Hierarchical Infinite Relational Model

Feras A. Saad[1] **Vikash K. Mansinghka**[1]

[1] Massachusetts Institute of Technology
Cambridge, MA, USA

Abstract

This paper describes the hierarchical infinite relational model (HIRM), a new probabilistic generative model for noisy, sparse, and heterogeneous relational data. Given a set of relations defined over a collection of domains, the model first infers multiple non-overlapping clusters of relations using a top-level Chinese restaurant process. Within each cluster of relations, a Dirichlet process mixture is then used to partition the domain entities and model the probability distribution of relation values. The HIRM generalizes the standard infinite relational model and can be used for a variety of data analysis tasks including dependence detection, clustering, and density estimation. We present new algorithms for fully Bayesian posterior inference via Gibbs sampling. We illustrate the efficacy of the method on a density estimation benchmark of twenty object-attribute datasets with up to 18 million cells and use it to discover relational structure in real-world datasets from politics and genomics.

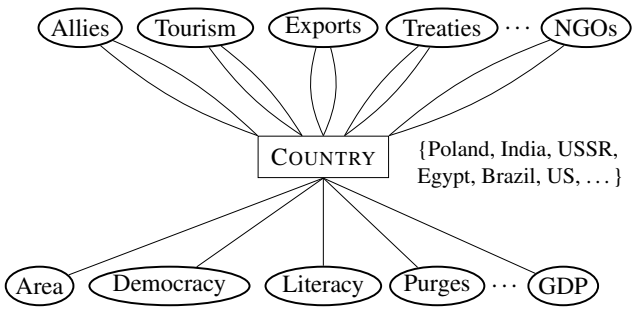
1 INTRODUCTION

Learning models for relational data is a widely studied problem that arises in a number of settings such as business intelligence (Chaudhuri et al., 2011), social networks (Carrington et al., 2005), bioinformatics (Rual et al., 2005), and recommendation systems (Su and Khoshgoftaar, 2009), amongst many others (Džeroski and Lavrač, 2001). In this setting, we observe attributes and interactions among a set of entities and our goal is to learn models that are useful for explaining or making predictions about the entities, their attributes, and/or their interactions. Fig. 1 shows two examples of relational systems for political and genomics data. For politics (Fig. 1a), one problem could be to discover what attributes of a particular country and interactions with other countries are likely to make it an attractive tourist hub. In genomics (Fig. 1b), our goal might be to predict what complexes a particular gene is likely to form, given information about its motifs, functions, and interactions with other genes. This paper addresses the problem of automatically learning probabilistic models for a variety of relational systems given a dataset of noisy and possibly sparse observations.

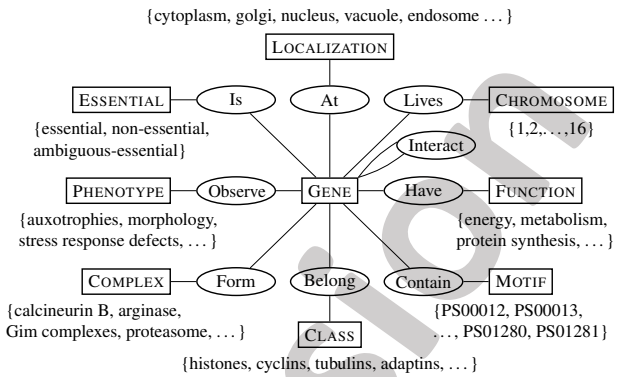
Learning probabilistic structure is an exceptionally difficult task (Daly et al., 2011). One approach to simplifying the learning problem is to posit a collection of hidden variables that both explain and decouple the relationships between observed variables. Using Bayesian nonparametrics, both the values and dimensionality of these hidden variables can be automatically inferred from data. This approach is commonly used for modeling relational data (Kemp et al., 2006; Xu et al., 2006; Roy and Teh, 2009; Sutskever et al., 2009; Kim et al., 2013; Nakano et al., 2014; Xuan et al., 2017; Fan et al., 2018): refer to Fan et al. (2020) for a recent survey on developments in the field. Our paper builds on the infinite relational model (IRM; Kemp et al., 2006; Xu et al., 2006), a widely used and flexible Bayesian nonparametric method that applies to a variety relational systems. The IRM is a cluster-based model: informally, to decide whether a binary relation R holds between a pair of entities i and j, the IRM flips a coin whose weight depends on the (latent) cluster assignments of i and j. A strength of the IRM, which we review in Sec. 2, is its ability to extract meaningful partitions from observational data. However, as we identify in Sec. 3, two limitations inherent to the IRM's inductive bias make the model (i) susceptible to combinatorial over-clustering; and (ii) fail to discover certain predictive structure between dependent but non-identically distributed relations, which can both result in an inaccurate overall model of the data.

To address these limitations, this paper introduces the hierarchical infinite relational model (HIRM) in Sec. 4, a new method that combines the flexibility of the IRM with a structure learning prior that infers subsets of relations that are probably independent of one another. By allowing different relations to be explained by different partitions, the HIRM

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).



(a) Relational system for political data (Rummel, 1999)



(b) Relational system for genomics data (Cheng et al., 2002)

Figure 1: Two relational systems that we analyze using the HIRM in this paper. Domains are in boxes, relations in ellipses, and domain entities between curly braces. (a) One domain, five unary relations, and five binary relations. (b) Nine domains and nine binary relations. Unary relations represent "attributes" while binary relations and higher represent "interactions".

specifies a large hypothesis space that includes the standard IRM in addition to compact models of the data that can only be approximated by an IRM using a combinatorially large number of clusters. The evaluations in Sec. 5 show that the HIRM makes more accurate predictions and discovers more fine-grained clustering structure as compared to the IRM, while retaining a flexible framework for automatic Bayesian structure discovery in a variety of relational systems.

2 INFINITE RELATIONAL MODEL

We begin with a review of the IRM, using a slightly more general definition of "relations" than was originally described in Kemp et al. (2006) or Xu et al. (2006).

Definition 2.1. A *relational system* S consists of n domains $D_1, \ldots, D_n$ and m relations $R_1, \ldots, R_m$. Each *domain* D_i $(1 \leq i \leq n)$ is a countably infinite set of distinct *entities* $\{e_1^i, e_2^i, \ldots\}$. Each *relation* R_k $(1 \leq k \leq m)$ is a map from the Cartesian product of t_k domains to an arbitrary codomain C_k. The symbol d_{ki} $(1 \leq k \leq m, 1 \leq i \leq t_k)$ denotes the domain index of the i-th argument of R_k.

Example 2.2. Suppose system S has $n = 4$ domains D_1, D_2, D_3, D_4, and $m = 3$ relations R_1, R_2, R_3; with

$$R_1 : D_1 \times D_1 \to \{0, 1\},$$
$$R_2 : D_1 \times D_3 \times D_4 \to \{1, 2, \ldots\},$$
$$R_3 : D_2 \to (-\infty, \infty).$$

In this system, we have

$$t_1 = 2; \quad d_{11} = 1, d_{12} = 1; \qquad C_1 = \{0, 1\};$$
$$t_2 = 3; \quad d_{21} = 1, d_{22} = 3, d_{23} = 4; \qquad C_2 = \{1, 2, \ldots\};$$
$$t_3 = 1; \quad d_{31} = 2; \qquad C_3 = (-\infty, \infty).$$

R_1 is a binary relation taking binary values, R_2 is a ternary relation taking positive integer values, and R_3 is a unary relation taking real values.

Remark 2.3. To simplify notation, for a given relation $R : D_1 \times \cdots \times D_n \to C$ and entity indexes $i_1, \ldots, i_n \in \mathbb{N}$, we will write $R(i_1, \ldots, i_n)$ to mean $R(e_{i_1}^1, \ldots, e_{i_n}^n)$.

