Relational Learning with One Network: An Asymptotic Analysis

Rongjing Xiang
Department of Computer Science
Purdue University

Jennifer Neville
Department of Computer Science and Statistics
Purdue University

Abstract

Theoretical analysis of structured learning methods has focused primarily on domains where the data consist of *independent* (albeit structured) examples. Although the statistical relational learning (SRL) community has recently developed many classification methods for graph and network domains, much of this work has focused on modeling domains where there is a *single* network for learning. For example, we could learn a model to predict the political views of users in an online social network, based on the friendship relationships among users. In this example, the data would be drawn from a single large network (e.g., Facebook) and increasing the data size would correspond to acquiring a larger graph. Although SRL methods can successfully improve classification in these types of domains, there has been little theoretical analysis addressing the issue of single network domains. In particular, the asymptotic properties of estimation are not clear if the size of the model grows with the size of the network. In this work, we focus on outlining the conditions under which learning from a single network will be asymptotically consistent and normal. Moreover, we compare the properties of maximum likelihood estimation (MLE) with that of generalized maximum pseudolikelihood estimation (MPLE) and use the resulting understanding to propose novel MPLE estimators for single network domains. We include empirical analysis on both synthetic and real network data to illustrate the findings.

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

1 Introduction

Statistical relational learning methods aim to exploit the relationships among instances, which naturally exist in network data, to improve both descriptive and predictive modeling performance compared to conventional learning methods that assume independent and identically distributed (i.i.d.) instances. Relational modeling techniques have been applied in many application domains, such as social networks, the worldwide web, and citation analysis, with great empirical success. However, there has been relatively little theoretical analysis of the properties of relational models, nor has there been much focus on precisely defining the estimation and learning tasks.

From the breadth of past work on real-world applications, it is clear that there are two distinct learning scenarios that are implicitly considered by relational learning researchers and analysts. In the first scenario, the domain consists of a population of independent graph samples (e.g., chemical compounds). Here each graph is structured, but the domain consists of a set of independent graphs, so we can reason theoretically about the characteristics of algorithms in the limit, as the number of available graph samples increases. In the second scenario, the domain consists of a *single*, large graph (e.g., Facebook). In this case, an increase in dataset size corresponds to acquiring a larger sample from the network. Since relational models typically focus on modeling the joint distribution of attributes (e.g., class labels) over a specific network structure, the asymptotic properties of the methods (as the size of the network grows) are not clear. Although many relational applications focus on learning and prediction in a single network, theoretical analysis of the models has focused on the former scenario (i.e., with multiple, independent network samples). In this work, we attempt to close this gap, by outlining the conditions under which learning from a single network will be asymptotically consistent and normal.

The purpose of our asymptotic analysis is two fold. First, the asymptotic consistency argument provides theoretical justification for current relational learning

approaches that practitioners have been applying to network data comprised of a single graph. Moreover, our analysis provides some theoretical insight into a central question in relational learning as to whether joint learning over the full data graph outperforms disjoint learning over subgraphs from the original network, and if so, to what extent, and in which situations? Our asymptotic normality and efficiency results illustrate how disjoint and joint learning approaches exploit the relational dependencies in the data differently, and this understanding points to a spectrum of learning approaches in between, by varying the level of joint learning.

More specifically, we investigate this tradeoff through a comparative analysis of maximum likelihood estimators (MLE) and generalized maximum pseudolikelihood estimators (MPLE). To the best of our knowledge, this is the first formal analysis of MLE and MPLE in the machine learning community for the context of single-network estimation. Although the asymptotic behavior of MLE and MPLE has been well studied for i.i.d. data, the analysis for network data requires different techniques and careful specification of modeling assumptions. Since probabilistic limit theorems do not hold in full generality for dependent data, asymptotic consistency and normality of relational learners can only be established under certain assumptions about the relational structure. Previous work in the statistical physics community has analyzed the asymptotic behavior of estimators for Gibbs measures on lattice data under various structural assumptions (see e.g. Comets, 1992). However, the typical assumptions made for lattice data, such as shift invariance, are not applicable to relational learning on heterogenous networks. Alternatively, we base our analysis on two assumptions that are more suitable for single-network domains. First, we assume that node degrees remain bounded as the size of the data graph grows. Second, we formalize the intuition that correlation decays rapidly in the model as graph distance increases, with a notion of *weak dependence* which requires that the total correlation of each clique with all other cliques across the network is finite.

The rest of this paper is organized as follows. In Section 2, we describe a templated Markov network model for relational data, state our modeling assumptions, and prove basic convergence. In Section 3, we prove the asymptotic consistency and normality for MLE and MPLE in the single-network scenario. We also compare the asymptotic efficiency of MLE with different MPLE's. In Section 4, we illustrate the findings with an empirical study on both synthetic data and a real world social network. Finally, we conclude the paper with a review of related work and a discussion.

2 Model Formulation

In this paper, our analysis is based on the framework of Markov networks. We chose a Markov network framework for the following reasons: (1) their formulations are used widely in relational learning, (2) their rich representation ability (e.g., Markov networks can naturally translate data relationships into a model of possibly cyclic probabilistic dependencies), (3) their ability to derive a consistent global distribution from local specifications, which is essential for consistent learning from subgraphs drawn from an unknown, larger population graph, and (4) their amenability to theoretical analysis due to the desirable analytical properties of exponential families. While we will use the general terminology of Markov networks throughout this paper, our formulation encompasses a rich class of undirected graphical model-based relational learners, including relational Markov networks (Taskar et al., 2002), Markov logic networks (Richardson and Domingos, 2006), relational dependency networks (Neville and Jensen, 2007), conditional random fields for relational learning (Sutton and Mccallum, 2006), and P* models (Robins et al., 2007).

2.1 Templated Markov Networks

Templated Markov Networks for relational data can be generally written as: $P_G(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\prod_{T\in\mathcal{T}}\prod_{C\in\mathcal{C}_T(G)}\Phi_T(\mathbf{x}^C,\mathbf{y}^C;\boldsymbol{\theta}_T)$, where $\mathcal{T}$ is the set of clique templates, and Z is the partition function. Each clique C is defined on a small subgraph of the whole data graph, while the parameters of cliques within the same template T are tied, denoted by $\boldsymbol{\theta}_T$. Therefore, we use a single potential function Φ_T for each template T. Each potential is further formulated as a log-linear function of a set of features $\boldsymbol{\phi}_T$. $\boldsymbol{\phi}_T$ are computed from the vector of attributes($\mathbf{x}$) and labels($\mathbf{y}$) within the corresponding clique C: $\Phi_T = \exp\left\{\left\langle\boldsymbol{\theta}_T, \boldsymbol{\phi}_T(\mathbf{x}^C,\mathbf{y}^C)\right\rangle\right\}$.

Throughout this paper, to keep the notation simple, we assume that only one template is defined, although the analysis can easily be generalized for a finite number of templates. As such, we drop the subscript T in the above formulation. Furthermore, let n denote the number of cliques in graph G, i.e., $n = |\mathcal{C}(G)|$. Then the model can be compactly written as follows.

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_n(\boldsymbol{\theta};\mathbf{x})}\exp\left\{\left\langle\boldsymbol{\theta}, \sum_{C\in\mathcal{C}(G)}\boldsymbol{\phi}(\mathbf{x}^C,\mathbf{y}^C)\right\rangle\right\} \quad (1)$$

where the partition function is $Z_n(\boldsymbol{\theta};\mathbf{x}) = \int_{\mathcal{Y}^G}\exp\left\langle\boldsymbol{\theta}, \sum_{c\in\mathcal{C}(G)}\boldsymbol{\phi}(\mathbf{x}^C,\mathbf{y}^C)\right\rangle d\mathbf{y}$.

