Psychological Theories of Categorization as Probabilistic Models

David Danks*
Department of Philosophy, Carnegie Mellon University; and
Institute for Human & Machine Cognition

1. Introduction

A wide variety of psychological models have been proposed to explain people's categorization of novel instances (e.g., Medin & Schaffer, 1978; Minda & Smith, 2001, 2002; Nosofsky, 1984, 1986, 1991; Nosofsky & Zaki 2002; Zaki, *et al*., 2003). That is, given that I have a range of possible categories, C_1, …, C_n and I observe a novel case X, what is the probability that I will classify X into one or another of the n possibilities? Essentially all of these theories share an important structural similarity: they all first calculate some "similarity" or "representativeness" measure of X for each C_i, and then combine those measures in some way to produce a probability (possibly of 1.0 on one particular category). I focus here on the first step in this process, and take no particular stance on the ways in which these functions are integrated to produce a precise behavioral response. The principal advantage to this approach is that it enables us to focus on single categories, and avoid the theoretically (and experimentally) challenging task of determining the second-stage behavior. In addition, results about equality between theories in the first stage imply that those theories will remain equal for particular versions of the second stage.

More precisely, suppose we have a list of m binary properties for a particular case. The first-stage functions, I will refer to them as "similarity measures," are thus functions from m-dimensional binary vectors to non-negative real numbers. The similarity measures in all of these

* **Acknowledgements:** Thanks to Bob Rehder for discussions about the two-stage nature of categorization theories. The feature matching equivalence results were originally presented at a February, 2004 Workshop on Causation and Categorization held at the Center for Advanced Study in the Behavioral Sciences. This research was partly supported by grants NCC2-1399 and NCC2-1377 from the National Aeronautics and Space Administration.

psychological theories are intended to be scale-invariant: for a given similarity measure $g(X)$, identical predictions would be obtained if we instead used $K \times g(X)$, for all positive K. Similarity measures that have been normalized so that the sum over all possible cases is 1 can be interpreted as probability distributions over the possible cases. Moreover, any probability distribution over the possible cases can also be interpreted as a (normalized) similarity measure. Thus, the focus here will be on connecting similarity measures and probability distributions.

There are essentially three classes of similarity measures that have been proposed in the psychological literature: exemplar, feature-matching, and causal model theories. The first two have dominated the psychological landscape for the past twenty years, and there has been some prior work on establishing the conditions in which they behave similarly (Nosofsky, 1990). In this paper, I will focus on exemplar and feature-matching models, but will not directly try to show when the theories are identical or differ. Instead, I connect them with probabilistic models proposed elsewhere, primarily computer science and statistics. In particular, I will show that the similarity measures defined by the theories are isomorphic to sets of probability distributions that are easily characterized using probabilistic models. By placing these theories into that general framework, we can then quickly and easily establish connections between these two classes of theories, and also causal model theory, which uses the framework of Bayesian networks.

2. Feature Matching Models and Markov Random Fields

In this section, I first provide a precise definition of a prototype model, here called a feature matching model, then introduce the framework of Markov Random Fields (MRFs), and show that feature matching models are isomorphic to a subset of MRFs. More precisely, the probability distributions defined by the normalized similarity measures for feature matching

models are the same probability distributions describable using MRFs and an additional constraint.

Feature matching models (Minda & Smith, 2001, 2002) assume that a category can be identified with some prototypical instance, though they do not assume that the prototype must ever have been actually observed. The intuition behind a feature matching model is that novel instances will be judged more likely to be a member of a category as we increase the number of properties (features) shared by the novel instance and the prototype. In this section, we consider a generalization of this model that allows for interactions between the features. In particular, we might think that certain *combinations* of properties are particularly important, and so our similarity measure must be able to incorporate those second-order features as well.

