Analogy as Nonparametric Bayesian Inference over Relational Systems

Ruairidh M. Battleday[1] **and Thomas L. Griffiths**[1,2]
[1]Department of Computer Science, Princeton University
[2]Department of Psychology, Princeton University
{battleday,tomg}@princeton.edu

Abstract

Much of human learning and inference can be framed within the computational problem of relational generalization. In this project, we propose a Bayesian model that generalizes relational knowledge to novel environments by analogically weighting predictions from previously encountered relational structures. First, we show that this learner outperforms a naive, theory-based learner on relational data derived from random- and Wikipedia-based systems when experience with the environment is small. Next, we show how our formalization of analogical similarity translates to the selection and weighting of analogies. Finally, we combine the analogy- and theory-based learners in a single nonparametric Bayesian model, and show that optimal relational generalization transitions from relying on analogies to building a theory of the novel system with increasing experience in it. Beyond predicting unobserved interactions better than either baseline, this formalization gives a computational-level perspective on the formation and abstraction of analogies themselves.

Keywords: generalization; inference; analogy; Bayesian models, nonparametric statistics.

Introduction

The problem of relational generalization—how a learner may use previously acquired relational knowledge to infer the nature of unobserved relationships—is fundamental to cognition. Indeed, much of the knowledge and inferential ability we regard as quintessentially human—learning and acting with little experience in unfamiliar environments, formal reasoning and discovery in mathematics and the sciences, and our intricate social and artistic interactions and understandings—are inherently relational (Euclid, trans. 1956; Hofstadter, 1979; Kuhn, 2012; Lakoff & Johnson, 2008; Law et al., 1999; Tenenbaum, Griffiths, & Kemp, 2006).

It is possible to view the problem of relational generalization as one of probabilistic inference. This is the approach taken in several recent probabilistic models of "theory learning" (Kemp, Griffiths, & Tenenbaum, 2004; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006; Kemp, Tenenbaum, Niyogi, & Griffiths, 2010). These models take the perspective of a naive learner, gradually inferring a class-level theory of a novel system by observing an increasing number of interactions within it. The paradigmatic example is of a child playing with different objects—composed of magnetic, ferrous, and plastic materials, say—and learning the relationships between them. However, we often want to consider situations in which a learner has acquired some relevant knowledge about relational systems that they can bring to the current inferential task. A second influential account of relational generalization has been proposed with this aim, taking the perspective of a mature learner making inferences about a novel relational system based on abstract knowledge of a specific set of underlying structural forms—for example, tree hierarchies or psychological spaces (Kemp & Tenenbaum, 2008). Here, the paradigmatic example is of a scientist, inferring the best organizing principles for a set of data they are analyzing from a possible set defined by the community.

In both of these approaches, successful relational generalization results from inferring the best organizing relational structure for a set of data, guided by prior assumptions about the likelihood of different structures. Instead of specifying such a prior distribution in advance, we can use nonparametric Bayesian statistics to induce one directly. That is, instead of beginning from general constraints, like the theory-based learner, or a fixed set of specific generative rules, like the form-based learner, we define a learner that updates inductive biases over the space of latent relational structures throughout their lifetime. The paradigmatic example becomes an adult in an unfamiliar environment using their prior experience with related relational systems as their guide. Interestingly, thinking about relational generalization in this way leads us to something that resembles analogy: we are re-using experience from past systems in proportion to their match with the current one (Gentner, 2010; Halford, Wilson, & Phillips, 2010; Holyoak, 2012).

In this paper, we develop an analogy-based model of relational generalization that instantiates these principles and provides a framework to unify existing probabilistic approaches. Analogy itself has been well investigated at Marr's algorithmic level (Marr, 1982), and we discuss how these concepts relate to our computational-level perspective. We then compare the performance of our strategy to the theory-based learner using relational data from random systems and the Wikipedia hyperlink graph, and examine the analogies it returns. By making the model fully nonparametric we are able to capture the benefits of both models, and show the conditions under which using an analogy or learning a theory is best. This unification gives an interesting perspective on how analogies are formed, and how inductive biases over relational systems might transform through the lifetime of a learner.

