Probabilistic team semantics

Arnaud Durand[1], Miika Hannula[2], Juha Kontinen[3], Arne Meier[4], and Jonni Virtema[5]

[1] Institut de Mathématiques de Jussieu - Paris Rive Gauche, CNRS UMR 7586 - Université Paris Diderot, durand@math.univ-paris-diderot.fr
[2] Department of Computer Science, University of Auckland, m.hannula@auckland.ac.nz
[3] Department of Mathematics and Statistics, University of Helsinki, juha.kontinen@helsinki.fi
[4] Leibniz Universität Hannover, Institut für Theoretische Informatik, meier@thi.uni-hannover.de
[5] Databases and Theoretical Computer Science, Hasselt University, jonni.virtema@uhasselt.be

Abstract. Team semantics is a semantical framework for the study of dependence and independence concepts ubiquitous in many areas such as databases and statistics. In recent works team semantics has been generalised to accommodate also multisets and probabilistic dependencies. In this article we study a variant of probabilistic team semantics and relate this framework to a Tarskian two-sorted logic. We also show that very simple quantifier-free formulae of our logic give rise to NP-hard model checking problems.

1 Introduction

Team semantics is the modern approach for the study of logics of dependence and independence. The systematic development of team semantics began by the introduction of Dependence Logic in 2007 [20] although the key ingredients of the new semantics were already introduced by Hodges 1997 [14]. In team semantics, satisfaction of formulae is defined not via single assignments but via sets of assignments (teams). Sets of assignments enables one to introduce a multitude of interesting atoms to the logic such as dependence, independence, and inclusion atoms:

$$=(\boldsymbol{x}, y),\ \boldsymbol{y} \perp_{\boldsymbol{x}} \boldsymbol{z} \text{ and } \boldsymbol{x} \subseteq \boldsymbol{y}$$

that do not make sense with respect to a single assignment. Independence logic, introduced by Grädel and Väänänen [10], extends first-order logic with independence atoms. The independence atom $\boldsymbol{y} \perp_{\boldsymbol{x}} \boldsymbol{z}$ holds if the value of $\boldsymbol{z}$ does not tell us anything new about the value of $\boldsymbol{y}$ when the value of $\boldsymbol{x}$ is fixed. By viewing a team X with domain $\{x_1, \ldots, x_n\}$ as a database table over attributes $x_1, \ldots, x_n$, dependence, inclusion, and independence atoms correspond exactly to functional, inclusion, and embedded multivalued dependencies (EMVDs), see, e.g., [18,13,12]. Moreover EMVDs and probabilistic conditional independence $\boldsymbol{Y} \perp \boldsymbol{Z}|\boldsymbol{X}$ have significant connections, confer, e.g., [11,21,1]. Multiteam semantics, introduced by Durand et al. [3], is the multiset analogue of team semantics. This setting enables the logical study of probabilistic dependencies such as the probabilistic conditional independence atoms $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z}$ that inherit their semantics from the corresponding notion $\boldsymbol{Y} \perp \boldsymbol{Z}|\boldsymbol{X}$ from statistics. One of the advantages of multiteam semantics is that it allowes to study the interplay of atoms such as $=(\boldsymbol{x}, y)$, $\boldsymbol{y} \perp_{\boldsymbol{x}} \boldsymbol{z}$, and $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z}$ in a unified framework.

In this paper, we focus on probabilistic team semantics. A probabilistic team is a set of assignments endowed with a probability distribution that maps each assignment of the set to a ratio. There is a vast literature on probabilistic logics but so far only few works study probabilistic team semantics. The teams that arise from applications (e.g., database tables) often contain duplicate rows leading naturally to multiteams (i.e., multiset analogues of teams). Furthermore, finite multiteams can be viewed as probabilistic teams endowed with the counting measure induced by the multiplicities. Importantly, in many applications, duplicate rows can store relevant information; e.g., if a table is used to store an outcome

of a poll or a collection of outcomes of measurements. In these cases the interest lies in the distribution of the data and not so much in the size of the sample. Hence it makes sense to abstract from the concrete data (multiteams) to the distribution of data (probabilistic teams). We consider a logic that uses probabilistic independence $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z}$ and marginal identity atoms $\boldsymbol{x} \approx \boldsymbol{y}$ as primitives in the setting of probabilistic team semantics. These atoms were recently introduced by Durand et al. [3] in the context of multiteam semantics. The marginal identity atom $\boldsymbol{x} \approx \boldsymbol{y}$ expresses that in a team the distribution of values for the variables $\boldsymbol{x}$ coincides with that of $\boldsymbol{y}$. We relate this logic to a natural variant of (two-sorted) existential second-order logic with quantification over rational distributions. We also consider the complexity of model checking and show that very simple formulae using $\boldsymbol{x} \approx \boldsymbol{y}$ give rise to NP-hard model checking problems.

Example 1. Consider a database table that lists results of experiments. The data can be regarded either as a multiteam or as the related probabilistic team using the counting measure; both interpretations having its own advantages. Each record corresponds to outcomes of measurements obtained simultaneously in two locations. The table has four attributes $\mathsf{Test1}$ and $\mathsf{Test2}$ that range over the possible types of measurements and $\mathsf{Outcome1}$ and $\mathsf{Outcome2}$ that range over outcomes of the measurements. The probabilistic independence atom $\mathsf{Test1} \perp\!\!\!\perp \mathsf{Test2}$ expresses that the types of measurements are independently picked in the two locations. The marginal identity atom $(\mathsf{Test1}, \mathsf{Outcome1}) \approx (\mathsf{Test2}, \mathsf{Outcome2})$ expresses that the distributions of results are the same in both test sites. The formula $\mathsf{Test1} = \mathsf{Test2} \vee (\mathsf{Test1} \neq \mathsf{Test2} \wedge \mathsf{Outcome1} \perp\!\!\!\perp \mathsf{Outcome2})$ expresses that there is no correlation between outcomes of the different measurements.

Example 2. Consider a database table that describes voting behaviour in two different elections by some sample of voters. Attributes of the table are $\mathsf{Election1}$ and $\mathsf{Election2}$ that range over political parties. Each record corresponds to a voting behaviour of a voter in the sample. The table then gives rise to a probabilistic team that approximates the voting behaviour of the population. The complex formula $\mathsf{Election1} = \mathsf{Election2} \vee (\mathsf{Election1} \neq \mathsf{Election2} \wedge \mathsf{Election1} \approx \mathsf{Election2})$ expresses that each party obtained the same portion of swing voters in the second election that it got in the first election.

It is well known that the satisfaction relation of team semantics can be formalised in (existential) second-order logic when the team is encoded by an additional relation. This result gives an upper bound and a "yardstick" for the expressive power of many of the logics studied in the team semantics literature. One of the motivations for the current article is to develop an analogous yardstick of expressivity for logics over multiteams and probabilistic teams. We use a variant of existential second-order logic over two-sorted structures for this purpose whose first sort encodes the first-order structure and whose second sort consists of the closed interval $[0, 1]$ of rational numbers $\mathbb{Q}_{[0,1]}$ over which arithmetic operations of multiplication and sum can be applied. Distributions from the first sort ranging over the second sort $\mathbb{Q}_{[0,1]}$ encode probabilistic teams.

In the second part of the article we consider the complexity of model-checking in probabilistic and multiteam semantics and show that, over multiteams, very simple formulae using $\boldsymbol{x} \approx \boldsymbol{y}$ give rise to NP-hard model checking problems. This result is in drastic contrast with the influential result of Galliani and Hella [7] that inclusion atoms in the ordinary team semantics give rise to a logic equivalent with (a fragment of) the least fixed point logic and accordingly is contained in PTIME. Interestingly our reduction does not work under the slightly different probabilistic interpretation of disjunction. It is an open question whether the data-complexity of $\mathsf{FO}(\boldsymbol{x} \approx \boldsymbol{y})$ is in PTIME for the probabilistic semantics.

Previous work on probabilistic team semantics: Probabilistic versions of dependence logic (and IF-logic) have been previously studied by Galliani, Mann, Sevenster, and Sandu [5,8,19]. Moreover, Hyttinen et al. [15,16] consider so-called quantum team and measure team logics over probabilistic teams and give complete axiomatisation for them. It is worth noting, as regards to the connectives and quantifiers, our semantics is similar to the one defined by Galliani [5] and that the atoms $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z}$ and $\boldsymbol{x} \approx \boldsymbol{y}$ were introduced only later by Durand et al. [3] in the multiteam semantics context.

2 A variant of existential second-order logic with quantification over rational distributions

First-order variables are denoted by x, y, z and tuples of first-order variables by $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$. The length of the tuple $\boldsymbol{x}$ is denoted by $|\boldsymbol{x}|$, and for two tuples $\boldsymbol{x}, \boldsymbol{y}$ we denote by $\boldsymbol{x} \setminus \boldsymbol{y}$ any tuple that lists those elements of $\boldsymbol{x}$ that do not appear in $\boldsymbol{y}$. By $\mathrm{Var}(\boldsymbol{x})$ we denote the set of variables that appear in the variable sequence $\boldsymbol{x}$. A *vocabulary* τ is a set of relation symbols and function symbols with prescribed arities. We mostly denote relation symbols by R and function symbols by f, and the related arities by $\mathrm{ar}(R)$ and $\mathrm{ar}(f)$, respectively. A vocabulary is *relational* (resp., *functional*) if it consists of only relation (resp., function) symbols. Similarly, a structure is *relational* (resp., *functional*) if it is defined over a relational (resp., functional) vocabulary. We let Var_1 and Var_2 denote disjoint countable sets of first-order and function variables (with prescribed arities), respectively. The set of rational numbers in the closed interval $[0, 1]$ is denoted by $\mathbb{Q}_{[0,1]}$. Given a finite set A, a function $f\colon A \to \mathbb{Q}_{[0,1]}$ is called a *(probability) distribution* if $\sum_{s\in A} f(s) = 1$. In addition, the empty function is a *distribution*.

A relational τ-structure is a tuple $\mathfrak{A} = (A, (R_i^{\mathfrak{A}})_{R_i\in\tau})$, where A is a nonempty set and each $R_i^{\mathfrak{A}}$ is a relation on A (i.e., $R_i^{\mathfrak{A}} \subseteq A^{\mathrm{ar}(R_i)}$). In this paper, we consider structures that enrich finite relational τ-structures by adding $\mathbb{Q}_{[0,1]}$ as a second domain sort and functions that map tuples from A to $\mathbb{Q}_{[0,1]}$.

Definition 3. *Let τ and σ be a relational and a functional vocabulary, respectively. A probabilistic $\tau \cup \sigma$-structure is a tuple*

$$\mathfrak{A} = (A, \mathbb{Q}_{[0,1]}, (R_i^{\mathfrak{A}})_{R_i\in\tau}, (f_i^{\mathfrak{A}})_{f_i\in\sigma}),$$

where A (i.e. the domain of $\mathfrak{A}$) is a finite nonempty set, each $R_i^{\mathfrak{A}}$ is a relation on A (i.e., a subset of $A^{\mathrm{ar}(R_i)}$), and each $f_i^{\mathfrak{A}}$ is a probability distribution from $A^{\mathrm{ar}(f_i)}$ to $\mathbb{Q}_{[0,1]}$ (i.e., a function such that $\sum_{\boldsymbol{a}\in A^{\mathrm{ar}(f_i)}} f_i(\boldsymbol{a}) = 1$).

