Irrelevance and Conditioning in First-Order Probabilistic Logic*

Daphne Koller
Computer Science Department
Gates Building 1A
Stanford, CA 94305-9010
koller@cs.stanford.edu

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120–6099
halpern@almaden.ibm.com

Abstract

First-order probabilistic logic is a powerful knowledge representation language. Unfortunately, deductive reasoning based on the standard semantics for this logic does not support certain desirable patterns of reasoning, such as indifference to irrelevant information or substitution of constants into universal rules. We show that both these patterns rely on a first-order version of *probabilistic independence*, and provide semantic conditions to capture them. The resulting insight enables us to understand the effect of conditioning on independence, and allows us to describe a procedure for determining when independencies are preserved under conditioning. We apply this procedure in the context of a sound and powerful inference algorithm for reasoning from statistical knowledge bases.

1 Introduction

First-order logic is widely recognized as being a fundamental building block in knowledge representation. As is well known, however, first-order logic does not have the necessary expressive power to deal with many situations of interest (Bacchus 1990). For example, while first-order logic allows us to express statements like "all birds fly", it does not allow us to assert in a natural way that most birds fly, or that any given bird is likely but not certain to fly. The standard way to deal with such uncertainty is via a probability distribution over the possibilities that we envision. If we are interested in doing probabilistic first-order reasoning, then we can take the set of possibilities, or *possible worlds*, to be first-order models.

Having a logic with a semantics immediately gives us a notion of deductive entailment. We deduce a formula φ from a (probabilistic) knowledge base KB if all the (probabilistic) models that satisfy KB also satisfy φ. Unfortunately, as in many logics, deductive entailment is often inadequate as an inference procedure. Many desirable patterns of reasoning, particularly those involving irrelevance, are simply not sound in all probabilistic models. Since it considers all models, entailment cannot validate these patterns.

To obtain these patterns of reasoning, we must restrict attention to those models that support them. To accomplish this, we must first understand the underlying semantics of these reasoning patterns; i.e., when do we get them and why? This theme also appears in the work on *belief networks* (Pearl 1988), which utilize independence to obtain a concise and intuitive representation of a probability distribution. The power and convenience of Bayesian networks has resulted in a resurgence of work on using probability as a knowledge representation paradigm and has led to a large number of applications. However, up to now it has mostly been applied in an *attribute-based* (essentially propositional) context, with the resulting limitations on expressive power. For example, using probabilistic first-order logic, we can easily express the effect of penicillin on a bacterial infection — $\forall x, i(\Pr(\mathit{Cured}(x) \mid \mathit{Infected}(x,i) \wedge \mathit{Bacterial}(i) \wedge \mathit{Treated}(x, \mathit{Penicillin})) \geq .8$. Using Bayesian networks, it is difficult to express binary predicates like *Infected* and *Treated*, and universal statements that hold for all patients and all infections. Note that the universal quantifier allows the rule to be applied to any bacterial infection, even one that the designer did not originally include in the knowledge base.

In this paper, we attempt to integrate the idea of probabilistic independence as a foundation for irrelevance into first-order probabilistic logic. In future work, we intend to use the resulting framework as a semantic basis for a first-order version of belief networks.

One perhaps surprising outcome of our analysis of independence is its connection to a seemingly unrelated property: substitution. In classic first-order logic, the following axiom is valid: $\forall x \varphi(x) \Rightarrow \varphi(c)$, where c is a constant. It is well known that in modal logics, this substitution property does not hold in general (Garson 1977). Assume we are told, as above, that $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x) = 0.9)$. Can we deduce that $\Pr(\mathit{Fly}(\mathit{Tweety}) \mid \mathit{Bird}(\mathit{Tweety})) = 0.9$? We can substitute *Tweety* for x if *Tweety* is a *rigid designator,* i.e., if it denotes the same thing in all possible worlds. But making *Tweety* a rigid designator can be problematic, because then the formula $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x)) = 0.9)$ is inconsistent with the observation $\mathit{Bird}(\mathit{Tweety}) \wedge \neg \mathit{Fly}(\mathit{Tweety})$

*Some of this research was performed while Daphne Koller was at U.C. Berkeley and at the IBM Almaden Research Center. Work supported in part by the Air Force Office of Scientific Research (AFSC) under Contract F49620-91-C-0080, and by a University of California President's Postdoctoral Fellowship.

(where an observation is assumed to hold with probability 1). Luckily, rigidity is not a necessary condition for such substitution to be possible. We show that the constant *Tweety* is substitutable for x if the formula $x = \mathit{Tweety}$ is independent of $\mathit{Fly}(x)$ given $\mathit{Bird}(x)$. Moreover, under a reasonable restriction on probability distributions, independence is necessary and sufficient for substitutability. Thus, substitution is simply a special case of irrelevance!

For these results to be applicable, we must believe that $\forall x(\Pr(\mathit{Fly}(x)|\mathit{Bird}(x)) = 0.9)$. When is this reasonable? A primary source for such beliefs is statistical information: if we observe that 90% of birds fly, we might well believe, for an arbitrary bird, that it flies with probability 0.9. But when does a statistical fact of this type imply the corresponding probabilistic statement? That is, does the statement "90% of birds fly" force us to believe that a particular domain element d flies with probability 0.9, when all we know is that d is a bird? Subjective Bayesians may say that there is no connection; our degree of belief $\Pr(\mathit{Fly}(d) \,|\, \mathit{Bird}(d))$ can be arbitrary. We show that, while this is true for a single bird d, it is false in general. If we know that 90% of birds fly, we *cannot* believe, for example, that $\Pr(\mathit{Fly}(d) \,|\, \mathit{Bird}(d)) = 0.7$ for every d in the domain. Our semantics forces a connection between the objective and subjective probabilities. However, in general, this connection is not strong enough to imply that $\Pr(\mathit{Fly}(d) \mid \mathit{Bird}(d)) = 0.9$ for every d: it is consistent with the information that 90% of birds fly that there is a particular domain element d whom we know to be a non-flying bird in all possible worlds. But in those models where the domain elements all "behave the same way" in an appropriate sense, we can show that the statistical statement does imply $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x)) = 0.9)$.