The Computational Problem of Relational Generalization

We can use the intuition above to construct a general framework for the computational problem of relational generalization. First, we can consider an environment to comprise a set of entities, all of which can "interact" under a number of relational types. For example, these might be inanimate objects, the relational predicate might be "moves when brought into contact with", and an interaction might be whether the object(s) move or not (in other words, it is binary). The goal of the agent is to make the optimal inference as to the nature of an unobserved interaction, r, based on a set of previously observed interactions, $\mathcal{R}$:

$$p(r|\mathcal{R}). \tag{1}$$

In Bayesian statistics, this is known as the marginal likelihood of the (unobserved) data, and may be found by marginalizing across all possible values of a set of parameters, θ:

$$p(r|\mathcal{R}) = \int_{\theta} p(r|\theta)\, p(\theta|\mathcal{R})\, d\theta. \tag{2}$$

We see that the insertion of parameters and assumption of independence of interactions given those parameters allows us to separate our inference into two parts: a prediction of the unobserved interaction given a set of parameters, and the evaluation of the posterior probability of those parameters given our observed data. It is in this second term that our various models differ, and the prior experience over relational structures can be defined.

Theory-based Relational Generalization

In probabilistic models of theory formation, the goal of the learner is to identify the best "theory" of the relational system underlying a set of observed interactions between entities (Kemp et al., 2004, 2006, 2010). A learner begins with general prior knowledge, and must then evaluate every possible theory of the data in order to make the inference. A theory itself is an assignment of entities to classes and a class-level relational system that gives the probability of each potential relation between members of different classes—a representation that has strong support from the prevalent use of folk and framework theories in the developmental and adult psychological literatures (Carey, 1986; Gopnik, Meltzoff, & Bryant, 1997). The key modeling insight is that although the number of possible theories grows combinatorially with the number of interacting entities, a preference for simpler theories allows the learner to make meaningful relational generalizations that accord well with human behaviour and relational systems of knowledge—the combinatorial equivalent of Occam's razor.

Mathematically, a theory is formalized as a matrix, η, that gives the likelihood of entities from different classes interacting, and a vector $\mathbf{z}$ that assigns each element to a latent class.

A learner can then infer the best theory for a system by evaluating the posterior probability over theories given a set of relational data, $\mathcal{R}$:

$$\arg\max_{\eta,\mathbf{z}} p(\eta, \mathbf{z}|\mathcal{R}). \tag{3}$$

We can invert this posterior probability by an application of Bayes' rule, giving us a hypothesis space of generative models to search across:

$$p(\eta, \mathbf{z}|\mathcal{R}) \propto p(\mathcal{R}|\eta, \mathbf{z}) p(\eta, \mathbf{z}). \tag{4}$$

Here, the prior over relational systems, $p(\mathbf{z}, \eta)$, is typically derived from the Chinese Restaurant Process (CRP; Aldous, 1985). This means that it is defined over all possible clustering of entities into latent classes—or, possible seating arrangements of patrons at a restaurant—and is *nonparametric*, in that it grows freely with the number of entities—the restaurant has infinite capacity. The preference for simplicity is encoded such that theories that assign entities to fewer and larger classes are weighted more heavily; in the restaurant analogy, this corresponds to preferring seating arrangements involving a small number of large tables. This ability to consider a potentially infinite number of relational structures is explicit in the name of one of these models: the Infinite Relational Model (IRM; Kemp et al., 2006).

Analogy-based Relational Generalization

The IRM gives us a way to interpret the basic types and quantities needed for a probabilistic model of relational generalization. First, the minimal parameters needed to define a relational structure are a class-level relation matrix and a class-assignment vector to map entities to those classes (where the latter is really a latent variable). That is, $\theta = (\eta, \mathbf{z})$. Second, the theory-based model can be thought of as one way of defining and evaluating a prior over those structures, where every structure is possible and allocated some probability.

