Discovering Latent Classes in Relational Data

Charles Kemp, Thomas L. Griffiths & Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139
{ckemp,gruffydd,jbt}@mit.edu

Abstract

We present a framework for learning abstract relational knowledge, with the aim of explaining how people acquire intuitive theories of physical, biological, or social systems. Our algorithm infers a generative relational model with latent classes, simultaneously determining the kinds of entities that exist in a domain, the number of these latent classes, and the relations between classes that are possible or likely. This model goes beyond previous category-learning models in psychology, which consider the attributes associated with individual categories but not the relationships that can exist between categories. We apply this domain-general framework in two specific domains: learning the structure of kinship systems and learning causal theories.

1 Introduction

Imagine a hotel employee serving drinks at the general convention of the Episcopal church. All the guests are in casual clothes, and at first he finds it difficult to identify the people who hold the most influential positions within the church. Eventually he notices that a group of guests is treated with deference by everyone – call them the archbishops. Another group is treated with deference by everyone except the archbishops – call them the bishops. Just by observing the guests mingle, he might be able to guess the office that each person holds.

Imagine now a child who stumbles across a set of small metallic bars: magnets, although she does not know it. As she plays with the bars she notices that some some surfaces attract each other and some repel each other. She should soon realize that there are two types of surfaces – call them North poles and South poles. Each surface repels others of the same type and attracts surfaces of the opposite type.

Learning to reason about social and physical systems, as in the above scenarios, rests on the ability to discover latent structure in relational data. Both scenarios highlight latent classes (bishops and archbishops, North and South poles) that influence the relations (deference, attraction) which manifest among a set of objects (guests, metallic bars). These latent classes provide the building blocks of our intuitive domain theories, allowing us to understand and predict the interactions between an indefinite number of novel objects in each domain. While cognitive scientists have developed successful computational models [1, 2] of how people learn categories defined by the *attributes* of objects – their observable features, such as the color or mass of metallic bars – such models cannot explain how people infer the relationally defined classes of intuitive domain theories.

This paper explores one approach to explaining the acquisition of intuitive theories, in the form of a rational model for the discovery of latent classes in relational data. We define a relational generative model in which a particular relation holds for any pair of objects with some independent probability that depends only on the classes of those objects. Statisticians and sociologists have used a model of this kind, called the *stochastic blockmodel*, to analyze social networks. However, the stochastic blockmodel assumes a fixed, finite number of classes. When discovering the latent structure of a

domain, people learn the number of classes at the same time as they learn the class assignments. Our model, the *infinite blockmodel*, allows an unbounded number of classes. We provide an algorithm that can be used to simultaneously learn the number of classes and the class assignments, and use this model to explain human inferences in two important settings: learning kinship systems and learning causal theories.

2 A generative model for relational data

Suppose we are interested in a system with N objects and a single directed relation R among those objects (we will allow multiple relations later). We can represent the relation as a graph G with labelled edges, where edge g_{ij} between objects i and j has value 1 if R holds between i and j and value 0 otherwise. We want to identify the latent classes Z of the N objects, using the information contained in G. We can do this by defining a process by which Z and G are generated, and using Bayesian inference to infer Z for an observed graph G. We will define our generative model in two stages, first defining how G is generated given Z, and then defining a process by which Z is generated.

2.1 Generating relations from classes

Assume that each potential relation between two objects is generated independently, and $p(g_{ij} = 1)$, the probability that the relation holds between i and j, depends only on z_i and z_j. Given a set of assignments Z, we can write the probability of G as

$$p(G|Z, \eta) = \prod_{A,B} \eta_{AB}^{m_{AB}^1} (1 - \eta_{AB})^{m_{AB}^0} \tag{1}$$

where A and B range over all classes, η_{AB} is the probability of the relation holding between a member of class A and a member of class B, and m_{AB}^1 is the number of members of class A and class B for which the relation holds.