What do we gain by having formal semantic foundations for irrelevance and for the connection between objective and subjective probabilities? For one thing, we can check whether these semantic properties of our model are compatible with our intuitions and information we might have about our domain. If they are not, perhaps we do not want the corresponding reasoning patterns. Even more importantly, we can determine the effect of *new* information on these semantic properties, and hence on our reasoning patterns. After all, our model of the world is almost never a static one. As new information comes in, we must incorporate it into our model; in a probabilistic framework, this is done by conditioning our probability distribution on this new information. Our semantic characterization of irrelevance allows us to provide a general methodology for determining when irrelevancies are preserved in the conditioning process. We believe that the ideas underlying this methodology will be applicable in many other contexts.

We present one further application of this semantic characterization. Philosophers (starting with Reichenbach (1949)) have long been interested in the problem of going from statistical information to subjective degrees of belief. Recently, Bacchus *et al.* (1994) presented an approach where one starts with a uniform probability distribution over the set of possible worlds and conditions on a statistical knowledge base. They show that this approach, called *random worlds*, supports many of the reasoning patterns that have been viewed as desirable in both default reasoning and statistical inference. These properties include: *direct inference*, which allows us to go from a statistical statement to a conclusion about a particular individual; *preference for more specific information*; and the ability to ignore *irrelevant information*. Our results allow us to analyze the success of this approach, and to pinpoint those characteristics that made it work. In fact, we can prove that these properties hold for a large and interesting class of priors containing the uniform prior. But our results do far more than allow us to prove theorems about specific properties. They also provide us with a general algorithm for reasoning about irrelevance in first-order statistical knowledge bases. This algorithm is particularly easy to apply when the knowledge base is derived from a statistical *semantic network* (Touretzky 1986; Sowa 1991).

2 Syntax and Semantics

We briefly review the syntax and semantics of first-order probabilistic logic, as introduced by Halpern (1990). Subjective probabilities, or degrees of belief, are expressed using a modal operator Pr. The agent's degree of belief in a formula ψ (with or without free variables), written $\Pr(\psi)$, is a numeric term, which is interpreted as a number between 0 and 1. Semantically, $\Pr(\psi)$ denotes the probability of the set of worlds where ψ holds.

Starting with a vocabulary Φ and a set of variables $\mathcal{X}$, let $\mathcal{L}(\Phi \cup \mathcal{X})$ be all the formulas that we get by closing off under the standard first-order operators and applications of Pr. Note that we allow interleaving of first-order quantifiers and modal operators. In particular, a formula such as $\forall x(\Pr(\mathit{Fly}(x)) = 0.9)$ says that, for each domain element d, the agent believes that the probability that d flies is 0.9. Let $\mathcal{L}_{obj}(\Phi \cup \mathcal{X})$ be the subset of $\mathcal{L}(\Phi \cup \mathcal{X})$ consisting of *objective* formulas, i.e., those that do not mention the Pr operator.

Our logic is expressive enough to represent conditional probability expressions $\Pr(\varphi \,|\, \psi)$. We simply treat a formula of the form $\Pr(\varphi \,|\, \psi) = \alpha$ as an abbreviation for $\Pr(\varphi \wedge \psi) = \alpha \cdot \Pr(\psi)$.

We now sketch the semantics of the logic. The truth of formulas in $\mathcal{L}_{obj}(\Phi \cup \mathcal{X})$ is completely determined by a standard finite first-order structure and a valuation.[1] That is, once we provide an interpretation for all the symbols in $\Phi \cup \mathcal{X}$, the standard rules of first-order logic allow us to assign truth values to arbitrary formulas in $\mathcal{L}_{obj}$. To deal with subjective probability, we need a set of possible worlds, with a distribution over them. Thus, we take a *probability structure* to be a tuple $M = (D, W, \pi, \mu)$, where D is a finite *domain*, W is a finite set of *possible worlds*, for each world $w \in W$, $\pi(w)$ is an *interpretation of the symbols in* Φ *over the domain* D (that is, $\pi(w)$ assigns to the predicate and function symbols in Φ predicates and functions of the right arity over D), and μ is a discrete probability distribution

[1] Recall that a valuation specifies an assignment of domain elements to the free variables in the formula.

over W. The function π allows us to view each world $w \in W$ as a first-order model. Note that we assume that all the worlds in W have the same domain D although they may be associated with different interpretations over D.[2]

We assign truth values to formulas given a structure M, a world w, and a valuation v. The definition is fairly standard (see (Halpern 1990)). For example, $(M, w, v) \models \forall x(\Pr(\varphi(x) \geq \alpha)$ if for all $d \in D$, $(M, w, v[x/d]) \models \Pr(\varphi(x)) \geq \alpha)$, where the numerical term $\Pr(\varphi(x))$ is interpreted as $\mu(\{w : (M, w, v[x/d]) \models \varphi(x)\})$. As usual, we write $M \models \varphi$ if $(M, w, v) \models \varphi$ for all worlds w and valuations v. Notice that if $M \models \varphi$ then $M \models \Pr(\varphi) = 1$. Also note that the truth of a sentence $\varphi \in \mathcal{L}_{obj}$ (i.e., an objective formula with no free variables) is fully determined by a world w, so we can write $w \models \varphi$ rather than $(M, w, v) \models \varphi$.

For the remainder of the paper, to simplify notation, we assume that we are dealing with fixed finite vocabulary Φ and a fixed finite domain D. We take the set W of worlds to consist of all interpretations of the symbols in Φ over the domain D, so that $w(a)$ for $a \in \Phi$ is the interpretation of a in w. Thus, since all components in M besides μ are fixed, we write $\mu \models \varphi$ rather than $M \models \varphi$.

Combining identical worlds into one can always be done (in this context) without loss of generality. However, we often need to make an additional assumption, which relates worlds that are "essentially identical." Formally, we say that two worlds w and w' are *isomorphic* if they agree on all formulas in $\mathcal{L}_{obj}$; i.e., the agent cannot distinguish isomorphic worlds by any formula in his language. If the agent's language encodes all the agent's information about a world, then the agent cannot distinguish between isomorphic worlds. In this case, it seems reasonable to assign such worlds the same probability. We say that a probability distribution μ is *exchangeable* if $\mu(w) = \mu(w')$ whenever w and w' are isomorphic.[3] Note that the exchangeability property does *not* entail a uniform distribution on the set of worlds. Only worlds that are essentially identical are forced to be equally likely. Furthermore, note that the worlds where we just exchange the properties of two elements are isomorphic. Hence, the exchangeability assumption forces the domain elements to be interchangeable. That is, the agent can only "recognize" a domain element via its "observable" properties. The agent cannot use properties for which there is no term in its language to distinguish between two domain elements. Exchangeability implies that $\Pr(\varphi(d) \mid \psi(d))$ is the same for all d, a property which simplifies many of our results.

