

A Complete Characterization of Projectivity for Statistical Relational Models

Manfred Jaeger^{1*} and Oliver Schulte²

¹Computer Science Department, Aalborg University, Aalborg, Denmark

²School of Computing Science, Simon Fraser University, Burnaby, Canada

jaeger@cs.aau.dk, oschulte@cs.sfu.ca

Abstract

A generative probabilistic model for relational data consists of a family of probability distributions for relational structures over domains of different sizes. In most existing statistical relational learning (SRL) frameworks, these models are not projective in the sense that the marginal of the distribution for size- n structures on induced substructures of size $k < n$ is equal to the given distribution for size- k structures. Projectivity is very beneficial in that it directly enables lifted inference and statistically consistent learning from sub-sampled relational structures. In earlier work some simple fragments of SRL languages have been identified that represent projective models. However, no complete characterization of, and representation framework for projective models has been given. In this paper we fill this gap: exploiting representation theorems for infinite exchangeable arrays we introduce a class of directed graphical latent variable models that precisely correspond to the class of projective relational models. As a by-product we also obtain a characterization for when a given distribution over size- k structures is the statistical frequency distribution of size- k substructures in much larger size- n structures. These results shed new light onto the old open problem of how to apply Halpern et al.’s “random worlds approach” for probabilistic inference to general relational signatures.

1 Introduction

Many types of generative models have been proposed for relational data in several fields, including machine learning and statistics. For i.i.d. data, a parametrized model defines a distribution over samples of a fixed size n , for every n . The analogue for generative relational models is a distribution $Q^{(n)}$ over complex multi-relational graphs (“worlds” in logical terminology) of a fixed size n , for every n . Research in statistical theory and discrete mathematics on the one hand, and AI and machine learning on the other hand has focussed on somewhat different aspects of relational models: the former is

mostly concerned with internal model properties such as exchangeability, projectivity and behavior in the limit, whereas the latter is focussed on learning and inference tasks for one size n at a time.

It is well known that in many popular statistical relational learning (SRL) frameworks the dependence of $Q^{(n)}$ on n exhibits sometimes counter-intuitive and hard to control behavior. Most types of SRL models are not projective in the sense that the distribution $Q^{(n)}$ for n nodes is the marginal distribution derived from the Q^{n+1} distribution [Shalizi and Rinaldo, 2013; Jaeger and Schulte, 2018]. For exponential random graph and Markov logic network (MLN) models it has also been observed that the $Q^{(n)}$ tend to become degenerate as n increases in the sense that the probability becomes concentrated on a few “extreme” structures [Rinaldo et al., 2009; Chatterjee and Diaconis, 2013; Poole et al., 2014]. Some authors have proposed to better control the behavior of MLNs by adjusting the model parameters as a function of n [Jain et al., 2010]; however, no strong theoretical guarantees have yet been derived for such approaches.

In this paper we focus on projectivity as a very powerful condition to control the behavior of $Q^{(n)}$. In projective models, inferences about a fixed set of individuals are not sensitive to population size. This implies that inference trivially becomes *domain-lifted* [Van den Broeck, 2011], convergence of query probabilities becomes trivial, and certain statistical guarantees for learning from sub-sampled relational structures can be obtained [Jaeger and Schulte, 2018]. These benefits come at a certain cost in terms of expressivity: projective models are necessarily “dense” in the sense that, e.g., the expected number of edges in a projective random graph model is quadratic in n . In spite of these limitations, there exist projective model types such as the stochastic block model and the infinite relational model [Xu et al., 2006; Kemp et al., 2006] that have been proven very useful in practice. It thus seems very relevant to fully exploit the capabilities of projective models by developing maximally expressive projective representation, learning and inference frameworks. In this paper we take an important step in this direction by deriving a complete characterization of projective models as a certain class of directed latent variable models.

While the characterization we obtain is completely general, we approach our problem from the perspective that knowl-

*Contact Author

edge about the distributions $Q^{(n)}$ is given in the form of statistical frequencies of substructures of a small size k . For example, k could be the maximal number of variables in an MLN formula, in which case the substructure frequencies are a sufficient statistics for learning the MLN parameters. In a somewhat different setting, k can be the number of variables used in a Halpern/Bacchus-style statistical probability formula forming a statistical knowledge base [Halpern, 1990; Bacchus, 1990]. In all cases the question arises of how to generalize this knowledge to infer probabilities for specific instances (“beliefs”), either by statistical model estimation (as in most current SRL frameworks), or by inferring plausible beliefs based on invariance or maximum entropy principles, as in the random worlds approach of Bacchus et al. [1992], and more recently in [Kern-Isberner and Thimm, 2010] and [Kuzelka et al., 2018]. A fundamental question that then arises is whether the given substructure frequencies can actually be the marginal distribution of $Q^{(n)}$ for large n . Results about the random worlds method need to be conditioned on the assumption that the statistical knowledge is “eventually consistent” [Halpern, 2017, Chapter 11]. Similar assumptions are made in [Kuzelka et al., 2018]. As a by-product of our characterization of projective models we obtain that the same characterization also describes the distributions that can be induced as marginals of arbitrary $Q^{(n)}$.

2 Related Work

We discuss work on generative graph models related to exchangeability and projectivity, the two key properties in our study.

Exchangeability. Exchangeability requires that a generative model should assign the same probability to graphs that differ only in node labellings. This is true for the large class of template-based relational models, because typical model discovery methods do not introduce templates that reference individual nodes [Kimmig et al., 2014]. For example, they may only construct first-order logic formulas with no constant symbols. This includes most structure learning algorithms for Markov Logic Networks (e.g., [Schulte and Khosravi, 2012]).¹ Similarly, the sufficient statistics of exponential random graph models (e.g., the number of triangles in a graph) are typically defined without special reference to any particular node. Niepert and Van den Broeck [2014] have exploited the weaker notion of *partial exchangeability* to obtain tractable inference for certain SRL models.