Note that if f is a 0-ary function symbol, then $f^{\mathfrak{A}}$ is the constant 1. Next, we define a variant of functional existential second-order logic with numerical terms ($\mathsf{ESOf}_{\mathbb{Q}}$) that is designed to describe properties of the above probabilistic structures. As first-order terms we have only first-order variables. For a set σ of function symbols, the set of numerical σ-terms i is defined via the following grammar:

$$i ::= f(\boldsymbol{x}) \mid i \times i \mid \mathrm{SUM}_{\boldsymbol{x}} i,$$

where $\boldsymbol{x}$ is a tuple of first-order variables from Var_1 and $f \in \sigma$. The value of a numerical term i in a structure $\mathfrak{A}$ under an assignment s is denoted by $[i]_s^{\mathfrak{A}}$. We have the following rules for the numerical terms:

$$[f(\boldsymbol{x})]_s^{\mathfrak{A}} := f^{\mathfrak{A}}(s(\boldsymbol{x})), \qquad [i \times j]_s^{\mathfrak{A}} := [i]_s^{\mathfrak{A}} \cdot [j]_s^{\mathfrak{A}},$$
$$[\mathrm{SUM}_{\boldsymbol{x}} i(\boldsymbol{x}, \boldsymbol{y})]_s^{\mathfrak{A}} := \sum_{\boldsymbol{a}\in A^{|\boldsymbol{x}|}} [i(\boldsymbol{a}, \boldsymbol{y})]_s^{\mathfrak{A}},$$

where $\cdot$ and $\sum$ are the multiplication and sum of rational numbers, respectively. In this context, $i(\boldsymbol{x}, \boldsymbol{y})$ is a numerical term over variables in $\boldsymbol{x}$ and $\boldsymbol{y}$. Note that, in the semantics of $\mathrm{SUM}_{\boldsymbol{x}} i$ the tuple $\boldsymbol{y}$ could be empty. Furthermore let τ be a relational vocabulary. The set of $\tau \cup \sigma$-formulae of $\mathsf{ESOf}_{\mathbb{Q}}$ is defined via the following grammar:

$$\phi ::= x = y \mid x \neq y \mid i = j \mid i \neq j \mid R(\boldsymbol{x}) \mid \neg R(\boldsymbol{x}) \mid \phi \wedge \phi \mid \phi \vee \phi \mid \exists x \phi \mid \forall x \phi \mid \exists f \psi,$$

where i is a numerical σ-term, $R \in \tau$ is a relation symbol, $f \in \mathrm{Var}_2$ is a function variable, $\boldsymbol{x}$ is a tuple of first-order variables, and ψ is a $\tau \cup (\sigma \cup \{f\})$-formula of $\mathsf{ESOf}_{\mathbb{Q}}$. Note that the syntax of $\mathsf{ESOf}_{\mathbb{Q}}$ admits of only first-order subformulae to appear in negation normal form. This restriction however does not restrict the expressiveness of the language.

Semantics of $\mathsf{ESOf}_{\mathbb{Q}}$ is defined via probabilistic structures and assignments analogous to first-order logic; note that first-order variables are always assigned to a value in A whereas functions map tuples from A to $\mathbb{Q}_{[0,1]}$. In addition to the clauses of first-order logic, we have the following semantical clauses:

$$\mathfrak{A} \models_s i = j \Leftrightarrow [i]_s^{\mathfrak{A}} = [j]_s^{\mathfrak{A}}, \qquad \mathfrak{A} \models_s i \neq j \Leftrightarrow [i]_s^{\mathfrak{A}} \neq [j]_s^{\mathfrak{A}},$$
$$\mathfrak{A} \models_s \exists f \phi \Leftrightarrow \mathfrak{A}[h/f] \models_s \phi \text{ for some probability distribution } h \colon A^{\mathrm{ar}(f)} \to \mathbb{Q}_{[0,1]},$$

where $\mathfrak{A}[h/f]$ denotes the expansion of $\mathfrak{A}$ that interprets f to h.

Note that the property of h being a probability distribution can be expressed by the formula $\mathrm{SUM}_{\boldsymbol{x}} h(\boldsymbol{x}) = 1$ suggesting that it is not vital whether the restriction to probability distributions is in the semantics or not; in this case, however, $\mathbb{Q}_{[0,1]}$ would not suffice as a second sort and the set of (non-negative) rationals should be used instead. Furthermore, for relating $\mathsf{ESOf}_{\mathbb{Q}}$ to our probabilistic team logic this assumption is essential. Recall that the constant 1 is defined by the unique 0-ary function and is thus essentially included in the language. In structures of size at least 2, the constant 0 can be defined by $g(y)$ by the use of the formula

$$\exists g \exists x \exists y \, (x \neq y \wedge g(x) = 1).^{6}$$

In order to get some idea of the expressive power of $\mathsf{ESOf}_{\mathbb{Q}}$, we note that the uniformity of a distribution f can be expressed with

$$\phi(f) := \forall \boldsymbol{x} \boldsymbol{y} (f(\boldsymbol{x}) = 0 \vee f(\boldsymbol{y}) = 0 \vee f(\boldsymbol{x}) = f(\boldsymbol{y})).$$

Furthermore, let $\frac{p}{q}$ be an arbitrary rational number. For $k \leq p$, denote by $\hat{k}$ the length $\log(p+1)$ bit sequence that encodes k, and denote by $\boldsymbol{y}_{\hat{k}}$ the variable sequence obtained from $\hat{k}$ by replacing bits 0 and 1 with variables y_0 and y_1, respectively. For $l \leq q-p$, define $\boldsymbol{z}_{\hat{l}}$ analogously in terms of bit sequences of length $\log((q-p)+1)$. For instance, $(y_0, y_0, \ldots, y_0, y_0)$ is $\boldsymbol{y}_{\hat{0}}$ and $(y_0, y_0, \ldots, y_0, y_1)$ is $\boldsymbol{y}_{\hat{1}}$. Let $E := \{\boldsymbol{y}_{\hat{k}} \boldsymbol{z}_{\hat{0}} \mid 1 \leq k \leq p\} \cup \{\boldsymbol{y}_{\hat{0}} \boldsymbol{z}_{\hat{l}} \mid 1 \leq l \leq q-p\}$. Note that E is not part of the syntax of our logic, but is used as a shorthand in the following formula. Now $i(\boldsymbol{x}) = \frac{p}{q}$ can be described by

$$\phi_{\frac{p}{q}}(\boldsymbol{x}) := \exists y_0 y_1 \exists f \big(y_0 \neq y_1 \wedge \bigwedge_{\boldsymbol{y}\boldsymbol{z}, \boldsymbol{y}'\boldsymbol{z}' \in E} f(\boldsymbol{y}\boldsymbol{z}) = f(\boldsymbol{y}'\boldsymbol{z}') \wedge$$
$$\forall \boldsymbol{y}\boldsymbol{z} (\boldsymbol{y}\boldsymbol{z} \notin E \leftrightarrow f(\boldsymbol{y}\boldsymbol{z}) = 0) \wedge i(\boldsymbol{x}) = \mathrm{SUM}_{\boldsymbol{y}} \boldsymbol{y}\boldsymbol{z}_{\hat{0}} \big).$$

Note that, by construction, E is finite, and consequently $\phi_{\frac{p}{q}}$ is an $\mathsf{ESOf}_{\mathbb{Q}}$-formula.

[6] $f(\boldsymbol{x}) = 0$ is always false for probability distributions f in structures of size 1.

3 Probabilistic Team Semantics

In this section, we present probabilistic team semantics for probabilistic team logics. Before going to probabilistic semantics, we quickly review the basics of (multi)team semantics.

3.1 Team and Multiteam Semantics

Syntactically, team logics are extensions of first-order logic FO given by the grammar rules:

$$\phi ::= x = y \mid x \neq y \mid R(\boldsymbol{x}) \mid \neg R(\boldsymbol{x}) \mid (\phi \wedge \phi) \mid (\phi \vee \phi) \mid \exists x\phi \mid \forall x\phi,$$

where $\boldsymbol{x}$ is a tuple of first-order variables.

Let D be a finite set of first-order variables and A be a nonempty set. A function $s\colon D \to A$ is called an *assignment*. The set D is the *domain* of s, and the set A the *codomain* of s. For a variable x and $a \in A$, the assignment $s(a/x)\colon D \cup \{x\} \to A$ is equal to s with the exception that $s(a/x)(x) = a$.

A *team* is a finite set of assignments with a common domain and codomain. Let X be a team with codomain A, and let $F\colon X \to \mathcal{P}(A)\setminus\{\emptyset\}$ be a function. We denote by $X[A/x]$ the modified team $\{s(a/x) \mid s \in X, a \in A\}$, and by $X[F/x]$ the team $\{s(a/x) \mid s \in X, a \in F(s)\}$. Let $\mathfrak{A}$ be a τ-structure and X a team with codomain A, then we say that X is a team of $\mathfrak{A}$.

Definition 4. *Let $\mathfrak{A}$ be a τ-structure and X a team of $\mathfrak{A}$. The satisfaction relation $\models_X$ for first-order logic is defined as follows:*

$$\begin{aligned}
\mathfrak{A} \models_X x = y &\Leftrightarrow \text{for all } s \in X : s(x) = s(y)\\
\mathfrak{A} \models_X x \neq y &\Leftrightarrow \text{for all } s \in X : s(x) \neq s(y)\\
\mathfrak{A} \models_X R(\boldsymbol{x}) &\Leftrightarrow \text{for all } s \in X : s(\boldsymbol{x}) \in R^{\mathfrak{A}}\\
\mathfrak{A} \models_X \neg R(\boldsymbol{x}) &\Leftrightarrow \text{for all } s \in X : s(\boldsymbol{x}) \notin R^{\mathfrak{A}}\\
\mathfrak{A} \models_X (\psi \wedge \theta) &\Leftrightarrow \mathfrak{A} \models_X \psi \text{ and } \mathfrak{A} \models_X \theta\\
\mathfrak{A} \models_X (\psi \vee \theta) &\Leftrightarrow \mathfrak{A} \models_Y \psi \text{ and } \mathfrak{A} \models_Z \theta \text{ for some } Y, Z \subseteq X \text{ s.t. } Y \cup Z = X\\
\mathfrak{A} \models_X \forall x\psi &\Leftrightarrow \mathfrak{A} \models_{X[A/x]} \psi\\
\mathfrak{A} \models_X \exists x\psi &\Leftrightarrow \mathfrak{A} \models_{X[F/x]} \psi \text{ holds for some } F\colon X \to \mathcal{P}(A) \setminus \{\emptyset\}.
\end{aligned}$$

Multiteams are multiset analogues of teams. Below we give a short introduction to multiteam semantics, as defined by Durand et al. [3], adjusted to the notation used later in this paper.

Definition 5. *A* multiset *is a function $\mathcal{A}\colon A \to \mathbb{N}$. The set $\{a \in A \mid \mathcal{A}(a) \geq 1\}$ is the set of* elements *of the multiset $\mathcal{A}$, and $\mathcal{A}(a)$ is the multiplicity of the element a. A* multiteam *is a multiset $\mathcal{X}\colon X \to \mathbb{N}$ where X is a team. The domain (codomain, resp.) of $\mathcal{X}$ is defined as the domain (codomain, resp.) of X.*

For a multiset $\mathcal{A}$, we define the *canonical set representative* $[\mathcal{A}]_{\text{cset}}$ of $\mathcal{A}$ by

$$[\mathcal{A}]_{\text{cset}} := \{\, (a, i) \mid a \in A, i \in \mathbb{N},\ 0 < i \leq \mathcal{A}(a) \,\}.$$

We say that a multiset $\mathcal{A}$ is a submultiset of a multiset $\mathcal{B}$, and write $\mathcal{A} \subseteq \mathcal{B}$, if and only if $[\mathcal{A}]_{\text{cset}} \subseteq [\mathcal{B}]_{\text{cset}}$. We write $\mathcal{A} = \mathcal{B}$ if and only if both $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{B} \subseteq \mathcal{A}$ hold. The *disjoint union* $\mathcal{A} \uplus \mathcal{B}$ of $\mathcal{A}$ and $\mathcal{B}$ is the function $A \cup B \to \mathbb{N}$ defined by

$$\mathcal{A} \uplus \mathcal{B}(s) := \begin{cases} \mathcal{A}(s) + \mathcal{B}(s) & \text{if } s \in A \text{ and } s \in B,\\ \mathcal{A}(s) & \text{if } s \in A \text{ and } s \notin B,\\ \mathcal{B}(s) & \text{if } s \notin A \text{ and } s \in B. \end{cases}$$

We write $|\mathcal{A}|$ to denote the size of $\mathcal{A}$, i.e., $|\mathcal{A}| := \sum_{a\in A} \mathcal{A}(a)$. Let $\mathcal{X}$ be a multiteam, A a finite set, and $F\colon [\mathcal{X}]_{\text{cset}} \to \mathcal{P}(A) \setminus \emptyset$ a function. We denote by $\mathcal{X}[A/x]$ the modified multiteam defined as

$$\biguplus_{s\in X} \biguplus_{a\in A} \{\big(s(a/x), \mathcal{X}(s)\big)\}.$$

By $\mathcal{X}[F/x]$ we denote the multiteam defined as

$$\biguplus_{s\in X} \biguplus_{1\le i\le \mathcal{X}(s)} \{\big(s(b/x), 1\big) \mid b \in F\big((s,i)\big)\}.$$

A multiteam $\mathcal{X}$ *over* $\mathfrak{A}$ is a multiteam with codomain A. We are now ready to define multiteam semantics for first-order logic. In the semantical clauses below, we use the lax semantics for existential quantifier and strict semantics for disjunction as defined by Durand et. al [3].