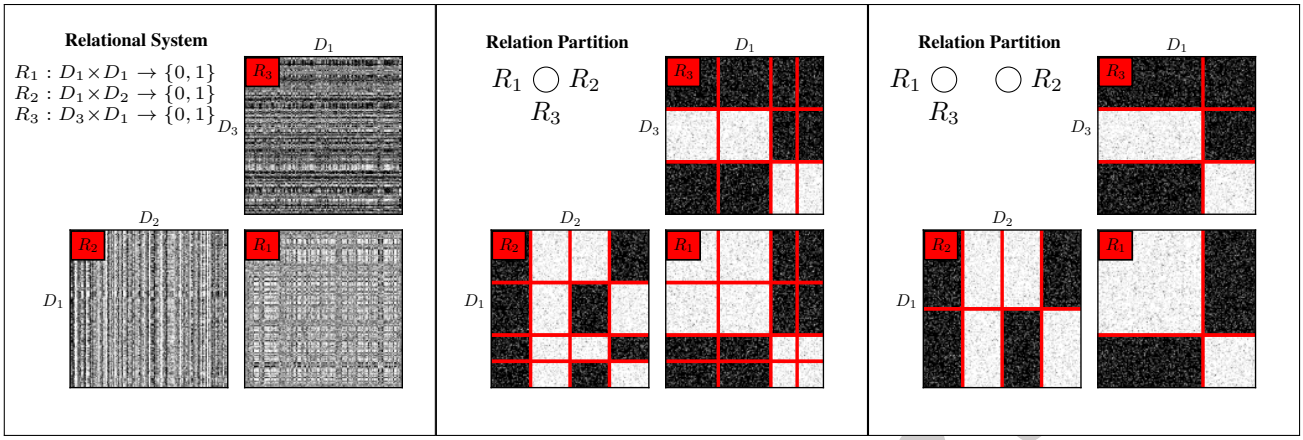
Consider a system S with n domains and m relations. For each $i = 1, \ldots, n$, the IRM assumes that entities $\{e_1^i, e_2^i, \ldots\}$ in domain D_i are associated with integer cluster assignments $\{z_1^i, z_2^i, \ldots\} =: z^i$. The IRM defines a joint probability distribution over cluster assignments and relation values with the following factorization structure:

$$P(z^1, \ldots, z^n, R_1, \ldots, R_m) \tag{1}$$
$$= \prod_{i=1}^{n} P(z^i) \prod_{k=1}^{m} P(R_k \mid z^1, \ldots, z^n).$$

To allow the IRM to discover an arbitrary number of clusters for each domain D_i, the cluster assignments z^i for the entities are given a nonparametric prior that assigns a positive probability to all possible partitions using the Chinese restaurant process (CRP; Aldous, 1985). For each $i = 1, \ldots, n$, the cluster assignment probabilities $P(z^i) = P(z_1^i, z_2^i, \ldots)$ in Eq. (1) are defined inductively with $z_1^i := 1$, and for $l \geq 2$

$$P(z_l^i = j \mid z_1^i, \ldots, z_{l-1}^i) \propto \begin{cases} n_j & \text{if } 1 \leq j \leq M \\ \gamma & \text{if } j = M + 1, \end{cases} \tag{2}$$

where $n_j := \sum_{c=1}^{l-1} \mathbf{1}[z_c^i = j]$ is the number of previous entities at cluster j; $M := \max\{z_1^i, \ldots, z_{l-1}^i\}$ is the number of clusters among the first $l - 1$ entities; and $\gamma > 0$ is a concentration parameter. The cluster assignment vectors $z^1, \ldots, z^n$ across the n domains are mutually independent, each drawn from a CRP (Eq. (2)). Next, for each relation R_k $(1 \leq k \leq m)$, a set of parameters $\theta_k(j_1, \ldots, j_{t_k})$ is used to dictate the distribution of $R_k(i_1, \ldots, i_{t_k})$, where $j_1, \ldots, j_{t_k}, i_1, \ldots, i_{t_k} \in \mathbb{N}$. The value of a relation depends only the cluster assignments, i.e., $R_k(i_1, \ldots, i_{t_k})$ and $R_k(i'_1, \ldots, i'_{t_k})$ share the same parameter whenever



(a) Relational System and Dataset (b) IRM (No Structural Independences) (c) HIRM ($(R_1, R_3) \perp R_2$)

Figure 2: Comparing posterior inferences from the standard IRM (Kemp et al., 2006) and HIRM. (a) Three relation signatures and an observed dataset. (b) The IRM forces all relations to be dependent and learns, for each domain, a single clustering (vertical/horizontal red lines) that is identical across all the relations. This conservative assumption by the IRM leads to an over-clustering of the domain D_1 that participates in all three relations, which in turn creates combinatorially many spurious clusters than are required to explain the data. (c) The HIRM finds sufficient evidence for the probable independence of R_2 with (R_1, R_3) and learns separate a domain clustering for D_1 within each of the two blocks of the relation partition, leading to a more succinct explanation of the data. While the clusterings in both (b) and (c) are in the HIRM hypothesis space, the clustering in (c) is $\approx 7{\times}10^{13}$ times more likely under the posterior, indicating that the independence structure and domain clusterings inferred by the HIRM is substantially more likely than the full dependence structure imposed by the IRM in (b).

$z^{d_{kl}}_{i_l} = z^{d_{kl}}_{i'_l}$ for each $l = 1, \ldots, t_k$. Thus, for domain index $i = 1, \ldots, n$; relation index $k = 1, \ldots, m$; entity indexes $i_1, \ldots, i_{t_k} \in \mathbb{N}$; and cluster indexes $j_1, \ldots, j_{t_k} \in \mathbb{N}$, the generative model of the IRM is given by:

$$\{z^i_1, z^i_2, \ldots\} \sim \text{CRP}(\gamma_i) \tag{3}$$

$$\theta_k(j_1, \ldots, j_{t_k}) \sim \pi_k(\lambda_k) \tag{4}$$

$$R_k(i_1, \ldots, i_{t_k}) \sim L_k(\theta_k(z^{d_{k1}}_{i_1}, \ldots, z^{d_{kt_k}}_{i_{t_k}})), \tag{5}$$

where $(\{\gamma_i\}_{i=1}^n, \{\lambda_k\}_{k=1}^m)$ are model hyperparameters. Eq. (5) ensures items within a cluster are generated by the same parameter. The prior π_k and likelihood L_k distributions in Eqs. (4) and (5) can be set depending on the codomain C_k of R_k (e.g., beta-Bernoulli for binary data, gamma-Poisson for counts, chisquare-normal for real values, etc.). Kemp et al. (2006) used the IRM to discover structure in a variety of real-world relational systems that appear quite different on the surface, including:

(a) Random graphs, with one domain D for the vertices and one relation $R : D \times D \to \{0, 1\}$ for the edges.

(b) Object-attribute data, with one relation $R : D_1 \times D_2 \to \{0, 1\}$, where $R(i, j) = 1$ iff item e^1_i has attribute e^2_j.

(c) Systems with multiple attributes and interactions, where, for example, D_1 are countries, D_2 are attributes; and D_3 are interactions; so that $R_1 : D_1 \times D_2 \to \{0, 1\}$ models attributes and $R_2 : D_1 \times D_1 \times D_3 \to \{0, 1\}$ models interactions, where $R_2(i, j, k) = 1$ iff countries e^1_i and e^1_j perform interaction e^3_k.

3 LIMITATIONS OF THE IRM

We next describe two limitations in the standard IRM that arise when using the model in practice, motivating the hierarchical structure learning prior that we introduce in Sec. 4.

3.1 ENFORCING SHARED DOMAIN CLUSTERINGS LEADS TO OVERFITTING

A key assumption of the IRM is that each domain D_i has a single clustering $z^i = \{z^i_1, z^i_2, \ldots\}$ that globally dictates the partition of its entities $\{e^i_1, e^i_2, \ldots\}$. The same cluster assignments z^i are used for all of relations $R_1, \ldots, R_m$ in which D_i participates, which can lead to substantial over-clustering and a failure to accurately model data in the presence of structural independences between relations. Fig. 2 illustrates and discusses this limitation in further detail.