We would like to be able to induce relations in the same manner as the IRM, but define a prior distribution over relational structures that captures a learner's previously internalized experiences with relational systems. Nonparametric Bayesian statistics allows us to do this by considering each of these experiences as a point in the space of possible relational structures that may grow in number over time. The learner can consider the posterior probability of parameter values given observed interactions in the context of these analogies:

$$p(\theta|\mathcal{R}) \propto \sum_{k=1}^{K} p(\mathcal{R}|\theta^{(k)}) p(\theta^{(k)}) \tag{5}$$

$$\propto \sum_{k=1}^{K} p(\mathcal{R}|\eta^{(k)}, \mathbf{z}^{(k)}) p(\eta^{(k)}, \mathbf{z}^{(k)}), \tag{6}$$

where we have K previously encountered systems, and have replaced the general parameters for a system, $\theta^{(k)}$, by a class-level interaction matrix and class-assignment vector, $\eta^{(k)}$ and $\mathbf{z}^{(k)}$, respectively. This effectively defines a mixture model over relational kernels, with weights given by their prior probability. The predictions made based on this model will be

weighted by the posterior probability of each system's parameters. We interpret this posterior probability as the analogical similarity between current and previous systems, and observe that it decomposes into the likelihood of the current environment's observed interactions under the previously encountered system, along with the prior probability of that system.

It is interesting to compare the assumptions and implications of the theory- and analogy-based models. First, the theory-based model, the IRM, is nonparametric in that the set of relational structures it considers grows with the number of entities it assigns to classes. The analogy-based model is nonparametric in that it grows with the number of relational systems available for analogical comparison. Second, the theory-based model is flexible—given enough data it can learn the optimal relational structure for any set of interacting entities. However, this flexibility comes at a cost: when few interactions in a novel environment have been observed its inductive biases may prove too general to support accurate predictions. By contrast, although the analogy model considers only a subset of possible relational structures, this allows it to make strong predictions with fewer samples; provided at least one is analogically relevant.

Monte Carlo Inference

The principles above give the optimal way of making relational generalizations. However, because the space of relational configurations grows combinatorially with the number of entities under consideration, exact inference is often intractable. Instead, we can use sampling techniques to approximate the distributions given above, based on the Monte Carlo principle. For our problem, this means that a prediction about the unobserved interaction, r, based on samples of parameters from the posterior distributions described above, $\theta^{(q)}$, will come arbitrarily close to the true model prediction as the number of samples, Q, grows:

$$\int_{\theta} p(r|\theta)\, p(\theta|\mathcal{R})\, d\theta \approx \frac{1}{Q}\sum_{l=1}^{Q} p(r|\theta^{(q)}). \tag{7}$$

In practice, these samples are generated via Markov chain Monte Carlo (MCMC; Neal, 1992). Making this approximation for the IRM is simple: draw samples of parameters from the posterior defined by the model, make the prediction regarding the unobserved interaction based on these samples, and take the average, relying on the samples to be provided at a proportion determined by the underlying posterior density:

$$p(r|\mathcal{R})_{IRM} \approx \frac{1}{Q}\sum_{l=1}^{Q} p(r|\eta^{(q)}, \mathbf{z}^{(q)}). \tag{8}$$

For the analogy model, the situation is more involved. We can consider each system as providing samples of parameters from the underlying mixture components. However, calculating the ratio of samples from each system that should be used for each prediction is difficult: it involves the true ratio of stored system posteriors, exactly the quantities that are intractable to compute. Instead, we can take an equal number of predictions from each system, and weight these predictions by an estimate of how much each system contributes to the posterior distribution over parameters. This is equivalent to our estimate of that system's analogical similarity to the current learning environment. Using hat notation to denote estimators, this translates to the following expression:

$$w_k = \frac{\hat{p}(\mathcal{R}|S^{(k)})p(S^{(k)})}{\sum_{k'}\hat{p}(\mathcal{R}|S^{(k')})p(S^{(k')})} \tag{9}$$

Provided we use a uniform prior over systems, we can use importance sampling to form an estimator of analogical similarity (or, model evidence), based on the samples that we have already obtained. This relies on the following harmonic mean estimator (Kass & Raftery, 1995):

$$\hat{p}(\mathcal{R}|S^{(k)}) = \Big[\frac{1}{Q}\sum_{q=1}^{Q} p(\mathcal{R}|\theta^{(qk)})^{-1}\Big]^{-1}. \tag{10}$$

Some intuition can be gained about this estimator by considering that terms with small likelihoods will contribute more towards the sum, and decrease the model evidence.