Definition 6 (Multiteam semantics). *Let $\mathfrak{A}$ be a τ-structure and $\mathcal{X}$ a multiteam over $\mathfrak{A}$. The satisfaction relation $\models_{\mathcal{X}}$ is defined as follows:*

$\mathfrak{A} \models_{\mathcal{X}} x = y \quad \Leftrightarrow$ *for all* $s \in X:$ *if* $\mathcal{X}(s) \ge 1$ *then* $s(x) = s(y)$
$\mathfrak{A} \models_{\mathcal{X}} x \neq y \quad \Leftrightarrow$ *for all* $s \in X:$ *if* $\mathcal{X}(s) \ge 1$ *then* $s(x) \neq s(y)$
$\mathfrak{A} \models_{\mathcal{X}} R(\boldsymbol{x}) \quad \Leftrightarrow$ *for all* $s \in X:$ *if* $\mathcal{X}(s) \ge 1$ *then* $s(\boldsymbol{x}) \in R^{\mathfrak{A}}$
$\mathfrak{A} \models_{\mathcal{X}} \neg R(\boldsymbol{x}) \Leftrightarrow$ *for all* $s \in X:$ *if* $\mathcal{X}(s) \ge 1$ *then* $s(\boldsymbol{x}) \notin R^{\mathfrak{A}}$
$\mathfrak{A} \models_{\mathcal{X}} (\psi \wedge \theta) \Leftrightarrow \mathfrak{A} \models_{\mathcal{X}} \psi$ *and* $\mathfrak{A} \models_{\mathcal{X}} \theta$
$\mathfrak{A} \models_{\mathcal{X}} (\psi \vee \theta) \Leftrightarrow \mathfrak{A} \models_{\mathcal{Y}} \psi$ *and* $\mathfrak{A} \models_{\mathcal{Z}} \theta$ *for some multisets* $\mathcal{Y}, \mathcal{Z} \subseteq \mathcal{X}$ *s.t.* $\mathcal{X} = \mathcal{Y} \uplus \mathcal{Z}$.
$\mathfrak{A} \models_{\mathcal{X}} \forall x\psi \quad \Leftrightarrow \mathfrak{A} \models_{\mathcal{X}[A/x]} \psi$
$\mathfrak{A} \models_{\mathcal{X}} \exists x\psi \quad \Leftrightarrow \mathfrak{A} \models_{\mathcal{X}[F/x]} \psi$ *holds for some function* $F\colon [\mathcal{X}]_{\text{cset}} \to \mathcal{P}(A) \setminus \emptyset$.

Using the counting measure, a multiteam $\mathcal{X}$ can be seen as a probability distribution over X; let $p_{\mathcal{X}}$ denote the distribution defined as follows:

$$p_{\mathcal{X}}(s) := \frac{\mathcal{X}(s)}{\sum_{t\in X} \mathcal{X}(t)}.$$

Conversely, every probability distribution p over a team X can be seen as a class $\mathcal{C}(p)$ of multiteams with that distribution as its counting measure:

$$\mathcal{C}(p) := \{\mathcal{X} \mid p_{\mathcal{X}} = p\}.$$

Teams in $\mathcal{C}(p)$ can be seen as discrete approximations of the probability distribution p. In the section below we abandon the discrete approach and device team based logics that take probability distributions of teams as primitive. Intuitively, the semantics of these probabilistic logics is defined such that satisfaction of formulae with respect to probabilistic teams and their *large enough* discrete approximations coincide.

3.2 Probabilistic teams

Let D be a finite set of variables, A a finite set, and X a finite set of assignments from D to A. A *probabilistic team* $\mathbb{X}$ is a distribution $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$. We call D and A the variable domain and value domain of $\mathbb{X}$, respectively. Let $\mathfrak{A}$ be a τ-structure and $\mathbb{X}$ a probabilistic team such that the domain of $\mathfrak{A}$ is the value domain of $\mathbb{X}$. Then we say that $\mathbb{X}$ is a probabilistic team of $\mathfrak{A}$. In the following, we will define two notations $\mathbb{X}[A/x]$ and $\mathbb{X}[F/x]$, similar to $\mathcal{X}[A/x]$

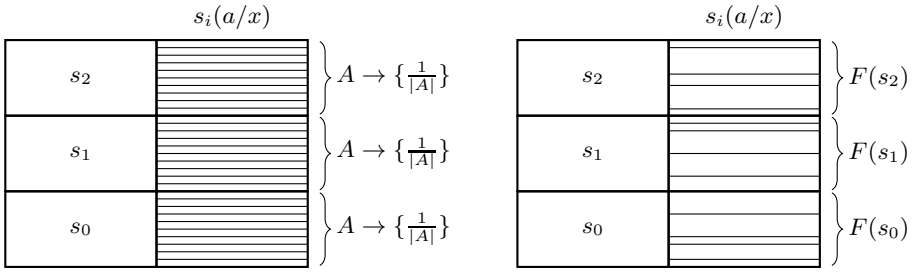


Fig. 1. Intuition of universal quantification of x (i.e., the set $\mathbb{X}[A/x]$) is depicted on the left side. The intuition of existential quantification of x (i.e., the set $\mathbb{X}[F/x]$) is depicted of the right side. The height of a box labelled by an assignment corresponds to the assignments probability. E.g., on left the probability of s_0 is $\frac{1}{3}$ whereas the probability of $s_0(a/x)$ (for any $a \in A$) is $\frac{1}{3|A|}$.

and $\mathcal{X}[F/x]$ of the previous section, in order to define the semantics of the universal and existential quantification of variables. Their intuition is depicted in Figure 1.

Let $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ be a probabilistic team, A a finite set, p_A the set of all probability distributions $d\colon A \to \mathbb{Q}_{[0,1]}$, and $F\colon X \to p_A$ a function. We denote by $\mathbb{X}[A/x]$ the probabilistic team $X[A/x] \to \mathbb{Q}_{[0,1]}$ such that

$$\mathbb{X}[A/x](s(a/x)) = \sum_{\substack{t \in X \\ t(a/x)=s(a/x)}} \mathbb{X}(t) \cdot \frac{1}{|A|},$$

for each $a \in A$ and $s \in X$. Note that if x is a fresh variable then the righthand side of the above equation is simply $\mathbb{X}(s) \cdot \frac{1}{|A|}$. By $\mathbb{X}[F/x]$ we denote the probabilistic team $X[A/x] \to \mathbb{Q}_{[0,1]}$ defined such that

$$\mathbb{X}[F/x](s(a/x)) = \sum_{\substack{t \in X \\ t(a/x)=s(a/x)}} \mathbb{X}(t) \cdot F(t)(a),$$

for each $a \in A$ and $s \in X$. Again, if x is a fresh variable, $\sum$ can be dropped from the above equation.

Let $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ and $\mathbb{Y}\colon Y \to \mathbb{Q}_{[0,1]}$ be probabilistic teams with common variable and value domains, and let $k \in \mathbb{Q}_{[0,1]}$ be a rational number. We denote by $\mathbb{X} \sqcup_k \mathbb{Y}$ the k-scaled union of $\mathbb{X}$ and $\mathbb{Y}$, that is, the probabilistic team $\mathbb{X} \sqcup_k \mathbb{Y}\colon X \cup Y \to \mathbb{Q}_{[0,1]}$ defined such that for each $s \in X \cup Y$,

$$(\mathbb{X} \sqcup_k \mathbb{Y})(s) := \begin{cases} k \cdot \mathbb{X}(s) + (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in X \text{ and } s \in Y, \\ k \cdot \mathbb{X}(s) & \text{if } s \in X \text{ and } s \notin Y, \\ (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in Y \text{ and } s \notin X. \end{cases}$$

We may now define probabilistic team semantics for first-order formulae.

Definition 7. *Let $\mathfrak{A}$ be a probabilistic τ-structure over a finite domain A, and $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ a probabilistic team of $\mathfrak{A}$. The satisfaction relation $\models_{\mathbb{X}}$ for first-order logic is defined as follows:*

$$\begin{aligned}
&\mathfrak{A} \models_{\mathbb{X}} x = y && \Leftrightarrow \textit{for all } s \in X : \textit{if } \mathbb{X}(s) > 0, \textit{ then } s(x) = s(y) \\
&\mathfrak{A} \models_{\mathbb{X}} x \neq y && \Leftrightarrow \textit{for all } s \in X : \textit{if } \mathbb{X}(s) > 0, \textit{ then } s(x) \neq s(y) \\
&\mathfrak{A} \models_{\mathbb{X}} R(\boldsymbol{x}) && \Leftrightarrow \textit{for all } s \in X : \textit{if } \mathbb{X}(s) > 0, \textit{ then } s(\boldsymbol{x}) \in R^{\mathfrak{A}} \\
&\mathfrak{A} \models_{\mathbb{X}} \neg R(\boldsymbol{x}) && \Leftrightarrow \textit{for all } s \in X : \textit{if } \mathbb{X}(s) > 0, \textit{ then } s(\boldsymbol{x}) \notin R^{\mathfrak{A}}
\end{aligned}$$

$\mathfrak{A} \models_{\mathbb{X}} (\psi \wedge \theta) \Leftrightarrow \mathfrak{A} \models_{\mathbb{X}} \psi$ *and* $\mathfrak{A} \models_{\mathbb{X}} \theta$
$\mathfrak{A} \models_{\mathbb{X}} (\psi \vee \theta) \Leftrightarrow \mathfrak{A} \models_{\mathbb{Y}} \psi$ *and* $\mathfrak{A} \models_{\mathbb{Z}} \theta$ *for some* $\mathbb{Y}, \mathbb{Z}, k$ *s.t.* $\mathbb{Y} \sqcup_k \mathbb{Z} = \mathbb{X}$
$\mathfrak{A} \models_{\mathbb{X}} \forall x\psi \quad \Leftrightarrow \mathfrak{A} \models_{\mathbb{X}[A/x]} \psi$
$\mathfrak{A} \models_{\mathbb{X}} \exists x\psi \quad \Leftrightarrow \mathfrak{A} \models_{\mathbb{X}[F/x]} \psi$ *holds for some* $F\colon X \to p_A$.

Next we define the semantics of probabilistic atoms considered in this paper: marginal identity and probabilistic independence atom. They were first introduced in the context of multiteam semantics in [3]. We define $|\mathbb{X}_{\boldsymbol{x}=\boldsymbol{a}}|$ where $\boldsymbol{x}$ is a tuple of variables and $\boldsymbol{a}$ a tuple of values, as the rational

$$|\mathbb{X}_{\boldsymbol{x}=\boldsymbol{a}}| := \sum_{\substack{s(\boldsymbol{x})=\boldsymbol{a} \\ s\in X}} \mathbb{X}(s).$$

If ϕ is some first-order formula, then $|\mathbb{X}_{\phi}|$ is defined analogously as the total sum of weights of those assignments in X that satisfy ϕ.