3.2 RESTRICTIONS WHEN CLUSTERING MULTIPLE RELATIONS

Kemp et al. (2006) applied the IRM to clustering multiple *relations*, by treating the relations themselves as entities within a new domain. More specifically, for a system with relations $R_1, \ldots, R_m$, all defined on same domain and codomain (say D and C), the key idea is to encode the system using one higher-order relation $R' : D' \times D \to C$, where the entities of D' are relations over D, i.e., $R'(j, i) := R_j(i)$ (for $1 \leq j \leq m$, $i \in D$). While an IRM for R' will learn a

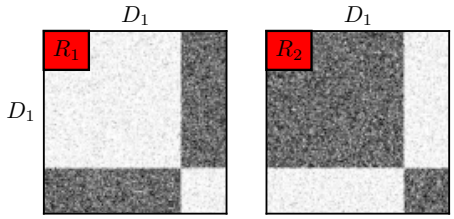


Figure 3: When used to cluster relations, the standard IRM (Kemp et al., 2006) uses a higher-ordering encoding that requires relations to be defined on the same domain and assumes all relations within a cluster are identically distributed. While R_1 and R_2 use identical partitions of D_1 and are anti-correlated, they are not identically distributed and are thus assigned different clusters by the IRM. In contrast, the HIRM can learn clusters of relations defined on different domains and can assign non-identically distributed relations to the same cluster (Fig. 5b shows a real-world example).

clustering of both D' (the relations) and D, there are at least two restrictions with this approach: (i) it only applies to relations defined on identical domains and codomains; and (ii) it clusters relations R_i and R_j together only if they are both dependent and identically distributed (Eq. (5)). Fig. 3 illustrates and discusses this limitation in further detail.

4 HIERARCHICAL INFINITE RELATIONAL MODEL

We now present the HIRM, which addresses the aforesaid limitations of the IRM by using a structure learning prior to infer probable independences among relations that cannot be represented structurally in a standard IRM.

Given a system S with domains $D_1, \ldots, D_n$ and relations $R_1, \ldots, R_m$, the HIRM first nonparametrically partitions the m *relations* using a CRP (Eq. (2)), where the cluster assignments of the relations are denoted by $y := \{y_1, \ldots, y_m\}$. This partition induces a random number $K := \max\{y_1, \ldots, y_m\}$ of subsystems $S_1, \ldots, S_K$ of S. For each $\ell = 1, \ldots, K$, the relations $\{R_i \mid 1 \leq i \leq m, y_i = \ell\}$ assigned to subsystem S_ℓ are modeled jointly by an IRM (Eqs. (3)–(5)), independently of all relations assigned to another subsystem $S_{\ell'}$ ($\ell' \neq \ell$). The HIRM thus defines a probability distribution over relation clusters, domain entity clusters, and relation values with the following factorization:

$$P(y_1, \ldots, y_m, \{z^{\ell 1}, \ldots, z^{\ell n}\}_{\ell=1}^{K}, R_1, \ldots, R_m) \tag{6}$$
$$= P(y) \prod_{\ell=1}^{K} \prod_{i=1}^{n} P(z^{\ell i}) \prod_{k \mid y_k = \ell} P(R_k \mid z^{\ell 1}, \ldots, z^{\ell n}).$$

For each subsystem index $\ell = 1, \ldots, K$; domain index $i = 1, \ldots, n$; relation index $k = 1, \ldots, m$; entity indexes $i_1, \ldots, i_{t_k}$; and cluster indexes $j_1, \ldots, j_{t_k}$, the generative specification of the HIRM is given by the following process:

$$\{y_1, \ldots, y_m\} \sim \text{CRP}(\gamma_0) \tag{7}$$
$$\{z_1^{\ell i}, z_2^{\ell i}, \ldots\} \sim \text{CRP}(\gamma_{\ell i}) \tag{8}$$
$$\theta_k(j_1, \ldots, j_{t_k}) \sim \pi_k(\lambda_k) \tag{9}$$
$$R_k(i_1, \ldots, i_{t_k}) \sim L_k(\theta_k(z_{i_1}^{y_k, d_{k1}}, \ldots, z_{i_{t_k}}^{y_k, d_{k t_k}})), \tag{10}$$

where $(\gamma_0, \{\{\gamma_{\ell i}\}_{i=1}^{n}\}_{\ell=1}^{K}, \{\lambda_k\}_{k=1}^{m})$ are model hyperparameters, possibly endowed with their own hyperpriors.

The HIRM generalizes and extends the IRM. First, it recovers the standard IRM when $\gamma_0 = 0$. For $\gamma_0 > 0$, Eq. (7) specifies a CRP partition prior over relations, where relations in the same block are modeled jointly using a standard IRM (Eqs. (8)–(10)). In Eq. (8), each domain D_i is associated with a different partition $z^{\ell i}$ for each subsystem S_ℓ in which it participates. This inductive bias allows the HIRM to express structural independences between relations and avoid modeling a Cartesian product of domain partitions when the data for (a subset of) relations in the system are not well-aligned (Sec. 3.1 and Fig. 2).

Additionally, Eq. (7) allows the HIRM to directly cluster dependent relations together, without using higher-order encodings that are limited to relations defined on the same domain as in the IRM (Sec. 3.2). Further, Eqs. (9) and (10) imply that relations R_k and $R_{k'}$ that are clustered together in a subsystem S_ℓ need not be identically distributed (resp. Fig. 3), as they each have their own parameters θ_k and $\theta_{k'}$, respectively. The dependence is instead modeled by the shared domain partitions $\{z^{\ell 1}, \ldots, z^{\ell n}\}$ within subsystem S_ℓ. In sum, the nonparametric structure learning prior Eq. (7) retains the benefits of the standard IRM while addressing the limitations discussed in Sec. 3, all within a Bayesian nonparametric model discovery framework.

4.1 POSTERIOR INFERENCE

An observed dataset $\{r_1, \ldots, r_m\}$ for a relational system consists of a finite number of realizations of relation values, i.e., observations of random variables of the form $\{R_k(i_1, \ldots, i_{t_k}) = r_k(i_1, \ldots, i_{t_k})\}$. For notational ease and without loss of generality, we assume that the relation values are fully observed for $N_i \geq 1$ entities $\{e_1^i, \ldots, e_{N_i}^i\}$ of each domain D_i ($i = 1, \ldots, n$), across all relations that it participates in (our reference implementations of the HIRM handles arbitrary index combinations with missing data).

Posterior inference in the HIRM is carried out by simulating an ergodic Markov chain that converges to the distribution obtained by conditioning Eq. (6) on the observed dataset. The chain initializes a state $\mathcal{S}$ by sampling it from the prior (Eqs. (7)–(9)) and iterates the state using Gibbs sampling. Algorithm 1 shows one full Gibbs scan through all the variables in the state $\mathcal{S}$. We next describe transition operators for the updates in lines 2, 8 and 15 of Algorithm 1.