By making the above approximations, we arrive at the following expression for predictions from the analogy-based model:

$$p(r|\mathcal{R})_{analogy} \approx \frac{1}{Q}\sum_{q=1}^{Q}\sum_{k=1}^{K} p(r|\eta^{(kq)}, \mathbf{z}^{(kq)}) \cdot w_k. \tag{11}$$

Correspondence with Previous Theories of Analogy

An early and influential theory of analogy in cognitive science was Gentner's structure mapping theory (SMT; Gentner, 1983), which explores the process of how a learner might best map two relational structures to one another—a theory at Marr's algorithmic level (Marr, 1982). Although the underlying computational problem of finding isomorphisms between two graphs is itself computationally intractable (Garey & Johnson, 1979), the approach achieves notable success in finding efficient mappings and matching human intuition and behavior (see Gentner and Forbus, 2011). Keane, Ledgeway, and Duff (1994) identify the following computational-level assumptions that allow these models to do this, and that justify calling a comparison between two domains "analogical": only making matches between entities of the same type; leveraging structural consistency across representations; and, favoring systematic matches. It is regarding these that our account can offer some insight. First, the IRM and our analogy model can be extended to simultaneously cluster relations *and* features (Kemp et al., 2010), and we expect that when the latter are taken into account matching between entities of the same type will arise naturally. Next, in SMT the use of a systematicity score is justified by an appeal to intuition: a match is more analogical the deeper the relational correspondence

between the two structures (Gentner, 1983). In our formulation, an analogical match is more useful for inference if the posterior probability of the base system is higher given the data we have observed from the target. This will be greater if the hierarchical depth of both models is consequential for predicting sample-based variability.

The literatures on schema-based learning and production systems also propose a prominent role for analogy at the algorithmic level. These often take relational generalization to be the implicit or explicit goal of the system being presented, during, for example, learning and reasoning based on prior structural information (Pirolli & Anderson, 1985; Tse et al., 2007), schema-induction and analogical mapping (Halford, Bain, Maybery, & Andrews, 1998; Halford & Wilson, 1980), skill learning and problem solving (Anderson & Thompson, 1989), and in modeling relational understanding during cognitive development (Doumas, Hummel, & Sandhofer, 2008; Leech, Mareschal, & Cooper, 2008). Again, many of their assumptions and empirical findings are relevant to our approach, and we hope to be able to analyze them in subsequent work. Finally, there is already at least one computational-level Bayesian model of relational generalization, "Bayesian analogy with relational transformations" (Lu, Chen, & Holyoak, 2012), that has been applied to a more limited subdomain: predicting comparative judgments of relations between vector-space embeddings of animal concepts. By proposing the general modeling framework above, we aim to provide a pathway to formally connect with these approaches and their results.

Simulating Relational Generalization

The statistical discussion above gives two main predictions. The first is that the theory-based model should perform increasingly well as the number of interactions it observes increases. The second is that the analogy-based model should perform increasingly well as the number of stored systems increases. We can test both of these predictions, and examine the interplay between the two, by comparing the ability of both models to predict unobserved interactions in random simulated systems and systems derived from the hyperlink graph of Wikipedia. Wikipedia is an online open-source and community-maintained information repository that uses web pages to explain concepts and reference facts, and hyperlinks between pages to specify relationships. It is a promising candidate for approximating human relational knowledge because it encodes rich relational structure through the hyperlinks between pages, and comes with a categorization framework that allow us to cluster classes into systems. We assess each model by the (log) probability of model predictions on a set of held-out interactions. We vary the number of observed interactions between $10-90\%$ and use the remaining 10% as test interactions. We then compare the performance of the IRM to learners that use 2, 5, 10, and 100 stored systems for analogical inference. We simulated systems of 30 entities, giving 900 possible binary interactions.