While this is a simple model, it is capable of expressing rich relational structure. The matrix η can be seen as specifying a *class graph*: a graph over the classes where the edge between class A and class B has weight η_{AB}, expressing which relations can hold among objects of different classes. Different kinds of relational structure correspond to different class graphs. Figure 1 shows several examples of class graphs η and object graphs G that can be defined using this model, which express a range of complex relational structures: a graph with community structure, a ring, a hierarchy and a fully connected graph. Multiple relations can be handled by assuming that each relation is conditionally independent of the others given class assignments z. Attribute information can be incorporated similarly if we assume that each attribute or relation is conditionally independent of all other attributes and relations given a set of class assignments.

2.2 Generating classes

Statisticians and sociologists have defined a model for relational data using Equation 1, assuming that the z_i are drawn from a fixed multinomial distribution over a finite number of classes [3]. This model, called the *stochastic blockmodel*, has been used to analyze the structure of various social networks. However, it does not capture one of the most important aspects of human learning: the discovery that the latent structure of a domain involves a certain number of classes.

We can define a model in which the number of classes are not fixed by choosing a different method for generating the z_i. An intuitive means of doing this is to allow the number of classes to "grow" as more objects are added to the system. Given one object, we have only a single class. As each object is added, we randomly decide whether that object is of the same class as some object we have seen before, or if it represents a new class. If the probability that a new object is of a particular class is directly proportional to the number of objects of that class seen before, the distribution over class assignments Z is that of a Chinese restaurant process (CRP). Under the CRP, the probability distribution over classes for the ith object, conditioned on the classes of the previous objects $1, \ldots, i-1$

is

$$p(z_i = A|z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_A}{i-1+\alpha} & n_A > 0 \\ \frac{\alpha}{i-1+\alpha} & A \text{ is a new class} \end{cases} \quad (2)$$

where n_A is the number of objects already assigned to class A, and α is a parameter of the distribution.

The CRP prior on Z can generate partitions with as many classes as objects, and thus potentially create countably infinitely many classes given a countably infinite number of objects. We thus call the model in which Z is generated according to Equation 2 and G is generated according to Equation 1 the *infinite blockmodel*. Other infinite models have been proposed by machine learning researchers [4, 5] using a similar construction.

2.3 Model inference

Having defined a generative model for G and Z, we can use Bayesian inference to compute a posterior distribution over Z given G:

$$p(Z|G) \propto p(G|Z)p(Z) \quad (3)$$

where $p(G|Z)$ can be derived from Equation 1, and $p(Z)$ follows from Equation 2. For the finite stochastic blockmodel, Snijders and Nowicki [3] describe a Gibbs sampler in which η and the distribution over classes are explicitly represented. We will define a Gibbs sampler for the infinite blockmodel, integrating out η and using the CRP to sample Z.

Gibbs sampling is a form of Markov chain Monte Carlo, a standard statistical tool for Bayesian inference with otherwise intractable distributionsA Gibbs sampler is a Markov chain in which the state corresponds to the variables of interest, in our case Z, and transitions result from drawing each variable from its distribution when conditioned on all other variables, in our case the conditional probability of z_i given all other assignments Z_{-i}, $p(z_i|Z_{-i})$. It follows from Equation 3 that this is

$$p(z_i|z_{-i}, G) \propto p(G|Z)p(z_i|Z_{-i}). \quad (4)$$

To compute the first term on the right hand side, we integrate out the parameters η and in Equation 1 using a symmetric Beta prior over every $\eta_A B$:

$$p(G|Z) = \prod_{A,B} \frac{\text{Beta}(m^1_{AB} + \beta, m^0_{AB} + \beta)}{\text{Beta}(\beta, \beta)} \quad (5)$$

where β is a hyperparameter. The second term follows from the fact that the CRP is exchangeable, meaning that the indices of the z_i can be permuted without affecting the probability of Z. As a consequence, we can treat z_i as the last object to be drawn from the CRP. The resulting conditional distribution follows directly from Equation 2.

To facilitate mixing, we supplement our Gibbs sampler with two Metropolis-Hastings updates. First, we consider proposals that attempt to split a class into two or to merge two existing classes [6]. Split-merge proposals allow sudden large-scale changes to the current state rather than the incremental changes characteristic of Gibbs sampling. Second, we run a Metropolis-coupled Markov Chain Monte Carlo simulation: we run several Markov chains at different temperatures and regularly consider swaps between the chains. If the coldest chain becomes trapped in a mode of the posterior distribution, the chains at higher temperatures are free to wander the state space and find other regions of high probability if they exist. To avoid free parameters, we sample the hyperparameters α and β using a Gaussian proposal distribution and an (improper) uniform prior over each.

