscr{H}$ is a scalar-valued RKHS with positive-definite kernel $k(\cdot, \cdot)$. Denote by $\mathscr{H}^d = \mathscr{H} \times \cdots \times \mathscr{H}$ the Hilbert space of vector-valued functions $\boldsymbol{f} = \{f_l : f_l \in \mathscr{H}\}_{l=1}^d$, equipped with an inner-product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathscr{H}^d} = \sum_{l=1}^d \langle f_l, g_l \rangle_{\mathscr{H}}$ for $\boldsymbol{f} = \{f_\ell\}_{\ell=1}^d$ and $\boldsymbol{g} = \{g_\ell\}_{\ell=1}^d$, and induced norm $\|\boldsymbol{f}\|_{\mathscr{H}} = \sqrt{\sum_\ell \|f_\ell\|_{\mathscr{H}}^2}$. Then $\mathscr{H}^d$ is a vector-valued RKHS, with a matrix-valued positive-definite kernel $\boldsymbol{K}(x, y) = k(x, y)\boldsymbol{I}_d$. The reproducing property for this vector-valued RKHS is

$$\boldsymbol{c}^\top \boldsymbol{f}(x) = \langle \boldsymbol{f}, \boldsymbol{c} k(\cdot, x) \rangle_{\mathscr{H}^d}$$

for any $\boldsymbol{f} \in \mathscr{H}^d$ and $\boldsymbol{c} \in \mathbb{R}^d$.

3.2 Maximum Mean Discrepancy and Two-Sample Tests

In this section, we review a well-known kernel two-sample test statistic, known as the *maximum mean discrepancy*, proposed by Gretton et al. (2012).

Let X and Y be random variables on a topological space $\mathcal{X}$ with respective Borel probability measures p and q. In the *two-sample testing* problem, we are given *i.i.d.*

observations $\{x_i\}_{i=1}^m \sim p$ and $\{y_j\}_{j=1}^n \sim q$, and would like to test the hypotheses $H_0 : p = q$ vs. $H_1 : p \neq q$.

One approach of constructing a test statistic is to design an *integral probability metric* (IPM) (Müller, 1997) of the form

$$\gamma_{\mathscr{F}}(p,q) := \sup_{f \in \mathscr{F}} \left| \mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{y\sim q}[f(y)] \right|, \tag{3.2}$$

for some class $\mathscr{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$. Many well-known metrics can be recovered under the IPM framework using different choices of $\mathscr{F}$ (Sriperumbudur et al., 2012); see Table 3.1 for some examples. In particular, by taking $\mathscr{F} = \{f : \|f\|_{\mathscr{H}} \leq 1\}$ in Eq. (3.2) to be the unit-ball in an RKHS with kernel k, we obtain the maximum mean discrepancy (MMD) (Gretton et al., 2012).

Table 3.1.: Examples of integral probability metrics.

$\mathscr{F}$	Metric
$\{f : \|f\|_\infty \leq 1\}$	Total variation distance
$\{\mathbf{1}_{(-\infty,t]} : t \in \mathbb{R}\}$	Kolmogorov distance
$\{f : \|f\|_L \leq 1\}$	Kantorovich metric (L_1-Wasserstein distance)[1]
$\{f : \|f\|_\infty + \|f\|_L \leq 1\}$	Dudley metric
$\{f : \|f\|_{\mathscr{H}} \leq 1\}$	Maximum mean discrepancy

Define the *mean embedding* of p to be an element μ_p in $\mathscr{H}$ such that $\mathbb{E}_p[f] = \langle f, \mu_p \rangle_{\mathscr{H}}$ for all $f \in \mathscr{H}$. Assuming that $k(\cdot,\cdot)$ is measurable and $\mathbb{E}_p[\sqrt{k(x,x)}] < \infty$, it can be shown that $\mu_p \in \mathscr{H}$ exists, and is given by $\mu_p = \mathbb{E}_{x\sim p}[k(\cdot,x)]$. The mean embedding provides an alternative representation of the MMD:

$$\mathrm{MMD}_{\mathscr{H}}(p,q) = \sup_{\|f\|_{\mathscr{H}}\leq 1} \left| \mathbb{E}_p[f] - \mathbb{E}_q[f] \right| = \sup_{\|f\|_{\mathscr{H}}\leq 1} \left| \langle \mu_p - \mu_q, f \rangle_{\mathscr{H}} \right| = \|\mu_p - \mu_q\|_{\mathscr{H}}. \tag{3.3}$$

[1] Here, $\|f\|_L := \sup_{x\neq y\in\mathcal{X}} \frac{|f(x)-f(y)|}{d(x,y)}$ denotes the Lipschitz semi-norm of a bounded continuous real-valued function f on a metric space $(\mathcal{X}, d)$.

If $\mathscr{H}$ is *universal*,[2] then the mean embedding μ is injective, and $\mathrm{MMD}(p,q)$ is a metric on the space of Borel probability measures on $\mathcal{X}$.

Using the reproducing property of $\mathscr{H}$ (*cf.* Eq. (3.1)), one can further express the squared population MMD as

$$\mathrm{MMD}^2_{\mathscr{H}}(p,q) = \mathbb{E}_{x,x'\sim p}\left[k(x,x')\right] + \mathbb{E}_{y,y'\sim q}\left[k(y,y')\right] - 2\,\mathbb{E}_{x\sim p,\,y\sim q}\left[k(x,y)\right]. \tag{3.4}$$

Given *i.i.d.* samples $\{x_i\}_{i=1}^m \sim p$ and $\{y_j\}_{j=1}^n \sim q$, Eq. (3.4) suggests an unbiased estimate of $\mathrm{MMD}^2_{\mathscr{H}}(p,q)$ via the sum of two U-statistics and a sample average:

$$\begin{aligned}\mathrm{MMD}^2_u(X,Y) = \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}^m k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^m\sum_{j\neq i}^m k(y_i,y_j) \\ - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(x_i,y_j).\end{aligned} \tag{3.5}$$

As an unbiased estimate of $\mathrm{MMD}^2_{\mathscr{H}}(p,q)$, MMD^2_u may be negative. A non-negative, but biased, estimate can be obtained by replacing the U-statistics in Eq. (3.5) by V-statistics:

$$\mathrm{MMD}^2_b(X,Y) = \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(x_i,x_j) + \frac{1}{n^2}\sum_{i=1}^m\sum_{j=1}^m k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(x_i,y_j). \tag{3.6}$$

Both the unbiased and biased estimates could be computed in $\mathcal{O}(mn)$ time.

The asymptotic distribution of MMD^2_u under the null hypotheses $H_0 : p = q$ is given by the following theorem:

Theorem 3.2.1 (Theorem 12 of Gretton et al. (2012)). *Denote the centered-kernel between feature space mappings from which the mean embedding of p has been subtracted,*

$$\begin{aligned}\widetilde{k}(x_i,x_j) &:= \left\langle \phi(x_i) - \mu_p, \phi(x_j) - \mu_p \right\rangle \\ &= k(x_i,x_j) - \mathbb{E}_{x\sim p}\left[k(x_i,x)\right] - \mathbb{E}_{x\sim p}\left[k(x,x_j)\right] + \mathbb{E}_{x,x'\sim p}\left[k(x,x')\right].\end{aligned}$$

Assume that $\widetilde{k}(\cdot,\cdot)$ is square-integrable and $\lim_{m,n\to\infty}\frac{m}{m+n} = \rho \in (0,1)$. Then under H_0, MMD^2_u converges in distribution according to

$$(m+n)\,\mathrm{MMD}^2_u(X,Y) \xrightarrow{\mathcal{D}} \frac{1}{\rho(\rho-1)}\sum_{\ell=1}^{\infty}\lambda_\ell(\chi^2_{1\ell} - 1)$$

[2] Universality requires that $k(\cdot,\cdot)$ is continuous, and $\mathscr{H}$ is dense in $\mathcal{C}(\mathcal{X})$ *w.r.t.* the L_∞-norm, where $\mathcal{C}(\mathcal{X})$ denotes the space of bounded continuous functions on $\mathcal{X}$.

where $\{\chi^2_{1\ell}\}_{\ell=1}^{\infty}$ *is an infinite sequence of independent* χ^2 *random variables with one degree of freedom, and* $\{\lambda_\ell\}_{\ell=1}^{\infty}$ *are the eigenvalues of*

$$\int_{\mathcal{X}} \widetilde{k}(x, x')\,\psi_\ell(x)\,\mathrm{d}p(x) = \lambda_\ell\,\psi_\ell(x').$$

In practice, to determine the critical value of the two-sample test, the $(1-\alpha)$-th quantile of the null distribution can be estimated using approximations based on moment-matching, or by bootstrapping on the aggregated data following the method of Arcones and Gine (1992).

3.3 Stein Discrepancy and Goodness-of-Fit Tests

In the previous section, we studied two-sample tests for which we are given *i.i.d.* observations $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ from two *unknown* distributions p and q, respectively, and would like to test the hypotheses $H_0 : p = q$ vs. $H_1 : p \neq q$. Another important class of statistical tests are *goodness-of-fit tests*, which measure how well a model distribution $p(x)$ fits observed data $\{x_i\}_{i=1}^n$. In other words, we assume that the model distribution p is given, and we have samples $\{x_i\}_{i=1}^n$ from an unknown *data-generating* distribution q. The goal is still to test the hypotheses $H_0 : p = q$ vs. $H_1 : p \neq q$, with H_1 indicating that the model does not provide a good description of the data.

As we discussed in Section 1.2, classical goodness-of-fit tests typically assume that the model distribution $p(x)$ is fully specified. In modern applications, however, $p(x)$ is often specified only up to an intractable normalization constant. Recently, a new line of research (Gorham and Mackey, 2015; Oates et al., 2017; Chwialkowski et al., 2016; Liu et al., 2016; Jitkrittum et al., 2017) has developed goodness-of-fit tests which work directly with *unnormalized* model distributions. By combining *Stein's method* (Stein, 1972, 1986) from probability theory with techniques from reproducing kernel Hilbert spaces (*cf.* Section 3.1), one can obtain computationally tractable goodness-of-fit tests based on the notion of *kernelized Stein discrepancy* (Liu et al., 2016; Chwialkowski et al., 2016). In this section, we review these recent developments, which form the basis of our investigation in Chapters 5 and 6.

3.3.1 Stein's Method

In probability theory, Stein's method (Stein, 1972, 1986) is a sophisticated technique for proving approximations to probability distributions and characterizing convergence rates. In this section, we briefly review several elements of Stein's method that are central to the development in this dissertation. For details on this fast-growing topic, we refer the reader to the many books and surveys available in the literature (Stein, 1986; Diaconis and Holmes, 2004; Barbour and Chen, 2005; Chen et al., 2011; Ross, 2011; Barbour and Chen, 2014; Chatterjee, 2014; Ley et al., 2017).

Stein operator. Let p be a probability distribution on a measurable space $\mathcal{X}$. To apply Stein's method, one begins by identifying an operator $\mathcal{A}_p$, which acts on real- or vector-valued functions $f : \mathcal{X} \to \mathbb{R}^d$ in some function space $\mathscr{F}$, that *characterizes* the distribution p in the sense that

$$\mathbb{E}_{x\sim q}\big[\mathcal{A}_p f(x)\big] = 0,\ \forall f \in \mathscr{F} \quad \text{if and only if} \quad p = q. \tag{3.7}$$

In this case, the operator $\mathcal{A}$ is called the *Stein operator*, and we have the *Stein identity*:

$$\mathbb{E}_{x\sim p}\big[\mathcal{A}_p f(x)\big] = 0,\ \forall f \in \mathscr{F}. \tag{3.8}$$

As an important example, consider a continuously differentiable (smooth) density $p(\boldsymbol{x})$ supported on $\mathcal{X}^d \subseteq \mathbb{R}^d$. The *score function* of p is given by

$$\boldsymbol{s}_p(\boldsymbol{x}) := \nabla \log p(\boldsymbol{x}) = \frac{\nabla p(\boldsymbol{x})}{p(\boldsymbol{x})}, \tag{3.9}$$

where ∇ takes the gradient with respect to $\boldsymbol{x}$. Define the Stein operator

$$\mathcal{A}_p f(\boldsymbol{x}) := \boldsymbol{s}_p(\boldsymbol{x}) f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}). \tag{3.10}$$

For real-valued smooth functions $f : \mathcal{X} \to \mathbb{R}$ satisfying a boundary condition,[3] it is easy to verify (using integration-by-parts) that the Stein identity of Eq. (3.8) holds. For

[3] Specifically, $\oint_{\partial\mathcal{X}} p(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}S = 0$ when $\mathcal{X}$ is bounded or $\lim_{r\to\infty} \oint_{\mathcal{B}_r} p(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}S = 0$ when $\mathcal{X} = \mathbb{R}^d$, where $\mathcal{B}_r$ denotes the unit-ball of radius r centered at the origin.

instance, one could take $p(x)$ to be the standard (one-dimensional) Gaussian density, with score function $s_p(x) = -x$, and obtain

$$\mathcal{A}_p f(x) = f'(x) - x f(x),$$

which recovers the operator Stein (1972) introduced to prove normal approximations. The fact that if $Z \sim \mathcal{N}(0,1)$, then $\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$ for all absolutely continuous functions $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}[f(Z)] < \infty$, is also known as *Stein's lemma*.

Generator method. The *generator method* of Barbour (1988) provides a general approach for constructing Stein operators. Let $\{X_t : t \geq 0\}$ denote a Markov process on $\mathcal{X}$ with stationary distribution p. The infinitesimal generator

$$\mathcal{A}f(x) := \lim_{t \to 0} \frac{\mathbb{E}[f(X_t) \,|\, X_0 = x] - f(x)}{t}$$

satisfies $\mathbb{E}_{x \sim p}[\mathcal{A}f(x)] = 0$ under mild conditions and thus gives rise to a Stein operator. By applying the generator method to the *overdamped Langevin diffusion*, Gorham and Mackey (2015) obtained the *Langevin Stein operator* (Gorham et al., 2016) for vector-valued functions $\boldsymbol{f} : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^d$,

$$\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x}) := \boldsymbol{f}(\boldsymbol{x})^\mathsf{T} \boldsymbol{s}_p(\boldsymbol{x}) + \nabla \cdot \boldsymbol{f}(\boldsymbol{x}) \tag{3.11}$$

where $\boldsymbol{s}_p(\boldsymbol{x})$ is the score function of Eq. (3.9), and $\nabla \cdot$ is the divergence operator. We note that Eq. (3.11) also appeared in Oates et al. (2017). For any smooth function $\boldsymbol{f} : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ satisfying the boundary condition

$$\boldsymbol{f}(x)^\mathsf{T} \boldsymbol{n}(\boldsymbol{x}) = 0, \quad \forall x \in \partial \mathcal{X}, \tag{3.12}$$

where $\boldsymbol{n}(x)$ represents the outward unit normal vector to the boundary $\partial \mathcal{X}$, it follows from integration-by-parts that the Stein identity holds:

$$\mathbb{E}_{\boldsymbol{x} \sim p}\left[\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x})\right] = 0. \tag{3.13}$$

3.3.2 Stein Discrepancy

Assume that we have identified a Stein operator $\mathcal{A}_p$ which characterizes the probability distribution p on $\mathcal{X}$. By (3.7), when $\mathcal{A}_p$ is applied to any *test function* $f \in \mathscr{F}$, the resulting function $\mathcal{A}_p f$ has zero-expectation under p. Moreover, the expectation of $\mathcal{A}_p f$ under any other distribution $q \neq p$ should be non-zero for at least some function zf in $\mathscr{F}$. This motivates taking

$$\mathbb{D}_{\mathscr{F}}(q \,\|\, p) := \sup_{f \in \mathscr{F}} \mathbb{E}_{\boldsymbol{x} \sim q}\left[\mathcal{A}_p f(\boldsymbol{x})\right] \tag{3.14}$$

as a discrepancy measure between the distributions q and p, called the *Stein discrepancy* (Gorham and Mackey, 2015). Different choices of the function class $\mathscr{F}$ have been examined in the literature: Gorham and Mackey (2015) considered functions in the $\mathcal{W}^{2,\infty}$ Sobolev space, while Chwialkowski et al. (2016); Liu et al. (2016); Oates et al. (2017) proposed taking $\mathscr{F}$ to be the unit-ball of a reproducing kernel Hilbert space (RKHS; *cf.* Section 3.1). The advantage of the latter approach is that the resulting discrepancy can be computed in closed-form, as we shall see next.

For the remainder of this section, let $p(\boldsymbol{x})$ denote a continuously differentiable (smooth) density supported on $\mathcal{X}^d \subseteq \mathbb{R}^d$, and $\mathcal{A}_p$ the Langevin Stein operator defined in Eq. (3.11). For our purposes, it is important to note that the score function $\boldsymbol{s}_p(\boldsymbol{x})$ (*cf.* Eq. (3.9)), and thus the Stein operator $\mathcal{A}_p$, can be evaluated even if p is only known up to a normalization constant: if $p(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})/Z$, then $\boldsymbol{s}_p(\boldsymbol{x}) = {}^{\nabla p(\boldsymbol{x})}/_{p(\boldsymbol{x})} = {}^{\nabla \tilde{p}(\boldsymbol{x})}/_{\tilde{p}(\boldsymbol{x})}$ does not depend on Z.

Let $\mathscr{H}$ be an RKHS of functions $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ with kernel $k(\cdot,\cdot)$, and consider the vector-valued RKHS $\mathscr{H}^d$ of functions $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$. In Eq. (3.14), setting $\mathscr{F} = \{\boldsymbol{f} \in \mathscr{H}^d : \|\boldsymbol{f}\|_{\mathscr{H}^d} \leq 1\}$ to be the unit-ball of $\mathscr{H}^d$, we obtain the *kernelized Stein discrepancy* (KSD) between q and p (Chwialkowski et al., 2016; Liu et al., 2016):

$$\mathbb{D}(q \,\|\, p) := \sup_{\boldsymbol{f} \in \mathscr{H}, \|\boldsymbol{f}\|_{\mathscr{H}^d} \leq 1} \mathbb{E}_{\boldsymbol{x} \sim q}\left[\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x})\right] \tag{3.15}$$

Using the reproducing property, it can be shown that (Liu et al., 2016):

$$\mathbb{D}^2(q \,\|\, p) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}' \sim q}\left[\kappa_p(\boldsymbol{x},\boldsymbol{x}')\right], \tag{3.16}$$

where $\kappa_p(\cdot,\cdot)$ is a "Steinalized" kernel obtained by successively applying the Stein operator $\mathcal{A}_p$ to each argument of $k(\cdot,\cdot)$:

$$\kappa_p(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{s}_p(\boldsymbol{x})^\mathsf{T} k(\boldsymbol{x},\boldsymbol{x}')\,\boldsymbol{s}_p(\boldsymbol{x}') + \boldsymbol{s}_p(\boldsymbol{x})^\mathsf{T}\nabla_{\boldsymbol{x}'}k(\boldsymbol{x},\boldsymbol{x}') + \nabla_{\boldsymbol{x}}k(\boldsymbol{x},\boldsymbol{x}')^\mathsf{T}\boldsymbol{s}_p(\boldsymbol{x}') + \nabla_{\boldsymbol{x}}^\mathsf{T}\nabla_{\boldsymbol{x}'}k(\boldsymbol{x},\boldsymbol{x}'). \tag{3.17}$$

In addition, Liu et al. (2016) showed that if the kernel $k(\cdot,\cdot)$ is *integrally strictly positive-definite* (Stewart, 1976), that is,

$$\int g(\boldsymbol{x})\,k(\boldsymbol{x},\boldsymbol{x}')\,g(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}' > 0$$

for any function $g \in L_2(\mathcal{X})$, and if $\|q(\boldsymbol{x})(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x}))\|_2^2 < \infty$, then

$$\mathbb{D}(q\,\|\,p) = 0 \quad \text{if and only if} \quad p = q.$$

In practice, one could take $k(\cdot,\cdot)$ to be the Gaussian RBF kernel:

$$k(\boldsymbol{x},\boldsymbol{x}') = \exp\left\{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2\sigma^2}\right\} \tag{3.18}$$

which is integrally strictly positive-definite. Gorham and Mackey (2017) further recommends the *inverse multiquadric* (IMQ) kernel

$$k(\boldsymbol{x},\boldsymbol{x}') = \left(\alpha + \|\boldsymbol{x}-\boldsymbol{x}'\|_2^2\right)^\beta$$

for $\alpha > 0$ and $\beta \in (-1,0)$, which yields a KSD that provably determines weak convergence of a sequence of probability measures to its target.

Goodness-of-Fit Testing via KSD

Recall that in goodness-of-fit testing, we are given an unnormalized model distribution p, and observe *i.i.d.* samples $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathcal{X}$ from an unknown data-generating distribution q. We wish to test the hypotheses $H_0 : p = q$ vs. $H_1 : p \neq q$.

We can obtain a test statistic by estimating the squared population KSD $\mathbb{S}(q\,\|\,p) := \mathbb{D}^2(q\,\|\,p)$ via an unbiased U-statistic:

$$\widehat{\mathbb{S}}(q\,\|\,p) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \kappa_p(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

which provides a minimum-variance unbiased estimator (Hoeffding, 1948), although $\widehat{\mathbb{S}}(q \,\|\, p)$ may be negative. The asymptotic behavior of $\widehat{\mathbb{S}}(q \,\|\, p)$ is characterized in the following theorem:

Theorem 3.3.1 (Theorem 4.1 of Liu et al. (2016)). *Let $k(\boldsymbol{x}, \boldsymbol{x}')$ be a integrally strictly positive definite kernel on $\mathcal{X}^d$, and assume that $\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\left[\kappa_p(\boldsymbol{x}, \boldsymbol{x}')^2\right] < \infty$. We have the following two cases:*

(i) If $q \neq p$, then $\widehat{\mathbb{S}}(q \,\|\, p)$ is asymptotically normal:

$$\sqrt{n}\left(\widehat{\mathbb{S}}(q \,\|\, p) - \mathbb{S}(q \,\|\, p)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \mathrm{Var}_{\boldsymbol{x}\sim q}(\mathbb{E}_{\boldsymbol{x}'\sim q}\left[\kappa_p(\boldsymbol{x}, \boldsymbol{x}')\right]) > 0$.

(ii) If $q = p$, then $\sigma^2 = 0$, and the U-statistic is degenerate:

$$n\widehat{\mathbb{S}}(q \,\|\, p) \xrightarrow{\mathcal{D}} \sum_{j=1}^{\infty} c_j (Z_j^2 - 1),$$

where $\{Z_j\} \overset{iid}{\sim} \mathcal{N}(0, 1)$ and $\{c_j\}$ are the eigenvalues of the kernel $\kappa_p(\cdot, \cdot)$ under q:

$$\int \kappa_p(\boldsymbol{x}, \boldsymbol{x}')\, \phi_j(\boldsymbol{x})\, q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = c_j\, \phi_j(\boldsymbol{x}').$$

In practice, to determine the critical value of the test, one can adopt the generalized bootstrap method for degenerate U-statistics (Arcones and Gine, 1992; Huskova and Janssen, 1993) approximate the null distribution of the test statistic.[4] Specifically, to obtain a bootstrap sample, we draw random multinomial weights $w_1, \ldots, w_n \sim \mathrm{Mult}(n; 1/n, \ldots, 1/n)$, set $\widetilde{w}_i = (w_i - 1)/n$, and compute

$$\widehat{\mathbb{S}}_b^*(q \,\|\, p) = \sum_{i=1}^{n} \sum_{j\neq i}^{n} \widetilde{w}_i \widetilde{w}_j \kappa_p(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{3.19}$$

Upon repeating this procedure m times, we calculate the critical value of the test by taking the $(1-\alpha)$-th quantile of the bootstrapped statistics $\{\widehat{\mathbb{S}}_b^*\}_{b=1}^m$.

[4] An alternative *wild bootstrap* (Shao, 2010) procedure was used in Chwialkowski et al. (2016), which allows the observations $\{\boldsymbol{x}_i\}_{i=1}^n$ to be weakly dependent.

Connections to Integral Probability Metrics

We conclude this section by discussing the connections between the (kernelized) Stein discrepancy and integral probability metrics (IPMs) such as the maximum mean discrepancy (MMD) discussed in Section 3.2.

By Eq. (3.8), $\mathbb{E}_{x\sim p}\left[\mathcal{A}_p f(\boldsymbol{x})\right] = 0$ for all $f \in \mathscr{F}$. Thus, the Stein discrepancy defined in Eq. (3.14) can be rewritten as

$$\mathbb{D}_{\mathscr{F}}(q \,\|\, p) := \sup_{f\in\mathscr{F}} \mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p f(\boldsymbol{x})\right] = \sup_{f\in\mathscr{F}} \left|\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p f(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p f(\boldsymbol{x})\right]\right|,$$

which resembles the form of an IPM $\gamma_{\mathscr{F}_p}(q,p)$ (*cf.* Eq. (3.2)) with $\mathscr{F}_p := \{\mathcal{A}_p f : f \in \mathscr{F}\}$. However, this dependency of $\mathscr{F}_p$ on p renders the Stein discrepancy asymmetric, resulting in a crucial distinction from IPMs such as MMD (*cf.* Eq. (3.3)), which are by definition symmetric in their arguments p and q.

For the kernelized Stein discrepancy of Eq. (3.15), Liu et al. (2016) observed that KSD can be viewed as a special case of MMD under the "Steinalized" kernel $\kappa_p(\cdot,\cdot)$ of Eq. (3.17). Specifically, let $\mathscr{H}_p$ denote the RKHS associated with kernel κ_p. By Eq. (3.4),

$$\begin{aligned}\mathrm{MMD}^2_{\mathscr{H}_p}(p,q) &= \mathbb{E}_{x,x'\sim p}\left[\kappa_p(x,x')\right] + \mathbb{E}_{y,y'\sim q}\left[\kappa_p(y,y')\right] - 2\,\mathbb{E}_{x\sim p,\,y\sim q}\left[\kappa_p(x,y)\right] \\ &= \mathbb{E}_{y,y'\sim q}\left[\kappa_p(y,y')\right] = \mathbb{D}^2(q \,\|\, p),\end{aligned}$$

where the second equality makes use of the observation that

$$\mathbb{E}_{\boldsymbol{x}\sim p}\left[\kappa_p(\boldsymbol{x},\boldsymbol{x}')\right] = \mathbb{E}_{\boldsymbol{x}'\sim p}\left[\kappa_p(\boldsymbol{x},\boldsymbol{x}')\right] = 0,$$

and the last equality follows from Eq. (3.16). Again, note that because $\mathscr{H}_p$ depends on p, KSD is asymmetric in its arguments, whereas the MMD of Eq. (3.3) is symmetric.

3.4 Summary

In this chapter, we have reviewed the basic definitions and properties of reproducing kernel Hilbert spaces, examined a kernel two-sample test based on the maximum mean discrepancy, and discussed the recent developments in applying Stein's method to goodness-of-fit testing.

The notion of Stein discrepancy provides a promising framework for establishing goodness-of-fit tests for unnormalized distributions, yet due to their reliance on the Langevin Stein operator of Eq. (3.11), the (kernelized) Stein discrepancy and goodness-of-fit test discussed in Section 3.3.2 apply exclusively to continuous distributions with smooth density functions. In Chapter 5, we will extend these notions by constructing a Stein operator for discrete spaces; we will also provide a characterization of Stein operators that encompasses both continuous and discrete distributions. In Chapter 6, we will discover that the notions of Stein operators and (kernelized) Stein discrepancy could be further generalized to point processes (*cf.* Section 2.2), which form infinite-dimensional distributions over *sets* containing an arbitrary number of points, eventually giving rise to a nonparametric goodness-of-fit test for general point processes.

We conclude this chapter by mentioning a number of kernel-based hypothesis tests that have been omitted in our discussions. Examples include fast two-sample tests using characteristic functions (Chwialkowski et al., 2015; Jitkrittum et al., 2016), tests for (conditional) independence (Gretton et al., 2008; Zhang et al., 2011, 2018), and tests of *relative* goodness-of-fit (Bounliphone et al., 2016; Jitkrittum et al., 2018). For Stein discrepancy tests, in addition to the kernelized Stein discrepancy test of Chwialkowski et al. (2016); Liu et al. (2016) we discussed, Jitkrittum et al. (2017) further proposed a linear-time adaptive test based on a *finite-set Stein discrepancy* statistic.

4. DECOUPLING HOMOPHILY AND RECIPROCITY WITH LATENT SPACE NETWORK MODELS

Networks form useful representations of data arising in various physical and social domains. In this chapter, we consider dynamic networks such as communication networks in which links connecting pairs of nodes appear over continuous time. We adopt a point process-based approach, and study latent space models which embed the nodes into Euclidean space. We propose models to capture two different aspects of dynamic network data: (*i*) communication occurs at a higher rate between individuals with similar features (*i.e.*, *homophily*), and (*ii*) individuals tend to reciprocate communications from other nodes, but in a manner that varies across individuals. Our framework marries ideas from point process models, including Poisson and Hawkes processes, with ideas from latent space models of static networks. We evaluate our models over a range of tasks on real-world datasets and show that a *dual* latent space model, which accounts for heterogeneity in both reciprocity and homophily, significantly improves performance for both static and dynamic link prediction.

4.1 Introduction

Latent space models are a valuable tool for modeling social network data. By incorporating an embedding over nodes (*e.g.*, people), such models can account for unobserved preferences, interests, attitudes, etc. Typically, the likelihood of edges (interactions) between two nodes depends on their distance in the embedding space: the closer they are, the more likely they are to be linked. This reflects the notion of *homophily* (McPherson et al., 2001) that has been observed in many social domains: similar entities are more likely to form a tie than two randomly selected entities. As such, including latent spaces in models of static social networks has often improved

descriptive and predictive accuracy with respect to modeling the link structure (*e.g.*, Hoff et al. 2002).

In this chapter, we focus on modeling the structure of dynamic networks, where interactions occur among entities over time. Dynamic networks are more complex than static networks because the temporal interactions can be varied and bursty, reflecting new, repeated, or correlated events. Much of the recent work in modeling temporal networks has typically represented the input networks as a sequence of snapshots taken at discrete time-points and often used Markov assumptions to restrict temporal correlations to the previous time-step.

It is much more natural to model the network dynamics using point processes, particularly when the interactions indicate events that occur in continuous time (*e.g.*, each pair of nodes has a sequence of interactions over time). Previous work have applied various point processes such as Poisson processes (*e.g.*, Iwata et al. 2013), renewal processes (*e.g.*, Min et al. 2011), and Hawkes processes (*e.g.*, Blundell et al. 2012) to modeling network data. Hawkes processes in particular have attracted a great deal of recent interest due to their capability to capture *reciprocity* in interaction data. Reciprocity refers to the act of responding to a particular action with the same type of action (Ekeh, 1974). For example, in social network interactions, if one person sends another a message, the likelihood that the other person will respond and send a message back in the near future increases. However, recent work has focused more on modeling reciprocity with specific individuals (or clusters of people) rather than modeling the dependencies among individuals that may influence reciprocity.