To make all of this precise, let X by the case we have just been shown, and let E be a prototypical case for category C. Since we are assuming that all of the properties are binary, we can assume (without loss of generality) that the prototypical case is "all features are present."[1] Define $\delta(q) = 1$ if X and E have the same value for the q-th property, and 0 otherwise.[2] We then define a feature matching model S as any model of the form: $S(X) = \prod_{i=1}^{m}\prod_{j=i}^{m}\left(s_{ij}\right)^{[1-\delta(i)\delta(j)]}$.[3] (Note that this particular characterization is not normalized.) That is, the similarity of X is the product of s_{ij} for every i, j pair (possibly $i = j$) in which X and E differ on at least one of the dimensions. We assume only that all s_{ij} are positive, though the model is only sensibly applied to categorization data if $s_{ij} \leq 1$. s_{ij} values smaller than 1.0 indicate increasingly important i, j interactions; $s_{ij} = 1$ indicates no i, j interaction at all. We are here concerned with models of

[1] This assumption is entirely one of convenience for notation. All of the claims and proofs can straightforwardly be extended for different prototypical cases.
[2] Under the assumption that E = "all present," this is just the standard delta-function for the q-th dimension of X.
[3] Sometimes, weighted city block distance measures are used, instead of a multiplicative measure. These versions are isomorphic to the logs of probability distributions.

representativeness ratings for all possible cases, where the ratings are only defined up to a multiplicative constant. Therefore, we will say that "there is a feature matching model for ratings R" iff there exists a feature matching model S such that, for all X, $S(X) \propto R(X)$. Similarly, there is a probability distribution for the ratings R iff for all X, $P(\mathrm{X}) \propto R(\mathrm{X})$.

A Markov Random Field (MRF) is composed of two components: an undirected graph over the variables, and a probability distribution over those same variables. For a particular undirected graph, we define a *clique* to be a set of variables **S** such that every pair of variables in **S** is connected by an edge. A *maximal clique* is a clique **T** such that $\mathbf{T} \cup \{A\}$ is not a clique for all $A \notin \mathbf{S}$. The Markov assumption for an MRF states: the probability distribution factors into the product of functions over the maximal cliques in the undirected graph. That is, if $G_1, \ldots, G_q$ are the maximal cliques for a particular undirected graph G, we can express the distribution as $P(X) = \frac{1}{Z}\prod_{i=1}^{q} g_i(X)$, where $g_i(X)$ is a function only of the X-values of the variables in clique G_i, and Z is a normalization constant. These g functions are frequently called the *clique potentials*.

Because we can always renormalize the Markov decomposition for an MRF, the clique potentials in an MRF are defined only up to a multiplicative constant. Therefore, we can assume (without loss of generality) that $g(\textit{all present}) = 1$ for all clique potentials.[4] We further define X- to be the set of variables that are absent in case X, and $g(\mathbf{S}+)$ to be the clique potential value for the case in which only the variables in **S** are present (and so $g(\varnothing+)$ is the clique potential value for the "all absent" case). A clique potential then has the "g-decomposition property" iff:

$$\forall X\left[g(X) = \prod_{i \in X-} \frac{g(\varnothing+)}{g(i+)} \prod_{i \in X-} \prod_{j=i+1}^{m} \frac{g(i+)g(j+)}{g(\{i,j\}+)g(\varnothing+)} \right]$$

[4] More generally, if E is not the "all present" case, then we assume $g(E) = 1$; that is, every clique potential equals 1 for the prototype case. The g-decomposition property must then be relativized to the particular E in obvious ways. All of the proofs hold under this generalization.

This property appears quite complex, and might seem quite stringent. We can prove, however, that it holds of some important classes of cliques. (All proofs are provided in the Appendix.)

> Theorem 1: The *g*-decomposition property holds for all cliques with 1 or 2 variables. It holds for only some cliques with 3 or more variables.

Given all of these definitions, the keystone theorem of this section is:

> Theorem 2: [There exists a feature matching model S for ratings R] iff [there exists a probability distribution P for R such that (i) there exists an MRF U that perfectly represents P; and (ii) the *g*-decomposition property holds for every (maximal) clique potential in U]. Moreover, for a given feature matching model S, there is a unique corresponding probability distribution, and *vice versa*.

That is, if there is a feature matching model for those ratings R, then we can also find a MRF U such that $P(X \mid U)$ is an equally good model of the ratings. Thus, there is not (from a theoretical point of view) any distinction between these theories.

Causal model theory predicts that ratings should be well-modeled by $P(X \mid B)$, where B is a Bayesian network over the variables (not defined in this paper). Therefore, we also derive three important corollaries about the relationship between them and feature matching models:

> Corollary 1: If B is a Bayes net tree, then there is a feature matching model S such that $S(X) \propto P(X \mid B)$ for all X.

