The discovery of structural form

Charles Kemp*† and Joshua B. Tenenbaum‡

*Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213; and ‡Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 30, 2008 (received for review March 17, 2008)

Algorithms for finding structure in data have become increasingly important both as tools for scientific data analysis and as models of human learning, yet they suffer from a critical limitation. Scientists discover qualitatively new forms of structure in observed data: For instance, Linnaeus recognized the hierarchical organization of biological species, and Mendeleev recognized the periodic structure of the chemical elements. Analogous insights play a pivotal role in cognitive development: Children discover that object category labels can be organized into hierarchies, friendship networks are organized into cliques, and comparative relations (e.g., “bigger than” or “better than”) respect a transitive order. Standard algorithms, however, can only learn structures of a single form that must be specified in advance: For instance, algorithms for hierarchical clustering create tree structures, whereas algorithms for dimensionality-reduction create low-dimensional spaces. Here, we present a computational model that learns structures of many different forms and that discovers which form is best for a given dataset. The model makes probabilistic inferences over a space of graph grammars representing trees, linear orders, multidimensional spaces, rings, dominance hierarchies, cliques, and other forms and successfully discovers the underlying structure of a variety of physical, biological, and social domains. Our approach brings structure learning methods closer to human abilities and may lead to a deeper computational understanding of cognitive development.

Keywords: cognitive development | structure discovery | unsupervised learning

Discovering the underlying structure of a set of entities is a fundamental challenge for scientists and children alike (1–7). Scientists may attempt to understand relationships between biological species or chemical elements, and children may attempt to understand relationships between category labels or the individuals in their social landscape, but both must solve problems at two distinct levels. The higher-level problem is to discover the form of the underlying structure. The entities may be organized into a tree, a ring, a dimensional order, a set of clusters, or some other kind of configuration, and a learner must infer which of these forms is best. Given a commitment to one of these structural forms, the lower-level problem is to identify the instance of this form that best explains the available data.

The lower-level problem is routinely confronted in science and cognitive development. Biologists have long agreed that tree structures are useful for organizing living kinds but continue to debate which tree is best—for instance, are crocodiles better grouped with lizards and snakes or with birds (8)? Similar issues arise when children attempt to fit a new acquaintance into a set of social cliques or to place a novel word in an intuitive hierarchy of category labels. Inferences like these can be captured by standard structure-learning algorithms, which search for structures of a single form that is assumed to be known in advance (Fig. L4). Clustering or competitive-learning algorithms (9, 10) search for a partition of the data into disjoint groups, algorithms for hierarchical clustering (11) or phylogenetic reconstruction (12) search for a tree structure, and algorithms for dimensionality reduction (13, 14) or multidimensional scaling (15) search for a spatial representation of the data. Higher-level discoveries about structural form are rarer but more fundamental, and often occur at pivotal moments in the development of a scientific field or a child’s understanding (1, 2, 4). For centuries, the natural representation for biological species was held to be the “great chain of being,” a linear structure in which every living thing found a place according to its degree of perfection (16). In 1735, Linnaeus famously proposed that relationships between plant and animal species are best captured by a tree structure, setting the agenda for all biological classification since. Modern chemistry also began with a discovery about structural form, the discovery that the elements have a periodic structure. Analogous discoveries are made by children, who learn, for example, that social networks are often organized into cliques, that temporal categories such as the seasons and the days of the week can be arranged into cycles, that comparative relations such as “longer than” or “better than” are transitive (17, 18) and that category labels can be organized into hierarchies (19). Structural forms for some cognitive domains may be known innately, but many appear to be genuine discoveries. When learning the meanings of words, children initially seem to organize objects into nonoverlapping clusters, with one category label allowed per cluster (20); hierarchies of category labels are recognized only later (19). When reasoning about comparative relations, children’s inferences respect a transitive ordering by the age of 7 but not before (21). In both of these cases, structural forms appear to be learned, but children are not explicitly taught to organize these domains into hierarchies or dimensional orders.