Even though η is integrated out, it is simple to recover the class graph given Z. The maximum likelihood value of η_{AB} given z is $\frac{m^1_{AB}+\beta}{m^0_{AB}+m^1_{AB}+2\beta}$. Predictions about missing edges are also simple to compute. The probability that an unobserved edge between objects i and j has value 1 is $p(g_{ij} = 1) = \frac{m^1_{z_i z_j}+\beta}{m^0_{z_i z_j}+m^1_{z_i z_j}+2\beta}$. If some edges in graph G are missing at random, we can ignore them and maintain counts m^0_{AB} and m^1_{AB} over only the observed part of the graph.

We do not claim that the MCMC simulations used to fit our model are representative of cognitive processing. The infinite block model addresses the question of what people know about relational systems, and our simulations will show that this knowledge can be acquired from data, but we do not address the process by which this knowledge is acquired.

3 Relational and attribute models on artificial data

We ran the infinite blockmodel on the relational structures shown in Figure 1, which represent some of the structures encountered in the real world. Our algorithm solves each of these cases perfectly, finding the correct number of classes and the correct assignment of objects to classes.

To further explore our model's ability to recover the true number of classes, we gave it graphs based on randomly-generated η matrices of different dimensions. When the hyperparameter β is small, the average connectivity between blocks is usually very high or very low. As β increases, the blocks of objects are no longer so cleanly distinguished. Figure 2 shows that the model makes almost no mistakes when the β is small but recovers the true number of classes less often as β increases.

For comparison, we also evaluated the performance of a model defined on attributes rather than relations. The analogous model for attributes uses an $N \times K$ matrix F rather than the $N \times N$ relation graph G, where f_{ik} is 1 if object i possesses attribute k and 0 otherwise. Assuming that attributes are generated independently and that $p(f_{ik} = 1)$, the probability that object i has attribute k, depends only on z_i, we have

$$p(F|Z, \theta) = \prod_{A,k} \theta_{Ak}^{n_A^{1k}} (1 - \theta_{Ak})^{n_A^{0k}}$$

where the product over A, k is a product over all classes A and features k, n_A^{1k} denotes the number of objects i for which $z_i = A$ and $f_{ik} = 1$, and θ_{Ak} is the probability that feature k takes value 1 for class A. Using a CRP prior on Z, we can apply Gibbs sampling as in Equation 4 to infer $P(Z|F)$, except now we use

$$p(F|z) = \prod_{A,k} \frac{\text{Beta}(n_A^{1k} + \beta, n_A^{0k} + \beta)}{\text{Beta}(\beta, \beta)}$$

in place of Equation 5. This model is an infinite mixture model [5], and is equivalent to Anderson's rational model of categorization [1].

The infinite mixture model can be applied to these data if we convert the relational graph G into an attribute matrix F. We flattened each N by N adjacency matrix into an attribute matrix with $K = 2N$ features, one for each row and column of the matrix. For example, a matrix for the social relation "defers to" is flattened into an attribute matrix with two features corresponding to each person P: "defers to P" and "is deferred to by P". This model does well when β is small, but its performance falls off more sharply than that of the blockmodel as β increases.

4 Kinship Systems

Australian tribes are renowned among anthropologists for the complex relational structure of their kinship systems. For instance, several of these kin systems are isomorphic to the dihedral group of order eight. Even trained field workers find these systems difficult to understand [7] which raises an intriguing question of cognitive development: how do children discover the social structure of their tribe? The learning problem is particularly interesting since many communities appear to have no explicit representations of kinship rules, let alone cultural transmission of such rules. [1] We focus here on the Alyawarra, a Central Australian tribe studied extensively by Denham [8]. Using Denham's data we show that our model is able to discover some of the properties of the Alyawarra kinship system.