Algorithm 1 MCMC Gibbs Scan for HIRM (Sketch)

Require: Markov chain state $\mathcal{S}$ containing relation cluster assignments $\{y_1, \ldots, y_m\}$, entity cluster assignments $\{z_1^{\ell i}, \ldots, z_{N_i}^{\ell i}\}$, and parameters $\{\theta_k(j_1, \ldots, j_{t_k})\}$, for $1 \le i \le m$, $1 \le \ell \le K$, and $1 \le k \le m$; dataset r.

for $k = 1, \ldots, m$ **do**
  resample y_k given $(r, \mathcal{S} \setminus \{y_k\})$
for $\ell = 1, \ldots, \max(y_1, \ldots, y_m)$ **do**
  $\triangleright$ I_ℓ is set of domains in subsystem S_ℓ
  $I_\ell \leftarrow \{d_{kj} \mid 1 \le k \le m, 1 \le j \le t_k, y_k = \ell\}$
  for $i \in I_\ell$ **do**
    for $j = 1, \ldots, N_i$ **do**
      resample $z_j^{\ell i}$ given $(r, \mathcal{S} \setminus \{z_j^{\ell,i}\})$
  $\triangleright$ T_ℓ is set of relations in subsystem S_ℓ
  $T_\ell \leftarrow \{k \mid 1 \le k \le m, y_k = \ell, (\pi_k, L_k)$ nonconjugate$\}$
  for $k \in T_\ell$ **do**
    for $j_1 = 1, \ldots, \max(z_1^{\ell d_{k1}}, \ldots, z_{N_{d_{k1}}}^{\ell d_{k1}})$ **do**
      ...
        for $j_{t_k} = 1, \ldots, \max(z_1^{\ell d_{kt_k}}, \ldots, z_{N_{d_{kt_k}}}^{\ell d_{kt_k}})$ **do**
          resample $\theta_k(j_1, \ldots, j_{t_k})$ given $(r, \mathcal{S})$

Reampling relation cluster assignments y_k : This kernel uses the auxiliary Gibbs sampler (Neal, 2000, Algorithm 8). Let $C_\ell := |\{k \mid 1 \le k \le K, y_k = \ell\}|$ be the number of relations in S_ℓ and $W_{\ell i} := \max\{z_1^{\ell i}, \ldots, z_{N_i}^{\ell i}\}$ be the number of clusters for domain D_i within S_ℓ $(1 \le \ell \le K)$.

Case 1: If y_k is a singleton $(C_k = 1)$, then it is resampled to take a new value $\ell \in \{1, \ldots, K\}$ with probability

$$c_{k\ell} \prod_{j_1=1}^{W_{\ell d_{k1}}} \cdots \prod_{j_{t_k}=1}^{W_{\ell d_{kt_k}}} w_{k\ell}(\mathbf{j}, \theta_k), \tag{11}$$

where $\mathbf{j} := (j_1, \ldots, j_{t_k})$ and

$$c_{k\ell} := \begin{cases} \gamma_0/(m-1+\gamma_0) & \text{if } \ell = y_k \\ C_\ell/(m-1+\gamma_0) & \text{otherwise,} \end{cases} \tag{12}$$

$$w_{k\ell}(\mathbf{j}, \theta_k) := \prod_{\mathbf{i} \in A_{k\ell}(\mathbf{j})} L_k(r_k(\mathbf{i}); \theta_k(\mathbf{j})). \tag{13}$$

Eq. (12) is the conditional probability from the CRP prior (Eq. (2)), and in Eq. (13) the symbol

$$A_{k\ell}(\mathbf{j}) := \{\mathbf{i} \mid z_{i_1}^{\ell d_{k1}} = j_1, \ldots, z_{i_{t_k}}^{\ell d_{kt_k}} = j_{t_k}\} \tag{14}$$

denotes the set of entity indexes $\mathbf{i} := (i_1, \ldots, i_{t_k})$ for domains $(d_{k1}, \ldots, d_{kt_k})$ that are assigned to cluster $\mathbf{j}$ of subsystem S_ℓ (where $1 \le k \le m$; $1 \le \ell \le M$; $1 \le j_1 \le W_{\ell d_{k1}}$; ...; $1 \le j_{t_k} \le W_{\ell d_{kt_k}}$). Note that if (π_k, L_k) is a conjugate pair, the parameters θ_k can be analytically integrated out, and Eq. (13) becomes

$$w_{k\ell}(\mathbf{j}) := \int_\theta \Big[\prod_{\mathbf{i} \in A_{k\ell}(\mathbf{j})} L_k(r_k(\mathbf{i}); \theta) \Big] \pi_k(\theta; \lambda_k) d\theta. \tag{15}$$

Case 2: If y_k is not a singleton $(C_k > 1)$, then

1. For domain indexes $i = 1, \ldots, n$, draw cluster assignments for a fresh entity partition, i.e.,

$$\{z_1^{K+1,i}, \ldots, z_{N_i}^{K+1,i}\} \sim \text{CRP}(\gamma), \tag{16}$$

$$W_{K+1,i} := \max\{z_1^{K+1,i}, \ldots z_{N_i}^{K+1,s}\}. \tag{17}$$

2. Draw parameters $\theta_k(j_1, \ldots, j_{t_k})$ for relation indexes $k = 1, \ldots, m$ and cluster indexes $j_1 = 1, \ldots, W_{K+1,d_1}$; ...; $j_{t_k} = 1, \ldots, W_{K+1,d_{kt_k}}$.

Next, resample y_k to take a new value $\ell \in \{1, \ldots, K+1\}$ using the same terms in Eqs. (11)–(13) from the previous case, except that the CRP weight $c_{k\ell}$ in Eq. (12) is instead

$$c_{k\ell} := \begin{cases} (C_\ell - 1)/(m-1+\gamma_0) & \text{if } \ell = y_k \\ C_\ell/(m-1+\gamma_0) & \text{if } \ell \ne y_k, \ell \le K \\ \gamma_0/(m-1+\gamma_0) & \text{if } \ell = K+1. \end{cases} \tag{18}$$

Resampling entity cluster assignments $z_j^{\ell i}$: Within each subsystem S_ℓ, the entity cluster assignments are transitioned using the collapsed Gibbs sampler (Neal, 2000, Alg. 3). Alternatively, the split-merge algorithm can be used (Jain and Neal, 2004). Xu et al. (2007) discuss additional sampling-based and variational approaches for these variables.

Resampling cluster parameters $\theta_k(j_1, \ldots, j_{t_k})$: Sample $\theta_k'(\mathbf{j}) \sim q_k(\theta_k(\mathbf{j}))$ from a proposal distribution (e.g., the prior $\pi_k(\lambda_k)$ or Gaussian drift $\mathcal{N}(\theta_k(\mathbf{j}), \sigma_k)$) and accept the move according to the Metropolis-Hastings probability

$$\min\left(1, \frac{\pi_k(\theta_k'(\mathbf{j}); \lambda_k) w_{k\ell}(\mathbf{j}, \theta_k') q_k(\theta_k(\mathbf{j}); \theta_k'(\mathbf{j}))}{\pi_k(\theta_k(\mathbf{j}); \lambda_k) w_{k\ell}(\mathbf{j}, \theta_k) q_k(\theta_k'(\mathbf{j}); \theta_k(\mathbf{j}))}\right). \tag{19}$$

where is $w_{k\ell}$ (Eq. (13)) is the data likelihood for cluster $\mathbf{j}$.

Resampling hyperparameters: Broad exponential hyperpriors are used for all the model hyperparameters $\gamma_0, \{\gamma_{\ell i}\}, \{\lambda_k\}$ that appear in Eqs. (7)–(9), which are resampled using gridded-Gibbs (Ritter and Tanner, 1992). It is also possible to instead use slice sampling (Neal, 2003).

5 EVALUATION

We implemented a prototype of the HIRM[1] and evaluated it in three settings: solving density estimation tasks in object-attribute data; discovering relational structure in political data; and learning relationships between gene properties.

5.1 OBJECT-ATTRIBUTE BENCHMARKS

We assessed the predictive performance of the HIRM on a benchmark of 20 object-attribute datasets (Gens and Domingos, 2013) and compared the results to two Bayesian nonparametric baselines. In Table 1, the first four columns

[1]Reference implementations of the HIRM in C++ and Python are available at https://github.com/probcomp/hierarchical-irm.

Table 1: Prediction accuracy of HIRM and Bayesian nonparametric baselines on a benchmark of 20 object-attribute datasets.