Within this model, we consider learning over training graphs with increasing number of cliques $n\to\infty$. Depending on the model specification, as the size of the graph grows, n can grow in the order of the number of nodes (e.g., node-centric clique specifications),

$E_{\boldsymbol{\theta}}$: expectation taken w.r.t. $P_{\boldsymbol{\theta}}$
G: training graph
$C \in \mathcal{C}(G)$: template potential cliques
n: number of cliques($\lvert\mathcal{C}(G)\rvert$)
$S \in \mathcal{S}(G)$: template pseudolikelihood components
m: number of pseudolikelihood components($\lvert\mathcal{S}(G)\rvert$)
$L_n(\boldsymbol{\theta};\mathbf{x})$: the conditional log-likelihood
$\tilde{L}_m(\boldsymbol{\theta};\mathbf{x})$: the conditional log-pseudolikelihood
$\mathbf{f}_n$ ($\tilde{\mathbf{f}}_m$): the gradient of the $L_n(\tilde{L}_m)$
$\mathbf{y}^R$ ($\mathbf{x}^R$): joint instantiation of $\mathbf{y}(\mathbf{x})$ over the node set R
$\boldsymbol{\phi}^C = \boldsymbol{\phi}(\mathbf{x}^C, \mathbf{y}^C)$: features in the potential function
$\hat{\boldsymbol{\theta}}_n$: the maximum likelihood estimate
$\tilde{\boldsymbol{\theta}}_m$: the maximum pseudolikelihood estimate

Table 1: Notations used in this paper

the number of edges (e.g., pairwise clique specifications), or higher. The MLE $\hat{\boldsymbol{\theta}}$ can be written as: $\hat{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} L_n(\boldsymbol{\theta};\mathbf{x})$, where the normalized[1] log-likelihood function is:

$$L_n(\boldsymbol{\theta};\mathbf{x}) = \left\langle \boldsymbol{\theta}, \frac{1}{n} \sum_{C\in\mathcal{C}(G)} \boldsymbol{\phi}^C \right\rangle - \frac{1}{n} \log Z_n(\boldsymbol{\theta};\mathbf{x}) \quad (2)$$

Instead of optimizing the joint distribution over the whole training graph, the generalized maximum pseudolikelihood estimator (MPLE) partitions the graph into small subgraph components. In this paper, we use a templating formulation, similar to that of the potential cliques $\mathcal{C}(G)$, for the MPLE components $\mathcal{S}(G)$, where $\{S : \cup S = G\}$ and each component S intersects with m_S cliques. Let m denote the number of pseudolikelihood components in G, i.e., $m = |\mathcal{S}(G)|$. Since $m_S << n$, $m = O(n)$. Let ∂S be the set of cliques which intersect with S's Markov blanket, i.e., $\partial S = \{C : C \in \mathcal{C}(G)$ and $C \cap S \neq \emptyset$ and $\exists v \notin S$ s.t. $v \in C\}$. The MPLE objective is defined as a product of local conditional probabilities:

$$PL(\boldsymbol{\theta};\mathbf{x}) = \prod_{S\in\mathcal{S}(G)} P(\mathbf{y}^S|\mathbf{y}^{G\setminus S},\mathbf{x}) = \prod_{S\in\mathcal{S}(G)} P(\mathbf{y}^S|\mathbf{y}^{\partial S},\mathbf{x})$$

$$= \prod_{S\in\mathcal{S}(G)} \frac{1}{\tilde{Z}_S(\boldsymbol{\theta};\mathbf{x})} \exp\left\langle \boldsymbol{\theta}, \sum_{C:C\cap S\neq\emptyset} \boldsymbol{\phi}(\mathbf{x}^C,\mathbf{y}^C) \right\rangle \quad (3)$$

where the partition function is $\tilde{Z}_S(\boldsymbol{\theta};\mathbf{x}) = \int_{\mathcal{Y}^S} \exp\left\langle \boldsymbol{\theta}, \sum_{C:S\cap C\neq\emptyset} \boldsymbol{\phi}(\mathbf{x}^C,\mathbf{y}^C)\right\rangle d\mathbf{y}$.

The MPLE can be written as: $\tilde{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} \tilde{L}_n(\boldsymbol{\theta};\mathbf{x})$, where the normalized log-peudolikelihood function is:

$$\tilde{L}_m(\boldsymbol{\theta};\mathbf{x}) = \frac{1}{m} \sum_{S\in\mathcal{S}(G)} \sum_{C:C\cap S\neq\emptyset} \left(\left\langle \boldsymbol{\theta}, \boldsymbol{\phi}^C \right\rangle - \log \tilde{Z}_S(\boldsymbol{\theta};\mathbf{x}) \right) \quad (4)$$

Clearly, the key difference of the two estimators lies in the partition function. The MLE does not partition the graph and needs to enumerate all y variables jointly in G in the partition function. In general, the space on which the MLE integral is taken grows exponentially with the training data size n—which alludes to the computational infeasibility of MLE for relational learning, distinguishing it from i.i.d. learning. In contrast, the MPLE partitions the graph into components of bounded sizes and only enumerates the variables within each component. Therefore, the domain of the MPLE integral remains fixed as the training data size grows. The common practice in relational learning of considering subgraphs from the training data as i.i.d. samples and learning from them independently is thus, in effect, MPLE learning.

By standard results on exponential families (see, e.g. Wainwright and Jordan, 2008), the gradient of L_n and $\tilde{L}_m$, denoted by $\mathbf{f}_n$ and $\tilde{\mathbf{f}}_m$ respectively, are:

$$\mathbf{f}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{C\in\mathcal{C}(G)} \boldsymbol{\phi}^C - E_{\boldsymbol{\theta}}[\boldsymbol{\phi}|\mathbf{x}] \quad (5)$$

$$\tilde{\mathbf{f}}_m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{S\in\mathcal{S}(G)} \sum_{C:C\cap S\neq\emptyset} \left(\boldsymbol{\phi}^C - E_{\boldsymbol{\theta}}\left[\boldsymbol{\phi}|\mathbf{y}^{\partial S},\mathbf{x}\right]\right) \quad (6)$$

2.2 Weak Dependence and Convergence

We provide theoretical justification of learning from subgraphs by studying the asymptotic behavior of the estimators. To this end, we first need the notion of Markov networks of infinite size. Based on standard arguments in probability theory (see also Singla and Domingos, 2007), the following local finiteness assumptions are sufficient for our model (1) to be well defined at infinity.

Definition 1 (Local finiteness)**.** *The infinite Markov network is locally finite if the number of variables in each variable v's Markov blanket is bounded, i.e., there exists $0 < d < \infty$ such that $\sum_{C:v\in C} |C| \leq d$ for every $v \in G$. In addition, the feature values must be bounded as well: $\boldsymbol{\phi}(\cdot) \leq M < \infty$.*

A further sufficient condition for the learners to be well-behaved is a notion of weak dependence. The data are weakly dependent if the total covariance of any clique with all other cliques in the network is finite.

Definition 2 (Weak dependence)**.** *The Markov network satisfies the weak dependence condition if $\lim_{n\to\infty} E_{\boldsymbol{\theta}}[V[\boldsymbol{\phi}]]$ exists, where $V[\boldsymbol{\phi}] = \frac{1}{n}\sum_{C_1\in\mathcal{C}(G)}\sum_{C_2\in\mathcal{C}(G)} Cov\left[\boldsymbol{\phi}^{C_1},\boldsymbol{\phi}^{C_2}\right]$ is the auto-covariance of feature values across the whole network.*

In the literature, various forms of weak dependence (see e.g. Dedecker et al., 2007) have been used to prove limit theorems for dependent data. The definition presented herein differs from past definitions in two main aspects. First, previous definitions usually involve testing all bounded functions, while we

[1] We normalize the log-likelihood because the unnormalized value goes to infinity as $n \to \infty$. The normalized version is also convenient to work with for proving consistency.

relax it to only consider the covariance of the feature function $\boldsymbol{\phi}$, which suffices for our purpose. Second, in related work, weak dependence conditions are primarily considered for time series and lattice data. These conditions can be adapted for graphs of subexponential growth to give sufficient conditions of our weak dependence assumption. For example, exponential decay in correlation (with respect to graph distance) implies that Definition 2 holds. While it is computationally infeasible to check weak dependence conditions empirically, part of our current work is deriving sufficient conditions for weak dependence which are convenient to work with while remains wide applicability.