> Corollary 2: There exists a probability distribution that can be modeled by a Bayes net (e.g., if the Bayes net has an unshielded collider), but for which there is no feature matching model such that $S(X) \propto P(X \mid B)$ for all X.

Corollary 3: There exists a probability distribution P that cannot be modeled by a Bayes net, but for which there is a feature matching model such that $S(X) \propto P(X)$ for all X.

That is, the probability distributions that can be captured by Bayes nets and by MRFs/feature matching models are overlapping, without inclusion in either direction. There are sets of ratings for which there are both MRF and Bayes net models, and ratings that can only be modeled by one or the other model.

3. Exemplar Models

An exemplar model (e.g., Medin & Schaffer, 1978; Nosofsky, 1984, 1986, 1991; Nosofsky & Zaki 2002; Zaki, *et al*., 2003) presumes that we have a list of exemplar instances, and the rating is the weighted average similarity measure for X compared to each exemplar as though the exemplar is the prototype for a feature matching model and the s_i-parameters are held fixed. More precisely, define $E_a(i)$ to be the value of the i-th dimension for exemplar E_a; W_a to be the (positive) relative weight of exemplar E_a; and $eq(A, B) = 1$ if A and B are equal, and 0 otherwise. Then the similarity measure for novel instance X given exemplars $E_1, \ldots, E_r$ is given by: $F(X) = \frac{1}{r}\sum_{a=1}^{r} W_a \prod_{i=1}^{m} (s_i)^{[1-eq(X_i, E_a(i))]}$.[5]

There is a straightforward interpretation of exemplar similarity measures in terms of noisy measurements. Define a *noisy measurement model* to be one in which our measurements of each property's value are subject to some noise, and we correctly measure the i-th dimension with probability q_i. Further suppose that we have a probability distribution P^* over a set of

[5] Again, weighted city block distances can be used instead of a multiplicative measure.

possibly occurring instances. For this situation, we can provide an alternate characterization of the exemplar similarity measure:

> Theorem 3: [There exists an exemplar model F for ratings R] iff [there exists a noisy measurement model M and probability distribution $P^*(E)$ over the exemplars such that $P(X \mid P^*, M)$ is appropriate for R].

That is, there is a straightforward equivalence between exemplar similarity measures and models that calculate the probability of observing X given that (i) one of the exemplars (drawn from probability distribution P^*) actually occurred, and (ii) our measurements are noisy. As shown in the proof, the noisy measurement model M corresponds to the s-parameters in the exemplar model F, and the probability distribution P^* corresponds to the weights in F.

Exemplar similarity measures have also been extended to include fragments of exemplars. Specifically, we allow exemplars to be undefined on certain dimensions, and we exclude those dimensions from the similarity product for that exemplar fragment. Note that this extension is *not* equivalent to simply expanding the set of exemplars to include all possible completions. In the context of a noisy measurement model, an undefined dimension implies that we do not know the true value, and so we should exclude that dimension from the calculation of the probability. Therefore, we can straightforwardly extend Theorem 3 to hold for exemplar fragment similarity measures. These models are equivalent to determining the probability of observing the current instance, given noisy measurements and a probability distribution over the exemplars, but where we may not have complete information about all of the exemplars.

4. Conclusion

This paper has focused on formal results connecting these similarity measures to other probabilistic models. These connections can now be used to support various cross-theoretical comparisons, as well as improved design of experiments to (potentially) distinguish the theories. A broader discussion of the impact of these results can be found in Danks (forthcoming), but I address one important issue in closing. Prototype/feature matching and exemplar models of categorization are often presented as stark contrasts to one another. One striking feature about the theoretical results presented here is that the underlying structure of both of these psychological theories is fundamentally the same. Both are isomorphic to computing $P(X$ occurring | Category model). They simply disagree about the nature of the category model: it is an MRF for feature matching models, and noisy measurement of cases in exemplar models. Thus, we can perhaps think of human categorization not as feature matching *versus* exemplar (versus causal model theory), but rather as the same process operating on different category concepts, depending on background knowledge, historical experiences, and so on. Moreover, there is a plausible intuitive story to be told about why our models for a particular category might change over time: we start with only a few instances of a category (exemplar model); as we observe more instances, we learn a correlation structure for the features (an MRF/feature matching model); and in some cases, we finally come to understand the asymmetric causal relationships underlying the category (a causal Bayes net/causal model theory). Important details (both conceptual and computational) are, of course, missing from this story; an important future research project is to attempt to fill in the gaps and test it empirically.