Projectivity. The importance of projectivity for graph modelling has been discussed previously [Shalizi and Rinaldo, 2013; Jaeger and Schulte, 2018]. Chatterjee and Diaconis [2013] discuss how estimation and inference in exponential random graph models depends on the sample size. Shalizi and Rinaldo [2013] give necessary and sufficient projectivity conditions for an exponential random graph model; they show that these are satisfied only in rare conditions. Jaeger and Schulte [2018] discuss a number of common SRL models, including MLNs and Relational Bayesian Networks, and

¹An exception is the Boostr system [Khot et al., 2013], which constructs first-order MLN formulas with constants.

show that they are projective only under restrictive conditions. Projective models used in practice factor a graph into independent components given a set of latent variables. Popular examples include the stochastic block model and generalizations [Hoff et al., 2002], the infinite relational model [Orbanz and Roy, 2014], and recent graph neural network models such as the graph variational auto-encoder [Kipf and Welling, 2016]. Our work shows that a latent conditional independence representation is not only sufficient for projectivity, but also necessary. We prove this result for a very large class of structured data, essentially general finite multi-dimensional arrays (tensors) with no restrictions on their dimensionality. Our results heavily depend on the theory of infinite exchangeable multi-dimensional arrays [Hoover, 1979; Aldous, 1981; Kallenberg, 2006; Orbanz and Roy, 2014]. The question of realizability of a given frequency distribution as a relational marginal has also been raised by Kuzelka et al. [2018], who then focus on approximate realizability, rather than characterizations of exact realizability.

3 Background

3.1 Basic Definitions

We use the following basic notation. The set of integers $\{1, \dots, n\}$ is denoted $[n]$. For any $d \geq 1$, we write $[n]_{\neq}^d$ for the set of d -tuples containing d distinct elements from $[n]$. The subset of $[n]_{\neq}^d$ containing tuples in which the elements appear in their natural order is denoted $\langle n \rangle^d$ (so that $\langle n \rangle^d$ corresponds to a standardized representation for the set of all d -element subsets of $[n]$). Extending this notation to the infinite case, we can also write $[\mathbb{N}]_{\neq}^d$ and $\langle \mathbb{N} \rangle^d$.

Relations and Possible Worlds. A relational *signature* S contains relations of varying arities. We refer to the maximal arity of relations contained in S as the *arity of* S , denoted $\text{arity}(S)$. A *possible world* ω (for S) specifies 1) a finite domain $D = \{d_1, \dots, d_n\}$, 2) for each m -ary relation from S an m -dimensional binary adjacency matrix. We refer to n as the *size* of ω , and also call ω an n -world. For most purposes, we can assume that $D = [n]$, or at least $D \subset \mathbb{N}$. However, even if we make this assumption for convenience of presentation, we do not generally assume that the integer label of a randomly observed domain element can also be observed. We denote by $\Omega^{(n)}$ the set of all possible worlds for a given signature S with domain $[n]$. The relevant signature is usually implicit from the context, and not made explicit in the notation. Finally, $\Omega := \cup_n \Omega^{(n)}$.

Relational Substructures. We also require notation to refer to different types of substructures of a possible n -world ω . For a subset $I \subset [n]$ of size $|I| = m < n$ we denote with $\omega \downarrow I$ the m -world induced by I , i.e., the possible world with domain I , and the relations of ω restricted to arguments from I . For a tuple $i \in [n]_{\neq}^m$ we denote with $\omega \downarrow i$ the world over the domain $[m]$ obtained by relabeling the domain elements in the sub-world induced by the set i as $i_h \mapsto h$ (cf. Figure 1, top row). A little less conventional is the following concept, that will become important for our main theorem: for $m = 1, \dots, \text{arity}(S)$ we define $D_m(\omega)$ as the *arity- m data* of ω . Informally speaking, $D_m(\omega)$ collects

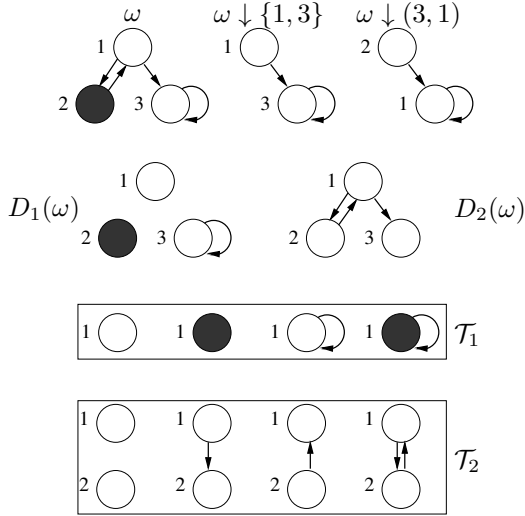


Figure 1: Top left: world ω with one unary relation (black/white) and one binary (edge) relation; top middle/right: sub-worlds induced by $I = \{1, 3\}$ and $i = (3, 1)$; second row: unary and binary data parts; bottom: spaces $\mathcal{T}_1, \mathcal{T}_2$ for the given signature.

all the information from all adjacency arrays of ω that refers to exactly m distinct elements. For example (cf. Figure 1), $D_1(\omega)$ contains the data (adjacency arrays) of all unary relations of S , but also the information contained on the diagonal of a two-dimensional adjacency array for a binary (edge) relation, i.e., the information about self-loops of that relation. A possible world can then also be described by the tuple $(D_m)_{m=1, \dots, \text{arity}(S)}$. Furthermore, $D_m(\omega)$ can be decomposed into the factors $D_m(\omega \downarrow i)$, where i ranges over $\langle n \rangle^m$. We denote with \mathcal{T}_m the space of possible values of $D_m(\omega \downarrow i)$ ($|i| = m$). A possible world $\omega \in \Omega^{(n)}$ then also is given by an assignment of a value in \mathcal{T}_m for all $i \in \langle n \rangle^m$ ($m = 1, \dots, \text{arity}(S)$).

