Approximate Sampling for Doubly-intractable Distributions and Modeling Choice Interdependence in a Social Network

by

Jing Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2011

Doctoral Committee:

Assistant Professor Yves A. Atchadé, Co-Chair
Assistant Professor Anocha Aribarg, Co-Chair
Professor Peter J. Lenk
Professor George Michailidis
Associate Professor Kerby A. Shedden

ACKNOWLEDGEMENTS

Foremost, I owe my deepest gratitude to my co-advisors Prof. Yves Atchadé and Prof. Anocha Aribarg for their continuous guidance and support during my doctoral study. I'm grateful for their inspiration, encouragement, patience, and enormous knowledge, without which this thesis would not have been possible.

I would like to thank the rest of my committee members: Prof. George Michailidis, Prof. Kerby Shedden and Prof. Peter Lenk, for their insightful comments.

My thanks also go to Stefan Witwicki for programming the survey and raffle websites. It is his hard work that made our field experiment feasible.

I wish to show my gratitude to Dr. Brenda Gunderson, Prof. Edward Rothman, and Prof. Ji Zhu for offering teaching and research assistant opportunities.

I'm indebted to my fellow student colleagues in the Department of Statistics and colleagues at Center for Statistical Consultation and Research, for stimulating discussions and all the fun we had together. I'm especially thankful to Yang Yang, Hao Zhou, Fangyi Liu, Jian Guo, Joel Vaughan, Eric Laber, Giselle Kolenic, Joe Kazemi, Kathy Welch, and Lingling Zhang.

Last but not least, I would like to thank my parents Jiaxiang Wang and Xuqin Jiang, my boyfriend Xuhuai Zhang, and friends Jie Gong and Shaoru Ge, for their love and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

Interdependent data occur in many real-world applications, for example, clustered data such as scores of students in the same classroom, repeated measures of the same subject under different experiment conditions, and longitudinal data. Here we focus on the scenarios where interdependence arises due to the proximity in some space. These dependent data constitute a random field. Examples are abundant in spatial statistics, image restoring, speech processing, and marketing. Risks of a certain disease tend to be more similar for neighboring states than distant states. Nearby pixels in a photo are more likely to have the same values. Note that space here is not restricted to physical space. Vicinity in social space could also give rise to dependence. People are connected with each other through different relations, which constitute social networks. By exchanging information and resources, oftentimes connected individuals have dependent behavior, such as purchase decisions.

In this thesis, we are particularly interested in the problem of modeling interdependent choices of consumers connected in a social network. With the emergence and growth of social media (e.g., Facebook, Twitter), innovative marketing tools have been invented to capitalize on the social networks (among the users) embedded on these websites. The functionality of the websites further facilitates users' ability to observe their friends' behavior and choices, and thus increases their chance of being influenced. This raises the question of how to quantify the effect of social network in this process. A significant network effect will justify the effectiveness and necessity for new advertising strategies, and will encourage and inspire the design of more marketing tools that are tailored for social media.

Random field models represent a general and flexible class of models that can be adopted to study interdependent choice data. However, the applicability of many of these models is hampered by some computational difficulties, which are of primary interest here. The main contributions of the thesis are the following.

1. We propose in chapter II a general framework for approximate sampling from the posterior distribution of these models. We also develop a new Metropolis-Hastings algorithm to efficiently sample from exponential random graph models, which are special cases of random field models for social networks.

2. In chapter IV, we study the problem of estimating ratio of normalizing constants, which for instance can be applied to calculate Bayes factor. We propose a flexible implementation of the path sampling identity. The preliminary simulation shows a good potential of the method.

3. We develop plausible models for consumers' interdependent choices in chapter III, with specifications that accommodate multiple sources of dependence and asymmetric social interactions. We design and conduct a novel field experiment to mimic the interactive advertising on Facebook.

The remaining of this chapter is organized as follows. To better understand the existing random field models, we offer a literature review. Then we describe in further details the outstanding questions and contributions of the thesis in sections 1.2 and 1.3.

1.1 Literature Review

Most of the random field models fall into two major frameworks: Markov random fields and mixed models. We will review the formulation of both frameworks and the extension to incorporate heterogeneous interdependence. Markov random fields directly model dependence on the observations, which encourages neighboring sites to have similar data. Whereas in mixed models, dependence is imposed on the random effect in the (transformed) mean, which induces marginal dependence of the data.

1.1.1 Markov Random Fields

Besag (1974) first advocated the use of Markov random fields as statistical models for spatial data, which later became prevalent in spatial statistics, image analysis, and etc. Before going into the details of the model, we first introduce some terminology. The two core concepts of a Markov random field (MRF) are 'neighbor' and 'clique'.

Assume the random variables we study are $X_1, \cdots, X_n$, with $1, \cdots, n$ indexing their positions in some space, which we call 'sites' (Besag 1974). A neighboring system defines the neighbors of each site, and satisfies the non-reflexive and symmetric constraints, i.e., any site is not a neighbor of itself, and if i is a neighbor of j then j is also i's neighbor. For instance, two people are neighbors if they are connected in a network. $X_1, \cdots, X_n$ constitute a Markov random field wrt a pre-defined neighboring system, if they satisfy the Markovian assumption,

$$p(x_i|x_{-i}) = p\left(x_i|\{x_j, j \in N(i)\}\right), 1 \leq i \leq n,$$

where $X_{-i} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$ and $N(i)$ is the set of all i's neighbors.

In other words, given neighbors' values, X_i is independent of random variables of all other sites. A clique consists of a single site or a set of sites which are neighbors of each other. The size of a clique is defined as the number of sites in that clique.

Assuming the positive condition, i.e., if $p(x_i) > 0$ for $1 \leq i \leq n$, then $p(x_1, \cdots, x_n) > 0$, the Hammersley-Clifford theorem (Hammersley and Clifford 1971) states that only the potentials within cliques can contribute to the likelihood. The likelihood of a Markov random field belongs to exponential family and takes the following form,

$$p(x_1, \cdots, x_n|\theta) = \frac{1}{Z(\theta)} \exp\left\{\sum_{c \in \mathcal{C}} U_c\left(x_i, i \in c; \theta\right)\right\}, \tag{1.1}$$

where θ represents the unknown parameters, $\mathcal{C}$ is the set of all possible cliques, U_c called potential is a function characterizing the feature of or interaction within clique c, and $Z(\theta) = \sum_{x_1, \cdots, x_n} \exp\left\{\sum_{c \in \mathcal{C}} U_c\left(x_i, i \in c; \theta\right)\right\}$ is a normalizing constant. The theorem greatly simplifies the specification of an MRF, while still retaining generality. There are no restrictions on the particular functional form U_c can take, and it can vary from clique to clique.

MRF is not only used to directly model interdependent data, but also serve as a prior distribution in the mixed models to introduce dependence. The remaining of this section is dedicated to introducing some common specifications of an MRF. We will touch upon how to use it as a prior in section 1.1.2.

Auto-models

Besag (1974) proposed a subclass of MRF's, called 'auto-models'. He took a conditional probability approach to construct the joint probability under some assumptions. For each scheme of auto-models, the conditional distribution of X_i only involves a weighted sum of its neighboring site values in additional to X_i's own characteristics. Therefore, in the joint distribution, there are only cliques of size 1 and

2. We will introduce two schemes of auto-models.

Auto-normal Model

For continuous data, if a multivariate Gaussian distribution can be assumed, then auto-normal models will apply. In an auto-normal model, the conditional distribution of X_i is

$$p(x_i|x_{-i}) = (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp\Big\{ - \frac{1}{2\sigma_i^2}[x_i - A_i^T\beta - \sum_{j=1}^{n} c_{ij}(x_j - A_j^T\beta)]^2 \Big\},$$

where β, σ_i, c_{ij} are parameters, T means matrix transpose, and A_i is vector of predictors for i. This formulation (Cressie 1993) is slightly more general than the one introduced in Besag (1974), which has the same conditional variance σ^2 for all i. $c_{ij} = 0$, for $j \notin N(i)$, and thus X_i only depends on its neighbors. In other words, conditional distribution of X_i is

$$X_i|X_{-i} \sim N(A_i^T\beta + \sum_{j\in N(i)} c_{ij}(x_j - A_j^T\beta), \sigma_i^2).$$

Let $X = (X_1, \cdots, X_n)$, $A = (A_1, \cdots, A_n)$, C be a matrix whose (i, j)-th entry is $C(i, j) = c_{ij}$, $M = \text{diag}\{\sigma_1^2, \cdots, \sigma_n^2\}$, meaning a diagonal matrix with σ_i^2 as i-th diagonal entry, I_n be an $n \times n$ identity matrix. Then X jointly follows a multivariate normal distribution MVN$(A^T\beta, (I_n - C)^{-1}M)$, provided that $I_n - C$ is invertible and $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$. To ensure a valid joint distribution, $(I_n - C)^{-1}M$ as a covariance matrix needs to be symmetric and positive definite. This formulation is also called conditional autoregressive (CAR) model.

Auto-logistic Model

In binary data scheme, the conditional probability for i follows a Bernoulli distribution,

$$p(x_i|x_{-i}) = \exp\{A_i^T\beta x_i + \sum_{j=1}^{n} c_{ij}x_ix_j\}/\{1 + \exp(A_i^T\beta + \sum_{j=1}^{n} c_{ij}x_j)\}.$$

Again, $c_{ij} = 0$ if $i \notin N(j)$, and $c_{ij} = c_{ji}$ is imposed to ensure the existence of a joint distribution. The corresponding joint probability is

$$p(x) = \frac{1}{Z(\alpha, C)} \exp\Big\{ \sum_{i=1}^{n} A_i^T \beta x_i + \sum_{1 \leq i < j \leq n} c_{ij} x_i x_j \Big\}.$$

In practice, often a homogeneous MRF parameterization is preferred to maintain a parsimonious model, where the parameters are the same for cliques of the same size. In this case, it means all c_{ij}'s are equal (Haran 2010). Auto-logistic model is widely used in spatial statistics and image analysis, and was also introduced to marketing by Russell and Petersen (2000).

Apart from the above models, auto-models can be readily derived for other schemes, such as binomial, poisson (Besag 1974) and multinomial data (Yang et al 2009).

1.1.2 Mixed Models

Mixed models provide another general framework to account for interdependence. Different from the conventional mixed models, the random effect captures the interdependence due to proximity in some space. In this section, we first review linear Gaussian process model for continuous data, which by itself is not necessary a mixed model, but often used as a distribution for the random effect in mixed models. Then mixed models for non-Gaussian data are introduced with different choices of random effect distributions.

Linear Gaussian Process Model

In a linear Gaussian process model, each X_i is expressed as

$$X_i = A_i^T \beta + w_i. \tag{1.2}$$

Let $w = (w_1, \cdots, w_n)$. Dependence is imposed through $w \sim \text{MVN}(0, \Sigma)$, where Σ is the covariance matrix. This model is widely adopted in spatial statistics, where each site is associated with a location s_i. To ensure that the distribution is proper, Σ needs to be specified as a symmetric and positive definite matrix. Many common parameterizations in practice fall into a general family called Matérn covariance (Hancock and Stein 1993).

Matérn covariance between X_i and X_j is a function of the distance between s_i and s_j. The concepts of location and distance are intuitively clear in spatial statistics. In the general context here (interdependence of the data originates from the proximity in some space), the ideas can still be carried over to more general space. The locations do not have to be longitudes and latitudes, but any finite-dimension coordinates indicating the positions in the space. It is actually not necessary to know the locations, as long as we have a clear criterion to calculate distances. For example, for people connected in a social network, the distance between any two individuals could be the length of the shortest path between them, or infinite if there is no path connecting them.

Matérn covariance with parameters $(\psi, \kappa, \phi, \nu)$ between two random variables with distance d is

$$\text{cov}(d; \psi, \kappa, \phi, \nu) = \begin{cases} \frac{\kappa}{2^{\nu-1}\Gamma(\nu)} \left(2\nu^{1/2} d/\phi\right)^{\nu} K_{\nu}\left(2\nu^{1/2} d/\phi\right) & \text{if } d > 0 \\ \psi + \kappa & \text{if } d = 0 \end{cases}$$

$K_\nu(d)$ is a modified Bessel function of order ν. ν controls the smoothness of the function. Larger ν leads to a smoother function. Two special cases of Matérn covariance function are exponential and gaussian covariance functions, corresponding to $\nu = .5$ and ∞, respectively.

If the exponential covariance function is applied, Σ can also be written as $\psi I_n +$

$\kappa H(\phi)$, with $H(\phi)$ an $n \times n$ matrix. The (i, j)-th entry of $H(\phi)$ is $\exp(-d_{ij}/\phi)$, d_{ij} representing the distance between sites i, j. Euclidean distance is often used, but other measures of distance are also popular depending on the applications, such as Manhattan distance and geodesic distance. Exponential covariance functions decompose the error into two parts: dependence due to proximity, $\kappa H(\phi)$, and independent error of individual sites, ψ. The formulation of $H(\phi)$ implies that the correlation decreases as the distance increases. In other words, sites far apart are less correlated.

Non-Gaussian Data

The linear Gaussian process model with exponential covariance function can also be framed into a mixed model.

$$X_i = A_i^T \beta + u_i + v_i, \tag{1.3}$$

where v_i is i.i.d. $N(0, \psi)$, u_i independent of v_i is a random effect capturing dependence, and u jointly follows $\mathrm{MVN}(0, \kappa H(\phi))$.

The same idea gives rise to the mixed model framework for dependent non-Gaussian data (Diggle et al. 1998). Recall that the usual generalized linear models connect the linear predictors and the expectation with a link function g: $\mathbb{E}(X_i|A_i) = \mu_i$, $\eta_i = g(\mu_i) = A_i^T \beta$. Here a random effect can be used to introduce dependence. For generality, we return to the earlier parameterization (1.2). We have

$$\eta_i = g(\mu_i) = A_i^T \beta + w_i, \tag{1.4}$$

where w_i yields dependence of the data. There are two major specifications of w: Gaussian process and Gaussian MRF (CAR and intrinsic conditional autoregressive model, Besag et al. 1991). More specifically, if we use Gaussian process, then $w \sim \mathrm{MVN}(0, \Sigma)$, with Σ parameterized as in section 1.1.2. If we use Gaussian MRF, formulation similar to (1.3) is employed (Besag et al. 1991), with $w_i = u_i + v_i$. $u = (u_1, \cdots, u_n) \sim \text{MVN}(0, (I_n - C)^{-1}M)$ or $\text{MVN}(0, Q^{-1})$, $v = (v_1, \cdots, v_n) \sim \text{MVN}(0, \psi I_n)$, and v is independent of u. Q is the precision matrix of intrinsic conditional autoregressive (ICAR) model. Its off-diagonal entry (i, j) is $-c_{ij}/\sigma^2$ with $c_{ij} = 1$ if i and j are neighbors, and 0 otherwise. Its i-th diagonal entry is $\sum_{j \neq i} c_{ij}/\sigma^2$. Q has rank $n - 1$, and thus is not positive definite. A common practice is to use ICAR as a prior for u in a Bayesian framework, and in many situations the posterior still maintains property despite a improper prior. See Sun et al. (1999) for conditions which ensure the property of posterior distributions with Gaussian MRF priors.

According to the description above, it is straightforward to build models for different types of non-Gaussian data. We take binary data as an illustration. Yang and Allenby (2003) provides a good example, where a probit formulation is used for the binary outcome. Each X_i is associated with a latent variable/random effect w_i, s.t. $p(x_i = 1) = p(A_i^T \beta + w_i > 0)$, and a Gaussian process prior is assigned to w. More specifically,

$$w_i = u_i + v_i, \qquad u = \rho B u + \epsilon,$$

where B is some weight matrix reflecting the proximity of the sites (consumers in their paper), $v \sim \text{MVN}(0, I_{n\times n})$, and $\epsilon \sim \text{MVN}(0, \sigma^2 I_{n\times n})$. With some algebra, it can be shown that

$$w \sim \text{MVN}\left(0, I_{n\times n} + \sigma^2 (I_{n\times n} - \rho B)^{-1} (I_{n\times n} - \rho B^T)^{-1}\right).$$

Thus the prior of w is essentially a Gaussian process.

1.1.3 Models for Heterogeneous Interactions

With homogeneous interaction, a global parameter is used to capture dependence. For instance, in auto-logistic, all non-zero c_{ij} has the same value. In linear Gaussian process, $H(\phi)$ is specified by only one parameter. These models may suffer from oversmoothing the dependence.

Inhomogeneous MRF

MRF is flexible enough to deal with inhomogeneous interactions. The question remaining is how to specify the parameters so that the model is not over-parameterized while still retaining enough flexibility. Aykroyd and Zimeras (1999) provides an example. For a general MRF, assume that each potential involves a pairwise interaction between the neighbors with c_{ij} capturing the strength. Let $c_{ij} = \gamma_i + \gamma_j$. Then c_{ij}'s sharing a site will be correlated. γ_i's can be specified as i.i.d. from some simple distribution. Aykroyd (1998) provides an alternative formulation. Inhomogeneous MRF can also be used as the distribution for random effect in a mixed model to account for heterogeneous interdependence.

Allocation Models

In the mixed model framework, allocation models (Knorr-Held and Raber 2000, Denison and Holmes 2001, Green and Richardson 2002, Fernández and Green 2002) offer an approach to study the heterogeneity of random effect. The basic assumption is that there are k values for the random effect, $w_1, \cdots, w_k$, k known or unknown. An allocation variable will assign random effect for each i to one of the k values. Models in the literature mainly differ by how they model the allocation variable.

As an extension, allocation models also offer a class of flexible priors for heterogeneous interactions in inhomogeneous MRF, e.g., prior for c_{ij} in the above example.

Different from the parameterization above, utilizing allocation models allows for discontinuity in the interaction strengths.

1.2 Computational Contributions

The contributions of the thesis in tackling computational issues are three-fold. Firstly, while theoretically sound, it may not be computationally applicable to apply some of the random field models to study interdependent choices. For example, in MRF, the biggest obstacle comes from its intractable normalizing constant, $Z(\theta)$ in (1.1), where $Z(\theta) = \sum_{x_1,\cdots,x_n} \exp\left\{\sum_{c\in\mathcal{C}} U_c\left(x_i, i\in c;\theta\right)\right\}$. For binary data, the summation involves 2^n terms. For moderate n, it is already impossible to calculate $Z(\theta)$. The intractable normalizing constant makes inference difficult for both frequentist and Bayesian approaches. This problem does not only occur in MRF, but many other statistical models including exponential random graph models for social network analysis.

Frequentist methods to tackle this problem are abundant in the literature (Besag 1974, Younes 1988, Geyer and Thompson 1992, Minka 2001, Snijders 2002, Hunter and Handcock 2006), whereas much less has been done for Bayesian inference. Moreover, the asymptotic properties of the MLE for this kind of models are poorly understood, which makes Bayesian inference particularly attractive. In Bayesian paradigm, the convergence condition for algorithms based on loopy belief propagation (Qi et al. 2005, Welling and Parise 2006) is not well understood either. Asymptotically consistent Markov chain Monte Carlo (MCMC) algorithms (Møller et al. 2006, Murray et al. 2006, Atchadé et al. 2008) for the posterior distribution were developed recently. However, the first two are restricted by the need of exact sampling, and all three are computationally intensive for large data sets.

In chapter II, we propose an efficient MCMC algorithm that improves on both aspects by replacing the exact sampling with approximate sampling. The theoretical study shows asymptotic consistency of our algorithm. We showcase its superior performance on large data sets in real data analysis.

Secondly, in this thesis, we are also interested in the Bayesian inference of large social networks. After introducing the approximate MCMC algorithm in the context of social network modeling in chapter II, we will address the slow mixing issue due to the sparsity of large networks. We develop a refined Metropolis-Hastings proposal used in simulating networks, which outperforms existing proposals in our real data example.

Lastly, along the lines of intractable normalizing constant, estimating the ratio of normalizing constants can be of major interest in some problems. For instance, the marginal density can be considered as a normalizing constant of a posterior density. Bayes factor, a key quantity in Bayesian inference, then becomes the ratio of two normalizing constants. Bridge sampling (Meng and Wong 1996) and path sampling (Gelman and Meng 1998) are two popular methods in the literature. In chapter IV, we propose a flexible implementation of the path sampling identity with a non-homogeneous Markov chain. The new method generates a consistent estimator of the ratio, and in particular performs better when the two distributions, the ratio of whose normalizing constants we compute, are 'far apart'.

1.3 Modeling Interdependent Choices in a Social Network

Motivated by the interactive advertising on Facebook, we designed and conducted a novel field experiment to better understand how consumers make choices in a social network. Two types of data were collected to take into account of possible practical constraints in the real-world data collection.

In chapter III, we apply a discrete-time Markov chain and an MRF to study these two types of data, respectively. As mentioned in the beginning of section 1.1, MRF offers a more direct way of modeling discrete choice data, and thus is the framework adopted here. Moreover, we are able to build a theoretical connection between Markov chain and MRF. We present some conditions for a Markov chain to converge to an MRF, which further suggests that MRF is a natural model that emerges from the Markov chain dynamic choice process.

We propose specifications for both models to accommodate multiple sources of dependence and asymmetric social interactions. The approximate sampling introduced in chapter II is employed in the estimation of MRF, and compared with the pseudolikelihood method. Our results show that participants rely on choices observed both at the micro (i.e., closely-connected friends) and macro (i.e., a reference group such as male vs. female) levels in making their decisions. We also discuss two avenues of model extensions in section 3.6.

CHAPTER II

Approximate Sampling for Doubly-intractable Distributions

2.1 Problem Description

Intractable normalizing constant occurs in many statistical models, including spatial point process (Møller and Waagepetersen 2003), image analysis (Ibanez and Simo 2003), protein design (Kleinman et al. 2006), social network modeling (Hunter and Handcock 2006), and a wide range of other statistical models. Although we may know the analytical form of the normalizing constants, it is computationally infeasible to calculate them (e.g., $Z(\theta)$ of MRF in section 1.2). This imposes a great challenge in both finding the maximum likelihood estimate (MLE) and carrying out Bayesian analysis based on the posterior distribution. In this chapter, we focus on the context of social networks modeling with exponential random graph (ERG) models (Wasserman and Pattison 1996), and propose an efficient algorithm for sampling from the posterior distribution. We want to stress that the algorithm can be applied to models with intractable normalizing constants in other areas apart from social networks. Moreover, we will touch upon another computational issue in social network modeling, slow mixing in simulating large sparse networks, and suggest a refined algorithm.

Networks are used to represent relations or connections among various types of entities, e.g., interconnected websites, airline networks, and electricity networks. Network analysis has a broad range of applications including social science, epidemics, language processing, and etc. Here we focus on social networks where the nodes of the network typically represent individuals or other social entities and the edges capture relationships such as friendship and collaboration.

A network x with n nodes is a $n \times n$ matrix, where entry x_{ij} represents the strength of the connection between the ordered pair of nodes (i, j), and can take a finite number of values. Let $\mathcal{X}$ be the space of all such networks (we omit the dependence on n). In the simplest case where the network is undirected and binary, $x_{ij} = x_{ji} = 1$ if there is an edge between nodes i and j (in other words, i and j are neighbors), and 0 otherwise. In this case, $\mathcal{X}$ is the space of all triangular matrices $\{x_{ij},\ 1 \leq i < j \leq n\}$ with 0-1 entries. As mentioned above, the specific modeling framework for social networks considered here is the ERG model. It assumes a probability distribution for x given by

$$p_\theta(x) = \frac{f_\theta(x)}{Z(\theta)} = \frac{1}{Z(\theta)} \exp\left\{\sum_{k=1}^{K} \theta_k S_k(x)\right\}, \quad x \in \mathcal{X}, \tag{2.1}$$

where θ represents the unknown parameters, $f_\theta(x)$ is the un-normalized likelihood, and $Z(\theta)$ is the normalizing constant,

$$Z(\theta) = \sum_{x \in \mathcal{X}} \exp\left\{\sum_{i=k}^{K} \theta_k S_k(x)\right\}. \tag{2.2}$$

$S_k(x)$ is a network statistic of interest, for example, the number of edges $E(x) = \sum_{i,j} x_{ij}$ to capture network density; the number of triangles $\sum_{i,j,h} x_{ij} x_{jh} x_{hi}$ to capture transitivity; the number of 2-stars $\sum_{i,j,h} x_{ih} x_{jh}$, where a k-star ($k \geq 2$) is a node with k neighbors or a node of degree k. See Wasserman and Pattison (1996), Snijders et al. (2006) and Hunter and Handcock (2006) for more examples.

We propose a flexible and efficient algorithm to sample from the posterior distribution of the parameter θ in ERG models. The algorithm also applies to other statistical models where the intractable normalizing constant is an issue. We also propose a new Metropolis-Hastings (M-H) sampler to sample efficiently from the ERG distribution p_θ on the space of networks $\mathcal{X}$. The new algorithm is particularly effective in dealing with large sparse social networks.

2.1.1 MCMC for Models with Intractable Normalizing Constants

One of the main issue with ERG models is the intractability of the normalizing constant $Z(\theta)$. Evaluating $Z(\theta)$ is simply infeasible for most networks in practice. For an undirected binary network of 10 people, the calculation of $Z(\theta)$ may involve 2^{45} terms. This problem, which becomes more severe for large networks, implies that the likelihood function and the posterior distribution (in a Bayesian framework) cannot be evaluated, even up to a normalizing constant.

Because of the intractable function $Z(\theta)$, computing the maximum likelihood estimator (MLE) for this model is not straightforward. A solution dating back to Besag (1974) is pseudo-likelihood methods where the likelihood function is replaced by a more tractable approximation (Strauss and Ikeda 1990). But this often does not work well in practice. More recently, algorithms based on stochastic approximation (Younes 1988), MCMC (Geyer and Thompson 1992, Hunter and Handcock 2006), and loopy belief propagation (Minka 2001) have been developed that make it possible to compute the MLE reliably. However, the behavior (particularly the asymptotic behavior) of the MLE for this type of statistical models still poorly understood. This fact makes the Bayesian inference for ERG models particularly attractive.

Here we take a Bayesian approach. We assume that the parameter space Θ is a subset of the p-dimensional Euclidean space $\mathbb{R}^p$. Let us denote the observed network by $\mathcal{D}$, and μ the density (wrt the Lebesgue measure on Θ) of the prior distribution for θ on the parameter space Θ. Then the posterior distribution has density

$$\pi(\theta|\mathcal{D}) = \frac{1}{\pi(\mathcal{D})}\frac{1}{Z(\theta)}f_\theta(\mathcal{D})\mu(\theta), \quad \theta \in \Theta,$$

where $\pi(\mathcal{D})$ is the normalizing constant of the posterior distribution (or the marginal distribution of $\mathcal{D}$). There are two intractable normalizing constants in the posterior $\pi(\theta|\mathcal{D})$, and Murray et al. (2006) coined the term doubly-intractable distribution to refer to this type of distributions. Conventional MCMC methods can get rid of $\pi(\mathcal{D})$ but not $Z(\theta)$. For example, in the Metropolis-Hastings algorithm, given the current θ and a new value θ' proposed from $Q(\theta, \theta')$, the acceptance ratio is

$$\min\left[1, \frac{\pi(\mathcal{D})}{\pi(\mathcal{D})}\frac{f_{\theta'}(\mathcal{D})}{f_\theta(\mathcal{D})}\frac{Z(\theta)}{Z(\theta')}\frac{\mu(\theta')}{\mu(\theta)}\frac{Q(\theta',\theta)}{Q(\theta,\theta')}\right] = \min\left[1, \frac{f_{\theta'}(\mathcal{D})}{f_\theta(\mathcal{D})}\frac{Z(\theta)}{Z(\theta')}\frac{\mu(\theta')}{\mu(\theta)}\frac{Q(\theta',\theta)}{Q(\theta,\theta')}\right]. \quad (2.3)$$

While $\pi(\mathcal{D})$ disappears, $Z(\theta)$ remains in the Hastings ratio.

An early attempt to deal with this issue is to estimate $Z(\theta)$ beforehand with methods such as bridge sampling (Meng and Wong 1996) or path sampling (Gelman and Meng 1998), and then plug it into the acceptance ratio (e.g., Green and Richardson 2002). However, the Markov chain may not converge to the target distribution because of the estimation error, and computation is unnecessarily intensive. Another way is to replace the likelihood with pseudolikelihood, and carry out inference on the induced pseudo-posterior. Yet the pseudo-posterior is not the posterior of interest, and it may not be a proper distribution. There are also methods based on loopy belief propagation for Bayesian inference (Qi et al. 2005, Welling and Parise 2006). However, their convergence condition is not well understood.

Asymptotically correct MCMC methods were developed recently. Møller et al. (2006) ingeniously used an Auxiliary Variable Method (AVM) to replace the intractable normalizing problem with an exact sampling (or perfect sampling, Propp and Wilson 1996) problem from p_θ. Murray et al. (2006) proposed exchange algorithms to further improve its mixing. However, exact sampling is quite computationally expensive and is simply infeasible for many useful models (e.g., the ERG model (2.26) due to alternating k-triangle statistic (2.25)). Atchadé et al. (2008) developed an adaptive MCMC method which estimates $Z(\theta)$ within the MCMC procedure. Their algorithm does not require perfect sampling but remains computer-intensive, especially for large networks.