References

Danks, David. Forthcoming. "Predicting Instances and Inferring Structures in Categorization." In A. Gopnik, ed. *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press.

Medin, D. L., & M. M. Schaffer. 1978. "Context Theory of Classification Learning." *Psychological Review*, 85: 207-238.

Minda, John Paul, & J. David Smith. 2001. "Prototypes in Category Learning: The Effects of Category Size, Category Structure, and Stimulus Complexity." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27: 775-799.

Minda, John Paul, & J. David Smith. 2002. "Comparing Prototype-based and Exemplar-based Accounts of Category Learning and Attentional Allocation." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28: 275-292.

Nosofsky, Robert M. 1984. "Choice, Similarity, and the Context Theory of Classification." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10: 104-114.

Nosofsky, Robert M. 1986. "Attention, Similarity, and the Identification-categorization Relationship." *Journal of Experimental Psychology: General*, 115: 39-57.

Nosofsky, Robert M. 1990. "Relations Between Exemplar-similarity and Likelihood Models of Classification." *Journal of Mathematical Psychology*, 34 (4): 393-418.

Nosofsky, Robert M. 1991. "Tests of an Exemplar Model for Relating Perceptual Classification and Recognition Memory." *Journal of Experimental Psychology: Human Perception and Performance*, 17: 3-27.

Nosofsky, Robert M., & Safa R. Zaki. 2002. "Exemplar and Prototype Models Revisited: Response Strategies, Selective Attention, and Stimulus Generalization." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28: 924-940.

Zaki, Safa R., Robert M. Nosofsky, Roger D. Stanton, & Andrew L. Cohen. 2003. "Prototype and Exemplar Accounts of Category Learning and Attention Allocation: A Reassessment." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29 (6): 1160-1173.

Appendix

Proofs of Feature Matching Model Theorems:

Theorem 1: The *g*-decomposition property holds for all cliques with 1 or 2 variables. It holds for only some cliques with 3 or more variables.

Proof of Theorem 1: The *g*-decomposition property states that the clique potential can be expressed as:

$$\forall X\left[g(X) = \prod_{i \in X-} \frac{g(\varnothing +)}{g(i+)} \prod_{i \in X-} \prod_{j=i+1}^{m} \frac{g(i+)g(j+)}{g(\{i, j\}+)g(\varnothing +)} \right]$$

. For all clique potentials, we set *g*(*all present*) = 1 by choice, which satisfies this property. Therefore, we need only concern ourselves with those *X* such that at least one variable is absent.

Clique with 1 variable: Let *g* be an arbitrary clique potential. The only relevant case is (0): the variable is absent. The second product is necessarily empty, and the first product is just *g*(0)/*g*(1). Since *g*(1) = 1, the *g*-decomposition property holds for this clique potential.

Clique with 2 variables: Let *g* be an arbitrary clique potential. The right-hand side of the *g*-decomposition property for (1,0) is equal to: $\frac{g(00)}{g(01)} \times \frac{g(01)g(10)}{g(11)g(00)}$. Since *g*(11) = 1, this product equals *g*(10). Similarly, the *g*-decomposition property holds for (0,1). Now consider (0,0). The right-hand side equals:

$\frac{g(00)}{g(01)} \times \frac{g(00)}{g(10)} \times \frac{g(01)g(10)}{g(11)g(00)}$. This product must equal *g*(00). Therefore, the *g*-decomposition property holds for *g*.

Clique with 3 variables: Let *g* be an arbitrary clique potential. The right-hand side of the *g*-decomposition property for (0,0,0) equals: $\frac{g(000)}{g(100)} \times \frac{g(000)}{g(010)} \times \frac{g(000)}{g(001)} \times \frac{g(100)g(010)}{g(110)g(000)} \times \frac{g(100)g(001)}{g(101)g(000)} \times \frac{g(010)g(001)}{g(011)g(000)}$,

which equals: $\frac{g(100)g(010)g(001)}{g(110)g(101)g(011)}$. Clearly, this value need not equal *g*(000), and so the *g*-decomposition property does not hold for all *g*. At the same time, the *g*-decomposition property must hold for some *g*, since there must be some clique potential for the probability distribution defined by the feature matching model with three features and all possible second-order features.