4 Worldlet Frequency Distributions

Many graph analysis methods examine frequent characteristic subgraphs to provide information about a larger graph. We can think of a subgraph as a template that can be instantiated multiple times in a large graph. For example, in a social network we can count the number of friendship triangles among women. Depending on the framework, such templates go by different names (e.g., graphlets, motifs, frequency subgraphs) and are represented using different syntax (e.g., SQL queries, first-order logic, semantic relationships). We observe that subgraph templates can be represented in a general syntax-independent way as the collection of fully specified graphs $\Omega^{(k)}$ of a fixed size k , where we think of k as a small number (typically in the range $k = 2, \dots, 5$). When seen as a subgraph pattern, we refer to a world $\omega \in \Omega^{(k)}$ as a *worldlet*. We assume that for every worldlet, the frequency of its occurrence in a larger world is available, through learning or expert elicitation (cf. [Bacchus, 1990]). As a notational convention, we use k and n to denote domain sizes of (small) worldlets and large “real” worlds, respectively. This convention only

is intended to support intuitions, and does not have any strict mathematical implications.

Statistical Frequency Distributions. The intuitive idea of observing random worlds by sampling subsets of larger domains can be formalized in slightly different ways, e.g. by assuming sampling with or without replacement, or by interpreting the observation as a unique world, or only an isomorphism class [Diaconis and Janson, 2007; Kuzelka *et al.*, 2018]. In many aspects alternative sampling models become essentially equivalent as $n \rightarrow \infty$ [Diaconis and Janson, 2007]. We here adopt a sampling model in which an ordered sample is drawn without replacement. Thus, a sample from a world $\omega \in \Omega^{(n)}$ is given by one of the $n!/(n-k)!$ tuples $i \in [n]_{\neq}^k$, and the observed worldlet then is $\omega \downarrow i$. Note that this sampling method does not rely on observing the original labels of elements drawn from ω to obtain the labeling of elements in the sampled worldlet, and therefore also makes sense when the elements of ω can not be assumed to have (observable) integer labels. The frequency distribution obtained through this sampling method is denoted $P^{(k)}(\cdot|\omega)$.

Example 4.1 Let $S = \{e\}$ consist of a single binary relation. Let $\omega \in \Omega^{(n)}$ be a “star” with center 1, i.e., e consists of the edges $\{1 \rightarrow l : l = 2, \dots, n\}$. The probability that a random draw of 2 elements contains the node 1 then is $2/n$, with equal probability that 1 is the first or second drawn element. The three worldlets $1 \bullet \bullet 2$, $1 \bullet \rightarrow 2$ and $1 \bullet \leftarrow 2$ then have probabilities $1 - 2/n$, $1/n$, $1/n$ (in this order) under $P^{(k)}(\cdot|\omega)$.

Every world ω defines a frequency distributions $P^{(k)}(\cdot|\omega)$. If first a random ω is selected, we obtain a two-step sampling procedure that was first described in a more general context by Fenstad [1967].

Fenstad Sampling. Given a possible world distribution $Q^{(n)}$, we define the *expected statistical frequency* distribution $P^{(k)} \circ Q^{(n)}$ for k -worlds ω' as follows:

$$(P^{(k)} \circ Q^{(n)})(\omega') := \sum_{\omega \in \Omega^{(n)}} Q^{(n)}(\omega) P^{(k)}(\omega' | \omega). \quad (1)$$

We denote with $\Delta_n^{(k)}$ the set of distributions on $\Omega^{(k)}$ that have a representation of the form (1) for some $Q^{(n)}$. If $k < l < n$, then $P^{(k)} \circ (P^{(l)} \circ Q^{(n)}) = P^{(k)} \circ Q^{(n)}$, and thus $\Delta_n^{(k)} \subseteq \Delta_l^{(k)}$.

Example 4.2 In this example and some of the following, we take S to contain a single undirected edge relation e . In order to comply with our general definitions, which are based on directed relations, we consider an undirected edge $i \bullet - \bullet j$ to be a shorthand for the conjunction $i \bullet \rightarrow \bullet j$ and $i \bullet \leftarrow \bullet j$, and we assume that all worlds with uni-directional edges ($i \bullet \rightarrow \bullet j$ but not $i \bullet \leftarrow \bullet j$) or self-loops ($i \bullet \rightarrow \bullet i$) have probability zero. Disregarding these probability zero worlds, $\Omega^{(3)}$ then contains 8 possible worlds belonging to 4 different isomorphism classes. The top row of Table 1 depicts these isomorphism classes, together with the count of worlds in each class.

Figure 2 illustrates for $n = 3, 4, 5, 6$ the worldlet frequency distributions $P^{(k)}(\cdot|\omega)$ defined by the worlds $\omega \in \Omega^{(n)}$.