We propose a general framework of approximate MCMC algorithms for doubly-intractable distributions. The idea is built around the AVM of Møller et al. (2006) and the exchange algorithm of Murray et al. (2006). But we avoid the exact sampling from p_θ which is replaced by samples from a Markov kernel that is close to p_θ. The resulting algorithm is significantly faster and we prove that it is stochastically stable in the sense that it admits an invariant distribution. Furthermore the distance between the invariant distribution of the algorithm and the target posterior distribution can be controlled by the user and be arbitrary small (of course, at the expense of more computing time).

The idea formalizes a practice that is already common among practitioners dealing with doubly-intractable distributions. For example, it has been advocated recently by Caimo and Friel (2011) in the context of social network modeling. We will carefully analyze these approximate schemes and present some theoretical results to guide their use in practice.

2.1.2 Simulating Large Social Networks

An important tool in computing with ERG models is the ability to sample from p_θ, the ERG model itself for a given parameter value. This applies to both frequentist and Bayesian approaches. Various Metropolis-Hastings (MH) algorithms are often used for the task. The mixing of these algorithms depends upon the choice of proposals or transition kernels used to generate a new network. See Snijders (2002) for a detailed discussion and a variety of proposals. For small to moderate networks, existing proposals usually work well. However, for large and sparse networks, without proper adjustment of the proposals, these algorithms may suffer from slow mixing.

One problem documented in Morris et al. (2008) has to do with the widely used random sampler proposal for simulating ERG models. In random sampler, one randomly selects a dyad (i.e. a pair of nodes) to toggle (from edge to no edge, or vice versa). Sparse networks have much more disconnected dyads than connected ones, thus oftentimes the proposal will add an edge to the network. If the parameter corresponding to network density is negative (which is usually the case for real-world large networks), then most of the time the proposed networks will be rejected, and the Markov chain will stay on the same state for a long time. A refined proposal called TNT (tie-no tie, Morris et al. 2008) improves the mixing by holding a probability 0.5 of choosing an edge and proposing to delete it. The rest of the time it picks an empty dyad to toggle.

Another problem resides in models with higher order network statistics (as opposed to nodal or dyadic statistics), for instance the number of triangles and alternating k-triangles. In the simulation of sparse large networks, proposals such as random sampler or TNT cannot vary the values of these statistics efficiently. To construct a triangle from three disconnected nodes, all three pairwise dyads need to be chosen, which takes many steps to achieve. If the statistic value rarely changes, the corresponding parameter has limited effect on the simulation. It is not hard to imagine that the time to reach convergence of the ERG model will be quite long.

We are particularly interested in the mixing of ERG models involving triangles. Triangles are usually used to characterize transitivity, which is generally of interest in social network modeling. Recently, model degeneracy (Handcock 2003, Snijders et al. 2006) has caught researchers' attention, where the model places disproportionate probability on a small set of outcomes. An important statistic to help push back model degeneracy, alternating k-triangle (Snijders et al. 2006), also consists of triangles. Here we introduce a new M-H proposal called OTNT (open triangle-tie-no tie) to sample more efficiently from p_θ for large sparse social networks.

The remaining of the chapter is organized as follows. In section 2.2, we propose a general framework for approximate MCMC algorithms and explore the theoretical properties. In section 2.3, we describe the M-H move OTNT. A simulation study and real data analysis are presented in section 2.4. We give detailed proofs in the appendix.

2.2 Approximate MCMC Simulation from Doubly-intractable Distributions

Let $\{p_\theta(\cdot),\ \theta \in \Theta\}$ be a statistical model with sample space $\mathcal{X}$ and parameter space Θ. As in the exponential random graph model, we assume that $\mathcal{X}$ is a finite set. But for notational convenience, we will still write $\int_{\mathcal{X}} f(x)dx$ to denote the summation $\sum_{x\in\mathcal{X}} f(x)$. We also assume that for all $\theta \in \Theta$, $p_\theta(\cdot) = Z(\theta)^{-1} f_\theta(\cdot)$, for some normalizing constant $Z(\theta)$ which is intractable. The parameter space Θ is a subset of the p-dimensional Euclidean space $\mathbb{R}^p$ endowed with the Lebesgue measure.

Suppose that we observe a data set $\mathcal{D} \in \mathcal{X}$ and choose a prior density μ (wrt the Lebesgue measure on Θ). The posterior distribution is $\pi(\theta|\mathcal{D}) \propto f_\theta(\mathcal{D})\mu(\theta)/Z(\theta)$. As $Z(\theta)$ cannot be evaluated, direct simulation from $\pi(\theta|\mathcal{D})$ is infeasible. The goal is to propose Monte Carlo methods that are consistent and perform well in applications.

As mentioned in section 2.1.1, the acceptance probability (2.3) of a straightforward M-H to sample from $\pi(\theta|\mathcal{D})$ is intractable as it involves the intractable normalizing constants. Møller et al. (2006) circumvent the problem by introducing a joint distribution $\tilde{\pi}(\theta, x|\mathcal{D}) = \pi(\theta|\mathcal{D})\tilde{f}(x|\theta, \mathcal{D})$, on $\Theta \times \mathcal{X}$, where $\tilde{f}(x|\theta, \mathcal{D})$ is a tractable distribution (up to a normalizing constant that does not depend on θ). The AVM of Møller et al. (2006) generates a Markov chain $\{(\theta_n, X_n),\ n \geq 0\}$ on the joint space $\Theta \times \mathcal{X}$ that can be described as follows. Given $(\theta_n, X_n) = (\theta, x)$, one proposes a new point (θ', Y) from the proposal $Q(\theta, d\theta')p_{\theta'}(dy)$, where Q is a transition kernel on Θ. This proposed value is accepted with the M-H ratio

$$\min\left[1, \frac{\tilde{\pi}(\theta', Y)}{\tilde{\pi}(\theta, x)} \frac{Q(\theta', \theta)}{Q(\theta, \theta')} \frac{f_\theta(x)}{f_{\theta'}(Y)} \frac{Z(\theta')}{Z(\theta)}\right] = \min\left[1, \frac{f_{\theta'}(\mathcal{D})}{f_\theta(\mathcal{D})} \frac{\tilde{f}(Y|\theta', \mathcal{D})}{\tilde{f}(x|\theta, \mathcal{D})} \frac{\mu(\theta')}{\mu(\theta)} \frac{Q(\theta', \theta)}{Q(\theta, \theta')} \frac{f_\theta(x)}{f_{\theta'}(Y)}\right].$$

We see that $Z(\theta)/Z(\theta')$ cancels out. Clearly, the invariant distribution of this M-H algorithm is $\tilde{\pi}(\theta, x|\mathcal{D})$, whose θ-marginal is $\int_x \tilde{\pi}(\theta, x|\mathcal{D})dx = \pi(\theta|\mathcal{D})$.

There is a nice simplification on the AVM introduced by Murray et al. (2006), which obtained a valid Θ-valued MCMC sampler with target distribution $\pi(\cdot|\mathcal{D})$. Given θ_n, one generates θ' from $Q(\theta_n, d\theta')$ and Y from $p_{\theta'}(dy)$. Then θ' is accepted with probability

$$\alpha(\theta, \theta', Y) = \min\left[1, \frac{f_{\theta'}(D)\mu(\theta')Q(\theta', \theta)}{f_\theta(D)\mu(\theta)Q(\theta, \theta')} \frac{f_\theta(Y)}{f_{\theta'}(Y)}\right]. \tag{2.4}$$

The auxiliary variable Y is used to estimate the ratio $Z(\theta)/Z(\theta')$ by a one-sample importance sampling estimator,

$$\frac{Z(\theta)}{Z(\theta')} = \mathbb{E}_{\theta'}\left[\frac{f_\theta(Y)}{f_{\theta'}(Y)}\right] \approx \frac{f_\theta(Y)}{f_{\theta'}(Y)}.$$

Y is discarded after that. Although not a M-H algorithm, Murray et al. (2006) showed that this algorithm (called single variable exchange algorithm or SVEA) generates a Markov chain $\{\theta_n,\ n \geq 0\}$ on Θ with a transition kernel that is reversible (and thus invariant) wrt $\pi(\cdot|\mathcal{D})$. We will denote by T_0 the transition kernel of SVEA:

$$T_0(\theta, A) = \int_{A\times\mathcal{X}} \alpha(\theta,\theta',x)Q(\theta,d\theta')p_{\theta'}(dx) + \mathbf{1}_A(\theta)\int_{\Theta\times\mathcal{X}}[1-\alpha(\theta,\theta',x)]Q(\theta,d\theta')p_{\theta'}(dx),$$

where $\mathbf{1}_A$ is an indicator function of the set A, and is equal to 1 if $\theta \in A$ and 0 otherwise.

But as in the case of AVM, this algorithm also requires an exact simulation from $p_{\theta'}(\cdot)$, which in many models is impossible or can be excruciatingly slow to achieve.

We develop an approximate MCMC sampler for $\pi(\cdot|\mathcal{D})$, which does not require exact sampling. The basic idea is to replace the exact sampling from $p_{\theta'}$ by approximately sampling from a distribution close to $p_{\theta'}$.

We phrase this idea in a more general framework. For $\theta, \theta' \in \Theta$ and a positive integer κ, let $P_{\kappa,\theta,\theta'}$ be a transition kernel on $\mathcal{X}$ that depends measurably on θ, θ', s.t. for some measurable function $B_\kappa(x,\theta,\theta')$,

$$\|P_{\kappa,\theta,\theta'}(x,\cdot) - p_{\theta'}(\cdot)\|_{\mathrm{TV}} \leq B_\kappa(x,\theta,\theta').$$

We will naturally impose that $B_\kappa(x,\theta,\theta')$ converges to zero as $\kappa \to \infty$. In the above equation, the total variation distance between two (finite state space) probability mass functions is $\|\mu - \nu\|_{\mathrm{TV}} = \sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)|$. We consider the following algorithm that generates a Markov chain $\{(\theta_n, X_n),\ n \geq 0\}$ on $\Theta \times \mathcal{X}$.

Algorithm 2.2.1. At time 0, choose $\theta_0 \in \Theta$ and $X_0 \in \mathcal{X}$.

At time n, given $(\theta_n, X_n) = (\theta, x)$:

1. Generate $\theta' \sim Q(\theta_n, \cdot)$.

2. Set $Y_0 = x$. Generate $Y_\kappa \sim P_{\kappa,\theta,\theta'}(Y_0,\cdot)$. Compute

$$\alpha(\theta,\theta',Y_\kappa) = \min\Big[1, \frac{f_{\theta'}(\mathcal{D})\mu(\theta')Q(\theta',\theta_n)}{f_{\theta_n}(\mathcal{D})\mu(\theta_n)Q(\theta_n,\theta')}\frac{f_\theta(Y_\kappa)}{f_{\theta'}(Y_\kappa)}\Big].$$

3. With probability $\alpha(\theta,\theta',Y_\kappa)$, set $(\theta_{n+1},X_{n+1}) = (\theta',Y_\kappa)$; with probability $1 - \alpha(\theta,\theta',Y_\kappa)$, set $(\theta_{n+1},X_{n+1}) = (\theta,x)$.

Although this algorithm generates a Markov chain on $\mathcal{X} \times \Theta$, we are only interested in the marginal chain $\{\theta_n, n \geq 1\}$. Typically, at the end of the simulation, $\{X_n, n \geq 1\}$ is discarded.

The transition kernel of our approximate MCMC algorithm is $\bar{T}_\kappa$ given by

$$\begin{aligned}\bar{T}_\kappa\Big((\theta,x), A\times B\Big) &= \int_{A\times B}\alpha(\theta,\theta',y_\kappa)Q(\theta,d\theta')P_{\kappa,\theta,\theta'}(x,dy_\kappa)\\ &\quad +\mathbf{1}_{A\times B}(\theta,x)\int_{\Theta\times\mathcal{X}}[1-\alpha(\theta,\theta',y_\kappa)]Q(\theta,d\theta')P_{\kappa,\theta,\theta'}(x,dy_\kappa).\end{aligned}$$

The limiting kernel of the Markov kernel implemented by Algorithm 2.2.1 is $\bar{T}_\star$ defined on $\Theta \times \mathcal{X}$ as

$$\begin{aligned}\bar{T}_\star\Big((\theta,x), A\times B\Big) &= \int_{A\times B}\alpha(\theta,\theta',y)Q(\theta,d\theta')p_{\theta'}(dy)\\ &\quad +\mathbf{1}_{A\times B}(\theta,x)\int_{\Theta\times\mathcal{X}}[1-\alpha(\theta,\theta',y)]Q(\theta,d\theta')p_{\theta'}(dy).(2.5)\end{aligned}$$

Notice the similarity between $\bar{T}_\star$ and T_0 (the transition kernel of SVEA). $\bar{T}_\star$ is also very closely related to the AVM when $\tilde{f}(x|\theta,\mathcal{D}) = Z^{-1}(\theta)f_\theta(x)$, although this choice does not lead to a tractable algorithm. We have the following result which shows that the limiting kernel $\bar{T}_\star$ will sample correctly from the posterior distribution. Since $\bar{T}_\kappa$ is close to $\bar{T}_\star$ for large κ, this implies that Algorithm 2.2.1 approximately sample from the posterior distribution. We make this intuition rigorous below.

Proposition 2.2.1. *Suppose that the posterior distribution $\pi(\cdot|\mathcal{D})$ is the unique invariant distribution of T_0, and that $\bar{T}_\star$ admits an invariant distribution with density $\bar{\pi}_\star$ on $\Theta \times \mathcal{X}$. Then $\int \bar{\pi}_\star(x,\theta)dx = \pi(\theta|\mathcal{D})$. That is, the θ-marginal of $\bar{\pi}_\star$ is $\pi(\cdot|\mathcal{D})$.*

Proof. Let $\pi_\star$ be the θ-marginal of $\bar{\pi}_\star$. That is, $\pi_\star(\theta) = \int \bar{\pi}_\star(\theta, x)dx$. Notice that from the definitions, $\bar{T}_\star((\theta, x), A \times \mathcal{X}) = T_0(\theta, A)$. Now let $\{(\theta_n, X_n),\ n \geq 0\}$ be a Markov chain with transition kernel $\bar{T}_\star$ in stationarity. Therefore $(\theta_{n-1}, X_{n-1}) \sim \bar{\pi}_\star$ and for any measurable subset A of Θ,

$$\begin{aligned}\pi_\star(A) &= \mathbb{P}\left(\theta_n \in A, X_n \in \mathcal{X}\right) = \int \bar{\pi}_\star(d\theta, dx)\bar{T}_\star\left((\theta, x), A \times \mathcal{X}\right) \\ &= \int \bar{\pi}_\star(d\theta, dx)T_0(\theta, A) = \int \mathbb{P}\left(\theta_{n-1} \in d\theta\right) T_0(\theta, A) = \int \pi_\star(d\theta)T_0(\theta, A).\end{aligned}$$

In other words, $\pi_\star$ is also an invariant distribution for T_0. The result then follows easily from the assumptions. □

It is known (Murray et al. 2006) that the kernel T_0 has invariant distribution $\pi(\cdot|\mathcal{D})$, and in practice it is usually easy to construct T_0 such that it has a unique invariant distribution. We will see below that under some regularity conditions, $\bar{T}_\star$ admits an invariant distribution $\bar{\pi}_\star$. Thus in many applications the conclusion of Proposition 4.2.2 holds.

Building the Kernel $P_{\kappa,\theta,\theta'}$

It remains to describe how we build the approximating kernel $P_{\kappa,\theta,\theta'}$. A natural candidate is

$$P_{\kappa,\theta,\theta'}(x, \cdot) \overset{\text{def}}{=} P^{\kappa}_{\theta'}(x, \cdot), \tag{2.6}$$

for some integer κ, where $P_{\theta'}$ is a Markov kernel with invariant distribution $p_{\theta'}$. We refer to this choice of $P_{\kappa,\theta,\theta'}$ as the homogeneous instrument kernel.

If we assume that $X_n \sim p_\theta$ and p_θ and $p_{\theta'}$ are far apart, instead of κ iterations from $P_{\theta'}$ starting from X_n, a possibly useful alternative is to build a bridge of distributions between p_θ and $p_{\theta'}$. Let $\{f_t,\ 0 \leq t \leq 1\}$ be a smooth path of un-normalized densities between f_θ and $f_{\theta'}$ such that $f_0 = f_\theta$ and $f_1 = f_{\theta'}$. Define $Z(t) = \int_{\mathcal{X}} f_t(x)dx$ to be the normalizing constant for f_t, $p_t(\cdot) = f_t(\cdot)/Z(t)$, and P_t a transition kernel (e.g., M-H kernel) with invariant distribution p_t. We then pick κ intermediate distributions between p_θ and $p_{\theta'}$ by choosing $0 \leq t_1 \ldots \leq t_\kappa \leq 1$. Step 2 of Algorithm 2.2.1 is then implemented as follows. Given $Y_\ell, \ell = 0, \ldots, \kappa-1$, we generate $Y_{\ell+1}$ from $P_{t_\ell+1}(Y_\ell, \cdot)$. So the transition kernel from x to y_κ is

$$P_{\kappa,\theta,\theta'}(x, \cdot) \overset{\text{def}}{=} \int_{\mathcal{X}} P_{t_1}(x, dy_1) \ldots \int_{\mathcal{X}} P_{t_{\kappa-1}}\left(y_{t_{\kappa-2}}, dy_{t_{\kappa-1}}\right) P_{t_\kappa}\left(y_{t_{\kappa-1}}, \cdot\right). \tag{2.7}$$

For large κ, the intermediate distributions will be very close to each other, making the transition from one distribution to another easier. With properly chosen $t_1, \ldots, t_\kappa$, as κ increases, we expect Y_κ to converge to $p_{\theta'}$. We refer to this choice of $P_{\kappa,\theta,\theta'}$ as the nonhomogeneous instrument kernel, as each P_{t_ℓ} $(\ell = 1, \ldots, \kappa)$ is different.

One generic choice of path is the geometric path defined as $f_t(x) = f_\theta^{1-t}(x) f_{\theta'}^t(x)$. Throughout the thesis, we will use this path for the nonhomogeneous instrument kernel. For an ERG model in its canonical form, path defined through $\theta(t) = \theta + t(\theta' - \theta)$ with $f_t(x) = f_{\theta(t)}(x)$ is equivalent to the geometric path.

In the simulations that we have performed, the nonhomogeneous instrument kernel does not show any particular advantage over the homogeneous one. This might be due to the fact that the conditional distribution of X_n given (X_{n-1}, θ_{n-1}) in Algorithm 2.2.1 is not $p_{\theta_{n-1}}(\cdot)$. Nevertheless, we found the bridging idea interesting and it might be useful for some other models.

Choosing the Parameter κ

It is possible to run Algorithm 2.2.1 with κ held fixed. In this case, the invariant distribution of the algorithm will not be precisely the posterior distribution $\pi(\cdot|\mathcal{D})$. The larger κ, the longer the computing time, and the closer to $\pi(\cdot|\mathcal{D})$ the corresponding invariant distribution gets. The total variation norm between the two distributions is bounded in Theorem II.1. From experience, we found that κ in the range $100 - 200$ yields a reasonable approximating distribution for problems of the size of those considered here.

On the other hand, if more computing effort is allowed, we can make κ increase with n, using $\kappa = \kappa_n$ at the n-th iteration of Algorithm 2.2.1. For instance, $\kappa_n = \lceil d \log(n+1) \rceil$ or $\kappa_n = \lceil d\sqrt{n} \rceil$, where $\lceil a \rceil$ is the smallest integer that is no less than a. The advantages is intuitively clear but we also show in Theorem II.2 that with increasing κ, the limiting distribution of θ_n is precisely $\pi(\cdot|\mathcal{D})$. Furthermore, a strong law of large numbers and central limit theorem also hold. Of course, these limiting arguments with $\kappa_n \to \infty$ come at the expense of greater computing effort. This is nevertheless the approach taken below in our numeric examples, where we use $\kappa_n = \lceil d \log(1+n) \rceil$, with d mostly in the range of $10 - 20$.

2.2.1 Theory

We now give some theoretical justification of the method. First, we define some terminology and notations to be used in the entire section. For a transition kernel P, we denote by P^n, $n \geq 0$, its n-th iterate, with $P^0(x, A) = \mathbf{1}_A(x)$. The total variation normal between two measures λ and μ is defined as $\|\lambda - \mu\|_{\text{TV}} \overset{\text{def}}{=} \frac{1}{2} \sup_{\{|h| \leq 1\}} |\lambda(h) - \mu(h)|$, where h is any measurable function. $Ph(x) = \int P(x,y)h(y)dy$, and $\mu h = \int \mu(x)h(x)dx$. Let $\lfloor a \rfloor$ be the greatest integer not exceeding a.

For simplicity, we assume that Θ is a compact subset of a p-dimensional Euclidean space $\mathbb{R}^p$ equipped with its Borel σ-algebra. We recall also that the sample space $\mathcal{X}$ is finite. Furthermore, we introduce the following assumptions.

A1 There exist $\epsilon_\kappa > 0$ and a positive integer n_κ, s.t. for all $(\theta, x) \in \Theta \times \mathcal{X}$, and for all $(\theta_1, x_1), \ldots, (\theta_{n_\kappa}, x_{n_\kappa}) \in \Theta \times \mathcal{X}$,

$$P_{\kappa,\theta,\theta_1}(x, x_1) \cdots P_{\kappa,\theta_{n_\kappa-1},\theta_{n_\kappa}}(x_{n_\kappa-1}, x_{n_\kappa}) \geq \epsilon_\kappa. \tag{2.8}$$

The compactness of Θ makes it possible to check A1 for many examples. See Proposition 2.2.3-2.2.4 below.

A2 $B_\kappa(x, \theta, \theta')$ does not depend on x, and denote it by $B_\kappa(\theta, \theta')$. Moreover,

$$\sup_{\theta\in\Theta} \int_\Theta Q(\theta, \theta') B_\kappa(\theta, \theta') \to 0, \quad \text{as } \kappa \to \infty.$$

Theorem II.1. *Assume A1-A2 and suppose also that Q, μ and f_θ are positive and continuous functions of θ (and θ' in Q). Then $\bar{T}_\kappa$ and $\bar{T}_\star$ have unique invariant distributions denoted by $\bar{\pi}_\kappa$ and $\bar{\pi}_\star$ respectively. Furthermore,*

$$\|\bar{\pi}_\kappa - \bar{\pi}_\star\|_{\text{TV}} \leq C \sup_{\theta\in\Theta} \int_\Theta Q(\theta, \theta') B_\kappa(\theta, \theta') d\theta', \tag{2.9}$$

where C is some constant that does not depend on θ or κ.

Proof. The basic idea is to bound the distance between the $\bar{T}_\kappa$ and $\bar{T}_\star$. This distance depends on the total variation norm between $P_{\kappa,\theta,\theta'}$ and $p_{\theta'}$, which is bounded by $B_\kappa(\theta, \theta')$.

Firstly, we will show that both $\bar{T}_\kappa$ and $\bar{T}_\star$ satisfy a uniform minorization condition. This implies the existence of a unique invariant distribution, and the kernel converges to the invariant distribution uniformly at a geometric rate (see e.g., Meyn and Tweedie 2003, Theorem 16.02).

We start with $\bar{T}_\star$ which is easier. Under the assumptions of the theorem and the compactness of Θ, we can easily find a probability density λ_Q on Θ, s.t. $\epsilon_Q \overset{\text{def}}{=}$

$\inf_{\theta,\theta'\in\Theta}\frac{Q(\theta,\theta')}{\lambda_Q(\theta')} > 0$. Also, since Θ is compact and $\mathcal{X}$ is finite, there exists $C_\alpha \overset{\text{def}}{=} \inf_{\theta,\theta'\in\Theta}\inf_{y_\kappa\in\mathcal{X}}\alpha(\theta,\theta',y_\kappa) > 0$. Therefore

$$\bar{T}_\star\Big((\theta,x),A\times B\Big) \geq C_\alpha\epsilon_Q\int_A \lambda_Q(d\theta')\int_B p_{\theta'}(dx).$$

The argument is the same for $\bar{T}_\kappa$ but we need A1. We have

$$\bar{T}_\kappa\Big((\theta,x),A\times B\Big) \geq C_\alpha\epsilon_Q\int_{A\times B}\lambda_Q(d\theta')P_{\kappa,\theta,\theta'}(x,dy_\kappa).$$

Under A1,

$$\bar{T}_\kappa^{n_\kappa}\Big((\theta,x),A\times B\Big) \geq \left(C_\alpha|\mathcal{X}|\right)^{n_\kappa}\epsilon_Q\epsilon_\kappa\lambda_Q(A)\frac{|B|}{|\mathcal{X}|},$$

where $|B|$ denotes the number of elements of the set B. By Theorem 16.02 of (Meyn and Tweedie 2003), $\bar{T}_\kappa$ and $\bar{T}_\star$ admit invariant distributions called $\bar{\pi}_\kappa$ and $\bar{\pi}_\star$ respectively, and for all $(\theta,x)\in\Theta\times\mathcal{X}$ and $n\geq 0$,

$$\left\|\bar{T}_\kappa^n\Big((\theta,x),\cdot\Big) - \bar{\pi}_\kappa(\cdot)\right\|_{\text{TV}} \leq (1-\bar{\epsilon}_\kappa)^{\lfloor n/n_\kappa\rfloor}, \tag{2.10}$$

$$\left\|\bar{T}_\star^n\Big((\theta,x),\cdot\Big) - \bar{\pi}_\star(\cdot)\right\|_{\text{TV}} \leq (1-\bar{\epsilon}_\star)^{n}, \tag{2.11}$$

where $\bar{\epsilon}_\kappa \overset{\text{def}}{=} \left(C_\alpha|\mathcal{X}|\right)^{n_\kappa}\epsilon_Q\epsilon_\kappa$, $\bar{\epsilon}_\star = C_\alpha\epsilon_Q$. We will use (2.10) repeatedly in the proof Theorem II.2.

To prove (2.9), we first bound $\|\bar{T}_\kappa - \bar{T}_\star\|_{\text{TV}}$.

$$\begin{aligned}
&\left\|\bar{T}_\kappa\Big((\theta,x),\cdot\Big) - \bar{T}_\star\Big((\theta,x),\cdot\Big)\right\|_{\text{TV}} \\
\leq\ & \sup_{|h|\leq 1}\int_\Theta Q(\theta,d\theta')\left|\int_{\mathcal{X}}\left[P_{\kappa,\theta,\theta'}(x,dy_\kappa) - p_{\theta'}(dy_\kappa)\right]\alpha(\theta,\theta',y_\kappa)h(\theta',y_\kappa)\right| && (2.12)\\
\leq\ & M_\alpha\int_\Theta Q(\theta,\theta')B_\kappa(\theta,\theta')d\theta', && (2.13)
\end{aligned}$$

where $M_\alpha \overset{\text{def}}{=} \sup_{\theta,\theta'\in\Theta}\sup_{y_\kappa\in\mathcal{X}}\alpha(\theta,\theta',y_\kappa)$. Finally,

$$\begin{aligned}\|\bar\pi_\kappa - \bar\pi_\star\|_{\mathrm{TV}} &= \frac{1}{2}\sup_{|h|\leq 1}|\bar\pi_\kappa(\bar T_\star^n - \bar\pi_\star)h + \sum_{j=1}^{n}\bar\pi_\kappa(\bar T_\kappa - \bar T_\star)(\bar T_\star^{j-1} - \bar\pi_\star)h|,\\ &\leq \sum_{j\geq 0}(1-\bar\epsilon_\star)^j \sup_{\theta\in\Theta,x\in\mathcal{X}}\|\bar T_\kappa - \bar T_\star\|_{\mathrm{TV}} \qquad \text{by letting } n\to\infty \text{ and (2.11)},\\ &\leq C\sup_{\theta\in\Theta}\int_\Theta Q(\theta,\theta')B_\kappa(\theta,\theta')d\theta', \qquad \text{by (2.13), where } C = \frac{M_\alpha}{1-\bar\epsilon_\star}.\end{aligned}$$

□

Let us denote $\pi_\kappa(\cdot) \overset{\text{def}}{=} \bar\pi_\kappa(\cdot\times\mathcal{X})$ the θ-marginal of $\bar\pi_\kappa$. Under the assumptions of the Theorem, it is straightforward to check that the transition kernel T_0 of the SVEA algorithm of (Murray et al. 2006) has $\pi(\cdot|\mathcal{D})$ as unique invariant distribution. Therefore by Proposition 4.2.2 and the bound (2.9), we conclude that

$$\|\pi_\kappa(\cdot) - \pi(\cdot|\mathcal{D})\|_{\mathrm{TV}} \leq C\sup_{\theta\in\Theta}\int_\Theta Q(\theta,\theta')B_\kappa(\theta,\theta')d\theta'. \tag{2.14}$$

Remark 2.2.1. *The bound (2.14) tells us that if we run Algorithm 2.2.1 under the conditions of the theorem with κ held fix, then the marginal distribution of θ_n will converge to the distribution π_κ, which is within $C\sup_{\theta\in\Theta}\int_\Theta Q(\theta,\theta')B_\kappa(\theta,\theta')d\theta'$ of the posterior distribution.*

Suppose now that we run Algorithm 2.2.1 using $\kappa = \kappa_n$ at the n-th iteration for some nondecreasing sequence $\{\kappa_n\}$, s.t. $\kappa_n \to \infty$. Since the transition kernel now is $\bar T_{\kappa_n}$, different for each n, then Algorithm 2.2.1 generates a nonhomogeneous Markov chain $\{(\theta_n, X_n),\ n\geq 0\}$. Given $\mathcal{F}_n = \sigma\{(\theta_0,X_0),\dots,(\theta_n,X_n)\}$, the conditional distribution of (θ_{n+1},X_{n+1}) is $\bar T_{\kappa_n}\Big((\theta_n,X_n), A\times B\Big)$. We can say the following about this process.