	Dataset Statistics			Average Test Log-Likelihood		
Dataset	N_{cols}	$N_{\text{rows}}^{\text{train}}$	$N_{\text{rows}}^{\text{test}}$	HIRM	IRM	DPMM
NLTCS	16	18338	3236	-06.00	-06.01 •	-06.01 •
MSNBC	17	330212	58265	-06.19	-06.27 •	-06.22 •
KDDCup 2000	64	199999	34955	-02.13	-02.13 •	-02.13 •
Plants	69	19733	3482	-13.75	-14.23 •	-13.81
Audio	100	17000	3000	-39.99	-40.34	-40.02
Jester	100	10000	4116	-52.91	-52.96	-52.92
Netflix	100	3500	3000	-56.96	-57.48 •	-56.96
Accidents	111	14458	2551	-33.85	-39.43 •	-38.93 •
Retail	135	24979	4408	-10.90	-10.99	-10.92
Pumsb-star	163	13897	2452	-32.77	-38.95 •	-38.02 •
DNA	180	2000	1186	-87.65	-97.44 •	-97.62 •
Kosarek	190	37825	6675	-10.91	-10.99	-10.95
MSWeb	294	62191	5000	-10.23	-11.20 •	-10.26
Book	500	9859	1739	-34.43	-34.52	-34.76
EachMovie	500	5526	591	-52.23	-52.09	-54.86
WebKB	839	3361	838	-156.67	-157.27	-158.26
Reuters-52	889	7560	1540	-90.22	-90.06	-89.34
20 Newsgroup	910	15057	3764	-153.52	-156.46 •	-153.95
BBC	1058	1895	330	-253.36	-253.86	-254.59
Ad	1556	2788	491	-45.19	-46.17	-52.40 •

• indicates significantly worse than HIRM ($p = 0.05$ Mann-Whitney U test).

summarize the dataset statistics (16–1556 columns, 2000–330212 rows). The last three columns show the test log-likelihood from the HIRM, IRM (Kemp et al., 2006; Xu et al., 2006), and Dirichlet process mixture model (DPMM; Lo, 1984). As in Kemp et al. (2006), the IRM encodes object-attribute data using one binary relation $R : \text{Attr} \times \text{Obj} \to \{0, 1\}$. The HIRM encodes each dataset using N_{cols} unary relations $\{R_i : \text{Obj} \to \{0, 1\} \mid i \in \text{Attr}\}$ with structure learning (Eq. (7)) over the dependence between the attributes. The DPMM uses the same encoding as the HIRM but without structure learning (i.e., all attributes are modeled jointly). Dots indicate significantly worse values than the HIRM ($p = 0.05$, Mann–Whitney U test on the $N_{\text{rows}}^{\text{test}}$ predictions from each model). Table 1 shows that the HIRM consistently outperforms these baselines—it is significantly better in 17 cases and worse in zero cases. Fig. 4 shows a plot of runtime vs. held-in data log score for two runs of the HIRM and IRM on four of the benchmarks. Despite using a structure learning prior, the runtime of the HIRM matches or outperforms the IRM; in fact, the HIRM often infers simpler partitions within the independent subsystems, which can improve both the runtime scaling and model fit.

To further assess the density estimation results, we compared the HIRM test log-likelihood to those obtained from probabilistic deep learning baselines for object-attribute data: LearnSPN (Gens and Domingos, 2013) and RAT-SPN (Peharz et al., 2019). Table 2 summarizes the comparison (tie means statistically insignificant differences). The results show that the HIRM, which is a relatively shallow Bayesian model (Eqs. (7)–(10)), is competitive on object-attribute data with higher capacity probabilistic deep learning baselines that fit the data using greedy search. The HIRM is distinguished by being additionally applicable to far more general relational systems, as we next demonstrate.

Table 2: Summary of no. of wins, ties, and losses of HIRM on benchmarks from Table 1, compared to two Bayesian nonparametric baselines and two probabilistic deep learning baselines.

	IRM	DPMM	LearnSPN	RAT-SPN
HIRM # win	11	7	6	4
HIRM # tie	9	13	8	13
HIRM # lose	0	0	6	3

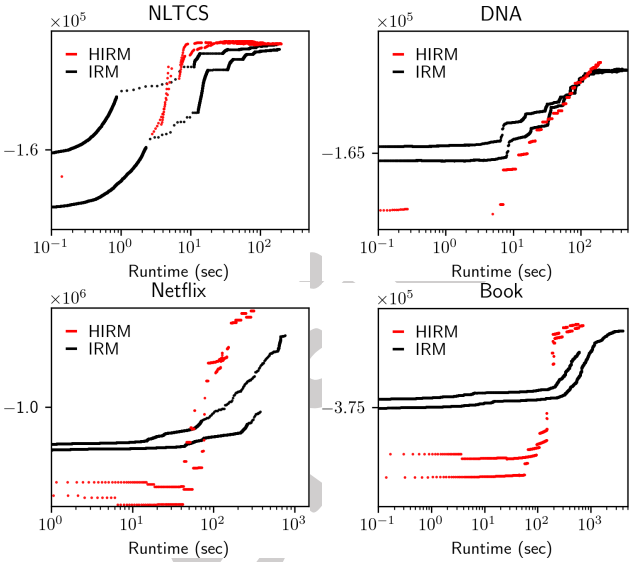


Figure 4: Runtime vs. held-in data log score for HIRM and IRM in four representative benchmarks from Table 1. For each method, two independent runs of inference are plotted.

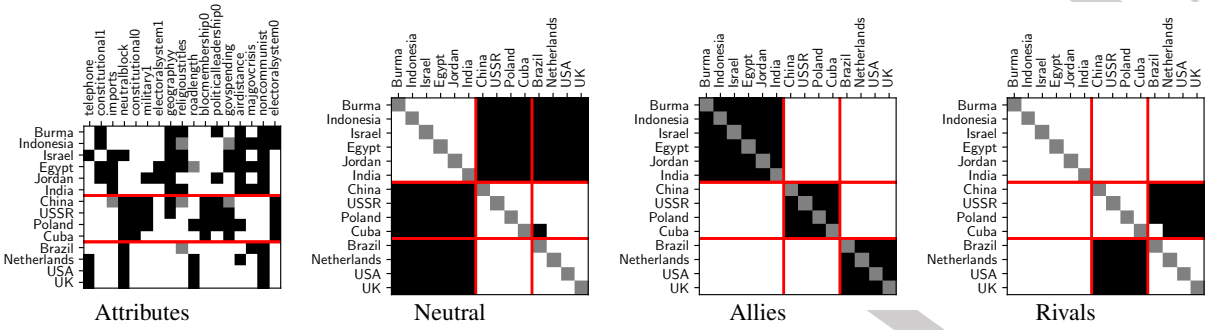
5.2 POLITICAL INTERACTIONS

We next applied the HIRM to the "Dimensionality of Nations" project (Rummel, 1999), using the version dataset from Kemp et al. (2006) for years 1950–1965. Figs. 1a and 5a show a subset of the 15 countries, 111 attributes and 56 interactions. Figs. 5b–5e show a collection of independent subsystems of relations discovered by the HIRM (gray cells indicate missing values). Each inferred subsystem reflects a different partition of the countries that explains the attribute and interactions within the subsytem. For example, in Fig. 5b, the HIRM finds that the geopolitcal bloc interactions are associated with attributes such as "electoral system", "political leadership", and "constitutional".[2] In Fig. 5c, which represents economic and cultural ties and includes attributes such as "absolute income", "agricultural population", and "arts and culture NGOs", the data shows that tourists from the UK and USA travel to countries from all clusters and all countries translate books from the USA