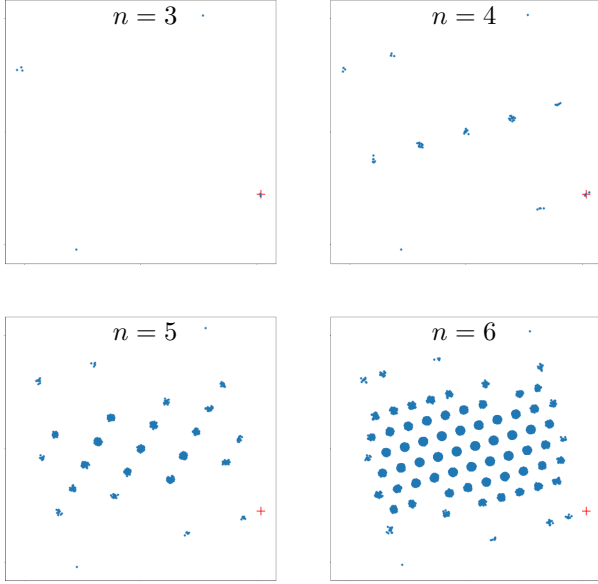


Figure 2: Illustration of $\Delta_n^{(k)}$ for $k = 3$ and $n = 3, 4, 5, 6$. Cf. examples 4.2 and 5.6

Each (blue) dot is the distribution defined by one world after projecting its 8-dimensional probability vector into 2-dimensional space. Some jitter is applied to exhibit the multiplicities of n -worlds defining the same distribution on worldlets of size 3. The sets $\Delta_n^{(k)}$ are the convex hulls of these points. The distribution marked by the (red) + in Table 1 and Figure 2 belongs to $\Delta_n^{(k)}$ for $n = 3, 4$, but not for $n = 5, 6$.

5 Relational Models and Distribution Families

As our goal is to examine properties of relational models that are independent of a particular model syntax, we use a family of distributions as a semantic view of a parametrized model. The two key properties of families in our study are exchangeability and projectivity.

5.1 Distribution Families: Exchangeability, and Projectivity

Definition 5.1 A family of distributions $\{Q^{(n)} : n \in \mathbb{N}\}$ specifies, for each finite domain size n , a distribution $Q^{(n)}$ on the possible world set $\Omega^{(n)}$.

Definition 5.2 A probability distribution $Q^{(n)}$ on $\Omega^{(n)}$ is *exchangeable*, if $Q^{(n)}(\omega) = Q^{(n)}(\omega')$ whenever ω and ω' are isomorphic. A family is exchangeable, if every member of the family is exchangeable.

Intuitively a distribution family is projective if its members are mutually consistent in the sense that the world distribution over a smaller domain size is the marginal distribution over a larger one. For a precise definition, we follow our notation for relational substructures, and for each n -world ω , write $\omega \downarrow [m]$ for the size- m subworld that results from restricting ω to





 ($\times 1$)	 ($\times 3$)	 ($\times 3$)	 ($\times 1$)	Name
1	0	0	0	1_{E_3}
0	0	0	1	1_{K_3}
0	1/3	0	0	+
1/4	0	1/4	0	bipart

Table 1: Some example worldlet distributions

the first m elements. A distribution $Q^{(n)}$ over n -worlds then induces a *marginal* probability for an m -world ω' as follows:

$$Q^{(n)} \downarrow [m](\omega') = \sum_{\omega \in \Omega^{(n)} : \omega \downarrow [m] = \omega'} Q^{(n)}(\omega)$$

Projectivity is the central concept for our investigation:

Definition 5.3 An exchangeable family $(Q^{(n)})_{n \in \mathbb{N}}$ is *projective*, if for all $m < n$: $Q^{(n)} \downarrow [m] = Q^{(m)}$.

Note that in contrast to more general notions of projectivity found in the theory of stochastic processes, we here define projectivity only for exchangeable families. Exchangeability implies that the marginal distribution $Q^{(n)} \downarrow I$ is the same for all subsets I of size m , and therefore we only need to consider the marginal $Q^{(n)} \downarrow [m]$ as a prototype.

Example 5.4 Statistical frequency distributions $P^{(k)}(\cdot \mid \omega)$ always are exchangeable. As a special case, if $\omega \in \Omega^{(n)}$, then $P^{(n)}(\cdot \mid \omega)$ samples a random permutation of ω , i.e., is the uniform distribution on the isomorphism class of ω . It follows that distributions defined by Fenstad sampling (1) also are exchangeable, for any $Q^{(n)}$.

We approach the question of how to characterize and represent projective families through the more specific question of whether a given distribution $Q^{(k)}$ can be embedded in a projective family. The following definition provides the necessary terminology.

Definition 5.5 Let $Q^{(k)}$ be an exchangeable distribution on $\Omega^{(k)}$. $Q^{(k)}$ is called

- *n*-*extendable*, if $Q^{(k)} \in \Delta_n^{(k)}$; any $Q^{(n)}$ that induces $Q^{(k)}$ via (1) is called an *extension* of $Q^{(k)}$.
- *extendable*, if it is *n*-extendable for all $n > k$;
- *projective extendable* if there exists a projective family $(Q^{(n)})_n$ of extensions of $Q^{(k)}$.

Example 5.6 The rows in Table 1 specify several exchangeable distributions on $\Omega^{(3)}$ (in the undirected graph setting, as described in Example 4.2). The numbers in the table specify the probabilities of each world in a given isomorphism class, not the total probability of the isomorphism class. The first two are the point masses on the empty graph (denoted E_3) and complete graph (K_3), respectively. If 1_{E_n} denotes the point mass on the empty graph of size n , then $(1_{E_n})_n$ is a projective family. Similarly for the family $(1_{K_n})_n$, and the family of mixtures $(0.5 \cdot 1_{E_n} + 0.5 \cdot 1_{K_n})_n$.

The row labeled + is the distribution marked by the (red) + in the plots of Figure 2. If $\omega \in \Omega^{(4)}$ is the graph that contains

the two edges $1 \bullet \bullet 2$ and $3 \bullet \bullet 4$, then this distribution is equal to $P^{(3)}(\cdot|\omega)$. Thus, it is 4-extendable, which is also visible in the top right panel of Figure 2 showing that ‘+’ coincides with sampling distributions induced by 4-worlds. However, ‘+’ is not n -extendable for any $n \geq 5$. This is visible in Figure 2 as for $n = 5, 6$ ‘+’ lies outside the convex hull of the worldlet frequency distributions. Proposition 7.1 below will provide a simple tool for proving the non-extendability of ‘+’.