In this chapter, we bring the strength of latent space models for static networks to point process models for dynamic networks. We make the key observation that the latent dimensions of users which influence link formation may be different from the latent dimensions of users which influence reciprocity. We refer to the former as the user's *homophily* latent space—dimensions which include preferences, interests, attitudes, etc. We refer to the latter as the user's *reciprocal* latent space—dimensions which include prosociality, agreeableness, level of self-monitoring, adherence to social norms, etc. Since

the set of temporal interactions in dynamic networks consists of various types of events (some new, some instigated by other events), it is unlikely that all such interactions are governed by the same process. We conjecture that different latent space embeddings will help models to distinguish bursty events due to reciprocation from other types of interactions in a conversation.

To explore these issues, we propose a set of latent space point process models including a Poisson process-based model and multiple Hawkes process-based models with different latent space embeddings. We evaluate the utility of the various models both quantitatively and qualitatively through a set of carefully designed experiments on real-world datasets. Our results show that a *dual* latent space Hawkes process model, which contains latent spaces for both homophily and reciprocity, are more accurate for both dynamic and static link prediction. Moreover, the embeddings themselves can be used for subjective evaluation and provide insights on how various pairs of entities interact in the network.

4.2 Problem Definition

We consider network data with the following properties:

- There exists a fixed set of vertices $V = \{1, \ldots, n\}$ throughout an observation time period $[0, T)$.
- For each ordered pair of vertices (u, v), we observe a set of event-times, corresponding to a sequence of directed links or messages from u to v. We write the overall observed data as $\{(u, v, \mathcal{H}_{uv})\}_{u,v \in V}$, where $\mathcal{H}_{uv} := \{t_i^{uv}\}_{i=1}^{n_{uv}}$ records the set of all time-points at which u sent v a message. We write $n_{uv} \geq 0$ for the total number of messages from u to v.
- A node never sends a message to itself; and the granularity of measurements is fine enough that the probability of two simultaneous events is zero.

These properties naturally motivate a point process-based approach, and we model the arrival times $\{t_i^{uv}\}$ of each link from node u to v as realizations of a point process

$N_{uv}(t)$, $t \in [0, T)$. The dynamic network evolving over time consists of $n^2 - n$ point processes $N_{uv}(t)$, which if treated as independent, involves $\mathcal{O}(n^2)$ parameters. Such an independence assumption however ignores important structure in the dynamics of the point processes. We assume two sources of dependency:

Static dependencies due to homophily, where baseline event rates vary between pairs of nodes because of shared features. Among other things, this accounts for the fact that the two processes $N_{uv}(t)$ and $N_{uw}(t)$ have a node u in common and therefore will share statistical properties. In general, homophily reflects how similarity in node-level properties (such as preferences, interests, and attitudes) affects link formation.

Dynamic interactions due to reciprocity, where activity between pairs of nodes is a function of previous history. At its simplest, this might account for reciprocity in communications between a pair of individuals. More generally, this accounts for how social influence, charisma, and the user-role affects the dynamics in a sequence of interactions. The nature of this reciprocation might depend on shared features between two nodes different from the features relevant to homophily.

Inspired by the work of Hoff et al. (2002), we model these phenomena by assigning to each node $v \in V$ a set of latent features. For the first effect, we write its feature vector as $\boldsymbol{z}_v \in \mathbb{R}^d$, and assume that the intensity function $\lambda_{uv}(t)$ underlying the process $N_{uv}(t)$ depends on the Euclidean distance between $\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2$. To account for the second effect, we cannot assume that $\lambda_{uv}(t)$ is fixed in time given these latent features, and instead must allow it to depend on previous network activity. This dependency will again be described by latent features associated with each node, but a different set which we write as $\boldsymbol{x}_v \in \mathbb{R}^d$.

While we do not explicitly assume the availability of any observed features for each node, they can be directly incorporated in our models by augmenting the $\boldsymbol{x}$- and $\boldsymbol{z}$-vectors. We also do not assume any additional information (such as message text or topic) for each link apart from its time-stamp, but we note that such information can be utilized by augmenting the hierarchical generative models with another level, as demonstrated in *e.g.*, He et al. (2015); Tan et al. (2016).

4.3 Latent Space Point Process Models of Dynamic Networks

In this section we present a series of latent space point process models for dynamic network data. We begin with the most straightforward model that only captures homophily, and proceed through a sequence of models of increasing complexity.

4.3.1 Poisson Latent Space Model

Perhaps the simplest latent space network point process model treats messages from a node u to v as a time-homogeneous Poisson process whose intensity is a function of the Euclidean distance between them in a latent feature space. In equations:

Poisson-rate latent space (PLS) model

$$\begin{aligned} \boldsymbol{z}_v &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d\times d}) && \forall v \in V \\ \lambda_{uv}(t) &= \gamma\, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} && \forall u \neq v \\ N_{uv}(\cdot) &\sim \mathsf{PoissonProcess}(\lambda_{uv}(\cdot)) && \forall u \neq v \end{aligned}$$

Here, we have placed independent Gaussian priors on the latent features for each node, resulting in a collection of correlated doubly-stochastic Poisson processes. The parameter γ can be assigned a prior if we have node-level or edge-level covariates available, but for identifiability we tie the parameter across all pairs of nodes.

4.3.2 Hawkes Latent Space Models

The remaining models augment the latent-space representation with additional non-Poissonian dynamics that capture reciprocity in communications across a network.

Hawkes Process Model

At its simplest, a node v is much more likely to send node u a message if u had just sent v a message earlier. To incorporate such reciprocity, the intensity function $\lambda_{uv}(t)$ governing the $N_{uv}(t)$ process can be modeled to depend on the events history of the reciprocal process $N_{vu}(t)$. Hawkes processes provide a simple mathematical tool to achieve this.

Specifically, for nodes $u, v \in V$, $u \neq v$, we model the pair of processes $N_{uv}(t)$ and $N_{vu}(t)$, as a bivariate Hawkes process, with intensity depending on the event histories, $\mathcal{H}_{uv} := \{t_i^{uv}\}_{i=1}^{n_{uv}}$ and $\mathcal{H}_{vu} := \{t_i^{vu}\}_{i=1}^{n_{vu}}$:

$$\lambda_{uv}(t|\mathcal{H}_{uv}, \mathcal{H}_{vu}) = \gamma_{uv} + \sum_{k:\, t_k^{vu} < t} \phi_{uv}(t - t_k^{vu}). \tag{4.1}$$

We have removed the self-excitation component since we do not consider self-loops in the network. Similar approaches have appeared in previous work (*e.g.*, Blundell et al. 2012), but we will comment on these in Section 4.5. While it is standard to parametrize the triggering function $\phi_{uv}(\cdot)$ as an exponential kernel with time scale τ, we found that learning τ suffered from identifiability issues. Instead, we model $\phi_{uv}(\cdot)$ as a weighted combination of basis kernels:

$$\phi_{uv}(t) = \sum_{b=1}^{B} \xi_b^{uv} \phi_b(t) \tag{4.2}$$

where ξ_b^{uv} is the weight of the kernel ϕ_b. We consider two possible forms for the basis kernel ϕ_b: (*i*) exponential kernels with length-scale τ, $\phi_b(t) = e^{-t/\tau}$; and (*ii*) locally periodic kernels with period p and length-scale τ, $\phi_b(t) = e^{-t/\tau} \sin^2\left(\frac{\pi t}{\tau}\right)$. In our experiments, we utilize kernels with time-scales of an hour, a day, and a week, which are interpretable and realistic for our applications.

We summarize the Hawkes process model below:

Hawkes process (HP) model

$$\lambda_{uv}(t) = \gamma + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, \phi_b(t - t_k^{vu}) \qquad \forall u \neq v$$

$$N_{uv}(\cdot) \sim \mathsf{HawkesProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$

We have again tied the parameters $\gamma_{uv} \equiv \gamma$ and $\xi_b^{uv} \equiv \xi_b$ across all node-pairs to avoid identifiability issues.

Hawkes Base-Rate Latent Space Model

The most straightforward way of modeling both homophily and reciprocity is to add the Hawkes triggering function term to the intensity functions of the previous PLS model:

Hawkes base-rate latent space (BLS) model

$$\boldsymbol{z}_v \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d \times d}) \qquad \forall v \in V$$

$$\lambda_{uv}(t) = \gamma \, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, \phi_b(t - t_k^{vu}) \qquad \forall u \neq v$$

$$N_{uv}(\cdot) \sim \mathsf{HawkesProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$

Here, and with the Poisson-rate latent space model, we shall refer to the $\boldsymbol{z}$-space as the *homophily latent space*, with the distance between $\boldsymbol{z}_u$ and $\boldsymbol{z}_v$ reflecting how dissimilar u and v are, regardless of their communication history. This distance sets a baseline rate of communication between the two nodes—*i.e.*, the rate at which one node initiates communication with the other. The Hawkes component captures the fact that having initiated a new communication, subsequent messages in that thread will follow different dynamics.

Hawkes Reciprocal Latent Space Model

The previous model assumes heterogeneity only in the rate at which different node-pairs initiate communications, and the Hawkes dynamics are themselves assumed to be homogeneous across all pairs. Our next model modifies Eq. (4.1) to reverse this assumption, associating latent features with reciprocity rather than the base-rate:

Hawkes reciprocal latent space (RLS) model

$$\boldsymbol{x}_v \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d\times d}) \qquad \forall v \in V$$

$$\lambda_{uv}(t) = \gamma + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, e^{-\|\boldsymbol{x}_u - \boldsymbol{x}_v\|_2^2} \, \phi_b(t - t_k^{vu}) \qquad \forall u \neq v$$

$$N_{uv}(\cdot) \sim \mathsf{HawkesProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$

We shall refer to the $\boldsymbol{x}$-space as the *reciprocal latent space*, since it modulates the magnitude of excitation triggered by each message between the pair of nodes.

Hawkes Dual Latent Space Model

As a final model, we combine the ideas of homophily and reciprocal latent spaces into a single model. Our final Hawkes process latent space model accounts for heterogeneity both in how two users initiate communications, as well as in the dynamics within a particular exchange. Using a mixture of exponential and periodic kernels with various length-scales, we can investigate whether a message sent from node u to v is more likely to trigger an immediate response, a response sometime over a week, or whether communications have a periodic nature.

Hawkes dual latent space (DLS) model

$$\begin{aligned}
\boldsymbol{z}_v &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d\times d}) && \forall v \in V \\
\boldsymbol{\mu}_v &\sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_{d\times d}) && \forall v \in V \\
\boldsymbol{\varepsilon}_v^{(b)} &\sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{d\times d}) && \forall v \in V,\ b = 1, \dots, B \\
\boldsymbol{x}_v^{(b)} &\sim \boldsymbol{\mu}_v + \boldsymbol{\varepsilon}_v^{(b)} && \forall v \in V,\ b = 1, \dots, B \\
\lambda_{uv}(t) &= \gamma\, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \beta\, e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2}\, \phi_b(t - t_k^{vu}) && \forall u \neq v \\
N_{uv}(\cdot) &\sim \mathsf{HawkesProcess}(\lambda_{uv}(\cdot)) && \forall u \neq v
\end{aligned}$$

Notice that we have also placed a hierarchical prior on the B reciprocal latent spaces to enforce consistency between the learned latent features across different kernels.

4.3.3 Inference

For all the models we have discussed, we perform maximum a posteriori (MAP) inference over the unknown parameters. Recall that we place independent standard Gaussian priors on the latent space vectors $\{\boldsymbol{z}_v\}_{v\in V}$ and $\{\boldsymbol{x}_v\}_{v\in V}$. Additionally, we place Gamma priors on the base rate γ and triggering magnitudes $\{\xi_b\}_{b=1}^B$ and β. Inference is tractable since it follows from Eq. (2.4) that the log-likelihood of all communications observed over the entire network $\left\{(u, v, \{t_i^{uv}\}_{i=1}^{n_{uv}})\right\}_{u,v\in V}$ can be written as

$$\log \mathcal{L} = \sum_{\substack{u,v=1 \\ u\neq v}}^{n} \left\{ -\Lambda_{uv}(0, T) + \sum_{i=1}^{n_{uv}} \log \lambda_{uv}(t_i^{uv}) \right\} \tag{4.3}$$

where the intensities $\lambda_{uv}(t)$ are specified for each model, and the cumulative intensities can be found in closed-form by noticing that for the basis kernel ϕ_b, we have

$$\int_0^T \sum_{k:\, t_k^{vu} < t} \phi_b(t - t_k^{vu}) \mathrm{d}t = \sum_{k=1}^{n_{vu}} \left[\Phi_b(T - t_k^{vu}) - \Phi_b(0) \right]$$

where $\Phi_b(t) := \int_0^t \phi_b(s)\mathrm{d}s$. In fact, the right-hand side, as well as the quantities $\left\{ \sum_{k:\, t_k^{vu} < t_i^{uv}} \phi_b(t_i^{uv} - t_k^{vu}) \right\}_{i=1}^{n_{uv}}$, are data statistics that can be pre-computed and cached for

each pair of nodes $u, v \in V$ and kernel ϕ_b. Furthermore, the gradients of the log-posterior function are also available in closed form, and the optimization can be carried out using L-BFGS-B (Byrd et al., 1995). Detailed calculations are provided in Appendix A.1.

We conclude this section with a brief summary of the complexity of each proposed model. Assuming that the number of nodes n and the dimensionality of the latent spaces d are both much larger than the number of basis-kernels B, the HP model has $\mathcal{O}(B)$ parameters, while the PLS, BLS, and RLS models have $\mathcal{O}(n \cdot d)$ parameters, and the DLS model has $\mathcal{O}(n \cdot d \cdot B)$ parameters.

4.4 Empirical Evaluation

In this section, we evaluate the proposed models of Section 4.3, both quantitatively and qualitatively, on three real-world datasets.[1] For the quantitative component, we evaluate model performance across multiple tasks, including predictive log-likelihood, dynamic link prediction, and (static) link prediction using the learned embeddings. For the qualitative component, we visualize the learned network embeddings, and demonstrate how the reciprocal latent spaces in the DLS model can be used to characterize different reciprocation patterns.

Dataset description. We perform experiments on three real-world communication networks:

ENRON This is the "core" network of the Enron email dataset (Klimmt and Yang, 2004) consisting of communications among 155 Enron executives. Each node represents an employee, while each link corresponds to an email message. We consider the period between January 2000 and April 2002, during which the vast majority of communications occurred. The resulting dataset contains 9,646 email messages spanning a period of 453 days.

[1] Code for the experiments is available at `https://github.com/jiaseny/lspp`.

EMAIL This dataset contains email communications within Purdue University from July 2011 to February 2012. Each node in the network represents an email address, while each link corresponds to an email message. We cleaned up this dataset by filtering out mailing-lists, and extracted the one hundred nodes with largest total degree. The resulting network consists of 34,438 email messages spanning a period of 237 days.

FACEBOOK This dataset contains Facebook wall messages among students of Purdue University from March 2007 to March 2008. Each node in the network represents an anonymized user account, and each link corresponds to a wall message. To focus on the "core" part of the network, we take a subset of the one hundred accounts with largest total degree. The resulting network consists of 18,865 wall messages posted over a period of 385 days.

Experiment setup. For each network dataset, we sort the messages according to their timestamps, and split the dataset into a training set consisting of the first 70% messages, and a test set consisting of the remaining 30% messages. All models are trained on the training set, and all evaluation tasks are performed on the test set.

For all Hawkes process-based models (*cf.*Section 4.3.2), we utilize $B = 4$ basis kernels—three exponential kernels with length-scales one hour, one day, and one week, respectively: $\phi_1(t) = e^{-\frac{t}{1/24}}$, $\phi_2(t) = e^{-t}$, $\phi_3(t) = e^{-t/7}$; and a locally periodic kernel with both period and length-scale set to one week: $\phi_4(t) = e^{-t/7} \sin^2\left(\frac{\pi t}{7}\right)$ (all units are in days). For all latent space models, we set the dimensionality of the latent vectors to be $d = 100$.[2] For the BLS, RLS, and DLS models, we set $\sigma^2 = \sigma_\mu^2 = \sigma_\epsilon^2 = 1$. For MAP inference, we perform optimization using the L-BFGS-B solver in the SciPy package with analytical gradients derived for each model.

[2]Since the DLS model contains one homophily latent space and four reciprocal latent spaces, one could argue that the latent spaces in the other models should be $(5\,d)$-dimensional. We experimented with setting $d = 500$ for the other models, and obtained similar results to the $d = 100$ setting.

4.4.1 Predictive Log-Likelihood

We learn the parameters of all models on the training set, and compute their predictive log-likelihood values on the test-set. From Table 4.1, we observe that the Hawkes process-based models significantly outperform the Poisson-rate latent space model (PLS), while the test log-likelihood values improve as we move from the base-rate latent space (BLS) and reciprocal latent space (RLS) models to the dual latent space (DLS) model. The DLS model also comfortably outperforms the Hawkes process (HP) model on two of the three datasets. For the Enron dataset, the simple HP model slightly outperforms the other models: we believe this is because the dataset is relatively unstructured, with pairwise reciprocity dominating most exchanges. Also notice that BLS outperforms PLS, which indicates that going beyond homogeneous Poisson processes to model reciprocity in the network indeed yields better predictive performance.

Table 4.1.: Predictive log-likelihood.

Model	ENRON	EMAIL	FACEBOOK
HP	**-16226.155**	-2129.940	-7871.895
PLS	-37803.978	-112684.130	-66742.379
BLS	-21779.686	-9850.932	-12119.869
RLS	-16565.449	-2113.254	-7867.870
DLS	-16422.946	**185.264**	**-6421.609**

4.4.2 Dynamic Link Prediction

We further gauge the performance of the learned models in a temporal link prediction task. Specifically, we randomly sample 100 time-points t_i during the test period, and ask every model to predict the probability that a link will appear between each pair of nodes in the $[t_i, t_i + \delta)$ time window (we set δ to be two weeks). Note that all models

are equipped with parameters estimated from the training set, and for Hawkes process models we also condition on all the historical training and test events up to time t_i. For each time-point t_i, we then compute the area under the ROC curve (AUC) measured across all pairs of nodes according to the predicted probabilities given by each model.[3] Finally, we report the mean and standard deviations of the AUC scores across all 100 randomly sampled testing time-points in Table 4.2.

Table 4.2.: Dynamic link prediction AUC scores.

Model	ENRON	EMAIL	FACEBOOK
HP	0.750 (0.070)	0.881 (0.088)	0.931 (0.095)
PLS	0.681 (0.041)	0.843 (0.087)	0.874 (0.078)
BLS	0.738 (0.065)	0.868 (0.095)	0.927 (0.096)
RLS	0.750 (0.070)	0.881 (0.088)	0.931 (0.095)
DLS	**0.928 (0.018)**	**0.971 (0.006)**	**0.979 (0.008)**

4.4.3 Network Embedding

As discussed in Section 2.1, a major motivation for developing latent-space network models is that the learned latent feature vectors for each node effectively provide a mapping that embeds the observed network into Euclidean space. To evaluate the quality of the learned embeddings for each latent space model, we perform link prediction on the test set by collapsing the messages into a single undirected and unweighted graph, where there exists an edge between two nodes if at least one communication exists between them in the test set. Given the learned latent feature vectors $\{\boldsymbol{z}_v\}_{v\in V}$ (or $\{\boldsymbol{x}_v^{(b)}\}_{v\in V}$ for reciprocal latent spaces), we compute the predicted probability that an

[3]The predicted probability of node u sending v at least one message during the time interval $[t, t+\delta)$ can be computed as $1-\exp\left\{-\int_t^{t+\delta}\lambda_{uv}(s)\,\mathrm{d}s\right\}$.

edge exists in the test graph via $p_{uv} \propto e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2}$, $\forall u, v \in V$, and then measure the link prediction AUC scores for all pairs of nodes.

In addition to the latent space models proposed in this work, we also compare with two popular approaches for embedding static networks:

Spectral Laplacian eigenmaps are widely used in spectral clustering (see *e.g.*, von Luxburg, 2007). Given the adjacency matrix $\mathbf{A}$ of the training network, we compute the d eigenvectors corresponding to the smallest eigenvalues of the symmetric normalized Laplacian $\mathbf{L}^{\text{sym}} := \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix of node degrees.

node2vec This is a state-of-the-art deep learning approach to learning continuous feature representations for networks (Grover and Leskovec, 2016).[4]

For both Laplacian eigenmaps and node2vec, we form an adjacency matrix of the training network $\mathbf{A}$ by collapsing the messages in the training set into an undirected graph with each edge weighted by the number of communications between the corresponding pair of nodes.[5] We set $d = 100$ for fair comparison.

Table 4.3 shows the obtained AUC scores, and Figure 4.1 plots the corresponding ROC curves. Shown alongside the ROC curves are two-dimensional projections (obtained using PCA) of the 100-dimensional latent spaces learned using each method. We observe that DLS performs on par with node2vec, and outperforms all other approaches in terms of AUC score.[6]

Homophily and reciprocal latent spaces. For the DLS model, we found that the learned homophily latent spaces $\{\mathbf{z}_v\}_{v \in V}$ always perform much better than the reciprocal

[4]We utilize the publicly available implementation at http://snap.stanford.edu/node2vec/ .

[5]For both Laplacian eigenmaps and node2vec, we have also experimented with treating the adjacency matrix $\mathbf{A}$ as binary (unweighted), but both methods exhibit degraded performance.

[5]For the DLS model, the homophily latent space is used.

[6]Notice that the current experiment setup does not yield standard errors for the AUC scores, since there is only one training/test set split. To investigate the statistical significance of the results, we conduct a further experiment which shows that while DLS significantly outperforms node2vec on ENRON, their performance are comparable on EMAIL and FACEBOOK. See Appendix A.2.1 for details.

Table 4.3.: Static link prediction AUC scores.

Model	ENRON	EMAIL	FACEBOOK
PLS	0.512	0.483	0.505
BLS	0.512	0.483	0.505
RLS	0.601	0.295	0.445
DLS	**0.906**	**0.958**	0.947
Spectral	0.687	0.428	0.452
node2vec	0.829	**0.958**	**0.956**

latent spaces $\left\{\left\{\boldsymbol{x}_v^{(b)}\right\}_{v\in V}\right\}_{b=1}^{B}$ under the static link prediction setup, as shown in the ROC curves for DLS-$\boldsymbol{z}$ and DLS-$\boldsymbol{x}^{(1)}$ in Figure 4.1.[7] Moreover, simply augmenting the homophily latent space with the reciprocal latent spaces actually leads to degraded performance in link prediction AUC. However, notice that the BLS model actually corresponds to a DLS model where we have removed the reciprocal latent spaces, and the BLS results show that in that case the learned homophily latent space performs quite poorly in link prediction as well. This indicates that the reciprocal latent spaces may have a *denoising* effect—*i.e.*, that it "explains away" communications primarily due to reciprocity such that the remaining communications arising from intensities with low reciprocal component has to be due to the fact that the pair of nodes are inherently similar in some way, which is modeled by the homophily latent features.

We further visualize the estimated homophily and reciprocal latent spaces of the DLS model by computing the pair-wise similarities $e^{-\|\boldsymbol{z}_u-\boldsymbol{z}_v\|_2^2}$ for every pair of nodes $u, v \in V$, and then plotting a heat-map of the inferred similarity matrices. For the ENRON dataset, Figure 4.2 shows the heat-maps (colors on log-scale) for both the homophily latent space and the reciprocal latent space corresponding to an hourly exponential

[7] The other reciprocal latent spaces exhibit similar performance, and we omit them from the plots to reduce clutter.

kernel (ϕ_1).[8] For each similarity matrix, we performed hierarchical clustering on the rows to obtain a node-ordering and accordingly permuted the rows and columns of the matrix simultaneously. Notice that the similarity matrices exhibit distinct clustering block-structures, indicating that the user-interaction patterns are quite different across the homophily and reciprocal latent spaces.

4.4.4 Exploring Reciprocation Patterns

While the reciprocal latent spaces in the DLS model may not be directly useful in static link prediction, they do offer a unique tool for examining the varying reciprocation patterns exhibited across different triggering kernels. Specifically, for each pair of nodes u and v, we can compute their *relative* similarities in the b-th kernel via

$$p_{uv}^{(b)} := \frac{e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2}}{\sum_{h=1}^{B} e^{-\|\boldsymbol{x}_u^{(h)} - \boldsymbol{x}_v^{(h)}\|_2^2}}, \quad b = 1, \dots, B.$$

This allows us to embed each pair of nodes onto a probability simplex where each pair $u, v \in V$ is represented by a point $\left(p_{uv}^{(1)}, \dots, p_{uv}^{(B)}\right)^\mathsf{T}$. Note that this simplicial embedding is of a different nature than the latent spaces themselves—if two points are nearby on this simplex, it indicates that the two pairs of nodes exhibit similar relative behavior across the chosen kernels, regardless of the absolute intensities of their communications.

In Figure 4.3, we selected two nodes in the ENRON network, and for each node v we plot the simplicial embeddings of each pair (v, u), $\forall u \in V$.[9] Figure 4.3 also plots node v's total outgoing intensity $\lambda_v(t) := \sum_{w \in V} \lambda_{vw}(t)$ as well as histograms showing the distribution of the Euclidean distances between v and the remaining nodes in the network. We observe that the two employees exhibit different reciprocation patterns with other employees in the corporation in terms of their active triggering kernels. For instance, the employee shown on the left appears to reciprocate with other employees in much of a similar manner since the points are more tightly concentrated, while the

[8] The complete set of heat-maps for the remaining reciprocal latent spaces as well as those for EMAIL and FACEBOOK are provided in the supplementary material.

[9] For visualization, we have collapsed the kernels ϕ_3 and ϕ_4 onto the same axis since they both have length-scale one week.

one on the right exhibits much more variability. Also notice that different reciprocating kernels may be active at different time-points, motivating the need for a mixture of kernel functions in modeling reciprocity.

4.5 Related Work

Point processes. Recent work on point process models of structured temporal data include Simma and Jordan (2010); Perry and Wolfe (2013); DuBois et al. (2013); Guo et al. (2015); He et al. (2015); Farajtabar et al. (2015); Du et al. (2016); Tan et al. (2016). In Blundell et al. (2012), Hawkes processes were combined with the infinite relational model (Kemp et al., 2006) to perform nonparametric clustering of nodes. This forms a simplification to our models, with each node having a latent cluster index rather than a latent embedding. In this model, messages are observed by all nodes in a cluster rather than individual nodes, so that reciprocity operates at the cluster level. Blundell et al. (2012) also do not model heterogeneity in the reciprocating dynamics among users.

In Linderman and Adams (2014), the authors develop a framework that combines random graph priors on the latent network structure with a reciprocating point process observation model. This is roughly equivalent to our RLS model, which we use as a proxy for comparison to Linderman and Adams (2014). However, our focus is not on learning a latent network structure as much as on teasing apart complementary parts of an observed point process. Similar to Figure 4.2, our latent embeddings can be summarized with an associated graph; in this sense the DLS model can be thought to learn two complementary graph structures underlying events on a network.

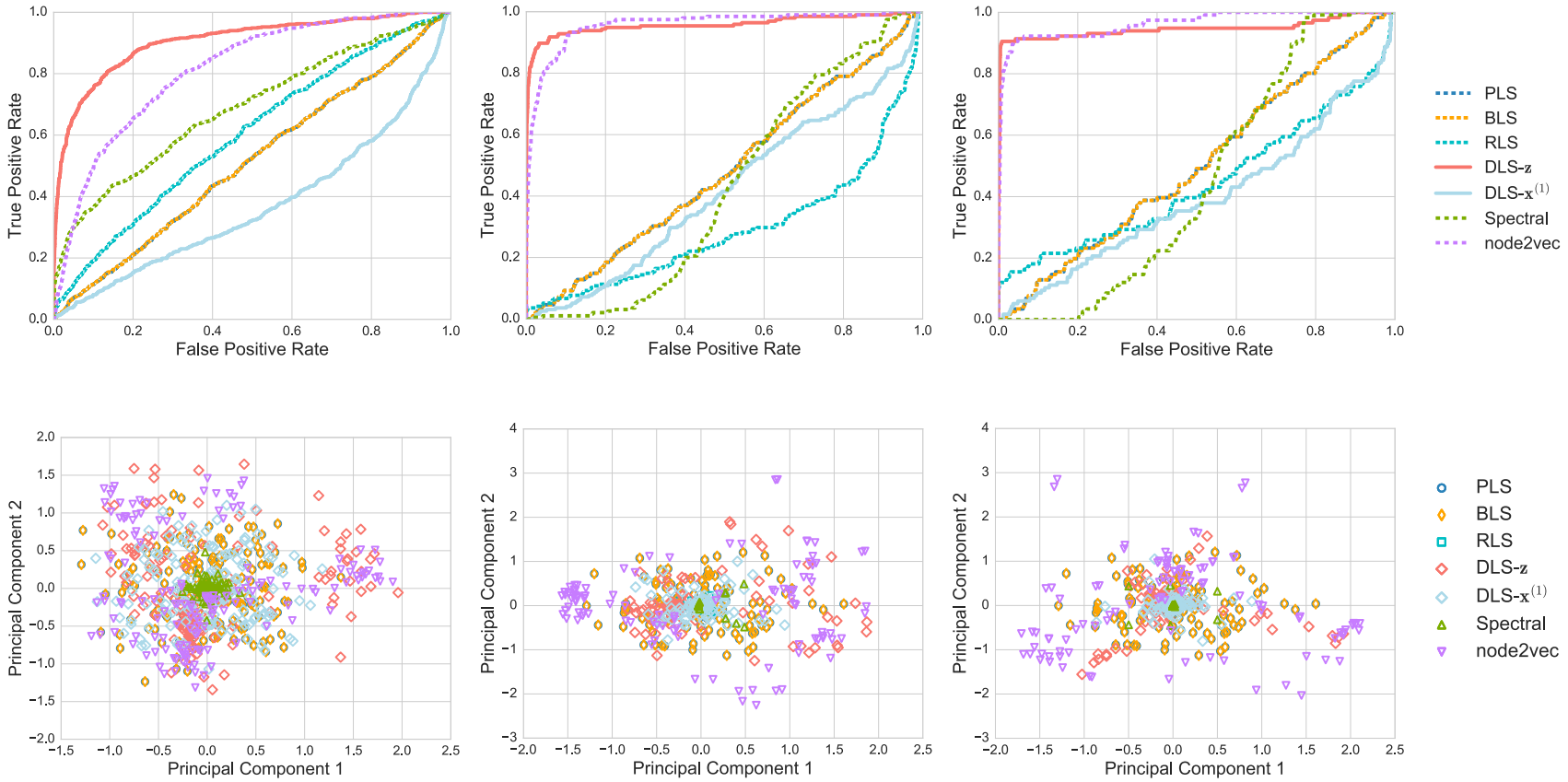
Graph embedding. Recent work in the graph mining community on learning feature representations for nodes in static networks include Perozzi et al. (2014); Tang et al. (2015); Grover and Leskovec (2016). The state-of-the-art approach is `node2vec` (Grover and Leskovec, 2016), which extends the skip-gram neural network architecture (Mikolov et al., 2013). Our experiments showed that by modeling both homophily and reciprocity

in temporal interactions, the DLS model performs comparably or superior to node2vec in static link prediction.