Throughout this paper, we focus on networks where weak dependence holds. We note that infinite Markov networks which fail to satisfy the weak dependence condition are not well behaved in general and the corresponding learning algorithm will tend to perform poorly. For example, perturbation of values at a single node may cause global label changes, which could make the algorithm oscillate between very different labelings. The break of weak dependence is closely related (although not identical) to the phenomenon of phase transition in statistical physics (Dobruschin, 1968). Loosely speaking, phase transitions in infinite Markov networks occur when different configurations at the boundary cause different probability distributions over inner nodes. In many application areas of relational learning like social networks and web graphs, such dramatic long range influence is not likely.

An immediate consequence of the weak dependence condition is the following law of large numbers for interdependent feature values.

Proposition 1. *Assuming weak dependence and the true parameter vector $\boldsymbol{\theta}$, the following holds:*

i) $\frac{1}{n}\sum_{C\in\mathcal{C}(G)} \boldsymbol{\phi}^C \xrightarrow{P} E_{\boldsymbol{\theta}}[\boldsymbol{\phi}]$.

ii) $\mathcal{V}(\boldsymbol{\theta}) \equiv \lim_{n\to\infty} E_{\boldsymbol{\theta}}[V_n[\boldsymbol{\phi}|\mathbf{x}]]$ *exists.*

iii) $E_{\boldsymbol{\theta}}\left[\boldsymbol{\phi}|\mathbf{x}^{G_n}\right] \xrightarrow{P} E_{\boldsymbol{\theta}}[\boldsymbol{\phi}]$.

Proof Sketch
i): For any $t>0$, by Markov's inequality,

$$P(||\frac{1}{n}\sum_{C\in\mathcal{C}(G)} \boldsymbol{\phi}^C - E_{\boldsymbol{\theta}}[\boldsymbol{\phi}]||^2 \geq t)$$
$$\leq \frac{1}{t}\int_{\mathcal{X},\mathcal{Y}} ||\frac{1}{n}\sum_{C\in\mathcal{C}(G)} \boldsymbol{\phi}(\mathbf{x}^C,\mathbf{y}^C) - E_{\boldsymbol{\theta}}[\boldsymbol{\phi}]||^2 d(\mathbf{x},\mathbf{y})$$
$$= \frac{1}{tn}V[\boldsymbol{\phi}]$$

Letting $n\to\infty$, $P(||\frac{1}{n}\sum_{C\in\mathcal{C}(G)} \boldsymbol{\phi}^C - E_{\boldsymbol{\theta}}[\boldsymbol{\phi}]||^2 \geq t)\to 0$.
ii): By the law of total covariance, $E_{\boldsymbol{\theta}}[V[\boldsymbol{\phi}|\mathbf{x}]] \leq E_{\boldsymbol{\theta}}[V[\boldsymbol{\phi}]]$. Thus, the increasing sequence $\{E_{\boldsymbol{\theta}}[V[\boldsymbol{\phi}|\mathbf{x}]]\}$ is bounded, and its limit exists.
iii) is proved in the same way as i). □

3 Asymptotic Analysis of Estimators

In this section, we establish asymptotic consistency and normality arguments for MLE and MPLE, in order to theoretically justify relational learning from a single network. We also analyze the efficiency of the learners by comparing their asymptotic variance. Throughout this section, compactness of the parameter space and identifiability of parameters[2] are assumed for the convenience of theoretical analysis.

3.1 Asymptotic Consistency

We use the following lemma, which appears in van der Vaart (2000, Theorem 5.7) to prove consistency of the estimators.

Lemma 1. *Let $\hat{\boldsymbol{\theta}}_n$ be a sequence of estimators, M_n be random functions, M be a fixed function of $\boldsymbol{\theta}$ such that for every $\delta>0$,*

$$\sup_{\boldsymbol{\theta}} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow{P} 0, \tag{7}$$

$$\sup_{\boldsymbol{\theta}:||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\geq\delta} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0) \tag{8}$$

$$M_n(\hat{\boldsymbol{\theta}}_n) \geq M_n(\boldsymbol{\theta}_0) - \epsilon(n) \tag{9}$$

where $\epsilon(n)\xrightarrow{P}0$. Then $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \hat{\boldsymbol{\theta}}_0$.

In order to discuss consistency, we need to consider the generative distribution for the data, which is the full Markov network parameterized by θ_0:

$$P_{\theta_0}(\mathbf{x},\mathbf{y}) = \frac{1}{Z_n(\boldsymbol{\theta}_0)}\exp\left\{\left\langle \boldsymbol{\theta}_0, \sum_{C\in\mathcal{C}(G)} \boldsymbol{\phi}(\mathbf{x}^C,\mathbf{y}^C)\right\rangle\right\} \tag{10}$$

where $Z_n(\boldsymbol{\theta}_0) = \int_{\mathcal{X}^G,\mathcal{Y}^G} \exp\left\langle \boldsymbol{\theta}_0, \sum_{c\in\mathcal{C}(G)} \boldsymbol{\phi}(\mathbf{x}^C,\mathbf{y}^C)\right\rangle d(\mathbf{x},\mathbf{y})$. Our general conditional model (1) differs from the generative distribution (10) in that while the generative distribution is over the whole space $(\mathcal{X}^G,\mathcal{Y}^G)$, the conditional learner only varies over $\mathcal{Y}^G$. We thus relate the estimate based on the conditional likelihood (resp. pseudolikelihood) to that based on the full generative likelihood (resp. pseudolikelihood). The log partition function of the generative likelihood (resp. pseudolikelihood) is denoted by $A(\theta)$ (resp. $\tilde{A}(\theta)$) in the following proofs. Weak dependence, local finiteness and the convexity of log partition functions are instrumental in establishing the conditions of Lemma 1.

[2]Nonidentifiable parameterization could be useful in practice due to interpretability. In such circumstances the validity of learning needs to be established based on equivalence class arguments.

Proposition 2. *Let $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimator based on graph G with n cliques, generated from $P_{\boldsymbol{\theta}_0}$ (as defined by Equation 10). Then $\hat{\boldsymbol{\theta}}_n$ is asymptotically consistent: $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.*

Proof Sketch
Let $A_n(\boldsymbol{\theta}) = \frac{1}{n}\log\int_{\mathcal{X}^G,\mathcal{Y}^G}\exp\left\langle\boldsymbol{\theta}, \sum_{c\in\mathcal{C}(G)}\boldsymbol{\phi}^C\right\rangle d(\mathbf{x},\mathbf{y})$. Given local finiteness, $\{A_n(\boldsymbol{\theta})\}$ is bounded and its limit exists, which we denote by $A(\boldsymbol{\theta}) = \lim_{n\to\infty} A_n(\boldsymbol{\theta})$. Furthermore, $A_n(\boldsymbol{\theta})$ is convex and we denote its conjugate function by A_n^*, i.e., $A_n^*(\boldsymbol{\phi}) = \sup_{\boldsymbol{\theta}}\langle\boldsymbol{\theta},\boldsymbol{\phi}\rangle - A_n(\boldsymbol{\theta})$. In Lemma 1, let $M_n(\boldsymbol{\theta}) = \left\langle\boldsymbol{\theta}, \frac{1}{n}\sum_{C\in\mathcal{C}(G)}\boldsymbol{\phi}^C\right\rangle - A_n(\boldsymbol{\theta})$, and $M(\boldsymbol{\theta}) = \langle\boldsymbol{\theta}, E_{\boldsymbol{\theta}}[\boldsymbol{\phi}]\rangle - A(\boldsymbol{\theta})$. We prove the consistency by checking each of the assumptions. Equation (7) follows from Proposition 1 and the compactness of parameter space. Moreover, for $\forall\delta>0, \forall\boldsymbol{\theta}$ s.t. $||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\geq\delta$:

$$M(\boldsymbol{\theta}_0) - M(\boldsymbol{\theta}) = \lim_{n\to\infty}\frac{1}{n}E_{\boldsymbol{\theta}_0}\Big[\log P_{\boldsymbol{\theta}_0}(\mathbf{y}^G|\mathbf{x}^G) - \log P_{\boldsymbol{\theta}}(\mathbf{y}^G|\mathbf{x}^G)\Big] \tag{11}$$