A similar procedure clearly demonstrates that some clique potentials for more than 3 variables do not have the f-decomposition property, but some do (namely, those that correspond to some feature matching model). ■

Theorem 2: [There exists a feature matching model S for ratings R] iff [there exists a probability distribution P for R such that (i) there exists a MRF U that perfectly represents P; and (ii) the g-decomposition property holds for every (maximal) clique potential in U]. Moreover, for a given feature matching model S, there is a unique corresponding probability distribution, and *vice versa*.

Proof of

Theorem 2: We prove this theorem using a series of lemmas. The first lemmas demonstrate conditional uniqueness. The later lemmas demonstrate existence (i.e., that the conditional holds). Note that the proofs demonstrate only that $S \propto P$, and exclude the ratings entirely. If they are proportional to each other, then if one is proportional to the ratings, so is the other.

Lemma 1: For every feature matching model S, if there is a probability distribution P such that $S(X) \propto P(X)$, then P is unique.

Proof of Lemma 1: Let P_1 and P_2 be such that $K_1 \times P_1(X) = S(X) = K_2 \times P_2(X)$, for all X. Therefore,

$\sum_X K_1 \times P_1(X) = \sum_X K_2 \times P_2(X)$. But since $\sum_X P_1(X) = 1 = \sum_X P_2(X)$, it must be the case that $K_1 = K_2$.

Therefore, for all X, $P_1(X) = P_2(X)$. ■

We will now use P_S to denote the (possibly non-existent) unique probability distribution for a particular S.

Lemma 2: For all probability distributions P, if there exists a feature matching model S such that $S(X) \propto P(X)$, then S is unique.

Proof of Lemma 2: Let S_1 and S_2 be such that $K_1 \times S_1(X) = P(X) = K_2 \times S_2(X)$, for all X. We first show that $K_1 = K_2$. If A is the "all present" instance, then $S_1(A) = 1 = S_2(A)$ by definition of feature matching model. Therefore, $K_1 = 1/P(A) = K_2$. Now, we must prove that, for all i, j, $s_{ij}^1 = s_{ij}^2$. Recall that $\mathbf{S}+$ is the case in which only the features in $\mathbf{S}$ are present, and so $\varnothing+$ is the "all absent" case. For all i, we have: $\frac{P(\varnothing+)}{P(i+)} = \frac{K_1 S_1(\varnothing+)}{K_1 S_1(i+)} = s_{ii}^1$, and similarly for s_{ii}^2. Therefore, since the left-hand side is the same for both s_{ii}^1 and s_{ii}^2, they are equal for all i. We also have, for all i,

q: $\frac{P(i+)}{P(\{i,q\}+)} = \frac{K_1 S_1(i+)}{K_1 S_1(\{i,q\}+)} = s_{ii}^{\ 1} s_{iq}^{\ 1}$, and similarly for $s_{iq}^{\ 2}$. Since $s_{ii}^{\ 1} = s_{ii}^{\ 2}$, that implies that $s_{iq}^{\ 1} = s_{iq}^{\ 2}$.

Therefore, S_1 and S_2 have identical parameters. ■

Lemma 3: If S is a feature matching model, then (i) there exists a Markov field U such that (ii) U perfectly represents P_S (that is, for all X, $S(X) \propto P(X \mid U)$), and (iii) every maximal clique potential has the *g*-decomposition property.

Proof of Lemma 3: Proof by construction: given a feature matching model, we will construct a Markov field U that has the necessary properties. Let U be the Markov field such that:

1. There exists node X_i in U iff feature F_i exists in S;
2. $X_i - X_j$ in U iff $S_{ij} \neq 1$;
3. Let N_i = # of (maximal) cliques containing X_i, and N_{ij} = # of (maximal) cliques containing both X_i and X_j.

   Then the clique potential for clique C_i is given by: $g(C_i) = \prod_{j \in C_i} (s_{jj})^{\frac{1}{N_j}[1-\delta(j)]} \prod_{j,k \in C_i} (s_{jk})^{\frac{1}{N_{jk}}[1-\delta(j)\delta(k)]}$.