4.6 Summary

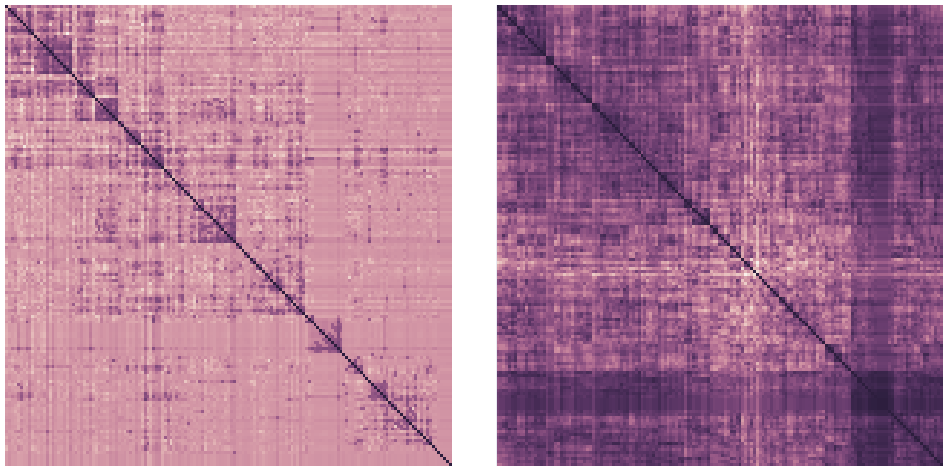
In this chapter, we have proposed latent space models for dynamic network data that embed the network nodes into Euclidean space. Our approach models heterogeneity across two important characteristics of such data—homophily and reciprocity—and connects latent space models of static networks to point process models including Poisson and Hawkes processes. The performance of our proposed dual latent space model shows that it is crucial to account for both characteristics to accurately model dynamic networks. In dynamic link prediction, we find that while the reciprocal latent space is important for accurate predictions, the inclusion of the homophily latent space produces a significant gain across all three real-world datasets. In static link prediction, while the reciprocal latent spaces are not directly useful for prediction, they greatly improve the quality of the estimated homophily latent space, by providing a denoising effect that filters out communications driven primarily by reciprocity.

Our findings shed further light on recent observations in Rudolph et al. (2016), who argue that modeling each observation *conditioned* on a set of other observations improves the quality of the learned embeddings. They refer to the conditioning set as *context* (*e.g.*, in natural language the context of a word is its surrounding words). Similarly, one might argue the context of a node in a network is its neighbors. Including reciprocal latent spaces in the model implicitly conditions on the set of reciprocating neighbors, and including homophily latent spaces implicitly conditions on the set of similar neighbors.



(a) Embedding (ENRON) (b) Embedding (EMAIL) (c) Embedding (FACEBOOK)

Figure 4.1.: Link prediction ROC curves (*top row*) and visualization of the learned embeddings (*bottom row*).



(a) $e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2}$ (b) $e^{-\|\boldsymbol{x}_u^{(1)} - \boldsymbol{x}_v^{(1)}\|_2^2}$

Figure 4.2.: Inferred node-similarity matrices in ENRON.

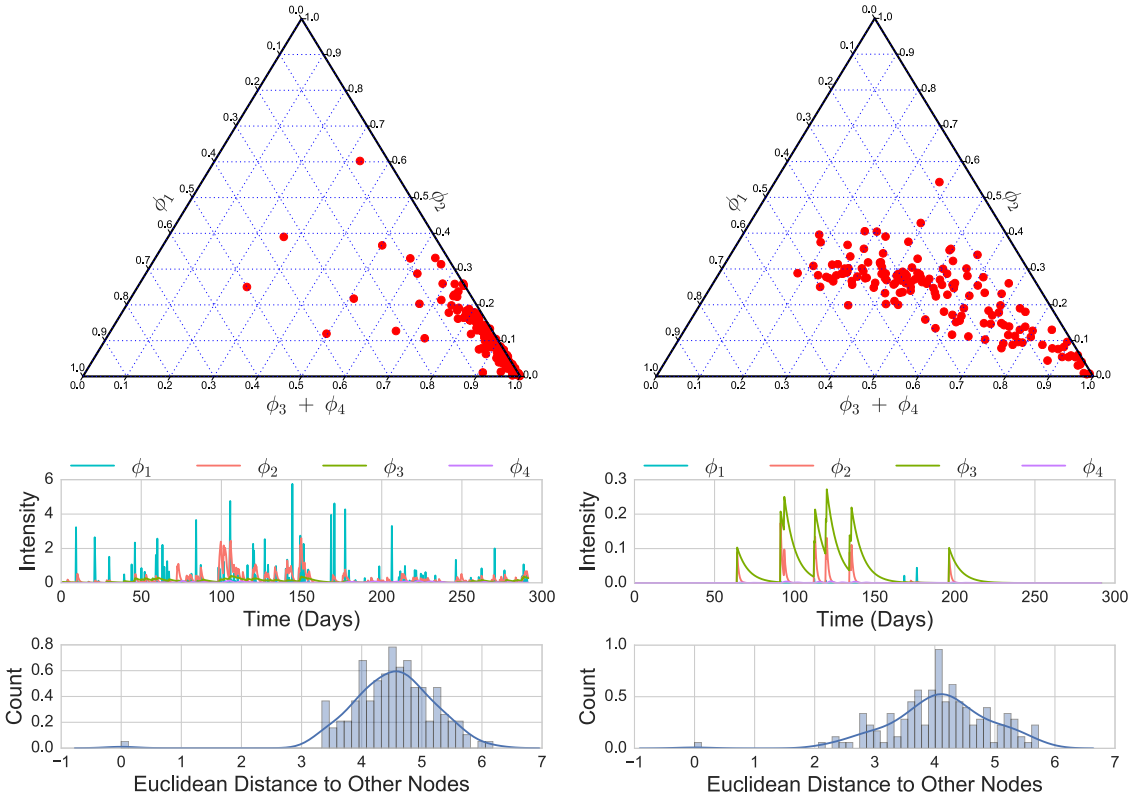


Figure 4.3.: Visualizing reciprocation patterns in ENRON.

5. GOODNESS-OF-FIT TESTING FOR DISCRETE DISTRIBUTIONS VIA STEIN DISCREPANCY

In Section 3.3, we discussed recent developments in nonparametric goodness-of-fit testing for unnormalized probability distributions based on Stein's method. However, the currently available tests apply exclusively to continuous distributions with smooth density functions. In this chapter, we introduce a kernelized Stein discrepancy measure for discrete spaces, and develop a nonparametric goodness-of-fit test for discrete distributions with intractable normalization constants. Furthermore, we propose a general characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators. We apply the proposed goodness-of-fit test to three statistical models involving discrete distributions, and our experiments show that the proposed test typically outperforms a two-sample test based on the maximum mean discrepancy.

5.1 Introduction

Let us begin by recalling our discussion in Section 3.3 on goodness-of-fit testing via Stein discrepancy. Given a distribution $p(\boldsymbol{x})$ on $\mathcal{X}^d$ and a class of test functions $f \in \mathscr{F}$ on $\mathcal{X}^d$, a Stein operator $\mathcal{A}_p$ satisfies $\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p f(\boldsymbol{x})\right] = 0$, so that when $\mathcal{A}_p$ is applied to any test function f, the resulting function $\mathcal{A}_p f$ has zero-expectation under p. Additionally, the expectation under any other distribution $q \neq p$ should be non-zero for at least some function f in $\mathscr{F}$. When $\mathscr{F}$ is sufficiently rich, the maximum value $\sup_{f\in\mathscr{F}} \mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p f(\boldsymbol{x})\right]$ serves as a discrepancy measure, called Stein discrepancy, between distributions p and q.

The properties of the Stein discrepancy measure depends on two objects: the Stein operator $\mathcal{A}_p$, and the set $\mathscr{F}$. Different authors have studied different choices of $\mathscr{F}$: Gorham and Mackey (2015) considered test functions in the $\mathcal{W}^{2,\infty}$ Sobolev space, and

the resulting test statistic requires solving a linear program under certain smoothness constraints. On the other hand, Oates et al. (2017); Chwialkowski et al. (2016); Liu et al. (2016) proposed taking $\mathscr{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS) (*cf.* Section 3.1), which leads to test statistics that can be computed in closed form and with time quadratic in n, the number of samples. Jitkrittum et al. (2017) further proposed a linear-time adaptive test that constructs test features by optimizing test power.

Regarding the choice of the Stein operator $\mathcal{A}_p$, all the aforementioned works consider the case when $\mathcal{X} \subseteq \mathbb{R}$ is a continuous domain, $p(\boldsymbol{x})$ is a smooth density on $\mathcal{X}^d$, and the Stein operator is defined in terms of the score function of p, $\boldsymbol{s}_p(\boldsymbol{x}) = \nabla \log p(\boldsymbol{x}) = {}^{\nabla p(\boldsymbol{x})}/_{p(\boldsymbol{x})}$, where ∇ is the gradient operator. Observe that any normalization constant in p cancels out in the score function, so that if the Stein operator $\mathcal{A}_p$ depends on p only through $\boldsymbol{s}_p$, then the discrepancy measure $\sup_{f \in \mathscr{F}} \mathbb{E}_{\boldsymbol{x} \sim q}\left[\mathcal{A}_p f(\boldsymbol{x})\right]$ can still be computed when p is unnormalized. However, constructing the Stein operator using the gradient becomes restrictive when one moves beyond distributions with smooth densities. For discrete distributions, even in the simple case of Bernoulli random variables, none of the aforementioned tests apply, since the probability mass function is no longer differentiable. This motivates more general constructions of tests based on Stein's method that would also be applicable to discrete domains.

In this chapter, we focus on the case where $\mathcal{X}$ is a finite set. The model distribution $p(\boldsymbol{x})$ is a probability mass function (pmf), whose normalization constant is computationally intractable. We note that examples of such intractable discrete distributions abound in statistics and machine learning, including the *Ising model* (Ising, 1924) in physics, the (Bernoulli) *restricted Boltzmann machine* (RBM) (Hinton, 2002) for dimensionality reduction, and the *exponential random graph model* (ERGM) (Frank and Strauss, 1986; Wasserman and Pattison, 1996) in statistical network analysis.

Our primary contribution is in establishing a kernelized Stein discrepancy measure between discrete distributions, using an appropriate choice of Stein operators for discrete spaces. Then, adopting a similar strategy as Chwialkowski et al. (2016); Liu et al. (2016),

we develop a nonparametric goodness-of-fit test for discrete distributions. Notably, the proposed test also applies to discrete distributions that were previously not amenable to classical tests due to the presence of intractable normalization constants. Furthermore, we propose a general characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators. For any Stein operator constructed as such, we could then define a kernelized Stein discrepancy measure to establish a valid goodness-of-fit test. Finally, we apply our proposed goodness-of-fit test to the Ising model, the Bernoulli RBM, and the ERGM, and our experiments show that the proposed test typically outperforms a two-sample test based on the maximum mean discrepancy (*cf.* Section 3.2) in terms of power while maintaining control on false-positive rate.

5.2 Discrete Stein Operators

We first propose a simple Stein operator for discrete distributions, and then provide a general characterization of Stein operators for both the discrete and continuous cases. In particular, we draw upon ideas in the literature on *score-matching* methods (Hyvärinen, 2005, 2007; Lyu, 2009; Amari, 2016), which we elaborate on further in Section 4.5.

5.2.1 Difference Stein Operator

Definition 5.2.1 (Cyclic permutation)**.** *For a set $\mathcal{X}$ of finite cardinality, a* cyclic permutation *$\neg : \mathcal{X} \to \mathcal{X}$ is a bijective function such that for some ordering $x^{[1]}, x^{[2]}, \ldots, x^{[|\mathcal{X}|]}$ of the elements in $\mathcal{X}$, $\neg x^{[i]} = x^{[(i+1) \bmod |\mathcal{X}|]}$, $\forall i = 1, 2, \ldots, |\mathcal{X}|$.*

Thus, starting with any element of x, repeated application of the $\neg$ operator generates the set $\mathcal{X}$: $\mathcal{X} = \{x, \neg x, \ldots, \neg^{(|\mathcal{X}|-1)} x\}$. In the simplest case, when $\mathcal{X}$ is a binary set, one can take $\mathcal{X} = \{\pm 1\}$ and define $\neg x = -x$.

The *inverse permutation* of $\neg$ is an operator $\neg : \mathcal{X} \to \mathcal{X}$ that satisfies $\neg(\neg x) = \neg(\neg x) = x$ for any $x \in \mathcal{X}$. Under the ordering of Definition 5.2.1, we have $\neg x^{[i]} = x^{[(i-1) \bmod |\mathcal{X}|]}$.

It is easy to verify that $\neg$ is also a cyclic permutation on $\mathcal{X}$. When $\mathcal{X}$ is a binary set, the inverse of $\neg$ is itself: $\neg = \neg$.

Definition 5.2.2 (Partial difference operator and difference score function)**.** *Given a cyclic permutation $\neg$ on $\mathcal{X}$, for any vector $\boldsymbol{x} = (x_1, \ldots, x_d)^\mathsf{T} \in \mathcal{X}^d$, write $\neg_i \boldsymbol{x} := (x_1, \ldots, x_{i-1}, \neg x_i, x_{i+1}, \ldots, x_d)^\mathsf{T}$. For any function $f : \mathcal{X}^d \to \mathbb{R}$, denote the* (partial) difference operator *as*

$$\Delta_{x_i} f(\boldsymbol{x}) := f(\boldsymbol{x}) - f(\neg_i \boldsymbol{x}), \quad i = 1, \ldots, d,$$

and write $\Delta f(\boldsymbol{x}) = (\Delta_{x_1} f(\boldsymbol{x}), \ldots, \Delta_{x_d} f(\boldsymbol{x}))^\mathsf{T}$. Define the (difference) score function *as $\boldsymbol{s}_p(\boldsymbol{x}) := {}^{\Delta p(\boldsymbol{x})}/_{p(\boldsymbol{x})}$, with*

$$(\boldsymbol{s}_p(\boldsymbol{x}))_i = \frac{\Delta_{x_i} p(\boldsymbol{x})}{p(\boldsymbol{x})} = 1 - \frac{p(\neg_i \boldsymbol{x})}{p(\boldsymbol{x})}, \quad i = 1, \ldots, d. \tag{5.1}$$

We will also be interested in the difference operator defined with respect to the inverse permutation $\neg$. To avoid cluttering notation, we shall use Δ and $\boldsymbol{s}_p$ to denote the difference operator and score function defined with respect to $\neg$, and use Δ^* to denote the difference operator with respect to $\neg$:

$$\Delta^*_{x_i} f(\boldsymbol{x}) := f(\boldsymbol{x}) - f(\neg_i \boldsymbol{x}), \quad i = 1, \ldots, d.$$

As in the continuous case, the score function $\boldsymbol{s}_p(\boldsymbol{x})$ can be easily computed even if p is only known up to a normalization constant: if $p(\boldsymbol{x}) = {}^{\tilde{p}(x)}/_{Z}$, then $\boldsymbol{s}_p(\boldsymbol{x}) = {}^{\Delta \tilde{p}(x)}/_{\tilde{p}(x)}$ does not depend on Z. For an exponential family distribution p with base measure $h(\boldsymbol{x})$, sufficient statistics $\boldsymbol{\phi}(\boldsymbol{x})$, and natural parameters $\boldsymbol{\theta}$:

$$p(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp\{\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\phi}(\boldsymbol{x})\},$$

the (difference) score function is given by

$$(\boldsymbol{s}_p(\boldsymbol{x}))_i = 1 - \frac{h(\neg_i \boldsymbol{x})}{h(\boldsymbol{x})} \exp\{\boldsymbol{\theta}^\mathsf{T} (\boldsymbol{\phi}(\neg_i \boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}))\}. \tag{5.2}$$

In the continuous case, it was obvious that two densities p and q are equal almost everywhere if and only if their score functions are equal almost everywhere. This still holds for the difference score function, but its proof is less trivial.

Theorem 5.2.1. *For any positive pmfs p and q on $\mathcal{X}^d$, we have that $\boldsymbol{s}_p(\boldsymbol{x}) = \boldsymbol{s}_q(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}^d$ if and only if $p = q$.*

The proof of Theorem 5.2.1 requires the following lemma, due to Brook (1964):

Lemma 5.2.2 (Brook, 1964)**.** *Assume that $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}^d$. The joint distribution $p(\boldsymbol{x})$ is completely determined by the collection of singleton conditional distributions $p(x_i|\boldsymbol{x}_{-i})$, where $\boldsymbol{x}_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$, $i = 1, \ldots, d$.*

Proof. Let $p(x_1, \ldots, x_d)$ and $p(y_1, \ldots, y_d)$ denote the joint densities (pmfs or pdfs) for $(x_1, \ldots, x_d)$ and $(y_1, \ldots, y_d)$, respectively. We can write

$$\begin{aligned}\frac{p(x_1, x_2, \ldots, x_d)}{p(y_1, y_2, \ldots, y_d)} &= \frac{p(x_1, x_2, \ldots, x_d)}{p(y_1, x_2, \ldots, x_d)} \cdot \frac{p(y_1, x_2, \ldots, x_d)}{p(y_1, y_2, \ldots, x_d)} \cdots \frac{p(y_1, y_2, \ldots, y_{d-1}, x_d)}{p(y_1, y_2, \ldots, y_{d-1}, y_d)} \\ &= \frac{p(x_1|x_2, \ldots, x_d)}{p(y_1|x_2, \ldots, x_d)} \cdot \frac{p(x_2|y_1, x_3, \ldots, x_d)}{p(y_2|y_1, x_3, \ldots, x_d)} \cdots \frac{p(x_d|y_1, \ldots, y_{d-1})}{p(y_d|y_1, \ldots, y_{d-1})}.\end{aligned}$$

Thus, the collection of all singleton conditional distributions completely determine the ratios of joint probability densities, which in turn completely determine the joint densities themselves, since they have to sum to one. □

Proof of Lemma 5.2.1. Clearly, $p = q$ implies that $\boldsymbol{s}_p(\boldsymbol{x}) = \boldsymbol{s}_q(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}^d$. It remains to be shown that the converse is true. By Eq. (5.1), $\boldsymbol{s}_p(\boldsymbol{x}) = \boldsymbol{s}_q(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}^d$ implies that $p(\neg_i \boldsymbol{x})/p(\boldsymbol{x}) = q(\neg_i \boldsymbol{x})/q(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}^d$ and all $i = 1, \ldots, d$. We show that the latter implies that all the singleton conditional distributions of p and q must match, *i.e.*, $p(x_i|\boldsymbol{x}_{-i}) = q(x_i|\boldsymbol{x}_{-i})$ for all $x_i \in \mathcal{X}$ and for all $i = 1, \ldots, d$, where $\boldsymbol{x}_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

Specifically, using the fact that $\neg$ is a cyclic permutation on $\mathcal{X}$, we can write

$$\begin{aligned}\frac{1}{p(x_i|\boldsymbol{x}_{-i})} &= \frac{\sum_{\xi_i \in \mathcal{X}} p(x_1, \ldots, x_{i-1}, \xi_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} \\ &= \sum_{\xi_i \in \mathcal{X}} \frac{p(x_1, \ldots, x_{i-1}, \xi_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} \\ &= \sum_{\ell=1}^{|\mathcal{X}|} \frac{p(x_1, \ldots, x_{i-1}, \neg^{(\ell)} x_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)}\end{aligned}$$

$$= \sum_{\ell=1}^{|\mathcal{X}|} \frac{p(\neg_i^{(\ell)}\boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{\ell=1}^{|\mathcal{X}|} \prod_{j=0}^{\ell-1} \frac{p(\neg_i^{(j+1)}\boldsymbol{x})}{p(\neg_i^{(j)}\boldsymbol{x})} = \sum_{\ell=1}^{|\mathcal{X}|} \prod_{j=0}^{\ell-1} \frac{p(\neg_i \boldsymbol{y}_{ij})}{p(\boldsymbol{y}_{ij})}, \tag{5.3}$$

where we adopted the convention that $\neg^{(0)}\boldsymbol{x} = \boldsymbol{x}$ and written $\boldsymbol{y}_{ij} := \neg_i^{(j)}\boldsymbol{x}$ in the last term. By Eq. (5.1), all the terms on the right-hand-side of Eq. (5.3) will be determined by the score function $\boldsymbol{s}_p(\boldsymbol{x})$, and thus $\boldsymbol{s}_p(\boldsymbol{x}) = \boldsymbol{s}_q(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}^d$ implies that all the singleton conditional distributions must match: $p(x_i|\boldsymbol{x}_{-i}) = q(x_i|\boldsymbol{x}_{-i})$, $\forall \boldsymbol{x} \in \mathcal{X}^d$. By Lemma 5.2.2, the joint probability distribution is fully specified by the collection of singleton conditional distributions, and thus we must have $p(\boldsymbol{x}) = q(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}^d$. □

In the literature on score functions (Hyvärinen, 2007; Lyu, 2009), such results, showing that a score function $\boldsymbol{s}_p(\boldsymbol{x})$ uniquely determines a probability distribution, are called *completeness* results. For our purposes, such completeness results provide a basis for establishing statistical hypothesis tests to distinguish between two distributions. We first introduce the concept of a difference Stein operator.

Definition 5.2.3 (Difference Stein operator)**.** *Let $\neg$ be a cyclic permutation on $\mathcal{X}$ and let $\multimap$ be its inverse permutation. For any function $f : \mathcal{X}^d \to \mathbb{R}$ and pmf p on $\mathcal{X}^d$, define the* difference Stein operator *of p as*

$$\mathcal{A}_p f(\boldsymbol{x}) := \boldsymbol{s}_p(\boldsymbol{x}) f(\boldsymbol{x}) - \Delta^* f(\boldsymbol{x}), \tag{5.4}$$

where $\boldsymbol{s}_p(\boldsymbol{x}) = {}^{\Delta p(\boldsymbol{x})}/_{p(\boldsymbol{x})}$ is the difference score function defined w.r.t. *$\neg$, and Δ^* is the difference operator* w.r.t. *$\multimap$.*

We note that any intractable normalization constant in p cancels out in evaluating the Stein operator $\mathcal{A}_p$. The Stein operator satisfies an important identity:

Theorem 5.2.3 (Difference Stein identity)**.** *For any function $f : \mathcal{X}^d \to \mathbb{R}$ and probability mass function p on $\mathcal{X}^d$,*

$$\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p f(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\boldsymbol{s}_p(\boldsymbol{x}) f(\boldsymbol{x}) - \Delta^* f(\boldsymbol{x})\right] = 0. \tag{5.5}$$

Proof. Notice that

$$\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p f(\boldsymbol{x})\right] = \sum_{\boldsymbol{x}\in\mathcal{X}^d} \left[f(\boldsymbol{x})\Delta p(\boldsymbol{x}) - p(\boldsymbol{x})\Delta^* f(\boldsymbol{x})\right].$$

To complete the proof, simply note that for each i,

$$\sum_{\boldsymbol{x}\in\mathcal{X}^d} f(\boldsymbol{x})\Delta_{x_i}p(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in\mathcal{X}^d} f(\boldsymbol{x})p(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\mathcal{X}^d} f(\boldsymbol{x})p(\neg_i\boldsymbol{x}),$$
$$\sum_{\boldsymbol{x}\in\mathcal{X}^d} p(\boldsymbol{x})\Delta^*_{x_i}f(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in\mathcal{X}^d} p(\boldsymbol{x})f(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\mathcal{X}^d} p(\boldsymbol{x})f(\shortminus_i\boldsymbol{x}).$$

The two equations are equal since $\neg$ and $\shortminus$ are inverse cyclic permutations on $\mathcal{X}$, with $\neg_i(\shortminus_i\boldsymbol{x}) = \shortminus_i(\neg_i\boldsymbol{x}) = \boldsymbol{x}$. □

Finally, we can extend the definition of the difference Stein operator to vector-valued functions $\boldsymbol{f}:\mathcal{X}^d\to\mathbb{R}^m$. In this case, $\Delta\boldsymbol{f}$ is an $d\times m$ matrix with $(\Delta\boldsymbol{f})_{ij} = \Delta_{x_i}f_j(\boldsymbol{x})$, and the Stein operator takes the form

$$\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x})\,\boldsymbol{f}(\boldsymbol{x})^\mathsf{T} - \Delta^*\boldsymbol{f}(\boldsymbol{x}).$$

Similar to Theorem 5.2.3, one can show that for any function $\boldsymbol{f}:\mathcal{X}^d\to\mathbb{R}^m$ and positive pmf p on $\mathcal{X}^d$,

$$\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\boldsymbol{s}_p(\boldsymbol{x})\,\boldsymbol{f}(\boldsymbol{x})^\mathsf{T} - \Delta^*\boldsymbol{f}(\boldsymbol{x})\right] = \boldsymbol{0}.$$

If $m=d$, taking the trace on both sides yields

$$\mathbb{E}_p\left[\operatorname{tr}\left(\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x})\right)\right] = \mathbb{E}_p\left[\boldsymbol{s}_p(\boldsymbol{x})^\mathsf{T}\boldsymbol{f}(\boldsymbol{x}) - \operatorname{tr}\left(\Delta^*\boldsymbol{f}(\boldsymbol{x})\right)\right] = 0.$$

The following result provides more convenient expressions for evaluating $\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x})\right]$ and $\mathbb{E}_{\boldsymbol{x}\sim p}\left[\operatorname{tr}\left(\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x})\right)\right]$. An analogous result for continuous distributions with smooth densities was provided in Ley and Swan (2013).

Lemma 5.2.4. *For positive pmfs p, q and any function $\boldsymbol{f}:\mathcal{X}^d\to\mathbb{R}^d$, we have*

$$\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim q}\left[(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x}))\,\boldsymbol{f}(\boldsymbol{x})^\mathsf{T}\right],$$
$$\mathbb{E}_{\boldsymbol{x}\sim q}\left[\operatorname{tr}\left(\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x})\right)\right] = \mathbb{E}_{\boldsymbol{x}\sim q}\left[(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x}))^\mathsf{T}\boldsymbol{f}(\boldsymbol{x})\right].$$

Proof. Theorem 5.2.3 states that $\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_q\boldsymbol{f}(\boldsymbol{x})\right] = 0$. Thus, writing $\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p\boldsymbol{f}(\boldsymbol{x}) - \mathcal{A}_q\boldsymbol{f}(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim q}[(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x}))\,\boldsymbol{f}(\boldsymbol{x})^\mathsf{T}]$ and taking the trace on both sides completes the proof. □

5.2.2 Characterization of Stein Operators

Generalizing our construction in the previous section, we can further identify a broad class of Stein operators which includes the difference Stein operator as a special case.

Let $\mathcal{L}$ be any operator defined on the space of functions $\mathscr{F} = \{f : \mathcal{X}^d \to \mathbb{R}\}$ that can be written in the form[1]

$$\mathcal{L}f(\boldsymbol{x}) = \sum_{\boldsymbol{x}' \in \mathcal{X}^d} g(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}'), \quad \forall f \in \mathscr{F} \tag{5.6}$$

for some bivariate (possibly vector-valued) function g on $\mathcal{X}^d \times \mathcal{X}^d$. Define a dual operator $\mathcal{L}^*$ via

$$\mathcal{L}^* f(\boldsymbol{x}) = \sum_{\boldsymbol{x}' \in \mathcal{X}^d} g(\boldsymbol{x}', \boldsymbol{x}) f(\boldsymbol{x}'), \quad \forall f \in \mathscr{F}. \tag{5.7}$$

In fact, when $\mathcal{X}$ is a finite set, any linear operator $\mathcal{L}$ on $\mathscr{F} = \{f : \mathcal{X}^d \to \mathbb{R}\}$ can be written in the form of Eq. (5.6). In this case, the operator $\mathcal{L}^*$ as defined in Eq. (5.7) is the *adjoint* operator of $\mathcal{L}$: $\langle \mathcal{L}f, g \rangle = \langle f, \mathcal{L}^* g \rangle$ for all $f, g \in \mathscr{F}$, where $\langle \cdot, \cdot \rangle$ is the appropriate inner-product on $\mathcal{X}^d$. If $g(\cdot, \cdot)$ is symmetric, then $\mathcal{L}$ is *self-adjoint*, *i.e.*, $\mathcal{L}^* = \mathcal{L}$.

Under these definitions, we have the following result which characterizes the Stein operators on a discrete space $\mathcal{X}^d$.

Theorem 5.2.5. *Denote $\mathscr{F} = \{f : \mathcal{X}^d \to \mathbb{R}\}$. For any positive pmf p on $\mathcal{X}^d$, a linear operator $\mathcal{T}_p$ satisfies the Stein identity*

$$\mathbb{E}_{\boldsymbol{x} \sim p}\left[\mathcal{T}_p f(\boldsymbol{x})\right] = 0 \tag{5.8}$$

for all functions $f \in \mathscr{F}$ if and only if there exist linear operators $\mathcal{L}$ and $\mathcal{L}^*$ of the forms (5.6) and (5.7), such that

$$\mathcal{T}_p f(\boldsymbol{x}) = \frac{\mathcal{L}p(\boldsymbol{x})}{p(\boldsymbol{x})} f(\boldsymbol{x}) - \mathcal{L}^* f(\boldsymbol{x}) \tag{5.9}$$

holds for all $\boldsymbol{x} \in \mathcal{X}^d$ and functions $f \in \mathscr{F}$.

[1] The notion can also be extended to vector-valued functions $\boldsymbol{f}$; we omit this generalization here for clarity.

Proof. Sufficiency: Suppose the linear operators $\mathcal{L}$ and $\mathcal{L}^*$ take the forms of Eqs. (5.6) and (5.7) for some function g, we show that the operator $\mathcal{T}_p$ defined via Eq. (5.9) satisfies the Stein identity of Eq. (5.8). We can write

$$\begin{aligned}\mathbb{E}_p\left[\mathcal{T}_p f(\boldsymbol{x})\right] &= \sum_{\boldsymbol{x}\in\mathcal{X}^d}\left[f(\boldsymbol{x})\,\mathcal{L}p(\boldsymbol{x}) - p(\boldsymbol{x})\,\mathcal{L}^* f(\boldsymbol{x})\right] \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^d}\sum_{\boldsymbol{x}'\in\mathcal{X}^d} f(\boldsymbol{x})g(\boldsymbol{x},\boldsymbol{x}')p(\boldsymbol{x}') - \sum_{\boldsymbol{x}\in\mathcal{X}^d}\sum_{\boldsymbol{x}'\in\mathcal{X}^d} p(\boldsymbol{x})g(\boldsymbol{x}',\boldsymbol{x})f(\boldsymbol{x}').\end{aligned}$$

The two terms in the last line cancel out since the double-summations are invariant under a swapping of summation indices $\boldsymbol{x}$ and $\boldsymbol{x}'$, giving $\mathbb{E}_p\left[\mathcal{T}_p f(\boldsymbol{x})\right] = 0$.