Theorem II.2. *Suppose that in Algorithm 2.2.1, we set $\kappa=\kappa_n$ and let $\{(\theta_n,X_n),\ n\geq 0\}$ be the resulting nonhomogeneous Markov chain. Assume A1-A2. Suppose also that*

$\inf_{\kappa\geq 1}\epsilon_\kappa > 0$, *and* $\sup_{\kappa\geq 1} n_\kappa < \infty$. *Denote the distribution of* (θ_n, X_n) *by* $\mathscr{L}_{(\theta_n,X_n)}$. *We have:*

(1)

$$\lim_{n\to\infty} \left\| \mathscr{L}_{(\theta_n,X_n)} - \bar{\pi}_\star \right\|_{\mathrm{TV}} = 0. \tag{2.15}$$

(2) For any measurable function $h : \Theta \times \mathcal{X} \to \mathbb{R}$,

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} h(\theta_i, X_i) = \bar{\pi}_\star h, \quad a.s., \tag{2.16}$$

where $\bar{\pi}_\star h = \int_\Theta \int_{\mathcal{X}} \bar{\pi}_\star(\theta, x) h(\theta, x) d\theta dx$. *Furthermore, if*

$$\sum_{i=1}^{\infty} i^{-1/2} \int_\Theta Q(\theta,\theta') B_{\kappa_i}(\theta,\theta') d\theta' < \infty, \text{ then} \tag{2.17}$$

$$n^{-1/2} \sum_{i=1}^{n} \Big(h(\theta_i, X_i) - \bar{\pi}_\star h \Big) \xrightarrow{d} N\Big(0, \sigma^2(h)\Big), \quad \text{as } n \to \infty, \tag{2.18}$$

where

$$\sigma^2(h) = Var\Big(h(\theta_0, X_0)\Big) + 2\sum_{i\geq 1} Cov\Big(h(\theta_0, X_0), h(\theta_i, X_i)\Big),$$

where the Var and Cov are computed under the assumption that $\{(\theta_n, X_n), n \geq 0\}$ *is a stationary Markov chain with kernel* $\bar{T}_\star$ *and invariant distribution* $\bar{\pi}_\star$.

Proof. We will first prove part (1) of the theorem. For $\kappa \geq 1$, we define $D_\kappa \overset{\text{def}}{=} C \sup_{\theta\in\Theta} \int_\Theta Q(\theta,\theta') B_\kappa(\theta,\theta') d\theta'$. We recall from the proof of Theorem II.1 the quantity $\bar{\epsilon}_\kappa \overset{\text{def}}{=} (C_\alpha \epsilon_Q |\mathcal{X}|)^{n_\kappa} \epsilon_\kappa$. By assumption, $\rho \overset{\text{def}}{=} \sup_{\kappa\geq 1} 1 - \bar{\epsilon}_\kappa < 1$. Let $n_0 = \sup_{\kappa\geq 1} n_\kappa < \infty$.

For any $1 \leq j \leq n$ and measurable function $|h(\theta, x)| \leq 1$, we have

$$\left| \mathbb{E}\Big(h(\theta_n, X_n)\Big) - \bar{\pi}_\star(h) \right| = \left| \mathbb{E}\mathbb{E}\Big(h(\theta_n, X_n) | \theta_{j-1}, X_{j-1}\Big) - \bar{\pi}_\star(h) \right|,$$

so we first work with the conditional distribution of $\mathbb{E}\left(h(\theta_n, X_n)|(\theta_{j-1}, X_{j-1})\right)$.

$$\begin{aligned}
&\sup_{|h|\leq 1} \frac{1}{2}\left|\mathbb{E}\Big(h(\theta_n, X_n)|\theta_{j-1}, X_{j-1}\Big) - \bar{\pi}_\star(h)\right| \\
\leq\ & \|\bar{T}_{\kappa_j}\ldots\bar{T}_{\kappa_n} - \bar{T}_{\kappa_n}^{n-j+1}\|_{\mathrm{TV}} + \|\bar{T}_{\kappa_n}^{n-j+1} - \bar{\pi}_{\kappa_n}\|_{\mathrm{TV}} + \|\bar{\pi}_{\kappa_n} - \bar{\pi}_\star\|_{\mathrm{TV}}.
\end{aligned}$$

By (2.10) and (2.9), the last two terms are bounded by $\rho^{\lfloor\frac{n-j+1}{n_0}\rfloor} + D_{\kappa_n}$.

According to the following decomposition

$$(\bar{T}_{\kappa_j}\ldots\bar{T}_{\kappa_n} - \bar{T}_{\kappa_n}^{n-j+1})h = \sum_{\ell=j-1}^{n-1} \bar{T}_{\kappa_j}\ldots\bar{T}_{\kappa_\ell}(\bar{T}_{\kappa_{\ell+1}} - \bar{T}_{\kappa_n})(\bar{T}_{\kappa_n}^{n-\ell-1} - \bar{\pi}_{\kappa_n})h, \tag{2.19}$$

we bound the first term by $\sum_{\ell=j-1}^{n-1} \rho^{\lfloor\frac{n-j+1}{n_0}\rfloor}\|\bar{T}_{\kappa_{\ell+1}} - \bar{T}_{\kappa_n}\|_{\mathrm{TV}}$.

Applying (2.13), we get $\|\bar{T}_{\kappa_{\ell+1}} - \bar{\pi}_\star\|_{\mathrm{TV}} \leq D_{\kappa_{\ell+1}}$ and $\|\bar{T}_{\kappa_n} - \bar{\pi}_\star\|_{\mathrm{TV}} \leq D_{\kappa_n}$. Therefore $\|\bar{T}_{\kappa_{\ell+1}} - \bar{T}_{\kappa_n}\|_{\mathrm{TV}} \leq D_{\kappa_{\ell+1}} + D_{\kappa_n}$. Finally, we have

$$\sup_{|h|\leq 1} \frac{1}{2}\left|\mathbb{E}\left(h(\theta_n, X_n)\right) - \pi_\star(h)\right| \leq \sum_{\ell=j-1}^{n-1} \rho^{\lfloor\frac{n-j+1}{n_0}\rfloor}(D_{\kappa_{\ell+1}} + D_{\kappa_n}) + \rho^{\lfloor\frac{n-j+1}{n_0}\rfloor} + D_{\kappa_n}.$$

Let $j = \lceil n/2 \rceil$, under A2, this bound will go to 0 as $n \to \infty$, which completes the proof of part (1).

To prove part (2), we start with some general properties of Markov kernels, and prove the theorem afterwards. Let T_1, T_2 be two Markov kernels on a measurable general state space $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. Suppose $T_i, i = 1, 2$ satisfies a uniform minorization condition: there exist $\epsilon_i > 0$, probability measure λ_i, and a positive integer n_i, s.t. $T_i^n(w, B) \geq \epsilon_i^{\lfloor n/n_i \rfloor}\lambda_i(B), \forall B \in \mathcal{B}(\mathbb{T}), w \in \mathbb{T}$. Denote the corresponding invariant distribution by π_i (a general distribution on $\mathbb{T}$).

Proposition 2.2.2. *We have*

a.

$$\|\pi_1 - \pi_2\|_{\mathrm{TV}} \leq \frac{n_2}{\epsilon_2}\|T_1 - T_2\|_{\mathrm{TV}}. \tag{2.20}$$

b. *Let* $M_{h_i} \overset{\text{def}}{=} \sup_{w\in\mathbb{T}} |h_i(w)| < \infty$ *and* $\pi_i(h_i) = 0$, $i = 1, 2$ *and define*

$$g_i = \sum_{k\geq 0} T_i^k h_i.$$

Then $\sup_{w\in\mathbb{T}} |g_i| \leq M_{h_i} n_i/\epsilon_i$ *and*

$$\sup_{w\in\mathbb{T}} |T_1 g_1 - T_2 g_2| \leq K \left(\sup_{w\in\mathbb{T}} |h_1 - h_2| + \|T_1 - T_2\|_{\text{TV}} \right), \tag{2.21}$$

where $K = \max\left(1 + \frac{n_1}{\epsilon_1}, \frac{M_{h_2} n_1 n_2}{\epsilon_1 \epsilon_2}(1 + \frac{n_2}{\epsilon_2})\right)$.

Proof. We first prove proposition a. We have

$$\begin{aligned} \|\pi_1 - \pi_2\|_{\text{TV}} &\leq \|\pi_1 - T_1^n\|_{\text{TV}} + \|T_1^n - T_2^n\|_{\text{TV}} + \|T_2^n - \pi_2\|_{\text{TV}} \\ &\leq (1-\epsilon_1)^{n/n_1} + \|T_1^n - T_2^n\|_{\text{TV}} + (1-\epsilon_2)^{n/n_2}. \end{aligned} \tag{2.22}$$

$$\begin{aligned} \|T_1^n - T_2^n\|_{\text{TV}} &= \sup_{|h|\leq 1} \left| \sum_{j=0}^{n-1} T_1^j (T_1 - T_2)(T_2^{n-j-1} h - \pi_2 h) \right| \\ \leq \ & \sum_{j=0}^{n-1} \sup_{|h|\leq 1} \left| T_1^j (T_1 - T_2) h \right| (1-\epsilon_2)^{(n-j-1)/n_2} \\ \leq \ & \sum_{j=0}^{n-1} \|T_1 - T_2\|_{\text{TV}} (1-\epsilon_2)^{(n-j-1)/n_2}. \end{aligned} \tag{2.23}$$

Now plugging (2.23) into (2.22) and letting $n \to \infty$, we complete the proof of proposition a.

For proposition b, firstly,

$$\sup_{w\in\mathbb{T}} |g_i| \leq \sup_{w\in\mathbb{T}} \left| \sum_{k\geq 0} (T_i^k - \pi) h_i \right| \leq M_{h_i} \sum_{k\geq 0} (1-\epsilon_i)^{k/n_i} \leq M_{h_i} n_i/\epsilon_i.$$

Then we have

$$\sup_{w\in\mathbb{T}} |T_1 g_1 - T_2 g_2| \leq \sup_{w\in\mathbb{T}} |g_1 - g_2| + \sup_{w\in\mathbb{T}} |h_1 - h_2|. \tag{2.24}$$

Now we bound the first part of (2.24).

$$\begin{aligned}\sup_{w\in\mathbb{T}}|g_1-g_2| &= \sup_{w\in\mathbb{T}}\left|\sum_{k\geq 0}(T_1^k-\pi_1)h_1-(T_2^k-\pi_2)h_2\right| \\ \leq\ \sup_{w\in\mathbb{T}}\left|\sum_{k\geq 0}(T_1^k-\pi_1)(h_1-h_2)\right| &+ \sup_{w\in\mathbb{T}}\left|\sum_{k\geq 0}\Big[(T_1^k-\pi_1)-(T_2^k-\pi_2)\Big]h_2\right|\end{aligned}$$

The first part of (2.25) can be bounded by $n_1\sup_{w\in\mathbb{T}}|h_1-h_2|/\epsilon_1$. The second part is further decomposed as

$$\begin{aligned} & \sup_{w\in\mathbb{T}}\left|\sum_{k\geq 1}\sum_{j=0}^{k-1}(T_1^j-\pi_1)\Big[(T_1-\pi_1)-(T_2-\pi_2)\Big](T_2^{k-j-1}-\pi_2)h_2\right| \\ = & \sup_{w\in\mathbb{T}}\left|\sum_{j\geq 0}\sum_{k'\geq 0}(T_1^j-\pi_1)\Big[(T_1-\pi_1)-(T_2-\pi_2)\Big](T_2^{k'}-\pi_2)h_2\right| \\ \leq & \sup_{w\in\mathbb{T}}\left|\sum_{j\geq 0}(T_1^j-\pi_1)\Big[(T_1-\pi_1)-(T_2-\pi_2)\Big]g_2\right| \\ \leq & \frac{M_{h_2}n_1n_2}{\epsilon_1\epsilon_2}(1+\frac{n_2}{\epsilon_2})\|T_1-T_2\|_{\mathrm{TV}}\end{aligned}$$

Plugging all the bounds we obtained so far into (2.24), (2.21) is proved. □

Now we are ready to prove part (2) of Theorem II.2.

$$n^{-1}\sum_{i=1}^n h(\theta_i,X_i)-\bar{\pi}_\star(h)=n^{-1}\sum_{i=1}^n h(\theta_i,X_i)-\bar{\pi}_{\kappa_i}(h)+n^{-1}\sum_{i=1}^n\bar{\pi}_{\kappa_i}(h)-\bar{\pi}_\star(h)$$

So we will prove both parts on the right hand side converge to 0 a.s., respectively.

First, we show $\lim_{n\to\infty}n^{-1}\sum_{i=1}^n h(\theta_i,X_i)-\bar{\pi}_{\kappa_i}(h)=0,$ a.s. Write $\bar{h}_i=h-\bar{\pi}_{\kappa_i}(h)$ and for any $\theta\in\Theta,x\in\mathcal{X}$, let $g_i(\theta,x)\stackrel{\text{def}}{=}\sum_{k\geq 0}\bar{T}_{\kappa_i}^k\bar{h}_i(\theta,x)$. Since g_i satisfies $g_i-\bar{T}_{\kappa_i}g_i=\bar{h}_i$, we have

$$\begin{aligned}\sum_{i=1}^n\bar{h}_i(\theta_i,X_i) &= \sum_{i=1}^n g_i(\theta_i,X_i)-\bar{T}_{\kappa_i}g_i(\theta_{i-1},X_{i-1})+\big(\bar{T}_{\kappa_1}g_1(\theta_0,X_0)-\bar{T}_{\kappa_n}g_n(\theta_n,x_n)\big) \\ &\quad+\sum_{i=2}^n\bar{T}_{\kappa_i}g_i(\theta_{i-1},X_{i-1})-\bar{T}_{\kappa_{i-1}}g_{i-1}(\theta_{i-1},X_{i-1}) \\ &= \sum_{i=1}^n g_i(\theta_i,X_i)-\bar{T}_{\kappa_i}g_i(\theta_{i-1},X_{i-1})+R_{1,n}.\end{aligned}$$

By Proposition 2.2.2, we have $\sup_{i\geq 1}|g_i| < \infty$. Combining (**??**), (2.21), and Kronecker's lemma (Hall and Heyde 1980), it follows that $n^{-1}R_{1,n} \to 0$. Let $D_i = g_i(\theta_i, X_i) - \bar{T}_{\kappa_i} g_i(\theta_{i-1}, X_{i-1})$. It is easy to see that each D_i is a martingale difference. By Proposition 2.2.2 and Minkowski's inequality, we have $\sum_{i=1}^{\infty} i^{-2}\mathbb{E}D_i^2 < \infty$. Therefore, $n^{-1}\sum_{i=1}^{n} D_i \to 0$, a.s. (Chow 1960). This proves that $n^{-1}\sum_{i=1}^{n} h(\theta_i, X_i) - \bar{\pi}_{\kappa_i} h \to 0$, a.s.

Next we show $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \bar{\pi}_{\kappa_i}(h) - \bar{\pi}_\star h = 0$. By Theorem II.1, we have

$$|\bar{\pi}_{\kappa_i} h - \bar{\pi}_\star h| \leq 2M_h D_{\kappa_i}.$$

Kronecker's lemma concludes the proof of the strong law of large numbers.

For central limit theorem, again we have the martingale approximation

$$\sum_{i=1}^{n}\Big(h(\theta_i, X_i) - \bar{\pi}_\star h\Big) = \sum_{i=1}^{n} g_i(\theta_i, x_i) - \bar{T}_{\kappa_i} g_i(\theta_{i-1}, x_{i-1}) + R_{2,n},$$

where $R_{2,n} = R_{1,n} + \sum_{i=1}^{n}(\bar{\pi}_{\kappa_i} h - \bar{\pi}_\star h)$. By (2.17) and Kronecker's lemma, we have $n^{-1/2}R_{2,n}$ converges to 0. The term $\sum_{i=1}^{n} D_i$ is a triangular martingale array. Since $\sup_{i\geq 1}\sup_{\theta,x} D_i^2 < \infty$, the conditional Lindeberg condition holds.

To show $n^{-1}\sum_{i=1}^{n} \bar{T}_{\kappa_i} D_i^2 \to \sigma^2(h)$ a.s., we use a similar logic as above. But instead of $h(\theta, x)$, we prove a strong law of large numbers for function $q_i(\theta, x) = \bar{T}_{\kappa_i} D_i^2 = \bar{T}_{\kappa_i} g_i^2(\theta, x) - \big(\bar{T}_{\kappa_i} g_i(\theta, x)\big)^2$. Since $\sup_{i\geq 1} M_{h_i} < \infty$, we can show

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} q_i(\theta_{i-1}, X_{i-1}) - \bar{\pi}_i q_i = 0, \quad \text{a.s.}$$

Moreover, it is not hard to show that for any $k \geq 1$,

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \bar{\pi}_{\kappa_i}(h\bar{T}_{\kappa_i}^k h) - \bar{\pi}_\star(h\bar{T}_\star^k h) = 0, \quad \text{a.s.}$$

Therefore, we have

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \bar{T}_{\kappa_i} g_i^2(\theta_{i-1}, X_{i-1}) - \big(\bar{T}_{\kappa_i} g_i(\theta_{i-1}, X_{i-1})\big)^2 = \sigma^2(h) \quad \text{a.s.},$$

where $\sigma^2(h)$ is defined in Theorem II.2. By Corollary 3.1 in Hall and Heyde (1980) we deduce that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\theta_i, X_i) - \bar{\pi}_\star h\right) \xrightarrow{d} N\left(0, \sigma^2(h)\right).$$

□

By marginalization, the above theorem implies the following about the marginal process $\{\theta_n,\ n \geq 1\}$ generated from Algorithm 2.2.1 with $\kappa = \kappa_n$: under the above assumptions, $\lim_{n\to\infty} \|\mathscr{L}_{\theta_n} - \pi_\star\|_{\mathrm{TV}} = 0$; $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} h(\theta_i) = \pi_\star h,\ a.s.$; and $n^{-1/2}\sum_{i=1}^{n}\left(h(\theta_i) - \pi_\star h\right) \xrightarrow{d} N\left(0, \sigma^2(h)\right),\ as\ n \to \infty$, where $\sigma^2(h)$ is as in Theorem II.2 with $h(\theta, x) = h(\theta)$.

Let us now show that the assumptions of the above theorems hold for the ERG model.

Proposition 2.2.3. *For an ERG model in the form of (2.1) for a binary symmetric network, if the parameter space is compact, $P_{\kappa,\theta,\theta'}$ in Algorithm 2.2.1 is the homogeneous instrument kernel (2.6), and $P_{\theta'}$ is the random sampler, then Theorems II.1 and II.2 hold.*

Proof. Please refer to section 2.3 for a detailed description of random sampler. Given x, let $N(x) \overset{\text{def}}{=} \{y$: triangular matrices y and x differ by one and only one entry$\}$. Assume that the network consists of n people, so there are $m = n(n-1)/2$ entries. For any $\theta \in \Theta$, the kernel of the random sampler is $P_\theta(x, y) = \frac{1}{m}\min[1, \frac{f_\theta(y)}{f_\theta(x)}]$ for $y \in N(x)$, and is 0 for all other $y \neq x$; $P_\theta(x, x) = 1 - \sum_{y\in N(x)} \frac{1}{m}\min[1, \frac{f_\theta(y)}{f_\theta(x)}]$. By the compactness of Θ, there exists $\epsilon_0 > 0$ such that for any $x \in \mathcal{X}$, any $y \in N(x)$, and any $\theta \in \Theta$, $P_\theta(x, y) \geq \epsilon_0$. For any $x, y \in \mathcal{X}$, one can change x to y by toggling only one dyad at a time in a maximum of m moves. Thus $P_\theta^m(x, y) \geq \epsilon_0^m$ for all $x, y \in \mathcal{X}$ and all $\theta \in \Theta$.

With the same idea, we can show that $P_{\kappa,\theta,\theta'} = P_{\theta'}^{\kappa}$ satisfies A1 with $n_\kappa = \lceil m/\kappa \rceil$, $\epsilon_\kappa = \epsilon_0^{n_\kappa \kappa}$. It follows directly that $\sup_\kappa n_\kappa \leq m < \infty$ and $\inf_\kappa \epsilon_\kappa = \epsilon_0^{2m-2} > 0$. Recall $B_\kappa(x,\theta,\theta') = \|P_{\theta'}^{\kappa} - p_{\theta'}\|_{\mathrm{TV}}$. By Meyn and Tweedie (2003) Theorem 16.02, $B_\kappa(x,\theta,\theta') \leq (1-\epsilon_0)^{\lfloor \kappa/m \rfloor}$, which does not depend on x. A2 and (2.17) can be then verified. All other conditions are trivial to check. □

Proposition 2.2.4. *For an ERG model in the form of (2.1) for a binary symmetric network, if the parameter space is compact, $P_{\kappa,\theta,\theta'}$ in Algorithm 2.2.1 is the nonhomogeneous instrument kernel (2.7), and both P_θ and $P_{\theta'}$ are the random sampler, then Theorems II.1 and II.2 hold, except for the central limit theorem result (2.18). If $\sum_{n\geq 1} n^{-1/2}\kappa_n^{-1} < \infty$ then (2.18) holds as well.*

Proof. The same arguments as above show that A1 holds; in fact with the same constants ϵ_κ and n_κ. Deriving the bound $B_\kappa(x,\theta,\theta')$ requires some work. We have

$$\begin{aligned}
B_\kappa(x,\theta,\theta') &= \frac{1}{2}\sup_{|h(x)|\leq 1}\left|P_{t_1}\dots P_{t_{\kappa-1}}P_{\theta'}h - p_{\theta'}h\right| \\
&\leq \frac{1}{2}\sup_{|h|\leq 1}\left|\sum_{j=0}^{\kappa-1} P_{t_1}\dots P_{t_j}(P_{t_{j+1}} - P_{\theta'})(P_{\theta'}^{\kappa-j-1} - p_{\theta'})h\right| + \|P_{\theta'}^{\kappa} - p_{\theta'}\|_{\mathrm{TV}} \\
&\leq \frac{1}{\kappa}\sum_{i=0}^{\kappa-1} i(1-\epsilon_0 m)^{\lfloor i/m \rfloor} + (1-\epsilon_0 m)^{\lfloor \kappa/m \rfloor}.
\end{aligned}$$

Thus for large κ, $B_\kappa(x,\theta,\theta') = O(1/\kappa)$ and A2 holds. And (2.17) is valid if $\sum_{n\geq 1} n^{-1/2}\kappa_n^{-1} < \infty$. □

Remark 2.2.2. *While the idea of bridging between p_θ and $p_{\theta'}$ is appealing, it is theoretically challenging to obtain a good idea of $B_\kappa(x,\theta,\theta')$. As shown above, $B_\kappa(x,\theta,\theta') = O(1/\kappa)$. However, the bound is quite loose, and we believe that the chain converges at a much faster rate.*

2.3 Metropolis-Hastings Moves on Spaces of Large Networks

In this section, we restrict attention to the ERG model (2.1). Efficient sampling from the ERG distribution p_θ is important for successfully fitting the ERG model to data. For example, our approximate MCMC algorithm as applied to the ERG model is sensitive to the mixing of the $\mathcal{X}$-move instrument kernel $P_{\kappa,\theta,\theta'}$. For the two instrument kernels specified in section 2.2, their mixing depends further on that of $P_{\theta'}$ and P_{t_ℓ}, respectively. When the random sampler is used to build these kernels, the resulting algorithm does not perform well on large sparse networks. As mentioned in the introduction, the same issue arises in computing the MLE for the model.

Random sampler works as follows. Given the current network y, we choose two nodes uniformly at random, and toggle their connection status (i.e. break the edge if two people are connected, otherwise build an edge). All other edges remain the same. This new network y' which differs only by one edge from y, is accepted with probability $\min[1, \frac{f(y')}{f(y)}]$, where $f(y)$ is the un-normalized density of an ERG model, and we ignore parameters here for notation lucidity.

As discussed in the introduction, its mixing can be slow due to the sparsity of the real networks in practice. Most dyads in sparse networks are disconnected. So for random sampler, there is a high probability that an empty dyad will be chosen to toggle. But the parameter controlling network density is usually negative for sparse networks, which results in a high chance of rejecting the proposal. A refined kernel is TNT (Morris et al. 2008), where with probability .5 an existing edge is selected to toggle, and with probability .5 a disconnected dyad is selected to toggle. Then we accept it with probability $\min[1, \frac{f(y')}{f(y)}]$. This way we are able to explore the network space more efficiently.

When the model involves statistics that depict higher order structures (as opposed to nodal or dyadic structure) such as triangles, simple proposals may be inefficient for large networks, in the sense that the statistic value will not vary easily. For instance, in a sparse large network, it will take many steps of toggling dyads before a triangle is formed.

Transitivity is an important characteristic to study in social network modeling, as it captures the idea that people having common friends tent to become friends. It is often quantified by statistics involving triangles, e.g., the number of triangles and alternating k-triangle. Recent studies (Snijders 2002, Handcock 2003) found that some specifications of the ERG model will lead to degenerate distributions, meaning that the distributions put most of the probability mass on a small subset of the sample space. The number of triangles could contribute to model degeneracy in some cases. Alternating k-triangle statistic (Snijders et al. 2006) was introduced to prevent model degeneracy. For a network of n people, we introduce its analytical expression given in Hunter and Handcock (2006),

$$v(x;\theta) = e^{\theta} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\theta})^k\} D_k(x),$$

where $D_k(x)$ is the number of connected dyads that share exactly k neighbors in common. The intuition is that as k increases, we less favor the formation of k-triangles (a k-triangle consists of k triangles that share one common edge) to prevent the edge explosion.

Neither random sampler nor TNT can help form triangles efficiently. In both proposals, a dyad is selected uniformly at random. For sparse networks, most of the time, a triangle is built from three disconnected nodes. Then all three pairwise dyads need to be chosen, which will take a long time. An easier way is to choose an open triangle (or two-star) and complete the missing edge. Here we propose a new M-H kernel based on this idea. Notice that slow mixing does not only exist for models with triangles, but also with many higher order structures, which are hard to form or/and destroy in sparse networks. We take triangles as an example for its popularity in social network modeling. As long as readers are aware of the problem, it is not a difficult task to modify existing proposals accordingly.

2.3.1 OTNT Proposal

The kernel is built on TNT, and we take one step further to complete triangles with the help of k-stars ($k \geq 2$). We assume the network is an n by n binary symmetric network, but the same idea can be carried over to more complicated networks. Given the current network y, we propose a new y' by adding or deleting an edge as follows. With probability w_1 we pick a disconnected dyad uniformly at random and propose adding the edge between them. With probability w_2, we randomly pick a connected dyad and propose breaking the connection. The rest of the time, we randomly choose a node of degree at least two, i.e., k-star ($k \geq 2$), and then randomly select two of its neighbors. If the three nodes form an open triangle in y, i.e. the neighbors are disconnected, then we propose connecting them to form the triangle. Otherwise we do not change the network. The basic idea is to complete open triangles, and hence we name this kernel OTNT (open triangle-tie-no tie). When $w_1 + w_2 = 1$, it turns back to TNT.

We accept y' with probability $\min[1, \frac{f(y')q(y',y)}{f(y)q(y,y')}]$, where q is the transition kernel of OTNT. For the two instrument kernels in section 2.2, $f(y)$ refers to $f_{\theta'}$ and f_{t_ℓ}, respectively. $q(y, y') \overset{\text{def}}{=} w_1 q_1(y, y') + w_2 q_2(y, y') + (1 - w_1 - w_2) q_3(y, y')$. Define

$y' - y = 1$, if y' has one more edge than y and they only differ by this edge. We have

$$q_1(y, y') = \frac{1}{\frac{n(n-1)}{2} - \mathrm{E}(y)}, \text{ if } y' - y = 1, \text{ and 0 otherwise.}$$

$$q_2(y, y') = \frac{1}{\mathrm{E}(y)}, \text{ if } y - y' = 1, \text{ and 0 otherwise.}$$

$$q_3(y, y') = \frac{1}{m} \sum_{k=1}^{l} \frac{1}{\binom{n_k}{2}}, \text{ if } y' - y = 1,$$

where n is the number of nodes, $E(y)$ is the number of edges in the network, and m is the number of k-stars ($k \geq 2$) in y. For q_3, assuming y and y' differ by edge (i, j), then l is the number of common neighbors i and j have, and n_k is the degree of their k-th common neighbor. For q_3, if $y' = y$, we always accept y' and do not need to calculate $q_3(y, y')$ nor $q(y, y')$. For all other y', $q_3(y, y') = 0$.

In section 2.4, we will compare the performance of random sampler, TNT and OTNT on a large social network example, and the result is in favor of OTNT. The key point is to form triangles more often with structures that can be tracked easily, such as k-stars in OTNT.

2.4 Numerical Examples

In this section, we showcase the efficiency and accuracy of our approximate MCMC algorithms in Ising model and conditional random field. Although these models are not ERG models, they both suffer from the intractable normalizing constant problem. We also apply the algorithm to a large social network and compare the performance of random sampler, TNT and OTNT. In all the examples, both homogeneous and nonhomogeneous instrument kernels introduced in section 2.2 perform well. There is negligible difference in computing and programming effort between them, so we just report the computing time for the first one. Unless otherwise noted, we choose $\kappa_n = \lceil d \log(n+1) \rceil$ at time n of Algorithm 2.2.1 with d a fixed value.