3 Irrelevance and Substitution

Recall that we are primarily interested in the property of irrelevance. Assume that we are given a distribution μ that satisfies $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x)) = 0.9)$. Can we conclude that it also satisfies $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x) \wedge \mathit{Red}(x)) = 0.9)$? The answer is clearly no, as it should be. It may be that red birds are far more likely to fly than regular birds. But in general, we may wish to assume that things are "as irrelevant to each other as possible;" given no information that suggests that color is relevant to flying ability in birds, we may wish to assume that it is not. What properties should μ have in order to validate the inference of $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x) \wedge \mathit{Red}(x)) = 0.9)$ from $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x)) = 0.9)$?

As we observed in the introduction, this property seems closely tied to probabilistic independence. We now formalize this intuition.

Definition 3.1: Let φ, ψ, θ be formulas, and let μ be a probability distribution. We say that φ *is independent of* θ *given* ψ *in* μ, denoted $\mu \models \mathcal{I}(\varphi, \theta | \psi)$, if $\mu \models \Pr(\varphi | \psi \wedge \theta) = \Pr(\varphi \mid \psi)$. ∎

Note that this definition also applies to formulas that have free variables. According to our semantics, this means that $\mathcal{I}(\varphi, \theta \mid \psi)$ holds for every possible valuation, i.e., that $\forall x(\Pr(\varphi(x) \mid \psi(x)) = \Pr(\varphi(x) \mid \psi(x) \wedge \theta(x)))$. The irrelevance property now follows immediately:

Proposition 3.2 : *If* $\mu \models \mathcal{I}(\varphi(x), \theta(x) \mid \psi(x))$, *then* $\mu \models \forall x(\Pr(\varphi(x) \mid \psi(x)) \in [\alpha, \beta]) \Leftrightarrow \forall x(\Pr(\varphi(x) \mid \psi(x) \wedge \theta(x)) \in [\alpha, \beta])$.

While this is fairly straightforward, it does have an unexpected application to the problem of *substitution*. The inability to substitute constants into universally quantified formulas is perhaps the most glaring deficiency of deduction in first-order probabilistic logic. After all, the ability to apply general rules to specific individuals is one of the characteristic features of first-order logic. Why do we lose this property in a probabilistic setting? The following example provides an explanation.

Example 3.3: Let $\Phi = \{\mathit{Bird}, \mathit{Tweety}\}$ and $D = \{d_1, d_2\}$, and consider the following four worlds: $w_1(\mathit{Bird}) = \{d_1\}$, $w_1(\mathit{Tweety}) = d_1$; $w_2(\mathit{Bird}) = \{d_1, d_2\}$, $w_2(\mathit{Tweety}) = d_1$; $w_3(\mathit{Bird}) = \{d_2\}$, $w_3(\mathit{Tweety}) = d_2$; $w_4(\mathit{Bird}) = \{d_1, d_2\}$, $w_4(\mathit{Tweety}) = d_2$. Suppose μ assigns probability $1/4$ to each of these worlds, and thus probability 0 to all other worlds. On the one hand, $\mu \models \forall x(\Pr(\mathit{Bird}(x)) = 3/4)$: We assign either d_1 or d_2 to x; in either case, $\mathit{Bird}(x)$ is false in precisely one of these worlds (e.g., $\mathit{Bird}(d_1)$ is false only in w_3). On the other hand, $\Pr(\mathit{Bird}(\mathit{Tweety})) = 1$: $\mathit{Bird}(\mathit{Tweety})$ holds in all four worlds.

The reason that we cannot substitute the constant *Tweety* into the universally quantified statement is that *Tweety* is not *rigid*: it has a different interpretation in each world. When interpreting a formula such as $\forall x(\Pr(\mathit{Bird}(x)) = 3/4)$, we first fix a particular domain element d, then consider the set of worlds where that particular domain element d has the property *Bird*. On the other hand, when interpreting $\Pr(\mathit{Bird}(\mathit{Tweety}))$, the denotation of the constant *Tweety* varies from world to world: it is d_1 in w_1, w_2 and d_2 in w_3, w_4. ∎

[2] As is well-known (Garson 1977), in modal logic we run into problems with *quantifying-in* if we do not make this assumption. In particular, it is difficult to give semantics to a formula such as $\exists x(\Pr(\mathit{Bird}(x)) = 0.1)$ when different worlds may have different domains.

[3] This is not the same as de Finetti's notion of exchangeability (1964), although the two notions are superficially similar.

While this inability to substitute may seem counterintuitive, as we saw in the introduction, at times it is a desirable feature. It is easy to see that we can substitute constants that are rigid designators. But if we make an assumption like exchangeability, then no constants are rigid designators (since if the constant c is interpreted as the domain element d in one world, for each d', there is an isomorphic world where c is interpreted as d'). The following result provides the key insight as to when it is safe to substitute, without assuming rigidity:

Proposition 3.4 : *(a) If μ is a distribution such that $\mu \models \forall x(\Pr(\varphi(x) \mid \psi(x) \wedge x = c) \in [\alpha, \beta])$ then $\mu \models \Pr(\varphi(c) \mid \psi(c)) \in [\alpha, \beta]$.*
(b) If μ is an exchangeable distribution, then $\mu \models \forall x(\Pr(\varphi(x) \mid \psi(x) \wedge x = c) = \Pr(\varphi(c) \mid \psi(c)))$.[4]

Intuitively, this tells us that substituting c is closely related to conditioning: substitution is essentially conditioning on the additional information that "x is also called c". Clearly, we can substitute if this additional information that $x = c$ is *irrelevant*. That is, we can substitute c for x in $\forall x(\Pr(\varphi(x) \mid \psi(x)) \in [\alpha, \beta])$ if $\Pr(\varphi(x) \mid \psi(x) \wedge x = c) = \Pr(\varphi(x) \mid \psi(x))$ for each x. But as we argued above, irrelevance is characterized by probabilistic independence. We can therefore apply Proposition 3.2 with $x = c$ as $\theta(x)$; combined with Proposition 3.4, we get:

Corollary 3.5: *Let $\mu \models \mathcal{I}(\varphi(x), (x = c) \mid \psi(x))$.*
(a) $\mu \models \forall x(\Pr(\varphi(x) \mid \psi(x)) \in [\alpha, \beta]) \Rightarrow \Pr(\varphi(c) \mid \psi(c)) \in [\alpha, \beta])$.
(b) If μ is exchangeable, then $\mu \models \forall x(\Pr(\varphi(x) \mid \psi(x)) = \Pr(\varphi(c) \mid \psi(c)))$.