If $\boldsymbol{x}, \boldsymbol{y}$ are variable sequences of length k, then $\boldsymbol{x} \approx \boldsymbol{y}$ is a *marginal identity atom* with the following semantics:

$$\mathfrak{A} \models_{\mathbb{X}} \boldsymbol{x} \approx \boldsymbol{y} \Leftrightarrow |\mathbb{X}_{\boldsymbol{x}=\boldsymbol{a}}| = |\mathbb{X}_{\boldsymbol{y}=\boldsymbol{a}}| \text{ for each } \boldsymbol{a} \in A^k \tag{1}$$

Note that the equality $|\mathbb{X}_{\boldsymbol{x}=\boldsymbol{a}}| = |\mathbb{X}_{\boldsymbol{y}=\boldsymbol{a}}|$ in (1) can be equivalently replaced with $|\mathbb{X}_{\boldsymbol{x}=\boldsymbol{a}}| \leq |\mathbb{X}_{\boldsymbol{y}=\boldsymbol{a}}|$ since the tuples $\boldsymbol{a}$ range over A^k. Due to this alternative formulation, marginal identity atoms were in [3] called probabilistic inclusion atoms.

If $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ are variable sequences, then $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z}$ is a *probabilistic conditional independence atom* with the satisfaction relation defined as

$$\mathfrak{A} \models_{\mathbb{X}} \boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z} \tag{2}$$

if for all $s\colon \mathrm{Var}(\boldsymbol{xyz}) \to A$ it holds that

$$|\mathbb{X}_{\boldsymbol{xy}=s(\boldsymbol{xy})}| \cdot |\mathbb{X}_{\boldsymbol{xz}=s(\boldsymbol{xz})}| = |\mathbb{X}_{\boldsymbol{xyz}=s(\boldsymbol{xyz})}| \cdot |\mathbb{X}_{\boldsymbol{x}=s(\boldsymbol{x})}|.$$

The logic $\mathsf{FO}(\perp\!\!\!\perp_{\mathrm{c}}, \approx)$ is now defined as the extension of FO with marginal identity and probabilistic conditional independence atoms. The following two examples demonstrate the expressivity of $\mathsf{FO}(\perp\!\!\!\perp_{\mathrm{c}}, \approx)$.

Example 8. The formula $\forall \boldsymbol{yx} \approx \boldsymbol{y}$ states that the probabilities for $\boldsymbol{x}$ are uniformly distributed over all value sequences of length $|\boldsymbol{x}|$.

Example 9. We define a formula $\phi(x) := \exists\alpha\beta\psi(x, \alpha, \beta)$ which expresses that the weight of a predicate $P(x)$ is at least two times that of a predicate $Q(x)$ in a probabilistic team over x. The subformula ψ in ϕ is given as

$$\psi := x\alpha \approx x\beta \wedge \alpha = 0 \leftrightarrow \beta \neq 0 \wedge \exists\gamma_P\gamma_Q\theta(x, \alpha, \beta, \gamma_P\gamma_Q), \text{ where} \tag{3}$$
$$\theta := \big((P(x) \wedge \alpha = 0) \leftrightarrow \gamma_P = 0\big) \wedge Q(x) \to \gamma_Q = 0 \wedge \gamma_P \approx \gamma_Q \tag{4}$$

Now $\mathfrak{A} \models_{\mathbb{X}} \phi(x) \iff |\mathbb{X}_{P(x)}| \geq 2 \cdot |\mathbb{X}_{Q(x)}|$ for any $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ where α, β, γ_P, and γ_Q are not in the variable domain of $\mathbb{X}$. The first two conjuncts in (3) indicate that the values of α must be chosen so that $\frac{1}{2} \cdot |\mathbb{Y}_{P(x)}| = |\mathbb{Y}_{P(x)\wedge\alpha=0}|$. Where $\mathbb{Y}$ denotes the team obtained form $\mathbb{X}$ by evaluating the quantifiers $\exists\alpha\beta$. The first conjunct in (4) implies that $|\mathbb{Z}_{P(x)\wedge\alpha=0}| = |\mathbb{Z}_{\gamma_P=0}|$ and the second that $|\mathbb{Z}_{Q(x)}| \leq |\mathbb{Z}_{\gamma_Q=0}|$, where $\mathbb{Z}$ is team obtained

from $\mathbb{Y}$ by evaluating the quantifiers $\exists\gamma_P\gamma_Q$. The third conjunct in (4) then indicates that $|\mathbb{Z}_{\gamma_P=0}| = |\mathbb{Z}_{\gamma_Q=0}|$. Put together, we have that

$$|\mathbb{X}_{Q(x)}| \stackrel{*}{=} |\mathbb{Z}_{Q(x)}| \leq |\mathbb{Z}_{\gamma_Q=0}| = |\mathbb{Z}_{\gamma_P=0}| = |\mathbb{Z}_{P(x)\wedge\alpha=0}| \stackrel{*}{=} |\mathbb{Y}_{P(x)\wedge\alpha=0}| = \frac{1}{2}|\mathbb{Y}_{P(x)}| \stackrel{*}{=} \frac{1}{2}|\mathbb{X}_{P(x)}|.$$

The equations $\stackrel{*}{=}$ follow from the fact that quantification of fresh variables do not change the distribution of assignments with respect to the old variables.

Our next example relates probabilistic conditional independence atoms and marginal identity atoms to Bayesian networks. A Bayesian network is a directed acyclic graph whose nodes represent random variables and edges represent dependency relations between these random variables. The applicability of Bayesian networks is grounded in the notion of conditional independence as the conditional independence relations encoded in the topology of such a network enable a factorization of the underlying joint probability distribution. Next we survey the possibility of refining Bayesian networks with information obtained from $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ formulae.

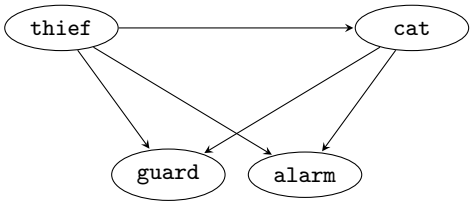
Example 10. Consider the Bayesian network $\mathbb{G}$ in Fig. 2 that models beliefs about house safety using four Boolean random variables. We note that the awakening of `guard` or `alarm` is conditioned upon both the presence of `thief` and `cat`. Furthermore, `cat` depends on `thief`, and `guard` and `alarm` are independent given `thief` and `cat`. From the network we obtain that the joint probability distribution for these variables can be factorized as

$$P(t, c, g, a) = P(t) \cdot P(c \mid t) \cdot P(g \mid t, c) \cdot P(a \mid t, c) \tag{5}$$

where, e.g., t abbreviates either $\texttt{thief} = T$ or $\texttt{thief} = F$, and $P(c \mid t)$ is the probability of c given t. The joint probability distribution (i.e., a team $\mathbb{X}$) can hence be stored as in Fig. 2.

Let t, c, g, a now refer to random variables `thief`, `cat`, `guard`, `alarm`. The dependence structure of a Bayesian network is characterized by the so-called local directed Markov property stating that each variable is conditionally independent of its non-descendants given its parents. For our network $\mathbb{G}$ the only non-trivial independence given by this property is $g \perp\!\!\!\perp_{tc} a$. Hence a probabilistic team $\mathbb{X}$ over t, c, g, a factorizes according to (5) iff $\mathbb{X}$ satisfies $g \perp\!\!\!\perp_{tc} a$. In this situation knowledge on various $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ formulae can further improve the decomposition of the joint probability distribution. Assume we have information suggesting that we may safely assume an $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ formula ϕ on $\mathbb{X}$:

- $\phi := t = F \to g = F$ indicates that `guard` never raises alert in absence of `thief`. In this case the two bottom rows of the conditional probability distribution for `guard` become superfluous.
- $\phi := tca \approx tcg$ indicates that `alarm` and `guard` have the same reliability for any given value of `thief` and `cat`. Consequently, the conditional distributions for `alarm` and `guard` are equal and one of the them can be removed.
- $\phi := \exists x(tcg \approx tcx \wedge tcga \perp\!\!\!\perp y \wedge x = T \leftrightarrow ay = TT)$ entails that `guard` is of a factor $P(y = T)$ less sensitive to raise alert than `alarm` for any given `thief` and `cat`. The formula introduces a fresh free variable y, independent of any random variable in $\mathbb{G}$, and such that that the probability of $ay = TT$ equals the probability of $g = T$ given tc. The latter property is expressed by introducing an auxiliary distribution for x. In this case it suffices to store the conditional probability table for `alarm` and the probability $P(y = T)$.



guard		
thief,cat	T	F
TT	0.8	0.2
TF	0.7	0.3
FT	0	1
FF	0	1

thief	
T	F
0.1	0.9

cat		
thief	T	F
T	0.1	0.9
F	0.6	0.4

alarm		
thief,cat	T	F
TT	0.9	0.1
TF	0.8	0.2
FT	0.1	0.9
FF	0	1

Fig. 2. Bayesian network $\mathbb{G}$ and its related conditional distributions

Next we connect probabilistic teams to multiteams. Denote by Prob the mapping that transforms a multiteam to its corresponding probabilistic team, i.e., given a multiteam $\mathcal{X}$, $\mathrm{Prob}(\mathcal{X})$ is the probabilistic team $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ such that

$$\mathbb{X}(s) = \frac{\mathcal{X}(s)}{\sum_{s' \in X} \mathcal{X}(s')}.$$

It follows from the definitions that Prob preserves the truth condition for marginal identity and probabilistic independence atoms.

Proposition 11. *Let ϕ be a marginal identity or a probabilistic independence atom, let $\mathcal{X}$ be a multiteam of a structure $\mathfrak{A}$, and let $\mathbb{X}$ be a probabilistic team of $\mathfrak{A}$ such that $\mathbb{X} = \mathrm{Prob}(\mathcal{X})$. Then $\mathfrak{A} \models_{\mathcal{X}} \phi \iff \mathfrak{A} \models_{\mathbb{X}} \phi$.*

The restriction of a team X to V is defined as $X \upharpoonright V = \{s \upharpoonright V \mid s \in X\}$ where $s \upharpoonright V$ denotes the restriction of the assignment s to V. The restriction of a probabilistic team $\mathbb{X}$ to V is then defined as the probabilistic team $Q\colon X \upharpoonright V \to \mathbb{Q}_{[0,1]}$ where

$$Q(s) = \sum_{s' \upharpoonright V = s} P(s').$$

The following locality property indicates that satisfaction of $\phi \in \mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ is determined by the restriction of a probabilistic team to the free variables of ϕ. The set of *free variables* $\mathrm{Fr}(\phi)$ of a formula $\phi \in \mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ is defined recursively as in first-order logic with the addition that for probabilistic independence and marginal identity atoms ϕ, $\mathrm{Fr}(\phi)$ consists of all variables that appear in ϕ.

Proposition 12 (Locality). *Let $\phi(\boldsymbol{x}) \in \mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ be a formula with free variables from $\boldsymbol{x} = (x_1, \ldots, x_n)$. Then for all structures $\mathfrak{A}$ and probabilistic teams $\mathbb{X} : X \to \mathbb{Q}_{[0,1]}$ where $\{x_1, \ldots, x_n\} \subseteq V \subseteq \mathrm{Dom}(X)$, $\mathfrak{A} \models_{\mathbb{X}} \phi \iff \mathfrak{A} \models_{\mathbb{X} \upharpoonright V} \phi$.*

Proof. For first-order atoms the claim is immediate. Furthermore, it is easy to check that the same holds for the atoms $\boldsymbol{x} \approx \boldsymbol{y}$ and $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z}$ (for multiteam semantics this has been discussed in [3]).