2.4.1 Ising Model

We have a lattice with N^2 nodes. Each node $x_{ij}, 1 \leq i, j \leq N$ is connected with its horizontal and vertical neighbors, and is a $\{-1, 1\}$-valued random variable. Let $x = (X_{ij})$ be an $N \times N$ matrix. Assume that these random variables jointly have an Ising distribution with probability mass function

$$p_\theta(x) = \exp\left\{\theta\left(\sum_{i=1}^{N}\sum_{j=1}^{N-1} x_{ij}x_{i,j+1} + \sum_{i=1}^{N-1}\sum_{j=1}^{N} x_{ij}x_{i+1,j}\right)\right\} / Z(\theta). \tag{2.25}$$

The normalizing constant cannot be computed unless the size of N^2 is small. This type of model is fairly common in image analysis.

We simulate the data by perfect sampling using Propp-Wilson algorithm (Propp and Wilson 1996) with $\theta = 0.25$ for $N = 20, 50, 100$. The prior is $\mu(\theta) \sim \text{Uniform}(0, 1)$, and the posterior is

$$\pi(\theta|x) \propto \exp\left\{\theta\left(\sum_{i=1}^{N}\sum_{j=1}^{N-1} x_{ij}x_{i,j+1} + \sum_{i=1}^{N-1}\sum_{j=1}^{N} x_{ij}x_{i+1,j}\right)\right\} \mathbf{1}_{(0,1)}(\theta).$$

At time n of Algorithm 2.2.1, we propose θ' using a Random Walk Metropolis sampler with proposal distribution $N(\theta_n, \sigma^2)$ ($\sigma = .05$ for $N = 20$, $\sigma = .01$ for $N = 50, 100$). The random sampler is used to sample the networks in Step 2 of Algorithm 2.2.1. For the nonhomogeneous instrument kernel, $f_\ell = f_\theta^{1-\ell/\kappa} f_{\theta'}^{\ell/\kappa}$, for $\ell = 1, \ldots, \kappa$.

The goal of this simulation is twofold. Firstly, we implement Algorithm 2.2.1 with different fixed values of κ, and examine how far away π_κ is from $\pi(\cdot|x)$. Secondly, we take $\kappa = \kappa_n = \lceil d \log(1 + n) \rceil$ and see how well the algorithm performs compared to the AVM. For the comparisons, we need to obtain samples from the true posterior distribution $\pi(\cdot|x)$. We do this by running Algorithm 2.2.1 for a long time with increasing κ_n (at a fast rate). We use the homogeneous instrument kernel with $\kappa_n = \lceil 100\sqrt{n} \rceil$, and run the algorithm for 100,000 iterations.

We run Algorithm 2.2.1 for $100,000$ iterations for different values of κ. We use the output to estimate the Kolmogorov-Smirnov distance between π_κ and $\pi(\cdot|x)$ defined as

$$\mathsf{KS}\left(\pi_\kappa, \pi(\cdot|x)\right) = \sup_{\theta\in[0,1]} \left| \int_0^\theta \pi_\kappa(u)du - \int_0^\theta \pi(u|x)du \right|.$$

In calculating the KS distance, we discard the first 20% of the outcomes as burn-in and a certain thinning (100 for $N = 20, 50$ and 200 for $N = 100$) is applied to the remaining data to reduce the auto-correlation. The results are presented in table 2.1. As expected from Theorem II.1, the KS distance between π_κ and $\pi(\cdot|x)$ decreases as κ increases. The drop in the KS distance is slower for larger N as expected. Also the homogeneous and nonhomogeneous instrument kernels give similar KS distances.

One thing worth mentioning is that the magnitude of σ^2 in proposing θ' affects the algorithm nonnegligibly when κ is small. The uniform bound of $B_\kappa(x, \theta, \theta')$ we get in the theory section is loose for small κ. A tighter bound depends on if x is from a distribution close to $p_{\theta'}$, as well as the distance between θ and θ' (for the nonhomogeneous instrument kernel). Large σ^2 thus leads to a large $B_\kappa(x, \theta, \theta')$. We use a smaller σ for $N = 50$ than $N = 20$. It is then not surprising to observe almost no difference in KS distance between them for small κ. When κ gets large, π_κ is closer to $\pi_\star$ for $N = 20$ than $N = 50$.

Now we allow κ to increase with n ($\kappa_n = \lceil d\log(n+1)\rceil$) so that the limiting distribution of θ_n is $\pi_\star$. We carry out 10,000 iterations. The first 2000 are discarded, and every 10-th of the remaining points are used to do posterior inference (same in all examples below). For comparison, we also run the AVM of Møller et al. (2006) in the setting $N = 20$, where θ' is also proposed from $N(\theta_n, .05^2)$. To reduce the computational load of perfect sampling, the prior of the θ is set to be Uniform$(0, .4)$, and the initial value is $\theta = .2$. The 'true' posterior mean of $\pi_\star$ is obtained using the

Table 2.1: Ising model example: Kolmogorov-Smirnov distance between π_κ and $\pi_\star$ for different size of the lattice, $N^2 = 20^2, 50^2, 100^2$. The homogeneous and nonhomogeneous instrument kernels are differentiated by subscripts 'h' and 'n', respectively.

	20^2		50^2		100^2	
κ	KS$_h$	KS$_n$	KS$_h$	KS$_n$	KS$_h$	KS$_n$
10	.42	.44	.44	.45	.56	.56
30	.31	.32	.34	.36	.47	.49
50	.28	.30	.29	.34	.47	.46
80	.24	.23	.26	.29	.42	.44
100	.20	.19	.27	.26	.44	.45
300	.15	.17	.21	.20	.31	.33
500	.07	.10	.14	.17	.33	.32
800	.07	.07	.13	.15	.26	.28
1000	.05	.07	.11	.14	.28	.28

above sample from $\pi_\star$ with burn-in of 20,000 and thinning of 10. The results are listed in table 2.2.

Again, the performance of both kernels is similar. It is clear from the table that our algorithm recovers the true posterior mean well. As d increases and hence κ_n, our posterior mean will be closer to the true value, and the posterior standard deviation is smaller. For the same reason as above, larger d or smaller σ is needed as N increases. The computing speed of our algorithm is remarkably faster than AVM.

The trace plots, histograms and lag-50 auto-correlation plots for our algorithm with both instrument kernels and AVM in the case of $N = 20, d = 100, \sigma = .05$ are given in figure 2.1. Our algorithm converges to the posterior distribution with good mixing. AVM sampler mixes slowly and often gets stuck at large θ, which gives the false peak in the posterior distribution.

Table 2.2: Ising model example: the posterior mean of the parameter $\bar{\theta}$, its asymptotic standard deviation $\sigma(\theta)$ (taking the auto-correlation into account), the values of d, the proposal standard deviation σ and the computing time for different N^2. The homogeneous and nonhomogeneous instrument kernels are distinguished by subscripts 'h' and 'n', respectively. The computing time is for homogeneous instrument kernel.

N^2	True $\bar{\theta}$	$\bar{\theta}_h$	$\bar{\theta}_n$	$\sigma(\theta)_h$	$\sigma(\theta)_n$	d	σ	Time (mins)
20^2	.252	.256	.253	.037	.013	100	.05	5.03
20^2 (AVM)	.252	.247		.027		NA	.05	577
50^2	.249	.255	.254	.044	.038	5	.01	1.02
50^2	.249	.250	.250	.021	.015	15	.01	1.25
100^2	.252	.254	.255	.031	.026	20	.01	3.26
100^2	.252	.252	.253	.012	.016	100	.01	7.04

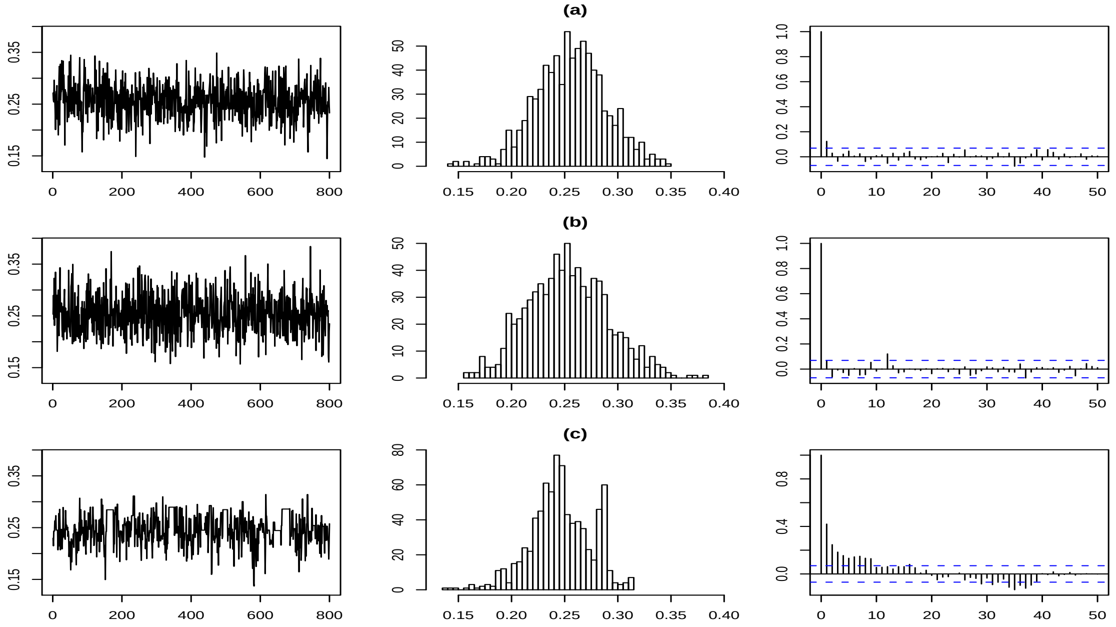


Figure 2.1: Ising model example: (a) Trace plot, histogram, and auto-correlation plot for Algorithm 2.2.1 with homogeneous instrument kernel, (b) with nonhomogeneous instrument kernel, (c) for AVM. Data are generated from Ising model with $N^2 = 20^2, \theta = 0.25$.

2.4.2 Conditional Random Field

We have N^2 individuals on the square lattice $\{1, \cdots, N\} \times \{1, \cdots, N\}$. Assume each individual has a vector of p covariates $A_{ij}^T = (A_{ij,1}, \cdots, A_{ij,p})$ and a dependent variable $X_{ij} \in \{-1, 1\}$, $i, j \in \{1, \cdots, N\}$. Further assume that the dependent variables $(X_{11}, X_{12}, \cdots, X_{NN})$ form a conditional random field given $A = (A_{11}, A_{12} \cdots, A_{NN})$ with distribution

$$\begin{aligned} & p_{\beta,\gamma}(x_{11}, x_{12}, \cdots, x_{NN}|A) \\ \propto \quad & \exp\left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ij} A_{ij}^T \beta + \gamma_1 \sum_{i=1}^{N} \sum_{j=1}^{N-1} x_{ij} x_{i,j+1} + \gamma_2 \sum_{i=1}^{N-1} \sum_{j=1}^{N} x_{ij} x_{i+1,j} \right\}, \end{aligned}$$

for parameters $\beta \in \mathbb{R}^p$ and $\gamma_1, \gamma_2 > 0$. Conditional random field models have a wide range of applications, such as speech processing and computational biology.

We simulate the data through perfect sampling, with $\beta = (.1, .3, .5, .7)$, $\gamma_1 = .2$, $\gamma_2 = .5$, and two different $N = 20, 50$. Each A_{ij} is generated from $N(0, I_4)$. We use prior $\beta \sim N(0, I_4)$ and $\gamma_1, \gamma_2 \sim \text{IG}(1, 1)$. Each parameter is drawn in turn from its full conditional distribution with approximate sampling, and random sampler is used to make a transition from Y_ℓ to $Y_{\ell+1}$. The proposal standard deviations σ for parameters are .05 and .01 for $N = 20, 50$, respectively. In both cases, we use $\kappa_n = \lceil d \log(1 + n) \rceil$ where $d = 10$. Results (after burn-in of 2000, and thinning of 10) are shown in table 2.3. Again, posterior means of our algorithm are very close to the true values obtained in a similar way as in Ising example.

2.4.3 Real Data Analysis - Large Social Network Modeling

The posterior distributions of some ERG models are doubly intractable (e.g., models with alternating k-triangle statistic), and cannot be handled by exact sampling. The improved efficiency of our algorithm makes it feasible to do inference on large networks in a Bayesian framework.

Table 2.3: Conditional random field example: the posterior means of the parameters $\bar{\theta}$ and their asymptotic standard deviations $\sigma(\theta)$ (taking the autocorrelation into account) for different N^2. The homogeneous and nonhomogeneous instrument kernels are distinguished by subscripts 'h' and 'n', respectively.

	20^2					50^2				
θ	True $\bar{\theta}$	$\bar{\theta}_h$	$\bar{\theta}_n$	$\sigma(\theta)_h$	$\sigma(\theta)_n$	True $\bar{\theta}$	$\bar{\theta}_h$	$\bar{\theta}_n$	$\sigma(\theta)_h$	$\sigma(\theta)_n$
β_1	.081	.073	.084	.199	.133	.092	.097	.093	.031	.115
β_2	.295	.298	.296	.074	.054	.313	.317	.311	.116	.040
β_3	.565	.559	.580	.046	.106	.481	.479	.480	.034	.122
β_4	.715	.718	.718	.167	.081	.717	.715	.716	.085	.091
γ_1	.196	.204	.204	.056	.042	.233	.229	.231	.017	.078
γ_2	.525	.536	.533	.063	.037	.478	.481	.478	.027	.044

The Faux Magnolia High school data set (Goodreau et al. 2008) includes a symmetric binary friendship network x of $n = 1461$ students and their demographic information A. We follow the ERG model specification in Goodreau et al. (2008):

$$p_{\theta_1,\theta_2,\beta}(x|A) = \exp\left\{\theta_1 \mathrm{E}(x) + \theta_2 v(x) + \beta^T S(x,A)\right\}/Z(\theta_1,\theta_2,\beta).$$

$\mathrm{E}(x) = \sum_{1\leq i<j\leq n} x_{ij}$ is the number of edges capturing network density. $v(x) = e^{.2}\sum_{i=1}^{n-2}\{1-(1-e^{-.2})^i\}D_i(x)$ is the alternating k-triangle statistic measuring friendship transitivity, and its ratio parameter is fixed to be .2 to facilitate the results comparison with MLE in Goodreau et al. (2008). $S(x,A)$ is a 3-dimensional vector of statistics capturing the similarities between friends on grade, race and gender, e.g., $\sum_{1\leq i<j\leq n} x_{ij}\mathbf{1}(grade_i = grade_j)$, where $\mathbf{1}(grade_i = grade_j) = 1$ if students i and j are in the same grade, and 0 otherwise. The prior we use is $N(0, 100I_5)$. We found it more efficient to work with edge list (a 1461 × 2 matrix whose rows recode dyads with edges) rather than a 1461 × 1461 matrix in programming. Again, we do a burn-in of 2000 and thinning of 10 for all results below.

To compare the performance of random sampler, TNT and OTNT, we carried out approximate sampling with each of them, respectively, starting with the same initial parameter values. Again we used $\kappa_n = \lceil d \log(1+n) \rceil$ with $d = 10$. For OTNT, we let $w_1 = .5, w_2 = w_3 = .25$. Figure 2.2 shows that in estimating alternating k-triangle parameter θ_2, OTNT produces much better mixing. The posterior sample of θ_2 is all over the place in random sampler, ranging from -10 to 30. While improved greatly comparing to random sampler, the mixing of TNT is still bad due to the lack of variation in the alternating k-triangle statistic value. On the contrary, OTNT mixes fast with a maximum κ_n as small as 93.

Table 2.4 gives the posterior means and 95% credible sets of the parameters. To save space, we only report results of the homogeneous instrument kernel, as both kernels yield very similar results. Approximate sampling with OTNT gives posterior means consistent with MLE in Goodreau et al. (2008). The Random sampler gives results that are completely off for all parameters, whereas the posterior mean of θ_2 for TNT is far from MLE as well. We believe that with much larger κ_n, TNT will eventually give a similar estimate, but at the expense of more computing power. We also present the posterior distributions for network density θ_1 and grade homophily β_1 generated with OTNT in figure 2.3. Again, the mixing is good, and posterior distributions are roughly normal with MLE close to posterior modes.

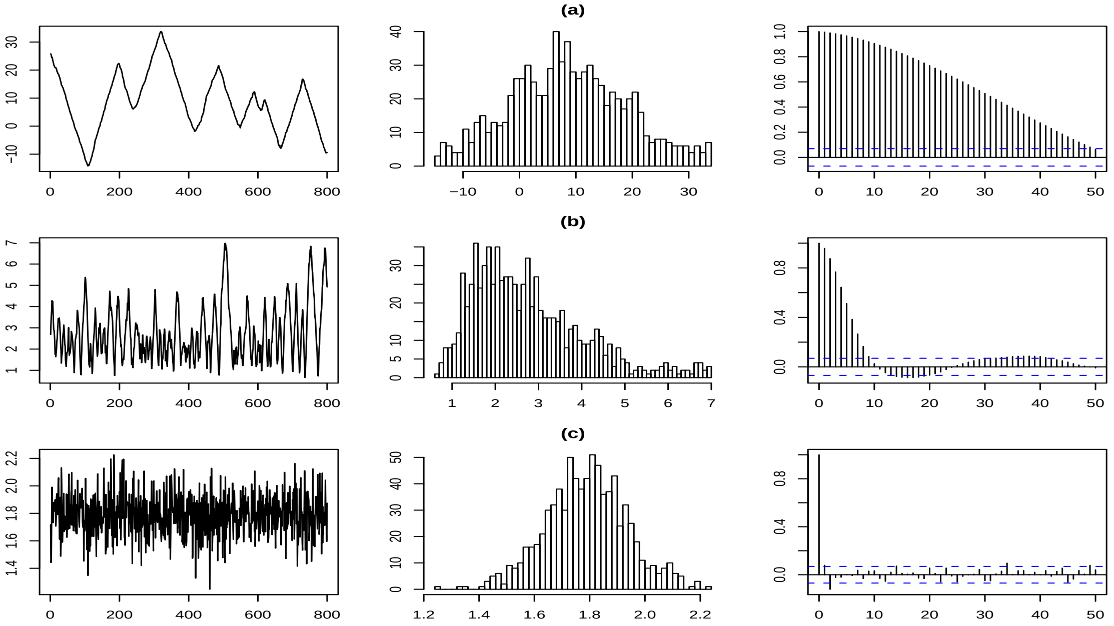


Figure 2.2: Social network example: trace plots, histograms and auto-correlation plots for θ_2 using different M-H moves (used in Algorithm 2.2.1 with the homogeneous instrument kernel). (a)-(c) are random sampler, TNT and OTNT, respectively.

Table 2.4: Social network results

Parameter	MLE	OTNT (95% Credible Set)	TNT	RS*
Density (θ_1)	-9.79	-9.83 (-10.4, -9.28)	-9.88	-127
Alternating k-triangle (θ_2)	1.82	1.79 (1.50, 2.09)	2.75	8.65
Homophily grade effect (β_1)	2.76	2.83 (2.38, 3.29)	2.86	77.6
Homophily race effect (β_2)	.918	.926 (.421, 1.43)	.934	11.0
Homophily gender effect (β_3)	.766	.780 (.299, 1.25)	.806	6.94

The posterior means of Algorithm 2.2.1 with the homogeneous instrument kernel for different M-H moves compared with MLE in Goodreau et al. (2008).
*RS represents random sampler.

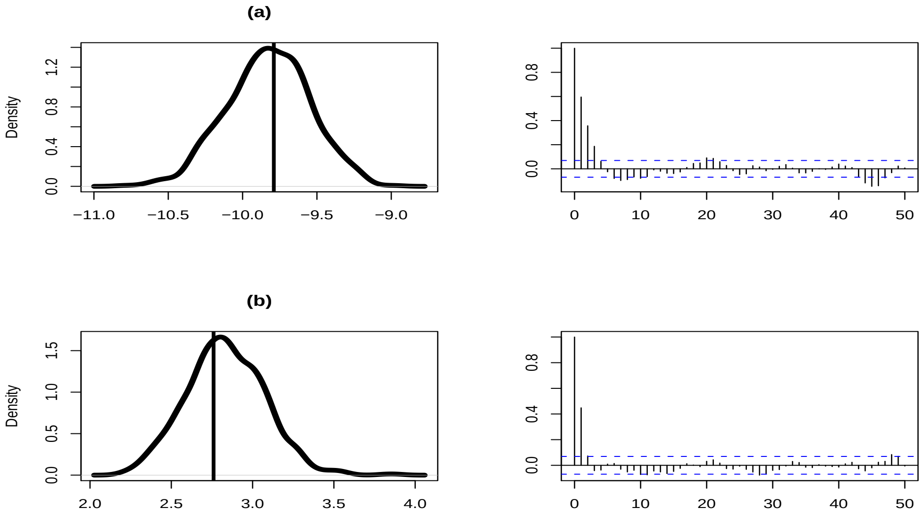


Figure 2.3: Social network example: density and autocorrelation plots of parameters corresponding to (a) network density θ_1 and (b) grade homophily β_1. Results are produced by algorithm 2.2.1 with homogeneous instrument kernel and OTNT. The vertical line indicates MLE in Goodreau et al. (2008).

CHAPTER III

Modeling Interdependent Choices in a Social Network

3.1 Introduction

In the past decade, the market place has observed the emergence and continuous growth of social media such as Facebook, Linkedin and Twitter. Although the primary purpose of social media is to facilitate interactions among their users, many of these sites rely on advertisers, not users, as their sole source of revenue generating mechanism. The primary key selling point for advertising on social media is the sites' ability to help advertisers target banner ads to consumers based on their detailed demographic and psychographic profiles. Interestingly, some social media sites, such as Facebook, go the extra mile to introduce a new form of interactive advertising that also takes advantage of the social network embedded in their websites. Specifically, these social media sites provide the capability to automatically pass along information about a user's response to an advertisement to his/her friends. For example, Facebook offers an interactive advertising feature called engagement advertising. After target users are exposed to an engagement ad, they may be asked to perform a certain activity, such as RSVP to attend Starbuck's donation event for AIDS foundation or becoming a fan of a brand page. Once the users respond to such a solicitation, their responses will be shown on their friends' news feeds, which is the first page users will see once they log on. This news feed feature clearly makes observation of friends' behavior more transparent and facilitates the propagation of the social interaction process. Thus the impact of targeted advertising may not only be limited to its intended consumers, but also to others with whom these consumers are connected.

In order to accurately measure the impact of and effectively design such interactive advertising, it is crucial to have a modeling framework that captures how consumers' preferences can be influenced and dynamically changed over time within a social network. Here we primarily focus on modeling choice interdependence as a reflection of changes in consumers' preferences induced by their observations of choices made by others in a social network. Our work deviates from the previous related literature in several important ways. Firstly, unlike the prior literature that attempts to model the impact of online word-of-mouth (Godes and Mayzlin 2004, Chevalier and Mayzlin 2006, Zhu and Zhang 2010), viral marketing activities (Leskovec et al. 2007, Toubia et al. 2009), and online community participation (Algesheimer et al. 2010, Stephen and Toubia 2010) on aggregate sales, our interest lies in modeling choice interdependence at the individual level. Secondly, our focus is on capturing how an individual consumer's choice is influenced by others' choices in a social network, and not on deriving group choices from diverse individual preferences (Aribarg et al. 2002, Aribarg et al. 2010) or on modeling coordinated choices (Hartmann 2010, Yang et al 2009) among a small group of individuals. Thirdly, our source of interdependence is *choices* made by others, not *preferences* of others (Yang and Allenby 2003, Yang et al. 2006) in a network. Finally, similar to Nair et al. (2010), our models can accommodate asymmetric effects of social influence.

However, our approach aims at modeling the joint distribution of individual choices within a network, and not the marginal distribution (i.e., an ego-centric approach, Trusov et al 2010) of individual quantity decisions.

Modeling choice interdependence in a social network is a challenging problem. To properly account for the dependence structure in the data (i.e., the joint probability of individuals' choices in a network), we propose two modeling frameworks for two types of choice data researchers may collect. With the social media technology, it is feasible to collect longitudinal choice data where choices made in a social network are recorded real time, and thus researchers also observe the sequence of choices (hereafter referred to as complete data). Given its complete information on the choice evolution, we apply a discrete-time Markov chain model for this type of data.

On the other hand, often time researchers may only observe a cross-sectional snap-shot of realized choices in a network at a particular time point (hereafter referred to as snap-shot data). In this case, although actual choices are made sequentially by different people, the choice sequence cannot be observed (e.g., researchers can only observe choices of hybrid vehicles in a neighborhood at a particular point in time). We propose a Markov random field (MRF) model which can also accommodate multiple relations within a network. Yet a big hurdle of fitting an MRF model is its well-known intractable normalizing constant problem, that is, the normalizing constant is a function of parameter values and thus cannot simply be ignored in the derivation of the posterior density. We demonstrate how either a pseudolikelihood (Besag 1974, Besag 1975) or the approximate sampling algorithm introduced in chapter II can be used to simplify the estimation of an MRF model. As snap-shot data are a result of a dynamic process, we also show necessary and sufficient conditions for the discrete-time Markov chain to converge to the MRF. This result suggests that MRF is a natural model that emerges from the Markov chain dynamic choice process. When complete data are unavailable in practice due to technical constraints, we do not lost too much information by collecting snap-shot data.

To test our proposed models, we design a field experiment where we first collect information about participants' social connections within a specific social network, as well as measuring each participant's intrinsic preferences. We later observe in real time how they make product choices given they observe other participants' choices. We fit a discrete-time Markov chain model to the complete data and an MRF model to the snap-shot data (i.e., realization of the complete data). Controlling for individuals' heterogeneous preferences, our results show that participants rely on choices observed both at the micro (i.e., closely-connected friends) and macro (i.e., a reference group such as male vs. female) levels in making their decisions. We also find asymmetry in choice interdependence across different types of consumers (i.e., expert vs. novice). Our choice predictive exercise underscores the considerable impact of choice interdependence in a social network on individuals' choices. Findings from our simulation also suggest that for small to medium-sized networks estimating an MRF model using the approximate sampling method leads to better choice prediction than using the pseudolikelihood method. However, as networks grow in size the difference in predictive performance dissipates, and thus using the pseudolikelihood method which is less computationally intensive may be more desirable.

This chapter is structured as follows. We start with a brief literature review, and then present our modeling framework in section 3.3. This is followed by a simulation study to evaluate different estimation methods in section 3.4. Section 3.5 provides a detailed description of our experimental design and data collection procedure, as well as estimation results and model comparison. We close with a discussion about our contributions and future research.

3.2 Literature Review

Marketing researchers have long realized that consumers' preferences are not independent. How the interactions among consumers within their social circles can influence their behavior is thus an important topic in marketing research. In the introduction chapter, we have reviewed existing statistical models for interdependent data. Here we narrow down to the marketing literature, and discuss some related prior work in more details and contrast our method with them.

Most of the existing choice models that factor in choice interdependence deal with small networks of consumers, such as a household and a small group of people (Yang et al. 2006, Hartmann 2010, Yang et al 2009). However, consumers can often be influenced by a larger group of people, such as a cohort of students, a group of colleagues in the same department, and co-op neighbors etc. With an exception of Yang et al (2009), extending these existing models to accommodate large networks is practically implausible. For example, Hartmann (2010) relies on the theory of simultaneous-move games (Bresnahan and Reiss 1991) to model simultaneous choice decisions (i.e., coordinated choices) made by multiple players. Yang et al (2009) discuss the difficulty in applying such an ”unconditional approach” in Hartmann (2010) to situations with more than two choice alternatives or two game players. Yang et al (2009) opt for a ”conditional approach” and introduce an auto-model for multinomial choices, which extends the auto-logistic model (Besag 1974, Russell and Petersen 2000, Moon and Russell 2008) from binary choices to multinomial choices. However, the auto-model proposed in Yang et al (2009) also encounters the intractable normalizing constant problem if it were to be applied to a larger-sized network, as opposed to a network of size three used in their paper.

In fact, as we introduced in chapter I, auto-model is a special case of MRF. Denote individual i's choice by y_i, $i = 1, \cdots, n$. In Yang et al (2009), each choice is dummy coded as a vector of 0 or 1, e.g., $Y_i = (0, 1, 0)$ means among the three choice options, i chooses option 2. The joint distribution of multinomial auto-model takes the form of $P(y_1, \cdots, y_n) = \exp\left\{\sum_{1 \le i \le n} y_i U(y_i) + \sum_{1 \le i < j \le n} \beta_{ij} y_i^T y_j\right\} / Z(\beta)$. $U_i(y_i)$ is a function of model parameters (e.g., a utility function) and $\beta_{ij} = 0$ only when decisions of individual i and j are independent. Given such restrictive form of $\beta_{ij} y_i^T y_j$, the auto-model can only accommodate symmetric social interaction effects ($\beta_{ij} = \beta_{ji}$). The MRF model is more general, as it allows for any functional form to capture choice interdependence and for asymmetric interaction effects. We add another layer of flexibility by allowing multiple relations (e.g., closely connected friends vs. a reference group) for the same participants in the model.