Not surprisingly, Corollary 3.5 does not apply in the examples above. There, the distribution μ did not make $\varphi(x)$ and $x = c$ independent. In fact, in Example 3.3, the two events *Bird*(x) and $x =$ *Tweety* were completely correlated. Hence, this result gives us the tools to decide when substitution is appropriate.

4 Independency Mappings

The results of the previous section show us how we can use the independence properties of a distribution to conclude that certain things are irrelevant and, in particular, to substitute constants into universal rules. To use these techniques, we need to know whether $\mu \models \mathcal{I}(\varphi, \theta \mid \psi)$ for possibly many formulas φ, θ, ψ. In this section, we describe a technique that allows us a concise and intuitive graphical representation of some of the independencies that hold in μ. This representation allows a simple procedure for answering a wide range of independence queries.

Our representation utilizes a standard tool from the literature: *Markov networks* (Pearl 1988). A Markov network is an undirected graph G that encodes the independencies that hold for a distribution μ. In general, the nodes of such a network correspond to *random variables*, while the edges represent dependencies between them. Until now, Markov networks have been applied purely in an attribute-based (essentially propositional) setting. We apply this idea in the context of first-order logic by viewing each vocabulary symbol a in Φ as a random variable, whose values are the possible interpretations of a over the domain D. Formally, the graph has a node for each symbol in Φ. The existence of an edge between two symbols a and b in Φ means that choosing an interpretation for a may directly influence the probabilities with which we choose an interpretation for b. Conversely, the absence of this edge implies that there is no such direct influence.

Example 4.1 : Consider again the situation in Example 3.3. A Markov network G for a distribution over this vocabulary has two nodes: one for *Bird* and one for *Tweety*. The symbol *Bird* has four possible interpretations—$\emptyset, \{d_1\}, \{d_2\}$ and $\{d_1, d_2\}$—and the symbol *Tweety* has two—d_1 and d_2. In the distribution μ in that example, the interpretations of *Tweety* and of *Bird* were correlated: $\mu(w(\textit{Bird}) = \{d_2\} \mid w(\textit{Tweety}) = d_1) = 0$ while $\mu(w(\textit{Bird}) = \{d_2\} \mid w(\textit{Tweety}) = d_2) = 1/2$ (where $w(\textit{Bird}) = \{d_2\}$ denotes the event $\{w : w(\textit{Bird}) = \{d_2\}\}$). Hence, a graph G that represents the independencies in μ *must* have an edge between *Bird* and *Tweety*. On the other hand, consider a distribution μ' that makes the interpretations of *Tweety* and *Bird* independent. If we have, for example, $\mu'(w(\textit{Tweety}) = d_1) = 1/2$ and $\mu'(w(\textit{Bird}) = \{d_1\}) = 2/7$, then the world w_1 in Example 3.3 has $\mu'(w_1) = 1/2 \times 2/7 = 1/7$. Similarly, if $\mu'(w(\textit{Tweety}) = d_2) = 1/2$ and $\mu'(w(\textit{Bird}) = \{d_1, d_2\}) = 3/7$, then the world w_4 has $\mu'(w_4) = 1/2 \times 3/7 = 3/14$. A graph with no edge between *Bird* and *Tweety* represents the independencies of μ' in this case. ∎

As usual, Markov networks can also encode more complex conditional independencies between the vocabulary symbols (random variables). Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ denote disjoint subsets of nodes in G. We say that $\mathcal{C}$ *separates* $\mathcal{A}$ from $\mathcal{B}$ in G if every path from a node in $\mathcal{A}$ to a node in $\mathcal{B}$ passes through a node in $\mathcal{C}$. Intuitively, G encodes the fact that $\mathcal{A}$ can only influence $\mathcal{B}$ via $\mathcal{C}$. Hence, if we fix a particular interpretation for the symbols in $\mathcal{C}$, the symbols in $\mathcal{B}$ can no longer influence $\mathcal{A}$. We say that G is an *independency mapping (I-map) for* μ (Pearl 1988) if, whenever $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$ in G, the distribution μ makes the interpretations of the symbols in $\mathcal{A}$ conditionally independent of the interpretations of the symbols in $\mathcal{B}$ given an interpretation for the symbols in $\mathcal{C}$. Note that an I-map is a sound but incomplete representation of independencies: many distributions have independencies that cannot be expressed in a graphical structure of this type.

An I-map G for μ encodes conditional independencies that hold in μ. Is this really useful? After all, the I-map encodes independencies of interpretations of vocabulary symbols, while we are interested in formulas. As we now show, there is a natural way in which we can translate the former to the latter. Essentially, we can make any formula in $\mathcal{A}$ independent of any formula in $\mathcal{B}$ given a formula that contains all the relevant information about $\mathcal{C}$.

[4] Without exchangeability, we cannot conclude that $\forall x(\Pr(\varphi(x) \mid \psi(x) \wedge x = c) = \Pr(\varphi(c) \mid \psi(c)))$, the property required to prove (b) from (a).

Definition 4.2: Let $\mathcal{D}$ be a set of symbols (that may include both symbols in Φ and variables). We say that $\sigma \in \mathcal{L}_{obj}(\mathcal{D})$ is a *maximally descriptive formula over* $\mathcal{D}$ if all (world,valuation) pairs that satisfy σ agree on all formulas in $\mathcal{L}_{obj}(\mathcal{D})$. ∎

For example, if σ_k is the formula asserting that there are exactly k birds, then $\sigma_k \wedge \mathit{Bird}(x)$ is a maximally descriptive formula over $\{\mathit{Bird}, x\}$.

Theorem 4.3: *Let μ be an exchangeable distribution, and let G be an I-map for μ. Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be disjoint sets of symbols in Φ such that $\mathcal{C}$ separates $\mathcal{A}$ and $\mathcal{B}$ in G, and let $\mathcal{X}$ be a set of variables. If $\varphi \in \mathcal{L}_{obj}(\mathcal{A} \cup \mathcal{X})$, $\theta \in \mathcal{L}_{obj}(\mathcal{B} \cup \mathcal{X})$, and ψ is a maximally descriptive formula over $\mathcal{C} \cup \mathcal{X}$, then $\mu \models \mathcal{I}(\varphi, \theta \mid \psi)$.*

This theorem allows us to "read" an I-map in a useful way. By looking at an I-map of an exchangeable distribution μ, we can determine many conditional independencies $\mathcal{I}(\varphi, \theta \mid \psi)$ that hold for μ.