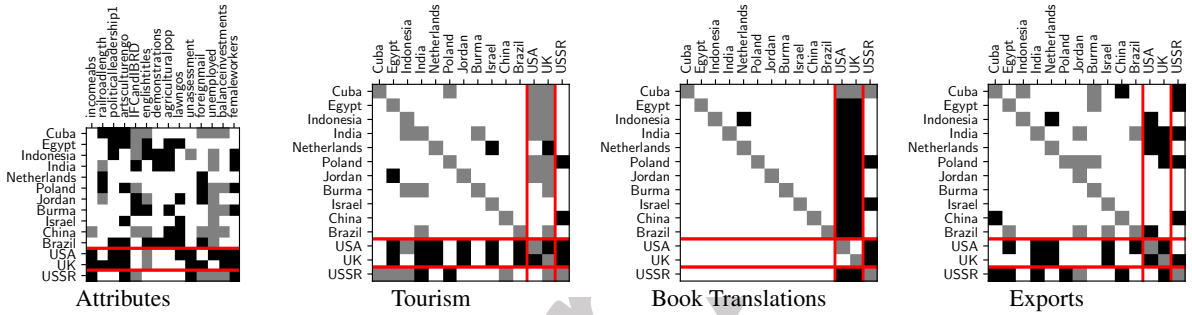
[2] In Fig. 5b, the Cuba–Brazil relationship is neutral despite the countries belonging to rival geopolitcal blocs, which is detected by the HIRM as probabilistically unlikely. This outlier is explained by the so-called the American–Brazilian–Cuban "triangular diplomacy" during the 1962 missile crisis (Hershberg, 2004).

(a) Subset of countries (15 total), attributes (111 total), and interactions (56 total) in the relational system.

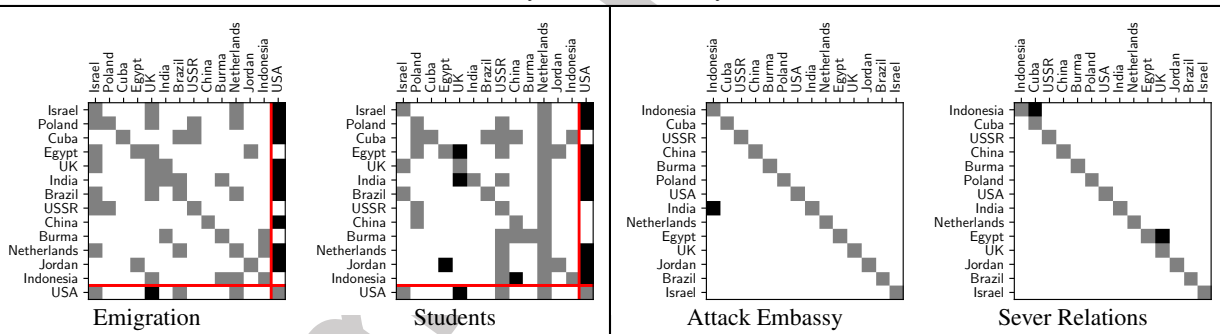
Countries (Domain)	Indonesia, Jordan, Burma, India, Israel, Egypt, Poland, USSR, UK, USA, Brazil
Attributes (Unary Relations)	Area, Telephone Users, Communist, Literacy, Protests, Purges, Democracy, ...
Interactions (Binary Relations)	Exports, Enemies, Allies, Economic Aid, Book Translations, Treaties, Tourism, ...



(b) Inferred Subsystem 1 (Geopolitical Blocs)



(c) Inferred Subsystem 2 (Economy and Culture)



(d) Inferred Subsystem 3 (USA Outlier)

(e) Inferred Subsystem 4 (Sparse Interactions)

Figure 5: Systems of concepts inferred by the HIRM on the "Dimensionality of Nations" data (schema in Fig. 1a).

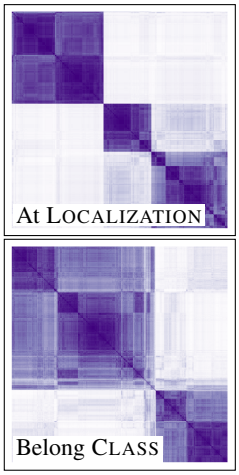
and UK, who in turn translate books from the USSR. Fig. 5d represents a subsystem of relations in which the USA is a clear outlier due to its unusually high number of immigrants and foreign students: the HIRM has inferred that these interactions are independent of the fact that China and Russia, for example, are geopolitical rivals of the USA (Fig. 5b). Fig. 5e contains sparse relations such as "Attack Embassy" and "Sever Relations", which form a subsystem with one country cluster and a small probability for the hostile event.

In contrast to the HIRM, the IRM cannot detect subsystem structure of this form since it uses a single country partition for all interactions, which is an inaccurate explanation of the data in light of the widely varying interaction patterns in the subsystems (Figs. 5b–5e) discovered by the HIRM.

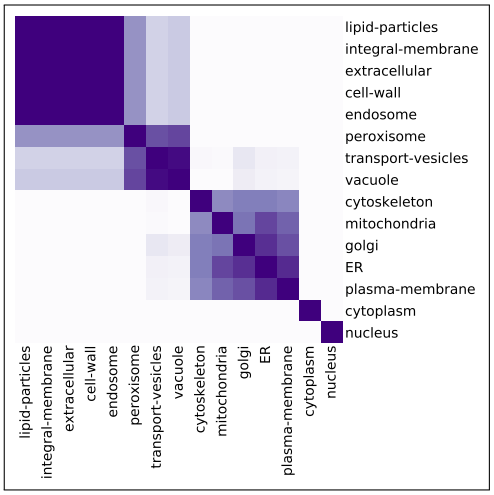
5.3 GENOMIC PROPERTIES

Our third application of the HIRM is to structure discovery in a widely used dataset of yeast genomes (Cheng et al., 2002). Fig. 1b shows a diagram of the relational system. There are nine domains: the GENE domain has 1,243 unique identifiers and the remaining domains represent gene properties. There is one binary relation between GENE and each of the eight other domains, as well as one binary relation (Interact) on GENE. A single gene is typically involved in multiple relations with the COMPLEX, PHENOTYPE, CLASS, MOTIF, and FUNCTION domains, but has only one value for ESSENTIAL and CHROMOSOME. Table 3 shows an example record for gene G235131: some characteristics of this gene are that the CLASS is missing, it forms two COMPLEX, has

(a) Inferred GENE Clusters (for two contexts)



(b) Inferred LOCALIZATION Clusters



(c) Inferred CLASS Clusters

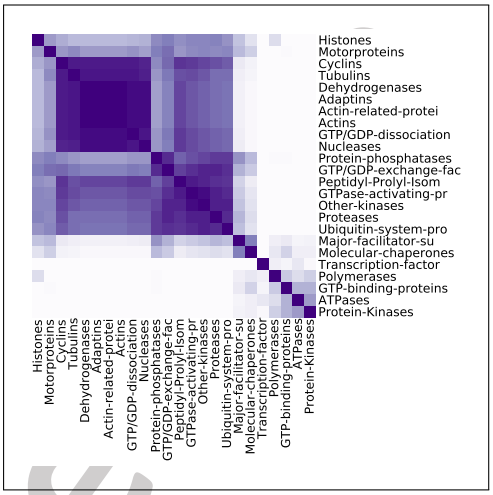


Figure 6: Posterior co-clustering probabilities for various relational domains in yeast genome data (schema in Fig. 1b).

Table 3: Example Gene

Field	Value
GENE	G235131
ESSENTIAL	Non-Essential
CLASS	?
COMPLEX	Histone Acetyltransferase
—	Transcriptosome
PHENOTYPE	Auxotrophies
—	Carbohydrate & Lipid Biosynth.
—	Conditional Phenotypes
—	Mating & Sporulation Defects
—	Nucleic Acid Metab. Defects
MOTIF	PS00633
CHROMOSOME	2
FUNCTION	Transcription
—	Cellular Organization
LOCALIZATION	Nucleus
Interactions	G234980, G235780, G235278, …

two FUNCTION; there are five observed PHENOTYPE; and it interacts with 11 other genes (three of which are listed).