Necessity: Assume that a linear operator $\mathcal{T}$ satisfies Eq. (5.8); we show that it can be written in the form of Eq. (5.9) for some linear operators $\mathcal{L}$ and $\mathcal{L}^*$ of the forms (5.6) and (5.7). Recall that for a finite set $\mathcal{X}$, any function $f : \mathcal{X}^d \to \mathbb{R}$ can be represented by a vector $\boldsymbol{f} \in \mathbb{R}^{|\mathcal{X}|^d}$, and any linear operator $\mathcal{T}$ on the set of functions f can be represented via a matrix $\mathbf{T} \in \mathbb{R}^{|\mathcal{X}|^d \times |\mathcal{X}|^d}$ under the standard basis of $\mathbb{R}^{|\mathcal{X}|^d}$. Under these notations, $\mathcal{T}f$ can be represented by $\mathbf{T}\boldsymbol{f}$, and Eq. (5.8) can be rewritten in matrix form as

$$\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{T}_p f(\boldsymbol{x})\right] = \sum_{\boldsymbol{x}\in\mathcal{X}^d} p(\boldsymbol{x})\mathcal{T}_p f(\boldsymbol{x}) = \boldsymbol{p}^\mathsf{T}(\mathbf{T}_p\boldsymbol{f}) = 0\,,$$

which holds for any function f (*i.e.*, for any vector $\boldsymbol{f}$) if and only if $\boldsymbol{p}^\mathsf{T}\mathbf{T}_p = \mathbf{0}$. We can always find a diagonal matrix $\mathbf{D}$ and a matrix $\mathbf{L}$ such that $\mathbf{T}_p = \mathbf{D} - \mathbf{L}$. Observe that $\boldsymbol{p}^\mathsf{T}\mathbf{T}_p = \mathbf{0}$, *i.e.*, $\boldsymbol{p}^\mathsf{T}\mathbf{D} = \boldsymbol{p}^\mathsf{T}\mathbf{L}$ if and only if $d_{ii} = \boldsymbol{p}^\mathsf{T}\mathbf{L}_{*i}/p_i$ for all i, where d_{ii} is the i-th diagonal element of $\mathbf{D}$ and $\mathbf{L}_{*i}$ is the i-th column of $\mathbf{L}$. Thus, Eq. (5.8) holds if and only if

$$\mathbf{T}_p = \operatorname{diag}\{\boldsymbol{p}\}^{-1}\operatorname{diag}\left\{\boldsymbol{L}^\mathsf{T}\boldsymbol{p}\right\} - \mathbf{L}$$

for some matrix $\mathbf{L}$, where $\operatorname{diag}\{\boldsymbol{p}\}$ denotes the diagonal matrix whose i-th diagonal entry equals p_i. Rewriting, we have

$$\operatorname{diag}\{\boldsymbol{p}\}\,\mathbf{T}_p = \operatorname{diag}\left\{\mathbf{L}^\mathsf{T}\boldsymbol{p}\right\} - \operatorname{diag}\{\boldsymbol{p}\}\,\mathbf{L}\,.$$

Right-multiplying both sides by an arbitrary vector $\boldsymbol{f} \in \mathbb{R}^{|\mathcal{X}|^d}$, we obtain

$$\boldsymbol{p} \odot (\mathbf{T}_p\boldsymbol{f}) = (\mathbf{L}^\mathsf{T}\mathbf{p}) \odot \boldsymbol{f} - \boldsymbol{p} \odot (\mathbf{L}^\mathsf{T}\boldsymbol{f})\,, \tag{5.10}$$

where $\odot$ denotes the Hadamard product. Let $\mathcal{L}$ and $\mathcal{L}^*$ be the linear operators with matrices $\mathbf{L}^\mathsf{T}$ and $\mathbf{L}$ under the standard basis, Eq. (5.10) can be re-written as

$$p(\boldsymbol{x})\,\mathcal{T}_p f(\boldsymbol{x}) = \mathcal{L}p(\boldsymbol{x})\,f(\boldsymbol{x}) - p(\boldsymbol{x})\,\mathcal{L}^* f(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathcal{X}^d$. Finally, dividing by $p(\boldsymbol{x})$ on both sides yields Eq. (5.9). □

We note that the sufficiency part of Theorem 5.2.5 remains valid when $\mathcal{X}$ is a continuous space, p is a density, $\mathscr{F} \subseteq \{f : \mathcal{X}^d \to \mathbb{R}\}$ is some family of functions for which $\mathcal{T}_p f$ and $\mathcal{L}f$ are well-defined, and the summations in Eqs. (5.6) and (5.7) are replaced by integrations. However, the necessity part requires further conditions on the expressiveness of $\mathscr{F}$.

Theorem 5.2.5 essentially states that (for a fixed p) given any pair of adjoint operators $\mathcal{L}$ and $\mathcal{L}^*$, one can construct a linear operator $\mathcal{T}_p$ satisfying Stein's identity; conversely, any Stein operator $\mathcal{T}_p$ can be expressed using a pair of adjoint operators $\mathcal{L}$ and $\mathcal{L}^*$. This connection between adjoint operators and Stein operators enables us to unify different forms of Stein operators for discrete and continuous distributions (see also Ley et al. (2017) for related discussions).

Remark 5.2.6 (Continuous case)**.** *For a continuous space $\mathcal{X} \subseteq \mathbb{R}$, consider a smooth density p on $\mathcal{X}^d$. Take $\mathcal{L} = \nabla$ to be the gradient operator, and let $\mathscr{F}$ consist of smooth functions $f : \mathcal{X}^d \to \mathbb{R}$ for which $f(\boldsymbol{x})\,p(\boldsymbol{x})$ vanishes on the boundary $\partial\mathcal{X}$. Using integration-by-parts, one can show that the adjoint operator of $\mathcal{L}$ is $\mathcal{L}^* = -\nabla$. Then, applying Eq. (5.9) of Theorem 5.2.5 recovers the continuous Stein operator of Eq. (3.10):*

$$\mathcal{A}_p f(\boldsymbol{x}) = \nabla \log p(\boldsymbol{x})\,f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}).$$

More generally, for vector-valued functions $\boldsymbol{f} : \mathcal{X}^d \to \mathbb{R}^d$, the adjoint operator of $\mathcal{L} = \nabla$ is $\mathcal{L}^* = -(\nabla\cdot)$, the negative divergence. Let $\mathscr{F}$ consist of smooth functions $\boldsymbol{f} : \mathcal{X}^d \to \mathbb{R}^d$ satisfying the boundary condition of Eq. (3.12). By modifying Eq. (5.9) accordingly to

$$\mathcal{T}_p \boldsymbol{f}(\boldsymbol{x}) = \frac{\mathcal{L}p(\boldsymbol{x})}{p(\boldsymbol{x})}\boldsymbol{f}(\boldsymbol{x})^\mathsf{T} - \mathcal{L}^*\boldsymbol{f}(\boldsymbol{x}) \tag{5.11}$$

and taking the trace on both sides, one recovers the Langevin Stein operator of Eq. (3.11):

$$\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x}) = \operatorname{tr}\big(\mathcal{T}_p \boldsymbol{f}(\boldsymbol{x})\big) = \boldsymbol{f}(\boldsymbol{x})^\mathsf{T}(\nabla \log p(\boldsymbol{x})) + \nabla\cdot\boldsymbol{f}(\boldsymbol{x}).$$

Remark 5.2.7 (Discrete case)**.** *In Eqs.* (5.6) *and* (5.7)*, define the vector-valued function* $\mathbf{g} : \mathcal{X}^d \times \mathcal{X}^d \to \mathbb{R}^d$ *with*

$$(\mathbf{g}(\boldsymbol{x}, \boldsymbol{x}'))_i = \mathbb{I}\{\boldsymbol{x}' = \boldsymbol{x}\} - \mathbb{I}\{\boldsymbol{x}' = \neg_i \boldsymbol{x}\} \tag{5.12}$$

where $\mathbb{I}\{\cdot\}$ *is the indicator function. Then, we have*

$$(\mathcal{L}f(\boldsymbol{x}))_i = \sum_{\boldsymbol{x} \in \mathcal{X}^d} (\mathbf{g}(\boldsymbol{x}, \boldsymbol{x}'))_i f(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\neg_i \boldsymbol{x}),$$

which recovers the difference operator Δ*. Similarly, define* $\mathbf{g}^*$ *by replacing* $\neg$ *with its inverse permutation* $\neg$ *in Eq.* (5.12)*. Notice that* $\mathbf{g}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{g}^*(\boldsymbol{x}', \boldsymbol{x})$*, and thus the adjoint of* $\mathcal{L}$ *is given by* $\mathcal{L}^* = \Delta^*$*. In this case, Eq.* (5.9) *boils down to the difference Stein operator defined in Eq.* (5.4)*.*

Note that if $\mathcal{X}$ *is binary, then* $\neg = \neg$*, and* $\mathcal{L}$ *is self-adjoint. When* $\mathcal{L}$ *is self-adjoint, in addition to Stein's identity, the Stein operator defined via Eq.* (5.9) *also satisfies* $\mathcal{T}_p\, p(\boldsymbol{x}) = 0$*.*

Graph-based discrete Stein operators. Extending the form of Eq. (5.12), we can obtain a more general recipe for constructing g, which, upon applying Theorem 5.2.5, gives rise to other Stein operators on $\mathcal{X}^d$. Specifically, suppose we have identified a simple graph $\mathcal{G} = (\mathcal{X}^d, \mathcal{E})$ on $|\mathcal{X}|^d$ vertices, with each vertex corresponding to a possible configuration $\boldsymbol{x} \in \mathcal{X}^d$. Then, it is natural to define g such that it respects the structure of $\mathcal{G}$, in the sense that $g(\boldsymbol{x}, \boldsymbol{x}') = 0$ if $\boldsymbol{x}' \notin \mathcal{N}_{\boldsymbol{x}} \cup \{\boldsymbol{x}\}$, where $\mathcal{N}_{\boldsymbol{x}} := \{\boldsymbol{x}' : (\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{E}\}$ is the set of neighbors of $\boldsymbol{x}$ in $\mathcal{G}$. If $\mathcal{G}$ is undirected, one would also make g symmetric, in which case $\mathcal{L} \equiv \mathcal{L}^*$ is self-adjoint.

Revisiting the difference Stein operator in this light, notice that $\neg$ defines a d-dimensional (undirected) lattice graph $\mathcal{G}$ on $\mathcal{X}^d$, in which two vertices $\boldsymbol{x}$ and $\boldsymbol{x}'$ are connected if and only if $\boldsymbol{x}' = \neg_i \boldsymbol{x}$ for some $i \in \{1, \ldots, d\}$. In this case, every vertex $\boldsymbol{x}$ has exactly d neighbors in $\mathcal{G}$: $\mathcal{N}_{\boldsymbol{x}} = \{\neg_1 \boldsymbol{x}, \ldots, \neg_d \boldsymbol{x}\}$. We then set $\mathbf{g}(\boldsymbol{x}, \neg_i \boldsymbol{x}) = -\boldsymbol{e}_i$ for each i, $\mathbf{g}(\boldsymbol{x}, \boldsymbol{x}) = \boldsymbol{e}$, and $\mathbf{g}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{0}$ for $\boldsymbol{x}' \notin \mathcal{N}_{\boldsymbol{x}} \cup \{\boldsymbol{x}\}$, where $\boldsymbol{e}_i \in \mathbb{R}^d$ is the i-th standard basis vector, and $\boldsymbol{e} \in \mathbb{R}^d$ is the all-ones vector. This recovers the form of $\mathbf{g}$ in Eq. (5.12).

As another example, one could take $g(\boldsymbol{x}, \boldsymbol{x}') = -|\mathcal{N}_{\boldsymbol{x}}|^{-1}$ for $\boldsymbol{x}' \in \mathcal{N}_{\boldsymbol{x}}$ and set $g(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{I}[\boldsymbol{x} = \boldsymbol{x}']$ otherwise. Then, Eq. (5.6) becomes

$$\mathcal{L}f(\boldsymbol{x}) = \frac{1}{|\mathcal{N}_{\boldsymbol{x}}|} \sum_{\boldsymbol{x}' \in \mathcal{N}_{\boldsymbol{x}}} \big[f(\boldsymbol{x}) - f(\boldsymbol{x}') \big],$$

which recovers the normalized Laplacian of $\mathcal{G}$ (see also Amari, 2016). Thus, by specifying an arbitrary graph structure $\mathcal{G}$ on $\mathcal{X}^d$, one could also utilize its Laplacian $\mathcal{L}$ to define a corresponding Stein operator $\mathcal{T}$ by applying Theorem 5.2.5.

5.3 Kernelized Discrete Stein Discrepancy

We can now proceed similarly as in the continuous case (Liu et al., 2016; Chwialkowski et al., 2016) to define the discrete Stein discrepancy and its kernelized counterpart. While all results in this section hold for the general Stein operators discussed in Section 5.2.2, for clarity we state them for the difference Stein operator described in Section 5.2.1.

Definition 5.3.1 (Discrete Stein discrepancy)**.** *Let $\mathcal{X}$ be a finite set. For a family $\mathcal{F}$ of functions $\boldsymbol{f} : \mathcal{X}^d \to \mathbb{R}^d$, define the* discrete Stein discrepancy *between two positive pmfs p, q as*

$$\mathbb{D}(q \,\|\, p) := \sup_{\boldsymbol{f} \in \mathcal{F}} \mathbb{E}_{\boldsymbol{x} \sim q} \big[\operatorname{tr}\big(\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x}) \big) \big],$$

where $\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x})^\mathsf{T} - \Delta^* \boldsymbol{f}(\boldsymbol{x})$ is the difference Stein operator w.r.t. p. Taking $\mathcal{F}$ to be the unit ball in an RKHS $\mathcal{H}^d$ of vector-valued functions $\boldsymbol{f} : \mathcal{X}^d \to \mathbb{R}^d$, we obtain the kernelized discrete Stein discrepancy *(KDSD):*

$$\mathbb{D}(q \,\|\, p) = \sup_{\boldsymbol{f} \in \mathcal{H}^d, \|\boldsymbol{f}\|_{\mathcal{H}^d} \leq 1} \mathbb{E}_{\boldsymbol{x} \sim q} \big[\operatorname{tr}\big(\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x}) \big) \big]. \tag{5.13}$$

Although Eq. (5.13) involves solving a variational problem, the next two results show that the kernelized discrete Stein discrepancy can actually be computed in closed-form. Due to space constraints, we defer their proofs to the Appendix.

Theorem 5.3.1. *The kernelized discrete Stein discrepancy as defined in Eq. (5.13) admits an equivalent representation:*

$$\mathbb{D}(q \,\|\, p)^2 = \mathbb{E}_{\boldsymbol{x}, \boldsymbol{x}' \sim q} \big[\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^\mathsf{T} k(\boldsymbol{x}, \boldsymbol{x}') \, \boldsymbol{\delta}_{p,q}(\boldsymbol{x}') \big], \tag{5.14}$$

where $\boldsymbol{\delta}_{p,q}(\boldsymbol{x}) := \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$ *is the score-difference between* p *and* q.

Proof. Observe that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim q}\left[\operatorname{tr}\left(\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x})\right)\right] &= \sum_{\ell=1}^{d} \mathbb{E}_{\boldsymbol{x}\sim q}\left[s_p^\ell(\boldsymbol{x}) f_\ell(\boldsymbol{x}) - \Delta^*_{x_\ell} f_\ell(\boldsymbol{x})\right] \\ &= \sum_{\ell=1}^{d} \mathbb{E}_{\boldsymbol{x}\sim q}\left[s_p^\ell(\boldsymbol{x}) \left\langle f_\ell, k(\cdot,\boldsymbol{x})\right\rangle_{\mathscr{H}} - \left\langle f_\ell, \Delta^*_{x_\ell} k(\cdot,\boldsymbol{x})\right\rangle_{\mathscr{H}}\right] \\ &= \sum_{\ell=1}^{d} \left\langle f_\ell, \mathbb{E}_{\boldsymbol{x}\sim q}\left[s_p^\ell(\boldsymbol{x}) k(\cdot,\boldsymbol{x}) - \Delta^*_{x_\ell} k(\cdot,\boldsymbol{x})\right]\right\rangle_{\mathscr{H}},\end{aligned}$$

where we used the reproducing property $\langle f_\ell, k(\cdot,\boldsymbol{x})\rangle_{\mathscr{H}} = f_\ell(\boldsymbol{x})$ and the fact that

$$\begin{aligned}\Delta^*_{x_j} f_i(\boldsymbol{x}) &= f_i(\boldsymbol{x}) - f_i(\shortminus_j \boldsymbol{x}) = \langle f_i, k(\cdot,\boldsymbol{x})\rangle - \left\langle f_i, k(\cdot,\shortminus_j \boldsymbol{x})\right\rangle = \left\langle f_i, k(\cdot,\boldsymbol{x}) - k(\cdot,\shortminus_j \boldsymbol{x})\right\rangle \\ &= \left\langle f_j, \Delta^*_{x_j} k(\cdot,\boldsymbol{x})\right\rangle.\end{aligned}$$

Denoting $\boldsymbol{\beta}(\cdot) := \mathbb{E}_{\boldsymbol{x}\sim q}\left[\boldsymbol{s}_p(\boldsymbol{x}) k(\cdot,\boldsymbol{x}) - \Delta^* k(\cdot,\boldsymbol{x})\right] \in \mathscr{H}^d$, we have

$$\mathbb{E}_{\boldsymbol{x}\sim q}\left[\operatorname{tr}\left(\mathcal{A}_p \boldsymbol{f}(\boldsymbol{x})\right)\right] = \sum_{\ell=1}^{d} \langle f_\ell, \beta_\ell\rangle_{\mathscr{H}} = \langle \boldsymbol{f}, \boldsymbol{\beta}\rangle_{\mathscr{H}^d}.$$

Thus, we can rewrite the kernelized discrete Stein discrepancy as

$$\mathbb{D}(q \,\|\, p) = \sup_{\boldsymbol{f}\in\mathscr{H}^d, \|\boldsymbol{f}\|_{\mathscr{H}^d}\le 1} \langle \boldsymbol{f}, \boldsymbol{\beta}\rangle_{\mathscr{H}^d},$$

which immediately implies that $\mathbb{D}(q \,\|\, p) = \|\boldsymbol{\beta}\|_{\mathscr{H}^d}$ since the supremum will be attained by $\boldsymbol{f} = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_{\mathscr{H}^d}$.

By Lemma 5.2.4, we have

$$\boldsymbol{\beta}(\cdot) = \mathbb{E}_{\boldsymbol{x}\sim q}\left[\boldsymbol{s}_p(\boldsymbol{x}) k(\cdot,\boldsymbol{x}) - \Delta^* k(\cdot,\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim q}\left[(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})) k(\cdot,\boldsymbol{x})\right].$$

Writing $\boldsymbol{\delta}_{p,q}(\boldsymbol{x}) := \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$, we have

$$\begin{aligned}\mathbb{D}(q \,\|\, p)^2 = \|\boldsymbol{\beta}\|^2_{\mathscr{H}^d} &= \sum_{\ell=1}^{d} \langle \beta_\ell, \beta_\ell\rangle_{\mathscr{H}} \\ &= \sum_{\ell=1}^{d} \left\langle \mathbb{E}_{\boldsymbol{x}\sim q}\left[\delta^\ell_{p,q}(\boldsymbol{x}) k(\cdot,\boldsymbol{x})\right], \mathbb{E}_{\boldsymbol{x}'\sim q}\left[\delta^\ell_{p,q}(\boldsymbol{x}') k(\cdot,\boldsymbol{x}')\right]\right\rangle_{\mathscr{H}}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\ell=1}^{d} \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\Big[\delta^{\ell}_{p,q}(\boldsymbol{x})\left\langle k(\cdot,\boldsymbol{x}), k(\cdot,\boldsymbol{x}')\right\rangle_{\mathscr{H}} \delta^{\ell}_{p,q}(\boldsymbol{x}')\Big] \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\Big[\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^{\mathsf{T}}\left\langle k(\cdot,\boldsymbol{x}), k(\cdot,\boldsymbol{x}')\right\rangle_{\mathscr{H}} \boldsymbol{\delta}_{p,q}(\boldsymbol{x}')\Big] \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\Big[\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^{\mathsf{T}} k(\boldsymbol{x},\boldsymbol{x}')\,\boldsymbol{\delta}_{p,q}(\boldsymbol{x}')\Big],
\end{aligned}
$$

where we used the reproducing property, $k(\boldsymbol{x},\boldsymbol{x}') = \langle k(\cdot,\boldsymbol{x}), k(\cdot,\boldsymbol{x}')\rangle_{\mathscr{H}}$. This concludes the proof. □

Theorem 5.3.2. *Define the kernel function*

$$
\begin{aligned}
\kappa_p(\boldsymbol{x},\boldsymbol{x}') = {} & \boldsymbol{s}_p(\boldsymbol{x})^{\mathsf{T}} k(\boldsymbol{x},\boldsymbol{x}')\,\boldsymbol{s}_p(\boldsymbol{x}') - \boldsymbol{s}_p(\boldsymbol{x})^{\mathsf{T}}\Delta^{*}_{\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}') - \Delta^{*}_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}')^{\mathsf{T}}\boldsymbol{s}_p(\boldsymbol{x}') \\
& + \operatorname{tr}\Big(\Delta^{*}_{\boldsymbol{x},\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}')\Big),
\end{aligned} \tag{5.15}
$$

then

$$
\mathbb{D}(q\,\|\,p)^2 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\big[\kappa_p(\boldsymbol{x},\boldsymbol{x}')\big]. \tag{5.16}
$$

Proof. Expanding the expression for $\boldsymbol{\delta}_{p,q}(x)$ and applying Lemma 5.2.4 twice, we obtain

$$
\begin{aligned}
\mathbb{D}(q\,\|\,p)^2 &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\big[\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^{\mathsf{T}} k(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{\delta}_{p,q}(\boldsymbol{x}')\big] \\
&= \mathbb{E}_{\boldsymbol{x}\sim q}\big[\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^{\mathsf{T}}\mathbb{E}_{\boldsymbol{x}'\sim q}\big[k(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{\delta}_{p,q}(\boldsymbol{x}')\big]\big] \\
&= \mathbb{E}_{\boldsymbol{x}\sim q}\big[\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^{\mathsf{T}}\mathbb{E}_{\boldsymbol{x}'\sim q}\big[k(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{s}_p(\boldsymbol{x}') - \Delta^{*}_{\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}')\big]\big] \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\big[\boldsymbol{s}_p(\boldsymbol{x})^{\mathsf{T}} k(\boldsymbol{x},\boldsymbol{x}')\,\boldsymbol{s}_p(\boldsymbol{x}') - \boldsymbol{s}_p(\boldsymbol{x})^{\mathsf{T}}\Delta^{*}_{\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}') - \Delta^{*}_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}')^{\mathsf{T}}\boldsymbol{s}_p(\boldsymbol{x}') \\
&\qquad + \operatorname{tr}\Big(\Delta^{*}_{\boldsymbol{x},\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}')\Big)\big] \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\big[\kappa_p(\boldsymbol{x},\boldsymbol{x}')\big],
\end{aligned}
$$

which completes the proof. □

The next result justifies $\mathbb{D}(q\,\|\,p)$ as a divergence measure.

Lemma 5.3.3. *For a finite set $\mathcal{X}$, let p and q be positive pmfs on $\mathcal{X}^d$. Let $\mathscr{H}$ be an RKHS on $\mathcal{X}^d$ with kernel $k(\cdot,\cdot)$, and let $\mathbb{D}(q\,\|\,p)$ be defined as in Eq. (5.13). Assume that the Gram matrix $\mathbf{K} = [k(\boldsymbol{x},\boldsymbol{x}')]_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}^d}$ is strictly positive definite, then $\mathbb{D}(q\,\|\,p) = 0$ if and only if $p = q$.*

Proof. By Theorem 5.3.1, we have

$$\begin{aligned}\mathbb{D}(q \,\|\, p)^2 &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\left[\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^\mathsf{T} k(\boldsymbol{x},\boldsymbol{x}')\,\boldsymbol{\delta}_{p,q}(\boldsymbol{x}')\right] \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^d}\sum_{\boldsymbol{x}'\in\mathcal{X}^d} q(\boldsymbol{x})\boldsymbol{\delta}_{p,q}(\boldsymbol{x})^\mathsf{T} k(\boldsymbol{x},\boldsymbol{x}')\,\boldsymbol{\delta}_{p,q}(\boldsymbol{x}')q(\boldsymbol{x}'),\end{aligned}$$

where $\boldsymbol{\delta}_{p,q}(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x}) \in \mathbb{R}^d$. Denote the ℓ-th element of $\boldsymbol{\delta}_{p,q}$ by $\delta^\ell_{p,q}$, and write $\boldsymbol{g}_\ell := [q(\boldsymbol{x})\delta^\ell_{p,q}(\boldsymbol{x})]_{\boldsymbol{x}\in\mathcal{X}^d}$ for $\ell = 1,\ldots,d$. Then, $\mathbb{D}(q \,\|\, p)^2 = \sum_{\ell=1}^d \boldsymbol{g}_\ell^\mathsf{T}\mathbf{K}\boldsymbol{g}_\ell$. Since $\mathbf{K}$ is strictly positive-definite, $\mathbb{D}(q \,\|\, p)^2 = 0$ if and only if $\boldsymbol{g}_\ell = \mathbf{0}$ for all ℓ. Therefore, $\boldsymbol{\delta}_{p,q}(\boldsymbol{x}) = \mathbf{0}$ for all $\boldsymbol{x} \in \mathcal{X}^d$. By Theorem 5.2.1, this holds if and only if $p = q$. □

5.4 Goodness-of-Fit Testing via KDSD

Given a (possibly unnormalized) model distribution p and *i.i.d.*samples $\{\boldsymbol{x}_i\}_{i=1}^n$ from an unknown data distribution q on $\mathcal{X}^d$, we would like to measure the goodness-of-fit of the model distribution p to the observed data $\{\boldsymbol{x}_i\}_{i=1}^n$. To this end, we perform the hypothesis test $H_0 : p = q$ vs. $H_1 : p \neq q$ using the kernelized discrete Stein discrepancy (KDSD) measure. Denote $\mathbb{S}(q \,\|\, p) := \mathbb{D}(q \,\|\, p)^2$; we can estimate $\mathbb{S}(q \,\|\, p)$ via a U-statistic (Hoeffding, 1948) which provides a minimum-variance unbiased estimator:

$$\widehat{\mathbb{S}}(q \,\|\, p) = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \kappa_p(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{5.17}$$

As in the continuous case (*cf.* Section 3.3), the U-statistic $\widehat{\mathbb{S}}(q \,\|\, p)$ is asymptotically normal under the alternative hypothesis $H_1 : p \neq q$, but becomes degenerate under the null hypothesis $H_0 : p = q$. More precisely, we have the following result adapted from Theorem 3.3.1; its proof follows from standard asymptotic results of U-statistics.

Theorem 5.4.1 (Adapted from Theorem 4.1 of Liu et al. (2016))**.** *Let $k(\boldsymbol{x},\boldsymbol{x}')$ be a strictly positive definite kernel on $\mathcal{X}^d$, and assume that $\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\left[\kappa_p(\boldsymbol{x},\boldsymbol{x}')^2\right] < \infty$. We have the following two cases:*

(i) If $q \neq p$, then $\widehat{\mathbb{S}}(q \,\|\, p)$ is asymptotically normal:

$$\sqrt{n}\left(\widehat{\mathbb{S}}(q \,\|\, p) - \mathbb{S}(q \,\|\, p)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2),$$

where $\sigma^2 = \mathrm{Var}_{\boldsymbol{x}\sim q}(\mathbb{E}_{\boldsymbol{x}'\sim q}\left[\kappa_p(\boldsymbol{x},\boldsymbol{x}')\right]) > 0$.

(ii) If $q = p$*, then* $\sigma^2 = 0$*, and the U-statistic is degenerate:*

$$n\widehat{\mathbb{S}}(q\,\|\,p) \xrightarrow{\mathcal{D}} \sum_j c_j(Z_j^2 - 1),$$

where $\{Z_j\} \overset{iid}{\sim} \mathcal{N}(0,1)$ *and* $\{c_j\}$ *are the eigenvalues of the kernel* $\kappa_p(\cdot,\cdot)$ *under* q*.*

Since the asymptotic distribution of $\widehat{\mathbb{S}}(q\,\|\,p)$ under the null hypothesis cannot be easily calculated, we follow Liu et al. (2016) and adopt the bootstrap method for degenerate U-statistics (Arcones and Gine, 1992; Huskova and Janssen, 1993) to draw samples from the null distribution of the test statistic. Specifically, to obtain a bootstrap sample, we draw random multinomial weights $w_1,\ldots,w_n \sim \mathrm{Mult}(n; 1/n,\ldots,1/n)$, set $\widetilde{w}_i = (w_i - 1)/n$, and compute

$$\widehat{\mathbb{S}}^*(q\,\|\,p) = \sum_{i=1}^{n}\sum_{j\neq i}^{n} \widetilde{w}_i\widetilde{w}_j\kappa_p(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{5.18}$$

Upon repeating this procedure m times, we calculate the critical value of the test by taking the $(1-\alpha)$-th quantile of the bootstrapped statistics $\{\widehat{\mathbb{S}}_b^*\}_{b=1}^m$.

The overall goodness-of-fit testing procedure is summarized in Algorithm 1. Computing the test statistic in Eq. (5.17) takes $\mathcal{O}(n^2)$ time, where n is the number of observations, and the bootstrapping procedure takes $\mathcal{O}(mn^2)$ time, where m is the number of bootstrap samples used.

Kernel choice. A practical question that arises when performing the KDSD test is the choice of the kernel function $k(\cdot,\cdot)$ on $\mathcal{X}^d$. For continuous spaces, the RBF kernel might be a natural choice; Gorham and Mackey (2017) also provide further recommendations. For discrete spaces, a naive choice is the δ-kernel, $k(\boldsymbol{x},\boldsymbol{x}') = \mathbb{I}\{\boldsymbol{x} = \boldsymbol{x}'\}$, which suffers from the curse of dimensionality. A more sensible choice is the *exponentiated Hamming kernel*:

$$k(\boldsymbol{x},\boldsymbol{x}') = \exp\{-H(\boldsymbol{x},\boldsymbol{x}')\}, \tag{5.19}$$

where $H(\boldsymbol{x},\boldsymbol{x}') := \frac{1}{d}\sum_{i=1}^{d}\mathbb{I}\{x_i \neq x_i'\}$ is the normalized Hamming distance. The next lemma shows that Eq. (5.19) defines a positive definite kernel.