Example 4.4: We construct a distribution μ^+ over the vocabulary $\{\mathit{Fly}, \mathit{Bird}, \mathit{Tweety}\}$ that extends μ' from Example 4.1 and satisfies $\forall x(\Pr(\mathit{Fly}(x) \mid \mathit{Bird}(x)) = 0.9)$. We define μ^+ so as to make $\Pr(\mathit{Fly}(d) \mid \mathit{Bird}(d)) = 0.9$, and $\Pr(\mathit{Fly}(d) \mid \mathit{Bird}(d)) = 0.5$, where $\mathit{Fly}(d)$ is chosen independently for each domain element d. More precisely, let w be a world over this vocabulary, and let w' be that world over the vocabulary $\{\mathit{Bird}, \mathit{Tweety}\}$ which agrees with w about the interpretations of *Tweety* and *Bird*. For our two domain elements d_1, d_2, define q_i to be 0.9 if $d_i \in w(\mathit{Bird}) \cap w(\mathit{Fly})$ (i.e., if d_i in w is a flying bird), $q_i = 0.1$ if $d_i \in w(\mathit{Bird}) - w(\mathit{Fly})$, and $q_i = 0.5$ otherwise (i.e., if d_i is not a bird). If we define $\mu^+(w) = \mu'(w') \cdot q_1 \cdot q_2$, then the graph G where the only edge is between *Bird* and *Fly* is an I-map for μ^+. Since $\emptyset$ separates *Tweety* and $\{\mathit{Fly}, \mathit{Bird}\}$, and *true* is a maximally descriptive formula over $\{x\}$, we conclude from Theorem 4.3 that $x = \mathit{Tweety}$ is unconditionally independent of both $\mathit{Fly}(x)$ and $\mathit{Bird}(x)$. It follows trivially that $\mu^+ \models \mathcal{I}(\mathit{Fly}(x), x = \mathit{Tweety} \mid \mathit{Bird}(x))$, so that Corollary 3.5 allows us to conclude that $\mu^+ \models \Pr(\mathit{Fly}(\mathit{Tweety}) \mid \mathit{Bird}(\mathit{Tweety})) = 0.9$. ∎

5 The Effects of Conditioning

In the previous sections, we showed that certain desirable patterns of reasoning hold, given certain semantic conditions on distributions. But our reasoning process is rarely static: we continually get new information, and we must update our probability distribution accordingly, by conditioning on the new information. In this section, we study to what extent the conditioning process preserves the desirable patterns of reasoning we have been studying. As we shall see, having a semantic characterization of these properties greatly facilitates this investigation.

Given a distribution μ on W and a sentence $\psi \in \mathcal{L}_{obj}$, let $\mu(\psi) = \mu(\{w : w \models \psi\})$; if $\mu(\psi) > 0$, the distribution $\mu|\psi$ obtained by conditioning on ψ has $(\mu|\psi)(w) = \mu(w)/\mu(\psi)$ if $w \models \psi$ and 0 otherwise. Note that this process is defined only for objective formulas ψ. Also note that conditioning interacts as we would hope with our semantics: $\mu \models \Pr(\varphi \mid \psi) = \alpha$ iff $\mu|\psi \models \Pr(\varphi) = \alpha$.

It is easy to show that exchangeability is always maintained under conditioning.

Proposition 5.1: *If μ is exchangeable and $\psi \in \mathcal{L}_{obj}$, then $\mu|\psi$ is exchangeable.*

Determining which (conditional) independencies are maintained is somewhat more complex. As the following example shows, conditioning on a formula ψ may create some dependencies; we need to make precise which ones.

Example 5.2: Consider again the distribution μ^+ from Example 4.4, and recall that μ^+ satisfies both $\forall x(\Pr(\mathit{Fly}(x)|\mathit{Bird}(x)) = 0.9$ and $\Pr(\mathit{Fly}(\mathit{Tweety}) \mid \mathit{Bird}(\mathit{Tweety})) = 0.9$. It is not too hard to see $\mu^+|\mathit{Bird}(\mathit{Tweety})$ satisfies the same two formulas, although *Bird* and *Tweety* are clearly not independent in $\mu^+|\mathit{Bird}(\mathit{Tweety})$. Since *Bird* and *Tweety* are not independent, there must be an edge between *Tweety* and *Bird* in the I-map G for $\mu^+|\mathit{Bird}(\mathit{Tweety})$. But, intuitively, we made no direct connection between *Fly* and *Tweety*; the only connection goes through *Bird*. We might hope that G supports this intuition, and does not have a direct edge between *Fly* and *Tweety*. The fact that we continue to get $\Pr(\mathit{Fly}(\mathit{Tweety})) = 0.9$ suggests that this is the case.

What happens if we condition μ^+ on the assertion $\mathit{Bird}(\mathit{Tweety}) \wedge \neg \mathit{Fly}(\mathit{Tweety})$? The resulting distribution must satisfy $\Pr(\mathit{Fly}(\mathit{Tweety})) = 0$; hence, it can no longer be the case that $\mathcal{I}(\mathit{Fly}(x), x = \mathit{Tweety} \mid \mathit{Bird}(x))$ (since this would imply a probability of 0.9 for $\mathit{Fly}(\mathit{Tweety})$). Intuitively, the assertion $\mathit{Fly}(\mathit{Tweety})$ makes a direct connection between *Fly* and *Tweety*, one that is not mediated by *Bird*. Therefore, in the I-map for this new distribution, *Bird* will no longer separate *Fly* from *Tweety*. ∎

The next result makes the intuitive arguments used in this example more precise, by giving us a formal procedure for determining how I-maps change as a result of conditioning.

Theorem 5.3: *Let G be an I-map for μ. Suppose $\psi = \theta_1 \wedge \ldots \wedge \theta_n$ is a formula in $\mathcal{L}_{obj}(\Phi)$ and $\mu(\psi) > 0$. Let G' be the graph obtained by adding to G edges between nodes a and b if, for some i, a and b both appear in θ_i. Then G' is an I-map for $\mu|\psi$.*

This theorem follows from a more general one, proved in the full paper, that holds for arbitrary Markov networks: Let μ be a distribution with the I-map G; let E be some event which refers only to some set $\mathcal{A}$ of nodes in G. Then the graph G' obtained by adding edges between all the nodes in $\mathcal{A}$ is an I-map for $\mu|E$. Unlike many of the results on Markov networks (particularly those discussed in (Pearl 1988)), this result does not require the distribution μ to be positive (i.e., that $\mu(w) > 0$ for every world w).