which is the limit of the KL divergence (rescaled by $\frac{1}{n}$) between the distributions parameterized by $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. To show this limit is strictly positive, pick a largest subset of cliques $\check{G}$ from G, such that $\forall C_i, C_j\in\check{G}, C_i\cap C_j = \emptyset$. Due to local finiteness, $|\check{G}|\geq\lceil n/d^2\rceil$. Define $\bar{G} = G\setminus\check{G}$. Furthermore, let $P_{\boldsymbol{\theta}}(R)$ denote $P_{\boldsymbol{\theta}}(\mathbf{x}^R,\mathbf{y}^R)$ and $P_{\boldsymbol{\theta}}(R|T)$ denote $P_{\boldsymbol{\theta}}(\mathbf{x}^R,\mathbf{y}^R|\mathbf{x}^T,\mathbf{y}^T)$ for any node sets R and T. We rewrite Equation (11) as

$$\begin{aligned}
&\frac{1}{n}\mathrm{KL}\left(P_{\boldsymbol{\theta}_0}(G)||P_{\boldsymbol{\theta}}(G)\right)\\
=&\frac{1}{n}\mathrm{KL}\left(P_{\boldsymbol{\theta}_0}(\check{G}|\bar{G})||P_{\boldsymbol{\theta}}(\check{G}|\bar{G})\right) + \mathrm{KL}(P_{\boldsymbol{\theta}_0}(\bar{G}||P_{\boldsymbol{\theta}}(\bar{G}))\\
\geq&\frac{1}{n}\mathrm{KL}\left(P_{\boldsymbol{\theta}_0}(\check{G}|\bar{G})||P_{\boldsymbol{\theta}}(\check{G}|\bar{G})\right)\\
\geq&\frac{1}{d^2}\mathrm{KL}\left(P_{\boldsymbol{\theta}_0}(C|\partial C)||P_{\boldsymbol{\theta}}(C|\partial C)\right)
\end{aligned}$$

Note that due to the identifiability of parameters, the KL divergence between the local probabilities is positive, i.e., $\exists\eta>0$ such that $M(\boldsymbol{\theta}_0) - M(\boldsymbol{\theta}) \geq \frac{1}{d^2}\mathrm{KL}\left(P_{\boldsymbol{\theta}_0}(C|\partial C)||P_{\boldsymbol{\theta}}(C|\partial C)\right) \geq \frac{\eta}{d^2} > 0$ i.e., Equation (8) holds. Finally, we check Equation (9). By Equation (5) $E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}|\mathbf{x}] = \frac{1}{n}\sum_{C\in\mathcal{C}(G)}\boldsymbol{\phi}^C \xrightarrow{P} E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]$, while $\frac{1}{n}\sum_{C\in\mathcal{C}(G)}\boldsymbol{\phi}^C \xrightarrow{P} E_{\boldsymbol{\theta}_0}$ by Proposition 1, so $E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}|\mathbf{x}] \xrightarrow{P} E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]$. On the other hand, $E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}|\mathbf{x}] \xrightarrow{P} E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}]$. Therefore, $E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}] \xrightarrow{P} E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]$. $M_n(\boldsymbol{\theta}_0) - M_n(\hat{\boldsymbol{\theta}}_n) = A_n^*(E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]) - A_n^*(E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}]) + \left\langle\hat{\boldsymbol{\theta}}_n, E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}]\right\rangle - \langle\boldsymbol{\theta}_0, E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]\rangle$. Since $E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}] \xrightarrow{P} E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]$, $A_n^*(E_{\hat{\boldsymbol{\theta}}_n}[\boldsymbol{\phi}]) \xrightarrow{P} A_n^*(E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}])$ by continuous mapping theorem. $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$ are bounded. Therefore, the RHS converge to 0. We thus get Equation (9). □

Proposition 3. *Let $\tilde{\boldsymbol{\theta}}_m$ be the maximum pseudolikelihood estimator based on graph G with m pseudolikelihood components, generated from $P_{\boldsymbol{\theta}_0}$. Then $\tilde{\boldsymbol{\theta}}_m$ is asymptotically consistent: $\tilde{\boldsymbol{\theta}}_m \xrightarrow{P} \boldsymbol{\theta}_0$.*

Proof Sketch
Similar to the previous proof, let $\tilde{A}_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{S\in\mathcal{S}(G)}\log\int_{\mathcal{X}^S,\mathcal{Y}^S}\exp\left\langle\boldsymbol{\theta}, \sum_{C:C\cap S\neq\emptyset}\boldsymbol{\phi}^C\right\rangle d(\mathbf{x},\mathbf{y})$, and $\tilde{A}(\boldsymbol{\theta}) = \lim_{m\to\infty}\tilde{A}_m(\boldsymbol{\theta})$. In Lemma 1, let $M_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{S\in\mathcal{S}(G)}\sum_{C:C\cap S\neq\emptyset}\boldsymbol{\phi}^C - \tilde{A}_m(\boldsymbol{\theta})$, and $M(\boldsymbol{\theta}) = \langle\boldsymbol{\theta}, E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]\rangle - \tilde{A}(\boldsymbol{\theta})$. Then Equation (7) and (9) are verified in the same way as the MLE case. We only check Equation (8) here. For $\forall\delta>0, \forall\boldsymbol{\theta}$ s.t. $||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\geq\delta$:

$$\begin{aligned}
&M(\boldsymbol{\theta}_0) - M(\boldsymbol{\theta})\\
=&\langle\boldsymbol{\theta}, E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]\rangle - \tilde{A}(\boldsymbol{\theta}_0) - \left(\langle\boldsymbol{\theta}, E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}]\rangle - \tilde{A}(\boldsymbol{\theta})\right)\\
=&\lim_{m\to\infty}\frac{1}{m}\sum_{S\in\mathcal{S}(G)}E_{\boldsymbol{\theta}_0}\left(\log P_{\boldsymbol{\theta}_0}(S|\partial S) - \log P_{\boldsymbol{\theta}}(S|\partial S)\right)\\
=&\lim_{m\to\infty}\frac{1}{m}\sum_{S\in\mathcal{S}(G)}KL\left(P_{\boldsymbol{\theta}_0}(S|\partial S)||P_{\boldsymbol{\theta}}(S|\partial S)\right)
\end{aligned}$$

where again the local KL divergence is positive due to identifiability of the parameters. Due to local finiteness of the graph and boundedness of pseudolikelihood components, as m increases, all possible pseudolikelihood components can be divided into a finite number of isometric classes. Since the number of different pseudolikelihood classes is finite, there exists η such that for any S, $KL\left(P_{\boldsymbol{\theta}_0}(S|\partial S)||P_{\boldsymbol{\theta}}(S|\partial S)\right) \geq \eta > 0$. Therefore, $M(\boldsymbol{\theta}_0) - M(\boldsymbol{\theta}) \geq \eta > 0$. □

Our consistency results justify both joint and disjoint learning over observed subgraphs from the underlying whole graph.

3.2 Asymptotic Normality

To establish asymptotic normality for relational estimators, we need appropriate forms of central limit theorems for dependent data. Lemma 2 and 3 serve for this purpose (see appendix for proofs).