Assume then that $\phi := \psi \vee \theta$, and that the claim holds for ψ and θ. Note first that for any probabilistic teams $\mathbb{X}$ and $\mathbb{Y}$ with common variable and value domains a simple calculation shows that

$$(\mathbb{X} \sqcup_k \mathbb{Y}) \upharpoonright V = \mathbb{X} \upharpoonright V \sqcup_k \mathbb{Y} \upharpoonright V. \tag{6}$$

Suppose that $\mathfrak{A} \models_{\mathbb{X}} \phi$. Then there are k, $\mathbb{Y}$, and $\mathbb{Z}$ such that $\mathbb{X} = \mathbb{Y} \sqcup_k \mathbb{Z}$, $\mathfrak{A} \models_{\mathbb{Y}} \psi$, and $\mathfrak{A} \models_{\mathbb{Z}} \theta$. By the induction assumption, it holds that $\mathfrak{A} \models_{\mathbb{Y}\restriction V} \psi$ and $\mathfrak{A} \models_{\mathbb{Z}\restriction V} \theta$. Now by (6), $\mathfrak{A} \models_{\mathbb{X}\restriction V} \phi$. The converse implication is proved analogously. The proof is similar for the cases $\phi := \exists x\psi$ and $\phi := \forall x\psi$. □

4 Translation from $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ to $\mathsf{ESOf}_{\mathbb{Q}}$

In this section, we show that any formula in $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ can be equivalently expressed as a sentence of $\mathsf{ESOf}_{\mathbb{Q}}$ that has exactly one free function variable for encoding probabilistic teams. The following lemma will be used to facilitate the translation. This lemma has been shown by Durand et al. [3] for multiteams and accordingly, by Proposition 11, it holds for probabilistic teams as well. The lemma entails that each probabilistic independence atom in $\phi \in \mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ can be assumed to be either of the form $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z}$ or of the form $\boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{y}$ for pairwise disjoint tuples $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$.

Lemma 13. *[3] Let $\mathfrak{A}$ be a structure and $\mathbb{X}$ a probabilistic team over $\mathfrak{A}$. Then*

(i) $\mathfrak{A} \models_{\mathbb{X}} \boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z} \quad \Leftrightarrow \quad \mathfrak{A} \models_{\mathbb{X}} \big(\boldsymbol{y} \setminus \boldsymbol{x} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z} \setminus \boldsymbol{x}\big)$,
(ii) $\mathfrak{A} \models_{\mathbb{X}} \boldsymbol{y} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z} \quad \Leftrightarrow \quad \mathfrak{A} \models_{\mathbb{X}} \big(\boldsymbol{y} \setminus \boldsymbol{z} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{z} \setminus \boldsymbol{y}\big) \wedge \big(\boldsymbol{y} \cap \boldsymbol{z} \perp\!\!\!\perp_{\boldsymbol{x}} \boldsymbol{y} \cap \boldsymbol{z}\big)$.

Theorem 14. *For every formula $\phi(\boldsymbol{x}) \in \mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ with free variables from $\boldsymbol{x} = (x_1, \ldots, x_n)$ there exists a formula $\phi^*(f) \in \mathsf{ESOf}_{\mathbb{Q}}$ with exactly one free function variable f such that for all structures $\mathfrak{A}$ and nonempty probabilistic teams $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$,*

$$\mathfrak{A} \models_{\mathbb{X}} \phi(\boldsymbol{x}) \iff (\mathfrak{A}, f_{\mathbb{X}}) \models \phi^*(f),$$

where $f_{\mathbb{X}} : A^n \to \mathbb{Q}_{[0,1]}$ is the probability distribution such that $f_{\mathbb{X}}(s(\boldsymbol{x})) = \mathbb{X}(s)$ for all $s \in X$.

Proof. We give a compositional translation * from $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ to $\mathsf{ESOf}_{\mathbb{Q}}$. For a subsequence $\boldsymbol{x}_i$ of $\boldsymbol{x}$, we denote by $\boldsymbol{x}_i^c$ a sequence $\boldsymbol{x} \setminus \boldsymbol{x}_i$, and by $\boldsymbol{x}(\boldsymbol{y}/\boldsymbol{x}_i)$ a sequence obtained from $\boldsymbol{x}$ by replacing $\boldsymbol{x}_i$ pointwise with $\boldsymbol{y}$.

If $\phi(\boldsymbol{x})$ is of the form $R(\boldsymbol{x_0})$, then $\phi^*(f) := \forall \boldsymbol{x}\big(f(\boldsymbol{x}) = 0 \vee R(\boldsymbol{x}_0)\big)$.
If $\phi(\boldsymbol{x})$ is of the form $\neg R(\boldsymbol{x_0})$, then $\phi^*(f) := \forall \boldsymbol{x}\big(f(\boldsymbol{x}) = 0 \vee \neg R(\boldsymbol{x}_0)\big)$.
If $\phi(\boldsymbol{x})$ is $\boldsymbol{x}_0 \approx \boldsymbol{x}_1$, then $\phi^*(f) := \forall \boldsymbol{z}\, \mathrm{SUM}_{\boldsymbol{x}_0^c} f(\boldsymbol{x}(\boldsymbol{z}/\boldsymbol{x}_0)) = \mathrm{SUM}_{\boldsymbol{x}_1^c} f(\boldsymbol{x}(\boldsymbol{z}/\boldsymbol{x}_1))$.
If $\phi(\boldsymbol{x})$ is $\boldsymbol{x}_1 \perp\!\!\!\perp_{\boldsymbol{x}_0} \boldsymbol{x}_2$ where $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2$ are disjoint, then $\phi^*(f) := \forall \boldsymbol{x_0 x_1 x_2}$
$\mathrm{SUM}_{(\boldsymbol{x}_0\boldsymbol{x}_1)^c} f(\boldsymbol{x}) \times \mathrm{SUM}_{(\boldsymbol{x}_0\boldsymbol{x}_2)^c} f(\boldsymbol{x}) = \mathrm{SUM}_{(\boldsymbol{x}_0\boldsymbol{x}_1\boldsymbol{x}_2)^c} f(\boldsymbol{x}) \times \mathrm{SUM}_{\boldsymbol{x}_0^c} f(\boldsymbol{x})$.
If $\phi(\boldsymbol{x})$ is of the form $\boldsymbol{x}_1 \perp\!\!\!\perp_{\boldsymbol{x}_0} \boldsymbol{x}_1$ where $\boldsymbol{x}_0, \boldsymbol{x}_1$ are disjoint, then
$\phi^*(f) := \forall \boldsymbol{x}_0 \boldsymbol{x}_1 \big(\mathrm{SUM}_{(\boldsymbol{x}_0\boldsymbol{x}_1)^c} f(\boldsymbol{x}) = 0 \vee \mathrm{SUM}_{(\boldsymbol{x}_0\boldsymbol{x}_1)^c} f(\boldsymbol{x}) = \mathrm{SUM}_{\boldsymbol{x}_0^c} f(\boldsymbol{x})\big)$.
If $\phi(\boldsymbol{x})$ is of the form $\psi_0(\boldsymbol{x}) \wedge \psi_1(\boldsymbol{x})$, then $\phi^*(f) := \psi_0^*(f) \wedge \psi_1^*(f)$.
If $\phi(\boldsymbol{x})$ is of the form $\psi_0(\boldsymbol{x}) \vee \psi_1(\boldsymbol{x})$, then $\phi^*(f) := \psi_0^*(f) \vee \psi_1^*(f)$

$$\vee \Big(\exists pghk\big(\forall \boldsymbol{x} \forall y (y = l \vee y = r \vee (p(y) = 0 \wedge k(\boldsymbol{x}, y) = 0)) \tag{7}$$

$$\wedge \forall \boldsymbol{x}(k(\boldsymbol{x}, l) = g(\boldsymbol{x}) \times p(l) \wedge k(\boldsymbol{x}, r) = h(\boldsymbol{x}) \times p(r)) \tag{8}$$

$$\wedge \forall \boldsymbol{x}\, (\mathrm{SUM}_y k(\boldsymbol{x}, y) = f(\boldsymbol{x})) \wedge \psi_0^*(g) \wedge \psi_1^*(h)\big)\Big). \tag{9}$$

If $\phi(\boldsymbol{x})$ is $\exists y \psi(\boldsymbol{x}, y)$, then $\phi^*(f) := \exists g\big((\forall \boldsymbol{x}\, \mathrm{SUM}_y g(\boldsymbol{x}, y) = f(\boldsymbol{x})) \wedge \psi^*(g)\big)$.
If $\phi(\boldsymbol{x})$ is of the form $\forall y \psi(\boldsymbol{x}, y)$, then $\phi^*(f) :=$
$\exists g\big(\forall \boldsymbol{x}(\forall y \forall z g(\boldsymbol{x}, y) = g(\boldsymbol{x}, z) \wedge \mathrm{SUM}_y g(\boldsymbol{x}, y) = f(\boldsymbol{x})) \wedge \psi^*(g)\big)$.

The claim now follows via a straightforward induction on the structure of the formula. The cases for first-order and dependency atoms, and likewise for conjunctions, follow directly from the semantical clauses.

The case for disjunctions requires a bit more care. First note that l (left) and r (right) denote distinct constant symbols than can be defined by $\exists l \exists r\, l \neq r$ in the beginning of the translation *. Recall that a probabilistic team $\mathbb{X}$ satisfies a disjunction $(\phi \vee \psi)$ if and only if $\mathbb{X}$ satisfies either ϕ or ψ, or there exists two nonempty probabilistic teams $\mathbb{Y}$ and $\mathbb{Z}$ and a ratio $q \in \mathbb{Q}_{[0,1]}$ such that $\mathbb{Y}$ satisfies ϕ, $\mathbb{Z}$ satisfies ψ, and, for each assignment s, it holds that $\mathbb{X}(s) = q \cdot \mathbb{Y}(s) + (1-q) \cdot \mathbb{Z}(s)$. In the translation, we encode the value of q by $p(l)$ and $(1-q)$ by $p(r)$. Line (7) expresses that p is such a function. We use $k(s(\boldsymbol{x}), l)$ and $k(s(\boldsymbol{x}), r)$ to encode the values of $q \cdot \mathbb{Y}(s)$ and $(1-q) \cdot \mathbb{Z}(s)$, respectively. Lines (7) and (8) together express that k is such a function. Finally, the first part of line (9) expresses that $\forall s : \mathbb{X}(s) = q \cdot \mathbb{Y}(s) + (1-q) \cdot \mathbb{Z}(s)$, whereas the latter part expresses that $\mathbb{Y}$ satisfies ϕ, $\mathbb{Z}$ satisfies ψ.

The cases for the quantifiers follow directly by the semantical clauses. □

5 Translation from $\mathsf{ESOf}_{\mathbb{Q}}$ to $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$

In this section, we construct a translation from $\mathsf{ESOf}_{\mathbb{Q}}$ to $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$. The proof utilises the observation that independence atoms and marginal identity atoms can be used to express multiplication and SUM in $\mathbb{Q}_{[0,1]}$, respectively. The translation then relates $\mathsf{ESOf}_{\mathbb{Q}}$ sentences in a certain normal form, presented in Lemma 16, to open $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ formulae. Before this, we start by stating a lemma which expresses that existential quantification of a constant probability distribution d can be characterised in $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$. Given a probabilistic team $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$, a tuple $\boldsymbol{x} = (x_1, \ldots, x_n)$ of fresh variables, and a probability distribution $d\colon A^n \to \mathbb{Q}_{[0,1]}$, we denote by $\mathbb{X}[d/\boldsymbol{x}]$ the probabilistic team $\mathbb{Y}$ where $\mathbb{Y}(s(\boldsymbol{a}/\boldsymbol{x})) = \mathbb{X}(s) \cdot d(\boldsymbol{a})$ for all $s \in X$.