[1] Findler describes a case where the "extremely forceful injunction against a male person having sexual relations with his mother-in-law" could only be expressed by naming the pairs who could and could not engage in this act [7]

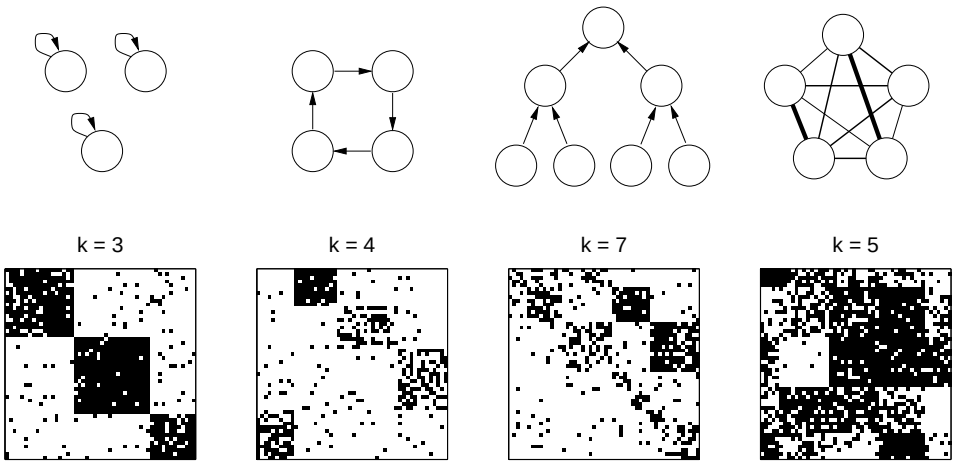


Figure 1: Class graphs (top row) and corresponding graphs over objects (bottom row). Only the edges in the class graphs with large weights are shown. Given an object graph, the infinite blockmodel perfectly recovers the true class assignments in each case.

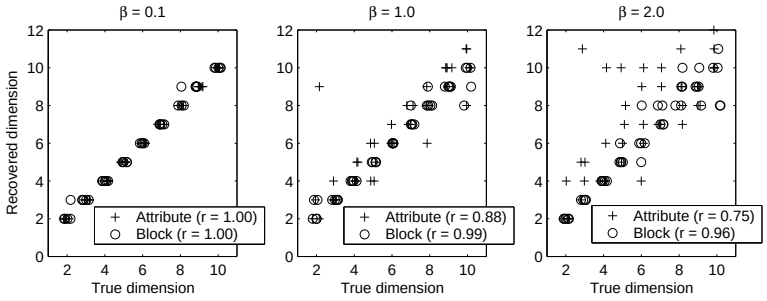


Figure 2: Model success at recovering the true number of latent classes in artificially generated data. The infinite blockmodel performs better than the infinite mixture model as the hyperparameter β increases.

Denham took photographs of 225 people, and asked 104 of them to provide a single kinship term for the subject of each photograph in the collection. We analyze the 104 by 104 square submatrix of the full 104 by 225 matrix of relations. Figure 3 shows three of 27 different kinship terms recorded. For each term, the (i, j) cell in the corresponding matrix is shaded if person i used that term to refer to person j. The Alyawarra have four kinship sections which are clearly visible in the first two matrices. 'Adiadya' refers to a classificatory younger brother or sister: that is, to a younger person in one's own section, even if he or she is not a biological sibling. 'Umbaidya' is used by female speakers to refer to a classificatory son or daughter, and by male speakers to refer to the child of a classificatory sister. We see from the matrix that women in section 1 have children in section 4, and vice versa. 'Anowadya' refers to a potential marriage partner. The eight rough blocks indicate that that men in section 1 may marry women from section 2, men in section 3 may marry women from section 4, and so on. These marriage restrictions are one example of the important behavioral consequences of the Alyawarra kinship system.

We fit the infinite blockmodel to all 27 kin-relation matrices simultaneously, treating each matrix as conditionally independent of all the others given an assignment of objects to classes. The maximum likelihood solution is represented in Figure 3. Denham recorded the age, gender and kinship section of each of his informants, and Figure 3 shows the composition of each class along each of these dimensions. The six age categories were chosen by Denham, and reflect his knowledge of Alyawarra terms for age groupings [8].