The last row in the table describes the distribution that in the limit for $n \rightarrow \infty$ is the worldlet frequency distribution defined by complete, balanced bipartite graphs, i.e., graphs whose edge set is equal to $\{i \bullet \bullet j : 1 \leq i \leq \lfloor n/2 \rfloor; \lfloor n/2 \rfloor + 1 \leq j \leq n\}$. It will follow from our main theorem that this distribution is projective extendable.

5.2 Domain Sampling Distributions

Extendable distributions $Q^{(k)}$ in the sense of Definition 5.5 are mixtures of worldlet frequency distributions. An important special case is when $Q^{(k)}$ is a pure worldlet frequency distribution $P^{(k)}(\cdot|\omega)$ defined by a single world ω . In that case, however, one cannot expect that $Q^{(k)}$ can be represented in this form with suitable ω for all n , because the sets $\{P^{(k)}(\cdot|\omega) : \omega \in \Omega^{(n)}\}$ for different n are concentrated on different grids of rational numbers, and therefore are largely disjoint (cf. Figure 2). Following the approach already taken by Bacchus et al. to give semantics to statistical probability terms in the random worlds approach [Bacchus *et al.*, 1992; Halpern, 2017] we therefore only require that $Q^{(k)}$ is approximately equal to some $P^{(k)}(\cdot|\omega)$, with an increasing accuracy in the approximation as the size of ω increases.

Definition 5.7 Let $Q^{(k)}$ be a probability distribution on $\Omega^{(k)}$. We say that $Q^{(k)}$ is a *domain sampling distribution* if the following holds: for every $\epsilon > 0$ there exists $n \in \mathbb{N}$, such that for every $n' \geq n$: there exists a possible n' -world ω , so that for all $\omega' \in \Omega^{(k)}$:

$$|P^{(k)}(\omega' | \omega) - Q^{(k)}(\omega')| < \epsilon. \quad (2)$$

Thus, the property of being a domain sampling distribution strengthens the property of extendability in that in the representation (1) only point masses $Q^{(n)} = 1_\omega$ are allowed, but weakens it in that (2) only requires approximate equality.

Example 5.8 For the worldlet distributions of Table 1 we have $1_{E_3} = P^{(3)}(\cdot|E_n)$ for all $n \geq 3$, so that 1_{E_3} is a domain sampling distribution (with zero approximation error). Similarly for 1_{K_3} . The mixture $0.5 \cdot 1_{E_3} + 0.5 \cdot 1_{K_3}$ is projective extendable, but not a domain sampling distribution. The distribution ‘+’ is not a domain sampling distribution. This is indicated by Figure 2, because already for $n = 6$ the distribution is separated by a distance $\epsilon > 0$ from the set $\Delta_6^{(3)}$. Because of the nested structure of the $\Delta_n^{(3)}$ there then also cannot be better approximations for larger $n > 6$. The last ‘bipart’ distribution in Table 1 again is a domain sampling distribution with a non-zero approximation error that only vanishes as $n \rightarrow \infty$.

6 A Representation Theorem

We now proceed to derive our main result, which is a comprehensive characterization of families $(Q^{(n)})_n$ and worldlet marginals $Q^{(k)}$ with the structural properties described in Section 5. We introduce a representation for projective families that is based on the analysis and representation theorems for infinite exchangeable arrays developed by Aldous [1981] and Hoover [1979]. The definitive treatment is given by Kallenberg [2006]. We therefore call the following an AHK model.

Definition 6.1 Let S be a signature with maximal $\text{arity}(S) = a \geq 1$. An AHK model for S is given by

- A family of i.i.d. random variables $\{U_i | i \in \langle \mathbb{N} \rangle^m, m = 0, \dots, a\}$, where each U_i is uniformly distributed on $[0, 1]$.
- A family of random variables $\{D_i | i \in \langle \mathbb{N} \rangle^m, m = 1, \dots, a\}$. For $i \in \langle \mathbb{N} \rangle^m$ the variable D_i takes values in \mathcal{T}_m .
- For each $m = 1, \dots, a$ a measurable function

$$f^m : [0, 1]^{2^m} \rightarrow \mathcal{T}_m \quad (3)$$

so that

- for $i = (i_1, \dots, i_m) \in \langle \mathbb{N} \rangle^m$ the value of D_i is defined as $f^m(U_i)$, where

$$U_i = (U_\emptyset, U_{i_1}, \dots, U_{i_m}, U_{(i_1, i_2)}, \dots, U_{(i_{m-1}, i_m)}, \dots, U_{(i_1, \dots, i_m)}), \quad (4)$$

is the vector containing all $U_{i'}$ -variables with $i' \subseteq i$ in lexicographic order.

- f^m is permutation equivariant, in the sense that for any permutation π of $[m]$

$$f^m(\pi U_i) = \pi f^m(U)$$

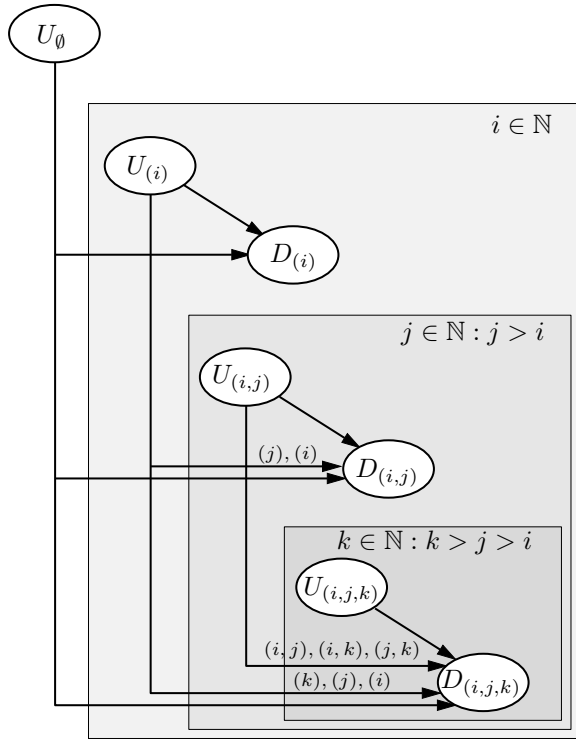
where πU_i is the permutation of U_i that in the place of $U_{i'}$ contains $U_{\pi i'}$ with $\pi i'$ the ordered tuple of the elements $\{\pi(i) : i \in i'\}$.