Lemma 15. *Let $\phi(\boldsymbol{x}) := \exists \boldsymbol{y}(\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \wedge \psi(\boldsymbol{x}, \boldsymbol{y}))$ be a $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$-formula with free variables from $\boldsymbol{x} = (x_1, \ldots, x_n)$. Then for all structures $\mathfrak{A}$ and probabilistic teams $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ where $\{x_1, \ldots, x_n\} \subseteq \mathrm{Dom}(X)$,*

$$\mathfrak{A} \models_{\mathbb{X}} \phi \iff \mathfrak{A} \models_{\mathbb{X}[d/\boldsymbol{y}]} \psi \text{ for some } d\colon A^{|\boldsymbol{y}|} \to \mathbb{Q}_{[0,1]}.$$

Proof. By the locality principle (Prop. 12) $\mathfrak{A} \models_{\mathbb{X}} \phi$ if and only if $\mathfrak{A} \models_{\mathbb{X} \upharpoonright \{x_1, \ldots, x_n\}} \phi$. Likewise it is straightforward to check that, for $d\colon A^{|\boldsymbol{y}|} \to \mathbb{Q}_{[0,1]}$

$$\mathfrak{A} \models_{\mathbb{X}[d/\boldsymbol{y}]} \psi \text{ if and only if } \mathfrak{A} \models_{\mathbb{X} \upharpoonright \{x_1, \ldots, x_n\}[d/\boldsymbol{y}]} \psi,$$

since $\mathbb{X}[d/\boldsymbol{y}] \upharpoonright \{x_1, \ldots, x_n, \boldsymbol{y}\} = \mathbb{X} \upharpoonright \{x_1, \ldots, x_n\}[d/\boldsymbol{y}]$. Accordingly, we may assume without loss of generality, that $\mathrm{Dom}(X) = \{x_1, \ldots, x_n\}$.

Now $\mathfrak{A} \models_{\mathbb{X}} \phi$ iff there is a function $F\colon X \to p_A$ such that $\mathfrak{A} \models_{\mathbb{Y}} \boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \wedge \psi(\boldsymbol{x}, \boldsymbol{y})$ where $\mathbb{Y} := \mathbb{X}[F/\boldsymbol{y}]$. Furthermore,

$$\mathfrak{A} \models_{\mathbb{Y}} \boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \text{ iff } |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{y}=s(\boldsymbol{x})\boldsymbol{a}}| = |\mathbb{Y}_{\boldsymbol{x}=s(\boldsymbol{x})}| \cdot |\mathbb{Y}_{\boldsymbol{y}=\boldsymbol{a}}| \text{ for all } s \in X \text{ and } \boldsymbol{a} \in A^n.$$

Since $\mathrm{Dom}(X) = \{x_1, \ldots, x_n\}$, the right-hand side of the above is equivalent to

$$\mathbb{X}(s) \cdot F(s)(\boldsymbol{a}) = \mathbb{X}(s) \cdot |\mathbb{Y}_{\boldsymbol{y}=\boldsymbol{a}}| \text{ for all } s \in X \text{ and } \boldsymbol{a} \in A^n.$$

This is equivalent with saying that $\mathbb{X}[F/\boldsymbol{y}] = \mathbb{X}[d/\boldsymbol{y}]$ for some distribution $d\colon A^n \to \mathbb{Q}_{[0,1]}$. □

Before proceeding to the translation, we construct the following normal form for $\mathsf{ESOf}_{\mathbb{Q}}$ sentences.

Lemma 16. *Every $\mathsf{ESOf}_{\mathbb{Q}}$ sentence ϕ is equivalent to a sentence ϕ^* of the form $\exists \boldsymbol{f} \forall \boldsymbol{x} \theta$, where θ is quantifier-free and such that its second sort identity atoms are of the form $f_i(\boldsymbol{uv}) = f_j(\boldsymbol{u}) \times f_k(\boldsymbol{v})$ or $f_i(\boldsymbol{u}) = \mathrm{SUM}_{\boldsymbol{v}} f_j(\boldsymbol{uv})$ for distinct f_i, f_j, f_k such that at most one of them is not quantified.*

Proof. First we define for each second sort term $i(\boldsymbol{x})$ a special formula θ_i defined recursively using fresh function symbols f_i as follows:

- If $i(\boldsymbol{u})$ is $g(\boldsymbol{u})$ where g is a function symbol, then θ_i is defined as $f_i(\boldsymbol{u}) = g(\boldsymbol{u})$. (We may intepret $g(\boldsymbol{u})$ as $\mathrm{SUM}_{\emptyset} g(\boldsymbol{u})$).
- If $i(\boldsymbol{uv})$ is $j(\boldsymbol{u}) \times k(\boldsymbol{v})$, then θ_i is defined as $\theta_j \wedge \theta_k \wedge f_i(\boldsymbol{uv}) = f_j(\boldsymbol{u}) \times f_k(\boldsymbol{v})$.
- If $i(\boldsymbol{u})$ is $\mathrm{SUM}_{\boldsymbol{v}} j(\boldsymbol{uv})$, then θ_i is defined as $\theta_j \wedge f_i(\boldsymbol{u}) = \mathrm{SUM}_{\boldsymbol{v}} f_j(\boldsymbol{uv})$.

The translation $\phi \mapsto \phi^*$ then proceeds recursively on the structure of ϕ.

(i) If ϕ is $i(\boldsymbol{u}) = j(\boldsymbol{v})$, then ϕ^* is defined as $\exists \boldsymbol{f}(f_i(\boldsymbol{u}) = f_j(\boldsymbol{v}) \wedge \theta_i \wedge \theta_j)$ where $\boldsymbol{f}$ is lists the function symbols f_k for each subterm k of i or j. If ϕ is $i(\boldsymbol{u}) \neq j(\boldsymbol{v})$, the translation is analogous.

(ii) If ϕ is an atom or negated atom of the first sort, then $\phi^* := \phi$.

(iii) If ϕ is $\psi_0 \circ \psi_1$ where $\circ \in \{\vee, \wedge\}$, ψ_0^* is $\exists \boldsymbol{f}_0 \forall \boldsymbol{x}_0 \theta_0$, and ψ_1^* is $\exists \boldsymbol{f}_1 \forall \boldsymbol{x}_1 \theta_1$, then ϕ_1^* is defined as $\exists \boldsymbol{f}_0 \boldsymbol{f}_1 \forall \boldsymbol{x}_0 \boldsymbol{x}_1 (\theta_0 \circ \theta_1)$.

(iv) If ϕ is $\exists y \psi$ where ψ^* is $\exists \boldsymbol{f} \forall \boldsymbol{x} \theta$, then ϕ^* is defined as $\exists g \exists \boldsymbol{f} \forall \boldsymbol{x} \forall y (g(y) = 0 \vee \theta)$.

(v) If ϕ is $\forall y \psi$ where ψ^* is $\exists \boldsymbol{f} \forall \boldsymbol{x} \theta$, then ϕ^* is defined as

$$\exists \boldsymbol{f}^* \exists \boldsymbol{f}_{\mathrm{id}} \exists d \forall y y' \forall \boldsymbol{x} (d(y) = d(y') \wedge \bigwedge_{f^* \in \boldsymbol{f}^*} \mathrm{SUM}_{\boldsymbol{x}} f^*(y, \boldsymbol{x}) = d(y) \wedge \theta^*)$$

where $\boldsymbol{f}^*$ is obtained from $\boldsymbol{f}$ by replacing each f from $\boldsymbol{f}$ with f^* such that $\mathrm{ar}(f^*) = \mathrm{ar}(f) + 1$, $\boldsymbol{f}_{\mathrm{id}}$ introduces new function symbol for each multiplication in θ, and θ^* is obtained by replacing all second sort identities α of the form $f_i(\boldsymbol{uv}) = f_j(\boldsymbol{u}) \times f_k(\boldsymbol{v})$ with

$$f_\alpha(y, \boldsymbol{uv}) = d(y) \times f_i^*(y, \boldsymbol{uv}) \wedge f_\alpha(y, \boldsymbol{uv}) = f_j^*(y, \boldsymbol{u}) \times f_k^*(y, \boldsymbol{v})$$

and $f_i(\boldsymbol{u}) = \mathrm{SUM}_{\boldsymbol{v}} f_j(\boldsymbol{uv})$ with $f_i^*(y, \boldsymbol{u}) = \mathrm{SUM}_{\boldsymbol{v}} f_j^*(y, \boldsymbol{uv})$

(vi) If ϕ is $\exists f \psi$ where ψ^* is $\exists \boldsymbol{f} \forall \boldsymbol{x} \theta$, then ϕ^* is defined as $\exists f \psi^*$.

It is straightforward to check that ϕ^* is of the correct form and equivalent to ϕ. What happens in (v) is that instead of guessing for all y some distribution f_y with arity $\mathrm{ar}(f)$, we guess a single distribution f^* with arity $\mathrm{ar}(f)+1$ such that $f^*(y, \boldsymbol{u}) = \frac{1}{|A|} \cdot f_y(\boldsymbol{u})$ where A is the underlying domain of the structure. This is described by the existential quantification of a unary uniform distribution d such that for all fixed y, $\mathrm{SUM}_{\boldsymbol{u}} f^*(y, \boldsymbol{u})$ is $d(y)$. Then note that $f_y(\boldsymbol{u}) = g_y(\boldsymbol{u}') \cdot h_y(\boldsymbol{u}'')$ iff $\frac{1}{|A|} \cdot f^*(y, \boldsymbol{u}) = g^*(y, \boldsymbol{u}') \cdot h^*(y, \boldsymbol{u}'')$ iff $d(y) \cdot f^*(y, \boldsymbol{u}) = g^*(y, \boldsymbol{u}') \cdot h^*(y, \boldsymbol{u}'')$. For identities over SUM, the reasoning is analogous. □

Theorem 17. *Let $\phi(p) \in \mathsf{ESOf}_{\mathbb{Q}}$ be a sentence of the form $\exists \boldsymbol{f} \forall \boldsymbol{x} \theta$ where θ is a quantifier-free $\mathsf{FOf}_{\mathbb{Q}}$ formula in which each second sort equality atom is of the form $f_i(\boldsymbol{x}_i) = f_j(\boldsymbol{x}_j) \times f_k(\boldsymbol{x}_k)$ or $f_i(\boldsymbol{x}_i) = \mathrm{SUM}_{\boldsymbol{x}_k} f_j(\boldsymbol{x}_k \boldsymbol{x}_j)$ for distinct f_i, f_j, f_k from $\{f_1, \ldots, f_n\} \cup \{p\}$. Then there is a formula $\Phi \in \mathsf{FO}(\perp\!\!\!\perp_{\mathrm{c}}, \approx)$ such that for all structures $\mathfrak{A}$ and probabilistic teams $\mathbb{X} := p^{\mathfrak{A}}$,*

$$\mathfrak{A} \models_{\mathbb{X}} \Phi \iff (\mathfrak{A}, p) \models \phi.$$

Proof. We define Φ as

$$\Phi := \forall \boldsymbol{x} \exists \boldsymbol{y}_1 \dots \boldsymbol{y}_n (\Theta \wedge \Psi)$$

where $\boldsymbol{x} = (x_1, \dots, x_m)$, $\boldsymbol{y}_i$ are sequences of variables of length $\mathrm{ar}(f_i)$, Θ is a compositional translation from θ, and

$$\Psi := \bigwedge_{i=1}^{n} \boldsymbol{x}\boldsymbol{y}_1 \dots \boldsymbol{y}_{i-1} \perp\!\!\!\perp \boldsymbol{y}_i. \tag{10}$$

By Lemma 15 it suffices to show that for all distributions $f_1, \dots, f_n$, subsets $M \subseteq A^m$, and probabilistic teams $\mathbb{Y} = \mathbb{X}[M/\boldsymbol{x}][f_1/\boldsymbol{y}_1] \dots [f_n/\boldsymbol{y}_n]$,

$$\mathfrak{A} \models_{\mathbb{Y}} \Theta \text{ iff } (\mathfrak{A}, p, f_1, \dots, f_n) \models \theta(\boldsymbol{a}) \text{ for all } \boldsymbol{a} \in M. \tag{11}$$

We show the claim by structural induction on the construction of Θ.