Model Specification

For the IRM, we have two elements to specify: the likelihood of an interaction given the model parameters, and the prior probability of those model parameters. For binary interactions, our likelihood model becomes as follows:

$$R_{ij}|\mathbf{z},\eta \sim \text{ Bernoulli}(\eta_{\mathbf{z}_i,\mathbf{z}_j}); \tag{12}$$

we also assume interactions are independent given these parameters. From this, we have a parameter for the likelihood of elements from each pair of classes interacting, $\eta_{A,B}$, and a class assignment latent variable for each element i, $\mathbf{z}_i$. We use the following prior distributions over these:

$$\eta_{A,B}|\alpha,\beta \sim \text{Beta}(\alpha,\beta), \tag{13}$$

$$\mathbf{z}|\gamma \sim \text{Chinese Restaurant Process}(\gamma). \tag{14}$$

The CRP is a discrete-time stochastic process that assigns a probability distribution over all possible class assignments of our known entities. It states that given a set of entities, entity i is assigned to a class based on the number of elements currently assigned to that class, N_A, or a new class with probability proportional to the hyperparameter γ:

$$P(z_{i+1}=A|z_1,\ldots,z_i,\gamma)=\begin{cases}\frac{N_A}{N+\gamma} \text{ if } N_A>0\\ \frac{\gamma}{N+\gamma} \text{ if A is a new class}\end{cases}; \tag{15}$$

this "rich-get-richer" property means it can be used as a complexity-limiting prior. The CRP is exchangable over arrival order, allowing us to sequentially base each class-assignment on the current assignment of all other entities (Aldous, 1985). We also use the following distributions to provide uncertainty on the entries of η and $\mathbf{z}$, with $\beta=\alpha$:

$$\alpha \sim \text{Improper }, p(\alpha)\propto \alpha^{-\frac{5}{2}}, \tag{16}$$

$$\gamma \sim \text{Exponential}(1), \tag{17}$$

For the analogy model, we define a stored relational structure as a pair (η,ζ), where η is a class-level relation matrix, and ζ is a class-probability vector. Given this (fixed) information, we can generate each $z_i^{(k)}$ independently for system k as follows:

$$z_i^{(k)}|\zeta^{(k)} \sim \text{ Multinomial }(\zeta^{(k)}). \tag{18}$$

We can then use the same likelihood model as the IRM for each system, and a uniform prior over system parameters:

$$p(\eta^{(k)},\zeta^{(k)})=\frac{1}{K}. \tag{19}$$

We conduct inference over all latent variables and parameters using Metropolis-Gibbs MCMC (Neal, 1992).

Synthetic Data

Following Kemp et al. (2004), we view a random system as the result of an inductive process of theory acquisition. We generate a class-assignment vector from the CRP, a class-level relational matrix from the IRM, and then a set of binary

interactions between entities based on these sampled parameters and our sampling model. The class-probability vector required for the analogy model may be derived from the proportion of entities in each class in the true assignment vector. We limited consideration to 101 systems of between three and six classes, inclusive, as this matches the Wikipedia data, detailed below.

Wikipedia Data

For the Wikipedia dataset, begin with the "Wikipedia network of top categories" database, which gives the largest strongly connected component of categories with over 100 constituent pages from Wikipedia in September, 2011 (Klymko, Gleich, & Kolda, 2014). We then use the Wikipedia API to find the supercategory for each of these categories (hereafter, "classes"), to serve as a system name, and limit our consideration to systems with between three and six classes (and with intelligible and non-self-referential class and system names). We then use this reduced graph to construct our representations of systems, and select 101 of these at random to conduct model inference over. We can then represent these systems by a class-level interaction matrix and class-probability vector, where $\eta_{m,n}$ represents the proportion of webpages from subcategory m that hyperlink to subcategory n, and θ_m is the proportion of pages that come from subcategory m. Finally, we can simulate data directly from this model by drawing class-labels for 30 entities (simulated pages) from the class-probability vector, and interactions (simulated hyperlinks) based on the class-relation matrix. We limit consideration to these systems because those with a fewer classes are frequently degenerate, having an extremely high concentration of probability mass in a single class or inter-class relation, and those with more are relatively few in number and underdetermined by the number of entities we conduct inference over. We leave overcoming these numeric issues to subsequent work.