An AHK model that does not contain the U_\emptyset variable is called an AHK[−] model.

Figure 3 gives an illustration of the structure of an AHK model in plate notation. An AHK model is fully determined by the functions $\mathbf{f} := (f^m)_{m=1, \dots, a}$, and we therefore write \mathbf{f} to refer to an AHK model. By a slight abuse of notation, we also use \mathbf{f} to denote the distribution defined by the model on the possible worlds over the infinite domain \mathbb{N} , and write $\mathbf{f} \downarrow [n]$ for the marginal on the induced sub-world over the domain $[n]$.

The following example gives a simple illustration of how the permutation equivariance condition for the functions f^m ensures exchangeability.

Example 6.2 We encode a version of the Erdős-Rényi random graph model in which any pair of nodes is connected with probability 1/2 by an edge, and that edge is given a random direction. Thus, the target distribution on worldlets of size 2 is $P(1 \bullet \leftarrow 2) = P(1 \bullet \rightarrow 2) = 0.25$, $P(1 \bullet \bullet 2) =$


 Figure 3: Plate representation of AHK model with $a = 3$

0.5. The state space \mathcal{T}_1 contains the two states “self-loop” and “no self-loop”. Since self-loops have probability zero, we simply let f^1 be the constant function that returns “no self-loop” regardless of the input U -variables. The state space \mathcal{T}_2 contains the four states $1 \bullet \bullet 2$, $1 \bullet \rightarrow \bullet 2$, $1 \bullet \leftarrow \bullet 2$, and $1 \bullet \leftrightarrow \bullet 2$, of which only the first three have non-zero probability. Let

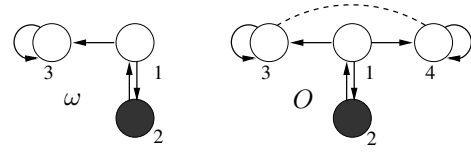
$$f^2(x_0, x_1, x_2, x_3) := \begin{cases} 1 \bullet \rightarrow \bullet 2 & \text{if } x_1 < x_2 \text{ and } x_3 < 0.5 \\ 1 \bullet \leftarrow \bullet 2 & \text{if } x_2 < x_1 \text{ and } x_3 < 0.5 \\ 1 \bullet \bullet 2 & \text{otherwise.} \end{cases}$$

For clarity we here use a notation that makes it clear that the functions f^m are defined on arrays of length 2^m , and their definition distinguishes arguments by their position in the input array, not by their semantic nature as a variable $U_{i'}$. For $\pi : 1 \mapsto 2, 2 \mapsto 1$ we then have $\pi(1 \bullet \rightarrow \bullet 2) = 1 \bullet \leftarrow \bullet 2$, and $f^2(\pi \mathbf{U}_{(1,2)}) = f^2(U_0, U_2, U_1, U_{(1,2)}) = \pi f^2(\mathbf{U}_{(1,2)})$. Together with the fact that the tuples $\mathbf{U}_{(1,2)}$ and $\pi \mathbf{U}_{(1,2)}$ have identical distribution, this implies that the two values $1 \bullet \rightarrow \bullet 2$, $1 \bullet \leftarrow \bullet 2$ of $D_{(1,2)}$ have the same probability.

Generalizing from this example, and also noting that the plate representation of the AHK models directly implies that marginals $\mathbf{f} \downarrow [n]$ simply are given by instantiating the plate model only for $i \subset [n]$, we can note the following proposition.

Proposition 6.3 Let \mathbf{f} be an AHK model. The marginals $\mathbf{f} \downarrow [n]$ are exchangeable, and the family $(\mathbf{f} \downarrow [n])_n$ is projective.

For a given worldlet distribution $Q^{(k)}$ with $k \geq \text{arity}(S)$


 Figure 4: Modularity of AHK models: if ω on the left has nonzero probability, then also the set of worlds O on the right.

we say that $Q^{(k)}$ has an AHK representation, if there exists an \mathbf{f} with $\mathbf{f} \downarrow [k] = Q^{(k)}$.

We can now formulate our main result.

Theorem 6.4 Let $Q^{(k)}$ be an exchangeable distribution on $\Omega^{(k)}$ with $k \geq \text{arity}(S)$. For the statements

- (A) $Q^{(k)}$ is a domain sampling distribution.
- (B) $Q^{(k)}$ has a AHK^- representation
- (C) $Q^{(k)}$ is a finite mixture of domain sampling distributions
- (D) $Q^{(k)}$ is extendable
- (E) $Q^{(k)}$ is projective extendable
- (F) $Q^{(k)}$ has an AHK representation

the following implications hold:

$$(A) \Leftrightarrow (B) \Rightarrow (C) \Leftrightarrow (D) \Leftrightarrow (E) \Leftrightarrow (F)$$

The full proof of the theorem is given in the extended online version of this paper (<http://arxiv.org/abs/2004.10984>).