1. If θ is an atom of the first sort, it clearly suffices to let $\Theta = \theta$.
2. Assume θ is of the form $f_i(\boldsymbol{x}_i) = f_j(\boldsymbol{x}_j) \times f_k(\boldsymbol{x}_k)$. Then Θ is defined as

$$\Theta := \exists \alpha \beta \Big((\alpha = 0 \leftrightarrow \boldsymbol{x}_i = \boldsymbol{y}_i) \wedge (\beta = 0 \leftrightarrow \boldsymbol{x}_j \boldsymbol{x}_k = \boldsymbol{y}_j \boldsymbol{y}_k) \wedge \boldsymbol{x}\alpha \approx \boldsymbol{x}\beta \Big).$$

   Assume that $(\mathfrak{A}, p, f_1, \dots, f_n) \models \theta(\boldsymbol{a})$ for any given $\boldsymbol{a} \in M$. Then we have $f_i(\boldsymbol{a}_i) = f_j(\boldsymbol{a}_j) \cdot f_k(\boldsymbol{a}_k)$. We define functions $F_\alpha, F_\beta \colon \mathbb{Y} \to \{0,1\}$ so that $F_\alpha(s) = 0$ iff $s(\boldsymbol{x}_i) = s(\boldsymbol{y}_i)$, and $F_\beta(s) = 0$ iff $s(\boldsymbol{x}_j \boldsymbol{x}_k) = s(\boldsymbol{x}_j \boldsymbol{x}_k)$. It suffices to show that $\mathfrak{A} \models_{\mathbb{Z}} \boldsymbol{x}\alpha \approx \boldsymbol{x}\beta$ where $\mathbb{Z} = \mathbb{Y}[F_\alpha/\alpha][F_\beta/\beta]$. By the construction of $\mathbb{Z}$, we have $|\mathbb{Z}_{\boldsymbol{x}\alpha=\boldsymbol{a}0}| = |\mathbb{Z}_{\boldsymbol{x}\boldsymbol{y}_i=\boldsymbol{a}\boldsymbol{a}_i}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot f_i(\boldsymbol{a}_i)$. Similarly, and using the hypothesis, we have $|\mathbb{Z}_{\boldsymbol{x}\beta=\boldsymbol{a}0}| = |\mathbb{Z}_{\boldsymbol{x}\boldsymbol{y}_j\boldsymbol{y}_k=\boldsymbol{a}\boldsymbol{a}_j\boldsymbol{a}_k}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot f_j(\boldsymbol{a}_j) \cdot f_k(\boldsymbol{a}_k) = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot f_i(\boldsymbol{a}_i)$. Furthermore, since we have $|\mathbb{Z}_{\boldsymbol{x}\alpha=\boldsymbol{a}1}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot (1 - f_i(\boldsymbol{a}_i)) = |\mathbb{Z}_{\boldsymbol{x}\beta=\boldsymbol{a}1}|$, it follows that $\mathfrak{A} \models_{\mathbb{Y}} \Theta$.
   Assume $\mathfrak{A} \models_{\mathbb{Y}} \Theta$, and let $\mathbb{Z}$ be the extension of $\mathbb{Y}$ to α, β where $Z_{\alpha=0} = Z_{\boldsymbol{x}_i=\boldsymbol{y}_i}$ and $\mathbb{Z}_{\beta=0} = \mathbb{Z}_{\boldsymbol{x}_j\boldsymbol{x}_k=\boldsymbol{y}_j\boldsymbol{y}_k}$. Then $\mathfrak{A} \models_{\mathbb{Z}} \boldsymbol{x}\alpha \approx \boldsymbol{x}\beta$ since $|\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot f_i(\boldsymbol{a}_i) = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot |\mathbb{Y}_{\boldsymbol{y}_i=\boldsymbol{a}_i}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{y}_i=\boldsymbol{a}\boldsymbol{a}_i}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{x}_i=\boldsymbol{a}\boldsymbol{y}_i}| = |\mathbb{Z}_{\boldsymbol{x}\alpha=\boldsymbol{a}0}| = |\mathbb{Z}_{\boldsymbol{x}\beta=\boldsymbol{a}0}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{x}_j\boldsymbol{x}_k=\boldsymbol{a}\boldsymbol{y}_j\boldsymbol{y}_k}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{y}_j\boldsymbol{y}_k=\boldsymbol{a}\boldsymbol{a}_j\boldsymbol{a}_k}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot |\mathbb{Y}_{\boldsymbol{y}_j=\boldsymbol{a}_j}| \cdot |\mathbb{Y}_{\boldsymbol{y}_k=\boldsymbol{a}_k}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot f_j(\boldsymbol{a}_j) \cdot f_k(\boldsymbol{a}_k)$ for all $\boldsymbol{a} \in M$.
3. Assume θ is of the form $f_i(\boldsymbol{x}_i) = \mathrm{SUM}_{\boldsymbol{x}_k} f_j(\boldsymbol{x}_k \boldsymbol{x}_j)$. We define Θ as

$$\Theta := \exists \alpha \beta \Big((\alpha = 0 \leftrightarrow \boldsymbol{x}_i = \boldsymbol{y}_i) \wedge (\beta = 0 \leftrightarrow \boldsymbol{x}_j = \boldsymbol{y}_j) \wedge \boldsymbol{x}\alpha \approx \boldsymbol{x}\beta \Big).$$

   Assume that $(\mathfrak{A}, p, f_1, \dots, f_n) \models \theta(\boldsymbol{a})$ for any given $\boldsymbol{a} \in M$. Then $f_i(\boldsymbol{a}_i) = \mathrm{SUM}_{\boldsymbol{x}_k} f_j(\boldsymbol{x}_k \boldsymbol{x}_j)$. We define functions $F_\alpha, F_\beta \colon \mathbb{Y} \to \{0,1\}$ such that $F_\alpha(s) = 0$ iff $s(\boldsymbol{x}_i) = s(\boldsymbol{y}_i)$, and $F_\beta(s) = 0$ iff $s(\boldsymbol{x}_j) = s(\boldsymbol{y}_j)$. Then $\mathfrak{A} \models_{\mathbb{Z}} \boldsymbol{x}\alpha \approx \boldsymbol{x}\beta$ because $|\mathbb{Z}_{\boldsymbol{x}\alpha=\boldsymbol{a}0}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{x}_i=\boldsymbol{a}\boldsymbol{y}_i}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{y}_i=\boldsymbol{a}\boldsymbol{a}_i}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot f_i(\boldsymbol{a}_i) = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot \mathrm{SUM}_{\boldsymbol{x}_k} f_j(\boldsymbol{x}_k \boldsymbol{a}_j) = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot |\mathbb{Y}_{\boldsymbol{y}_j=\boldsymbol{a}_j}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{y}_j=\boldsymbol{a}\boldsymbol{a}_j}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{x}_j=\boldsymbol{a}\boldsymbol{y}_j}| = |\mathbb{Z}_{\boldsymbol{x}\beta=\boldsymbol{a}0}|$. Furthermore, since $|\mathbb{Z}_{\boldsymbol{x}\alpha=\boldsymbol{a}1}| = |\mathbb{Z}_{\boldsymbol{x}\beta=\boldsymbol{a}1}|$ it follows that $\mathfrak{A} \models_{\mathbb{Y}} \Theta$.
   Assume that $\mathfrak{A} \models_{\mathbb{Y}} \Theta$, and let $\mathbb{Z}$ be the extension of $\mathbb{Y}$ to α, β where $Z_{\alpha=0} = Z_{\boldsymbol{x}_i=\boldsymbol{y}_i}$ and $\mathbb{Z}_{\beta=0} = \mathbb{Z}_{\boldsymbol{x}_j=\boldsymbol{y}_j}$. Analogously to the previous case, we obtain $\mathfrak{A} \models_{\mathbb{Z}} \boldsymbol{x}\alpha \approx \boldsymbol{x}\beta$ since $|\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot f_i(\boldsymbol{a}_i) = |\mathbb{Z}_{\boldsymbol{x}\alpha=\boldsymbol{a}0}| = |\mathbb{Z}_{\boldsymbol{x}\beta=\boldsymbol{a}0}| = |\mathbb{Y}_{\boldsymbol{x}\boldsymbol{y}_j=\boldsymbol{a}\boldsymbol{a}_j}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot |\mathbb{Y}_{\boldsymbol{y}_j=\boldsymbol{a}_j}| = |\mathbb{Y}_{\boldsymbol{x}=\boldsymbol{a}}| \cdot \mathrm{SUM}_{\boldsymbol{x}_k} f_j(\boldsymbol{a}_j)$ for all $\boldsymbol{a} \in M$.
4. Assume θ is $\theta_0 \wedge \theta_1$. Then we let $\Theta := \Theta_0 \wedge \Theta_1$, and the claim follows by a straightforward argument.
5. Assume θ is $\theta_0 \vee \theta_1$. Then we let

$$\Theta := \exists z \Big(z \perp\!\!\!\perp_{\boldsymbol{x}} z \wedge (\Theta_0 \wedge z = 0) \vee (\Theta_1 \wedge \neg z = 0) \Big).$$

Assume $(\mathfrak{A}, p, f_1, \ldots, f_n) \models \theta_0 \vee \theta_1$ for all $\boldsymbol{a} \in M$. Then we find $M_0 \cup M_1 = M$, $M_0 \cap M_1 = \emptyset$, such that $(\mathfrak{A}, p, f_1, \ldots, f_n) \models \theta_i$ for all $\boldsymbol{a} \in M_i$. We define $F : Y \to p_A$ so that $F_z(s) = c_i$ if $s(\boldsymbol{x}) \in M_i$; by c_i we denote the distribution

$$c_i(a) := \begin{cases} 1 \text{ if } a = i, \\ 0 \text{ otherwise.} \end{cases}$$

Letting $\mathbb{Z}_i = \mathbb{X}[M_i/\boldsymbol{x}][f_1/\boldsymbol{y}_1] \ldots [f_n/\boldsymbol{y}_n][c_i/z]$, it follows that $\mathbb{Z} = \mathbb{Y}[F/z] = \mathbb{Z}_0 \sqcup_k \mathbb{Z}_1$ for $k = \frac{M_0}{M}$. By the induction hypothesis $\mathfrak{A} \models_{\mathbb{Z}_i} \Theta_i$, and accordingly $\mathfrak{A} \models_{\mathbb{Z}_i} \Theta_0 \wedge z_i$. Since $\mathfrak{A} \models_{\mathbb{Z}} z \perp\!\!\!\perp_{\boldsymbol{x}} z$, we obtain by Proposition 12 that $\mathfrak{A} \models_{\mathbb{Y}} \Theta$.
Assume $\mathfrak{A} \models_{\mathbb{Y}} \Theta$, and let $F\colon Y \to p_A$ be such that $\mathfrak{A} \models_{\mathbb{Z}} z \perp\!\!\!\perp_{\boldsymbol{x}} z \wedge ((\Theta_0 \wedge z = 0) \vee (\Theta_1 \wedge \neg z = 0))$ for $\mathbb{Z} = \mathbb{Y}[F/z]$. Consequently, $\mathfrak{A} \models_{\mathbb{Z}'_0} \Theta_0$ and $\mathfrak{A} \models_{\mathbb{Z}'_1} \Theta_1$ where $k\mathbb{Z}'_0 = \mathbb{Z}_{z=0}$ and $(1-k)\mathbb{Z}'_1 = \mathbb{Z}_{z=1}$ for $k = |\mathbb{Z}_{z=0}|$. Since $\mathbb{Z}$ satisfies $z \perp\!\!\!\perp_{\boldsymbol{x}} z$, we have furthermore that either $\mathbb{Z}_{\boldsymbol{x}=\boldsymbol{a}} = \mathbb{Z}_{\boldsymbol{x}z=\boldsymbol{a}0}$ or $\mathbb{Z}_{\boldsymbol{x}=\boldsymbol{a}} = \mathbb{Z}_{\boldsymbol{x}z=\boldsymbol{a}1}$ for all $\boldsymbol{a} \in M$. This entails that $\mathbb{Z}_{z=0} = \mathbb{Z}_{\boldsymbol{x} \in M_0}$ for some $M_0 \subseteq M$. Therefore, $\mathbb{Z}'_0 = \frac{|M|}{|M_0|}(\mathbb{X}[M/\boldsymbol{x}][f_1/\boldsymbol{y}_1] \ldots [f_n/\boldsymbol{y}_n])_{\boldsymbol{x} \in M_0} = \mathbb{X}[M_0/\boldsymbol{x}][f_1/\boldsymbol{y}_1] \ldots [f_n/\boldsymbol{y}_n]$. By the induction hypothesis, we then obtain $(\mathfrak{A}, p, f_1, \ldots, f_n) \models \theta_0$ for all $\boldsymbol{a} \in M_0$, and by analogous reasoning that $(\mathfrak{A}, p, f_1, \ldots, f_n) \models \theta_1$ for all $\boldsymbol{a} \in M \setminus M_0$. Consequently, $(\mathfrak{A}, p, f_1, \ldots, f_n) \models \theta$ for all $\boldsymbol{a} \in M$ which concludes the proof. □