Fig. 6a shows two heatmaps that summarize the clusterings of genes learned by the HIRM under two different contexts. More specifically, each row and column in a heatmap represents a unique GENE and the color of a cell is the posterior probability (between 0 and 1) that the two genes are assigned to the same latent cluster (estimated by an ensemble of 100 posterior HIRM samples). The top (resp. bottom) heatmap in Fig. 6a shows posterior co-clustering probabilities conditioned on being in the subsystem that contains the "GENE At LOCALIZATION" (resp. "GENE Belong CLASS") relation, which we call a "context". These heatmaps reflect a key feature of the HIRM: it discovers context-specific clusters that are different across the learned subsystems. Table 4 lists various co-clustering probabilities between G235131 (Table 3) and other genes, which show that a pair of genes that are similar in the LOCALIZATION context need not be similar in the CLASS context. Further, even though G235131 belongs

Table 4: Example gene posterior co-clustering probabilities in each of the two contexts shown in Fig. 6a.

GENE 1	GENE 2	Co-clustering probability within subsystem containing		Pattern
		LOCALIZATION	CLASS	
G235131	G235278	0.98	0.87	LL
G235131	G239017	0.52	0.47	MM
G235131	G236063	0.03	0.13	UU
G235131	G235388	0.83	0.27	LU
G235131	G240065	0.03	0.68	UL

U = Unlikely 0–0.33; M = Medium 0.33–0.66; L = Likely 0.66–1

to an unknown CLASS, the HIRM is still able to compute its co-clustering probabilities within this context by using observations of its other properties (i.e., relation values) that are inferred to be predictive of the missing value.

We next computed posterior co-clustering probabilities for domains that represent gene properties. In Fig. 6b, the HIRM infers a likely cluster of LOCALIZATION entities that includes cell wall, extracellular, integral membrane, and lipid particles, whereas cytoplasm and nucleus are inferred as probable singletons. Fig. 6c shows co-clustering probabilities for CLASS, which reflect a probable cluster (cyclins, tublins, adaptins, …) embedded within a larger more noisy cluster, as well as singletons such as transcription factor and polymerases. These heatmaps show quantitative estimates of posterior uncertainty in the partition structures detected by the HIRM, which cannot be captured using inference approaches such as approximate maximum likelihood or maximum a posteriori estimation and highlight a key benefit of using fully Bayesian sampling approaches (Sec. 4.1) for probabilistic structure learning in complex domains.

6 RELATED WORK

Several variations of the standard IRM have been introduced in the literature on nonparametric relational Bayesian mod-

els (Ishiguro et al., 2012; Ohama et al., 2013; Jonas and Kording, 2015; Briercliffe, 2016). Our method is distinguished by being the first hierarchical extension that uses a nonparametric structure learning prior over the relations themselves to improve modeling capacity and address shortcomings of the IRM identified in Sec. 3, which include combinatorial over-clustering and failing to detect relationships between dependent but non-identically distributed relations. These limitations have not been addressed by previous variations of the IRM. A key advantage of our hierarchical approach is that it can be composed with several IRM variants that address other shortcomings of the standard IRM, including (i) the subset IRM (Ishiguro et al., 2012), which detects and filters out irrelevant observations in the case of extreme sparsity; and (ii) the logistic regression IRM (Jonas and Kording, 2015), which improves predictive accuracy for semi-supervised tasks that specify one or more target variables as well as exogenous (non-probabilistic) predictor variables.

Other approaches to relational modeling include relational extensions of Bayesian networks (Heckerman et al., 2004; Koller and Pfeffer, 1997; Friedman et al., 1999) and Markov random fields (Taskar et al., 2002; Richardson and Domingos, 2006). While these approaches are typically more expressive than the models we consider here, they inherit traditional challenges of structure learning and model selection for directed models (Daly et al., 2011) (e.g., there is a super-exponential number of graphs to consider (Robinson, 1977)); and can require tuning evaluation measures, clause construction operators, or search strategies (Kok and Domingos, 2005) for undirected models. We instead build on Bayesian nonparametric relational models (Fan et al., 2020) that (i) use latent variables to provide a layer of indirection and simplify the learning problem as compared to searching over arbitrary graphical structures; and (ii) can be learned using principled algorithms for Bayesian inference.

Deep generative models have also been developed for relational data (Kipf and Welling, 2016; Mehta et al., 2019; Fan et al., 2019; Qu et al., 2019). These methods either typically assume that there is one binary adjacency matrix being modeled (i.e., a random graph relation) or work in a semi-supervised setting of predicting labels. In contrast, we aim to discover generative models for datasets with richer relational schemas than a single binary matrix (e.g., Fig. 1) and operate in a fully unsupervised setting without assuming beforehand that there are specific labels to predict. This approach allows us in Sec. 5.1 to make predictions using inferred joint probabilities for up to 1556 variables, and in Secs. 5.2 and 5.3 to automatically model sparse and noisy systems with multiple entities, attributes, and interactions.

Using the Chinese restaurant process as a structure learning prior (Eq. (7)) has been considered in other settings, including non-relational tabular data (Mansinghka et al., 2016), multivariate time series (Saad and Mansinghka, 2018), topic modeling (Blei et al., 2010), and computer vision (Salakhutdinov et al., 2013), among others. The same insight of using an outer CRP to partition relations (used in this work to extend the IRM) can also be applied to other models that handle relational systems with multiple relations, such as the Mondrian process (Roy and Teh, 2009). More broadly, it would be particularly fruitful to investigate a representation theorem for the ergodic distributions of a relational system modeled by an HIRM within the framework of exchangeable random structures from Orbanz and Roy (2015).

In addition to the IRM, several other Bayesian nonparametric models are special cases of the HIRM, including the infinite hidden relational model (Xu et al., 2006), infinite mixture model (Rasmussen, 2000), Dirichlet process mixture model (Lo, 1984), and Cross-Categorization (Mansinghka et al., 2016). By generalizing the likelihood term in Eq. (10) to include regression on relation values that are endogenous to the system, the HIRM could be further extended to express a relational variant of Dirichlet process mixtures of generalized linear models (Hannah et al., 2011).

Finally, as a domain-general model for relational data, the HIRM can be used to extend previous methods for automatic Bayesian modeling of non-relational tabular data that synthesize probabilistic programs in domain-specific languages (Saad et al., 2019). Expressing the HIRM in probabilistic programming languages would simplify several end-user workflows for data analysis tasks such as imputation, outlier detection, dependence detection, and search (Saad and Mansinghka, 2016, 2017; Saad et al., 2017), as well as enable fast exact inference (Saad et al., 2021) for the broad range of probabilistic queries that the HIRM can handle.

7 CONCLUSION

This paper has presented the hierarchical infinite relational model (HIRM), a new method for discovering probabilistic structure in relational data. A key insight in our approach is to use a nonparametric prior that divides a system of relations into independent subsystems, each to be learned using a separate infinite relational model. This Bayesian nonparametric approach to structure learning generalizes the standard infinite relational model (Kemp et al., 2006) and addresses several limitations in its inductive biases.

While methods based on the IRM, such as the HIRM, specify relatively simple probabilistic theories for relational systems as compared to other approaches that specify more complex theories (Muggleton and de Raedt, 1994; Getoor et al., 2007), our evaluations illustrate the efficacy of our approach on density estimation tasks and show that it can discover meaningful structure in real-world politics and genomics datasets. The results also underscore the benefit of principled and fully Bayesian structure learning for inferring probable independences, which can improve scalability, interpretability, uncertainty characterization, and model fit.

REFERENCES

D. J. Aldous. Exchangeability and related topics. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*. Springer, 1985.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, Jan. 2010.

C. Briercliffe. Poisson process infinite relational model: A Bayesian nonparametric model for transactional data. Master's thesis, University of British Columbia, 2016.

P. J. Carrington, J. Scott, and S. Wasserman, editors. *Models and Methods in Social Network Analysis*. Number 27 in Structural Analysis in the Social Sciences. Cambridge University Press, 2005.