7 Discussion

In this section we consider some of the trade-offs between limitations in expressivity of projective models on the one hand, and gain in algorithmic and statistical tractability on the other hand. Limitations in expressivity can be considered in terms of what distributions $Q^{(n)}$, for a fixed n , can be represented, and in terms of the limitations for the family $\{Q^{(n)} | n \in \mathbb{N}\}$ as a whole. Considering a single distribution $Q^{(n)}$, we can observe a *modularity* property as described by the following proposition, and illustrated in Figure 4

Proposition 7.1 Let \mathbf{f} be an AHK model, $\omega \in \Omega^{(n)}$ with $\mathbf{f} \downarrow [n](\omega) > 0$. Let $O \subset \Omega^{(n+1)}$ be the set of $n+1$ worlds ω' for which $\omega' \downarrow [n] = \omega' \downarrow \{1, \dots, n-1, n+1\} = \omega$. Then $\mathbf{f} \downarrow [n](O) > 0$. Moreover, if \mathbf{f} is an AHK^- model, then $\omega' \downarrow [n] = \omega$ and $\omega' \downarrow \{1, \dots, n-1, n+1\} = \omega$ are independent events given $\omega' \downarrow [n-1]$.

Figure 4 illustrates the proposition with $n = 3$: if the world ω on the left has nonzero probability, then also the set of 4-worlds O on the right has nonzero probability. O is the set of 4-worlds for which the substructures induced by $\{1, 2, 3\}$ and $\{1, 2, 4\}$ are both isomorphic to ω . The dashed arc connecting nodes 3 and 4 on the right indicates that the value of $D_{(3,4)}$ determining the relations between nodes 3 and 4 can vary for different elements of O .

As an application of Proposition 7.1 we can see that the ‘+’ distribution of Table 1 does not have an AHK representation, and therefore cannot be extendable (cf. Example 5.6): letting

$n = 2$ and $\omega = 1 \bullet \bullet 2$, we obtain from the proposition that also 3-worlds with two edges $1 \bullet \bullet 2$ and $2 \bullet \bullet 3$ must have nonzero probability, which is not the case for '+'.

We now turn to structural limitations of the whole family $\{Q^{(n)} | n \in \mathbb{N}\}$ implied by an AHK representation. As already mentioned in the introduction, projective families generate structures that are “dense” in the limit. More precisely, if $\omega \in \Omega^{(k)}$ is a worldlet with $f \downarrow [k](\omega) > 0$, then the expected number of k -tuples in worlds of size n which induce sub-worlds isomorphic to ω grows linearly in n^k . Specifically, if graph edges have a nonzero probability at all, then the expected number of edges grows linearly in n^2 . It must be emphasized, though, that this only imposes limits on modeling the asymptotic behavior of evolving graphs. For any fixed domain size, an AHK model can fit any observed degree distribution:

Example 7.2 Let $n^* \in \mathbb{N}$, and let $f(d)$ ($d = 0, 1, \dots, n^*$) denote an out-degree distribution for directed graphs on $[n^*]$. For arbitrary n we can normalize out-degrees in graphs of size n via $d \mapsto d/n$. Let $F(\delta)$ ($\delta \in [0, 1]$) be the cumulative distribution function obtained from $f(\cdot)$ for the normalized degrees $d \mapsto d/n^*$. We now define

$$f^2(U_i, U_j, U_{(i,j)}) := \begin{cases} i \bullet \rightarrow j & \text{if } U_i \geq F(U_{(i,j)}) \text{ and } U_j < F(U_{(i,j)}) \\ i \bullet \leftarrow j & \text{if } U_j \geq F(U_{(i,j)}) \text{ and } U_i < F(U_{(i,j)}) \\ i \bullet \leftrightarrow j & \text{if } U_i \geq F(U_{(i,j)}) \text{ and } U_j \geq F(U_{(i,j)}) \\ i \bullet \bullet j & \text{otherwise} \end{cases}$$

Let δ_i denote the normalized out-degree of node i . Then for all $u \in [0, 1]$ we obtain the expected normalized out-degree:

$$E[\delta_i | U_i = u] = F^{-1}(u). \quad (5)$$

U_i being uniformly distributed, the right-hand side of (5) is distributed with cdf $F(\cdot)$, and so the expected normalized degree distribution follows $F(\cdot)$. In the special case $n = n^*$ then the expected absolute degree distribution is the original $f(\cdot)$.

On the positive side, we obtain significant computational and robustness advantages from the use of projective models: inference is *lifted* in the strongest possible sense that the complexity of computing a query probability for a query involving k named entities is independent of the size of the domain in which the entities are embedded. For learning, projectivity is a necessary condition for consistent estimation from substructures randomly sampled from domains of unknown size. However, further conditions beyond projectivity are required to formulate and derive precise consistency guarantees [Jaeger and Schulte, 2018]. Statistical consistency and robustness results can therefore not be directly given for AHK models in general without first identifying a suitable effectively representable and parameterizable class of functions from which the f^m can be constructed. Identifying rich and tractable such classes, and evaluating their learning capabilities empirically and theoretically is future work.

When evaluating the trade-offs of AHK models for a particular application, it must always be born in mind that the strengths of generative, projective models only come to bear when one needs to deal with diverse types of queries (so that

a discriminative model for a fixed prediction task would be inadequate), and when one has to deal with data from domains of different and/or uncertain sizes. We note that this is basically the opposite side of the task spectrum from where many current popular node classification and link prediction problems are situated, in which both learning and inference is conducted for a fixed task on a single given graph, e.g., [Wu *et al.*, 2020].