Results

We find that on the synthetic and Wikipedia data using around five and two analogies is sufficient to match the performance of the IRM, respectively (see Figure 1). When the number of interactions is very small, no model performs well; likely because there are not enough data to definitively group entities. After this, there is a slight and decreasing benefit from using analogies, as predicted. However, the ultimate benefits from using a theory are not as evident as predicted. We suspect that this, along with the extremely good performance of the 10- and 100-system analogy models, is because the space of possible relational structures generated from 30 entities is not large enough to require such flexibility, and is well covered by simple examples. Some evidence for this can be seen by inspecting the best analogies drawn by the full model (see Figure 2), which although often interesting seem to imply the interactional structure in the data was not rich enough to support more fine-grained matches and theories. Nevertheless, it is interesting that the analogy model based on human-structured relational knowledge (Wikipedia) requires fewer of those systems to compete with the IRM.

Unifying Models of Relational Generalization

From the above analyses, two further questions arise. The first is whether we can combine the benefits of both modeling approaches, with more accurate predictions in the small-sample regime and the flexibility to learn the structure of the new environment as observed data increases. Fortunately, because the posterior of both modeling approaches can be expressed in equivalent forms, it is straightforward to combine these models in a fully nonparametric manner. That is, we can consider generalizations based on $K+1$ systems, where the first K come from previously stored systems, and the last comes from a new system inferred by the theory-based model. Predictions based on samples from these systems can be weighted in the following manner, which builds upon the estimator given above by including a nonparametric prior:

$$p(r|\mathcal{R}) \approx \frac{1}{Q}\sum_{q=1}^{Q}\sum_{k=1}^{K+1} p(r|S^{(k)}) \cdot w_k \tag{20}$$

$$w_k = \frac{\hat{p}(\mathcal{R}|S^{(k)})p(S^{(k)})}{\hat{p}(\mathcal{R})} \tag{21}$$

$$p(S^{(k)}) = \begin{cases} \frac{1}{N-1+\tau} \text{ if } k \leq K \\ \frac{\tau}{N-1+\tau} \text{ if } k = K+1 \end{cases}, \tag{22}$$

where $\hat{p}(\cdot)$ is the harmonic mean estimator, and τ is the weight of predictions from the theory-based model. We optimize τ *post-hoc* by numeric maximization using Brent's algorithm.

When we examine results from this model, we find that it interpolates well between theory- and analogy-based models, and outperforms both across all systems and data partitions (Figure 3). For models using fewer analogies, we also see the predicted transition to the use of theories as the amount of observed data grows, captured in the weight assigned to the solution provided by the IRM. Although this is not as striking as predicted, we suspect that both the sharpness of the transition and the performance benefit of the nonparametric model will continue to grow as more entities are considered, and as our estimator of model evidence improves (this was often poorly aligned with MCMC performance, and is known to have stability issues; Kass and Raftery, 1995).

The second question is where these analogies might come from. The unification given above offers some insight, particularly through the nonparametric prior. Here we have considered a single inference, made by combining accrued relational experience with the flexibility to consider the system novel. Spread over a lifetime, this ability would allow a learner to induce the right level of abstraction over structures to support the types of relational generalizations they are likely to require in the future. This could explain how an abstract form like a tree hierarchy is induced from experience with many tree-like systems. To make this extension, we will need to