6 Complexity of FO(≈) in multiteams vs. probabilistic teams

One of the fundamental results in logics in team semantics state that, in contrast to dependence and independence logics that correspond to existential second-order logic (accordingly, NP), the expressivity of inclusion logic equals only that of positive greatest fixed-point logic and thus PTIME over finite ordered models [6,7,20]. In this section, we consider the complexity of $\mathsf{FO}(\approx)$ that can be thought of as a probabilistic variant of inclusion logic. We present a formula $\phi \in \mathsf{FO}(\approx)$ which captures an NP-complete property of multiteams (the example works under both strict and lax semantics introduced by Durand et al. [3]). The possibility of expressing similar properties in probabilistic teams is left open. It is worth noting that our reduction is similar to the ones presented for quantifier-free dependence and independence logic formulae under team semantics [17,2] (see also the recent survey on complexity aspects of logics in team semantics [4]).

The following example relates $\mathsf{FO}(\approx)$ to the exact cover problem, a well-known NP-complete problem [9]. Given a collection $\mathcal{S}$ of subsets of a set A, an *exact cover* is a subcollection $\mathcal{S}^*$ of $\mathcal{S}$ such that each element in A is contained in exactly one subset in $\mathcal{S}^*$.

Example 18. Consider an exact cover problem over $A = \{1, 2, 3, 4\}$ and $\mathcal{S} = \{S_1 = \{1, 2, 3\}, S_2 = \{2\}, S_4 = \{1, 3, 4\}\}$. We construct a multiteam $\mathcal{X}$ as follows. The multiteam $\mathcal{X}$, depicted in Fig. 3, is a constant function mapping all assignments to 1. For each element i of a subset S_j, we create an assignment that maps `element` to 0, `set` to s_j, `left` to i, and `right` to the next element in S_j (under some ordering). Also, if $S_j = \{i\}$, then right is mapped to i. In our example case these assignments appear above the solid line of the multiteam $\mathcal{X}$ in Fig. 3. Furthermore, for each element i of A we create an assignment that maps `element` to i and all other variables to 0. The answer to the exact cover problem is then positive iff $\mathcal{X}$ satisfies

$$\phi := \mathtt{set} \neq 0 \vee (\mathtt{element} \approx \mathtt{left} \wedge \mathtt{set}, \mathtt{right} \approx \mathtt{set}, \mathtt{left}). \tag{12}$$

Note that since ϕ consists only of variables and connectives, we do not need to concern structures; we write $\mathcal{X} \models \phi$ instead of $\mathfrak{A} \models_{\mathcal{X}} \phi$. Now $\mathcal{X} \models \phi$ if and only if $\mathcal{Z} \models \mathtt{set} \neq 0$

Multiteam $\mathcal{X}$

element	set	left	right	$\mathcal{X}(s)$
0	S_1	1	2	1
0	S_1	2	3	1
0	S_1	3	1	1
0	S_2	2	2	1
0	S_3	1	3	1
0	S_3	3	4	1
0	S_3	4	1	1
1	0	0	0	1
2	0	0	0	1
3	0	0	0	1
4	0	0	0	1

Probabilistic team $\mathbb{X}$

element	set	left	right	$\mathbb{X}(s)$	$\mathbb{Y}$	$\mathbb{Z}$
0	S_1	1	2	1/10	1/2	1/2
0	S_1	2	1	1/10	1/2	1/2
0	S_2	2	3	1/10	1/2	1/2
0	S_2	3	2	1/10	1/2	1/2
0	S_3	3	1	1/10	1/2	1/2
0	S_3	1	3	1/10	1/2	1/2
1	0	0	0	1/10		1
2	0	0	0	1/10		1
3	0	0	0	1/10		1
4	0	0	0	1/10		1

Fig. 3. A multiteam $\mathcal{X}$ and a probabilistic team $\mathbb{X}$

and $\mathcal{Y} \models \texttt{element} \approx \texttt{left} \wedge \texttt{set}, \texttt{right} \approx \texttt{set}, \texttt{left}$, for some $\mathcal{Z}$, $\mathcal{Y}$ such that $\mathcal{Z} \uplus \mathcal{Y} = \mathcal{X}$. Note that any subset of the assignments above the solid line in Fig. 3 satisfy $\texttt{set} \neq 0$ and could be a priori assigned to $\mathcal{Z}$. Note also that all of the assignments below the solid line must be assigned to the team $\mathcal{Y}$. Henceforth, the conjunct $\texttt{element} \approx \texttt{left}$ forces to select assignments from above the solid line to $\mathcal{Y}$ exactly one assignment for each element of A. Then $\texttt{set}, \texttt{right} \approx \texttt{set}, \texttt{left}$ enforces that this selection either subsumes a subset S_i or does not intersect it at all. In the example case, we can select the segments that corresponds to sets S_1 and S_2.

The same reduction does not work for probabilistic teams. The probabilistic team $\mathbb{X}$ in Fig. 3 corresponds to the exact cover problem defined over $A = \{1, 2, 3\}$ and $\mathcal{S} = \{S_1 = \{1, 2\}, S_2 = \{2, 3\}, S_4 = \{3, 1\}\}$. This instance does not admit an exact cover. However, for satisfaction of (12) by $\mathbb{X}$, taking half weights of the upper part for $\mathbb{Y}$ and all the remaining weights for $\mathbb{Z}$, we have $\mathfrak{A} \models_{\mathbb{Y}} \texttt{set} \neq 0$ and $\mathfrak{A} \models_{\mathbb{Z}} \texttt{element} \approx \texttt{left} \wedge \texttt{set}, \texttt{right} \approx \texttt{set}, \texttt{left}$ where $\mathbb{X} = \mathbb{Y} \sqcup_k \mathbb{Z}$ for $k = \frac{3}{10}$.

It is straightforward to generalise the previous example to obtain the following result.

Corollary 19. *Data complexity of the quantifier-free fragment of* $\mathsf{FO}(\approx)$ *under multiteam semantics is* NP*-hard. This remains true for very simple fragments as* $\texttt{set} \neq 0 \vee (\texttt{element} \approx \texttt{left} \wedge \texttt{set}, \texttt{right} \approx \texttt{set}, \texttt{left})$ *is such a formula for which model checking is hard for* NP.

The obvious brute force algorithm gives inclusion to NP.

Theorem 20. *Data complexities of* $\mathsf{FO}(\approx)$ *and the quantifier-free fragment of* $\mathsf{FO}(\approx)$ *under multiteam semantics are* NP*-complete.*

7 Conclusion

In this article, we have initiated a systematic study of probabilistic team semantics. Some features of our semantics have been discussed in the literature but the logic $\mathsf{FO}(\perp\!\!\!\perp_c, \approx)$ has not been studied before in the probabilistic framework. Probabilistic logics with team semantics have already been applied in the context of so-called Bell's Inequalities of quantum mechanics [15]. On the other hand, our work is in part motivated by the study of implication problems of database and probabilistic dependencies. Independence logic has recently been used to give a finite axiomatisation for the implication problem of independence atoms (i.e., EMVD's) and inclusion dependencies [12]. It is an interesting open question to apply our probabilistic logic to analyse the implication problem of conditional independence statements whose exact complexity is still open [11,21].

Acknowledgements

The second author was supported by grant 3711702 of the Marsden Fund. The third author was supported by grant 308712 of the Academy of Finland. This work was supported in part by the joint grant by the DAAD (57348395) and the Academy of Finland (308099). We also thank the anonymous referees for their helpful suggestions.

References

1. Corander, J., Hyttinen, A., Kontinen, J., Pensar, J., Väänänen, J.: A logical approach to context-specific independence. Proceedings of WoLLIC 2016.
2. Durand, A., Kontinen, J., de Rugy-Altherre, N., Väänänen, J.: Tractability Frontier of Data Complexity in Team Semantics. Proc. of GandALF 2015.
3. Durand, A., Hannula, M., Kontinen, J., Meier, A., Virtema, J.: Approximation and dependence via multiteam semantics. In: Gyssens, M., Simari, G.R. (eds.) Proceedings of FoIKS 2016.
4. Durand, A., Kontinen, J., Vollmer, H.: Expressivity and complexity of dependence logic. In: Dependence Logic: Theory and Applications. Springer (2016)
5. Galliani, P.: Probabilistic dependence logic (2008), manuscript
6. Galliani, P.: Inclusion and exclusion dependencies in team semantics - on some logics of imperfect information. Annals of Pure and Applied Logic 163(1), 68–84 (2012)
7. Galliani, P., Hella, L.: Inclusion logic and fixed point logic. In: Proc. CSL. pp. 281–295 (2013)
8. Galliani, P., Mann, A.L.: Lottery semantics: A compositional semantics for probabilistic first-order logic with imperfect information. Studia Logica 101(2), 293–322 (2013)
9. Garey, M.R., Johnson, D.S.: Computers and Intractability; A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY, USA (1990)
10. Grädel, E., Väänänen, J.A.: Dependence and independence. Studia Logica 101(2), 399–410 (2013)
11. Gyssens, M., Niepert, M., Gucht, D.V.: On the completeness of the semigraphoid axioms for deriving arbitrary from saturated conditional independence statements. Information Processing Letters 114(11), 628 – 633 (2014)
12. Hannula, M., Kontinen, J.: A finite axiomatization of conditional independence and inclusion dependencies. Inf. Comput. 249, 121–137 (2016),
13. Hannula, M., Kontinen, J., Link, S.: On the finite and general implication problems of independence atoms and keys. J. Comput. Syst. Sci. 82(5), 856–877 (2016)
14. Hodges, W.: Compositional semantics for a language of imperfect information. Logic Journal of the IGPL 5(4), 539–563 (electronic) (1997)
15. Hyttinen, T., Paolini, G., Väänänen, J.: Quantum team logic and Bell's inequalities. The Review of Symbolic Logic FirstView, 1–21 (2015)
16. Hyttinen, T., Paolini, G., Väänänen, J.: A logic for arguing about probabilities in measure teams. Arch. Math. Log. 56(5-6), 475–489 (2017).
17. Kontinen, J.: Coherence and computational complexity of quantifier-free dependence logic formulas. Studia Logica 101(2), 267–291 (2013)
18. Kontinen, J., Link, S., Väänänen, J.A.: Independence in database relations. In: Proc. 20th WoLLIC. LNCS, vol. 8071, pp. 179–193. Springer (2013)
19. Sevenster, M., Sandu, G.: Equilibrium semantics of languages of imperfect information. Ann. Pure Appl. Logic 161(5), 618–631 (2010),
20. Väänänen, J.: Dependence Logic - A New Approach to Independence Friendly Logic, London Mathematical Society student texts, vol. 70, 2007.
21. Wong, S.K.M., Butz, C.J., Wu, D.: On the implication problem for probabilistic conditional independency. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 30(6), 785–805 (2000)