Now, we must show that U, as constructed by the above method, defines a probability distribution such that, for all X, $S(X) = K * P(X \mid U)$. Clearly, the construction of the Markov field is well-defined, and each clique potential is well-defined. Therefore, we can normalize the product of clique potentials to get a well-defined probability distribution by dividing through by $Z = \sum_X \prod_i g_i(C_i)$.

Let $K = Z$, and consider an arbitrary case X: $K \times P(X \mid U) = K \times \frac{1}{Z} \prod_i g_i(C_i) = \prod_i g_i(C_i)$. Substituting in the clique potentials, this equals: $\prod_i \left[\prod_{j \in C_i} (s_{jj})^{\frac{1}{N_j}[1-\delta(j)]} \prod_{j,k \in C_i} (s_{jk})^{\frac{1}{N_{jk}}[1-\delta(j)\delta(k)]} \right]$. Every variable X_j is part of at least one clique, and X_j and X_k are part of the same clique iff $s_{jk} \neq 1$. Using these two facts, we can reorder the product to get:

$$\prod_j \left[\prod_{C_i : j \in C_i} (s_{jj})^{\frac{1}{N_j}[1-\delta(j)]} \right] \prod_{j,k : j<k \,\&\, s_{jk} \neq 1} \left[\prod_{C_i : j,k \in C_i} (s_{jk})^{\frac{1}{N_{jk}}[1-\delta(j)\delta(k)]} \right].$$

The term in the first set of brackets is either equal to 1 (if X_j is present) or $s_j^{[1/N\_j]}$ (if N_j is absent). Thus, when we multiply the N_j instances of that term together (since N_j is defined to be the number of distinct cliques in which j occurs), we get either 1 (if present) or s_j (if absent). By similar reasoning, the term in the second set of brackets is

always either 1 or s_{jk}. We can thus rewrite the product as: $\prod_j \left(s_{jj}\right)^{[1-\delta(j)]} \prod_{j,k:j<k\&s_{jk}\neq 1} \left(s_{jk}\right)^{[1-\delta(j)\delta(k)]}$. Now, notice that, if $s_{jk} = 1$, then $\left(s_{jk}\right)^{[1-\delta(j)\delta(k)]} = 1$ regardless of X_j's and X_k's presence or absence. Therefore, we can, without changing the value of the expression, multiply the product by $\prod_{j,k:j<k\&s_{jk}=1} \left(s_{jk}\right)^{[1-\delta(j)\delta(k)]}$, yielding a full product of: $\prod_{i=1}^{m}\prod_{j=i}^{m} \left(s_{ij}\right)^{[1-\delta(i)\delta(j)]}$, which is what $S(X)$ is defined to be.

Now, we must show that all of the maximal clique potentials have the *g*-decomposition property. By construction, the clique potential value for an arbitrary X is given by: $g(X) = \prod_{j\in C_i} \left(s_{jj}\right)^{\frac{1}{N_j}[1-\delta(j)]} \prod_{j,k\in C_i} \left(s_{jk}\right)^{\frac{1}{N_{jk}}[1-\delta(j)\delta(k)]}$, also representable as: $g(X) = \prod_{j\in X-} \left(s_j\right)^{1/N_j} \prod_{j\in X-;j<k\in C_i} \left(s_{jk}\right)^{1/N_{jk}}$. Using this latter representation, we can explicitly compute *g* for particular cases: $g(\varnothing+) = \prod_{j\in C_i} \left(s_j\right)^{1/N_j} \prod_{j,k\in C_i} \left(s_{jk}\right)^{1/N_{jk}}$, $g(q+) = \prod_{j\in C_i\setminus q} \left(s_j\right)^{1/N_j} \prod_{j,k\in C_i} \left(s_{jk}\right)^{1/N_{jk}}$, and $g(\{q,r\}+) = \prod_{j\in C_i\setminus\{q,r\}} \left(s_j\right)^{1/N_j} \frac{\prod_{j,k\in C_i} \left(s_{jk}\right)^{1/N_{jk}}}{\left(s_{qr}\right)^{1/N_{qr}}}$. Using these equations, we have $\frac{g(\varnothing+)}{g(q+)} = \left(s_q\right)^{1/N_q}$, and $\frac{g(q+)g(r+)}{g(\{q,r\}+)g(\varnothing+)} = \left(s_{qr}\right)^{1/N_{qr}}$. Substituting these values back into the earlier equation for $g(X)$, we can see that the *g*-decomposition property clearly holds for this (arbitrary, maximal) clique potential. ■

We now must prove the other direction of the biconditional. We do this in two steps: first, considering only a single maximal clique, and second, assembling the cliques into a full Markov field *U*.