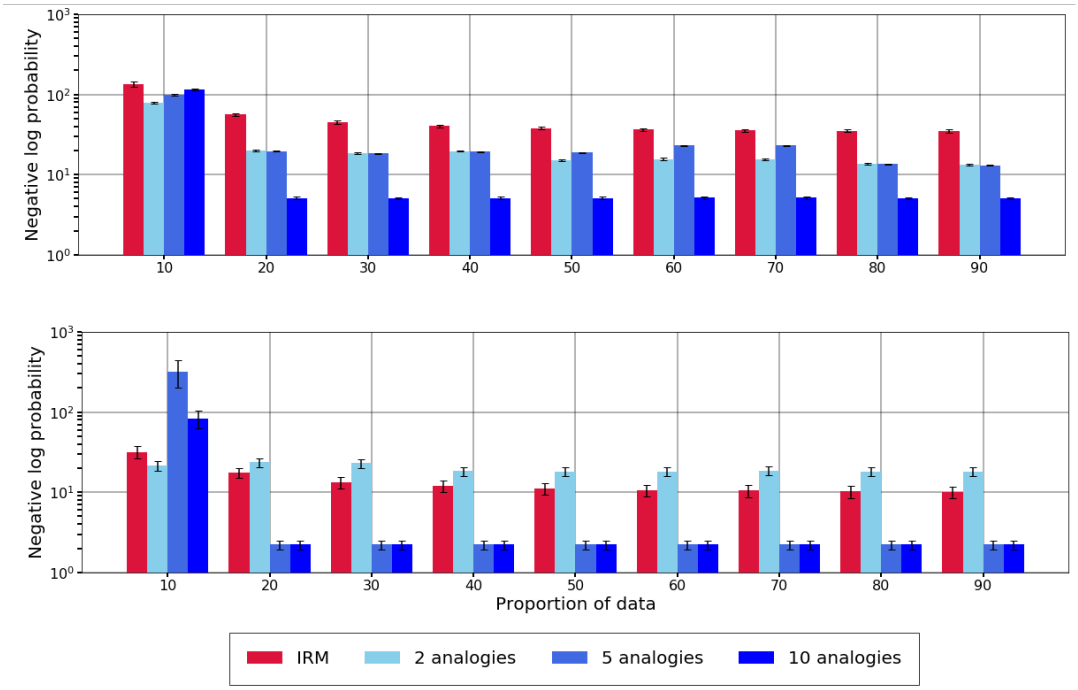

Figure 1: Results for IRM and analogy models over random (top) and Wikipedia systems (bottom; lower scores represent better performance).

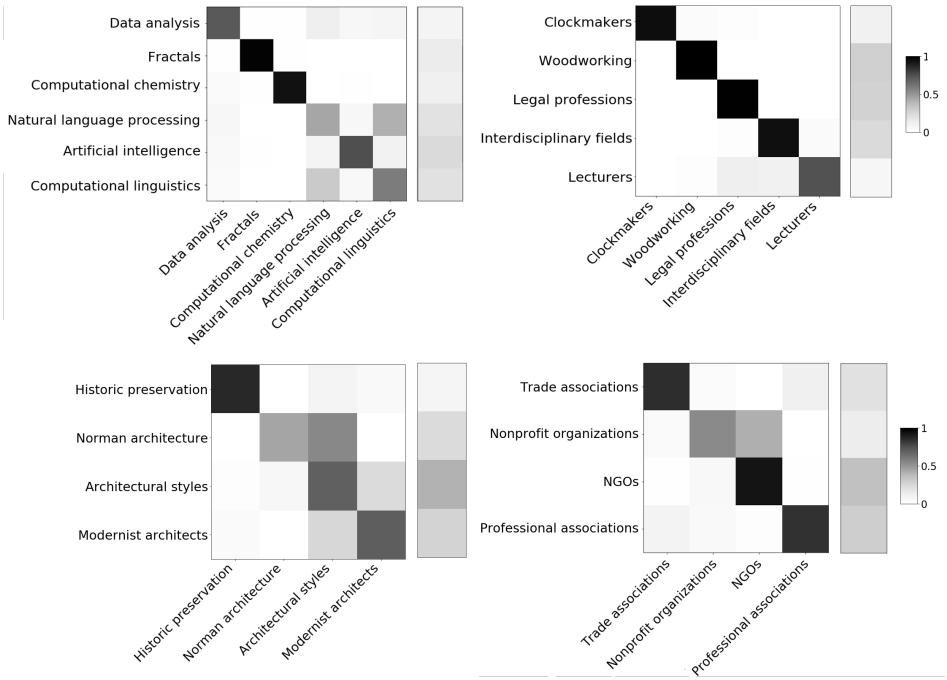

Figure 2: Two example analogies from the full analogy model with 90% of data (target systems left, most analogically similar stored system right). While some analogies exhibit interesting correspondences, others are often successful simply because they recapitulate one main feature; for example, a strong diagonal component in the class-level relation matrix.