Lemma 2. *Let $\mathbf{f}$ be the gradient of log-likelihood function, as defined in Equation 5, and $\mathcal{V}(\boldsymbol{\theta}_0)$ as defined in Proposition 1. Then $\sqrt{n}\mathbf{f}_n(\boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathcal{V}(\boldsymbol{\theta}_0))$.*

Proposition 4. *Assuming weak dependence, the maximum likelihood estimator is asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathcal{V}(\boldsymbol{\theta}_0)^{-1}\right) \tag{12}$$

Proof Sketch
We apply the method describe in van der Vaart (2000, Section 5.3). Apply Taylor expansion to $\mathbf{f}_n$, and note that the MLE $\hat{\boldsymbol{\theta}}_n$ is the zero point of $\mathbf{f}_n(\boldsymbol{\theta})$: $\mathbf{0} = \mathbf{f}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{f}_n(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\nabla\mathbf{f}_n(\boldsymbol{\theta}_0) + o(||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||^2)$ Reagrranging the terms, and multiplying both sides by $\sqrt{n}$, we get $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\nabla\mathbf{f}_n(\boldsymbol{\theta}_0)^{-1}\sqrt{n}\mathbf{f}_n(\boldsymbol{\theta}_0) - o(||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||^2)$. Since $\nabla\mathbf{f}_n(\boldsymbol{\theta}_0) \to -\mathcal{V}(\boldsymbol{\theta}_0)$, and $\mathbf{f}_n(\boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathcal{V}(\boldsymbol{\theta}_0))$ by Lemma 2, Equation (12) follows. □

Next, we consider the maximum pseudolikelihood estimator. The asymptotic normality of the MPLE is closely related to two covariance matrices. Intuitively, they can be viewed as the "within component" autocovariance and the "across network" covariance. The autocovariance of feature values within each component is: $\tilde{\mathcal{V}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\tilde{V}[\boldsymbol{\phi}|\mathcal{S},\mathbf{x}]\right]$, where:

$$\tilde{V}[\boldsymbol{\phi}|\mathcal{S},\mathbf{x}] = \frac{1}{m}\sum_{S\in\mathcal{S}}\sum_{C_1:C_1\cap S\neq\emptyset}\sum_{C_2:C_2\cap S\neq\emptyset}\mathrm{Cov}\left[\mathbf{u}^{C_1},\mathbf{u}^{C_2}\right]$$

and $\mathbf{u}^{C_i} = \boldsymbol{\phi}^{C_i} - E_{\boldsymbol{\theta}}\left[\boldsymbol{\phi}|\mathbf{y}^{\partial S},\mathbf{x}\right]$. We further denote the covariance of component conditional values of features across the network by $\mathcal{C}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[C[\boldsymbol{\phi}|\mathbf{x}]\right]$, where:

$$C[\boldsymbol{\phi}|\mathbf{x}] = \frac{1}{m}\sum_{S_1,S_2\in\mathcal{S}(G)}\sum_{C_1:C_1\cap S_1\neq\emptyset}\sum_{C_2:C_2\cap S_2\neq\emptyset}\mathrm{Cov}\left[\mathbf{u}^{C_1},\mathbf{u}^{C_2}\right]$$
$$= \frac{1}{m}\sum_{C_1,C_2:C_1\bowtie C_2}\mathrm{Cov}\left[\mathbf{u}^{C_1},\mathbf{u}^{C_2}\right]$$

Here $C_1 \bowtie C_2$ refers to the set of C where $\exists S \in \mathcal{S}(G)$ s.t. $C_1 \cap (S \cup \partial S) \neq \emptyset$ and $C_2 \cap (S \cup \partial S) \neq \emptyset$. The second equality follows from the fact that cliques outside of the Markov blanket of each other are conditionally independent.

We are now ready to state the following central limit theorem for $\tilde{\mathbf{f}}_m$, and consequently the asymptotic normality of MPLE.

Lemma 3. $\sqrt{m}\tilde{\mathbf{f}}_m(\boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathcal{C}(\boldsymbol{\theta}_0))$.

Proposition 5. *The maximum pseudolikelihood estimator is asymptotically normal:*

$$\sqrt{m}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \tilde{\mathcal{V}}(\boldsymbol{\theta}_0)^{-1}\mathcal{C}(\boldsymbol{\theta}_0)\tilde{\mathcal{V}}(\boldsymbol{\theta}_0)^{-1}\right) \quad (13)$$

Proof Sketch
Similar to the proof for MLE, we apply Taylor expansion to $\tilde{\mathbf{f}}_n$:

$$\mathbf{0} = \tilde{\mathbf{f}}_m(\tilde{\boldsymbol{\theta}}_m) = \tilde{\mathbf{f}}_m(\boldsymbol{\theta}_0) + (\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0)\nabla\tilde{\mathbf{f}}_m(\boldsymbol{\theta}_0) + o(||\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0||^2)$$

Observe that $\nabla\tilde{\mathbf{f}}_m(\boldsymbol{\theta}_0) \to -E_{\boldsymbol{\theta}_0}\left[\tilde{V}[\boldsymbol{\phi}|\mathcal{S},\mathbf{x}]\right] = -\tilde{\mathcal{V}}(\boldsymbol{\theta}_0)$. Combined with Lemma 3, this gives Equation 13. □

Although MLE can be regarded as MPLE with just one component, for which $\tilde{\mathcal{V}}$ and $\mathcal{C}$ are well defined, and the resulting asymptotic variance coincides with $\mathcal{V}^{-1}$, Proposition 4 and Proposition 5 are based on two extremely different conditions: the former a single component of infinite size, while the latter infinitely many components with fixed size. In practice, due to the intractability of MLE on large networks, we tend to use a finite number of small subgraphs for estimation. In this case, the asymptotic variance of MPLE suggests that partitioning the graph so as to maximize within-component variance $\tilde{V}$ would make the learner perform better as it reduces the asymptotic variance. This will be further discussed in Section 4.

In addition, when the pseudolikelihood components are independent of each other, then $\tilde{\mathcal{V}} = \frac{1}{m}\mathcal{C} = \frac{1}{n}\mathcal{V}$ and the asymptotic variance of the MLE and MPLE become identical. Thus, the asymptotic normality results of both MLE and MPLE naturally generalize those of the i.i.d. case. In general, however, the MLE is asymptotically more efficient than the MPLE in the sense of matrix determinant comparison. See appendix for proof.

Proposition 6. *The MLE is asymptotically more efficient than the MPLE, i.e.,*

$$det\left(\frac{1}{n}\mathcal{V}^{-1}\right) \leq det\left(\frac{1}{m}\tilde{\mathcal{V}}^{-1}\mathcal{C}\tilde{\mathcal{V}}^{-1}\right) \quad (14)$$

Finally, we compare maximum pseudolikelihood estimators with different component partitions. Consider two pseudolikelihood estimators $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_2$, which consist of components $\mathcal{S}_1(G)$ and $\mathcal{S}_2(G)$ respectively. We say that $\mathcal{S}_2(G)$ is a *finer partition* compared to $\mathcal{S}_1(G)$, if for every $S_2 \in \mathcal{S}_2$, there is some $S_1 \in \mathcal{S}_2$ s.t. $S_2 \subseteq S_1$. See appendix for proof.

Proposition 7. *If $\mathcal{S}_2(G)$ is finer than $\mathcal{S}_1(G)$, then $\tilde{\boldsymbol{\theta}}_1$ is asymptotically more efficient than $\tilde{\boldsymbol{\theta}}_2$, i.e.,*

$$det\left(\frac{1}{m_1}\tilde{\mathcal{V}}_1^{-1}\mathcal{C}_1\tilde{\mathcal{V}}_1^{-1}\right) \leq det\left(\frac{1}{m_2}\tilde{\mathcal{V}}_2^{-1}\mathcal{C}_2\tilde{\mathcal{V}}_2^{-1}\right) \quad (15)$$

4 Experiments

In this section, we investigate the empirical performance of relational learners on finite data. Since MLE is not tractable on networks of moderate to large sizes, we only demonstrate its performance on small synthetic data. Our main focus is on comparing different generalized MPLEs. For high order MPLE components, different constructions will result in different covariance structures, and thus their form is expected to impact the performance (i.e., convergence) of the model. According to Proposition 5, in order to maximize performance, component construction should aim to maximize the within-component covariance $\tilde{\mathcal{V}}$ and to minimize the between-component covariance $\mathcal{C}$.[3] However, it is computationally intractable to directly optimize either of these covariance matrices in practice because the evaluation would involve running inference across the whole network while varying each clique. Therefore, we need to resort to some heuristic choice of components to approximate the analytical objective. Since network links directly affect component covariance, it may appear that graph clustering could be used as a heuristic to select components that

[3]Minimizing $\mathcal{C}$ is generally a secondary concern, but usually it can be optimized simultaneously with $\tilde{\mathcal{V}}$.

satisfy this objective. However, we note that conventional graph clustering algorithms are not immediately applicable since we are mainly interested in small, and possibly overlapping, subgraphs for the MPLE components. As such, we investigate the following methods of component construction:

- *Singleton*: Each node in the network forms a component.
- *Edgewise*: Each pair of nodes connected by an edge in the network forms a component.
- *Random*: Starting from edgewise components, we randomly expand each component to a particular size. The component is expanded with a random choice of BFS or DFS at each step.
- *Heuristic*: Starting from edgewise components, we greedily expand each component by choosing, at each step, to add the node with maximum number of links to the nodes already in the component, with a discount to reflect the number of times that the node (and edge) have already appeared in other components. This approach is a heuristic designed to maximize $\tilde{\mathcal{V}}$.