Theorem 5.3 clearly validates our intuitive argument from Example 5.2: The process of conditioning on $\mathit{Bird}(\mathit{Tweety})$ adds an edge between *Bird* and *Tweety*, but the node *Bird* still separates *Fly* from *Tweety*. On the other hand, conditioning on $\mathit{Bird}(\mathit{Tweety}) \wedge \neg\mathit{Fly}(\mathit{Tweety})$ adds a direct edge between *Fly* and *Tweety*, thereby preventing *Bird* from separating these two symbols and eliminating the resulting conditional independency. As we show in Section 7, this seemingly simple analysis provides us with a very powerful tool for reasoning about irrelevance in statistical knowledge bases.

6 Objective and Subjective Probabilities

The nature of the connection between objective probabilities (frequencies) and subjective probabilities (degrees of belief) has generated extensive and often heated debate. On the one hand, devout *frequentists* believe that all probabilities derive from frequencies, and that subjective probabilities are meaningless, or at least irrelevant to the "real world." On the other hand, the more extreme *subjective Bayesians* believe that the concept of "the real world" is uninteresting and irrelevant, and that only subjective probabilities, which may be arbitrary, are meaningful. While we do not have the space to elaborate on this discussion, we show that the formal semantics of our language *forces* a connection between these two types of probabilities.

Example 6.1: In the depression era in the U.S., unemployment was such that, for any advertised job, there would be a line of applicants stretched around the block, waiting to interview for the job. There is a true story about a journalist who went up to one of the men standing in such a line, and asked him "There are a thousand people standing in line to interview for this one job; what do you think your chances are?" The man answered: "Oh, fifty-fifty, just like everybody else." Is it reasonable to believe that $\forall x(\Pr(\mathit{GetJob}(x)) = 0.5)$? *A priori*, it seems that the answer is "yes;" after all, you might say, "I may believe whatever I wish." Surprisingly, this is inconsistent with the basic semantics of first-order probabilistic logic. Let us simplify the analysis, and assume that the number of people interviewing for the job is three. Thus, if we ignore for the moment all other aspects of the world, there are three possible worlds over this domain, say w_1, w_2, and w_3, each with domain $\{d_1, d_2, d_3\}$, where in w_i, the person d_i is the one who ultimately gets the job. Suppose we take the probability of w_i to be p_i, $i = 1, 2, 3$. The formula $\forall x(\Pr(\mathit{GetJob}(x)) = 0.5)$ implies $\Pr(\mathit{GetJob}(d_1)) = 0.5$, and since d_1 gets the job just in w_1, we get that $p_1 = 0.5$. Similarly, we get that $p_2 = 0.5$ and $p_3 = 0.5$. But p_1, p_2, and p_3 must sum to 1, so this is inconsistent. Intuitively, the two statements "precisely one person will get the job" and "$\forall x(\Pr(\mathit{GetJob}(x)) = 1/2)$" both count occurrences of the event "$\mathit{GetJob}(d)$ holds in w": The first statement does so for a fixed w, and the second for a fixed d (weighted by $\mu(w)$). Hence, the outcomes must be related. ∎

To generalize this argument, we first introduce some notation. As in (Halpern 1990), we augment first-order logic with a statistical quantifier. Formally, we allow *proportion expressions* of the form $||\psi(x)||_x$. This is interpreted as a rational number between 0 and 1, that represents the proportion (or fraction) of domain elements satisfying $\psi(x)$. As for probabilities, we allow *conditional proportion expressions* of the form $||\varphi(x) \mid \psi(x)||_x$, interpreting each one as an abbreviation for the formula obtained by multiplying to clear the denominator. We can now show, using similar arguments to those above, that:

Theorem 6.2: *The two statements* $||\varphi(x) \mid \psi(x)||_x \geq \alpha$ *and* $\forall x(\Pr(\varphi(x) \mid \psi(x)) < \alpha)$ *are inconsistent.*

It follows that we cannot have statistical information that α of ψ's are φ's and believe that the probability that $\varphi(d)$ holds given $\psi(d)$ is $\beta \neq \alpha$ for each domain element d. We *can* believe that the probability of $\varphi(d)$ given $\psi(d)$ is β for *some* domain elements, just not all of them. In the above example, for instance, it is quite legitimate to have $p_1 = 3/4, p_2 = 1/4, p_3 = 0$. This would imply $\Pr(\mathit{GetJob}(d_1)) = 3/4$, $\Pr(\mathit{GetJob}(d_2)) = 1/4$, and $\Pr(\mathit{GetJob}(d_3)) = 0$.

When do we get a tighter connection between objective and subjective probabilities? Part of the answer is fairly clear: if we want to have $\forall x(\Pr(\varphi(x) \mid \psi(x)) = \alpha)$, we must first have $\Pr(\varphi(d) \mid \psi(d))$ be the same for all domain elements d. As we have already pointed out in Section 2, this is a consequence of the exchangeability assumption. It turns out that this is also sufficient to guarantee the desired connection: If $\Pr(\varphi(d) \mid \psi(d))$ takes the same value, say β, for all domain elements d, then Theorem 6.2 implies that necessarily $\beta = \alpha$.

Theorem 6.3 : *If* μ *is exchangeable and* $\mu \models ||\varphi(x) \mid \psi(x)||_x \in [\alpha, \beta]$, *then* $\mu \models \forall x(\Pr(\varphi(x) \mid \psi(x)) \in [\alpha, \beta])$.

7 From Statistics to Degrees of Belief

We now present one important application of the ideas presented in this paper, to the problem of going from statistical information to subjective degrees of belief. Assume that we have a knowledge base $KB \in \mathcal{L}_{obj}(\Phi)$ which can contain statistical as well as first-order statements. We would like to use our information to induce degrees of belief in statements concerning particular individuals. There are a number of properties that we might hope such an approach would have. We briefly describe the ones of interest here using simple examples.

Direct inference: Suppose KB_{fly} is

$$||\mathit{Fly}(x) \mid \mathit{Bird}(x)||_x = 0.9 \wedge \mathit{Bird}(\mathit{Tweety}).$$

Then we would like to conclude $\Pr(\mathit{Fly}(\mathit{Tweety})) = 0.9$; that is, we would like our degree of belief in $\mathit{Fly}(\mathit{Tweety})$ to be determined by the statistical information. This tight connection between statistical information and degrees of belief is known as *direct inference*.