Another close line of work is modeling interdependent preferences using simultaneous autoregressive models (Smith and LeSage 2004, Yang and Allenby 2003, Yang et al. 2006). A comparison between simultaneous autoregressive models and conditional autoregressive models (a special case of MRF) can be found in Besag (1974), Cressie (1993) and Wall (2004) for Gaussian data. For non-Gaussian choice data, the above models will fall into the category of mixed models introduced in section 1.1.2, with Gaussian process as the distribution of random effect. They inherit from mixed models the differences to MRF (section 1.1). In addition, the Markovian assumption in MRF ensures that given neighbors' preferences, one's preference is independent of other people's preferences. On the contrary, in simultaneous autoregressive models, given neighbors' preferences, one's preference can still depend on those of the neighbors' of his/her neighbors, and so on and so forth. This may not be a desirable property in parameter interpretation for some situations. On the other hand, even if this is a necessary property, we can always expand the neighboring system to incorporate higher-order neighbors in MRF.

Another conceptual distinction is that this type of model takes the form that one's preference is a function of others' preferences (e.g., in Yang and Allenby (2003), preferences of different consumers in a network are specified to be correlated through a spatially correlated error structure). In contrast, we conceptualize a consumer's preference to be influenced by his or her friends' observed choices, rather than preferences. In some contexts such as interactive advertising on a social media site, consumers are more likely to observe his or her friends' choice decisions rather than discuss with one another their underlying preferences. MRF provides a natural framework to study choice interaction. But the framework of MRF is flexible enough to also accommodate preferences interaction (together with choice interaction).

Social network and its influence on people's behavior have also sparked interest of researchers from other fields such as sociology, physics, computer science, statistics, etc. While social network modeling (Wasserman and Pattison 1996, Hunter and Handcock 2006) focuses on the connections among nodes (e.g., relationships among individuals) within a network, here we study the behavior of the nodes in a network. Recently, researchers studying the dynamics of network and behavior are able to collect snap-shot data at more than one time point (Koskinen and Snijders 2007, Snijders et al. 2007), e.g., every three months. However, for snap-shot data at two consecutive time points, there exist multiple processes that can give rise to the observed data. These papers use a continuous Markov process to fit the data and treat the possible processes between consecutive time points as latent variables to be augmented. In our context, given the availability of complete data, we adopt a simpler discrete-time Markov chain model. Due to the large volume of the data, in reality it may be difficult to keep track of the time between two consecutive choices. Thus, a discrete-time Markov chain is more appropriate. However, when time is available, a continuous time Markov process can be applied to study the factors that influence when people make a decision, and even how the length of the time affects the decision with a semi-Markov process.

3.3 Modeling Framework

To examine choice interdependence in a social network, two types of choice data can be collected: complete and snap-shot data. Complete data are longitudinal choice data where the sequence of choices is observed. With the information of the sequence, we can better understand the role of a social network in the entire dynamic choice process. That is, not only can we model which choice is made, but we can also model who is likely to makes a choice. Snap-shot data, on the other hand, are cross-sectional choice data collected at a fixed time point. This type of data are easier and less expensive to obtain. However, lack of information on the sequence of choices makes it difficult to understand the dynamic process, as different processes can give rise to the same snap-shot data. No matter what type of data are available, we need to model all consumers' choices jointly rather than independently to properly preserve the dependence structure in the data.

3.3.1 Discrete-Time Markov Chain for Complete Data

For complete data, we have information on both choices and the choices' sequence. A discrete-time Markov chain can be applied to take advantage of the extra information, which can capture "who" makes the choice in addition to "which" choice is made. Following Koskinen and Snijders (2007), we assume that each time at most one member may change the choice (i.e., no multiple changes are allowed simultaneously). This assumption simplifies the modeling as we do not need to handle the possible dependence of the transitions made at the same time. The state of the Markov chain consists of two parts: who intends to change (e.g., individual who makes a choice for the first time changes the choice from 0 to a different value; individual who changes his/her last choice) the choice and everyone's choice. Again, we emphasize that the state of the Markov chain involves all members' choices rather than a single person's due to the interdependence. We denote "who intends to change the choice" by c^t, which takes value from 1 to n. Denote consumer i's choice at time t by Y_i^t, taking nominal values $0, \cdots, K$. The state at time point t is $S^t = (c^t, Y_1^t, \cdots, Y_n^t)$. The first-order Markovian property dictates that:

$$P(s^{t+1}|s^1, ..., s^t, \theta) = P(s^{t+1}|s^t, \theta).$$

Under the assumption that only one transition is allowed, only one person's choice $Y_{c^{t+1}}^{t+1}$ can be different from $Y_{c^{t+1}}^t$, while all other choices remain the same $Y_j^{t+1} = Y_j^t$ for $j \neq c^t$. Then we have:

$$\begin{aligned} P(s^{t+1}|s^t, \theta) &= P(c^{t+1} = i, y_i^{t+1}|c^t, y^t, \theta)\mathbf{1}(y_{-i}^{t+1} = y_{-i}^t) \\ = & \ P(c^{t+1} = i|c^t, y^t, \theta)P(y_i^{t+1}|c^{t+1} = i, c^t, y^t, \theta)\mathbf{1}(y_{-i}^{t+1} = y_{-i}^t). \end{aligned} \tag{3.1}$$

The likelihood is $P(s^1, \cdots, s^T|s^0, \theta) = \prod_{t=0}^{T-1} P(s^{t+1}|s^t, \theta)$, where $y^0$and c^0 are to be specified later. There is great flexibility in specifying the choice part $P(y_i^{t+1}|c^{t+1} = i, c^t, y^t, \theta)$ of the transition probability. For multinomial choices, a multinomial logit model can be used to fit the choice part. Researchers can postulate any probability form of interest for $P(c^{t+1} = i|c^t, y^t, \theta)$. We include $\mathbf{1}(y_{-i}^{t+1} = y_{-i}^t)$ for completion to assure that only one choice can be made at a time.

We make several remarks about the model. First, a transition or change does not necessarily mean making a different choice. We allow each individual i to keep the same choice (e.g., an individual can log on to the website but then log out keeping the same choice he or she has chosen before in our experiment). Second, modeling everyone's choice together preserves the interdependence within the system. More importantly, we can model who likely makes the next choice decision. This model specification provides more insights of the mechanism of the choice dynamics. Finally, it is easier to identify asymmetric dependence structure in a discrete-time Markov chain model, as in each transition only one person's choice is influenced by others', and thereby statistics corresponding to the parameters for asymmetry always have different values (see conditions to identify asymmetry in section 3.3.2).

3.3.2 Markov Random Field (MRF) with Multiple Relations for Snapshot Data

For snap-shot data, we can specify a Markov random field (MRF) to capture the dependence among consumers' choices. Consumers' choices $Y = (Y_1, \cdots, Y_n)$, with Y_i taking nominal values $0, \cdots, K$, are the variables constituting a random field, and their dependence comes from the presence of relationships among consumers in a social network, e.g., figure 3.1. The circles and arrows in figure 3.1 depict a directed network of consumers, where relationships are not necessarily reciprocal. Recall that to define an MRF, we first need to construct a neighboring system. In an undirected network, it is straightforward to define neighbors as people who are connected by an edge in the network. For a directed network, two people are neighbors if there is an arrow between them regardless of the direction. It is legitimate to do so, since being neighbors represents conditional dependence in an MRF and should be reciprocal. The directionality of a network can be used in potentials to capture asymmetric

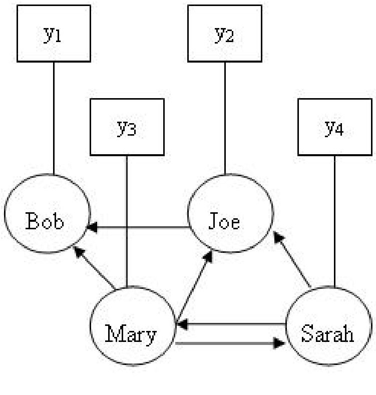


Figure 3.1: A random field of consumers' choices in a social network

social interactions. We will show an example in the data analysis.

We have introduced the general framework of a Markov random field in section 1.1.1. Individual i's neighbors will include anyone whose choice Y_i may be conditionally dependent on. If there is only one type of relation, and thus one network among the n individuals, then the set of i's neighbors $N(i)$ contains everyone who is connected with i in the network. The joint distribution can be determined by the interactions within cliques determined by this neighboring system. Recall that the joint distribution of Y is given by:

$$P(y_1, \cdots, y_n|\theta) = \frac{1}{Z(\theta)} \exp\left\{\sum_{c \in \mathcal{C}} U_c(y_i, i \in c; \theta)\right\}. \tag{3.2}$$

The choice interdependence in real life is usually complicated and may result from different types of relations. Therefore for each relation, we define a separate set of neighbors (individuals connected in each network), which results in different sets of cliques for different networks. For example, we can conceptualize that each individual is influenced by both micro (i.e., closely-connected friends) and macro (e.g., a reference group) levels of relationships.

We specify the MRF with multiple relations as follows:

$$P(y_1, \cdots, y_n|\theta) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{m=1}^{M} \sum_{c \in \mathcal{C}_m} U_c(y_i, i \in c; \theta) \right\}, \tag{3.3}$$

where M is the number of relations and $\mathcal{C}_m$ are possible cliques corresponding to relation m. It is easy to see that $P(y_i|y_{-i}, \theta) = P(y_i|y_j, j \in \bigcup_{m=1}^{M} N_m(i), \theta)$ with $N_m(i)$ i's neighbors in network m, which means the conditional distribution of Y_i only depends on i's neighbors.

Some remarks related to the MRF are worth mentioning here. First, an MRF can accommodate networks in various formats. A directed network allows relationships to be non-reciprocal. A directed and/or weighted network allows relationships to be asymmetric. Second, estimating an MRF does not require the inclusion of all possible cliques. However, in order to capture choice interdependence, some cliques of size greater than one need to be included, as cliques of size one are only related to individual intrinsic preference and unique characteristics.

Third, U_c can be any arbitrary function. An example of U_c in figure 3.1 between Sarah and Joe could be $U_c = \theta \mathbf{1}(Y_2 = Y_4)$. Then a positive coefficient reflects the tendency to purchase the same product, if a consumer shares the purchase decision with another consumer. We can use the same potential for cliques of the same size. On the other hand, we can use different potentials for cliques (i, j) and (j, i) if the network is directed.

Fourth, with some necessary conditions, we can identify asymmetric interactions with the MRF. The first condition involves a clearly defined identity of each node: at the individual level or at the role level. The second condition is that statistics corresponding to the parameters for asymmetry must have different values, which most of the time is satisfied in a directed network. To identify asymmetric interactions in the case of individuals, researchers also need to observe more than one choice for each individual. If nodes' identities are defined at the role level, one decision per individual will be sufficient as long as aggregate statistics associated with each role differ.

3.3.3 Estimation Methods for MRF

Despite the ease and flexibility of modeling choices jointly, an MRF model has an intractable normalizing constant $Z(\theta)$ (section 1.2). The problem is more severe for large K or n. We will introduce pseudolikelihood (Besag 1974, Besag 1975) as one way of getting around the computational issue. Pseudolikelihood replaces the likelihood $P(y|\theta)$ with a pseudolikelihood function $PL(y|\theta)$ that does not involve an intractable normalizing constant, and thus is easy to compute especially for very large networks. However, since it does not work with the exact likelihood function, the induced posterior distribution, $p_{\mathrm{PL}}(\theta|y)$, is also not the one of interest, $p(\theta|y)$. In contrast, the approximate MCMC approach in chapter II tries to approximate the likelihood function and is proved to be asymptotically consistent. In this section, we will introduce pseudolikelihood method, and details about approximate MCMC can be found in section 2.2.

The maximum pseudolikelihood estimator is proposed in Besag (1974) and Besag (1975) as an alternative to the maximum likelihood estimator. The pseudolikelihood function is defined as:

$$PL(y_1, \cdots, y_n|\theta) = \prod_{i=1}^{n} P(y_i|y_{-i}, \theta). \tag{3.4}$$

Each term on the right is the conditional probability of Y_i that only depends on i's neighbors. The intractable normalizing constant problem does not occur in this specification because of its conditional independent structure. In case of choice variable, $P(y_i|y_{-i}, \theta)$ has the same form as a logit model with y_{-i} entering as a covariate. To further illustrate the connection between the likelihood of an MRF and its corresponding pseudolikelihood, suppose we specify an MRF with three individuals facing four product options (0 to 3), where individual 1 declares a relationship with 2 but not vice versa, and individuals 2 and 3 have reciprocal relationships. One simple specification based only on cliques of size one (i.e., capturing the utility each individual i can obtain for each choice option) and size two (e.g., capturing choice interdependence) is $P(y|\theta) = \frac{\exp\left\{\sum_{1\leq i\leq 3}\beta u(y_i)+\theta[\mathbf{1}(y_1=y_2)+2\times\mathbf{1}(y_2=y_3)]\right\}}{\sum_{0\leq y_1,y_2,y_3\leq 3}\exp\left\{\sum_{1\leq j\leq 3}\beta u(y_i)+\theta[\mathbf{1}(y_1=y_2)+2\times\mathbf{1}(y_2=y_3)]\right\}}$. The associated pseudolikelihood entails a multiplication of three terms, each of which corresponds to each individual, is given by $PL(y|\theta) = \frac{\exp\{\beta u(y_1)+\theta\mathbf{1}(y_1=y_2)\}}{\sum_{0\leq y_1\leq 3}\exp\{\beta u(y_1)+\theta\mathbf{1}(y_1=y_2)\}} \times \frac{\exp\{\beta u(y_2)+\theta[\mathbf{1}(y_2=y_3)]\}}{\sum_{0\leq y_2\leq 3}\exp\{\beta u(y_2)+\theta[\mathbf{1}(y_2=y_3)]\}} \times \frac{\exp\{\beta u(y_3)+\theta[\mathbf{1}(y_3=y_2)]\}}{\sum_{0\leq y_3\leq 3}\exp\{\beta u(y_3)+\theta[\mathbf{1}(y_3=y_2)]\}}$. The normalizing constant for each component in the pseudolikelihood is much easier to compute, as it only involves a summation over four terms (i.e., four choice options) as opposed to 64 terms (i.e., there are 4^3 all possible values of y) in the MRF specification.

The pseudolikelihood function is not equivalent to its associated likelihood function and does not sum up to 1 except for the trivial case where all choices are independent of each other. However, the maximum pseudolikelihood estimator is proven to be consistent as n goes to infinity under certain conditions (Geman and Graffigne 1987). Yet, the efficiency of a maximum pseudolikelihood estimator as compared to that of a maximum likelihood estimator is not well established. With a pseudolikelihood[1], traditional Markov chain Monte Carlo (MCMC) procedures can then be implemented on the posterior distribution defined by a pseudolikelihood, $p_{\mathrm{PL}}(\theta|y) \propto \pi(\theta)PL(y|\theta)$.

[1] Before replacing a likelihood with a pseudolikelihood in the derivation of a posterior distribution, one needs to check whether the pseudolikelihood is proper (i.e., the integral or sum is finite) to ensure that the posterior is also proper.

3.3.4 Connection between MRF and Discrete-time Markov Chain

Despite the difference in the formulation and the type of data they handle, Markov chain and MRF are closely connected. The Markov chain in this subsection we refer to is the marginalized discrete-time Markov chain made up of choices only, the likelihood of which is $P(y^1, \cdots, y^T|y^0, \theta) = \sum_{c^0,\cdots,c^T} P(s^1, \cdots, s^T|s^0, \theta)P(c^0|y^0, \theta)$. The intuition is that as the complete data give rise to the snap-shot data, we speculate that under certain conditions, the Markov chain may converge to the MRF as the dynamic process continues. In this subsection, we give a sufficient condition for this speculation.

Theorem III.1. *In the Markov chain model, if (i) the choice part of the transition probability* (3.1) *has the same form as the conditional probability of MRF calculated based on equations* (3.2) *or* (3.3)*, i.e.,* $P(y_i^{t+1}|c^{t+1} = i, c^t, y^t, \theta) = P(y_i|y_{-i} = y_{-i}^t, \theta)$*; (ii)* c^{t+1} *is independent of* y^t*, i.e.,* $P(c^{t+1}|c^t, y^t, \theta) = P(c^{t+1}|c^t, \theta)$*, then the stationary distribution of the Markov chain is given by the MRF (*(3.2) *or* (3.3)*). If further (iii) the transition probability of the Markov chain produces an irreducible and aperiodic Markov chain, then its limiting distribution is* (3.2) *or* (3.3)*.*

A Markov chain is irreducible if any two states can reach each other within finite number of transitions, i.e., for any two states z, z' (can be the same), there exists $t \geq 0$ such that $P(y^t = z'|y^0 = z) > 0$. A state z is aperiodic if there is no integer $d > 1$ such that $P(y^t = z|y^0 = z) = 0$ whenever it is not divisible by d. If all states are aperiodic, the Markov chain is aperiodic. For an irreducible Markov chain, we only need to show that one state is aperiodic to obtain this property for the whole chain.

Proof. The key thing is that our Markov chain can be considered as being generated from a Gibbs sampler with stationary distribution of (3.2) or (3.3). We define the sampler as follows. In iteration $t+1$, we randomly pick an i according to $P(c^{t+1} = i|c^t, y^t, \theta)$, and then generate Y_i^{t+1} from $P(y_i^{t+1}|y^t, c^{t+1} = i, \theta)$. Our Markov chain can be obtained by this sampler. Condition (i) and (ii) ensures that it is a Gibbs sampler with invariant distribution (3.2) or (3.3). Condition (iii) guarantees that the limiting distribution of this sampler is the same as its stationary distribution (Ross 1996). Therefore, our Markov chain will converge to (3.2) or (3.3) as the number of transitions goes to infinity.

Now we will prove the invariant distribution of the Gibbs sampler is $P(y|\theta)$ defined in (3.2) or (3.3). We eliminate θ and denote the choices in consecutive two steps by x and y for notation lucidity. All we need to prove is $\sum_x P(x)Q(y|x) = P(y)$, where $Q(y|x)$ is the transition kernel of the sampler. Then we have

$$\begin{aligned}
& \sum_x P(x)Q(y|x) \\
= \; & \sum_x P(x) \sum_{i=1}^n P(c^{t+1} = i|c^t, x) P(y_i|c^{t+1} = i, c^t, x) \mathbf{1}(y_{-i} = x_{-i}) \\
= \; & \sum_x P(x) \sum_{i=1}^n P(c^{t+1} = i|c^t) P(y_i|x_{-i}) \mathbf{1}(y_{-i} = x_{-i}) \\
= \; & \sum_{i=1}^n P(c^{t+1} = i|c^t) \sum_x P(x_i|x_{-i}) P(x_{-i}) P(y_i|x_{-i}) \mathbf{1}(y_{-i} = x_{-i}) \\
= \; & \sum_{i=1}^n P(c^{t+1} = i|c^t) \sum_x P(x_i|x_{-i}) P(y_i, x_{-i}) \mathbf{1}(y_{-i} = x_{-i}) \\
= \; & \sum_{i=1}^n P(c^{t+1} = i|c^t) \sum_{x_i} P(x_i|y_{-i}) P(y) \\
= \; & P(y) \sum_{i=1}^n P(c^{t+1} = i|c^t) \\
= \; & P(y). \qquad (3.5)
\end{aligned}$$

□

Remark: Note condition (ii) in the theorem is a sufficient rather than necessary condition. The necessary condition takes a less intuitive form, which is $\sum_{i=1}^{n}\sum_{x_i} P(c^{t+1} = i|c^t, x_i, y_{-i})P(x_i|y_{-i}) = 1$. It is not difficult to prove the above conclusion after replacing condition (ii) with this necessary condition.

3.4 Simulation Study

To compare the performance of pseudolikelihood and approximate sampling methods, we conduct the following simulation exercise. Suppose each household in a community is given two different coupons to choose. For simplicity, we assume the houses are built on an N by N square lattice and define neighbors as households living in front of, behind, to the left and right of the house. We want to examine if households tend to choose the same or different coupons as their neighbors. The model formulation[2] is as follows. The household with coordinate (i, j) on the lattice picks either of the two coupons, i.e., $x_{ij} = \{-1, 1\}$. Let $X = (x_{ij})_{1\le i,j\le N}$, the likelihood is

$$P(x|\theta) = \exp\left\{\theta\left(\sum_{i=1}^{N}\sum_{j=1}^{N-1} x_{ij}x_{i,j+1} + \sum_{i=1}^{N-1}\sum_{j=1}^{N} x_{ij}x_{i+1,j}\right)\right\}/Z(\theta).$$

We simulate the data for $N = 5, 10, 20$ corresponding to 25, 100 and 400 households, and $\theta = .25, .4$. Since we know the true value of θ, mean square error is calculated for comparison of the two methods. Moreover, 20% of the data are randomly taken out of the sample, and we compare the out-of-sample hit rates as well. The results are presented in table 3.1. For mean square error, approximate sampling with homogeneous instrument kernel performs better in all cases. The difference is larger when the social effect is stronger. But as network size increases, the gap diminishes. The comparison in terms of hit rate demonstrates a similar pattern, although hit rate is less sensitive to difference. Hence in cases with small social

[2] This model is the Ising model in section 2.4.1.

Table 3.1: Simulation results

θ	N^2	MSE ($\times 10^{-2}$)		Out-of-sample hit rate (%)	
		PL	AS	PL	AS
.25	25	2.08 (.044)	1.64** (.047)	60.1 (.602)	61.3 (.540)
	100	.610 (.020)	.434** (.012)	57.1 (.100)	57.4 (.242)
	400	.152 (.004)	.102** (.005)	57.2 (.108)	57.2 (.124)
.4	25	1.90 (.084)	1.55** (.066)	78.0 (.575)	86.0** (.565)
	100	.825 (.016)	.537** (.011)	79.0 (.170)	81.9** (.104)
	400	.485 (.013)	.359** (.009)	74.0 (.038)	75.6** (.019)

Within parenthesis is the standard error. PL stands for pseudolikelihood. AS stands for approximate sampling.
** Approximate Sampling is significantly better then Pseudolikelihood at .05 level.

effect and large network, the pseudolikelihood and approximate sampling methods predict choices equally well. These results suggest that when dealing with a large network, due to computational consideration, the pseudolikelihood method is more desirable. Nonetheless, for a small and moderate-sized network the approximate sampling method provides more accurate inference.

3.5 Experimental Design and Data Collection

We design a field experiment that facilitates us to collect longitudinal choice data where choices made in a social network are recorded in real time (complete data). As a result, we at the same time obtain cross-sectional (snap-shot) data that are the realization of the complete data. There are several design issues we consider. First, we obtain data from a relatively well-connected social network. Our network involves the undergraduate students who took an introduction to marketing class in the fall of 2009 at a Midwest university. The majority of students knew each other. We expect this network to be denser than that of a randomly selected group of students, and close to a friendship network in reality.

Second, our experimental design circumvents possible confounds from homophily or endogenous group formation (i.e., individuals with similar preferences are more likely to be friends). In order to tease the influence of others' choices in a network on an individual's choice apart from homophily, we need to control for each individual's intrinsic preference. As such, we measured each participant's (initial) preferences for product attributes using a conjoint choice experiment at the same time we collected network information. Without the ability to observe longitudinal choice decisions within-subject and include individual fixed (Nair et al. 2010) or random effects (Hartmann 2010, Yang et al 2009) in the model, or to directly measure initial preferences (in our study), researchers may include individual covariates as proxies of intrinsic preferences.

Lastly, we choose two products of interest to students, a bundle of university sports paraphernalia and a Bluetooth wireless headset, to motivate them to reveal their true preferences. Our participants also had opportunities to win products of their choices (both the paraphernalia bundle and Bluetooth wireless headset) as an additional incentive to participate in our study. Our paraphernalia bundle contains a jacket, a t-shirt, and a hat with the university logo. We ran the experiment during the football season to further enhance the participants' interests.

Our data collection procedure consisted of three phases. Phase 1 was recruitment. Participants completed two conjoint choice tasks, one for the university paraphernalia and the other for a Bluetooth wireless headset. They also answered a series of social network questions, along with some demographic questions. The purpose was to obtain participants' initial product preferences. Effects of choice interdependence can be identified if participants' preferences for product attributes, as reflected through their choices, change as they observe other participants' choices. Each participant completed twelve conjoint tasks, each involving four options, for each product. We also included the fifth option of "an ear-plugged wired headset at $29" for the Bluetooth wireless headset conjoint experiment.

We included four attributes for the paraphernalia bundle: jacket style (four levels: style 1-blue, style 1-grey, style 2-blue and style 2-grey); t-shirt color (two levels: blue and white); hat style (two levels: style 1 and 2); and price (four levels: $39, $49, $59 and $69). We included six attributes for the Bluetooth wireless headset: brand (two levels: Motorola and Plantronics); color (two levels: black and silver); weight (two levels: 8 and 13 grams); talk time per battery charge (two levels: 5 and 8 hours); noise cancelation (two levels: yes and no); and price (four levels: $49, $59, $69 and $79). For each product, a blocked design involving ten sets of twelve quadruples were created using SAS OPTEX. Participants were randomly assigned to one of the ten sets.

For network information, we first asked participants to nominate ten students in the class he or she had interacted with most frequently[3]. To ensure privacy, we provided the subjects with an option to indicate to which nominees they were unwilling to disclose their choices. At the end, we asked participants to enter two online raffles. We collected demographic information such as name, gender, most frequently used email, time spent on email every day, confidence in apparel taste, knowledge about Bluetooth headset, and interest in winning the products for students who agreed to participate in the raffles.

In Phase 2, we launched the online raffle for the university paraphernalia bundle.

[3] To facilitate the nomination process, our JAVA program has a built-in search tool. Participants could easily locate the nominees' names by typing part of their first or last names. The search tool helps us avoid typos and solves the problem that participants may not necessarily remember the full name of all the nominees.

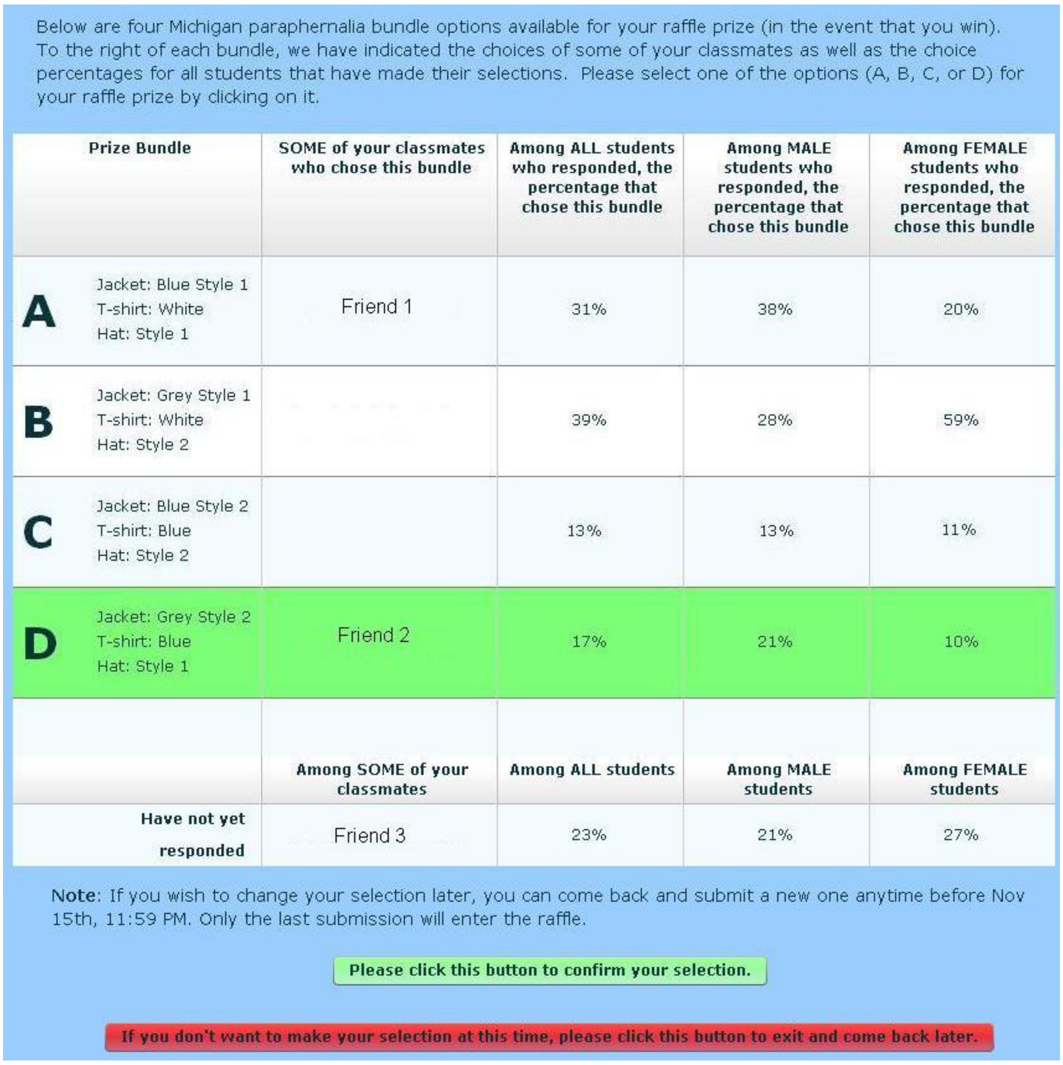
All participants received emails with the instruction to enter the raffle at the same time. A link included in the email directed them to a website, where they were asked to select a paraphernalia bundle they wish to receive should they win the raffle. We provided participants with four paraphernalia options describing by the same attribute levels as shown to them in Phase 1 (figure 3.2)[4]. The four options were chosen according to the results from a pre-test with a similar pool of participants. We expected the four options to be relatively close with regard to their utilities.