4.1 Experiments on Synthetic Data

We use synthetic data to directly evaluate the quality of parameter estimation under two levels of network connectivity and two levels of model dependence. These experiments are intended to provide further insights into model performance in practical situations when only limited data is available and to explore whether behavior in the finite data region matches that predicted by the asymptotic analysis.

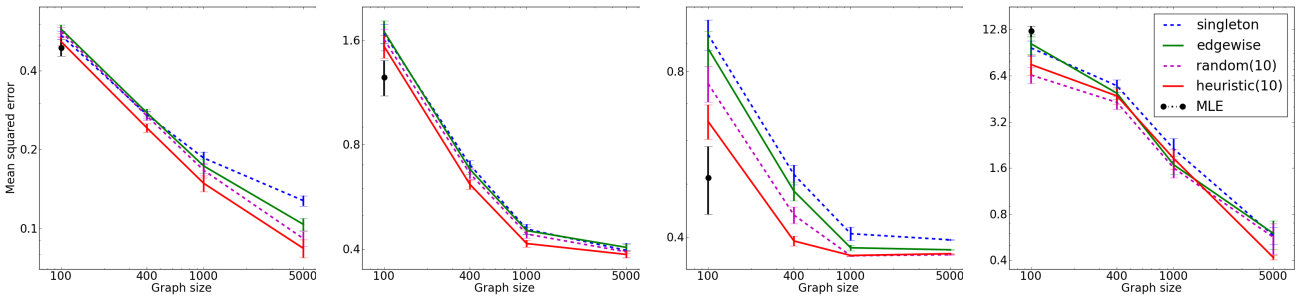
We use the Erdos-Renyi random graph model to generate synthetic network structures with high and low linkage (average degree of 10 and 5 respectively). Then we used a manually-specified Markov network model to generate a binary attribute and class label on the nodes, exploring both high and low interaction parameters. Details are in the appendix.

For all learners we use the BFGS optimizer. For MLE we use Gibbs Sampling to compute the gradient and value of likelihood function in each optimization step, while we perform exact computation by brute force enumeration for all MPLE components. We report mean squared error (MSE= $E||\theta - \theta_0||^2$) for evaluation. Figure 1 plots MSE as training data size increases, averaged over [200, 200, 80, 20] trials for network sizes [100, 400, 1000, 5000] respectively. In general, the learning algorithms converge more quickly when both the level of linkage and local interactions are low, while they converge most slowly when both the linkage and interaction levels are high. This is due to the fact that both denser graph structure and larger interaction parameters result in higher component variance. In all but the high linkage, high interaction case, the behavior of the learners is consistent with the asymptotic analysis. While the MLE attains lowest estimation error, it is only feasible for small networks. The heuristic component construction achieves the lowest error among all pseudoliklihood methods. In the exceptional high linkage, high interaction case, the performance of the learners in the small data regime seems to contradict the asymptotic prediction. While the heuristic approach still achieves best performance at large network sizes, the random construction attains the best performance for small to moderate network sizes. This is because when both linkage and interaction levels are high, the covariance across the network is too large relative to the network size and our assumption of weak dependence is violated, hence the asymptotic analysis is no longer applicable. While the empirical results seems to suggest that MPLE components based on random construction are more stable in such scenarios, this warrants further investigation before making this conclusion.

4.2 Experiments on Facebook Data

We next apply the relational learners to data from a large "University" Facebook network. The data include friendship links, transactions among user (e.g., wall posts), as well as user profile attributes. We divide the network into 8 articulated subnetworks of comparable sizes for training and testing (e.g., "Class of 2008") and evaluate performance on four different classification tasks (predicting profile attributes based on other profile attributes and connections in the friendship network). See appendix for details.

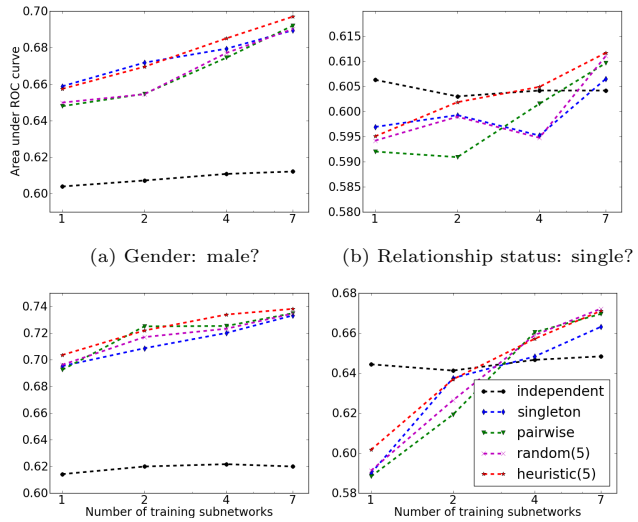
The classification results are shown in Figure 2. We use one test subnetwork for evaluation, while varying the number of subnetworks used for training, and report average results over the 8 subnetworks. As a baseline for comparison, we also include an *independent* learning approach which treats each user as an i.i.d. instance. As the size of the training graph grows, the independent learning curve remains relatively flat, while the relational learners improve their performance. This is due to the fact that relational models explore the covariance structure of the data, and take more training instances to converge to their optimum. When the autocorrelation in the data is high (as is the case for the tasks "gender" and "political view"), the relational learners perform constantly better than i.i.d. learning. When the autocorrelation is low (as is the case for the tasks "relationship status" and "religious view"), the relational estimates tend to be biased for small dataset sizes, but still outperform independent learning as more data becomes available. Similar to the synthetic data results, the heuristic component construction achieves superior performance in most cases among all the relational learners. While the



(a) Low linkage, low interaction (b) Low linkage, high interaction (c) High linkage, low interaction (d) High linkage, high interaction

Figure 1: Average estimation MSE over randomly-generated synthetic-data training networks (log-log scale).

computational cost is the same, the random construction performs much worse. This demonstrate the benefit of exploiting the structure of the graph in the component construction. The singleton and pairwise approaches are inferior to high-order constructions when only moderate amount of training data is available. However, we observe that they eventually converge to almost the same optimum as the high-order approaches at the maximum size of training network.



(a) Gender: male? (b) Relationship status: single?

(c) Political view: conservative? (d) Religious view: Christian?

Figure 2: Classification results on Facebook, as the number of training subnetworks K is varied.

5 Related Work

Markov networks that tie parameters through templating is a popular approach to model relational and network data. Our analysis builds upon the models and learning algorithms proposed by Taskar et al. (2002), Richardson and Domingos (2006), and Neville and Jensen (2007). These algorithms have been successfully applied in single-network settings. However to date, the theoretical issues of model convergence, consistency and efficiency in single-network domains with data interdependence have not been explored.

Asymptotic analysis for i.i.d. estimators and its implication in finite data learning have received considerable attention in the machine learning community. In (Liang and Jordan, 2008), the asymptotic analysis is motivated by comparing the performance of generative, discriminative and pseudolikelihood estimators. Dillion and Lebanon (2009) exploited an understanding of the asymptotic variance to make better choice of the pseudolikelihood component weights. In relational domains, analyzing the asymptotic covariance provides us with additional insights as to how the covariance structure of the network affects learning, and how to exploit that understanding to construct better pseudolikelihood components.