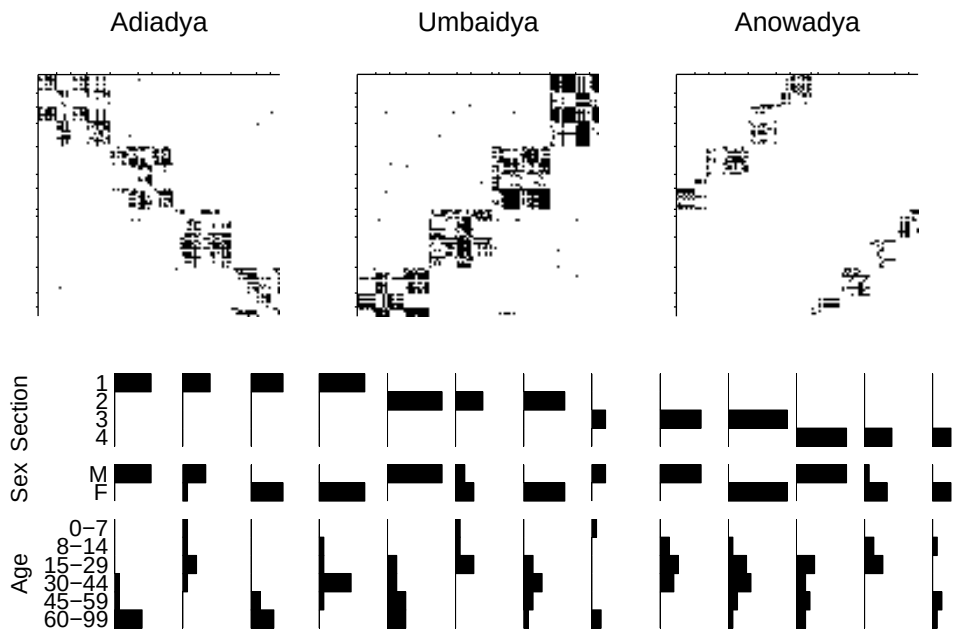


Figure 3: Top: Object graphs for three Alyawarra kinship terms described in the text. The 104 individuals are sorted by the 13 classes found by the infinite blockmodel (tick marks separate latent classes). Bottom: Breakdown of the 13 classes by kinship section, gender and age.

The blockmodel finds 13 classes, each of which includes members of just one kinship section. Section 1 is split into four classes corresponding to older men, older women, younger men and younger women. Section 3 is split into three classes: younger men, older men, and women. The remaining sections have one class for the younger people, and a class each for older men and older women. Note that none of the demographic data were used to fit the model — the 13 classes were discovered purely from the relational kinship data.

When given the same data, the maximum likelihood partition found by the purely attribute-based infinite mixture model is qualitatively worse. It includes only 5 classes: one for each of three kin sections, and two for Section 1 (split into older and younger people). We might expect that the true class structure has at least 16 classes (4 sections by 2 genders by 2 age categories) and probably more, since the age dimension might be broken into more than two categories. While the blockmodel comes much closer to this ideal, it clearly has limitations, such as failing to represent the higher-order relationships (hierarchical or factorial) between the classes. We are currently exploring such extensions.

5 Causal Theories

Tenenbaum and Niyogi (2003) studied people's ability to learn simple causal theories in situations similar to the magnetism example mentioned earlier. Here we use the infinite blockmodel to explain some of their findings. Their subjects were placed in a virtual world, where they were able to move around a set of identical-looking objects. Some objects "activate" other objects whenever they touch. If x activates y (denoted $x \rightarrow y$), then y lights up and beeps whenever x and y touch. In some worlds, activation is symmetric (denoted $x - y$): both x and y light up and beep. Unknown to the subjects, each object is in one of two classes, A or B, which determines its activation relations. Figure 5 shows the theories used in four different conditions of the experiment, expressed as class graphs, as well as graphs of the activation relations over objects generated by these theories. In the first two worlds, every A activates (asymmetrically or symmetrically) every B. In the remaining two