To facilitate the propagation of social influence, we showed the participants choices made by their "friends" (micro-level relation), as well as the aggregate proportions of participants choosing each option (macro-level relation: male vs. female). The aggregate proportions were computed based on: 1) all participants who had already chosen an option; 2) only female participants (i.e., a reference group) who had already chosen an option; and 3) only male participants who have already chosen an option at a particular time point. The term "friends" here are referred to as other participants who had agreed to share preferences with each participant in Phase 1. Participants could change their selections as often as they wished before the end of the raffle period.

The novelty of the experiment stems from the real-time delivery of friends' choices during the raffle period. The procedure works as follows. The first time a participant made a choice or every time him/her changed a choice, emails were sent out to his or her "friends". We recorded everyone's actions. Our database was updated real-time to assure that the website offered up-to-date information. To increase the response rate, we sent a reminder to enter the raffle in the middle of the raffle period.

Phase 2 lasted for a week, and then we launched the second raffle for the Bluetooth

[4] We omitted price, as including price may have prompted students to choose the most expensive bundle.



Below are four Michigan paraphernalia bundle options available for your raffle prize (in the event that you win). To the right of each bundle, we have indicated the choices of some of your classmates as well as the choice percentages for all students that have made their selections. Please select one of the options (A, B, C, or D) for your raffle prize by clicking on it.

Prize Bundle	SOME of your classmates who chose this bundle	Among ALL students who responded, the percentage that chose this bundle	Among MALE students who responded, the percentage that chose this bundle	Among FEMALE students who responded, the percentage that chose this bundle
A Jacket: Blue Style 1 T-shirt: White Hat: Style 1	Friend 1	31%	38%	20%
B Jacket: Grey Style 1 T-shirt: White Hat: Style 2		39%	28%	59%
C Jacket: Blue Style 2 T-shirt: Blue Hat: Style 2		13%	13%	11%
D Jacket: Grey Style 2 T-shirt: Blue Hat: Style 1	Friend 2	17%	21%	10%

	Among SOME of your classmates	Among ALL students	Among MALE students	Among FEMALE students
Have not yet responded	Friend 3	23%	21%	27%

Note: If you wish to change your selection later, you can come back and submit a new one anytime before Nov 15th, 11:59 PM. Only the last submission will enter the raffle.

Please click this button to confirm your selection.
If you don't want to make your selection at this time, please click this button to exit and come back later.

Figure 3.2: Raffle website for bundle of sports paraphernalia

wireless headset in Phase 3. The set-up of the second raffle is very similar to the first one, except that if participants did not make a selection, the best product they could receive was a wired headset at a much lower price than any of the Bluetooth wireless options. The raffle was closed after a week, and two reminders were sent during this period. Our experimental design makes it possible for us to obtain both snap-shot and complete data. Complete data come from our track of the logon activities and confirmed choice selections. Snap-shot data are all the selections we observed by the time raffles closed.

3.5.1 Empirical Analysis

Sample Profile

In Phase 1, we have 292 participants in total, among which 215 agree to participate in the raffles. All the analysis hereafter is based on these 215 participants. The demographic profile of all the participants (292) is very similar to those in the raffle (215) (figure 3.3). Thus, the raffle participants are considered a random sample of all the participants from Phase 1 with respect to most demographic information.

Among the 215 raffle participants, 39.1% are female and 18.6% are categorized as "experts" on the Bluetooth headset (we asked participants to rate themselves on a 1-7 scale in Phase 1 how knowledgeable they are about Bluetooth headset and experts are those with ratings higher than 5). More than half of the participants are "very interested" in winning a paraphernalia bundle with a rating of 7 on a 1-7 scale (1 is "not interested all at"), and the mean rating is 6.0. For the Bluetooth headset, students' interests are more evenly distributed, although one-third of the ratings are 7, and the mean rating is 4.5 (on a 1-7 scale).

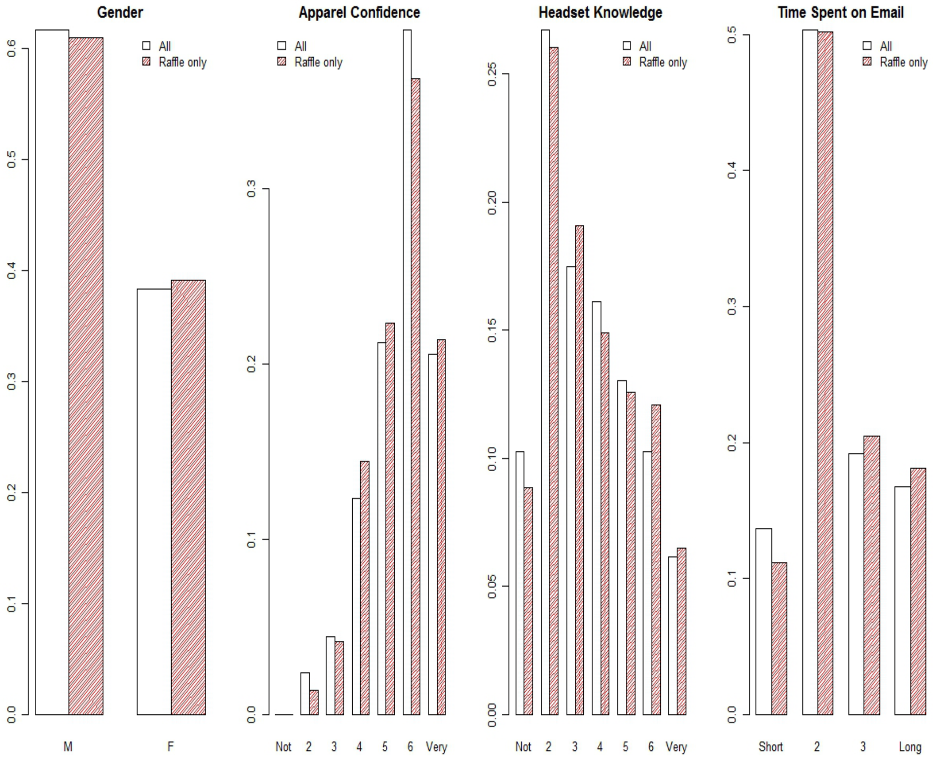


Figure 3.3: Demographic information

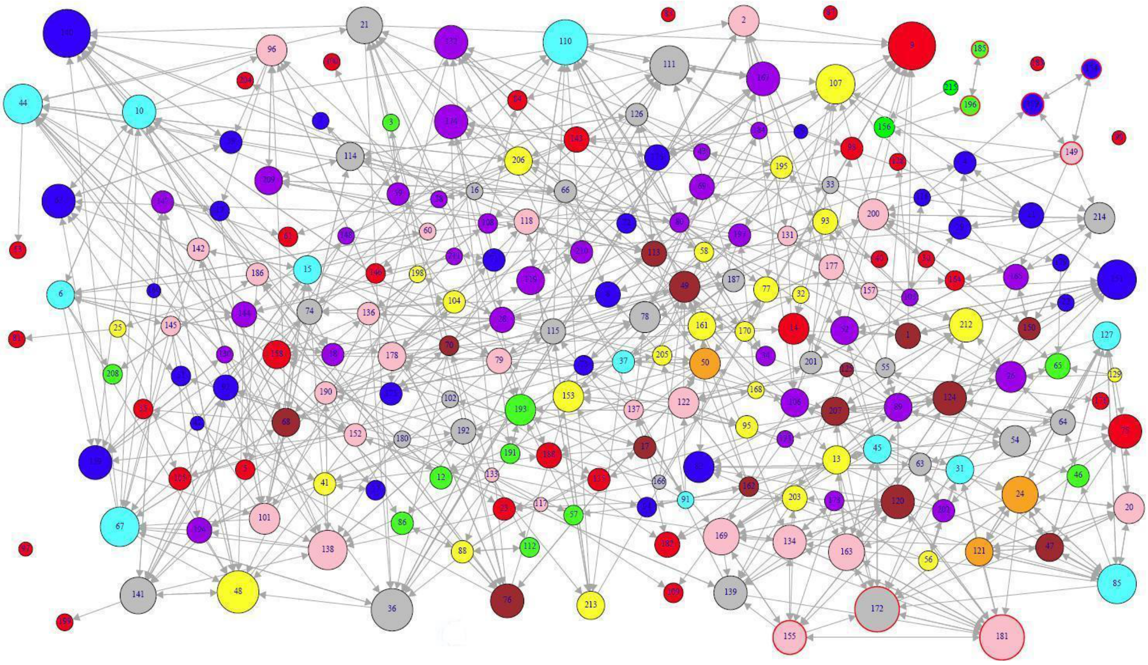


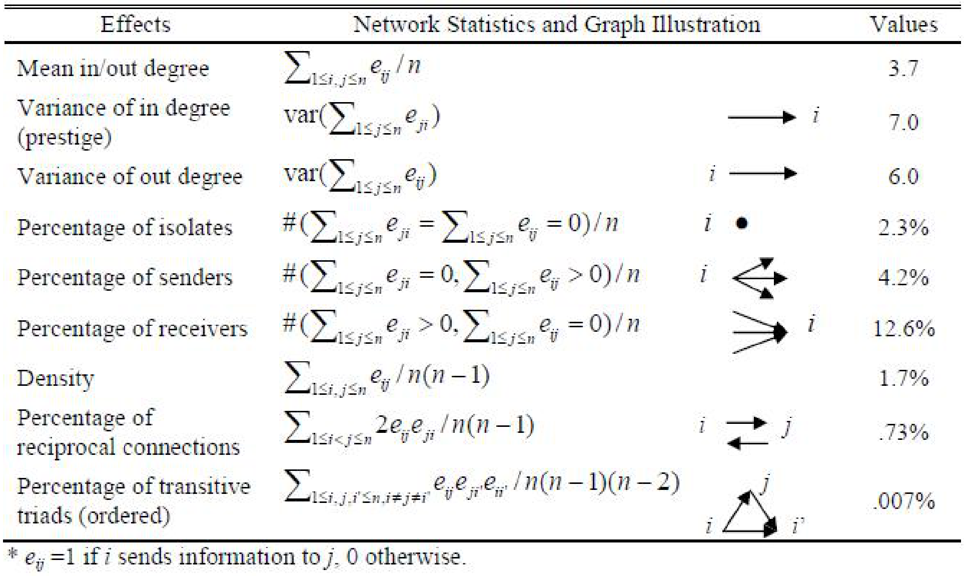
Figure 3.4: Network plot

Social Network

Figure 3.4 depicts the directed social network among the raffle participants. Each node represents a participant. An arrow represents a connection. A node with larger size indicates a more popular participant with more friends (i.e., receivers with arrows in), and nodes in the same color have the same number of arrows out (i.e., senders). Recall that in our experiment, ”friends” are defined as people who nominated a particular participant as his or her friends in Phase 1. Notice that there are five socially peripheral participants with no arrow in or out. Among the 628 pairs of connected participants (the same two individuals are only counted as one pair), 26.6% of them have arrows both in and out present (e.g., they have reciprocal connections).

On average, every participant has 3.7 friends. The most popular participants

Table 3.2: Network summary statistics



Effects	Network Statistics and Graph Illustration		Values
Mean in/out degree	$\sum_{1\le i,j\le n} e_{ij} / n$		3.7
Variance of in degree (prestige)	$\mathrm{var}(\sum_{1\le j\le n} e_{ji})$	$\rightarrow i$	7.0
Variance of out degree	$\mathrm{var}(\sum_{1\le j\le n} e_{ij})$	$i \rightarrow$	6.0
Percentage of isolates	$\#(\sum_{1\le j\le n} e_{ji} = \sum_{1\le j\le n} e_{ij} = 0) / n$	$i\ \bullet$	2.3%
Percentage of senders	$\#(\sum_{1\le j\le n} e_{ji} = 0, \sum_{1\le j\le n} e_{ij} > 0) / n$	i	4.2%
Percentage of receivers	$\#(\sum_{1\le j\le n} e_{ji} > 0, \sum_{1\le j\le n} e_{ij} = 0) / n$	i	12.6%
Density	$\sum_{1\le i,j\le n} e_{ij} / n(n-1)$		1.7%
Percentage of reciprocal connections	$\sum_{1\le i<j\le n} 2e_{ij}e_{ji} / n(n-1)$	$i \rightleftarrows j$	.73%
Percentage of transitive triads (ordered)	$\sum_{1\le i,j,i'\le n, i\ne j\ne i'} e_{ij}e_{ji'}e_{ii'} / n(n-1)(n-2)$	j i i'	.007%

* e_{ij} =1 if i sends information to j, 0 otherwise.

have 12 friends, while 14 participants have no friends. Table 2 provides some network structure statistics. As expected, the density, reciprocity and transitivity are low as we work with a relatively large network.

Raffle Choice Summary

For the paraphernalia bundle, 167 participants visit the website, and 62.3% of them are male, which is similar to the proportion of males among the 215 raffle participants. Among the 25 participants who visit the site multiple times, 17 do not change their choices in all visits, whereas the remaining eight first did not choose any option but then pick an option later. The percentages of people choosing options 1 to 4 and not choosing any of the options are 30.5%, 38.9%, 9.0%, 16.8%, and 4.8%, respectively.

For the Bluetooth headset, 97 participants visit the website, and 25.8% are ”experts”, which is higher than the percentage of experts among the 215 enrolled (18.6%). The main reason could be that experts are more interested in winning a

Bluetooth headset than novices, as the mean ratings of interest are 5.9 and 4.2 for experts and novices, respectively. Among the 97 participants, 92 also logged on to the bundle raffle website. Thirty-four participants visit the site multiple times, and 28 of them maintain the same choices in all the visits. The remaining 6 first did not choose an option but later choose one. The percentages of people choosing options 1to 4 and not choosing any of the options are 14.4%, 3.1%, 74.2%, 2.1%, and 6.2%, respectively.

Model Estimation

We fit three models to the experimental data: model 1, a baseline hierarchical Bayes logit model for Phase 1 data only; model 2 the same baseline model for Phase 1 data and the discrete-time Markov chain model (3.1) for complete raffle data from Phase 2 and 3, respectively; and model 3 the same baseline model for Phase 1 data and the MRF model with multiple relations (3.3) to bundle raffle data and asymmetric effects to the Bluetooth headset raffle data (3.2).

Hierarchical Bayes Model

We use the hierarchical Bayes logit model (Allenby and Rossi 1999) to fit the conjoint data from Phase 1. In the absence social network information, the best guess for a participant's choice in the raffle is based on his or her initial preferences for different product attributes. The model specification is as follows.

Participant i's utility of option k is $u^I_{ik} = \sum_{p=1}^{P} \beta_{ip} x^I_{pk} + \epsilon^I_{ik}$, where β_{ip} captures individual i's preference for the p-th P is the number of product attributes, x^I_{pk} indicates the p-th attribute's value for product k, ϵ^I_{ik} is an error term following type I extreme distribution, and the superscript I means it is related to initial preferences. All the attributes are dummy-coded including the price. In the Bluetooth headset case, the utility for the wired headset is u^I_{i0}=0, consequently those for Bluetooth wireless headsets have an intercept term β_0. Hereafter, we mainly use the model for the paraphernalia bundle as an illustration. Let $n = 215$, $\beta_i = (\beta_{i1}, \cdots, \beta_{iP})$, $\beta = (\beta_1, \cdots, \beta_n)$, i's choices in 12 conjoint tasks $Y^I_i = (Y^I_{i1}, \cdots, Y^I_{i12})$, the corresponding attributes for option k in q-th task $x^I_{iq,k} = (x^I_{iq,1k}, \cdots, x^I_{iq,Pk})$, $Y^I = (Y^I_1, \cdots, Y^I_n)$, and $x^I = (x^I_{11,1}, \cdots, x^I_{n12,4})$, then the likelihood is given by:

$$P(y^I|x^I, \beta) = \prod_{i=1}^{n}\prod_{q=1}^{12} \frac{\exp\{\beta_i^T x^I_{iq,y^I_{iq}}\}}{\sum_{k=1}^{4}\exp\{\beta_i^T x^I_{iq,k}\}}. \tag{3.6}$$

We follow the conventional prior setting of a hierarchical Bayes model: $b \sim N(0, 100I_{P\times P})$, $V_\beta \sim IW(P+2, I_{P\times P})$, and $\beta_i \sim N(b, V_\beta)$, where $I_{P\times P}$ is a $P \times P$ identity matrix.

Discrete-time Markov Chain

We fit a discrete-time Markov chain model to complete data. We only model the choices of participants who at least log on to the raffle website, and treat the rest as if they decline the raffle in Phase 1. However, the information that these participants did not choose any option still enters the model through the calculation of our network statistics. To model "who intends to change a choice" at time point t+1, c^{t+1}, we use a logit model with utility $u^{t+1}_i = \alpha e_i + \epsilon^{t+1}_i$, where α is the parameter, e_i is i's interest in winning the product taking values from 1 to 7, the error term ϵ^{t+1}_i follows a type I extreme value distribution. The intuition is that participants are more likely to make their choice if they are more interested in winning the product. Its likelihood is $P(c^{t+1} = i|c^t, y^t, x, \alpha) = \frac{\exp\{u^{t+1}_{c^{t+1}}\}}{\sum_{i=1}^{n}\exp\{u^{t+1}_i\}}$.

Conditional on $c^{t+1} = i$, we next specify the component of the model that capture "which choice a participant makes". For the paraphernalia bundle raffle, in case of participants not choosing any option, the attributes of this option are defined as the mean values of the four options because we inform participants that if they do not make a raffle choice, they will receive a randomly chosen option as their prize. In addition, we include an intercept to account for unobserved rationale behind this "no-choice" option and treat it as an "outside good" utility. g_i is i's gender.

We conjecture two sources of dependence: micro-level friends' influence and macro-level influence from some reference group. Our network is directed with senders sharing choice decisions with receivers. We expect the sender to influence the receiver, which is a micro-level relationship. To account for the influence of this relationship, we sum up the number of an individual's senders who make the same choice as the individual (receiver) by the time he/she makes the choice. To control for a different number of friends (i.e., senders), we normalize the summation by the number of friends each receiver has. Thus, this binary network assigns a weight of 1 to every connection.

On the raffle website, we also make available the summary information on the choices made among different groups (i.e., reference groups) of participants. We use the information about choices of the reference group to capture the influence from the macro-level relationship. In other words, we construct another network with people in the same reference group as neighbors. We expect that participants tend to choose the option that more members in the same reference group have chosen.

We further assume that not choosing any option is not affected by any network effect. Let 0 be the "no-choice" option, $x = (x_1, \cdots, x_4)$ with x_k the vector of attributes of option k in the raffle, $\bar{x} = \sum_{k=1}^{4} x_k/4$ be the mean values for all the attributes, and γ be a vector of parameters associated with the network effect. The probability of each choice made at time $t+1$, $P(y_i^{t+1}|c^{t+1} = i, c^t, y^t, x, \beta, \gamma, g, W)$, has a logit form with:

$$\begin{aligned} u^{t+1}_{iy_i^{t+1}} &= \mathbf{1}(y_i^{t+1}=0)(\gamma_0+\gamma_1\beta_i^T\bar{x}) + \mathbf{1}(y_i^{t+1}>0)\gamma_1\beta_i^T x_{y_i} + \\ &\gamma_2\mathbf{1}(y_i^{t+1}>0)\frac{\sum_{j=1}^m w_{ji}\mathbf{1}(y_i^{t+1}=y_j^t)}{\sum_{j'=1}^n w_{j'i}} + \\ &\gamma_3\mathbf{1}(y_i^{t+1}>0)\frac{\sum_{j=1,\neq i}^m \mathbf{1}(y_i^{t+1}=y_j^t, g_i=g_j)}{\sum_{j=1,\neq i}^m \mathbf{1}(y_j^t>0, g_i=g_j)} + \epsilon^{t+1}_{ty_i^{t+1}}, \end{aligned} \tag{3.7}$$

where γ_0 is the intercept for the "no-choice" option, γ_1 adjusts the magnitude of the initial preferences (i.e., a scale parameter), γ_2 is the local friends' influence (i.e., micro-level relationship), and γ_3 is the reference group influence (i.e., macro-level relationship). $W=(w_{ij})_{1\le i\le j\le n}$ is the matrix of binary directed network with $w_{ij}=1$ if person i sends information to j and 0 otherwise. We also define $\frac{0}{0}=0$ for all the ratios. Note that the denominator of the last summation is the number of participants in a reference group who chose a particular option.

For the Bluetooth headset raffle data, the likelihood is different, as our main focus is on examining the asymmetric interactions between students who are knowledgeable (i.e., experts) and not knowledgeable (i.e., novices) about Bluetooth. Also, selection of 0 is modeled differently as $-\gamma_0\beta_{i0}$, since it represents a wired headset at \$19.99. $P(y_i^{t+1}|c^{t+1}=i, c^t, y^t, x, \beta, \gamma, g, W)$ is:

$$\begin{aligned} u^{t+1}_{iy_i^{t+1}} &= -I(y_i^{t+1}=0)\gamma_0\beta_{i0} + I(y_i^{t+1}>0)\gamma_1\beta_i^T x_{y_i} + \\ &\gamma_2\mathbf{1}(y_i^{t+1}>0,\ \text{i is novice})\frac{\sum_{j=1}^m w_{ji}\mathbf{1}(y_i^{t+1}=y_j^t,\ \text{j is expert})}{\sum_{j'=1}^n w_{j'i}} + \\ &\gamma_3\mathbf{1}(y_i^{t+1}>0,\ \text{i is expert})\frac{\sum_{j=1}^m w_{ji}I(y_i^{t+1}=y_j^t,\ \text{j is novice})}{\sum_{j'=1}^n w_{j'i}} + \epsilon^{t+1}_{ty_i^{t+1}}. \end{aligned} \tag{3.8}$$

Thus γ_2 reflects novice's choice dependence on expert's, and γ_3 reflects expert's on novice's.

Observing the exact choices' sequence and assuming that only one choice can be made at a particular time point allows us to use a simple logit specification to model choice made at each time point. The full likelihood is specified as:

$$\begin{aligned}
&p(s^1, \cdots, s^T | s^0, x, \beta, \gamma, g, W) \\
= &\prod_{t=0}^{T-1} \frac{\exp\{u^{t+1}_{c^{t+1}}\}}{\sum_{i=1}^{n} \exp\{u_i^{t+1}\}} \frac{\exp\{u^{t+1}_{c^{t+1} y^{t+1}_{c^{t+1}}}\}}{\sum_{k=\mathbf{1}(y^t_{c^{t+1}}=0)}^{4} \exp\{u^{t+1}_{c^{t+1}k}\}} \mathbf{1}(y^{t+1}_{-c^{t+1}} = y^t_{-c^{t+1}}).
\end{aligned}$$

We specify $c^0 = 0$, meaning nobody intends to change before the raffle starts, and y^0 is a vector of 0, indicating that everyone starts without any selection. Note that the likelihood of choices y^{t+1} for the discrete-time Markov chain is very similar to the pseudolikelihood for the MRF, as they both involve the multiplication of choice probabilities across participants. The difference is in the network statistics. Specifically, statistics in the discrete-time Markov chain are computed based on choices up to the previous time point, whereas statistics in the pseudolikelihood are computed based on choices observed at the last time point.

We include all 215 people's data in Phase 1 to get better estimates of initial preferences. The full likelihood contains both conjoint and complete raffle data. $P(s^1, \cdots, s^T, y^I | s^0, x, x^I, \beta, \gamma, g, W) = P(s^1, \cdots, s^T | s^0, x, \beta, \gamma, g, W) P(y^I | x^I, \beta)$, where $p(y^I | x^I, \beta)$ is defined in (3.6). The prior for the new parameters is $\gamma \sim N(0, 100 I_{4\times 4})$, the rest is the same as those specified for the baseline model.

We carry out the Metropolis-within-Gibbs algorithm for 10,000 iterations for all the models (see Appendix 0.1.2 for details), and use every tenth draw from the last 8,000 for model estimation and inference. According to the trace plots and auto-correlation plots with 10,000 iterations, the Markov chain has converged and has good mixing. To perform a hypothesis test, we calculate the probability of a parameter greater than 0 based on the draws.

Markov Random Field with Multiple Relations

We fit the MRF to snap-shot data, which correspond to participants' final choices when the raffle is closed. For the paraphernalia bundle data (3.9), we fit the MRF with multiple relations (micro vs. macro level of relationships). For the Bluetooth headset data (3.10), we fit MRF with only the micro-level relationship, but we also account for asymmetry in social interactions.

Apparently, we have a directed network of sharing choice decisions. We believe the conditional probability of one's choice depends on both senders and receivers (i.e., two individuals are neighbors as long as at least one of them shares choice decisions with the other). When specifying the potentials (U_c), we focus on cliques of size one and two. Potentials related to cliques of size one capture the effects of each individual's initial preference. For cliques of size two, i.e., any sender and receiver, the micro-level interaction is whether the sender and receiver choose the same option, normalized by the number of friends the receiver has.

As mentioned above, to account for additional source of dependence, we build another network, with people in the same reference group as neighbors. The macro-level interaction in cliques of size two is again whether the two neighbors choose the same option, normalized by the number of friends they have who choose a none-zero option (notice individuals in the same reference group have the same number of friends).

Let $m = 167$, $y = (y_1, \cdots, y_m)$ representing 167 participants' final choice decisions. The specification of the MRF with multiple relations (fit to the paraphernalia bundle raffle data) described above is:

$$\begin{aligned}p(y|x,\beta,\gamma,g,W) = \exp\Big\{ &\sum_{i=1}^{m}[\mathbf{1}(y_i=0)(\gamma_0+\gamma_1\beta_i^T\bar{x}) + \mathbf{1}(y_i>0)\gamma_1\beta_i^T x_{y_i}] + \\ &\sum_{1\le i,j\le m}\gamma_2 w_{ij}\mathbf{1}(y_j>0)\frac{\mathbf{1}(y_j=y_i)}{\sum_{i'=1}^{n} w_{i'j}} + \\ &\sum_{h=1}^{2}\gamma_3\sum_{i=1}^{m}\mathbf{1}(y_i>0)\frac{\sum_{j=1,\neq i}^{m}\mathbf{1}(y_i=y_j,g_i=g_j=h)}{\sum_{j=1,\neq i}^{m}\mathbf{1}(y_j>0,g_i=g_j=h)}\Big\}/Z(\beta,\gamma). \qquad (3.9)\end{aligned}$$

We can find the counterpart of each parameter in the Markov chain model above. The first summation involves cliques of size one; the second summation involves cliques of size two that capture the micro-level relationship; and the last summation involves cliques of size two that capture the macro-level relationship.

For Bluetooth headset raffle, there are 97 participants in this data set. The specification allows asymmetric interactions between novices and experts:

$$\begin{aligned}p(y|x,\beta,\gamma,g,W) = \exp\Big\{ &\sum_{i=1}^{m}[-I(y_i=0)\gamma_0\beta_{i0} + I(y_i>0)\gamma_1\beta_i^T x_{y_i}] + \\ &\sum_{\text{i is expert, j is novice}}\gamma_2 w_{ij}\mathbf{1}(y_j>0)\frac{\mathbf{1}(y_j=y_i)}{\sum_{i'=1}^{n} w_{i'j}} + \\ &\sum_{\text{i is novice, j is expert}}\gamma_3 w_{ij}\mathbf{1}(y_j>0)\frac{\mathbf{1}(y_j=y_i)}{\sum_{i'=1}^{n} w_{i'j}}\Big\}/Z(\beta,\gamma). \qquad (3.10)\end{aligned}$$

We can also find the counterpart of each parameter in the Markov chain for Bluetooth data above.

The full likelihood contains both conjoint and final raffle choices, $p(y,y^I|x,x^I,\beta,\gamma,g,W) = p(y|x,\beta,\gamma,g,W)p(y^I|x^I,\beta)$, where $p(y^I|x^I,\beta)$ is defined in (3.6). The prior is the same as in Markov chain models. Note that although we estimate initial preferences at the individual level, we can only estimate the effects of choice interdependence at the aggregate level, as we only observe one choice (from the raffle) for each individual in each product category.

We have two algorithms to fit the MRF model. For the approximate sampling approach (with homogeneous instrument kernel), we sample from the posterior distribution derived from the full likelihood using the algorithm described above (see Appendix 0.1.4 for details), with $\kappa_i = 10\log(i+1)+100$. The pseudolikelihood approach involves replacing the complicated likelihood of MRF with the corresponding pseudolikelihood. With the same prior as above, we can obtain posterior draws from $p_{PL}(\theta|y)$ by running the Metropolis-within-Gibbs algorithm (Appendix 0.1).

Table 3.3 reports the posterior means of the model parameters. In general, both the pseudolikelihood and approximate sampling methods lead to very similar results for the MRF model. Since complete data contain more information than the snapshot data, the magnitude of the parameters are not really comparable for the MRF and the discrete-time Markov chain model. Nonetheless, the parameters from both models seem to yield similar conclusions about the patterns of choice interdependence in the social network.