Preference for more specific information: Suppose we have statistical information about φ within two sets, where one is more specific than (i.e., a subset of) the other. For example, suppose that, in addition to the information in KB_{fly}, we also know that $||\mathit{Fly}(x) \mid \mathit{Penguin}(x)||_x = 0.01 \wedge \forall x(\mathit{Penguin}(x) \Rightarrow \mathit{Bird}(x)) \wedge \mathit{Penguin}(\mathit{Tweety})$. In that case, we would hope to use the more specific statistics for $\mathit{Fly}(x)$, and conclude that $\Pr(\mathit{Fly}(\mathit{Tweety})) = 0.01$.

Irrelevant information: If, in addition to KB_{fly}, we also learn that $\mathit{Red}(\mathit{Tweety})$. We might hope to ignore the seemingly irrelevant information $\mathit{Red}(\mathit{Tweety})$ and still conclude that $\Pr(\mathit{Fly}(\mathit{Tweety})) = 0.9$.

(Bacchus *et al.* 1994) presents one approach, called *random worlds*, to dealing with this problem: Start with a uniform prior over the set of possible worlds, condition on the knowledge base KB, and use the resulting posterior distribution $\mu_0|KB$ to form degrees of belief. They show that all these properties hold for the random worlds approach. Is there anything special about the uniform prior that gives these results? Our analysis in the previous sections gives us the tools to answer this question. It is immediate from the definition of exchangeability that the uniform prior is exchangeable: Since all worlds have the same probability, then in particular so do isomorphic worlds. It is also easy to check that the uniform prior is "very independent": the interpretations of all the vocabulary symbols are chosen independently. Formally, we say that a distribution μ is *fully independent* if the graph G_0 that has no edges is an I-map for μ. Clearly, the uniform prior is fully independent. As we show below, full independence and exchangeability are the only properties of the uniform prior that are required to prove these properties. It follows that these results actually hold for a large class of priors. For example, the *random-propensities* distribution considered in (Bacchus *et al.* 1995) is also fully independent and exchangeable. So is the distribution μ' presented in Example 4.1. Thus, we have a large space in which we can look for a prior that would give us the benefits of the uniform prior without some of its disadvantages.

For example, a rather general theorem was proved in (Bacchus *et al.* 1994) from which direct inference and preference for more specific information followed quite easily. The following result is a restatement of that theorem, but for arbitrary fully independent and exchangeable prior distributions, rather than just the uniform prior.

Theorem 7.1: *Suppose μ is an exchangeable and fully independent distribution and let KB be a knowledge base of the form $KB' \wedge \psi(c)$. If $KB \models ||\varphi(x) \mid \psi(x)||_x \in [\alpha, \beta]$, and KB', $\varphi(x)$, and $\psi(x)$ do not mention c, then $\mu|KB \models$* $\Pr(\varphi(c)) \in [\alpha, \beta]$.

We give the main ideas for this proof using an example, noting that the full proof is essentially identical.

Example 7.2: Let $\Phi = \{Fly, Bird, Tweety\}$. Assume that we start from a fully independent and exchangeable prior μ, and condition on KB_{fly} (defined above).

(a) By Theorem 5.3, we can construct an I-map G for $\mu|KB_{fly}$ by adding an edge between *Bird* and *Fly* because of the conjunct $||Fly(x) \mid Bird(x)||_x = 0.9$ and an edge between *Bird* and *Tweety* because of the conjunct $Bird(Tweety)$. Since our original I-map was G_0, these are the only edges.

(b) Since *Bird* separates *Fly* and *Tweety*, we now apply Theorem 4.3 and conclude $\mathcal{I}(Fly(x), x = Tweety \mid \psi(x))$ where $\psi(x)$ is a maximally descriptive formula over $\{Bird, x\}$. As described immediately after Definition 4.2, $\sigma_k \wedge Bird(x)$ is an appropriate choice.

(c) By Proposition 5.1, $\mu|(KB_{fly} \wedge \sigma_k)$ is exchangeable, and so by Theorem 6.3, $\mu|(KB_{fly} \wedge \sigma_k) \models \forall x(\Pr(Fly(x) \mid Bird(x)) = 0.9)$. It follows that $\mu|KB_{fly}$ satisfies $\forall x(\Pr(Fly(x) \mid Bird(x) \wedge \sigma_k) = 0.9)$.

(d) Applying Corollary 3.5 to our conclusions from steps (b) and (c), we conclude that $\mu|KB_{fly} \models \Pr(Fly(Tweety) \mid Bird(Tweety) \wedge \sigma_k) = 0.9$ for every k. Since we know $Bird(Tweety)$, and the formulas σ_k are mutually exclusive and exhaustive, the desired conclusion follows easily. ∎

(Bacchus *et al.* 1994) also presents a theorem dealing with the treatment of irrelevant information. Similar arguments allow us to generalize that theorem, showing that it holds for all fully independent and exchangeable priors. In fact, the techniques of this paper give us a powerful new (sound but not complete) technique for testing when such irrelevance holds, by using I-maps. As we now demonstrate, this approach allows us to deal with quite complex knowledge bases.