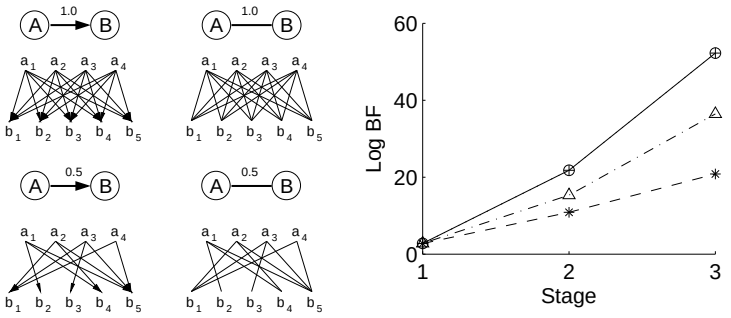
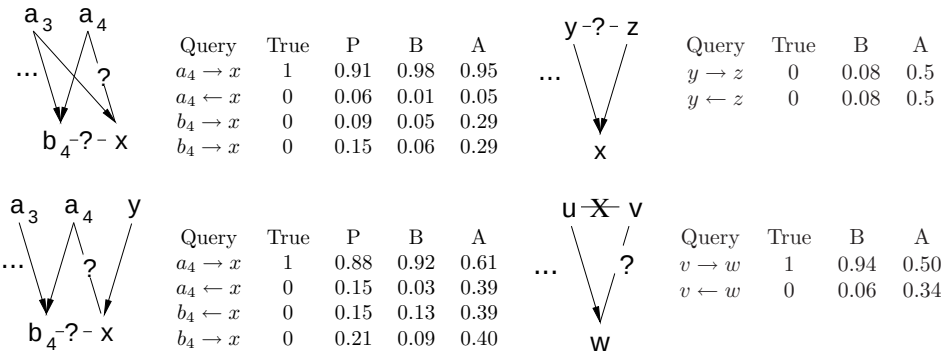


Figure 4: Left: the four class graphs used in the experiments of Tenenbaum and Niyogi. Right: Bayes factors (y-axis) for the first three stages (x-axis) of each experiment comparing the infinite blockmodel to a null hypothesis where each object is placed in its own class. ($\circ$: $A \rightarrow B$, $+$: $A - B$, $\triangle$: $A \overset{0.5}{\rightarrow} B$, $*$: $A \overset{0.5}{-} B$)



Query	True	P	B	A
$a_4 \rightarrow x$	1	0.91	0.98	0.95
$a_4 \leftarrow x$	0	0.06	0.01	0.05
$b_4 \rightarrow x$	0	0.09	0.05	0.29
$b_4 \rightarrow x$	0	0.15	0.06	0.29

Query	True	B	A
$y \rightarrow z$	0	0.08	0.5
$y \leftarrow z$	0	0.08	0.5

Query	True	P	B	A
$a_4 \rightarrow x$	1	0.88	0.92	0.61
$a_4 \leftarrow x$	0	0.15	0.03	0.39
$b_4 \leftarrow x$	0	0.15	0.13	0.39
$b_4 \rightarrow x$	0	0.21	0.09	0.40

Query	True	B	A
$v \rightarrow w$	1	0.94	0.50
$v \leftarrow w$	0	0.06	0.34

Figure 5: Predictions about new objects (v, w, y, z), after seeing old objects from the theory $A \rightarrow B$. Edges with question marks show activation relations to be predicted. The cross on the edge between u and v indicates that u and v have been observed not to activate each other. Tables show predictions of experimental subjects (P), of the infinite blockmodel (B) and of the infinite mixture model (A).

worlds, each A activates (asymmetrically or symmetrically) a random subset (on average, 50%) of B's. These four theories all correspond to stochastic blockmodels. They can be denoted using class graphs as $A \rightarrow B$, $A - B$, $A \overset{0.5}{\rightarrow} B$ and $A \overset{0.5}{-} B$, respectively.

Tenenbaum and Niyogi (2003) examined whether subjects can discover these simple theories after interacting with some subset of the objects. Their experiments had seven phases, and three new objects were addeded to the screen during each phase (see [9] for details). As new objects were added, subjects made predictions about how these objects would interact with old objects or with each other. At the end of the experiment, subjects also verbally described how the objects work. No mention of classes was made during the instructions, so inferring the existence of two classes and the relation between them constitutes a genuine discovery.