All the parameters related to individuals' heterogeneous intrinsic preferences (i.e., initial preferences) for different product attributes appear reasonable. The scale parameter for initial preference is significant for both categories suggesting that participants' intrinsic preferences still play a significant role in determine participants' choices of products in the raffles. For the paraphernalia bundle raffle, we find participants rely on the choices of others both at the micro and macro levels in making their decisions. For Bluetooth headset raffle, we account for potential asymmetry in social interactions between experts and novices. As expected, experts have stronger influence on novices' choices than vice versa. In the discrete-time Markov choice model, we also model "who intends to change a choice" and find that for the Bluetooth headset raffle, participants are more likely to make choices if they have stronger interest in winning a Bluetooth headset. Overall, the magnitude of choice interde-

Table 3.3: Parameter estimates

	Parameter	Markov Chain	Markov Random Field	
			PL	AS
Bundle	**Initial Preference (β):**			
	Style 1 blue jacket[1]	1.03**	1.00**	1.27**
	Style 1 grey jacket	1.49**	1.40**	1.81**
	Style 2 blue jacket	-.177	-.262	-.333
	Blue t-shirt[2]	.302**	.306**	.356**
	Style 1 hat[3]	.247**	.266**	.333**
	\$39[4]	3.29**	3.16**	3.98**
	\$49	2.59**	2.45**	3.13**
	\$59	1.30**	1.14**	1.43**
	Social Influence:			
	Outside good (γ_0)	.377	-.685**†	-.756**†
	Scale parameter for initial preference (γ_1)	.502**	.461**	.366**
	Local friends' influence (γ_2)	.828**	.632**	.566**
	Social norm effect (γ_3)	2.21**	1.23**	1.25**
	Product interest effect on who will change (α)	.002	n/a	n/a
Bluetooth	**Initial Preference (β):**			
	Ear-plug headset at \$29 ($-\beta_0$)	-1.18**†	-1.39	-.599
	Motorola[5]	.540**	.801**	.783**
	Black[6]	-.171**†	-.106	-.254**†
	8 grams[7]	1.40**	1.53**	1.36**
	8 hours[8]	1.55**	1.64**	1.64**
	Noise cancellation	1.71**	1.82**	2.01**
	\$49[9]	3.92**	4.27**	3.97**
	\$59	2.26**	2.40**	2.26**
	\$69	.700**	.919**	.611**
	Social Influence:			
	Outside good (γ_0)	-.003	-.020	-.060
	Scale parameter for initial preference (γ_1)	.467**	.353**	.333**
	Local friends' influence (experts to novices; γ_2)	6.92**	4.33**	4.62**
	Local friends' influence (novices to experts; γ_3)	5.96**	2.09*	3.34**
	Product interest effect on who will change (α)	.056*	n/a	n/a

* significant at .9 level, ** significant at .95 level, † probability of parameter less than or equal to 0.

Reference level: [1] style 2 grey jacket, [2] white t-shirt, [3] style 2 hat, [4] \$69, [5] Plantronics, [6] silver, [7] 13 grams, [8] 5 hours, [9] \$79.

pendence effects appears to be larger for the Bluetooth headset raffle than for the paraphernalia bundle raffle.

Model Comparison

In this section, we compare the three models described above using in-sample and out-of-sample hit rates. A hit rate is computed as the mean percentage of correctly predicted choices among all participants. We opt to use a hit rate instead of traditional fit measures such as log marginal density and DIC because these traditional measures require the calculation of the normalizing constant of the Markov random field, which is computationally expensive.

More specifically, for log marginal density, following Chib and Jeliazkov (2001), although we can pick some θ such that the normalizing constant is known, we still need to evaluate the acceptant ratio for different proposed θ'. If θ' is far away from θ, there is non-negligible error in the estimation. As a result, log marginal density may not be accurately estimated and hence not a good measure for comparison in this case. The computation of DIC has a similar problem. On the contrary, hit rate is easy to obtain and is usually of interest in many real world problems. Moreover, as shown in the simulation study, hit rate is less sensitive to the difference and therefore makes our conclusion more conservative.

To make the results more comparable across models, in the discrete-time Markov chain model the two hit rate measures are computed only for final choices, and in prediction, we condition on who intends to change the choice. We randomly pick 10% of the respondents and exclude all of their choices for the out-of-sample fit.

Making predictions with the baseline model is straightforward. Because people's choices are treated as independent, we can draw each prediction from its own posterior predictive probability $P(y_i^*|D) = \int P(y_i^*|\theta)p(\theta|D)d\theta \approx \sum_l P(y_i^*|\theta_l)$, where θ_l is the posterior draw. Prediction for the remaining models is quite different from the traditional in-sample or out-of-sample prediction for independent data. Interdependent choices should be predicted jointly rather than separately. When data are left out, we cannot fit the model to only the remaining data as it destroys the dependence structure in the network. We also cannot simply assign a "no-choice" option for the out-of-sample participants. To preserve the dependence structure in the data, we draw from $P(y^*|D) = \int P(y^*|\theta)p(\theta|D)d\theta$ for in-sample prediction. More specifically, we draw parameters along with the predictions from$P(\theta, y^*|D) \propto P(D|\theta)P(y^*|\theta)\pi(\theta)$. For out-of-sample prediction, we fit the model with all the data and treat the out-of-sample data as missing values. In other words, we draw parameters and predictions together from $P(\theta, y_{\text{out}}^*|D_{in}) \propto P(D_{in}, y_{out}^*|\theta)\pi(\theta)$, where y_{out}^* is the out-of-sample prediction and D_{in} is the in-sample data.

Results from the choice prediction exercise are shown in table 3.4. In most cases, discrete-time Markov chain and MRF models perform better than the baseline model, especially for the Bluetooth headset raffle. These findings suggest the significance of the effects of choice interdependence in a social network. We expect choice interdependence to play a more important role in the Bluetooth raffle because students are less likely to have strong intrinsic preference for such a high-tech product, as compared to a paraphernalia bundle, and hence rely on their peers' choices to make their own choice decisions. For Markov random field model, approximate sampling method performs better than pseudolikelihood, particularly in the Bluetooth headset raffle, where the social effect is stronger.

Table 3.4: Model comparison

Measure (%)		Markov Chain	Markov Random Field		Baseline
			PL	AS	
Bundle	In-sample hit rate	39.9 (.262)	41.1 (.278)	42.5 (.483)	39.7 (.071)
	Out-of-sample hit rate	41.8 (.234)	42.0 (.079)	42.3 (.267)	38.0 (.375)
Bluetooth	In-sample hit rate	37.9 (.674)	33.5 (.355)	37.5 (.140)	28.5 (.184)
	Out-of-sample hit rate	39.0 (.211)	34.5 (.277)	38.7 (.185)	26.2 (.405)

PL stands for pseudolikelihood. AS stands for approximate sampling.
Within parenthesis is the standard error of the hit rate.

3.6 Discussion and Extension

The work in this chapter makes several contributions. First, we derive necessary and sufficient conditions for the discrete-time Markov chain to converge to the Markov random field, and hence show that Markov random field is a natural model that emerges from the Markov chain dynamic choice process. Second, we propose a modeling framework for examining choice interdependence in a social network. Specifically, we propose a discrete-time Markov chain to model complete data and the Markov random field (MRF) which can accommodate multiple relations to model snap-shot data. Finally, we demonstrate how one can use either a pseudolikelihood or an approximate sampling method to circumvent the intractable normalizing constant problem in estimating the MRF.

Our results show that controlling for individuals' heterogeneous preferences, individuals rely on choices observed both at the micro (i.e., closely-connected friends) and macro (i.e., a reference group) levels in making their decisions. Choice interdependence is shown to play a more important role for products that are not familiar to consumers (e.g., a high-tech product such as Bluetooth headset). We also find asymmetry in social interactions across different types of consumers (i.e., expert vs. novice). With regard to model comparison, the discrete-time Markov chain provides more insights about the dynamic process of choice interdependence (”who” intends to change a choice and ”which” choice is made) in a social network but does not necessarily predict choices better than the MRF model. Our simulation shows that the approximate sampling method performs better than the pseudolikelihood method for small and medium-sized networks. The difference in their performance dissipates as size of the network increases. As such, the pseudolikelihood method which is less computationally intensive may be more desirable for large networks.

Here we propose a discrete-time Markov chain to model the sequences of choices with regard to ”who intends to change a choice.” However, additional information on the time between two consecutive choices will allow us to also model ”how long it takes for someone to change a choice.” It is generally difficult to track the time between two consecutive choices. Even with the social media technology, it can become cumbersome to recorded larger amount of data. However, if such data are available, a continuous-time, instead of a discrete-time, Markov chain model can be applied to study the factors that influence the time it takes for someone to change a choice. Modeling this type of data is beyond the scope of this thesis, but can be a fruitful avenue for future research.

Finally, we want to briefly mention two other avenues of future research. Firstly, in this chapter, we are mainly interested in the behavior interaction due to the nature of the interactive advertising. In other marketing problems, if we believe preference interaction is the major source of dependence, the mixed models in section 1.1.2 offer a general framework. Since choice data are non-Gaussian, the model specifications in section 1.1.2 will apply. In the marketing context, for multinomial choice data, we have individual i's utility for product option k is

$$u_{ik} = \beta_i^T x_k + v_{ik},$$

where the random effects v_{ik} follow a linear Gaussian process for each k (for binary choice data, a special case is Yang and Allenby 2003). Or

$$u_{ik} = \beta_i^T x_k + v_{ik} + \epsilon_{ik}, \tag{3.11}$$

where random effects v_{ik} follow a Gaussian MRF including CAR and ICAR for each k, and $(\epsilon_{i1}, \cdots, \epsilon_{iK})$ is independent Multivariate Gaussian random variable.

Secondly, the model can be extended to allow for potential heterogeneous choice or preference interactions. In this chapter, we already showed an example of asymmetric interactions from two different types of students. The identifiability here comes from the clearly defined distinct 'roles' (experts and novices) of individuals and a directed network, and the asymmetry is at the aggregated level, i.e., a global parameter for all experts' associations with novices. In section 1.1.3, we summarized some existing models for heterogeneous interactions. They can also be adopted in the marketing context. For choice interaction, an inhomogeneous MRF model can be used. For instance, we can have c_{ij} capture social interaction strength for each pair of neighbors (i, j). Then the parameterizations in 1.1.3 will apply. For preference interaction, we can employ the mixed model framework (3.11), but impose an allocation model on the random effects or allow them to follow an inhomogeneous Gaussian MRF.

CHAPTER IV

Dynamic Path Sampling

4.1 Introduction

As we mentioned in section 2.1.1, if Metropolis-Hasting algorithm is used to sample from the doubly-intractable distributions, we only need to know the ratio of normalizing constants. Methods such as bridge sampling (Meng and Wong 1996) and path sampling (Gelman and Meng 1998) were developed to estimate the ratio. In the case where normalizing constant $Z(\theta)$ itself rather than the ratio is desired, we can always find a particular θ_0 with $Z(\theta_0)$ easy to compute, such as $Z(0)$ in (2.1), and obtain $Z(\theta)$ through the ratio. Therefore, apart from doubly-intractable distributions, estimating the ratio of normalizing constants has a wide range of applications. For example, the likelihood ratio test involves the ratio of normalizing constants. In most cases, the normalizing constants are easy to compute, but could be quite cumbersome in the presence of missing or latent data. In some applications, marginal density can be considered as a normalizing constant. Its evaluation problem becomes the problem of estimating normalizing constants, and computing Bayes factor in a Bayesian framework turns to the question of estimating the ratio. We refer the readers to Meng and Wong (1996) and Gelman and Meng (1998) for a thorough literature review of methods to compute ratio of normalizing constants.

In this chapter, we propose a flexible way of implementing the path sampling identity via a non-homogeneous Markov chain, which we call dynamic path sampling. The method is particulary efficient when the two distributions, the ratio of whose normalizing constants we compute, are 'far apart'. We will review path sampling, propose dynamic path sampling, and study its theoretical properties in section 4.2. It is followed by a preliminary simulation study to demonstrate the performance of dynamic path sampling comparing to importance sampling and conventional implementation of path sampling.

4.2 A Dynamic Implementation of the Path Sampling Identity

Let $\{p_\theta,\ \theta \in \Theta\}$ be a family of probability distributions on a sample space $\mathcal{X}$. We assume that

$$p_\theta(dx) = \frac{f_\theta(x)\lambda(dx)}{Z(\theta)},$$

where f_θ is a positive un-normalized density and $Z(\theta) = \int_{\mathcal{X}} f_\theta(x)\lambda(dx)$ the normalizing constant. λ is a reference σ-finite measure on the sample space $\mathcal{X}$ that we assume equipped with a σ-algebra $\mathcal{B}$. We do not make any specific assumption on the parameter space Θ at this point. We are interested in the ratio of normalizing constants $\frac{Z(\theta_1)}{Z(\theta_0)}$ for some arbitrary points $\theta_0, \theta_1 \in \Theta$. Denote the inner product in $\mathbb{R}^d$ and the gradient operator with respect to θ by $\langle \cdot, \cdot \rangle$ and ∇_θ, respectively.

The path sampling identity (Gelman and Meng 1998) is a method to estimate the log-ratio of the normalizing constants, e.g. $\log \frac{Z(\theta_1)}{Z(\theta_0)}$. The idea is to construct a smooth path $\{f_t,\ 0 \le t \le 1\}$ of un-normalized densities between f_{θ_0} and f_{θ_1} such that $f_0 = f_{\theta_0}$ and $f_1 = f_{\theta_1}$. Define $Z(t) = \int_{\mathcal{X}} f_t(x)\lambda(dx)$ and $p_t(\cdot) = f_t(\cdot)/Z(t)$.

Assuming the legitimacy of interchange of integration with differentiation, we have

$$\begin{aligned}\frac{d}{dt}\log Z(t) &= \int_{\mathcal{X}} \frac{1}{Z(t)}\frac{d}{dt}f_t(x)\lambda(dx)\\ &= \mathbb{E}_t\Big[\frac{d}{dt}\log f_t(X)\Big],\end{aligned}$$

where $\mathbb{E}_t$ is with respect to $X \sim \pi_t$. If we write $r_{\theta_0,\theta_1} \overset{\text{def}}{=} \log\left(\frac{Z(\theta_1)}{Z(\theta_0)}\right)$ and $U(x,t) \overset{\text{def}}{=} \frac{d}{dt}\log f_t(x)$, then integrating both side of the above equation gives

$$r_{\theta_0,\theta_1} = \int_0^1 \mathbb{E}_t\Big(U(X,t)\Big)dt. \tag{4.1}$$

Equation (4.1) is the path sampling identity and gives a representation of the log-ratio of the normalizing constants as an integral of the expected log-derivative of the likelihood function along a path. This offers an effective approach to approximate r_{θ_0,θ_1} by Monte Carlo. The optimal path between any two un-normalized densities f_0 and f_1 can be found in Gelman and Meng (1998) Lemma 2. But this optimal path is rarely available in practice. One generic choice of path is the geometric path defined as

$$f_t(x) = f_{\theta_0}^{1-t}(x)f_{\theta_1}^{t}(x). \tag{4.2}$$

In this case the function U takes the form $U(x,t) = U(x) = \log f_{\theta_1}(x) - \log f_{\theta_0}(x)$ and (4.1) becomes

$$r_{\theta_0,\theta_1} = \int_0^1 \mathbb{E}_t\left(U(X)\right)dt. \tag{4.3}$$

In this paper we focus on the geometric path (4.2) and the identity (4.3). But the method developed works as well with other paths, so we will keep using $U(X,t)$ throughout the chapter. See Gelman and Meng (1998) for other examples of paths.

There are different ways to evaluate the integral in (4.3) (Gelman and Meng 1998). One way is to introduce a prior density $\mu(t)$ for $t \in [0,1]$, for example Uniform$(0,1)$, then (4.1) can be rewritten as

$$r_{\theta_0,\theta_1} = \mathbb{E}\left(\frac{U(X,\tau)}{\mu(\tau)}\right),$$

where the expectation is taken over $(X,\tau) \sim p_t(x)\mu(t)$. This yields an estimator of r:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{U(X_i,\tau_i)}{\mu(\tau_i)}$$

with draws $(X_i,\tau_i),\ i = 1,\cdots,n$, from $p(x,t) = p_t(x)\mu(t)$. The other way is to use various types of numerical integration methods. For instance, if we take the right Riemann sum with $t_1,\ldots,t_\kappa$, κ evenly spaced points in $[0,1]$, the corresponding estimator is

$$\frac{1}{\kappa}\sum_{\ell=1}^{\kappa}\mathbb{E}_t\left(U(X_{t_\ell},t_\ell)\right), \tag{4.4}$$

The expectations are estimated by Monte Carlo sampling from p_{t_l}.

In both case, we need draws from $p_t(\cdot)$. This is usually done by running a Markov chain with target distribution $p_t(\cdot)$ until its convergence. If n different points are selected in the interval $[0,1]$, then n Markov chains with different invariant distributions are needed, which is computationally intense.

4.2.1 Dynamic Path Sampling

We propose a direct approach to implement to path sampling identity (4.1) using a nonhomogeneous Markov chain. As above, let $\theta_0,\theta_1 \in \Theta$ and let $\{f_t,\ 0 \le t \le 1\}$ be the geometric path (4.2) between f_{θ_0} and f_{θ_1} and let p_t, $0 \le t \le 1$, be the probability measure on $\mathcal{X}$ with density (wrt λ) proportional to f_t. We assume that for any $t \in [0,1]$, we have a Markov kernel P_t with invariant distribution p_t. Let $\kappa > 0$ be a positive integer, $0 = t_0 < t_1 < \ldots < t_\kappa = 1$ a sequence and let $(X_0,\ldots,X_\kappa)$ be a nonhomogeneous Markov chain with transition kernels $\{P_{t_\ell}, 1 \le \ell \le \kappa\}$. This means that the conditional distribution $\mathbb{P}_x^{(\kappa)}$ of $(X_1, \ldots, X_\kappa)$ given $X_0 = x$ is

$$\mathbb{P}_x^{(\kappa)}\left(dx_1, \ldots, dx_\kappa\right) = \prod_{\ell=1}^{\kappa} P_{t_\ell}\left(x_{\ell-1}, dx_\ell\right).$$

Denote $\mathbb{E}_x^{(\kappa)}$ the expectation operator wrt $\mathbb{P}_x^{(\kappa)}$. Then we approximate r_{θ_0,θ_1} by

$$\hat{r}_{\theta_0,\theta_1} = \frac{1}{\kappa}\sum_{\ell=1}^{\kappa} U(X_\ell, t_\ell), \tag{4.5}$$

where $U(x,t) = \frac{d}{dt}\log f_t(x) = \log f_{\theta_1}(x) - \log f_{\theta_0}(x)$. Since $U(x,t)$ does not involve t, we will omit t and call it $U(x)$ hereafter.

This gives the following algorithm.

Algorithm 4.2.1. (i) Pick κ and choose $0 = t_0 < t_1 < \ldots < t_\kappa = 1$.

(ii) At time 0, choose $x_0 \in \mathcal{X}$. For $0 \leq \ell < \kappa$, given X_ℓ, generate $X_{\ell+1}$ from $P_{t_{\ell+1}}(X_\ell, \cdot)$.

(iii) Compute the estimator (4.5) as $\hat{r}_{\theta_0,\theta_1} = \frac{1}{\kappa}\sum_{\ell=1}^{\kappa} U(X_\ell)$.

For the implementation, one can choose P_t as a Metropolis-Hastings kernel with invariant distribution p_t. We stress again that the path (4.2) and the right Riemann sum (4.5) is just one way to implement the idea. We can choose other paths and numerical integration methods with no modification to the overall approach.

4.2.2 Theory

We now give some theoretical justification of the method. For a transition kernel P on a measurable space $\mathcal{X}$, we denote by P^n, $n \geq 0$, its n-th iterate. Any transition kernel P acts on bounded measurable functions h on $\mathcal{X}$ as $Ph(\cdot) \stackrel{\text{def}}{=} \int P(\cdot, dy)h(x)$ and on σ-finite measures λ as $\lambda P(\cdot) \stackrel{\text{def}}{=} \int \lambda(dx)P(x, \cdot)$. If $V : \mathcal{X} \to [1, +\infty)$ is a function, the V-norm of a function $h : \mathcal{X} \to \mathbb{R}$ is defined as $|h|_V \stackrel{\text{def}}{=} \sup_{x\in\mathcal{X}} |h(x)|/V(x)$.

When $V = 1$, this is the supremum norm. The set of functions with finite V-norm is denoted by $\mathcal{L}_V$. If λ is a signed measure the V-norm of λ is defined as $\|\lambda\|_V \overset{\text{def}}{=} \sup_{\{h, |h|_V \leq 1\}} |\lambda(h)|$.

We will make some basic ergodicity and smoothness assumption on $\{P_t\}$. We assume that there exists a measurable function $V : \mathcal{X} \to [1, \infty)$ such that for any $\beta \in (0, 1]$, there exist $\rho_\beta \in (0, 1)$, and a finite constant $C_{1\beta}$ such that

$$\sup_{0 \leq t \leq 1} \|P_t^n(x, \cdot) - p_t(\cdot)\|_{V^\beta} \leq C_{1\beta} \rho_\beta^n V^\beta(x), \tag{4.6}$$

and there exist $a \in (0, 1)$ and $b \in (0, \infty)$ such that

$$\sup_{0 \leq t \leq 1} P_t V(x) \leq a V(x) + b. \tag{4.7}$$

For any $\beta \in (0, 1]$ there exists a finite constant $C_{2\beta}$ such that for any $t, t' \in [0, 1]$

$$\sup_{|h|_{V^\beta} \leq 1} \sup_{x \in \mathcal{X}} \frac{|P_t h(x) - P_{t'} h(x)|}{V^\beta(x)} \leq C_{2\beta} |t - t'|. \tag{4.8}$$

Theorem IV.1. *Assume (4.6), (4.7) and (4.8). Suppose also that $U \in L_{V^\eta}$ for some $\eta \in [0, 1/2)$ and that for given κ, $\sup_{1 \leq i \leq \kappa} |t_i - t_{i-1}| \leq C\kappa^{-1}$ for some finite constant C. The dynamic path sampling has the following properties.*

a. There exists a finite constant $C(\theta_0, \theta_1)$ such that for any $\kappa > 1$,

$$\mathbb{E}_x^{(\kappa)} \left(\left| \frac{1}{\kappa} \sum_{j=1}^{\kappa} U(X_j) - r_{\theta_0, \theta_1} \right|^2 \right) \leq \frac{C(\theta_0, \theta_1)}{\kappa}. \tag{4.9}$$

b. As $\kappa \to \infty$,

$$\frac{1}{\sqrt{\kappa}} \sum_{j=1}^{\kappa} \left(U(X_j) - r_{\theta_0, \theta_1} \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

The asymptotic variance σ^2 is given by

$$\sigma^2 = \int_0^1 \pi_t \left(U^2\right) dt + 2 \sum_{j \geq 1} \int_0^1 \pi_t \left(U P_t^j U\right) dt. \tag{4.10}$$

Proof: We start the proof with some general properties of Markov kernels. Then we use them to prove Proposition 4.2.2, which is applied to prove this theorem in the end.

Let P_1, P_2 be two Markov kernels on a measurable general state space $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ with invariant distribution p_1, p_2 respectively. Assume that there exists a measurable function $V: \mathbb{T} \to [1, \infty)$, constants $\rho_i, a_i \in (0,1)$, $b_i, C_i \in (0, \infty)$ $(i = 1, 2)$ such that

$$\|P_i^n(x, \cdot) - p_i(\cdot)\|_V \leq C_i \rho_i^n V(x), \tag{4.11}$$

and

$$P_i V(x) \leq a_i V(x) + b_i. \tag{4.12}$$

Define

$$\delta(P_1, P_2) \overset{\text{def}}{=} \sup_{|h|_V \leq 1} \sup_{x \in \mathbb{T}} \frac{|P_1 h(x) - P_2 h(x)|}{V(x)}. \tag{4.13}$$

Proposition 4.2.1. *We have*

a.

$$\|p_1 - p_2\|_V \overset{\text{def}}{=} \sup_{|h|_V \leq 1} |p_1(h) - p_2(h)| \leq K_1 \delta(P_1, P_2), \tag{4.14}$$

where K_1 can be taken as $K_1 = C_2 \left(\inf_x V(x) + \frac{b_1}{1-a_1}\right) \left(\frac{1}{1-\rho_2}\right)$.

b. *Let $h_i \in L_V$ such that $p_i(h_i) = 0$, $i = 1, 2$ and define*

$$g_i(x) = \sum_{k \geq 0} P_i^k h_i(x).$$

Then $|g_i|_V \leq |h_i|_V C_i (1 - \rho_i)^{-1}$ and

$$|P_1 g_1 - P_2 g_2|_V \leq K_2 \left(|h_1 - h_2|_V + \delta(P_1, P_2)\right), \tag{4.15}$$

where K_2 can be taken as $K_2 = \left(1 + \frac{C_1}{1-\rho_1}\right) \vee \left(\frac{C_1 C_2 (1+K_1)|h_2|_V}{(1-\rho)^2}\right)$, where $\rho = \rho_1 \vee \rho_2$.

Proof. The proof is very similar to that of Propositions 2.2.2. First, we prove proposition a. Similarly we bound $\|p_1 - p_2\|_V$ as follows.

$$\begin{aligned} \|p_1 - p_2\|_V &\leq \|p_1 - P_1^n(x,)\|_V + \|P_1^n(x,) - P_2^n(x,)\|_V + \|p_2 - P_2^n(x,)\|_V \\ &\leq C_1\rho_1^n V(x) + \|P_1^n(x,) - P_2^n(x,)\|_V + C_2\rho_2^n V(x) \end{aligned}$$

Then $\|P_1^n(x,) - P_2^n(x,)\|_V$ is decomposed similarly.

$$\|P_1^n(x,) - P_2^n(x,)\|_V = \sup_{|h|_V \leq 1} \left| \sum_{j=0}^{n-1} P_1^j (P_1 - P_2)(P_2^{n-j-1}h - p_2 h) \right|$$

$$\begin{aligned} &\leq \sup_{|h|_V \leq 1} \left| \sum_{j=0}^{n-1} P_1^j (P_1 - P_2) C_2 \rho_2^{n-j-1} V(x) \right| \\ &\leq C_2 \delta(P_1, P_2) \sup_{|h|_V \leq 1} \left| \sum_{j=0}^{n-1} \rho_2^{n-j-1} \sup_j P_1^j V(x) \right|. \end{aligned}$$

It can be shown that $P_i^j V(x) \leq a_i^j V(x) + \frac{b_i}{1-a_i}$, so $\sup_j P_1^j V(x) \leq V(x) + \frac{b_i}{1-a_i}$. Then $\|P_1^n(x,) - P_2^n(x,)\|_V \leq C_2 \left(\inf_x V(x) + \frac{b_1}{1-a_1}\right)\left(\frac{1}{1-\rho_2}\right)\delta(P_1, P_2)$, which completes the proof.

Next we prove proposition b.

$$\begin{aligned} |g_i|_V = \sup_x \frac{\left|\sum_{k\geq 0} P_i^k h_i(x) - p_i h_i\right|}{V(x)} &\leq |h_i|_V \sup_x \sum_{k\geq 0} \sup_{|f|_V \leq 1} \frac{\left|P_i^k f(x) - p_i f\right|}{V(x)} \\ = |h_i|_V \sup_x \sum_{k\geq 0} C_i \rho_i^k = |h_i|_V \frac{C_i}{1-\rho_i}. \end{aligned} \tag{4.16}$$

To bound $|P_1 g_1 - P_2 g_2|_V$, similarly, we first bound $|g_1 - g_2|_V$.

$$|g_1 - g_2|_V = \left| \sum_{k\geq 0} (P_1^k - p_1)h_1 - (P_2^k - p_2)h_2 \right|_V$$

$$\leq \left| \sum_{k\geq 0} (P_1^k - p_1)(h_1 - h_2) \right|_V + \left| \sum_{k\geq 0} \left[(P_1^k - p_1) - (P_2^k - p_2) \right] h_2 \right|_V. \tag{4.17}$$

The first part of (4.17) can be bounded by $\frac{C_1}{1-\rho_1}|h_1 - h_2|_V$. The second part is further decomposed as

$$\begin{aligned}
&\left|\sum_{k\geq 1}\sum_{j=0}^{k-1}(P_1^j - p_1)\Big[(P_1-p_1)-(P_2-p_2)\Big](P_2^{k-j-1}-p_2)h_2\right|_V \\
\leq\; & \left|\sum_{j\geq 0}(P_1^j - p_1)\Big[(P_1-p_1)-(P_2-p_2)\Big]g_2\right|_V \\
\leq\; & \frac{C_1}{1-\rho_1}\Big[|(P_1-P_2)g_2|_V + |(p_1-p_2)g_2|_V\Big].
\end{aligned}$$

With proposition a and (4.16), (4.15) will follow easily. □

In the special setting of this thesis we use Proposition 4.2.1 to prove the following.