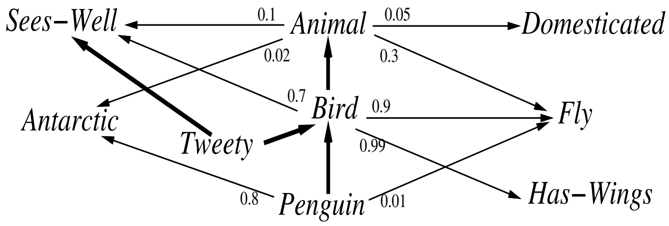

Figure 1: A graphical representation of a knowledge base

Example 7.3: Figure 1 is a graphical representation of a statistical knowledge base *KB*. The dark unlabeled arrows denote *is-a* arrows (Touretzky 1986); for example, the edge between *Tweety* and *Bird* corresponds to the statement $Bird(Tweety)$, while the edge between *Bird* and *Animal* corresponds to $\forall x(Bird(x) \Rightarrow Animal(x))$. The labeled arrows correspond to statistical statements; for example, the edge between *Fly* and *Bird* labeled with 0.9 represents the statement $||Fly(x) \mid Bird(x)||_x = 0.9$. What independencies hold after we condition some fully independent and exchangeable prior μ on *KB*? Figure 1 already gives us the answer: If we view the figure as an undirected graph, we get precisely the I-map for $\mu|KB$! Hence, we can conclude for example that $\{Bird, Animal\}$ separates *Tweety* from *Fly*. Using arguments as above (and the fact that $Bird(x)$ implies $Animal(x)$), it is not too hard to show that we can use the statistics for birds when reasoning about Tweety. In particular, $\mu|KB \models \Pr(Fly(Tweety)) = 0.9 \wedge \Pr(Has\text{-}Wings(Tweety)) = 0.99 \wedge \Pr(Domesticated(Tweety)) = 0.05$. Note that, due to the complexity of this knowledge base, none of these conclusions follow from the (Bacchus *et al.* 1994) theorems. If we now condition on $Antarctic(Tweety)$, we add an edge between *Antarctic* and *Tweety*. This creates a path from *Tweety* to *Fly* that is not blocked by *Bird*. Hence, we will no longer be able to conclude that $\Pr(Fly(Tweety)) = 0.9$, but our other two conclusions still hold. This is precisely the behavior we would expect. ∎

This type of analysis also applies to knowledge bases with non-unary predicates.

Example 7.4: Assume that our knowledge base *KB* is

$$\begin{array}{l} ||\mathit{Sells}(x,y) \mid \mathit{Supermarket}(x) \wedge \mathit{Cheese}(y)||_{x,y} = 0.6 \\ ||\mathit{Sells}(x,y) \mid \mathit{Supermarket}(x) \wedge \mathit{Specialty}(y)||_{x,y} = 0.2 \\ \mathit{Supermarket}(\mathit{Safeway}) \wedge \mathit{Cheese}(\mathit{Stilton}) \\ \mathit{English}(\mathit{Stilton}) \wedge \mathit{Stinky}(\mathit{Stilton}) \\ ||\mathit{Stinky}(x) \mid \mathit{Cheese}(x)||_x = 0.25. \end{array} .$$

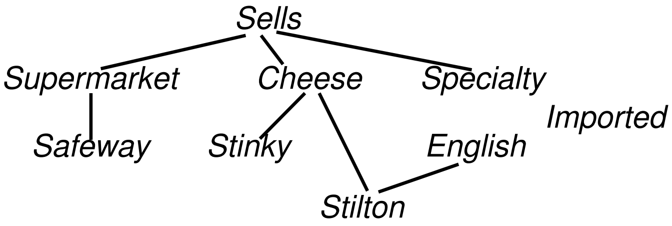

Figure 2: I-map for supermarket knowledge base

The I-map resulting from conditioning a fully independent and exchangeable prior μ on *KB* is shown in Figure 2. Using the independencies implied by this diagram, we can show that $\mu|\mathit{KB} \models \Pr(\mathit{Sells}(\mathit{Safeway},\mathit{Stilton}) \mid \mathit{KB}) = 0.6$. But if we add to the knowledge base the statements $\forall x(\mathit{English}(x) \Rightarrow \mathit{Imported}(x)) \wedge ||\mathit{Specialty}(x) \mid \mathit{Imported}(x)||_x = 0.9$, the resulting I-map will also have an edge from *English* to *Imported* and from *Specialty* to *Imported*, creating a path from *Stilton* to *Sells* which is not blocked by *Cheese*. This will prevent us from drawing the no-longer-desired conclusion that $\Pr(\mathit{Sells}(\mathit{Safeway},\mathit{Stilton})) = 0.6$. ∎

The knowledge bases in these examples were fairly complex, yet our analysis of independencies was easy to carry out. But there was nothing special about these examples: our analysis uses only simple syntactic tests, and can easily be applied to any knowledge base. The analysis is particularly easy for knowledge bases that correspond to a probabilistic (or statistical) *semantic network* (as did the one in the first example). In this case, as we showed, the I-map can simply be read off the network. As in steps (b), (c), and (d) in Example 7.2 above, our other theorems then allow us to derive probabilities for various queries.

8 Conclusions

In this paper, we used a semantic approach to analyze several important patterns of reasoning in first-order probabilistic logic. In particular, we presented a semantic characterization and a graphical representation language for irrelevance in this context. We showed that the deeper understanding gained by our analysis allows us to determine when certain irrelevancies are maintained as we gain new information. Perhaps the most important immediate payoff of these results is a sound and simple procedure for reasoning about irrelevance in statistical knowledge bases. Our graphical representation language immediately reveals that some facts are irrelevant to our query given our information, allowing us to ignore certain parts of the knowledge base entirely. When combined with our other results, we obtain a sound (although incomplete) inference algorithm for a large sublanguage of first-order probabilistic logic.

Our graphical representation of irrelevance is particularly well-matched to knowledge bases corresponding to statistical semantic networks. Given the popularity of semantic networks as a knowledge representation language, we feel that this new approach for sound irrelevance reasoning in such networks is a significant contribution of our work.

We conclude with one important direction in which our work should be extended. We defined independence for events defined by arbitrary formulas. However, our definition of Markov networks represents independence at the more coarse-grained level of vocabulary symbols. While this is still fairly powerful, it is insufficient for certain applications. For example, we cannot use such networks to represent the fact that $\mathit{Cancer}(x)$ is independent of $\mathit{Cancer}(y)$ if x and y are not related. In future work, we hope to provide a more refined representation for independencies that would allow us to capture independencies of this type.

References

F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. Technical Report 9855, IBM, 1994. To appear, *Artificial Intelligence*. Available by anonymous ftp from logos.uwaterloo.ca/pub/bacchus or via WWW at http://logos.uwaterloo.ca. A preliminary version of this work appeared in *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, 1993, pages 563–569.

F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. Reasoning with noisy sensors in the situation calculus. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, pages 1933–1940, 1995.

F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, Cambridge, Mass., 1990.

B. De Finetti. Foresight: Its logical laws, its subjective sources. In H. E. Kyburg, Jr. and H. Smokler, editors, *Studies in Subjective Probability*. John Wiley & Sons, New York, 1964.

J. W. Garson. Quantification in modal logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, Vol. II*, pages 249–307. Reidel, Dordrecht, Netherlands, 1977.

J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, Calif., 1988.

H. Reichenbach. *The Theory of Probability*. University of California Press, Berkeley, 1949. This is a translation and revision of the German edition, published as *Wahrscheinlichkeitslehre*, in 1935.

J. Sowa, editor. *Principles of Semantic Networks*. Morgan Kaufmann, 1991.

D. S. Touretzky. *The Mathematics of Inheritance Systems*. Morgan Kaufmann, 1986.