Algorithm 1 Goodness-of-fit testing via KDSD

1: **Input:** Difference score function $\boldsymbol{s}_p$ of p, data samples $\{\boldsymbol{x}_i\}_{i=1}^n \sim q$, kernel function $k(\cdot,\cdot)$, bootstrap sample size m, significance level α.
2: **Objective:** Test $H_0 : p = q$ vs. $H_1 : p \neq q$.
3: Compute test statistic $\widehat{\mathbb{S}}(q \,\|\, p)$ via Eq. (5.17).
4: **for** $b = 1, \ldots, m$ **do**
5: Compute bootstrap test statistic $\widehat{\mathbb{S}}_b^*$ via Eq. (5.18).
6: Compute critical value $\gamma_{1-\alpha}$ by taking the $(1-\alpha)$-th quantile of the bootstrap test statistics $\{\widehat{\mathbb{S}}_b^*\}_{b=1}^m$.
7: **Output:** Reject H_0 if test statistic $\widehat{\mathbb{S}}(q \,\|\, p) > \gamma_{1-\alpha}$, otherwise do not reject H_0.

Lemma 5.4.2. *The exponentiated Hamming kernel, as defined in Eq.* (5.19)*, is positive definite.*

Proof. Without loss of generality, assume that $\mathcal{X} = \{0, 1\}$ is a binary set; the general case can be easily accommodated by modifying the feature map to be described next. Define the feature map $\boldsymbol{\phi} : \mathcal{X}^d \to \mathcal{X}^{2d}$, $\boldsymbol{x} \mapsto \widetilde{\boldsymbol{x}}$, where $\widetilde{x}_{2i-1} = \mathbb{I}\{x_i = 0\}$ and $\widetilde{x}_{2i} = \mathbb{I}\{x_i = 1\}$ for $i = 1, \ldots, d$. Then, the normalized Hamming distance can be expressed as

$$H(\boldsymbol{x}, \boldsymbol{x}') = 1 - \frac{1}{d}\sum_{i=1}^{d} \mathbb{I}\{x_i = x_i'\} = 1 - \frac{1}{2d}\sum_{j=1}^{2d} \widetilde{x}_j \widetilde{x}_j' = 1 - \frac{1}{2d}\widetilde{\boldsymbol{x}}^\mathsf{T}\widetilde{\boldsymbol{x}}' = 1 - \frac{1}{2d}\boldsymbol{\phi}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x}').$$

Thus, $1 - H(\boldsymbol{x}, \boldsymbol{x}')$ is a positive definite kernel. By Taylor expansion, $\exp\{1 - H(\boldsymbol{x}, \boldsymbol{x}')\}$ (and hence $\exp\{-H(\boldsymbol{x}, \boldsymbol{x}')\}$) also constitutes a positive definite kernel on $\mathcal{X}^d$. □

When the inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$ encode additional structure about $\mathcal{X}^d$, the Hamming distance may no longer be appropriate. For instance, when $\boldsymbol{x} \in \{0, 1\}^{\binom{d}{2}}$ represents the (flattened) adjacency matrix of an undirected and unweighted graph on d vertices, two graphs $\boldsymbol{x}$ and $\boldsymbol{x}'$ may be isomorphic yet have non-zero Hamming distance. In this case, we can resort to the literature on graph kernels (Vishwanathan et al., 2010). Section 5.6 gives an example of using the *Weisfeiler-Lehman graph kernel* of Shervashidze et al. (2011) to test whether a set of graphs $\{\boldsymbol{x}_i\}_{i=1}^n$ comes from a specific distribution.

5.5 Related Work and Discussion

Stein's method. Related to our characterization via adjoint operators, Ley et al. (2017) also proposed the notion of a canonical Stein operator. Recently, Bresler and Nagaraj (2017); Reinert and Ross (2017) applied Stein's method to bound the distance between two stationary distributions of irreducible Markov chains in terms of their Glauber dynamics. Notably, they also make use of a difference operator for the binary case, and it is interesting to investigate whether their analysis techniques could be adopted for goodness-of-fit testing.

Goodness-of-fit tests. Closely related to our work is the kernelized Stein discrepancy test proposed independently by Chwialkowski et al. (2016); Liu et al. (2016) for smooth densities on continuous spaces. Our work further identifies and characterizes Stein operators for discrete domains, unifying them via Theorem 5.2.5 under a general framework for constructing Stein operators from adjoint operators. Under this framework, any Stein operator can be directly used to establish a KDSD test (under completeness conditions).

In addition to kernel-based tests, other forms of goodness-of-fit tests have also been examined for discrete distributions. Some recent examples include Valiant and Valiant (2016); Martín del Campo et al. (2017); Daskalakis et al. (2018). However, these tests are often model-specific, and typically assume that the normalization constant is easy to evaluate. In contrast, the KDSD test we propose is fully nonparametric, and applies to any unnormalized statistical model.

Score-matching methods. Proposed by Hyvärinen (2005), score-matching methods make use of score functions to perform parameter estimation in unnormalized models. Suppose we observe data $\{\boldsymbol{x}\}_{i=1}^{n}$ from some unknown density $q(\boldsymbol{x})$ which we would like to approximate using a parameterized model density $p(\boldsymbol{x};\boldsymbol{\theta})$. To estimate the parameters $\boldsymbol{\theta}$, score-matching methods minimize the *Fisher divergence*:

$$J(\boldsymbol{\theta}) = \int_{\xi\in\mathbb{R}^d} q(\xi)\,\|\nabla_\xi \log p(\xi;\boldsymbol{\theta}) - \nabla_\xi \log q(\xi)\|_2^2\,\mathrm{d}\xi.$$

Similar to the continuous KSD (Liu et al., 2016), if we set

$$k(\boldsymbol{x}, \boldsymbol{x}') = \frac{\mathbb{I}\{\boldsymbol{x} = \boldsymbol{x}'\}}{\sqrt{q(\boldsymbol{x})\, q(\boldsymbol{x}')}}$$

and apply Theorem 5.3.1, the KDSD statistic can be written as

$$\mathbb{D}^2(q \,\|\, p) = \mathbb{E}_{\boldsymbol{x} \sim q}\left[\|\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})\|_2^2\right],$$

which takes the same form as $J(\boldsymbol{\theta})$ with the continuous score function $\nabla \log p(\boldsymbol{x})$ replaced by the difference score function $\boldsymbol{s}_p(\boldsymbol{x})$.

Extensions of score-matching to discrete data have also been considered in Hyvärinen (2007); Lyu (2009); Amari (2016), and our work draws insights from these in the design of score functions for Stein operators. In particular, Lyu (2009) examined the connections between adjoint operators and Fisher divergence, and Amari (2016) discussed score functions for data from a graphical model. However, the connections to Stein operators and kernel-based hypothesis testing have not appeared in the score-matching literature.

Two-sample tests. Complementing goodness-of-fit tests (or one-sample tests) are two-sample tests, where we test if two collections of samples come from the same distribution. A well-known kernel two-sample test statistic is the *maximum mean discrepancy* (MMD) of Gretton et al. (2012) (see Section 3.2 for details). Given *i.i.d.* samples $\{\boldsymbol{x}_i\}_{i=1}^n \sim p$ and $\{\boldsymbol{y}_j\}_{j=1}^{n'} \sim q$, one could compute a U-statistic estimate of $\mathrm{MMD}(p, q)$ in $\mathcal{O}(nn')$ time (*cf.* Eq. (3.5)). The critical value of the test is calculated by bootstrapping on the aggregated data.

Two-sample tests can also be used as goodness-of-fit tests by comparing observed data with samples from the null model. For distributions with intractable normalization constants, obtaining exact samples from p could become very difficult or expensive. Further, approximate samples may introduce bias and/or correlation among the samples, violating the test assumptions, and leading to unpredictable test errors.

5.6 Applications

We apply the proposed KDSD goodness-of-fit test to three statistical models involving discrete distributions. We describe the models and derive their difference score functions in Section 5.6.1, and present experiments in Section 5.6.2.

5.6.1 Statistical Models

Ising model. The Ising model (Ising, 1924) is a canonical example of a Markov random field (MRF). Consider an (undirected) graph $G = (V, E)$, where each vertex $i \in V$ is associated with a binary *spin*. The collection of spins form a random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, whose components x_i and x_j $(i \neq j)$ interact directly only if $(i, j) \in E$. The pmf is

$$p_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\Theta})} \exp\left\{ \sum_{(i,j)\in E} \theta_{ij} x_i x_j \right\},$$

where θ_{ij} are the edge potentials and

$$Z(\boldsymbol{\Theta}) = \sum_{\boldsymbol{x}\in\mathcal{X}^d} \exp\left\{ \sum_{(i,j)\in E,\, i<j} \theta_{ij} x_i x_j \right\}$$

is the partition function which is prohibitive to compute when d is high. Recognizing the pmf as an exponential family distribution, we can apply Eq. (5.2) to obtain the difference score function:

$$(\boldsymbol{s}_p(\boldsymbol{x}))_i = 1 - \exp\left\{ -2x_i \sum_{j\in\mathcal{N}_i} \theta_{ij} x_j \right\},$$

where $\mathcal{N}_i := \{j : (i, j) \in E\}$ denotes the set of vertices adjacent to node i in graph G.

Bernoulli restricted Boltzmann machine (RBM). The RBM (Hinton, 2002) is an undirected graphical model consisting of a bipartite graph between visible units $\boldsymbol{v}$ and hidden units $\boldsymbol{h}$. In a Bernoulli RBM, both $\boldsymbol{v}$ and $\boldsymbol{h}$ are Bernoulli-distributed; $\mathcal{X} = \{0, 1\}$. The joint pmf of an RBM with M visible units $\boldsymbol{v}$ and K hidden units $\boldsymbol{h}$ is given by

$$p(\boldsymbol{h}, \boldsymbol{v}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})\},$$

with energy function

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = -(\boldsymbol{v}^\top \mathbf{W}\boldsymbol{h} + \boldsymbol{v}^\top \boldsymbol{b} + \boldsymbol{h}^\top \boldsymbol{c}),$$

where $\mathbf{W} \in \mathbb{R}^{M \times K}$ are the weights, $\boldsymbol{b} \in \mathbb{R}^M$ and $\boldsymbol{c} \in \mathbb{R}^K$ are the bias terms; $\boldsymbol{\theta} := (\mathbf{W}, \boldsymbol{b}, \boldsymbol{c})$, and

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp\{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})\}$$

is the partition function.

Marginalizing out the hidden variables $\boldsymbol{h}$, the pmf of $\boldsymbol{v}$ is given by

$$p(\boldsymbol{v}|\boldsymbol{\theta}) = \frac{1}{Z'(\boldsymbol{\theta})} \exp\{-F(\boldsymbol{v}; \boldsymbol{\theta})\},$$

where the *free energy* takes the form

$$F(\boldsymbol{v}; \boldsymbol{\theta}) = -\boldsymbol{v}^\top \boldsymbol{b} - \sum_{k=1}^{K} \log(1 + \exp\{\boldsymbol{v}^\top \mathbf{W}_{*k} + c_k\}),$$

and $Z'(\boldsymbol{\theta}) = \sum_{\boldsymbol{v}} \exp\{-F(\boldsymbol{v}; \boldsymbol{\theta})\}$ is another normalization constant. (Here, $\mathbf{W}_{*k}$ denotes the k-th column of $\mathbf{W}$.) Thus, we can write down the (difference) score function as

$$\begin{aligned}(\boldsymbol{s}_p(\boldsymbol{v}; \boldsymbol{\theta}))_i &= 1 - \exp\{F(\boldsymbol{v}; \boldsymbol{\theta}) - F(\neg_i \boldsymbol{v}; \boldsymbol{\theta})\} \\ &= 1 - e^{\tilde{v}_i b_i} \prod_{k=1}^{K} \frac{1 + \exp\{\boldsymbol{v}^\top \mathbf{W}_{*k} + \tilde{v}_i w_{ik} + c_k\}}{1 + \exp\{\boldsymbol{v}^\top \mathbf{W}_{*k} + c_k\}},\end{aligned}$$

where $\tilde{v}_i = \neg v_i - v_i$. Note that $\boldsymbol{s}_p(\boldsymbol{v}; \boldsymbol{\theta})$ is again free of normalization constants and can be easily evaluated.

Exponential random graph model (ERGM). The ERGM (Frank and Strauss, 1986; Wasserman and Pattison, 1996) is a well-studied statistical model for network data (see Section 2.1.1 for details). Recall that in a typical ERGM, the probability of observing an adjacency matrix $\boldsymbol{y} \in \{0, 1\}^{n \times n}$ is

$$p(\boldsymbol{y}) = \frac{1}{Z(\boldsymbol{\theta}, \tau)} \exp\left\{\sum_{k=1}^{n-1} \theta_k S_k(\boldsymbol{y}) + \tau T(\boldsymbol{y})\right\}.$$

Here, $S_k(\cdot)$ counts the number of edges ($k = 1$) or k-stars ($k \geq 2$), $T(\cdot)$ counts triangles, and $Z(\boldsymbol{\theta}, \tau)$ is the normalization constant.

We consider an ERGM distribution of undirected graphs $\boldsymbol{y}$ with three sufficient statistics: $S_1(\boldsymbol{y})$, the number of edges (1-stars); $S_2(\boldsymbol{y})$, the number of wedges (2-stars); and $T(\boldsymbol{y})$, the number of triangles.[2] The parameters for these sufficient statistics are θ_1, θ_2, and τ, respectively. The score function can be written as

$$(\boldsymbol{s}_p(\boldsymbol{y}))_{ij} = 1 - \exp\{\theta_1 \delta_1(\boldsymbol{y}) + \theta_2 \delta_2(\boldsymbol{y}) + \tau \delta_3(\boldsymbol{y})\},$$

with the *change statistics* given by

$$\begin{aligned}
\delta_1(\boldsymbol{y}) &:= [S_1(\neg_{ij}\boldsymbol{y}) - S_1(\boldsymbol{y})] = (-1)^{y_{ij}} \\
\delta_2(\boldsymbol{y}) &:= [S_2(\neg_{ij}\boldsymbol{y}) - S_2(\boldsymbol{y})] = (-1)^{y_{ij}}(|\mathcal{N}_i^{\backslash j}| + |\mathcal{N}_j^{\backslash i}|) \\
\delta_3(\boldsymbol{y}) &:= [T(\neg_{ij}\boldsymbol{y}) - T(\boldsymbol{y})] = (-1)^{y_{ij}}|\mathcal{N}_i \cap \mathcal{N}_j|
\end{aligned}$$

where $\mathcal{N}_i$ denotes the neighbor-set of node i, and $\mathcal{N}_i^{\backslash j} := \mathcal{N}_i \backslash \{j\}$.

5.6.2 Experiments

We apply the kernelized discrete Stein discrepancy (KDSD) test to the statistical models described in Sections 5.6.1.[3] In the absence of established baselines, we compare with a two-sample test based on the maximum mean discrepancy (MMD) (see Section 4.5). For both KDSD and MMD, we utilize the exponentiated Hamming kernel (Eq. (5.19)) for the Ising model and RBM, and the Weisfeiler-Lehman graph kernel (Shervashidze et al., 2011) for the ERGM.

Setup. Denote the null model distribution by p and the alternative distribution by q. For each distribution, we draw exact *i.i.d.*samples by running n independent Markov chains with different random initializations, each for 10^5 iterations, and collecting only the last sample of each chain. For KDSD, we draw n samples from q; for MMD, we draw n samples from q and another n samples from p. Under this setup, both KDSD and MMD

[2] Notice that the sufficient statistics are not independent: *e.g.*, $S_2(\boldsymbol{y}) > T(\boldsymbol{y})$ since every triangle contains three 2-stars.

[3] Code for the experiments is available at `https://github.com/jiaseny/kdsd`.

takes time $\mathcal{O}(mn^2)$, where m is the number of bootstrap samples used to determine the critical threshold. We set $m = 5000$ for both methods throughout.

For each model, we choose a "perturbation parameter" and fix its value for the null distribution p, while drawing data samples under various values of the perturbation parameter. We also vary the sample size n to examine the performance of the test as n increases. For each value of the perturbation parameter and each sample size n, we conduct 500 independent trials. In each trial, we first randomly flip a fair coin to decide whether to set the alternative distribution q to be the same as p or with a different value of the perturbation parameter. (In the former case, the null hypothesis $H_0 : p = q$ should not be rejected, and in the latter case it should be.) Then, we draw n independent samples from q (for KDSD) or both p and q (for MMD) and perform the hypothesis test $H_0 : p = q$ vs. $H_1 : p \neq q$ under significance level $\alpha = 0.05$. We evaluate the performance of the KDSD and MMD tests in terms of their false-positive rate (FPR; Type-I error) and false-negative rate (FNR; Type-II error), and report the results across 500 independent trials.

Ising model. We consider a periodic 10-by-10 lattice, with $d = 100$ random variables. We focus on the ferromagnetic setting and set $\theta_{ij} = 1/T$, where T is the temperature of the system. For $T_0 \in \{5, 20\}$ and various values of T', we test the hypotheses $H_0 : T = T_0$ vs. $H_1 : T \neq T_0$ using data samples drawn from the model under $T = T'$. To draw samples from the Ising model, we apply the Metropolis algorithm: in each iteration, we propose to flip the spin of a randomly chosen variable x_i, and adopt this proposal with probability $\min\left(1, \exp\left\{-2x_i \sum_{j \in \mathcal{N}_i} \theta_{ij} x_j\right\}\right)$.

Bernoulli RBM. We use $M = 50$ visible units and $K = 25$ hidden units. We draw the entries of the weight matrix $\mathbf{W}$ *i.i.d.* from a Normal distribution with mean zero and standard deviation $1/M$, and the entries of the bias terms $\boldsymbol{b}$ and $\boldsymbol{c}$ *i.i.d.* from the standard Normal distribution. We corrupt the weights in $\mathbf{W}$ by adding *i.i.d.* Gaussian noise with mean zero and standard deviation σ, and test the hypotheses $H_0 : \sigma = 0$ (no-corruption) vs. $H_1 : \sigma \neq 0$ using data samples drawn under $\sigma = \sigma'$ for various values of σ'. To draw

samples from the RBM, we perform block Gibbs sampling by exploiting the bipartite structure of the graphical model.

ERGM. We consider an ERGM distribution for undirected graphs on 20 nodes, with the dimension of each sample $d = \binom{20}{2} = 190$. We fix $\theta_1 = -2$ and $\tau = 0.01$, For various values of the 2-star parameter θ_2', we test the hypotheses $H_0 : \theta_2 = 0$ vs. $H_1 : \theta_2 \neq 0$ using data samples drawn under $\theta_2 = \theta_2'$. To draw MCMC samples from the ERGM, we utilize the `ergm` R package (Handcock et al., 2017).

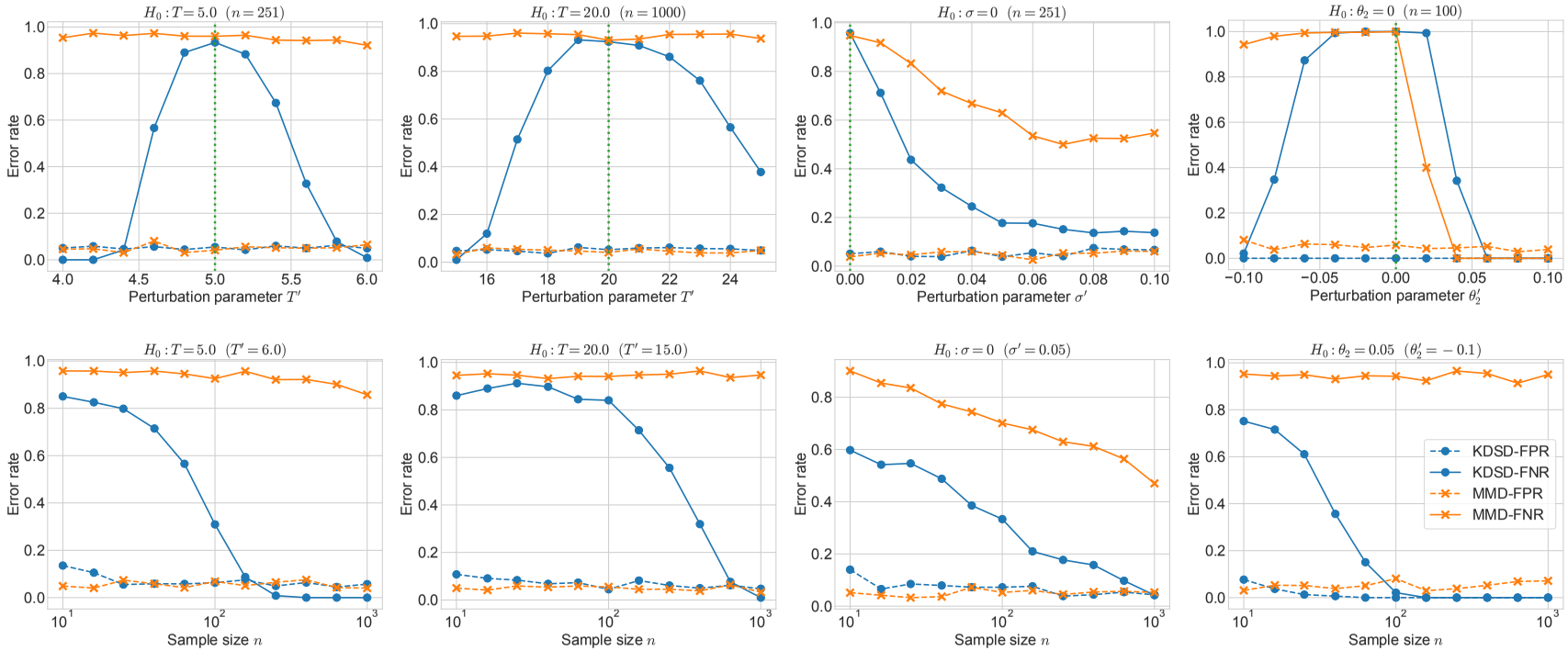
Results. In Figure 5.1, the top row plots the testing error rate vs. different values of the perturbation parameter in H_1, for a fixed H_0 and sample size; while the bottom row plots the error rate vs. sample size n for a fixed pair of H_0 and H_1. We observe that both KDSD and MMD maintain a false-positive rate (Type-I error) around or below the significance level $\alpha = 0.05$. In addition, KDSD consistently achieves lower false-negative rate (Type-II error) than MMD in most cases, indicating that KDSD, by utilizing the score function information of p, leads to a more powerful test.

It is interesting to note that in the ERGM example, MMD exhibits higher power than KDSD when the data samples were drawn from an ERGM distribution with $\theta_2' \in (0, 0.05)$ (roughly). We hypothesize that this may correspond to a regime in which a small change in θ_2 causes a subtle change in the *global* graph structure that can be more easily detected by MMD, while the difference Stein operator of Section 5.2.1 may be more adapt in detecting *local* differences. Thus, the performance of the KDSD test could be improved by constructing Stein operators (using the characterization of Section 5.2.2) that exploit higher-order structure in the graph samples, which shall be investigated in future work.

5.7 Summary

In this chapter, we have defined a difference Stein operator for discrete spaces, and introduced a kernelized Stein discrepancy measure for discrete probability distributions. This enabled us to establish a nonparametric goodness-of-fit test for discrete distributions with intractable normalization constants. Furthermore, we have proposed a general

characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators. We have applied the proposed goodness-of-fit test to three statistical models involving discrete distributions, and shown that it typically outperforms a two-sample test based on the maximum mean discrepancy.



(a) Ising model ($T_0 = 5$) (b) Ising model ($T_0 = 20$) (c) Bernoulli RBM (d) ERGM

Figure 5.1.: *Top row:* KDSD and MMD testing error rate vs. perturbation parameter (the vertical dotted lines indicate the value of the perturbation parameter under H_0). *Bottom row:* KDSD and MMD testing error rate vs. sample size.

6. A STEIN–PAPANGELOU GOODNESS-OF-FIT TEST FOR POINT PROCESSES

Point processes provide a powerful framework for modeling the distribution and interactions of events in time or space (see Section 2.2 for details). Their flexibility has given rise to a variety of sophisticated models in statistics and machine learning, yet model diagnostic and criticism techniques remain underdeveloped. In this chapter, we propose a general Stein operator for point processes based on the Papangelou conditional intensity function. We then establish a kernel goodness-of-fit test by defining a Stein discrepancy measure for general point processes. Notably, our test also applies to non-Poisson point processes whose intensity functions contain intractable normalization constants due to the presence of complex interactions among points. We apply our proposed test to several point process models, and show that it outperforms a two-sample test based on the maximum mean discrepancy.

6.1 Introduction

Point processes have been the subject of much recent activity in statistics and machine learning, and a spate of sophisticated probabilistic and deep neural network models have been developed (Reinhart, 2018; Linderman and Adams, 2014; Du et al., 2016; Zaheer et al., 2017; Xiao et al., 2017). While the complexity of such point process models has grown at a rapid pace, corresponding tools for model diagnostics, evaluation, and criticism have lagged behind, restricted mostly to the spatial statistics literature. Beyond Poisson-type processes (Daley and Vere-Jones, 2008; Brown et al., 2002), and residual-based analysis and diagnostic plots for some spatial processes (Baddeley et al., 2005), rigorous statistical tests to assess how well a point process model fits the observed data remains an important and under-studied topic (Coeurjolly and Lavancier, 2013).

In this chapter, we investigate an important model criticism technique—the goodness-of-fit test—for point processes. Well-established goodness-of-fit tests for point processes, are only available under the simplest scenarios—such as when the null model is a Poisson process. For more general point processes, the construction of such tests typically rely on pseudo-likelihood approximations (Strauss and Ikeda, 1990) which introduce biases and errors that are hard to quantify, or heuristic summary statistics (such as Ripley's K-function) which could only capture certain aspects of the observed data and may lead to a considerable loss of statistical power.

A major hurdle preventing the construction of rigorous statistical tests (such as those based on the likelihood-ratio statistic) for more sophisticated point processes is the presence of *intractable normalization constants* in the density/intensity functions. For many widely used models that capture pairwise or higher-order dependencies between points, these functions can often be evaluated only up to a normalization constant, because summing over all possible configurations leads to an intractable infinite-dimensional integral. This precludes the use of classical tests (such as the likelihood-ratio test) which require the fully specified model density.

In Section 3.3 and Chapter 5, we discussed recent developments in nonparametric goodness-of-fit testing based on Stein's method (Stein, 1972, 1986), which work directly with unnormalized probability distributions (Gorham and Mackey, 2015; Chwialkowski et al., 2016; Liu et al., 2016; Jitkrittum et al., 2017; Yang et al., 2018). Central to these tests is the notion of a *Stein operator* (*cf.* Section 3.3.1) $\mathcal{A}_p$ such that, for functions f in some family $\mathscr{F}$, the expectation $\mathbb{E}\left[\mathcal{A}_p f\right]$ equals zero only under the distribution-of-interest p. All the aforementioned works have considered distributions over *fixed-length* (d-dimensional) vectors residing in a space $\mathbb{X}$ that is either the Euclidean space $\mathbb{R}^d$ (for continuous distributions) or $\mathcal{X}^d$ where $\mathcal{X}$ is a finite set (for discrete distributions). These works have shown how to construct Stein operators (and goodness-of-fit tests) which only require unnormalized probability densities. On the other hand, a realization of a point process is a *set* containing an arbitrary number of points, and forms an element of

an infinite-dimensional space. Constructing a Stein operator for this setting does not follow easily from previous work, and requires a new set of tools.

A primary contribution of this chapter is to identify a suitable Stein operator for general point processes. While such constructions have been well-studied for Poisson process approximations in the probability literature (Barbour, 1988; Barbour and Brown, 1992), constructions for general point processes have been largely unexplored. Our key technical tool in constructing a general Stein operator is the *Papangelou conditional intensity* of a point process (see Section 2.2.2). Importantly, any (intractable) normalization constant in the density or intensity function of the point process cancels out when evaluating the Papangelou conditional intensity. Using our proposed Stein operator, along with a suitable kernel function on the space of point configurations, we proceed to define a kernelized Stein discrepancy measure between distributions, following a similar strategy as in Section 3.3 and Chapter 5. This allows us to develop a computationally feasible, nonparametric goodness-of-fit test for general point processes, including those with intractable normalization constants (*e.g.*, the Gibbs process). We apply our proposed goodness-of-fit test to the Poisson process, as well as two processes with inter-point interactions: the *Hawkes process* (Hawkes, 1971) exhibiting self-excitation, and the *Strauss process* (Strauss, 1975) featuring repulsion. Our experiments show that the proposed test outperforms a two-sample test based on the maximum mean discrepancy (*cf.* Section 3.2) in terms of power while maintaining control on false-positive rate.

6.2 Stein Operators for Point Processes

As discussed in Section 3.3.1, Stein's method involves identifying an operator $\mathcal{A}$ that satisfies Stein's identity: a random variable Φ is distributed according to the probability measure μ if and only if $\mathbb{E}_\mu[\mathcal{A}_\mu h(\Phi)] = 0$ for all functions h in some class $\mathcal{H}$. When Φ is real-valued, $\mathcal{A}$ can be characterized through a simple differential operator (the Langevin Stein operator of Eq. (3.11)), with Stein's identity following easily from integration-by-parts. When ϕ is discrete-valued, an alternative Stein operator using partial differences

was provided in Yang et al. (2018). However, when ϕ is a point process—a random variable taking values in an infinite-dimensional space $\mathcal{N}_{\mathbb{X}}$—we will require a new set of tools, based on the *generator method* of Barbour (1988).

We begin by reviewing the Stein operator for the Poisson process, and then propose a general Stein operator for arbitrary finite point processes. Our proposed Stein operator can be easily evaluated for point processes whose intensity functions contain intractable normalization constants, such as the Gibbs process.

6.2.1 Stein Operator for the Poisson Process

Stein's method for Poisson process approximation was pioneered by Barbour and Brown (1992), using the generator method of Barbour (1988). For a Poisson process Φ_μ on $\mathbb{X}$ with mean measure μ, Barbour and Brown (1992) considered an immigration-death process on $\mathbb{X}$ with immigration intensity μ and unit per-capita death rate. This process has stationary distribution Φ_μ, and infinitesimal generator $\mathcal{A}_\mu$ given by

$$(\mathcal{A}_\mu h)(\phi) = \int_{\mathbb{X}} [h(\phi + \delta_x) - h(\phi)]\,\mu(\mathrm{d}x) + \int_{\mathbb{X}} [h(\phi - \delta_x) - h(\phi)]\,\phi(\mathrm{d}x) \tag{6.1}$$

for any configuration $\phi \in \mathcal{N}_{\mathbb{X}}$. Notably, the infinitesimal generator $\mathcal{A}_\mu$ *characterizes* the Poisson process Φ_μ, as demonstrated by the following result:

Theorem 6.2.1 (Stein identity for the Poisson process (Barbour and Brown, 1992))**.** *Let $\mathcal{A}_\mu$ be the infinitesimal generator defined in Eq. (6.1). A point process Φ on $\mathbb{X}$ is a Poisson process with intensity measure μ if and only if for any measurable and bounded function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$,*

$$\mathbb{E}\big[\mathcal{A}_\mu h(\Phi)\big] = 0. \tag{6.2}$$

In the literature on Stein's method (Stein, 1986), an operator $\mathcal{A}$ that characterizes the distribution of Φ is called a *Stein operator*, and Eq. (6.2) a *Stein identity*.