Proposition 4.2.2. *Assume (4.6-4.8). Let $\{h_t,\ 0 \leq t \leq 1\}$ be a family of functions $h_t : \mathcal{X} \to \mathbb{R}$ such that $h_t \in L_{V^\beta}$ for some $\beta \in (0,1)$ and* $\sup_{t\in[0,1]} |h_t|_{V^\beta} < \infty$*. Let $\{t_0, t_1, \ldots, t_\kappa\}$ as in Theorem IV.1. Denote $h_i = h_{t_i}$ and suppose that for any $1 \leq i \leq \kappa$, $|h_i - h_{i-1}|_{V^\beta} \leq C\kappa^{-1}$ where C does not depend on i nor κ. Then for any $0 < m \leq \kappa$ and for $\alpha \in (1, \beta^{-1}]$,*

$$\mathbb{E}_x^{(\kappa)}\left[\left|\sum_{i=1}^m h_i(X_i) - p_i(h_i)\right|^\alpha\right] \leq c\, m^{1\vee(\alpha/2)} V^{\alpha\beta}(x), \tag{4.18}$$

where the constant c does not depend on κ nor m.

Proof. Write $\bar{h}_i = h_i - p_i(h_i)$ and $g_i = \sum_{k\geq 0} P_i^k \bar{h}_i$. By Proposition 4.2.1, $\sup_{1\leq i\leq\kappa} |g_i|_{V^\beta} < \infty$. Combining (4.15), (4.14), the assumption $|h_i - h_{i-1}|_{V^\beta} \leq C\kappa^{-1}$ and (4.8) it follows that

$$\sup_{2\leq i\leq\kappa} |P_i g_i - P_{i-1}g_{i-1}|_{V^\beta} \leq c\kappa^{-1},$$

for a finite constant c that does not depend on κ. Since g_i satisfies $g_i(x) - P_i g_i(x) = \bar{h}_i(x)$, we have

$$\begin{aligned}
\sum_{i=1}^m \bar{h}_i(X_i) &= \sum_{i=1}^m g_i(X_i) - P_i g_i(X_{i-1}) + (P_1 g_1(X_0) - P_m g_m(X_m)) \\
&\quad + \sum_{i=2}^m P_i g_i(X_{i-1}) - P_{i-1}g_{i-1}(X_{i-1}) \quad = \sum_{i=1}^m g_i(X_i) - P_i g_i(X_{i-1}) + R_\kappa.
\end{aligned}$$

We easily see that $\mathbb{E}_x^{(\kappa)}\left(|R_\kappa|^\alpha\right) \le cV^{\alpha\beta}(x)$ for some finite constant c that does not depend on κ. The term $\{(\sum_{i=1}^m g_i(X_i) - P_i g_i(X_{i-1}), \mathcal{F}_m)\,,\ 1 \le m \le \kappa\}$ is a martingale array and by Burkeholder inequality,

$$\mathbb{E}_x^{(\kappa)}\left(\left|\sum_{i=1}^m g_i(X_i) - P_i g_i(X_{i-1})\right|^\alpha\right) \le cm^{1\vee\alpha/2}V^{\alpha\beta}(x).$$

□

Now we are ready to prove Theorem IV.1. Note that $\frac{\partial}{\partial t}p_t\,(U) = p_t\,(U^2) - (p_t(U))^2 \le p_t\,(U^2)$, the supremum for $t \in [0,1]$ of which is finite as a consequence of (4.6) and (4.7) and the fact that $U^2 \in \text{Ł}_V$. We can write

$$\frac{1}{\kappa}\sum_{i=1}^\kappa U(X_i) - r_{\theta_0,\theta_1} = \frac{1}{\kappa}\sum_{i=1}^\kappa (U(X_i) - p_i(U)) + \frac{1}{\kappa}\sum_{i=1}^\kappa p_i(U) - r_{\theta_0,\theta_1}.$$

We apply Proposition 4.2.2 with $h_i(X) = U(X)$, $m = \kappa$, and $\alpha = 2$:

$$\mathbb{E}_x^{(\kappa)}\left[\left|\frac{1}{\kappa}\sum_{i=1}^\kappa U(X_i) - p_i(U)\right|^2\right] \le c\,V(x)^{2\eta}\kappa^{-1}.$$

By the properties of Riemann sums,

$$\left|\frac{1}{\kappa}\sum_{i=1}^\kappa p_i(U) - r_{\theta_0,\theta_1}\right| \le \frac{1}{\kappa}\sup_{0\le t\le 1}\left|\frac{\partial}{\partial t}p_t\,(h_t)\right| \le |U^2|_V \sup_{0\le t\le 1} p_t(V)\kappa^{-1}.$$

And we deduce that:

$$\mathbb{E}_x^{(\kappa)}\left[\left|\frac{1}{\kappa}\sum_{i=1}^\kappa U(X_i) - r_{\theta_0,\theta_1}\right|^2\right] \le \left(\frac{CV^\eta(x)}{\sqrt{\kappa}} + \frac{|U|_V^2 \sup_{0\le t\le 1} p_t(V)}{\kappa}\right)^2.$$

Notice that $U(X) = \log f_{\theta_1}(X) - \log f_{\theta_0}(X)$, so the constant may depend on θ_0 and θ_1.

Then by Proposition 4.2.2, and we have the martingale approximation

$$\sum_{i=1}^\kappa (U(X_i) - r_{\theta_0,\theta_1}) = \sum_{i=1}^\kappa g_i(X_i) - P_i g_i(X_{i-1}) + \bar{R}_\kappa,$$

where $\bar{R}_\kappa = R_\kappa + \sum_{i=1}^{\kappa} p_i(U) - r_{\theta_0,\theta_1}$ and $\mathbb{E}_x^{(\kappa)}\left(|\bar{R}_\kappa|^2\right) \le c$. Thus $\kappa^{-1/2}\bar{R}_\kappa$ converges in probability to zero. The term $\sum_{i=1}^{\kappa} g_i(X_i) - P_i g_i(X_{i-1})$ is a triangular martingale array (notice that i depends on κ). Given Proposition 4.2.1, we now apply Proposition 4.2.2 again to the function $h_i(X) = P_i g_i^2(X) - (P_i g_i(X))^2$ to obtain the weak law of large numbers. We conclude that

$$\frac{1}{\kappa}\sum_{i=1}^{\kappa} P_i g_i^2(X_{i-1}) - (P_i g_i(X_{i-1}))^2 - \sigma^2 \to 0 \quad \text{in probability} .$$

Then by (Hall and Heyde 1980) Corollary 3.1 we deduce that

$$\frac{1}{\sqrt{\kappa}}\sum_{i=1}^{\kappa} \left(U(X_i) - r_{\theta_0,\theta_1}\right) \xrightarrow{d} \mathcal{N}\left(0,\sigma^2\right) .$$

The asymptotic variance can be written

$$\begin{aligned}\sigma^2 = \int_0^1 \mathbb{E}_t\left(g_t^2(X_t) - (P_t g_t(X_t))^2\right) dt &= \int_0^1 \mathbb{E}_t\left(2U(X_t)g_t(X_t) - U^2(X_t)\right) dt \\ &= \int_0^1 \mathbb{E}_t\left\{U^2(X_t) + 2\sum_{j\ge1} U(X_t)P_t^j U(X_t)\right\} dt, \\ &= \int_0^1 \mathbb{E}_t\left(U^2(X_t)\right) dt + 2\sum_{j\ge1}\int_0^1 \mathbb{E}_t\left(U(X_t)P_t^j U(X_t)\right) dt.\end{aligned}$$

□

Ratios of normalizing constants are typically used in Statistics for models comparison and selection. It is thus important to assess the precision of methods computing these ratios. We will now investigate the problem of estimating the asymptotic variance σ^2 given in the above theorem. Firstly, the next proposition show that we can estimate the term $\int_0^1 \pi_t\left(UP_t^\ell U\right) dt$ by taking the usual empirical auto-covariance. For $l \ge 0$, define $\gamma_\kappa(\ell) = 0$ if $\ell \ge \kappa$ and

$$\gamma_\kappa(\ell) = \frac{1}{\kappa}\sum_{j=1}^{\kappa-\ell}\left(U(X_j) - \frac{1}{\kappa}\sum_{i=1}^{\kappa} U(X_i)\right)\left(U(X_{j+\ell}) - \frac{1}{\kappa}\sum_{i=1}^{\kappa} U(X_i)\right), \qquad (4.19)$$

otherwise.

Lemma 4.2.1. *Under the assumption of Theorem IV.1, for any $\ell \geq 0$,*

$$\lim_{\kappa\to\infty} \mathbb{E}_x^{(\kappa)} \left[\left| \gamma_\kappa(\ell) - \int_0^1 \mathbb{E}_t \left(U(X_t) P_t^\ell U(X_t) \right) dt \right| \right] = 0. \tag{4.20}$$

Although each covariance term $\int_0^1 \pi_t \left(U P_t^\ell U \right) dt$ can be consistently estimated using $\gamma_\kappa(\ell)$, it is well known that a simple summation $\gamma_\kappa(0) + 2\sum_{\ell=1}^{\kappa} \gamma_\kappa(\ell)$ is not a consistent estimate of σ^2. It is necessary to down weight the $\gamma_\kappa(\ell)$ for large ℓ which are anyway typically small and badly estimated. A consistent estimator can be obtained by downweighting covariance terms for large ℓ leading to estimators of the form

$$\Gamma_\kappa = \gamma_\kappa(0) + 2\sum_{\ell=1}^{b_\kappa^{-1}} w(\ell b_\kappa)\gamma_\kappa(\ell). \tag{4.21}$$

where $\{b_n,\ n \geq 1\}$ is a sequence of positive numbers such that $b_n \to 0$ but $nb_n \to \infty$ as $n \to \infty$. Typical choice of b_n is $b_n \propto n^{-1/2}$. The function $w:\ \mathbb{R} \to \mathbb{R}$ is a weight function such that $w(0) = 1$. A typical example of w is Parzen kernel function: $w(z) = 1 - 6z^2 + 6z^3$ if $z \in [0, .5]$, $w(z) = 2(1-z)^3$ if $z \in [.5, 1]$, and $w(z) = 0$ elsewhere.

4.3 A Toy Example

Here we give a toy example where the goal is to estimate the ratio of the normalizing constants of multivariate normal distributions. Since we know the true value, we are able to compare Algorithm 4.2.1 with a conventional implementation of path sampling (4.3) (we call path sampling for short in this subsection) in terms of accuracy and precision. Moreover, we will compare both methods with the importance sampling method based on the identity $\frac{Z(\theta_1)}{Z(\theta_0)} = \mathbb{E}_{\theta_0}\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}\right)$. In this importance sampling approach one runs a Markov chain with invariant distribution $p_{\theta_0}(\cdot)$ until convergence and use the log of the sample mean of $\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}$ to estimate $e^{r_{\theta_0,\theta_1}}$.

Consider a multivariate Gaussian distribution $N(\mu, \theta^2 I_{3\times3})$, where $I_{3\times3}$ is a 3 by 3 identity matrix. We estimate the ratio of the normalizing constants for two scenarios: $\mu_1 = \mu_2 = (0,0,0), \theta_1 = .1, \theta_2 = .2$, and $\mu_1 = (0,0,0), \mu_2 = (5,5,5), \theta_1 = 1, \theta_2 = 5$. The true values of the ratio are 8 and 125, respectively. We repeat the estimators 1000 times, and calculate mean and mean square error (MSE) based on the 1000 estimators for each method.

For all three algorithms, the transition kernel is a Random Metropolis kernel with proposal $T(x, \cdot) = N(x, \sigma^2)$. The scale parameter σ^2 is chosen so that the acceptance rate of the Markov chain is approximately 30%. All algorithms start with the same initial value (0,0,0). For dynamic path sampling, we try various value of $\kappa = 500, 10^3, 5 \times 10^3, 10^4, 5 \times 10^4, 10^5, 10^6$. For path sampling (4.4), we take K points along the path, and for each point t_k, we estimate $\mathbb{E}_{t_k}(U(X_k, t_k))$ by running a MCMC sampler with invariant distribution p_k for n iterations. We try $K = 5, 10, 20, 25, 50, 100, 200, 500, 1000$, $n = 100, 200, 500, 1000, 2000$, and compared the best result with dynamic path sampling of similar computing time, i.e. $Kn = \kappa$. As for importance sampling, to match the computing time, we run a chain of κ iterations. When estimating the mean, we discard the first 20% of the data.

The results are given in table 4.1. In each row, the total number of iterations is the same for all three methods, and hence the computing time is similar. In the first case where the two distributions have more overlap support, dynamic path sampling has a much smaller MSE and higher accuracy. It picks up the ratio with very small number of iterations, and its accuracy increases much faster. As we expect, path sampling catches up as the number of iterations becomes very large. Whereas importance sampling also has high accuracy, but its corresponding MSE is very high. Its MSE decreases as iteration number grows, however is still very big (more than 200 times

Table 4.1: The sample mean and MSE of estimators for dynamic path sampling (DPS), path sampling (PS), and importance sampling (IS). e^r is the true value.

κ	K	n	e^r	Mean			MSE		
				DPS	PS	IS	DPS	PS	IS
500	5	100	8	7.42	9.83	7.29	5.94	26.8	116
10^3	10	100	8	7.63	7.54	7.34	2.53	5.10	68.0
5×10^3	25	200	8	7.94	8.30	7.52	.496	1.16	33.5
10^4	50	200	8	7.98	7.96	7.63	.271	.447	23.3
5×10^4	100	500	8	7.99	8.15	7.70	5.06×10^{-2}	.106	11.5
10^5	100	10^3	8	7.99	8.14	7.80	2.57×10^{-2}	5.07×10^{-2}	8.90
10^6	10^3	10^3	8	8.00	8.01	7.84	2.51×10^{-3}	3.97×10^{-3}	.941
5×10^3	50	100	125	124.0	309.8	10.43	5.98×10^3	67.5×10^3	1.58×10^4
10^4	100	100	125	125.8	189.2	11.15	3.63×10^3	8.76×10^3	1.42×10^4
10^5	100	10^3	125	124.5	184.7	11.64	287.6	3.95×10^3	1.37×10^4
10^6	10^3	10^3	125	125.0	129.7	13.84	30.5	39.5	1.27×10^4

larger) comparing to the other two methods.

In the second case where the distributions are far apart from each other, dynamic path sampling still performs better than path sampling with small number of iterations. They become comparable when iteration number is large. As expected, importance sampling fails in this situation. The estimate is quite biased and its MSE is huge even for large n. This is because the sample from $p_{\theta_0}(\cdot)$ is concentrated around the $(0, 0, 0)$, while $p_{\theta_1}(\cdot)$ has most mass around $(5, 5, 5)$.

This example shows that dynamic path sampling provides a faster and more accurate approach to implement the path sampling identity particularly when the two distributions have reasonable overlap or the computing time is short. On the other hand, the importance sampling method performs very poorly and will not work well in practice unless the two distributions are very close.

APPENDICES

0.1 Estimation Procedures:

Here we use the specification of bundle raffle as an illustration. We assume that the students are sorted so that the first m students logged on to the website.

0.1.1 Hierarchical Bayes Model

The Metropolis-Hasting algorithm (Chib and Greenberg 1995) was used to generate the empirical draws for estimation. Draws for each parameter were obtained as follows:

1. Generate b

   $$f(b|\{\beta_i\}, V_\beta) = N\left([I + \frac{V_\beta}{100n}]^{-1}\sum_{i=1}^{n}\frac{\beta_i}{n}, \frac{V_\beta}{n}[I + \frac{V_\beta}{100n}]^{-1}\right)$$

2. Generate V_β

   $V_\beta|\{\beta_i, b\} \sim$ Inverted Wishart $\left[P + n + 2, I + \sum_{i=1}^{n}(\beta_i - b)(\beta_i - b)^T\right]$

3. Generate $\{\beta_i, i = 1, \cdots, n\}$

   $$f(\beta_i|y_i^I, x_i^I, b, V_\beta) \propto \exp\{-\frac{1}{2}(\beta_i - b)^T V_\beta^{-1}(\beta_i - b)\}p(y_i^I|x_i^I, \beta_i),$$

   where $p(y_i^I|x_i^I, \beta_i) = \prod_{q=1}^{12}\frac{\exp\{\beta_i^T x_{iq,y_{iq}^I}^I\}}{\sum_{k=1}^{4}\exp\{\beta_i^T x_{iq,k}^I\}}$.

0.1.2 Discrete-time Markov Chain

1. Generate b and V_β the same way as in hierarchical Bayes model.

2. Generate $\{\beta_i, i = 1, \cdots, n\}$

   (a) for $i \leq m$,

   $$f(\beta_i | y_i^I, x_i^I, s^0, \cdots, s^T, b, V_\beta, \gamma, g, W) \propto \exp\{-\frac{1}{2}(\beta_i - b)^T V_\beta^{-1}(\beta_i - b)\} \times$$
   $$P(y_i^I | x_i^I, \beta_i) \prod_{t:c^{t+1}=i} p(y_i^{t+1} | c^{t+1} = i, c^t, y^t, x, \beta_i, \gamma, g, W),$$

   where

   $$p(y_i^{t+1} | c^{t+1} = i, c^t, y^t, x, \beta_i, \gamma, g, W) = \frac{\exp\{u_{iy_i^{t+1}}^{t+1}\}}{\sum_{k=I(y_i^t=0)}^4 \exp\{u_{ik}^{t+1}\}},$$

   with u_{ik}^{t+1} defined in (3.7) and (3.8).

   (b) for $i > m$, β_i is generated as in hierarchical Bayes model.

3. Generate $\{\gamma_j, j = 0, \cdots, 3\}$

   $$f(\gamma_j | s^1, \cdots, s^T, x, \beta, \gamma_{-j}, g, W) \propto \exp\{-\frac{\gamma_j^T \gamma_j}{200}\} \prod_{t=0}^{T-1} p(s^{t+1} | s^t, x, \beta, \gamma, g, W),$$

   where

   $$p(s^{t+1} | s^t, x, \beta, \gamma, g, W) = \frac{1}{m} \mathbf{1}(y_{-c^{t+1}}^{t+1} = y_{-c^{t+1}}^t) \frac{\exp\{u_{c^{t+1} y_{c^{t+1}}^{t+1}}^{t+1}\}}{\sum_{k=\mathbf{1}(y_{c^{t+1}}^t=0)}^4 \exp\{u_{c^{t+1}k}^{t+1}\}},$$

   with $u_{c^{t+1}k}^{t+1}$ defined in (3.7) and (3.8).

0.1.3 Pseudolikelihood

(a) Generate b and V_β the same way as in hierarchical Bayes model.

(b) Generate $\{\beta_i, i = 1, \cdots, n\}$

i. for $i \leq m$,

$$f(\beta_i|y_i^I, y, x_i^I, x, b, V_\beta, \gamma, g, W) \propto$$
$$\exp\{-\frac{1}{2}(\beta_i - b)^T V_\beta^{-1}(\beta_i - b)\}p(y_i^I|x_i^I, \beta_i)p(y_i|y_{-i}, x, \beta_i, \gamma, g, W),$$

where $p(y_i|y_{-i}, x, \beta_i, \gamma, g, W)$ is the choice probability associated with individual i that depends only on his or her neighbors' choices.

ii. for $i > m$, β_i is generated as in hierarchical Bayes model.

(c) Generate $\{\gamma_j, j = 0, ..., 3\}$

$$f(\gamma_j|y, x, \beta, \gamma_{-j}, g, W) \propto \exp\{-\frac{\gamma_j^T\gamma_j}{200}\}\prod_{i=1}^{m} p(y_i|y_{-i}, x, \beta_i, \gamma, g, W)$$

0.1.4 Approximate Sampling

(a) Generate b and V_β the same way as in hierarchical Bayes model.

(b) Generate $\{\beta_i, i = 1, ..., n\}$

i. for $i \leq m$,

$$f(\beta_i|y_i^I, y, x_i^I, x, b, V_\beta, \gamma, g, W) \propto$$
$$\exp\{-\frac{1}{2}(\beta_i - b)^T V_\beta^{-1}(\beta_i - b)\}p(y_i^I|x_i^I, \beta_i)p(y|x, \beta, \gamma, g, W),$$

where $p(y|x, \beta, \gamma, g, W)$ is defined in (3.9) and (3.10). We need to apply approximate sampling to draw β_i. Please refer to Algorithm 2.2.1 for details.

ii. for $i > m$, β_i is generated as in hierarchical Bayes model.

(c) Generate $\{\gamma_j, j = 0, ..., 3\}$

$$f(\gamma_j|y, x, \beta, \gamma_{-j}, g, W) \propto \exp\{-\frac{\gamma_j^T\gamma_j}{200}\}p(y|x, \beta, \gamma, g, W),$$

approximate sampling is used to draw γ_j.

BIBLIOGRAPHY

Algesheimer, R., Borle, S., Dholakia, U.M., Singh, S.S., 2010. The impact of customer community participation on customer behaviors: an empirical investigation. Marketing Science. Forthcoming.

Allenby, G. M., Rossi, P.E., 1999. Marketing models of consumer heterogeneity. Journal of Econometrics 89 (1-2), 57-78.

Aribarg, A., Arora, N., Bodur, H.O., 2002. Understanding the role of preference revision and concession in group decisions. Journal of Marketing Research 39(3), 336-349.

Aribarg, A., Arora, N., Kang, M.Y., 2010. Predicting joint choice using individual data. Marketing Science 29(1), 139-157.

Atchadé, F.Y., Lartillot, N., Robert, C., 2008. Bayesian computation for statistical models with intractable normalizing constants. Technical Report, University of Michigan.

Aykroyd, R.G., 1998. Bayesian estimation for homogeneous and inhomogeneous Gaussian random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 533-539.

Aykroyd, R.G., Zimeras, S., 1999, Inhomogeneous prior models for image reconstruction. Journal of the American Statistical Association 94, 934-946.

Banerjee, S., Carlin, B., Gelfand, A., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall Ltd.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society, Series B 36 (2), 192-236.

Besag, J., 1975. Statistical analysis of non-lattice data. Journal of Royal Statistal Society, Series D 24(3), 179-195.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43 (1), 1-59.

Bresnahan, T.F., Reiss, P.C., 1991. Entry and competition in concentrated markets. Journal of Political Economy, 99 (5), 997-1009.

Caimo, A., Friel, N., 2011. Bayesian inference for exponential random graph models. Social Networks 33 41-55.

Chevalier, J.A., Mayzline, D., 2006. The effect of word of mouth on sales: online book reviews. Journal of Marketing Research 43 (3), 345-354.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. American Statistical Association 49 (4), 327-335.

Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis-Hastings output. Journal of American Statistical Association 96 (453), 270-281.

Chow, Y.S., 1960. A martingale inequality and the law of large numbers. Proceedings of the American Mathematical Society 11, 107-111.

Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York.

Denison, D.G.T., Holmes, C.C., 2001. Bayesian partitioning for estimating disease risk. Biometrics 57, 143-149.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics (with discussion). Journal of the Royal Statistical Society, Series C 47, 299-350.

Fernández, C., Green, P.J., 2002. Modelling spatially correlated data via mixtures: a Bayesian approach. Journal of the Royal Statistical Society, Series B 64, 805-826.

Geman, S., Graffigne, C., 1987. Markov random field image models and their applications to computer vision. In Proceedings of the International Congress of Mathematicians, 1496-1517, Berkeley, California, 1986.

Gelman, A., Meng, X.L., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Statistical Science 13 (2), 163-185.

Geyer, C.J., Thompson, E.A., 1992. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). Journal of the Royal Statistical Society, Series B 54 (3), 657-699.

Godes, D., Mayzlin, D., 2004. Using online conversations to study word-of-mouth communication. Marketing Science 23 (4), 545-560.

Goodreau, S.M., Handcock, M.S., Hunter, D.R., Butts, C.T., Morris M., 2008. A statnet tutorial. Journal of Statistical Software 24 (9).

Green, P.J., Richardson, S., 2002. Hidden Markov models and disease mapping. Journal of the American Statistical Association 97 (460), 1055-1070.

Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and Its Application. Academic Press, New York, NY.

Hammersley, J.M., Clifford, P., 1971. Markov field on finite graphs and lattices. Unpublished.

Handcock, M. S., 2003. Assessing degeneracy in statistical models of social networks. Working Paper no.39, Center for Statistics and the Social Sciences, University of Washington. Available from http://www.csss.washington.edu/Papers/

Handcock, M.S., Stein, M.L., 1993. A Bayesian analysis of kriging. Technometrics 35, 403-410.

Haran, M., 2010. Gaussian random field models for spatial data, to appear in Handbook of Markov chain Monte Carlo, Editors, Brooks, S.P., Gelman, A.E., Jones, G.L., Meng, X.L., Springer-Verlag.

Hartmann, W.R., 2010. Demand estimation with social interactions and the implications for targeted marketing. Marketing Science 29 (4), 585-601.

Hunter, D.R., Handcock, M.S., 2006. Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics 15 (3), 565-583.

Ibanez, M.V., Simo A., 2003. Parameter estimation in Markov random field image modeling with imperfect observations: a comparative study. Pattern Recognition Letters 24, 237-389.

Kleinman, A., Rodrigue, N., Bonnard, C., Philippe, H., 2006. A maximum likelihood framework for protein design. BMC Bioinformatics, 7.

Knorr-Held, L., Raber, G., 2000. Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56 (1), 13-21.

Koskinen, J.H., Snijders, T.A.B., 2007. Bayesian inference for dynamic social network data. Journal of Statistical Planning and Inference 137 (12), 3930-3938.

Leskovec, J., Adamic, L.A., Bernardo, H.A., 2007. The dynamics of viral marketing. ACM Transactions on the Web, 1(1).

Meng, X.L., Wong, W.H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistical Sinica 6, 831-860.

Meyn, S.P., Tweedie, R.L., 1993. Markov Chains and Stochastic Stability. Springer-Verlag London Ltd., London.

Minka, T.P., 2001. Expectation propagation for approximate Bayesian inference. Uncertainty in AI. http://www.stat.cmu.edu/minka/papers/ep/.

Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K., 2006. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika 93 (2), 451-458.

Møller, J., Waagepetersen, R.P., 2003. Statistical Inference and Simulation for Spatial Point Processes (Monographs on Statistics and Applied Probability), Chapman & Hall/CRC, Boca Raton, FL.

Moon, S., Russell, G.J., 2008. Predicting product purchase from inferred customer similarity: an autologistic model approach. Marketing Science 54 (1), 71-82.

Morris, M., Handcock, M.S., Hunter, D.R., 2008. Specification of exponential-family random graph models: terms and computational aspects. Journal of Statistical Software 24 (4).

Murray, I., Ghahramani, Z., MacKay, D.J.C., 2006. MCMC for doubly-intractable distributions. In: 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006), July 13-16, 2006, Cambridge, MA, US.

Nair, H., Manchanda, P., Bhatia, T., 2010. Asymmetric social interactions in physician prescription behavior: the role of opinion leaders. Journal of Marketing Ressearch 47(5) 883-895.

Propp, J.G., Wilson, D.B., 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures and Algorithms 9 (1-2), 223-252.

Qi, Y., Szummer, M., Minka, T., 2005. Bayesian conditional random fields. Artificial Intelligence and Statistics.

Ross, S.M., 1996. Stochastic Processes (2nd ed.). Wiley, New York.

Russell, G.J., Petersen, A., 2000. Analysis of cross category dependence in market basket selection. Journal of Retailing 76 (Fall), 367-392.

Smith, T.E., LeSage, J.P., 2004. A Bayesian probit model with spatial dependencies. Advances in Econometrics 18, 127-162.

Snijders, T.A.B., 2002. Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure 3 (2).

Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S., 2006. New specifications for exponential random graph models. Sociological Methodology, 99-153.

Snijders, Tom A.B., Christian Steglich and Michael Schweinberger. 2007. Modeling the Co- evolution of Networks and Behavior. Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, New Jersey.

Stephen, A.T., Toubia, O., 2010. Deriving value from social commerce networks. Journal of Marketing Research 47(2), 215-228.

Strauss, D., Ikeda, M., 1990. Pseudolikelihood estimation for social networks. Journal of the American Statistical Association 85 (409), 204-212.

Sun, D., Tsutakawa, R.K., Speckman, P.L., 1999. Posterior distribution of hierarchical models using CAR(1) distributions. Biometrika 86, 341-350.

Toubia, O., Stephen, A.T., Freud, A., 2009. Viral marketing: a large-scale field experiment. INSEAD working paper, available from http://www.insead.edu/facultyresearch/research/doc.cfm?did=42584

Trusov, M., Bodapati, A.V., Bucklin, R.E., 2010. Determining Influential Users in Internet Social Networks. Journal of Marketing Research 47 (4), 643-658.

Wall, M.M., 2004. A close look at the spatial structure implied by the CAR and SAR models. Journal of Statistical Planning and Inference 121, 311-324.

Wasserman, S., Pattison, P.E., 1996. Logit models and logistic regression for social networks: I. an introduction to Markov graphs and p^*. Psychometrika 61 (3), 401-425.

Welling, M., Parise, S., 2006. Bayesian random fields: The Bethe-Laplace approximation. In UAI.

Yang, S., Allenby, G.M., 2003. Modeling interdependent consumer preferences. Journal of Marketing Research 40 (3), 282-294.

Yang, S., Narayan, V., Assael, H., 2006. Estimating the interdependence of television program viewership between spouses: a Bayesian simultaneous equation model. Marketing Science 25 (4), 336-349.

Yang, S., Zhao, Y., Erdem, T., Zhao, Y., 2009. Modeling the intra-household behavioral interaction. Journal of Marketing Research. Forthcoming.

Younes, L., 1988. Estimation and annealing for gibbsian fields. Annales de l'Institut Henri Poincaré. Probabilité et Statistiques 24, 269-294.

Zhu, F., Zhang, X.Q., 2010. Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics. Journal of Marketing 74 (2), 133-148.