S. Chaudhuri, U. Dayal, and V. Narasayya. An overview of business intelligence technology. *Commun. ACM*, 54(8): 88–98, 2011.

J. Cheng, C. Hatzis, H. Hayashi, M.-A. Krogel, S. Morishita, D. Page, and J. Sese. KDD cup 2001 report. *SIGKDD Expl. Newsl.*, 3(2):47–64, 2002.

R. Daly, Q. Shen, and S. Aitken. Learning Bayesian networks: Approaches and issues. *The Knowledge Engineering Review*, 26(2):99–157, 2011.

S. Džeroski and N. Lavrač, editors. *Relational Data Mining*. Springer, 2001.

X. Fan, B. Li, and S. A. Sisson. The binary space partitioning-tree process. In *Proc. AISTATS 2018*, volume 84 of *Proceedings of Machine Learning Research*, pages 1859–1867. PMLR, 2018.

X. Fan, B. Li, C. Li, S. Sisson, and L. Chen. Scalable deep generative relational model with high-order node dependence. In *Proc. NeurIPS 2019*. Curran Associates, Inc., 2019.

X. Fan, B. Li, L. Luo, and S. A. Sisson. Bayesian nonparametric space partitions: A survey. Technical Report arXiv:2002.11394, arXiv, 2020.

N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI 1999*, 1999.

R. Gens and P. Domingos. Learning the structure of sum-product networks. In *ICML 2013*, volume 28 of *Proceedings of Machine Learning Research*, pages 873–880. PMLR, 2013.

L. Getoor, N. Friedman, D. Koller, A. Pfeffer, and B. Pfeffer. Probabilistic relational models. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

L. A. Hannah, D. M. Blei, and W. B. Powell. Dirichlet process mixtures of generalized linear models. *J. Mach. Learn. Res.*, 12(54):1923–1953, 2011.

D. Heckerman, C. Meek, and D. Koller. Probabilistic models for relational data. Technical Report MSR-TR-2004-30, Microsoft Research, 2004.

J. G. Hershberg. The United States, Brazil, and the Cuban missile crisis, 1962 (Part 1). *J. Cold War Stud.*, 6(2):3–20, 2004.

K. Ishiguro, N. Ueda, and H. Sawada. Subset infinite relational models. In *Proc. AISTATS 2012*, Proceedings of Machine Learning Research, pages 547–555. PMLR, 2012.

S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13 (1):158–182, 2004.

E. Jonas and K. Kording. Automatic discovery of cell types and microcircuitry from neural connectomics. *eLife*, 4: e04250, 2015.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI 2006*, pages 381–388. AAAI Press, 2006.

D. I. Kim, P. Gopalan, D. M. Blei, and E. B. Sudderth. Efficient online inference for Bayesian nonparametric relational models. In *Proc. NIPS 2013*. Curran Associates, Inc., 2013.

T. N. Kipf and M. Welling. Variational graph auto-encoders. Technical Report arXiv:1611.07308, arXiv, 2016.

S. Kok and P. Domingos. Learning the structure of Markov logic networks. In *Proc. ICML 2005*, pages 441–448. ACM, 2005.

D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In *UAI 1997: Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 1997.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.*, 12(1):351–357, 1984.

V. Mansinghka, P. Shafto, E. Jonas, C. Petschulat, M. Gasner, and J. B. Tenenbaum. CrossCat: A fully Bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *J. Mach. Learn. Res.*, 17(138):1–49, 2016.

N. Mehta, L. Carin, and P. Rai. Stochastic blockmodels meet graph neural networks. In *Proc. ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 4466–4474. PMLR, 2019.

S. Muggleton and L. de Raedt. Inductive logic programming: Theory and methods. *J. Log. Program.*, 19–20:629–679, 1994.

M. Nakano, K. Ishiguro, A. Kimura, T. Yamada, and N. Ueda. Rectangular tiling process. In *Proc. ICML 2014*, volume 32 of *Proceedings of Machine Learning Research*, pages 361–369. PMLR, June 2014.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.*, 9(2): 249–265, 2000.

R. M. Neal. Slice sampling. *Ann. Statist.*, 31(3):705–767, 2003.

I. Ohama, H. Iida, T. Kida, and H. Arimura. An extension of the infinite relational model incorporating interaction between objects. In *Proc. PAKDD 2013*, volume 79819 of *Lecture Notes in Artificial Intelligence*, pages 147–159. Springer, Apr. 2013.

P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):437–461, Feb. 2015.

R. Peharz et al. Random sum-product networks: A simple and effective approach to probabilistic deep learning. In *Proc. UAI 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 334–344. PMLR, 2019.

M. Qu, Y. Bengio, and J. Tang. GMNN: Graph Markov neural networks. In *Proc. ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 5241–5250. PMLR, 2019.

C. E. Rasmussen. The infinite Gaussian mixture model. In *Proc. NIPS 1999*, pages 554–560. MIT Press, 2000.

M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

C. Ritter and M. A. Tanner. Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler. *J. Am. Stat. Assoc.*, 87(419):861–868, 1992.

R. W. Robinson. Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics V*, volume 622 of *Lecture Notes in Mathematics*, pages 28–43. Springer, 1977.

D. M. Roy and Y. Teh. The Mondrian process. In *Proc. NIPS 2008*, pages 833–840. Curran Associates, Inc., 2009.

J. F. Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

R. J. Rummel. Dimensionality of nations project: Attributes of nations and behavior of nation dyads, 1950-1965. Technical Report ICPSR 5409, Inter-university Consortium for Political and Social Research, 1999.

F. Saad and V. Mansinghka. A probabilistic programming approach to probabilistic data analysis. In *Proc. NIPS 2016*, pages 2011–2019. Curran Associates, Inc., 2016.

F. Saad and V. Mansinghka. Detecting dependencies in sparse, multivariate databases using probabilistic programming and non-parametric Bayes. In *Proc. AISTATS 2017*, volume 54 of *Proceedings of Machine Learning Research*, pages 632–641. PMLR, 2017.

F. Saad, L. Casarsa, and V. Mansinghka. Probabilistic search for structured data via probabilistic programming and nonparametric Bayes. Technical Report arXiv:1704.01087, arXiv, 2017.

F. A. Saad and V. K. Mansinghka. Temporally-reweighted Chinese restaurant process mixtures for clustering, imputing, and forecasting multivariate time series. In *Proc. AISTATS 2018*, volume 84 of *Proceedings of Machine Learning Research*, pages 755–764. PMLR, 2018.

F. A. Saad, M. F. Cusumano-Towner, U. Schaechtle, M. C. Rinard, and V. K. Mansinghka. Bayesian synthesis of probabilistic programs for automatic data modeling. *Proc. ACM Program. Lang.*, 3(POPL):37:1–37:32, Jan. 2019.

F. A. Saad, M. C. Rinard, and V. K. Mansinghka. SPPL: probabilistic programming with fast exact symbolic inference. In *Proc. PLDI 2021*. ACM, 2021.

R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba. Learning with hierarchical-deep models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1958–1971, Aug. 2013.

X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. Artif. Intell.*, 2009(421425), 2009.

I. Sutskever, R. R. Salakhutdinov, and J. B. Tenenbaum. Modelling relational data using Bayesian clustered tensor factorization. In *Proc. NIPS 2009*. Curran Associates, Inc., 2009.

B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proc. UAI 2002*. AUAI Press, 2002.

Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel. Infinite hidden relational models. In *Proc. UAI 2006*. AUAI Press, 2006.

Z. Xu, V. Tresp, S. Yu, K. Yu, and H.-P. Kriegel. Fast inference in infinite hidden relational models. In *Proc. MLG 2007*. ACM, 2007.

J. Xuan, J. Lu, G. Zhang, R. Y. D. Xu, and X. Luo. Bayesian nonparametric relational topic model through dependent gamma processes. *IEEE Trans Knowl. Data Eng.*, 40(7), 2017.