Although Barbour (1988) derived the expression of $\mathcal{A}$ using the generator method, Theorem 6.2.1 can actually be viewed as a direct consequence of the Mecke formula

(Theorem 2.2.1). This hints at a possible generalization of the Stein operator for Poisson processes in Eq. (6.1) to general (finite) point processes, which we discuss next.

6.2.2 The Stein–Papangelou Operator

Stein's method for Poisson process approximation has been extensively studied since Barbour and Brown (1992), yet few works have considered more general point processes such as Hawkes processes and Gibbs processes (with the exceptions of Schuhmacher and Stucki (2014); Decreusefond and Vasseur (2018)). Here, we propose a generalization of the Stein operator in Eq. (6.1) to general (finite) point processes on $\mathbb{X}$. Our key insight is the analogy between the Mecke formula (Theorem 2.2.1) for Poisson processes and the GNZ formula (Theorem 2.2.2) for general point processes.

We begin by providing an interpretation of the right-hand side of Eq. (6.1). From the complete randomness of the Poisson process, $\mu(\mathrm{d}x) = \lambda(x)\,\mathrm{d}x$ gives the conditional intensity of an event at location x given the rest of the Poisson process realization ϕ. Then, the first integral equals the expected change in the value of the function h if a new event were added to the point process realization. Similarly, the second term gives the average change in h if one of the events were *removed* from ϕ. For a point process model with interactions, the conditional intensity at location x will depend on the rest of the point process realization; indeed, this is exactly the Papangelou conditional intensity $\rho(x|\phi)$. Thus, it is natural to consider substituting the intensity function $\lambda(x)$ with the Papangelou conditional intensity $\rho(x|\phi)$. Somewhat surprisingly, we can show that the resulting expression still gives a valid Stein operator for the associated point process.

To simplify presentation, let us define the 'inclusion' and 'exclusion' functionals $\mathcal{D}_x^+, \mathcal{D}_x^-$ at a point $x \in \mathbb{X}$ as

$$(\mathcal{D}_x^+ h)(\phi) := h(\phi + \delta_x) - h(\phi);$$
$$(\mathcal{D}_x^- h)(\phi) := h(\phi) - h(\phi - \delta_x),$$

for any measurable and bounded function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ and (finite) point configuration $\phi \in \mathcal{N}_{\mathbb{X}}$. Using these notations, we have the following definition:

Definition 6.2.1 (Stein–Papangelou operator for finite point processes)**.** *Let* $\rho : \mathbb{X}\times\mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ *be the Papangelou conditional intensity of a finite point process on* $\mathbb{X}$*. Define the* Stein–Papangelou operator $\mathcal{A}_\rho$ *via*

$$\begin{aligned}(\mathcal{A}_\rho h)(\phi) &= \int_{\mathbb{X}} (\mathcal{D}_x^+ h)(\phi)\rho(x|\phi)\,\mathrm{d}x - \int_{\mathbb{X}} (\mathcal{D}_x^- h)(\phi)\phi(\mathrm{d}x) \\ &= \int_{\mathbb{X}} [h(\phi+\delta_x) - h(\phi)]\,\rho(x|\phi)\,\mathrm{d}x + \sum_{x\in\phi} [h(\phi-\delta_x) - h(\phi)] \end{aligned} \tag{6.3}$$

for any function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ *and configuration* $\phi \in \mathcal{N}_{\mathbb{X}}$*.*

Notice that Eq. (6.3) reduces to Eq. (6.1) for a Poisson process, since its Papangelou conditional intensity equals its intensity function: $\rho(x|\phi)\,\mathrm{d}x = \lambda(x)\,\mathrm{d}x = \mu(\mathrm{d}x)$. A crucial advantage of Eq. (6.3) is that the Stein operator $\mathcal{A}_\rho$ now depends only on the Papangelou conditional intensity ρ of the point process, which is usually easy to obtain even when the point process likelihood itself is computationally intractable.

We conclude this section by showing that Eq. (6.3) indeed defines a valid Stein operator for general (finite) point processes—*i.e.*, that it satisfies a Stein identity.

Theorem 6.2.2 (Stein identity for finite point processes)**.** *Let* Φ *be a finite point process on* $\mathbb{X}$ *with Papangelou conditional intensity* $\rho : \mathbb{X}\times\mathcal{N}_{\mathbb{X}} \to \mathbb{R}$*, and let* $\mathcal{A}_\rho$ *be the operator defined via Eq.* (6.3)*. Then, we have*

$$\mathbb{E}\big[\mathcal{A}_\rho h(\Phi)\big] = 0 \tag{6.4}$$

for all measurable and bounded functions $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$*.*

Proof. To prove Eq. (6.4), it suffices to show that

$$\mathbb{E}\left[\int_{\mathbb{X}} (\mathcal{D}_x^+ h)(\Phi)\,\rho(x|\Phi)\,\mathrm{d}x\right] = \mathbb{E}\left[\sum_{x\in\Phi} (\mathcal{D}_x^- h)(\Phi)\right]$$

for any function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ and configuration $\phi \in \mathcal{N}_{\mathbb{X}}$. Notice that for any $x \in \phi$, $(D_x^- h)(\phi) = h(\phi) - h(\phi-\delta_x) = h(\phi-\delta_x+\delta_x) - h(\phi-\delta_x) = (D_x^+ h)(\phi-\delta_x)$. Thus, applying the GNZ formula (Theorem 2.2.2) with $h(x,\Phi) := (D_x^+ h)(\Phi)$ gives the desired result. □

A similar idea, but under a different context, has also been proposed in the probability literature (Schuhmacher and Stucki, 2014).

6.3 Stein Discrepancy and Goodness-of-Fit Testing

Equipped with a proper Stein operator, we are now ready to define a notion of *discrepancy* between two point processes with different intensity measures.

6.3.1 (Kernelized) Stein Discrepancy

Following a central observation made by Gorham and Mackey (2015) under the context of continuous distributions with smooth densities, we note that since the Stein identity of Eq. (6.4) holds when the point process Φ has Papangelou conditional intensity ρ (denoted $\Phi \sim \rho$), one could consider the *maximum violation* of Eq. (6.4) when $\Phi \sim \eta \neq \rho$ by choosing test functions within a function class $\mathscr{F}$. This leads to the following definition:[1]

Definition 6.3.1 (Stein discrepancy for point processes)**.** *Let Φ be a finite point process on $\mathbb{X}$ with Papangelou conditional intensity $\rho : \mathbb{X} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$, and let $\mathcal{A}_\rho$ be the Stein operator defined via Eq.* (6.3)*. For a family $\mathscr{F}$ of functions $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$, define the* Stein discrepancy *between Papangelou conditional intensities η and ρ as*

$$\mathbb{D}_{\mathscr{F}}(\eta \,\|\, \rho) := \sup_{h \in \mathscr{F}} \mathbb{E}_{\Phi \sim \eta}\big[\mathcal{A}_\rho h(\Phi)\big]. \tag{6.5}$$

Clearly, $\mathbb{D}_{\mathscr{F}}(\eta \,\|\, \rho) = 0$ when $\eta \equiv \rho$. While in principle the Stein discrepancy can be defined with respect to any family of functions $\mathscr{F}$, in practice we need to choose a function space that is both rich enough to ensure that the resulting Stein discrepancy has sufficient discriminative power, yet also suitably tractable such that Eq. (6.5) can be efficiently computed.

Toward this end, we follow Chwialkowski et al. (2016); Liu et al. (2016) and take $\mathscr{F}$ to be the unit-ball in a *reproducing kernel Hilbert space* (RKHS). Specifically, let $k : \mathcal{N}_{\mathbb{X}} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ be a positive definite (p.d.) kernel on the space of finite point configurations $\mathcal{N}_{\mathbb{X}}$ (Section 6.3.3 discusses various choices of k), and let $\mathscr{H}_k$ be its associated RKHS (consisting of functions $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$). We have the following definition:

[1] As Eq. (6.3) reduces to Eq. (6.1) for Poisson processes, we present all results using the Stein–Papangelou operator.

Definition 6.3.2. *The* kernelized Stein discrepancy *(KSD) between finite point processes with Papangelou conditional intensities η and ρ is*

$$\mathbb{D}_{\mathscr{H}_k}(\eta \,\|\, \rho) := \sup_{h \in \mathscr{H}_k,\, \|h\|_{\mathscr{H}_k} \leq 1} \mathbb{E}_{\Phi \sim \eta}\big[\mathcal{A}_\rho h(\Phi)\big], \tag{6.6}$$

where $\mathscr{H}_k$ is the RKHS of a p.d. kernel $k(\cdot,\cdot)$ on $\mathcal{N}_{\mathbb{X}}$.

Using the reproducing property of $\mathscr{H}_k$, our next result shows that Eq. (6.6) can actually be evaluated in closed-form. This follows directly from Liu et al. (2016); due to space constraints, we defer its proof to the Appendix.

Theorem 6.3.1. *The squared-KSD can be expressed as*

$$\mathbb{D}^2_{\mathscr{H}_k}(\eta \,\|\, \rho) = \mathbb{E}_{\Phi,\Psi \sim \eta}\big[\kappa_\rho(\Phi,\Psi)\big], \tag{6.7}$$

where $\kappa_\rho(\phi,\psi) := \mathcal{A}^\psi_\rho \mathcal{A}^\phi_\rho k(\phi,\psi)$ is a kernel function on $\mathcal{N}_{\mathbb{X}}$ obtained by applying the Stein operator $\mathcal{A}$ twice on each argument of the reproducing kernel $k(\cdot,\cdot)$ of $\mathscr{H}_k$. Its expression is given by

$$\begin{aligned}
&\kappa_\rho(\phi,\psi) \\
&= \int_{\mathbb{X}}\int_{\mathbb{X}} \Big[k(\phi+\delta_u,\psi+\delta_v) - k(\phi,\psi+\delta_v) - k(\phi+\delta_u,\psi) + k(\phi,\psi)\Big]\rho(u|\phi)\,\rho(v|\psi)\,\mathrm{d}u\,\mathrm{d}v \\
&\quad + \int_{\mathbb{X}}\Big[\sum_{x\in\phi}\big[k(\phi-\delta_x,\psi+\delta_v) - k(\phi-\delta_x,\psi)\big] - |\phi|\cdot\big[k(\phi,\psi+\delta_v) - k(\phi,\psi)\big]\Big]\rho(v|\psi)\,\mathrm{d}v \\
&\quad + \int_{\mathbb{X}}\Big[\sum_{y\in\psi}\big[k(\phi+\delta_u,\psi-\delta_y) - k(\phi,\psi-\delta_y)\big] - |\psi|\cdot\big[k(\phi+\delta_u,\psi) - k(\phi,\psi)\big]\Big]\rho(u|\phi)\,\mathrm{d}u \\
&\quad + \Big[\sum_{x\in\phi}\sum_{y\in\psi} k(\phi-\delta_x,\psi-\delta_y) - |\phi|\cdot\sum_{y\in\psi} k(\phi,\psi-\delta_y) - |\psi|\cdot\sum_{x\in\phi} k(\phi-\delta_x,\psi) + |\phi|\cdot|\psi|\cdot k(\phi,\psi)\Big].
\end{aligned} \tag{6.8}$$

Proof. By the reproducing property of $\mathscr{H}_k$, $h(\phi) = \langle h(\cdot), k(\phi,\cdot)\rangle_{\mathscr{H}_k}$. For any $x \in \mathbb{X}$, we have

$$\begin{aligned}
(\mathcal{D}^+_x h)(\phi) &= \langle h(\cdot), k(\phi+\delta_x,\cdot)\rangle_{\mathscr{H}_k} - \langle h(\cdot), k(\phi,\cdot)\rangle_{\mathscr{H}_k} = \big\langle h(\cdot), (\mathcal{D}^+_x k)(\phi,\cdot)\big\rangle_{\mathscr{H}_k}; \\
(\mathcal{D}^-_x h)(\phi) &= \langle h(\cdot), k(\phi,\cdot)\rangle_{\mathscr{H}_k} - \langle h(\cdot), k(\phi-\delta_x,\cdot)\rangle_{\mathscr{H}_k} = \big\langle h(\cdot), (\mathcal{D}^-_x k)(\phi,\cdot)\big\rangle_{\mathscr{H}_k}.
\end{aligned}$$

Thus, by Eq. (6.3),

$$\begin{aligned}
\mathbb{E}_{\Phi\sim\eta}\left[\mathcal{A}_\rho h(\Phi)\right] &= \mathbb{E}_{\Phi\sim\eta}\left[\int_{\mathbb{X}} (\mathcal{D}_x^+ h)(\Phi)\rho(x|\Phi)\,\mathrm{d}x - \int_{y\in\mathbb{X}} (\mathcal{D}_x^- h)(\Phi)\Phi(\mathrm{d}x)\right] \\
&= \mathbb{E}_{\Phi\sim\eta}\left[\int_{\mathbb{X}} \left\langle h(\cdot), (\mathcal{D}_x^+ k)(\Phi,\cdot)\right\rangle_{\mathscr{H}_k} \rho(x|\Phi)\,\mathrm{d}x - \int_{y\in\mathbb{X}} \left\langle h(\cdot), (\mathcal{D}_x^- k)(\Phi,\cdot)\right\rangle_{\mathscr{H}_k} \Phi(\mathrm{d}x)\right] \\
&= \mathbb{E}_{\Phi\sim\eta}\left[\left\langle h(\cdot), \int_{\mathbb{X}} (\mathcal{D}_x^+ k)(\Phi,\cdot)\,\rho(x|\Phi)\,\mathrm{d}x\right\rangle_{\mathscr{H}_k} - \left\langle h(\cdot), \int_{\mathbb{X}} (\mathcal{D}_x^- k)(\Phi,\cdot)\,\Phi(\mathrm{d}x)\right\rangle_{\mathscr{H}_k}\right] \\
&= \left\langle h(\cdot), \mathbb{E}_{\Phi\sim\eta}\left[\mathcal{A}_\rho k(\Phi,\cdot)\right]\right\rangle_{\mathscr{H}_k}.
\end{aligned}$$

Defining $\beta_{\eta,\rho} := \mathbb{E}_{\Phi\sim\eta}\left[\mathcal{A}_\rho k(\Phi,\cdot)\right]$, we can rewrite the kernelized Stein discrepancy as

$$\mathbb{D}_{\mathscr{H}_k}(\eta\,\|\,\rho) = \sup_{h\in\mathscr{H},\|h\|_{\mathscr{H}_k}\le 1} \left\langle h,\, \beta_{\eta,\rho}\right\rangle_{\mathscr{H}_k},$$

which immediately implies that $\mathbb{D}_{\mathscr{H}_k}(\eta\,\|\,\rho) = \|\beta_{\eta,\rho}\|_{\mathscr{H}_k}$ since the supremum will be obtained by $h = \beta_{\eta,\rho}/\|\beta_{\eta,\rho}\|_{\mathscr{H}_k}$. Therefore, we can write

$$\begin{aligned}
\mathbb{D}^2_{\mathscr{H}_k}(\eta\,\|\,\rho) = \|\beta_{\eta,\rho}\|^2_{\mathscr{H}_k} &= \left\langle \mathbb{E}_{\Phi\sim\eta}\left[\mathcal{A}_\rho^\Phi k(\Phi,\cdot)\right], \mathbb{E}_{\Psi\sim\eta}\left[\mathcal{A}_\rho^\Psi k(\Psi,\cdot)\right]\right\rangle_{\mathscr{H}_k} \\
&= \mathbb{E}_{\Phi,\Psi\sim\eta}\left[\left\langle \mathcal{A}_\rho^\Phi k(\Phi,\cdot), \mathcal{A}_\rho^\Psi k(\Psi,\cdot)\right\rangle_{\mathscr{H}_k}\right] \\
&= \mathbb{E}_{\Phi,\Psi\sim\eta}\left[\mathcal{A}_\rho^\Phi \mathcal{A}_\rho^\Psi \left\langle k(\Phi,\cdot), k(\Psi,\cdot)\right\rangle_{\mathscr{H}_k}\right] = \mathbb{E}_{\Phi,\Psi\sim\eta}\left[\mathcal{A}_\rho^\Phi \mathcal{A}_\rho^\Psi k(\Phi,\Psi)\right],
\end{aligned}$$

where we applied the reproducing property, $\left\langle k(\Phi,\cdot), k(\Psi,\cdot)\right\rangle_{\mathscr{H}_k} = k(\Phi,\Psi)$.

Deriving the expression in Eq. (6.8). Fixing ψ and applying Eq. (6.3) to $k(\phi,\psi)$ viewed as a function of ϕ, we have

$$\mathcal{A}_\rho^\phi k(\phi,\psi) = \int_{\mathbb{X}} [k(\phi+\delta_u,\psi) - k(\phi,\psi)]\,\rho(u|\phi)\,\mathrm{d}u + \sum_{x\in\phi}[k(\phi-\delta_x,\psi) - k(\phi,\psi)].$$

Now, fixing ϕ and applying Eq. (6.3) to $\mathcal{A}_\rho^\phi k(\phi,\psi)$ viewed as a function of ψ, we have

$$\begin{aligned}
&\mathcal{A}_\rho^\psi \mathcal{A}_\rho^\phi k(\phi,\psi) \\
&= \int_{\mathbb{X}} \left[\mathcal{A}_\rho^\phi k(\phi,\psi+\delta_v) - \mathcal{A}_\rho^\phi k(\phi,\psi)\right]\rho(v|\psi)\,\mathrm{d}v + \sum_{y\in\psi}\left[\mathcal{A}_\rho^\phi k(\phi,\psi-\delta_y) - \mathcal{A}_\rho^\phi k(\phi,\psi)\right] \\
&= \int_{\mathbb{X}} \left[\left(\int_{\mathbb{X}} [k(\phi+\delta_u,\psi+\delta_v) - k(\phi,\psi+\delta_v)]\,\rho(u|\phi)\,\mathrm{d}u + \sum_{x\in\phi}[k(\phi-\delta_x,\psi+\delta_v) - k(\phi,\psi+\delta_v)]\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - \Big(\int_{\mathbb{X}} [k(\phi + \delta_u, \psi) - k(\phi, \psi)]\, \rho(u|\phi)\, \mathrm{d}u + \sum_{x \in \phi} [k(\phi - \delta_x, \psi) - k(\phi, \psi)] \Big) \Big] \rho(v|\psi)\, \mathrm{d}v \\
&+ \sum_{y \in \psi} \Big[\Big(\int_{\mathbb{X}} \big[k(\phi + \delta_u, \psi - \delta_y) - k(\phi, \psi - \delta_y)\big] \rho(u|\phi)\, \mathrm{d}u + \sum_{x \in \phi} \big[k(\phi - \delta_x, \psi - \delta_y) - k(\phi, \psi - \delta_y)\big] \Big) \\
&\qquad - \Big(\int_{\mathbb{X}} [k(\phi + \delta_u, \psi) - k(\phi, \psi)]\, \rho(u|\phi)\, \mathrm{d}u + \sum_{x \in \phi} [k(\phi - \delta_x, \psi) - k(\phi, \psi)] \Big) \Big] \\
&= \int_{\mathbb{X}} \int_{\mathbb{X}} \Big[k(\phi + \delta_u, \psi + \delta_v) - k(\phi, \psi + \delta_v) - k(\phi + \delta_u, \psi) + k(\phi, \psi) \Big] \rho(u|\phi)\, \rho(v|\psi)\, \mathrm{d}u\, \mathrm{d}v \\
&+ \int_{\mathbb{X}} \Big[\sum_{x \in \phi} \big[k(\phi - \delta_x, \psi + \delta_v) - k(\phi - \delta_x, \psi)\big] - |\phi| \cdot \big[k(\phi, \psi + \delta_v) - k(\phi, \psi)\big] \Big] \rho(v|\psi)\, \mathrm{d}v \\
&+ \int_{\mathbb{X}} \Big[\sum_{y \in \psi} \big[k(\phi + \delta_u, \psi - \delta_y) - k(\phi, \psi - \delta_y)\big] - |\psi| \cdot \big[k(\phi + \delta_u, \psi) - k(\phi, \psi)\big] \Big] \rho(u|\phi)\, \mathrm{d}u \\
&+ \Big[\sum_{x \in \phi} \sum_{y \in \psi} k(\phi - \delta_x, \psi - \delta_y) - |\phi| \cdot \sum_{y \in \psi} k(\phi, \psi - \delta_y) - |\psi| \cdot \sum_{x \in \phi} k(\phi - \delta_x, \psi) + |\phi| \cdot |\psi| \cdot k(\phi, \psi) \Big],
\end{aligned}
$$

which recovers the expression in Eq. (6.8). This concludes the proof. □

To evaluate $\kappa_\rho(\phi, \psi)$ for a pair of configurations (ϕ, ψ) using Eq. (6.8), we need to compute one double integral and two single integrals over the domain $\mathbb{X} \subseteq \mathbb{R}^d$ as well as summations over the points in both ϕ and ψ.[2] Evaluating these integrals could require numerical integration techniques, but observe that we have reduced the problem of evaluating a normalization constant for a distribution on $\mathcal{N}_\mathbb{X}$ (an infinite-dimensional integral) to a finite-dimensional one. For most applications, d is small ($d = 1$ for temporal point processes and typically $d = 2$ for spatial point processes), and standard numerical quadrature methods should suffice.

While Theorem 6.2.2 implies that $\mathbb{D}_{\mathscr{H}_k}(\eta \,\|\, \rho) = 0$ for $\eta \equiv \rho$, we note that for non-Poisson processes, $\mathbb{D}_{\mathscr{H}_k}(\eta \,\|\, \rho) = 0$ may not be sufficient to guarantee that $\eta \equiv \rho$. This is due to the fact that while the Mecke formula fully characterizes a Poisson process, the GNZ formula (which was crucial in establishing our Stein operator) provides only a necessary condition for a point process to have a specific Papangelou conditional intensity.

[2] For concreteness, we provide example Python code for implementing Eq. (6.8) in Appendix B.

6.3.2 Goodness-of-Fit Testing via KSD

We now apply the kernelized Stein discrepancy measure of Definition 6.3.2 to construct a goodness-of-fit test for general (finite) point processes, including those with computationally intractable intensity functions.

Suppose we observe samples $\{\mathcal{X}_i\}_{i=1}^m$ from a point process with *unknown* Papangelou conditional intensity η, where each $\mathcal{X}_i := \{x_k\}_{k=1}^{n_i} \subseteq \mathbb{X}$ is a collection of points in $\mathbb{X}$ (note that the cardinalities n_i would vary). Given a statistical model which posits that the observed samples arose from a point process with (known) Papangelou conditional intensity ρ, we would like to quantify the 'goodness-of-fit' of the model ρ to the data $\{\mathcal{X}_i\}_{i=1}^m$. (Often we have only a single realization $\mathcal{X}$ of a point process, rather than many realizations of the process. In this case, it suffices to partition the space into a collection of blocks with equal volume, and treat the restriction of $\mathcal{X}$ to block i as the i-th realization $\mathcal{X}_i$.)

Formally, we perform the hypothesis test $H_0 : \rho = \eta$ vs. $H_1 : \rho \neq \eta$ using kernelized Stein discrepancy (KSD). For convenience, we omit the dependency on $\mathscr{H}_k$ and denote $\mathbb{S}(\eta \,\|\, \rho) := \mathbb{D}^2_{\mathscr{H}_k}(\eta \,\|\, \rho)$. Given observed samples $\{\mathcal{X}_i\}_{i=1}^m$ from a point process with (unknown) Papangelou conditional intensity η, by Eq. (6.8) we can estimate $\mathbb{S}(\eta \,\|\, \rho)$ via a U-statistic (Hoeffding, 1948) which gives a minimum-variance unbiased estimator:

$$\widehat{\mathbb{S}}(\eta \,\|\, \rho) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \kappa_\rho(\mathcal{X}_i, \mathcal{X}_j), \tag{6.9}$$

where the expression for $\kappa_\rho(\phi, \psi)$ is shown in Eq. (6.8).[3] By standard asymptotic results on U-statistics (analogous to Theorem 3.3.1), the U-statistic $\widehat{\mathbb{S}}(\eta \,\|\, \rho)$ is asymptotically normally distributed under the alternative hypothesis $H_1 : \rho \neq \eta$:

$$\sqrt{m}\,(\widehat{\mathbb{S}}(\eta \,\|\, \rho) - \mathbb{S}(\eta \,\|\, \rho)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 > 0$, but becomes degenerate under the null hypothesis $H_0 : \rho = \eta$.

Since the asymptotic distribution of $\widehat{\mathbb{S}}(\eta \,\|\, \rho)$ under H_0 is not available in closed-form, we follow Liu et al. (2016) and adopt the generalized bootstrap method for degenerate

[3] When evaluating Eq. (6.8), recall that $\phi + \delta_x$ and $\phi - \delta_x$ are equivalent to $\phi \cup \{x\}$ and $\phi \setminus \{x\}$, respectively.

U-statistics (Arcones and Gine, 1992; Huskova and Janssen, 1993) to approximate the distribution. To obtain a bootstrap sample, we draw random multinomial weights $w_1, \ldots, w_m \sim \text{Mult}(m; 1/m, \ldots, 1/m)$, set $\widetilde{w}_i = (w_i - 1)/m$, and compute

$$\widehat{\mathbb{S}}^*(\eta \,\|\, \rho) = \sum_{i=1}^{m} \sum_{j \neq i}^{m} \widetilde{w}_i \widetilde{w}_j \kappa_p(\mathcal{X}_i, \mathcal{X}_j). \tag{6.10}$$

Upon repeating this procedure $\widetilde{m}$ times, we calculate the critical value of the test by taking the $(1-\alpha)$-th quantile of the bootstrapped statistics $\{\widehat{\mathbb{S}}^*_b\}_{b=1}^{\widetilde{m}}$. We reject the null hypothesis H_0 if $\widehat{\mathbb{S}}(\eta \,\|\, \rho) > \gamma_{1-\alpha}$. The overall goodness-of-fit testing procedure is summarized in Algorithm 2.

As noted at the end of Section 6.3.1, $\mathbb{S}(\eta \,\|\, \rho) = 0$ may be insufficient to guarantee that $\eta \equiv \rho$. Thus, the KSD goodness-of-fit test may fail to reject H_0 even when the observed data arose from a point process with a Papangelou conditional intensity different from that specified by the null model, yielding Type-II errors. To the best of our knowledge, no necessary-and-sufficient condition for characterizing general (non-Poisson) point processes is known in the literature, and existing approaches (Baddeley and Turner, 2005; Coeurjolly and Lavancier, 2013) also only guarantee Type-I error control, and suffer from the same loss of power.

Algorithm 2 KSD goodness-of-fit test for point processes

1: **Input:** Papangelou conditional intensity ρ, point configurations $\{\mathcal{X}_i\}_{i=1}^m \sim \eta$, kernel function $k(\cdot,\cdot)$ on $\mathcal{N}_{\mathbb{X}}$, bootstrap sample size $\widetilde{m}$, significance level α.

2: **Objective:** Test $H_0 : \rho = \eta$ vs. $H_1 : \rho \neq \eta$.

3: Compute test statistic $\widehat{\mathbb{S}}(\eta \,\|\, \rho)$ via

$$\widehat{\mathbb{S}}(\eta \,\|\, \rho) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \kappa_\rho(\mathcal{X}_i, \mathcal{X}_j),$$

where

$$\begin{aligned}
\kappa_\rho(\phi,\psi) = &\int_{\mathbb{X}} \int_{\mathbb{X}} \Big[k(\phi \cup \{u\}, \psi \cup \{v\}) - k(\phi, \psi \cup \{v\}) - k(\phi \cup \{u\}, \psi) + k(\phi,\psi) \Big] \rho(u|\phi)\, \rho(v|\psi)\, \mathrm{d}u\, \mathrm{d}v \\
&+ \int_{\mathbb{X}} \Big[\sum_{x \in \phi} \big[k(\phi \backslash \{x\}, \psi \cup \{v\}) - k(\phi \backslash \{x\}, \psi) \big] - |\phi| \cdot \big[k(\phi, \psi \cup \{v\}) - k(\phi,\psi) \big] \Big] \rho(v|\psi)\, \mathrm{d}v \\
&+ \int_{\mathbb{X}} \Big[\sum_{y \in \psi} \big[k(\phi \cup \{u\}, \psi \backslash \{y\}) - k(\phi, \psi \backslash \{y\}) \big] - |\psi| \cdot \big[k(\phi \cup \{u\}, \psi) - k(\phi,\psi) \big] \Big] \rho(u|\phi)\, \mathrm{d}u \\
&+ \Big[\sum_{x \in \phi} \sum_{y \in \psi} k(\phi \backslash \{x\}, \psi \backslash \{y\}) - |\phi| \cdot \sum_{y \in \psi} k(\phi, \psi \backslash \{y\}) - |\psi| \cdot \sum_{x \in \phi} k(\phi \backslash \{x\}, \psi) + |\phi| \cdot |\psi| \cdot k(\phi,\psi) \Big].
\end{aligned}$$

4: **for** $b = 1, \ldots, \widetilde{m}$ **do**

5: Draw random multinomial weights $w_1, \ldots, w_m \sim \mathrm{Mult}(m; {}^1\!/_m, \ldots, {}^1\!/_m)$; set $\widetilde{w}_i = (w_i - 1)/m$.

6: Compute bootstrap test statistic $\widehat{\mathbb{S}}_b^*$ via

$$\widehat{\mathbb{S}}_b^*(\eta \,\|\, \rho) = \sum_{i=1}^{m} \sum_{j \neq i}^{m} \widetilde{w}_i \widetilde{w}_j \kappa_p(\mathcal{X}_i, \mathcal{X}_j).$$

7: Compute critical value $\gamma_{1-\alpha}$ by taking the $(1-\alpha)$-th quantile of the bootstrapped statistics $\{\widehat{\mathbb{S}}_b^*\}_{b=1}^{\widetilde{m}}$.

8: **Output:** Reject H_0 if $\widehat{\mathbb{S}}(\eta \,\|\, \rho) > \gamma_{1-\alpha}$, otherwise do not reject H_0.

Computational complexity. Calculating the test statistic $\widehat{\mathbb{S}}(\eta \,\|\, \rho)$ in Eq. (6.9) requires $\mathcal{O}(m^2)$ evaluations of $\kappa_\rho(\mathcal{X}_i, \mathcal{X}_j)$, where m is the number of data samples. Once the kernel matrix $[\kappa_\rho(\mathcal{X}_i, \mathcal{X}_j)]_{i,j=1}^m$ is cached, the bootstrapping procedure takes $\mathcal{O}(\widetilde{m} \cdot m^2)$ time, where $\widetilde{m}$ is the number of bootstrap samples.