8 Conclusion

In this paper we have laid theoretical foundations for the study and application of rich classes of projective families. Bringing together research strands in statistical graph theory and statistical relational learning we have derived an explicit characterization of projective families in the form of a directed graphical (plate) model. We have shown that closely linked to projectivity is the (approximate) realizability as a statistical frequency distributions of worldlet samples drawn from large domain. These results give us a characterization of the form of statistical knowledge to which the random worlds approach of Bacchus *et al.* [1992] can be applied.

Interestingly, the structure of AHK models has much in common with the “independent choice logic” family of SRL frameworks [Sato, 1995; Poole, 1997; Kimmig *et al.*, 2011] that also generate random relational structures as deterministic functions of a set of a-priori independent random variables. However, the continuous nature of the U_i variables in the AHK model, and the potential need of functions f^m not readily expressible in existing SRL languages pose significant challenges for the direct application of existing SRL techniques.

On the theoretical side, many interesting questions remain regarding statistical principles of model selection, and unbiasedness and consistency of estimation: for a given worldlet distribution $Q^{(k)}$ there will often be multiple AHK models that precisely fit $Q^{(k)}$ and therefore are indistinguishable based on likelihood scores. What invariance, parsimony, or plain parameter regularization principles are then most useful for model selection?

Acknowledgments

Oliver Schulte’s contribution was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

- [Aldous, 1981] David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.
- [Bacchus *et al.*, 1992] Fahiem Bacchus, Adam Grove, Joseph Y. Halpern, and Daphne Koller. From statistics to beliefs. In *Proc. of National Conference on Artificial Intelligence (AAAI-92)*, 1992.
- [Bacchus, 1990] Fahiem Bacchus. *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach to Probabilities*. MIT Press, Cambridge, MA, USA, 1990.

- [Chatterjee and Diaconis, 2013] Sourav Chatterjee and Persi Diaconis. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.
- [Diaconis and Janson, 2007] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, 2007.
- [Fenstad, 1967] Jens Erik Fenstad. Representations of probabilities defined on first order languages. In J. N. Crossley, editor, *Sets, Models and Recursion Theory*, pages 156–172. North Holland, Amsterdam, 1967.
- [Halpern, 1990] Joseph Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.
- [Halpern, 2017] Joseph Y Halpern. *Reasoning about uncertainty*. MIT press, 2017.
- [Hoff *et al.*, 2002] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.
- [Hoover, 1979] Douglas N. Hoover. Relations on probability spaces and arrays of random variables. *HPreprint, Institute for Advanced Study, Princeton, NJ*, 2, 1979.
- [Jaeger and Schulte, 2018] Manfred Jaeger and Oliver Schulte. Inference, learning, and population size: Projectivity for srl models. IJCAI-StarAI Workshop on Statistical-Relational AI, July 2018.
- [Jain *et al.*, 2010] Dominik Jain, Andreas Barthels, and Michael Beetz. Adaptive markov logic networks: Learning statistical relational models with dynamic parameters. In *ECAI*, pages 937–942, 2010.
- [Kallenberg, 2006] Olav Kallenberg. *Probabilistic symmetries and invariance principles*. Springer Science & Business Media, 2006.
- [Kemp *et al.*, 2006] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, 2006.
- [Kern-Isberner and Thimm, 2010] Gabriele Kern-Isberner and Matthias Thimm. Novel semantical approaches to relational probabilistic conditionals. In *Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, 2010.
- [Khot *et al.*, 2013] Tushar Khot, Jude Shavlik, and Sriraam Natarajan. Boost, 2013. URL = <http://pages.cs.wisc.edu/~tushar/Boost/>.
- [Kimmig *et al.*, 2011] Angelika Kimmig, Bart Demoen, Luc De Raedt, Vitor Santos Costa, and Ricardo Rocha. On the implementation of the probabilistic logic programming language problog. *Theory and Practice of Logic Programming*, 11(2-3):235–262, 2011.
- [Kimmig *et al.*, 2014] Angelika Kimmig, Lilyana Mihalkova, and Lise Getoor. Lifted graphical models: a survey. *Machine Learning*, pages 1–45, 2014.
- [Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
- [Kuzelka *et al.*, 2018] Ondrej Kuzelka, Yuyi Wang, Jesse Davis, and Steven Schockaert. Relational marginal problems: Theory and estimation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1–8. AAAI Press, 2018.
- [Niepert and Van den Broeck, 2014] Mathias Niepert and Guy Van den Broeck. Tractability through exchangeability: A new perspective on efficient probabilistic inference. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [Orbanz and Roy, 2014] Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461, 2014.
- [Poole *et al.*, 2014] David Poole, David Buchman, Seyed Mehran Kazemi, Kristian Kersting, and Sriraam Natarajan. Population size extrapolation in relational probabilistic modelling. In *International Conference on Scalable Uncertainty Management*, pages 292–305. Springer, 2014.
- [Poole, 1997] David Poole. The independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence*, 94(1-2):7–56, 1997.
- [Rinaldo *et al.*, 2009] Alessandro Rinaldo, Stephen E. Fienberg, and Yi Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Statist.*, 3:446–484, 2009.
- [Sato, 1995] Taisuke Sato. A statistical learning method for logic programs with distribution semantics. In *Proceedings of the 12th International Conference on Logic Programming (ICLP’95)*, pages 715–729, 1995.
- [Schulte and Khosravi, 2012] Oliver Schulte and Hassan Khosravi. Learning graphical models for relational data via lattice search. *Machine Learning*, 88(3):331–368, 2012.
- [Shalizi and Rinaldo, 2013] Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *Annals of statistics*, 41(2):508, 2013.
- [Van den Broeck, 2011] Guy Van den Broeck. On the completeness of first-order knowledge compilation for lifted probabilistic inference. In *Advances in Neural Information Processing Systems*, pages 1386–1394, 2011.
- [Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [Xu *et al.*, 2006] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Learning infinite hidden relational models. *Uncertainty in Artificial Intelligence (UAI2006)*, 2006.