We consider two aspects of these experiments: the relative difficulty of learning the four theories shown in Figure 5, and the specific predictions that people make about relations for new objects after they have learned one of these theories. Given experience with 18 objects, people had no difficulty learning the two deterministic theories (with $\eta_{AB} = 1$): $A \rightarrow B$ and $A - B$. The asymmetric nondeterministic structure, $A \overset{0.5}{\rightarrow} B$, was much more difficult; only about half of 18 subjects succeeded on this task. The symmetric nondeterministic structure, $A \overset{0.5}{-} B$, was the most difficult; only two out of 18 subjects attained even partial success.

These findings are consistent with the behavior of a Bayesian learner inferring the theory that best

explains the observed relations. The weight of the evidence that the world respects a block structure can be expressed as the marginal likelihood of the observed relational data under the infinite block model. We computed these likelihoods by enumerating then summing over all possible class assignments Z for up to 9 objects. Figure 5 plots Bayes factors (log ratio of evidence terms) for the infinite blockmodel relative to a "null hypothesis" where each object belongs to its own class. The Bayes factors increase in all cases as more objects and relations are observed, but the rate of increase varies across the four theories in accordance with their relative ease of learning.

Learning the correct causal theory based on a set of observed relations should allow people to infer the unobserved causal relations that will hold for a new object x in the same domain – as long as they observe sufficient data to infer the class membership of x. Figure 5 shows several kinds of relational prediction that human learners can perform. All of these examples assume a learner who has observed the objects and relations in Figure 5 generated by the $A \rightarrow B$ theory. Given a new object x which has just been activated by an old A object, a learner with the correct theory should classify x as a B, and thus predict that another A will activate x, but that nothing will happen between x and a B. Analogous predictions can be made if x is observed only to be activated by a new object y. Figure 5 shows that people make these predictions correctly after learning the theory [9], as does the infinite blockmodel. The infinite mixture model performs poorly on these tasks (Figure 5) as a consequence of treating relations like attributes. Under this model, learning about relations between new objects is identical to learning about entirely new features, and *none* of the learner's previous experience is relevant. By treating relations properly, the blockmodel offers a qualitative increase in representational power over previous attribute-based models of concept learning. Only the blockmodel thus accounts for a principle function of intuitive theories: to support generalizations from previous experience to wholly new systems in the same domain.

6 Conclusions and future directions

We have presented an infinite generative model for representing abstract relational knowledge and discovering the latent classes generating those relations. This analysis hardly begins to approach the richness and flexibility of people's intuitive domain theories, but may at least provide some of the critical building blocks. It may also be of use in other fields. Our framework for discovering latent classes de novo, when even their number is unknown, may be seen as an extension of relational models previously proposed in mathematical anthropology (stochastic block models [3]) and machine learning (probabilistic relational models (PRMs) [10] . We are also exploring the cognitive relevance of other kinds of relational structures proposed in machine learning and anthropology, such as the overlapping class model of Kubica et al. [11], or structures (where each object in class A can and must relate to exactly one object in some other class B). Developing a framework that can form spotaneous and flexible combinations of these structures remains a formidable open task.

References

[1] J. R. Anderson. The adaptive nature of human categorization. 98(3):409–429, 1991.

[2] R. M. Nosofsky. Attention, similarity, and the identification-categorization relationship. 115:39–57, 1986.

[3] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[4] R. M. Neal. Bayesian mixture modeling by monte carlo simulation. Technical Report 91-2, University of Toronto, 1991.

[5] C. E. Rasmussen. The infinite Gaussian mixture model. In *NIPS*, volume 13, 2002.

[6] S. Jain and R. M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. Technical report, University of Toronto, 2000.

[7] Nicholas Findler. Automated rule discovery for field work in anthropology.

[8] Woodrow Denham. The detection of patterns in alyawarra nonverbal behavior. 1973.

[9] J. B. Tenenbaum and S. Niyogi. Learning causal laws. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. 2003.

[10] B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clusterin in relational data. In *IJCAI*, volume 15, 2001.

[11] J. Schneider J. Kubica, A.Moore and Y. Yang. Stochastic link and group detection. In *IJCAI*, 2002.