6 Conclusion

In this paper, we have performed an asymptotic analysis of relational learning methods applied to a single network. We illustrate the findings with an empirical exploration of the performance of MPLEs with different component construction schemes in various conditions. This provides a starting point for more sophisticated approaches to constructing MPLE components and controlling the level of joint learning in network data. We believe that further exploration in this direction is important for the relational learning community, since relational models grow with the size of the training data, making full MLE intractable to apply in practice.

As part of future work, we plan to explore other important theoretical issues which have not been addressed in this paper. One of particular interest is the account for model mis-specification, i.e., how the learners perform when the true underlying distribution falls outside of the model family. While our experimental results on Facebook data seem to suggest MPLE based on heuristic component construction is quite robust, it is worth pursuing a formal analysis of this setting.

Acknowledgements

We thank anonymous reviewers for helpful comments. This research is supported by NSF under contract numbers SES-0823313, IIS-1017898, CCF-0939370. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

References

E. Bolthausen. On the central limit theorem for stationary mixing random fields. *Ann. Probab.*, 10(4):1047–1050, 1982.

Francis Comets. On consistency of a class of estimators for exponential families of markov random fields on the lattice. *The Annals of Statistics*, 20(1):pp. 455–468, 1992.

J. Dedecker, P. Doukhan, G. Lang, J.R. Leon, S. Louhichi, and C. Prieur. *Weak Dependence: With Examples and Applications*. Springer, 2007.

Joshua Dillion and Guy Lebanon. Statistical and computational tradeoffs in stochastic composite likelihood. In *AISTATS*, 2009.

P. L. Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probab. Appl.*, 13:197–224, 1968.

Percy Liang and Michael I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *ICML '08: Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

Jennifer Neville and David Jensen. Relational dependency networks. *J. Mach. Learn. Res.*, 8:653–692, 2007. ISSN 1532-4435.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Mach. Learn.*, 62(1-2):107–136, 2006. ISSN 0885-6125.

Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2): 173–191, May 2007.

P. Singla and P. Domingos. Markov logic in infinite domains. In *Proceeding of the 23rd conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2007.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob.*, 2:583–602, 1972.

Charles Sutton and Andrew Mccallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006.

Benjamin Taskar, Pieter Abbeel, and Daphne Koller. Discriminative probabilistic models for relational data. In *Proc. of the 18th Conference in Uncertainty in Artificial Intelligence*, 2002.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge, 2000.

Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.

A APPENDIX—SUPPLEMENTARY MATERIAL

A.1 Proofs of Central Limit Theorems

To prove the essential central limit theorems, we use a standard technique for dependent data, developed by Bolthausen (1982), which is based on the well-known Stein's method(Stein (1972)).

Proof Sketch of Lemma 2

Let $\mathbf{a}$ be any nonzero vector of the same dimension as $\mathbf{f}_n$, $u_k = \left\langle \mathbf{a}, \boldsymbol{\phi}(\mathbf{x}^{C_k}, \mathbf{y}^{C_k}) \right\rangle - E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}|\mathbf{x}]$, $z_n \equiv \frac{1}{\sigma_n}\sum_{k=1}^n u_k = \frac{1}{\sigma_n}\langle \mathbf{a}, n\mathbf{f}_n\rangle$, where $\sigma_n^2 \equiv E_{\boldsymbol{\theta}_0}\left[\left(\sum_{k=1}^n u_k\right)^2\right] \to n\mathbf{a}^T\mathcal{V}(\boldsymbol{\theta}_0;\mathbf{x})\mathbf{a}$. The proof then reduces to showing that z_n is asymptotically standard normal. By Lemma 2 in Bolthausen (1982), it suffices to show, for any $\lambda \in \mathcal{R}$,

$$\lim_{n\to\infty} E_{\boldsymbol{\theta}_0}\left[(i\lambda - z_n)\exp(i\lambda z_n)\right] = 0 \tag{16}$$

Apply the decomposition

$$(i\lambda - z_n)\exp(i\lambda z_n) = Q_1 + Q_2$$

where

$$Q_1 = i\lambda\exp(i\lambda z_n)\left(1 - \frac{1}{\sigma_n}\sum_{k=1}^n u_k z_n\right)$$

$$Q_2 = \frac{1}{\sigma_n}\exp(i\lambda z_n)\sum_{k=1}^n u_k(i\lambda z_n - 1)$$

We have

$$E_{\boldsymbol{\theta}_0}|Q_1| \le cE_{\boldsymbol{\theta}_0}\left|1 - \frac{1}{\sigma_n}\sum_{k=1}^n u_k z_n\right| = 0$$

and $E_{\boldsymbol{\theta}_0}|Q_2| \le cE_{\boldsymbol{\theta}_0}\left|\frac{1}{\sigma_n}\sum_{k=1}^n u_k z_n\right| + c'E_{\boldsymbol{\theta}_0}\left|\frac{1}{\sigma_n}\sum_{k=1}^n u_k\right|$. Note that $\sigma_n = O(n)$ and $E_{\boldsymbol{\theta}_0}\left|\sum_{k=1}^n u_k z_n\right| \to \mathbf{a}^T\mathcal{V}(\boldsymbol{\theta}_0;\mathbf{x})\mathbf{a}$. Therefore we get $E_{\boldsymbol{\theta}_0}|Q_2| \to 0$. □

Proof Sketch of Lemma 3

Similarly, let $u_k = \left\langle \mathbf{a}, \sum_{C:C\cap S_k\neq\emptyset}\boldsymbol{\phi}(\mathbf{x}^C, \mathbf{y}^C)\right\rangle - m_{S_k}E_{\boldsymbol{\theta}_0}[\boldsymbol{\phi}|\mathbf{y}^{\partial S_k}, \mathbf{x}]$, $z_m \equiv \frac{1}{\sigma_m}\sum_{k=1}^m u_k = \frac{1}{\sigma_m}\left\langle \mathbf{a}, m\tilde{\mathbf{f}}_m\right\rangle$, where $\sigma_m^2 \equiv E_{\boldsymbol{\theta}_0}\left[\left(\sum_{k=1}^m u_k\right)^2\right] \to m\mathbf{a}^T\mathcal{C}(\boldsymbol{\theta}_0;\mathbf{x})\mathbf{a}$. Furthermore, define $z_{m,k} = \frac{1}{\sigma_m}\sum_{j:S_j\cap S_k\neq\emptyset} u_k$. $z_{m,k}$ is bounded by local finiteness. We show Equation (16) based on the follow decomposition.

$$(i\lambda - z_m)\exp(i\lambda z_m) = Q_1 + Q_2 + Q_3$$

where

$$Q_1 = i\lambda\exp(i\lambda z_m)\left(1 - \frac{1}{\sigma_m}\sum_{k=1}^m u_k z_m\right)$$

$$Q_2 = \frac{1}{\sigma_m}\exp(i\lambda z_m)\sum_{k=1}^m u_k(i\lambda z_{m,k} + \exp(-i\lambda z_{m,k}) - 1)$$

$$Q_3 = -\frac{1}{\sigma_m}\sum_{k=1}^m u_k\exp\left(i\lambda(z_m - z_{m,k})\right)$$

Again, $E_{\boldsymbol{\theta}_0}|Q_1| = 0$ holds. Furthermore,

$$E_{\boldsymbol{\theta}_0}|Q_2| \le cE_{\boldsymbol{\theta}_0}\left|\frac{1}{\sigma_m}\sum_{k=1}^m u_k\frac{(\lambda z_{m,k})^2}{2}\right|$$

where we have used an identity $|iv + \exp(-iv) - 1| \leq \frac{v^2}{2}$. Since $z_{m,k} = O(\frac{1}{m})$,

$$E_{\boldsymbol{\theta}_0}|Q_2| \leq c' E_{\boldsymbol{\theta}_0} \left| \frac{1}{\sigma_m} \sum_{k=1}^{m} u_k \right| \to 0$$