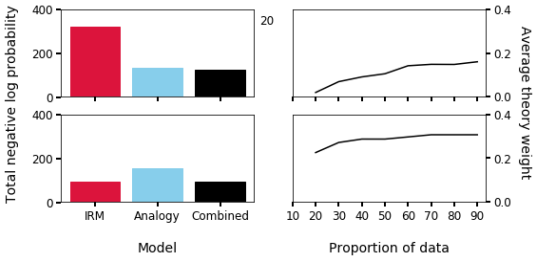

Figure 3: Nonparametric model results. Left: mean performance across systems summed over data partitions; Right: estimate of model evidence for IRM. Top: random systems; Bottom: Wikipedia. Results are for two-analogy model, and data omitted for smallest partition.

add even more flexibility to our model; for instance, by using a *hierarchical* Dirichlet process prior (Canini, Shashkov, & Griffiths, 2010).

Discussion

In the present work, we have given a small-scale assessment of our ideas about relational generalization, and how it relates to analogy. It will be of great future interest to deploy the model over a larger number of elements and systems, and explore the effectiveness of different estimators for the true fully Bayesian idea of analogical similarity given above, along with their coherence with human judgments. Finally, we look forward to more fully examining the theoretical and empirical correspondence with previous influential accounts of analogy in the literature, as well as with recent statistical work on generalization from the cognitive and neural sciences (Lake, Lawrence, & Tenenbaum, 2018; Whittington et al., 2019).

Acknowledgments

This work was supported by grant #61454 from the John Templeton Foundation.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour XIII—1983* (pp. 1–198). Springer.

Anderson, J. R., & Thompson, R. (1989). Use of analogy in a production system architecture. *Similarity and Analogical Reasoning*, 267–297.

Canini, K. R., Shashkov, M. M., & Griffiths, T. L. (2010). Modeling transfer learning in human categorization with the hierarchical Dirichlet process. In *ICML* (Vol. 27, pp. 151–158).

Carey, S. (1986). Cognitive science and science education. *American Psychologist*, *41*(10), 1123-1130.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1.

Euclid. (trans. 1956). *The thirteen books of Euclid's Elements* (T. L. Heath et al., Trans.). Courier Corporation.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability* (Vol. 174). Freeman: San Francisco.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752–775.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley interdisciplinary reviews: Cognitive Science*, *2*(3), 266–276.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories* (Vol. 1). MIT Press Cambridge, MA.

Halford, G. S., Bain, J. D., Maybery, M. T., & Andrews, G. (1998). Induction of relational schemas: Common processes in reasoning and complex learning. *Cognitive Psychology*, *35*(3), 201–245.

Halford, G. S., & Wilson, W. H. (1980). A category theory approach to cognitive development. *Cognitive Psychology*, *12*(3), 356–411.

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: the foundation of higher cognition. *Trends in Cognitive Sciences*, *14*(11), 497–505.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach*. Harvester press Hassocks, Sussex.

Holyoak, K. J. (2012). Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, 234–259.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, *18*(3), 387–438.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). *Discovering latent classes in relational data* (Tech. Rep. No. AI Memo 2004-019). Cambridge, MA: Massachusetts Institute of Technology.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI* (Vol. 3, p. 5).

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Klymko, C., Gleich, D., & Kolda, T. G. (2014). Using triangles to improve community detection in directed networks. *arXiv preprint arXiv:1404.5874*.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago press.

Lake, B. M., Lawrence, N. D., & Tenenbaum, J. B. (2018). The emergence of organizing structure in conceptual representation. *Cognitive science*, *42*, 809–832.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago Press.

Law, J., et al. (1999). Actor network theory and after.

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, *31*(4), 357–378.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological review*, *119*(3), 617.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, *56*, 71-113.

Pirolli, P. L., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology*, *39*(2), 240.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., ... Morris, R. G. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2019). The tolman-eichenbaum machine: Unifying space and relational memory through generalisation in the hippocampal formation. *bioRxiv*, 770495.