Lemma 4: For a given probability distribution *P* and single-clique MRF *U* that perfectly represents *P*, if the clique potential has the *g*-decomposition property, then there is a feature matching model *S* such that $S(X) \propto P(X)$ for all X.

Proof of Lemma 4: Proof by construction: we show how to construct an appropriate feature matching model:

1. F_i in *S* iff X_i in *U*;
2. Set the s_{ii} and s_{ij} values as: $s_{ii} = \frac{g(\varnothing+)}{g(X_i+)}$, and $s_{ij} = \frac{g(X_i+)g(X_j+)}{g(\varnothing+)g(\{X_i,X_j\}+)}$

Clearly, this feature matching model is well-defined. And since the *g*-decomposition property holds for *g*, we have:

$$g(X) = \prod_{i \in X-} \frac{g(\varnothing +)}{g(i+)} \prod_{i \in X-; j \in X} \frac{g(i+)g(j+)}{g(\{i, j\}+)g(\varnothing +)} = \prod_{i \in X-} s_{ii} \prod_{i \in X-; i < j \in X} s_{ij} = S(X).$$

Therefore, $S(X) \propto g(X) \propto P(X)$ for all X. ■

Lemma 5: For given probability distribution *P* and perfect Markov field *U*, if there is a feature matching model for every maximal clique in *U*, then there is a feature matching model for *P*.

Proof of Lemma 5: We use $s_{ii}(C_j)$ to denote the value of s_{ii} in the feature matching model for maximal clique C_j, and define $s_{ii}(C_j) = 1$ if $X_i \notin C_j$. We define $s_{ij}(C_k)$ similarly. Now, let $s_{ii} = \prod_j s_{ii}(C_j)$ and $s_{ik} = \prod_j s_{ik}(C_j)$. These parameters are clearly well-defined. We will show that the resulting feature matching model, denoted *S**, is appropriate for *P*.

Because there is a perfect MRF *U* for *P*, we have $P(X) = \frac{1}{Z} \prod_i g_i(C_i)$ for all *X*. Note that each clique potential in *U* is determined only up to a multiplicative constant, since we normalize the product to get a probability distribution. Therefore, we choose (without loss of generality) the clique potential functions such that g_i(all present) = 1 for all *i*. Let $K = Z$, and consider $K \times P(X)$. The *K* and $1/Z$ terms cancel out, and then—since there is a feature matching model for every g_i—we can substitute the clique feature matching models into the product to get:

$$\prod_{C_i} \left[\prod_{j \in C_i} \left(s_{jj}(C_i) \right)^{[1-\delta(j)]} \prod_{j,k \in C_i} \left(s_{jk}(C_i) \right)^{[1-\delta(j)\delta(k)]} \right].$$

Reordering the product some, we get: $\prod_j \left[\prod_{i: j \in C_i} \left(s_{jj}(C_i) \right)^{[1-\delta(j)]} \right] \prod_{j,k>j} \left[\prod_{i: j,k \in C_i} \left(s_{jk}(C_i) \right)^{[1-\delta(j)\delta(k)]} \right]$. Consider the product over *j*. For any arbitrary *j*, each term in the product is either 1 (if X_j is present) or $s_j(C_i)$ (if X_j is absent). A similar argument holds for the product over *j*, *k*. But since we are taking the product over all C_i for *j* (*k*), we can rewrite the product as: $\prod_j \left(s_{jj} \right)^{[1-\delta(j)]} \prod_{j,k>j} \left(s_{jk} \right)^{[1-\delta(j)\delta(k)]}$, which is simply the feature matching model prediction for *S** for arbitrary *X*. ■

Theorem **2** follows immediately from Lemmas 1-5. ■

Corollary 1: If B is a Bayes net tree, then there is a feature matching model S such that $S(X) \propto P(X \mid B)$ for all X.