Finally, since u_k is independent of $z_m - z_{m,k}$, $E_{\boldsymbol{\theta}_0}|\mathbf{Q}_3| = E_{\boldsymbol{\theta}_0} \left| \frac{1}{\sigma_m} u_k \right| E_{\boldsymbol{\theta}_0} \left| \exp\left(i\lambda(z_m - z_{m,k})\right) \right| = 0$. □

A.2 Asymptotic Efficiency Proofs

Proof Sketch of Proposition 6

We consider the autocovariance of the joint vector of likelihood and pseudolikelihood gradients $\mathbf{h} = \begin{pmatrix} \mathbf{f} \\ \tilde{\mathbf{f}} \end{pmatrix}$:

$$E\left[\mathrm{Var}\left[\mathbf{h}|\mathcal{S}, \mathbf{x}\right]\right] \tag{17}$$
$$= \begin{pmatrix} E[\mathrm{Var}[\mathbf{f}|\mathbf{x}]] & E[\mathrm{Cov}[\mathbf{f}, \tilde{\mathbf{f}}|\mathcal{S}, \mathbf{x}]] \\ E[\mathrm{Cov}[\tilde{\mathbf{f}}, \mathbf{f}|\mathcal{S}, \mathbf{x}]] & E[\mathrm{Var}[\tilde{\mathbf{f}}|\mathcal{S}, \mathbf{x}]] \end{pmatrix}$$

Since the above covariance matrix is positive semidefinite, we have:

$$det\left(E[\mathrm{Var}[\mathbf{f}|\mathcal{S}, \mathbf{x}]]\right) \geq \frac{det\left(E[\mathrm{Cov}[\mathbf{f}, \tilde{\mathbf{f}}|\mathcal{S}, \mathbf{x}]E[\mathrm{Cov}[\tilde{\mathbf{f}}, \mathbf{f}|\mathcal{S}, \mathbf{x}]\right)}{det\left(E[\mathrm{Var}[\tilde{\mathbf{f}}|\mathcal{S}, \mathbf{x}]]\right)} \tag{18}$$

Also recognize that:

$$\mathrm{Var}[\mathbf{f}|\mathbf{x}] = \frac{1}{n^2} \sum_{C_1, C_2} \mathrm{Cov}(\boldsymbol{\phi}^{C_1}, \boldsymbol{\phi}^{C_2}|\mathbf{x}) = \frac{1}{n} V[\boldsymbol{\phi}|\mathbf{x}] \tag{19}$$

$$\mathrm{Var}[\tilde{\mathbf{f}}|\mathcal{S}, \mathbf{x}] = \frac{1}{m^2} \mathrm{Var}\left[\sum_{S \in \mathcal{S}} \sum_{C: C \cap S \neq \emptyset} \mathbf{u}^C \right]$$
$$= \frac{1}{m^2} \sum_{C_1, C_2: C_1 \bowtie C_2} \mathrm{Cov}\left(\mathbf{u}^{C_1}, \mathbf{u}^{C_2}\right) = \frac{1}{m} C[\phi|x] \tag{20}$$

$$\mathrm{Cov}[\mathbf{f}, \tilde{\mathbf{f}}|\mathcal{S}, \mathbf{x}] = \mathrm{Cov}[\tilde{\mathbf{f}}, \mathbf{f}|\mathcal{S}, \mathbf{x}]$$
$$= \frac{1}{nm} \sum_{C_1 \in \mathcal{C}} \sum_{S \in \mathcal{S}} \sum_{C_2: C_2 \cap S \neq \emptyset} \mathrm{Cov}\left[\boldsymbol{\phi}^{C_1}, \mathbf{u}^{C_2}\right]$$
$$= \frac{1}{nm} \sum_{S \in \mathcal{S}} \sum_{C_1: C_1 \cap S \neq \emptyset} \sum_{C_2: C_2 \cap S \neq \emptyset} \mathrm{Cov}\left[\mathbf{u}^{C_1}, \mathbf{u}^{C_2}\right]$$
$$= \frac{1}{n} \tilde{V}[\boldsymbol{\phi}|\mathcal{S}, \mathbf{x}] \tag{21}$$

Plugging Equations (19)-(21) into (18) and rearranging factors, we obtain (14). □

Proof Sketch of Proposition 7

Again consider the covariance matrix of joint feature vector as in Equation (17), by the law of total covariance, we have $\mathrm{Var}[\mathbf{h}_1|\mathcal{S}_1, \mathbf{x}] \leq \mathrm{Var}[\mathbf{h}_2|\mathcal{S}_2, \mathbf{x}]$. Using a block matrix representation, we have:

$$\frac{det\left(E\left[\mathrm{Cov}[\mathbf{f}, \tilde{\mathbf{f}}|\mathcal{S}_1, \mathbf{x}]\right] E\left[\mathrm{Cov}[\tilde{\mathbf{f}}, \mathbf{f}|\mathcal{S}_1, \mathbf{x}]\right]\right)}{det\left(E\left[\mathrm{Var}[\tilde{\mathbf{f}}|\mathcal{S}_1, \mathbf{x}]\right]\right)}$$
$$\leq \frac{det\left(E\left[\mathrm{Cov}[\mathbf{f}, \tilde{\mathbf{f}}|\mathcal{S}_2, \mathbf{x}]\right] E\left[\mathrm{Cov}[\tilde{\mathbf{f}}, \mathbf{f}|\mathcal{S}_2, \mathbf{x}]\right]\right)}{det\left(E\left[\mathrm{Var}[\tilde{\mathbf{f}}|\mathcal{S}_2, \mathbf{x}]\right]\right)}$$

By applying Equation (20) and (21) and rearranging factors, we get (15). □

A.3 Dataset Information

Synthetic Data Generation

For the synthetic data experiments we used an Erdos-Renyi random graph model to generate synthetic network structures and then generated a binary attribute and class label on the nodes use Markov network model specified below. In the Erdos-Renyi model, we generate a network of N nodes by including an edge between each $N(N-1)/2$ pair of nodes independently at random with probability $p = \frac{d_{avg}}{N(N-1)/2}$. In this work, we considered $d_{avg} = 5$ for the low linkage setting and $d_{avg} = 10$ for the high linkage setting. In the Markov network, we specified two clique types: a singleton clique defined on each node: $\Phi_{y=k} = \exp\{\theta_k I(x = k)\}$ for $k = 1, 2$, and an interaction clique defined on each edge (i, j): $\Phi_{y_i=y_j} = \exp\theta_3$. The attributes $(\mathbf{x})$ and labels $(\mathbf{y})$ are then generated by Gibbs sampling of 1000000 iterations. For the low interaction setting we used $\theta_3 = 0.1$ and for the high interaction setting we use $\theta_3 = 0.3$. The singleton clique parameters were constant in all experiments at $\theta_1 = 1.2, \theta_2 = 0.8$.

Facebook Data and Model

The Facebook dataset consists of 7,315 nodes and 46,817 edges, which comprises all the students and alumni, with public profiles, from a large "University" network. The data includes user profile attributes (gender, relationship status, political view and religious view), as well as friendship links and transactions (wall posting and picture tagging) between users. We divide the network into 8 articulated subnetworks of comparable sizes for training and testing (e.g., "Class of 2008").

For classification, we build a pairwise Markov network based on friendship links. Two types of cliques are specified: the singleton clique is a linear weighting of node attributes: $\Phi_{y=k} = \exp\left\{\sum_{j=1}^{m} \theta_{jk}^{node} I(x = k)\right\}$, and the edgewise clique is a linear weighting of transactions between users i and j: $\Phi_{y_i=y_j=k} = \exp\left\{\sum_{l=1}^{t} \theta_{lk}^{edge} s_k(i, j) + \theta_{0k}^{edge}\right\}$ where $s_l(i, j)$ denotes the strength of the transaction l. We perform classification tasks based on the four aforementioned profile attributes. When classifying each attribute y, the model is conditioned on the remaining 3 attributes x_1, x_2, x_3. The two aforementioned transactions are applied in the model, and the strength $s_l(i, j)$ is computed as the logarithm of the total count of transaction type l occurred between i and j.