To be more precise, recall that a sample $\mathcal{X}_i$ consists of $n_i := |\mathcal{X}_i|$ points in $\mathbb{X}$. Evaluating $\kappa_\rho(\mathcal{X}_i, \mathcal{X}_j)$ for each pair of samples $(\mathcal{X}_i, \mathcal{X}_j)$ using Eq. (6.8) requires numerical integration. Assuming q quadrature points per dimension, the time complexity for a single evaluation of $\kappa_\rho(\mathcal{X}_i, \mathcal{X}_j)$ is given by

$$\mathcal{O}((q^{2d} + 2\,q^d \bar{n} + |\bar{n}|^2) \cdot t_k + 2\,q^d\, t_\rho) = \mathcal{O}((q^d + |\bar{n}|)^2 \cdot \bar{n}^2).$$

Here, $\bar{n}$ is the average cardinality of the observed samples, t_k is the time required to evaluate the kernel function $k(\cdot, \cdot)$ for a pair of samples with size $\bar{n}$, and t_ρ is the time needed for a single evaluation of the Papangelou conditonal intensity (typically, $t_k, t_\rho = \mathcal{O}(\bar{n}^2)$ in the worst case). Putting everything together, the overall time complexity of Algorithm 1 is

$$\mathcal{O}(m^2 \cdot (q^d + |\bar{n}|)^2 \cdot \bar{n}^2 + \widetilde{m} \cdot m^2).$$

Note that when d is large, one could apply Monte Carlo integration in lieu of numerical quadrature to avoid the curse of dimensionality, and the q^d term would be replaced by c, the number of Monte Carlo points.

6.3.3 Kernel Functions for Point Processes

Our theoretical development so far hold generally for any positive definite kernel on the space of finite counting measures $\mathcal{N}_\mathbb{X}$. There has been work on *set kernels* or *multi-instance kernels* (Gärtner et al., 2002), where the similarity of two sets is measured by their average pairwise point similarities, as well as kernels which make parametric assumptions on the distributions of the points (Kondor and Jebara, 2003; Bach, 2008; Carrière et al., 2017).

We argue that a proper kernel function $k(\mathcal{X}, \mathcal{Y})$ between two point configurations $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{X}$ should capture their similarities with regard to their *extrinsic* and/or *intrinsic*

characteristics as needed. Extrinsic characteristics refer to inhomogeneities in intensity, resulting in different expected counts for different point processes in the same parts of the $\mathbb{X}$-space. Intrinsic characteristics pertain to point interactions within a point process—*i.e.*, whether the points exhibit attraction or repulsiveness. Any prior knowledge regarding the nature of deviations from the null model could accordingly be incorporated into the kernel function. One simple approach would be to map each point configuration into a feature-vector, with components including *e.g.*, the number of points in different regions of the space, the number of points within some distance r of each other, the average distance from a point to their k-th nearest neighbor, etc.

As a flexible nonparametric alternative that takes both extrinsic and intrinsic features into consideration, we propose to use the maximum mean discrepancy (MMD) between two counting measures to define a p.d. kernel. Specifically, we have the following:

Proposition 6.3.1. *Given a positive definite kernel $k_{\mathbb{X}}(\cdot,\cdot)$ on the ground space $\mathbb{X}$, define the $\mathcal{M}$-*kernel*:*

$$k_{\mathcal{M}}(\phi,\psi) := \exp\{-\widehat{d^2}(\phi,\psi)\}, \tag{6.11}$$

where $\widehat{d^2}(\phi,\psi)$ denotes the V-statistic estimate of the squared-MMD between configurations $\phi,\psi \in \mathcal{N}_{\mathbb{X}}$:

$$\widehat{d^2}(\phi,\psi) := \frac{1}{|\phi|^2}\sum_{x\in\phi}\sum_{x'\in\phi} k_{\mathbb{X}}(x,x') + \frac{1}{|\psi|^2}\sum_{y\in\psi}\sum_{y'\in\psi} k_{\mathbb{X}}(y,y') - \frac{2}{|\phi|\cdot|\psi|}\sum_{x\in\phi}\sum_{y\in\psi} k_{\mathbb{X}}(x,y). \tag{6.12}$$

Then, $k_{\mathcal{M}}(\cdot,\cdot)$ is a positive definite kernel on $\mathcal{N}_{\mathbb{X}}$.[4]

The proof of Proposition 6.3.1 utilizes the following result of Schoenberg (1938):

Lemma 6.3.2 (Schoenberg, 1938)**.** *The function*

$$k(x,y) := \exp\left\{-\frac{f(x,y)}{\ell}\right\}$$

[4]If either ϕ or ψ is an empty configuration, we define $k(\phi,\psi) = 1$ if both are empty and $k(\phi,\psi) = 0$ otherwise.

defined on a domain $\mathcal{D}$ is a positive definite kernel for all $\ell > 0$ if and only if f is a conditionally negative definite *function, i.e., $\sum_{i,j=1}^{n} c_i c_j f(x_i, x_j) \leq 0$ for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{D}$, and $c_1, \ldots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} c_i = 0$.*

Proof of Proposition 6.3.1. Denote

$$\xi(\phi, \psi) := \sum_{x \in \phi} \sum_{y \in \psi} k_{\mathbb{X}}(x, y). \tag{6.13}$$

By Proposition 3.1 of Gärtner et al. (2002), $\xi(\cdot, \cdot)$ is a p.d. kernel on $\mathcal{N}_{\mathbb{X}}$ if $k_{\mathbb{X}}$ is a p.d. kernel on $\mathbb{X}$. By Schoenberg (1938) (Lemma 6.3.2 in the Appendix), to show that Eq. (6.11) defines a p.d. kernel, it suffices to show that $\widehat{d^2}$ is a conditionally negative-definite function. To this end, observe that for any $n \in \mathbb{N}$, $\phi_1, \ldots, \phi_n \in \mathcal{N}_{\mathbb{X}}$, and $c_1, \ldots, c_n \in \mathbb{R}$ satisfying $\sum_{i=1}^{n} c_i = 0$, we have

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \widehat{d^2}(\phi_i, \phi_j) = & \frac{1}{|\phi|^2} \left(\sum_{i=1}^{n} c_i \sum_{x \in \phi_i} \sum_{x' \in \phi_i} k_{\mathbb{X}}(x, x') \right) \left(\sum_{j=1}^{n} c_j \right) \\
& + \frac{1}{|\psi|^2} \left(\sum_{i=1}^{n} c_i \right) \left(\sum_{j=1}^{n} c_j \sum_{y \in \phi_j} \sum_{y' \in \phi_j} k_{\mathbb{X}}(y, y') \right) \\
& - \frac{2}{|\phi_i| \cdot |\phi_j|} \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \xi(\phi_i, \phi_j) \\
= & - \frac{2}{|\phi_i| \cdot |\phi_j|} \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \xi(\phi_i, \phi_j) \\
\leq & \, 0 \,,
\end{aligned}$$

where we used the fact that $\xi(\cdot, \cdot)$ is a p.d. kernel on $\mathcal{N}_{\mathbb{X}}$. This concludes the proof. □

6.4 Related Work

Classical diagnostic measures for point processes have largely been restricted to temporal point processes. For spatial point processes, traditional approaches (Diggle, 2003) primarily rely on heuristic summary statistics (*e.g.*, the 'K-function' of Ripley (1976)) to test for specific properties of the data, such as complete randomness or clustering.

Related to our work, and also motivated by the GNZ formula, Baddeley et al. (2005) defined the *h-weighted residual measure* for a parametric model $\widehat{\rho}$ fitted to an observed configuration ϕ on a bounded domain $B \subseteq \mathbb{X}$:

$$\gamma(B, h, \widehat{\rho}) := \sum_{x \in \phi \cap B} h(x, \phi \backslash \{x\}) - \int_B h(u, \phi)\, \widehat{\rho}(u|\phi)\, \mathrm{d}u,$$

where h is a user-specified weight function. Informally, our proposed KSD goodness-of-fit test statistic could be viewed as a *kernelization* of the h-weighted residuals, where we take the supremum over all test functions h in an RKHS. In doing so, we obtain a parsimonious and more powerful test capturing various aspects of the model intensity that would have been difficult for any specific h to fully cover. In addition, the KSD test allows users the flexibility to emphasize specific aspects of interest through the design of the kernel function.

6.5 Empirical Evaluation

We apply the kernelized Stein discrepancy (KSD) test to the point process models described in Section 2.2.2.[5] We also compare with a test based on the maximum mean discrepancy (MMD) (*cf.* Section 3.2), which draws samples from the null model, and performs a two-sample test between the drawn samples and the observed data. Note that here we are computing the MMD test statistic between two *collections* of *point configurations* in $\mathcal{N}_{\mathbb{X}}$, as opposed to Eq. (6.12) which estimates the MMD between two sets of *points* in $\mathbb{X}$. Given samples $\{\mathcal{X}_i\}_{i=1}^m, \{\mathcal{Y}_j\}_{j=1}^m$ from two point processes ρ and η, we compute the U-statistic estimate of $\mathrm{MMD}^2(\rho, \eta)$:

$$\widehat{\mathrm{MMD}^2}(\rho, \eta) := \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathcal{X}_i, \mathcal{X}_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathcal{Y}_i, \mathcal{Y}_j) - \frac{2}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{n} k(\mathcal{X}_i, \mathcal{Y}_j).$$

The critical value of the MMD test is calculated by bootstrapping on the aggregated data.

[5] Code for the experiments is available at https://github.com/jiaseny/stein-papangelou.

Setup. We adopt a similar experiment setup as in Chapter 5. Denote the Papangelou conditional intensities for the null and the alternative point process models by ρ and η, respectively. For KSD, we draw m *i.i.d.* samples (point configurations) from η; for MMD, we draw m samples from η and another m samples from ρ. For the kernel function $k(\cdot,\cdot)$ on $\mathcal{N}_{\mathbb{X}}$ (used in both KSD and MMD), we utilize the $\mathcal{M}$-kernel defined via Eqs. (6.11) and (6.12), where the *ground kernel* $k_{\mathbb{X}}(\cdot,\cdot)$ in Eq. (6.12) is set to a Gaussian RBF kernel. To ensure fair comparison, we set the bandwidth of the RBF kernel for both KSD and MMD to the median pairwise distance (Gretton et al., 2012) of the aggregated points in the samples drawn from η. We use $\widetilde{m} = 10,000$ bootstrap samples for both methods.

For each model, we choose a single parameter, fix its value for the null model ρ, and draw samples for η under different values of that parameter. For each value of the chosen parameter and sample size m, we conduct 500 independent trials. In each trial, we flip a fair coin to decide whether the alternative model η will be set to the same as ρ or with a different value of the chosen parameter (in the former case, the null hypothesis $H_0 : \rho = \eta$ should not be rejected, and in the latter case it should be). We conduct the hypothesis test $H_0 : \rho = \eta$ vs. $H_1 : \rho \neq \eta$ under significance level $\alpha = 0.01$, and evaluate the performance of KSD and MMD in terms of their false-positive rate (FPR; Type-I error) and false-negative rate (FNR; Type-II error).

Poisson process. We consider a Poisson process on the unit-square $[0,1]^2$ with intensity function $\lambda(x) = \gamma + \varepsilon \sin(2\pi(x+y))$, where γ is a base-rate, and ε represents the perturbation magnitude. We fix $\gamma = 50$ throughout, vary the perturbation magnitude ε, and test the hypotheses $H_0 : \varepsilon = 0$ vs. $H_1 : \varepsilon \neq 0$.

Hawkes process. We consider a Hawkes process on $[0,1]$ with intensity function given in Eq. (2.3) and set $g(t) = \beta e^{-t/\tau}$. We fix $\gamma = 20$ and $\beta = 2$ throughout, vary the time-scale parameter τ, and test the hypotheses $H_0 : \tau = 0.1$ vs. $H_1 : \tau \neq 0.1$. To simulate from a Hawkes process, we employ the thinning algorithm of Ogata (1981).

Strauss process. We consider Strauss processes on $[0,1]^d$ ($d = 1$ or 2) with conditional intensity given in Eq. (2.7). We fix $\beta = 20$ and $\gamma = 0.8$ ($d = 1$) or 0.9 ($d = 2$), vary the interaction radius r and test the hypotheses $H_0 : r = r_0$ vs. $H_1 : r \neq r_0$ with $r_0 = 0.2$ or

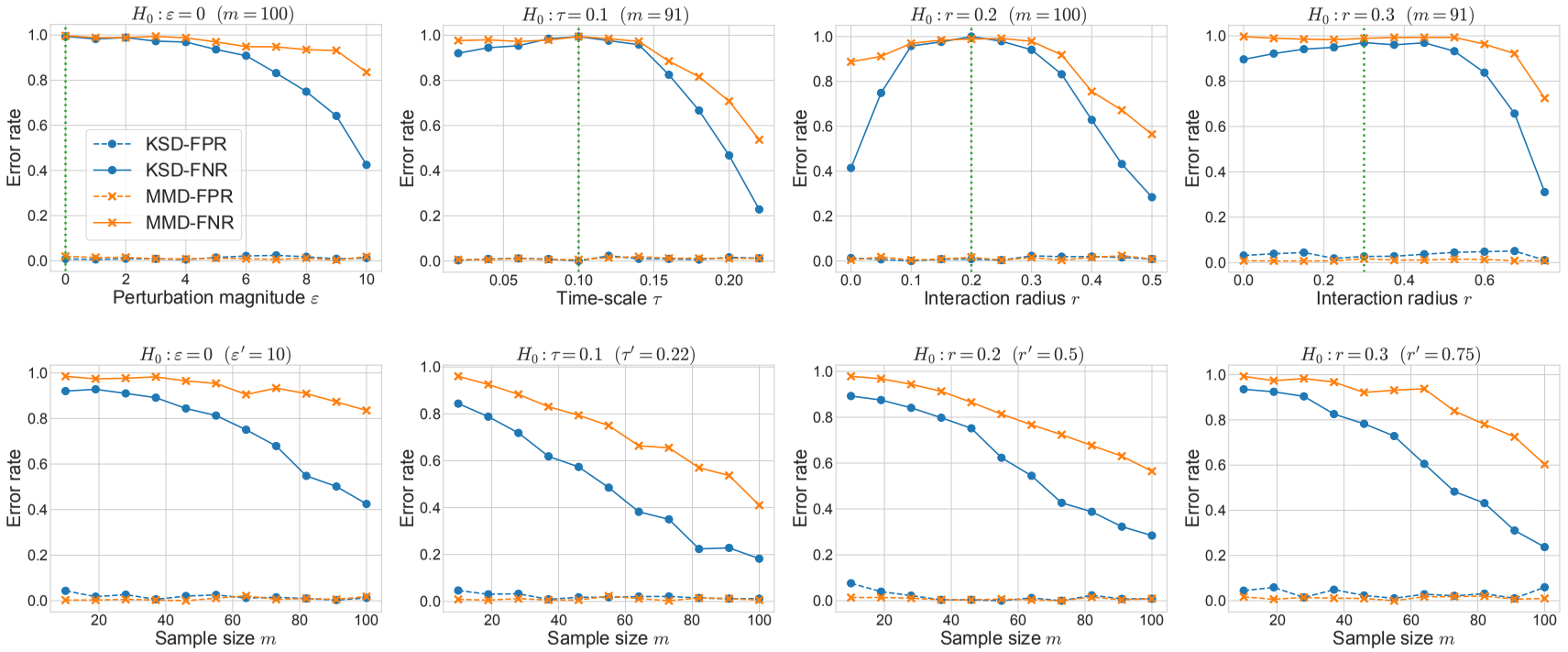
0.3. To simulate from a 1-D Strauss process, we apply rejection sampling to realizations of a Poisson process with intensity β. To simulate from a 2-D Strauss process, we use the MCMC sampler provided in the R package `spatstat` (Baddeley and Turner, 2005; Baddeley et al., 2015).

Results. In Figure 6.1, the top row plots the testing error rate vs. different values of the parameter we chose to vary for η, under a given sample size. The bottom row plots the error rate vs. sample size for a specific value of the chosen parameter. We observe that both methods generally maintain a false-positive rate (Type-I error) around the significance level, while KSD consistently achieves lower false-negative rate (Type-II error) than MMD across different parameter settings as well as sample sizes.[6] This indicates that KSD, by utilizing information from the Papangelou conditional intensity ρ of the null model, gives rise to a more powerful test. We emphasize that the MMD two-sample test requires generating exact samples from the null model, which could be computationally costly or intractable. Finally, we note that the statistical power of both methods could be improved by using more sophisticated constructions of kernel functions on the space of counting measures, which we leave for future work.

6.6 Summary

We have introduced a general Stein operator based on the Papangelou conditional intensity for point processes which can be evaluated even when the intensity function contains an intractable normalization constant. Using the proposed Stein operator, we have developed a kernelized Stein discrepancy test for measuring the goodness-of-fit of a point process model. We have applied the proposed test to several point process models, and showed that it outperforms a two-sample test based on the maximum mean discrepancy, which assumes the availability of exact samples from the null model.

[6] In Figure 6.1d, the Type-I error for KSD appears slightly higher than the nominal significance level 0.01. We found that this was due to numerical quadrature error involved in evaluating Eq. (6.8) under limited computational budget (since the double-integral over $\mathbb{X}$ is now four-dimensional). This issue could be alleviated using Monte Carlo integration techniques, which shall be investigated in future work.



(a) Poisson process (b) Hawkes process (c) Strauss process ($d = 1$) (d) Strauss process ($d = 2$)

Figure 6.1.: *Top row:* KSD and MMD testing error rate vs. varying parameter (the vertical dotted lines indicate the value of the parameter under H_0). *Bottom row:* KSD and MMD testing error rate vs. sample size.

7. CONCLUSION AND FUTURE DIRECTIONS

In this dissertation, we developed statistical models and model-criticism techniques for learning from data exhibiting relational, temporal, and/or spatial dependencies. On the one hand, we proposed latent space point process models to decouple the influences of homophily and reciprocity from dynamic interactions in communication networks. On the other hand, we developed goodness-of-fit tests for discrete distributions (including network models) and point processes involving intractable normalization constants, provide the first generally applicable and computationally feasible model-criticism approaches under those circumstances. In this chapter, we summarize the contributions of this dissertation, and outline several avenues for future research.

7.1 Summary of Contributions

The contributions of this dissertation fall into the following aspects:

Theoretical

- In Chapter 5, we proposed:
  - A difference Stein operator for discrete spaces;
  - A general characterization of Stein operators that encompasses both discrete and continuous distributions, with a recipe for constructing new Stein operators; and
  - A kernelized discrete Stein discrepancy measure for discrete distributions, including those containing intractable normalization constants.
- In Chapter 6, we constructed:
  - A Stein–Papangelou operator for general point processes based on the Papangelou conditional intensity function;

  - A positive-definite kernel on the space of point configurations using the maximum mean discrepancy; and
  - A kernelized Stein discrepancy measure for general point processes, including those with intractable intensity functions.

Methodological

- In Chapter 4, we developed:
  - A collection of latent space point process models, including a Poisson process latent space model, two single-latent space Hawkes process models, and a dual-latent space model, to capture and decouple the influences of homophily and reciprocity in temporal interactions; and
  - Methodology to evaluate the proposed models, including static and dynamic link prediction tasks, as well as exploration of the learned node embeddings.
- In Chapters 5 and 6, we developed:
  - A nonparametric goodness-of-fit test for unnormalized discrete distributions (including network models) based on the kernelized discrete Stein discrepancy; and
  - A nonparametric Stein–Papangelou goodness-of-fit test for general point processes (including those with intractable intensity functions).

  These goodness-of-fit tests provide the first generally applicable and computationally feasible model-criticism approaches under the aforementioned circumstances.

Empirical

- In Chapter 4, we
  - Evaluated the utility of our models both quantitatively and qualitatively on three real-world datasets; and
  - Showed that incorporating both homophily and reciprocal latent spaces improves predictive performance and gives rise to interpretable embeddings.
- In Chapter 5, we

  - Applied our kernelized discrete Stein discrepancy goodness-of-fit test to the Ising model, the Bernoulli restricted Boltzmann machine, and the exponential random graph model; and
  - Demonstrated that the proposed test typically outperforms a two-sample test based on the maximum mean discrepancy in terms of power while maintaining control on false-positive rate.

- In Chapter 6, we
  - Applied our Stein–Papangelou goodness-of-fit test to the Poisson process, the Hawkes process, and the Strauss process; and
  - Demonstrated that the proposed test outperforms a two-sample test based on the maximum mean discrepancy in terms of power while maintaining control on false-positive rate.

7.2 Future Directions

7.2.1 Stein's Method for Model Criticism and Bayesian Inference

The complexity of modern statistical and machine learning models typically culminates in likelihood functions containing intractable normalization constants, which preclude the use of conventional model criticism techniques and bring about challenges in performing inference for those models. As we have examined in Chapters 5 and 6. Stein's method provides a principled framework for developing goodness-of-fit tests and model criticism techniques under this scenario. Interestingly, the concept of Stein discrepancy can also be applied to develop a novel approach for approximate Bayesian inference in complex probabilistic models, providing an efficient and accurate alternative to conventional MCMC and variational methods.

Flexible, Interpretable, and Scalable Techniques for Model Criticism

Much progress remains to be made toward enhancing the flexibility, effectiveness, and scalability of existing model criticism techniques.

First, current Stein discrepancy tests could not be applied directly to models with latent variables, as their likelihood function involves marginalizing over these variables. However, many widely used models, such as probabilistic topic models (Blei et al., 2003), stochastic blockmodels (*cf.* Chapter 2), and the latent space models we studied in Chapter 4, fall into this category.

Second, current tests do not exploit specific *structure* in the data or model. For kernelized Stein discrepancy tests, their statistical power depends heavily on the choices of the Stein operator and the kernel function. For model distributions with additional structure (*e.g.*, conditional independencies encoded in a graphical model), one could design appropriate Stein operators and factorized kernel functions that exploit such structure to establish more powerful tests. Special care also needs to be taken when the distribution is supported on a highly constrained domain or manifold.

Third, variants of the kernelized Stein discrepancy statistic could be used to construct *interpretable features* (Jitkrittum et al., 2017, 2018) that reveal aspects of the data which the current model fails to capture, pointing out ways to improve the model fit. Future work along this direction is important for the development of interpretable diagnostic and criticism techniques for complex models.

Fourth, kernel-based hypothesis tests require the computation of a Gram matrix with cost quadratic in the number of data samples, which could become prohibitive for large datasets. While this computation is amenable to parallelization, more efficient tests that run in near-linear time could be obtained by *sketching* the kernel matrix (via *e.g.*, element-wise sampling (Achlioptas and Mcsherry, 2007), the Nyström method (Drineas and Mahoney, 2005), or random Fourier features (Rahimi and Recht, 2008)). By exploiting the special structure of positive semi-definite matrices, one could further develop improved approximation guarantees for the kernel-based test statistic.

Finally, practical considerations such as data censoring in survival analysis and differential privacy constraints present new challenges in the design of effective model criticism techniques. The recent work of Fernández and Gretton (2019) takes a step toward this direction.

Stein's Method for Approximate Bayesian Inference

Viewed as a distance measure between data samples and a model distribution, the kernelized Stein discrepancy statistic also provides a method to sample from a potentially intractable Bayesian posterior distribution (such as those arising from Bayesian neural networks), presenting an efficient and accurate alternative to conventional MCMC and variational methods. Specifically, one starts by randomly initializing a set of particles, and then iteratively refines these particles to minimize the discrepancy between their empirical distribution and the target distribution. when the particles are added in a greedy fashion, the optimization procedure could be carried out using an instance of the Frank–Wolfe algorithm known as kernel herding (Chen et al., 2010, 2018). A related idea is the *Stein variational gradient descent* algorithm of Liu and Wang (2016), which iteratively minimizes the KL divergence between the particles and the target distribution. In future work, one could explore similar algorithms for sampling from network models and point processes with intractable normalization constants, as computationally efficient alternatives to MCMC methods.

7.2.2 Invariance Principles for Networks and Point Processes

The study of networks models and point provides hinges upon fundamental symmetry and invariance principles in probability theory. Understanding the properties and limitations implied by these invariance principles is crucial for gaining a deeper understanding of these models, and also provides valuable insights for the development of new sampling, learning, and inference algorithms. While distinct in terms of their application domains, networks and point processes share many common characteristics

in terms of the invariance principles governing them. In particular, Caron and Fox (2017) introduced a point process representation of sparse graphs using exchangeable random measures; an immediate direction of future work is to construct Stein operators and goodness-of-tests for sparse graphs under this representation by building upon our development in Chapter 6. Other related topics include kernel methods for permutations (Jiao and Vert, 2015) and Fourier analysis on the symmetric group (Clausen and Baum, 1993; Kondor, 2008).

REFERENCES

Dimitris Achlioptas and Frank Mcsherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2), 2007. (Cited on page 109.)

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008. (Cited on pages 12 and 15.)

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016. (Cited on pages 62, 71, and 78.)

T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954. (Cited on page 3.)

Miguel A. Arcones and Evarist Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674, 1992. (Cited on pages 32, 37, 75, and 97.)

Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L. Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18:226:1–226:92, 2018. (Cited on page 16.)

G. J. Babu and E. D. Feigelson. *Astrostatistics*. Chapman and Hall, 1996. (Cited on page 1.)

Francis R. Bach. Graph kernels between point clouds. In *Proceedings of the 25th International Conference on Machine Learning*, pages 25–32, 2008. (Cited on page 99.)

A. Baddeley, R. Turner, J. Mller, and M. Hazelton. Residual analysis for spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (5):617–666, 2005. (Cited on pages 23, 86, and 102.)

Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. (Cited on pages 97 and 104.)

Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London, 2015. (Cited on page 104.)

A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286: 509–512, 1999. (Cited on page 12.)

A. D. Barbour. Stein's method and Poisson process convergence. *Journal of Applied Probability*, 25:175–184, 1988. (Cited on pages 6, 34, 88, and 89.)

A.D. Barbour and T.C. Brown. Stein's method and point process approximation. *Stochastic Processes and their Applications*, 43(1):9 – 31, 1992. (Cited on pages 6, 88, 89, and 90.)

A.D. Barbour and L.H.Y. Chen. An introduction to stein's method, 2005. (Cited on page 33.)

Andrew D. Barbour and Louis H. Y. Chen. Stein's (magic) method. *arXiv:1411.1179*, 2014. (Cited on page 33.)

David M. Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1(1):203–232, 2014. (Cited on pages viii, 3, and 4.)

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. (Cited on page 109.)

Charles Blundell, Katherine A. Heller, and Jeffrey M. Beck. Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems 25*, 2012. (Cited on pages 17, 41, 45, and 56.)

Wacha Bounliphone, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton. A test of relative similarity for model selection in generative models. In *ICLR*, 2016. (Cited on page 39.)

George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. (Cited on page 3.)

George E. P. Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383–430, 1980. (Cited on page 3.)

George E P Box and Norman R Draper. *Empirical Model-building and Response Surface*. John Wiley & Sons, Inc., New York, NY, USA, 1986. (Cited on page 3.)

Guy Bresler and Dheeraj Nagaraj. Stein's method for stationary distributions of markov chains and application to Ising models. *arXiv:1712.05736*, 2017. (Cited on page 77.)

D. Brook. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51(3/4): 481–483, 1964. (Cited on page 64.)

Emery N Brown, Riccardo Barbieri, Valérie Ventura, Robert E Kass, and Loren M Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346, 2002. (Cited on page 86.)

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208, 1995. (Cited on pages 49 and 124.)

François Caron and Emily B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017. (Cited on page 111.)

Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning*, pages 664–673, 2017. (Cited on page 99.)

Sourav Chatterjee. A short survey of Stein's method. In *Proceedings of the ICM*, 2014. (Cited on page 33.)

Sourav Chatterjee and Persi Diaconis. Estimating and understanding exponential random graph models. *Ann. Statist.*, 41(5):2428–2461, 10 2013. (Cited on page 13.)

Louis H.Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein's Method*. Springer, 2011. (Cited on page 33.)

Wilson Ye Chen, Lester Mackey, Jackson Gorham, Francois-Xavier Briol, and Chris Oates. Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 844–853, 2018. (Cited on page 110.)

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2010. (Cited on page 110.)

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016. (Cited on pages 4, 5, 9, 25, 26, 32, 35, 37, 39, 61, 71, 77, 87, and 92.)

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems 28*, pages 1981–1989. 2015. (Cited on page 39.)

Michael Clausen and Ulrich Baum. Fast Fourier transforms for symmetric groups: Theory and implementation. *Mathematics of Computation*, 61(204):833–847, 1993. (Cited on page 111.)

Jean-François Coeurjolly and Frédéric Lavancier. Residuals and goodness-of-fit tests for stationary marked Gibbs point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):247–276, 2013. (Cited on pages 86 and 97.)

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I*. Springer-Verlag, second edition, 2003. (Cited on pages 20 and 23.)

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes (Vol. II)*. Springer, second edition, 2008. (Cited on pages 20, 23, 24, and 86.)

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1989–2007, 2018. (Cited on page 77.)

Laurent Decreusefond and Aurélien Vasseur. Stein's method and Papangelou intensity for Poisson or Cox process approximation. *arXiv:1807.02453*, 2018. (Cited on page 90.)

Persi Diaconis and Susan Holmes. *Stein's Method: Expository Lectures and Applications*, volume 46 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 2004. (Cited on page 33.)

Peter J. Diggle. *Statistical analysis of spatial point patterns*. Edward Arnold, 2003. (Cited on pages 1 and 101.)

Petros Drineas and Michael W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, 2005. (Cited on page 109.)

N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*, 2016. (Cited on pages 56 and 86.)

C. DuBois, C. Butts, and P. Smyth. Stochastic blockmodeling of relational event dynamics. In *AISTATS*, 2013. (Cited on page 56.)

D. Durante and D. Dunson. Nonparametric bayes dynamic modelling of relational data. *Biometrika*, 101(4):883, 2014. (Cited on page 16.)

P. Ekeh. *Social exchange theory: The two traditions*. Heinemann London, 1974. (Cited on page 41.)

P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae*, 6:290–297, 1959. (Cited on page 12.)

Y. Fan and C. R. Shelton. Learning continuous-time social network dynamics. In *UAI*, 2009. (Cited on page 16.)

M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. COEVOLVE: A joint point process model for information diffusion and network co-evolution. In *NIPS*, 2015. (Cited on page 56.)

Tamara Fernández and Arthur Gretton. A maximum-mean-discrepancy goodness-of-fit test for censored data. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019. (Cited on page 110.)

O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81:395:832–842, 1986. (Cited on pages 12, 61, and 80.)

W. Fu and E. Xing. Dynamic mixed membership blockmodel for evolving networks. In *ICML*, 2009. (Cited on page 16.)

Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alex J. Smola. Multi-instance kernels. In *Proceedings of the International Conference on Machine Learning*, pages 179–186, 2002. (Cited on pages 99 and 101.)

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013. (Cited on page 3.)

Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2 (2):129–233, 2010. (Cited on pages 12 and 15.)

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems (NIPS) 28*, 2015. (Cited on pages 4, 5, 26, 32, 34, 35, 60, 87, and 92.)

Jackson Gorham and Lester W. Mackey. Measuring sample quality with kernels. In *Proceedings of The 34th International Conference on Machine Learning*, pages 1292–1301, 2017. (Cited on pages 4, 36, and 75.)

Jackson Gorham, Andrew B. Duncan, Sebastian J. Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *arXiv:1611.06972*, 2016. (Cited on page 34.)

Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. 2008. (Cited on page 39.)

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1): 723–773, 2012. (Cited on pages 6, 9, 26, 29, 30, 31, 78, and 103.)

A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *KDD*, 2016. (Cited on pages 53 and 56.)

F. Guo, C. Blundell, H. Wallach, and K. Heller. The Bayesian echo chamber: Modeling social influence via linguistic accommodation. In *AISTATS*, 2015. (Cited on page 56.)

Mark S. Handcock, Garry Robins, Tom Snijders, Jim Moody, and Julian Besag. Assessing degeneracy in statistical models of social networks. *Journal of the American Statistical Association*, 76:33–50, 2003. (Cited on page 13.)

Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007. (Cited on page 15.)

Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, and Martina Morris. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (http://www.statnet.org), 2017. URL https://CRAN.R-project.org/package=ergm. R package version 3.8.0. (Cited on page 83.)

S. Hanneke and E. Xing. Discrete temporal models of social networks. *Electron. J. Stat.*, 4:585–605, 2010. (Cited on page 16.)

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971. (Cited on pages 7, 17, 19, and 88.)

X. He, T. Rekatsinas, J. Foulds, L. Getoor, and Y. Liu. HawkesTopic: A joint model for network inference and topic modeling from text-based cascades. In *ICML*, 2015. (Cited on pages 43 and 56.)

Creighton Heaukulani and Zoubin Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, pages I–275–I–283, 2013. (Cited on page 17.)

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. (Cited on pages 6, 61, and 79.)

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. (Cited on pages 37, 74, and 96.)

P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems 20*, 2008. (Cited on page 15.)

Peter D Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295, 2005. (Cited on page 15.)

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098, 2002. (Cited on pages 12, 14, 15, 17, 41, and 43.)

Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373): 33–50, 1981. (Cited on page 12.)

David R Hunter and Mark S Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006. (Cited on page 13.)

David R Hunter, Steven M Goodreau, and Mark S Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008a. (Cited on page 14.)

David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29, 2008b. (Cited on pages viii, 13, and 14.)

Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811–1823, 1993. (Cited on pages 37, 75, and 97.)

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005. (Cited on pages 62 and 77.)

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007. (Cited on pages 62, 65, and 78.)

Ernst Ising. *Beitrag zur Theorie des Ferro- und Paramagnetismus*. PhD thesis, 1924. (Cited on pages 61 and 79.)

T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade Poisson processes. In *KDD*, 2013. (Cited on page 41.)

Yunlong Jiao and Jean-Philippe Vert. The kendall and mallows kernels for permutations. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1935–1944, 2015. (Cited on page 111.)

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems 29*, pages 181–189. 2016. (Cited on page 39.)

Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems (NIPS) 30*. 2017. (Cited on pages 4, 32, 39, 61, 87, and 109.)

Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. Informative features for model comparison. In *Advances in Neural Information Processing Systems 31*, pages 808–819. 2018. (Cited on pages 39 and 109.)

C. Kemp, J. Tenenbaum, and T. Griffiths. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006. (Cited on page 56.)

J.F.C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, 1992. (Cited on page 20.)

B. Klimmt and Y. Yang. Introducing the enron corpus. In *CEAS*, 2004. (Cited on pages 17 and 49.)

A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933. (Cited on page 3.)

Imre Risi Kondor. *Group Theoretical Methods in Machine Learning*. PhD thesis, 2008. (Cited on page 111.)

Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning*, pages 361–368, 2003. (Cited on page 99.)

Pavel N. Krivitsky, Mark S. Handcock, Adrian E. Raftery, and Peter D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213, 2009. (Cited on page 15.)

Günter Last and Mathew Penrose. *Lectures on the Poisson Process*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2017. (Cited on pages 20 and 21.)

Christophe Ley and Yvik Swan. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18:14 pp., 2013. (Cited on page 66.)

Christophe Ley, Gesine Reinert, and Yvik Swan. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017. (Cited on pages 33, 69, and 77.)

Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1413–1421, 2014. (Cited on pages 2, 56, and 86.)

Scott Linderman, Christopher H Stock, and Ryan P Adams. A framework for studying synaptic plasticity with neural spike train data. In *Advances in Neural Information Processing Systems 27*, pages 2330–2338. 2014. (Cited on page 2.)

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NIPS) 29*, 2016. (Cited on page 110.)

Qiang Liu, Jason D. Lee, and Michael I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016. (Cited on pages 4, 5, 6, 9, 25, 26, 32, 35, 36, 37, 38, 39, 61, 71, 74, 75, 77, 78, 87, 92, 93, and 96.)

Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2009. (Cited on pages 62, 65, and 78.)

Abraham Martín del Campo, Sarah Cepeda, and Caroline Uhler. Exact goodness-of-fit testing for the Ising model. *Scandinavian Journal of Statistics*, 44(2):285–306, 2017. (Cited on page 77.)

M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. (Cited on pages 14 and 40.)

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. (Cited on page 56.)

K. Miller, T. Griffiths, and M. Jordan. Nonparametric latent feature models for link prediction. In *NIPS*, 2009. (Cited on page 16.)

B. Min, K. Goh, and A. Vazquez. Spreading dynamics following bursty human activity patterns. *Phys. Rev. E*, 83(3):036102, 2011. (Cited on page 41.)

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. (Cited on page 30.)

K. Nowicki and T. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087, 2001. (Cited on pages 12 and 15.)

Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017. (Cited on pages 4, 5, 32, 34, 35, and 61.)

Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981. (Cited on page 103.)

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988. (Cited on page 2.)

A. O'Hagan. HSSS model criticism. In *Highly Structured Stochastic Systems*, pages 422–444. Oxford University Press, 2003. (Cited on page 3.)

F. Papangelou. The conditional intensity of general point processes and an application to line processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 28 (3):207–226, 1974. (Cited on pages 6 and 23.)

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900. (Cited on page 3.)

B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online learning of social representations. In *KDD*, 2014. (Cited on page 56.)

Patrick O. Perry and Patrick J. Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75 (5):821–849, 2013. (Cited on pages 17 and 56.)

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. 2008. (Cited on page 109.)

Gesine Reinert and Nathan Ross. Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *arXiv:1712.05743*, 2017. (Cited on page 77.)

Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018. (Cited on pages 19 and 86.)

Alessandro Rinaldo, Stephen E. Fienberg, and Yi Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Statist.*, 3:446–484, 2009. (Cited on page 13.)

B. D. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):255–266, 1976. (Cited on pages 4, 87, and 101.)

B. D. Ripley and F. P. Kelly. Markov point processes. *Journal of the London Mathematical Society*, s2-15(1):188–192, 1977. (Cited on page 22.)

G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29: 192–215, 2007a. (Cited on page 13.)

Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p^*) models for social networks. *Social Networks*, 29(2):173 – 191, 2007b. (Cited on page 12.)

Nathan Ross. Fundamentals of stein's method. *Probab. Surveys*, 8:210–293, 2011. (Cited on page 33.)

W. Rudin. *Functional Analysis*. McGraw-Hill, 1991. (Cited on pages 26 and 28.)

M. Rudolph, F. Ruiz, S. Mandt, and D. Blei. Exponential family embeddings. In *NIPS*, 2016. (Cited on page 57.)

Purnamrita Sarkar and Andrew W. Moore. Dynamic social network analysis using latent space models. In *Advances in Neural Information Processing Systems 18*, pages 1145–1152. 2006. (Cited on pages 16 and 17.)

I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938. (Cited on pages 100 and 101.)

Dominic Schuhmacher and Kaspar Stucki. Gibbs point process approximation: Total variation bounds using Stein's method. *Ann. Probab.*, 42(5):1911–1951, 09 2014. (Cited on pages 90 and 91.)

Dino Sejdinovic and Arthur Gretton. What is an RKHS?, 2012. (Cited on pages 26, 27, and 28.)

Sohan Seth, Iain Murray, and Christopher K. I. Williams. Model criticism in latent space. *Bayesian Anal.*, 2018. Advance publication. (Cited on page 3.)

Daniel K. Sewell and Yuguo Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657, 2015. (Cited on page 17.)

Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *The Annals of Statistics*, 41(2):508–535, 2013. (Cited on page 13.)

Xiaofeng Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235, 2010. (Cited on page 37.)

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011. (Cited on pages 76 and 81.)

A. Simma and M. Jordan. Modeling events with cascades of Poisson processes. In *UAI*, 2010. (Cited on page 56.)

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948. (Cited on page 3.)

T. Snijders, G. van de Bunt, and C. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60, 2010. (Cited on page 16.)

Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1): 99–153, 2006. (Cited on page 13.)

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electron. J. Statist.*, 6:1550–1599, 2012. (Cited on page 30.)

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 583–602. University of California Press, 1972. (Cited on pages 5, 26, 32, 33, 34, and 87.)

Charles Stein. Approximate computation of expectations. *Institute of Mathematical Statistics Lecture Notes–Monograph Series*, 7:i–164, 1986. (Cited on pages 5, 26, 32, 33, 87, and 89.)

James Stewart. Positive definite functions and generalizations, an historical survey. *Rocky Mountain J. Math.*, 6(3):409–434, 09 1976. (Cited on page 36.)

David Strauss and Michael Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990. (Cited on pages 4 and 87.)

David J. Strauss. A model for clustering. *Biometrika*, 62(2):467–475, 1975. (Cited on pages 7, 23, and 88.)

X. Tan, S. Naqvi, A. Qi, K. Heller, and V. Rao. Content-based modeling of reciprocal relationships using Hawkes and Gaussian processes. In *UAI*, 2016. (Cited on pages 43 and 56.)

J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: Large-scale information network embedding. In *WWW*, 2015. (Cited on page 56.)

Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pages 142–155, 2016. (Cited on page 77.)

Marijtje A. J. van Duijn, Tom A. B. Snijders, and Bonne J. H. Zijlstra. p_2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2):234–254, 2004. (Cited on page 12.)

Marijtje A.J. van Duijn, Krista J. Gile, and Mark S. Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52 – 62, 2009. (Cited on page 13.)

M. N. M. van Lieshout. *Markov Point Processes and Their Applications*. Imperial College Press, 2000. (Cited on page 22.)

S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010. (Cited on page 76.)

U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007. (Cited on page 53.)

Stanley Wasserman and Philippa Pattison. Logit models and logistic regressions for social networks: I. An introduction to markov graphs and p^*. *Psychometrika*, 61(3): 401–425, 1996. (Cited on pages 6, 12, 61, and 80.)

D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393: 440–42, 1998. (Cited on page 12.)

Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Le Song, and Hongyuan Zha. Wasserstein learning of deep generative point process models. In *Advances in Neural Information Processing Systems 30*, pages 3247–3257. 2017. (Cited on page 86.)

Jiasen Yang, Vinayak Rao, and Jennifer Neville. Decoupling homophily and reciprocity with latent space network models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017. (Cited on page 9.)

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. (Cited on pages 9, 87, and 89.)

Jiasen Yang, Vinayak Rao, and Jennifer Neville. A Stein–Papangelou goodness-of-fit test for point processes. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019. (Cited on page 9.)

S. Young and E. Scheinerman. Random dot product graph models for social networks. In *WAW*, 2007. (Cited on page 15.)

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems 30*, pages 3391–3401. 2017. (Cited on page 86.)

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI, pages 804–813, 2011. (Cited on page 39.)

Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018. (Cited on page 39.)

A. APPENDIX TO CHAPTER 4

A.1 MAP Estimation Details

As described in subSection 4.3.3, we perform maximum a posteriori (MAP) inference to estimate the parameters in all the discussed models. In this subsection, we present the MAP estimation details for the HP and DLS models by deriving the closed form expressions of the log-posterior function and its gradients; the optimization can then be carried out using L-BFGS-B (Byrd et al., 1995). The derivations for the PLS, BLS, RLS models follow analogously, since they can all be viewed as degenerate cases of the DLS model.

Before presenting the MAP estimation details, recall that the observed data $\{(u, v, \mathcal{H}_{uv})\}_{u,v \in V}$ are collected over a time period $[0, T)$, where $\mathcal{H}_{uv} := \{t_i^{uv}\}_{i=1}^{n_{uv}}$ records the set of all time-points at which u sent v a message.

A.1.1 Hawkes Process (HP) Model

Recall the Hawkes Process (HP) model:

$$\lambda_{uv}(t) = \gamma + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, \phi_b(t - t_k^{vu}) \qquad \forall u \neq v$$

$$N_{uv}(\cdot) \sim \mathsf{HawkesProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$

Notice that

$$\Lambda_{uv}(0, T) = \int_0^T \lambda_{uv}(t) \mathrm{d}t = \gamma \, T + \sum_{b=1}^{B} \xi_b \sum_{k=1}^{n_{vu}} \left[\Phi_b(T - t_k^{vu}) - \Phi_b(0) \right]$$

where $\Phi_b(t) := \int_0^t \phi_b(s) \mathrm{d}s$.

Placing Gamma$(1,1)$ priors on γ and each ξ_b, and denoting $\boldsymbol{\xi} := \{\xi_b\}_{b=1}^B$, the joint density can be written as

$$p(\{\mathcal{H}_{uv}\}_{u,v=1}^n, \gamma, \boldsymbol{\xi}) \propto \prod_{\substack{u,v=1 \\ u \neq v}}^{n} \left\{ e^{-\Lambda_{uv}(0,T)} \prod_{k=1}^{n_{uv}} \lambda_{uv}(t_i^{uv}) \cdot e^{-\gamma} \cdot \prod_{b=1}^{B} e^{-\xi_b} \right\}$$

and the log-posterior function is given by

$$\begin{aligned}
\log p(\gamma, \boldsymbol{\xi} \mid \{\mathcal{H}_{uv}\}_{u,v=1}^n) &= \sum_{\substack{u,v=1 \\ u \neq v}}^{n} \left\{ -\Lambda_{uv}(0,T) + \sum_{i=1}^{n_{uv}} \log \lambda_{uv}(t_i^{uv}) \right\} - \gamma - \sum_{b=1}^{B} \xi_b \\
&= \sum_{\substack{u,v=1 \\ u \neq v}}^{n} \left\{ -\gamma\, T - \sum_{b=1}^{B} \xi_b\, \Delta_{b,T}^{vu} + \sum_{i=1}^{n_{uv}} \log \left(\gamma + \sum_{b=1}^{B} \xi_b\, \delta_{b,i}^{uv} \right) \right\} - \gamma - \sum_{b=1}^{B} \xi_b
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner-product, and we have adopted the shorthand notations

$$\begin{aligned}
\Delta_{b,T}^{vu} &:= \sum_{k=1}^{n_{vu}} \left[\Phi_b(T - t_k^{vu}) - \Phi_b(0) \right] \\
\delta_{b,i}^{uv} &:= \sum_{k:\, t_k^{vu} < t_i^{uv}} \phi_b(t_i^{uv} - t_k^{vu})
\end{aligned}$$

to denote data statistics that can be pre-computed and cached for each pair of nodes $u, v \in V$ and kernel ϕ_b.

The gradients of the log-posterior are given by

$$\begin{aligned}
\frac{\partial \log p}{\partial \gamma} &= -(n^2 - n)\, T + \sum_{\substack{u,v=1 \\ u \neq v}}^{n} \sum_{i=1}^{n_{uv}} \left(\gamma + \sum_{b=1}^{B} \xi_b\, \delta_{b,i}^{uv} \right)^{-1} - 1 \\
\frac{\partial \log p}{\partial \xi_b} &= \sum_{\substack{u,v=1 \\ u \neq v}}^{n} \left[-\Delta_{b,T}^{vu} + \sum_{i=1}^{n_{uv}} \delta_{b,i}^{uv} \left(\gamma + \sum_{b=1}^{B} \xi_b\, \delta_{b,i}^{uv} \right)^{-1} \right] - 1\,.
\end{aligned}$$

A.1.2 Hawkes Dual Latent Space (DLS) Model

Recall the Hawkes Dual Latent Space (DLS) model:

$$\boldsymbol{z}_v \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d \times d}) \qquad \forall v \in V$$

$$\boldsymbol{\mu}_v \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_{d\times d}) \qquad \forall v \in V$$

$$\boldsymbol{\varepsilon}_v^{(b)} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{d\times d}) \qquad \forall v \in V,\ b = 1,\dots,B$$

$$\boldsymbol{x}_v^{(b)} \sim \boldsymbol{\mu}_v + \boldsymbol{\varepsilon}_v^{(b)} \qquad \forall v \in V,\ b = 1,\dots,B$$

$$\lambda_{uv}(t) = \gamma\, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \beta\, e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2}\, \phi_b(t - t_k^{vu})$$

$$N_{uv}(\cdot) \sim \mathsf{HawkesProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$

Placing Gamma(1, 1) priors on γ and β, setting $\sigma^2 = \sigma_\mu^2 = \sigma_\varepsilon^2 = 1$, and integrating out $\{\boldsymbol{\mu}_v\}_{v=1}^n$, the log-density function can be written as

$$\begin{aligned}
&\log p(\gamma, \beta, \{\boldsymbol{z}_v\}_{v=1}^n, \{\{\boldsymbol{x}_v^{(b)}\}_{b=1}^B\}_{v=1}^n \mid \{\mathcal{H}_{uv}\}_{u,v=1}^n) \\
&\quad = \sum_{\substack{u,v=1 \\ u\neq v}}^{n} \Bigg\{ -\gamma\, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2}\, T - \beta \sum_{b=1}^{B} \Delta_{b,T}^{vu}\, e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2} \\
&\qquad\qquad + \sum_{i=1}^{n_{uv}} \log\left(\gamma\, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} + \beta \sum_{b=1}^{B} \delta_{b,i}^{uv}\, e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2} \right) \Bigg\} \\
&\quad\quad - \frac{1}{2} \sum_{v=1}^{n} \sum_{b=1}^{B} \|\boldsymbol{x}_v^{(b)}\|_2^2 + \frac{B^2}{2(B+1)} \sum_{v=1}^{n} \|\bar{\boldsymbol{x}}_v\|_2^2 - \frac{1}{2} \sum_{v=1}^{n} \|\boldsymbol{z}_v\|_2^2 - \gamma - \beta
\end{aligned}$$

where $\bar{\boldsymbol{x}}_v := \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{x}_v^{(b)}$ denotes the mean latent position of node v across all basis-kernels.

The gradients of the log-posterior are given by

$$\frac{\partial \log p}{\partial \gamma} = \sum_{\substack{u,v=1 \\ u\neq v}}^{n} \left[-T\, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} + \sum_{i=1}^{n_{uv}} e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2}\, h^{-1}(u, v, i) \right] - 1$$

$$\frac{\partial \log p}{\partial \beta} = \sum_{\substack{u,v=1 \\ u\neq v}}^{n} \sum_{b=1}^{B} r(u, v, b)\, e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2} - 1$$

$$\nabla_{\boldsymbol{z}_v} \log p = \sum_{\substack{u=1 \\ u\neq v}}^{n} \left\{ \gamma \left[-2\, T + \sum_{i=1}^{n_{uv}} \left(h^{-1}(u, v, i) + h^{-1}(v, u, i) \right) \right] e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} \cdot 2(\boldsymbol{z}_u - \boldsymbol{z}_v) \right\} - \boldsymbol{z}_v$$

$$\nabla_{\boldsymbol{x}_v^{(b)}} \log p = \sum_{\substack{u=1 \\ u\neq v}}^{n} \left\{ \beta \left[r(u, v, b) + r(v, u, b) \right] e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2} \cdot 2(\boldsymbol{x}_v^{(b)} - \boldsymbol{x}_u^{(b)}) \right\} - \boldsymbol{x}_v^{(b)} + \frac{B}{B+1} \cdot \bar{\boldsymbol{x}}_v$$

where

$$h(u,v,i) := \gamma\, e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2} + \beta \sum_{b=1}^{B} \delta_{b,i}^{uv}\, e^{-\|\boldsymbol{x}_u^{(b)} - \boldsymbol{x}_v^{(b)}\|_2^2}$$

$$r(u,v,b) := -\Delta_{b,T}^{vu} + \sum_{i=1}^{n_{uv}} \delta_{b,i}^{uv}\, h^{-1}(u,v,i)\,.$$

A.2 Additional Experiment Results

A.2.1 Further Experiment on Static Link Prediction

In Section 4.4.3, we noted that the experiment setup for the static link prediction task did not yield standard errors for the AUC scores reported in Table 4.3, since there was only one training/test split. To investigate the statistical significance of the results, we conducted a follow-up experiment.

For each dataset, we computed confidence intervals by performing six trials on subsets of the data. Specifically, in the i-th trial, we let the training set to contain all events during the period between the $\lceil \frac{i-1}{10} \rceil$-th and the $\lfloor \frac{i+2}{10} \rfloor$-th event, and the test set to contain all events during the period between the $\lceil \frac{i+2}{10} \rceil$-th and $\lfloor \frac{i+4}{10} \rfloor$-th event. In this way, each trial used 30% training data and 20% test data, with the training and test data being non-overlapping.[1] As in Section 4.4.3, we fitted the model on the training set, and performed link prediction on the test set. The results are shown in Table A.1.[2]

By conducting two-sided t-tests at the 95% confidence level, we conclude that while DLS significantly outperforms node2vec on ENRON, their performance differences on EMAIL and FACEBOOK are not significant.

[1] Notice, however, that the training/test data across different trials may share common observations. Thus, strictly speaking, the trials are not independent, and the computed standard error estimates might under-estimate the "true" associated uncertainty.

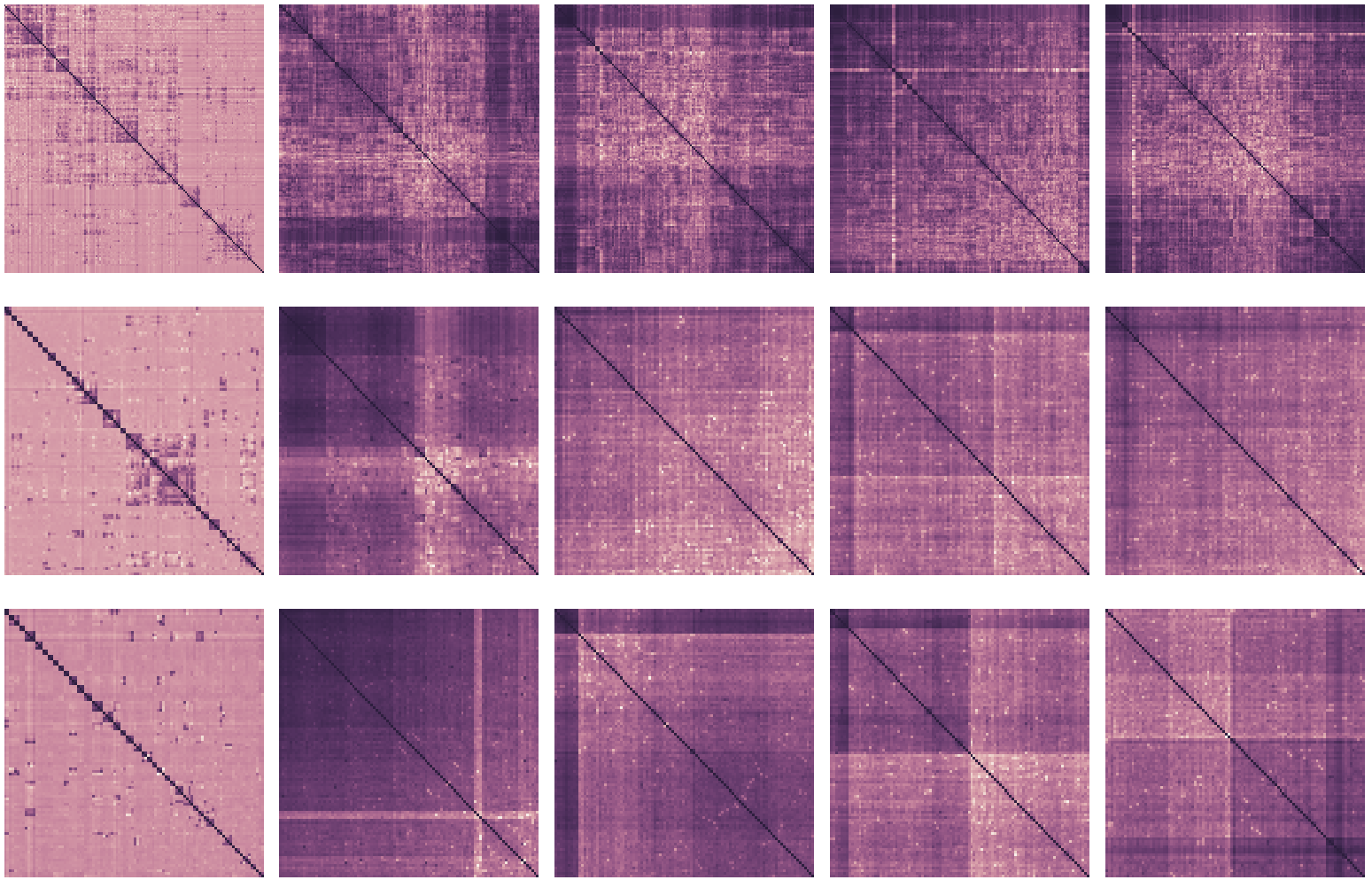
[2] Note that the overall performance for all methods are slightly degraded since we are only using subsets of the data.

Table A.1.: Static link prediction AUC scores and standard deviations.

Model	ENRON	EMAIL	FACEBOOK
PLS	0.510 (0.009)	0.496 (0.015)	0.491 (0.013)
BLS	0.510 (0.009)	0.496 (0.015)	0.491 (0.013)
RLS	0.439 (0.073)	0.386 (0.081)	0.456 (0.055)
DLS	**0.864 (0.016)**	**0.934 (0.016)**	**0.892 (0.040)**
Spectral	0.516 (0.020)	0.526 (0.032)	0.492 (0.021)
node2vec	0.749 (0.050)	**0.953 (0.007)**	**0.935 (0.033)**

A.2.2 Visualization of the Inferred Node-Similarity Matrices

We visualize the estimated homophily and reciprocal latent spaces of the DLS model by computing the pair-wise similarities $e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2}$ for every pair of nodes $u, v \in V$, and then plotting a heat-map of the inferred similarity matrices. Figure A.1 shows the heat-maps (colors on log-scale) for both the homophily latent space and the reciprocal latent spaces corresponding to the hourly (ϕ_1), daily (ϕ_2), weekly (ϕ_3) exponential kernels and the weekly locally periodic kernel (ϕ_4) on all three datasets. For each similarity matrix, we performed hierarchical clustering on the rows to obtain a node-ordering and accordingly permuted the rows and columns of the matrix simultaneously. Notice that the similarity matrices exhibit different clustering block-structures, indicating that the user-interaction patterns are quite different across the homophily and reciprocal latent spaces with different kernels and time-scales.



(a) $e^{-\|\boldsymbol{z}_u - \boldsymbol{z}_v\|_2^2}$ (b) $e^{-\|\boldsymbol{x}_u^{(1)} - \boldsymbol{x}_v^{(1)}\|_2^2}$ (c) $e^{-\|\boldsymbol{x}_u^{(2)} - \boldsymbol{x}_v^{(2)}\|_2^2}$ (d) $e^{-\|\boldsymbol{x}_u^{(3)} - \boldsymbol{x}_v^{(3)}\|_2^2}$ (e) $e^{-\|\boldsymbol{x}_u^{(4)} - \boldsymbol{x}_v^{(4)}\|_2^2}$

Figure A.1.: Inferred node-similarity matrices in ENRON (*top row*), EMAIL (*middle row*), and FACEBOOK (*bottom row*).

B. APPENDIX TO CHAPTER 6

For concreteness, we provide an example Python implementation of Eq. (6.8).

```
def kernel(X, Y):
    """
    Evaluates a kernel function for two point-sets X and Y.
    Args:
        X, Y: numpy arrays of shape (_,d), collections of d-dimensional points.
    Returns:
        float, value of k(X, Y).
    """

def papangelou(u, X):
    """
    Evaluates the Papangelou conditional intensity (u|X) of a point process
        at location u given observed point-set X.
    Args:
        u: numpy array of shape (d,), a location in the ground space.
        X: numpy array of shape (n, d), the given point-set.
    Returns:
        float, value of rho(u|X).
    """

def integrate(func, domain):
    """
    Integrates a (univariate or multivariate) function func over domain.
        See scipy.integrate for a list of numerical integration routines.
    Args:
        func: function, a univariate or multivariate function.
        domain: list of lists, the integration ranges for each variable.
    Returns:
        float, value of the definite integral.
    """

def kappa(X, Y, domain):
    """
    Evaluates Eq.(12) for point-sets X and Y using kernel() and papangelou().
    Args:
        X, Y: numpy arrays of shape (_,d), collections of d-dimensional points.
```

```
    domain: list of lists, ranges specifying the ground space.
Returns:
    float, value of kappa(X, Y).
"""
n = X.shape[0]
m = Y.shape[0]
k = kernel(X, Y)
k_X = sum(kernel(np.delete(X, i, axis=0), Y) for i in xrange(n))
k_Y = sum(kernel(X, np.delete(Y, j, axis=0)) for j in xrange(m))
k_X_Y = sum(kernel(np.delete(X, i, axis=0), np.delete(Y, j, axis=0))
            for i in xrange(n) for j in xrange(m))

def integrand_uv(u, v):
    # Double integrand over u and v
    k_uv = kernel(np.vstack((X, u)), np.vstack((Y, v)))
    k_v = kernel(X, np.vstack((Y, v)))
    k_u = kernel(np.vstack((X, u)), Y)
    c_u = papangelou(u, X)
    c_v = papangelou(v, Y)
    return (k_uv - k_v - k_u + k) * c_u * c_v

def integrand_v(v):
    # Integrand over v
    k_X_v = sum(kernel(np.delete(X, i, axis=0), np.vstack((Y, v)))
                for i in xrange(n))
    k_v = kernel(X, np.vstack((Y, v)))
    c_v = papangelou(v, Y)
    return ((k_X_v - k_X) - n*(k_v - k)) * c_v

def integrand_u(u):
    # Integrand over u
    k_Y_u = sum(kernel(np.vstack((X, u)), np.delete(Y, j, axis=0))
                for j in xrange(m))
    k_u = kernel(np.vstack((X, u)), Y)
    c_u = papangelou(u, X)
    return ((k_Y_u - k_Y) - m*(k_u - k)) * c_u

# Compute integrals
term1 = integrate(integrand_uv, domain)
term2 = integrate(integrand_v, domain)
term3 = integrate(integrand_u, domain)
term4 = k_X_Y - n*k_Y - m*k_X + m*n*k

return term1 + term2 + term3 + term4
```

VITA

Jiasen Yang was born in Beijing, China on May 21, 1995. He obtained a B.S. degree in Statistics from the Special Class for the Gifted Young at the University of Science and Technology of China in 2013, and a joint M.S. degree in Statistics and Computer Science from Purdue University in 2015. His research interests lie at the intersection of machine learning, statistics, and theoretical computer science, with specific topics including statistical network analysis, point processes, kernel and nonparametric methods, Stein's method, approximate Bayesian inference, and randomized sketching methods.