Proof of

Corollary 1: Let P denote the probability distribution defined by B. If B is a Bayes net tree, then since there are no unshielded colliders, there is a MRF U with the same skeleton that perfectly represents P. Since B is a tree, that skeleton has no triangles, and so every maximal clique has at most 2 variables. Therefore, by Theorem 1, the *g*-decomposition property holds of all of the maximal cliques in U. Therefore, by
Theorem **2**, there is a feature matching model that is proportional to U. Hence, there is a feature matching model that is proportional to P. ■

Corollary 2: There exists a probability distribution that can be modeled by a Bayes net (e.g., if the Bayes net has an unshielded collider), but for which there is no feature matching model such that $S(X) \propto P(X \mid B)$ for all X.

Proof of Corollary 2: Let P denote the probability distribution defined by B. If B contains an unshielded collider, then there is no MRF U such that U perfectly represents P. Therefore, there is no feature matching model such that $S(\mathrm{X}) \propto P(\mathrm{X})$ for all X. ■

Corollary 3: There exists a probability distribution that cannot be modeled by a Bayes net, but for which there is a feature matching model such that $S(X) \propto P(X)$ for all X.

Proof of Corollary 3: Let P be the probability distribution that is perfectly represented by the MRF $F_1 — F_3 — F_2$; $F_1 — F_4 — F_2$. There is no Bayes net for this distribution (since every possible joint edge orientation results in either a cycle or a collider). But the maximal cliques are all of size 2, and so they all have the *g*-decomposition property, and so there is a feature matching model for this distribution. ■

Proofs of Exemplar Model Theorems:

Theorem 3: [There exists an exemplar model F for ratings R] iff [there exists a noisy measurement model M and probability distribution $P^*(E)$ over the exemplars such that $P(X \mid P^*, M)$ is appropriate for R].

Proof of Theorem 3: For both directions of the biconditional, we will simply demonstrate that $F(X) \propto P(X \mid P^*, M)$.

(⇒) Assume there exists an exemplar model F with s-parameters and weights W. We will further assume that the weights are normalized so that $\Sigma W = 1$. Define the noisy measurement model to have noise parameters $q_i = 1 / (1 + s_i)$, and set $P^*(E_a) = W_a$. Since the s-parameters are positive, the q-parameters will fall in [0,1] as required. Since the weights are positive and normalized to sum to 1, they define a probability distribution. $P(X \mid P^*, M)$ is just the weighted average over E_a of $P(X \mid E_a, M)$, weighted by $P^*(E_a)$. In addition, because the noise parameters are independent of each other, we can factor each E_a-specific probability, resulting in: $P(X \mid P^*, M) = \frac{1}{r}\sum_{a=1}^{r} P^*(E_a)\prod_{i=1}^{m} P(X_i \mid E_a(i), M)$. If $X_i = E_a(i)$, then this probability is given by q_i; otherwise, it is given by $(1 - q_i)$. Therefore, if we also substitute in the $P^*(E_a)$ values, we can rewrite this expression as: $\frac{1}{r}\sum_{a=1}^{r} W_a \prod_{i=1}^{m} (q_i)^{eq(X_i, E_a(i))}(1 - q_i)^{[1-eq(X_i, E_a(i))]}$. Now divide this expression by $Z = \prod_{j=1}^{m} q_j$. (Since we are only concerned with proportionality, it doesn't matter if we divide by a positive constant.) Because q_i will cancel out when X and E_a agree on dimension i, we can rewrite this expression as: $\frac{1}{r}\sum_{a=1}^{r} W_a \prod_{i:X_i \neq E_a(i)} \frac{1 - q_i}{q_i}$. Substituting in the value for q_i yields: $\frac{1}{r}\sum_{a=1}^{r} W_a \prod_{i:X_i \neq E_a(i)} s_i$, which is just the exemplar similarity measure.

(⇐) Assume that there exists a noisy measurement model M and probability distribution P^*. Define the exemplar similarity measure using the same instances, let $s_i = \frac{1 - q_i}{q_i}$, and set $W_a = P^*(E_a)$. (Since q_i is a probability, it is less than or equal to 1, and so s_i will be non-negative.) Because this is simply the inverse of the q_i definition in the first half of the proof, and every step in the first half is reversible, this exemplar similarity measure must be proportional to $P(X \mid P^*, M)$. ■