Here, we show that discoveries about structural form can be understood computationally as probabilistic inferences about the organizing principles of a dataset. Unlike most structure-learning algorithms (Fig. L4), the model we present can simultaneously discover the structural form and the instance of that form that best explain the data (Fig. 1B). Our approach can handle many kinds of data, including attributes, relations, and measures of similarity, and we show that it successfully discovers the structural forms of a diverse set of real-world domains. Any model of form discovery must specify the space of structural forms it is able to discover. We represent structures using graphs and use graph grammars (22) as a unifying language for expressing a wide range of structural forms (Fig. 2). Of the many possible forms, we assume that the most natural are those that can be derived from simple generative processes (23). Each of the first six forms in Fig. 2A can be generated by using a single context-free production that replaces a parent node with two child nodes and specifies how to connect the children to each other and to the neighbors of

Author contributions: C.K. and J.B.T. designed research; C.K. performed research; and C.K. and J.B.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

†To whom correspondence should be addressed. E-mail: ckemp@cmu.edu.

FREELY AVAILABLE ONLINE THROUGH THE PNAS OPEN ACCESS OPTION.

© 2008 by The National Academy of Sciences of the USA
then search for the structure
otherwise [see supporting information (SI)]
definition of compatibility].
The parameter
about a finite set of entities, we want to find the form
suggests that they are useful for describing the world, and that
modeled as an order. The popularity of the forms in Fig. 2
can be modeled as a tree and a seriation structure that can be
thinking are founded on two forms: a classification structure that
recur again and again in formal models across many different
rings (28, 29), trees (1, 12, 30), hierarchies (31, 32) and grids (33)
data analysis. Partitions (9, 25), chains (26), orders (1, 25, 27),
ammoned by algorithms for unsupervised learning or exploratory
many of the structural forms discussed by psychologists (24) and
complex productions.
forms, including multidimensional spaces and cylinders, can be
generated by combining these basic forms or by using more
complex productions.
It is striking that the simple grammars in Fig. 2A generate
many of the structural forms discussed by psychologists (24) and
assumed by algorithms for unsupervised learning or exploratory
data analysis. Partitions (9, 25), chains (26), orders (1, 25, 27),
rings (28, 29), trees (1, 12, 30), hierarchies (31, 32) and grids (33)
recur again and again in formal models across many different
literatures. To highlight just one example, Inhelder and Piaget (1)
suggest that the elementary logical operations in children’s
thinking are founded on two forms: a classification structure that
can be modeled as a tree and a seriation structure that can be
modeled as an order. The popularity of the forms in Fig. 2
suggests that they are useful for describing the world, and that
they spring to mind naturally when scientists seek formal
descriptions of a domain.
The problem of form discovery can now be posed. Given data
D about a finite set of entities, we want to find the form F
and the structure S of that form that best capture the relationships
between these entities. We take a probabilistic approach, and
define a hierarchical generative model (34) that specifies how the
data are generated from an underlying structure, and how this
structure is generated from an underlying form (Fig. 1B). We
then search for the structure S and form F that maximize the
posterior probability

\[ P(S, F | D) \propto P(D | S) P(S | F) P(F). \]  

\[ P(F) \] is a uniform distribution over the forms under consideration
(Fig. 2). Structure S is a cluster graph, an instance of one
of the forms in Fig. 2, where the nodes represent clusters of entities (Fig. 4A shows a cluster graph with the form of an order).
The prior \( P(S | F) \) favors graphs where \( k \), the number of clusters,
is small: \( P(S | F) \propto \theta^k \) if \( S \) is compatible with \( F \), and \( P(S | F) = 0 \)
otherwise [see supporting information (SI) Appendix for
the definition of compatibility]. The parameter \( \theta \) determines the

\[ P(S, F | D) \propto P(D | S) P(S | F) P(F). \]  

\[ P(F) \] is a uniform distribution over the forms under consideration
(Fig. 2). Structure S is a cluster graph, an instance of one
of the forms in Fig. 2, where the nodes represent clusters of entities (Fig. 4A shows a cluster graph with the form of an order).
The prior \( P(S | F) \) favors graphs where \( k \), the number of clusters,
is small: \( P(S | F) \propto \theta^k \) if \( S \) is compatible with \( F \), and \( P(S | F) = 0 \)
otherwise [see supporting information (SI) Appendix for
the definition of compatibility]. The parameter \( \theta \) determines the

\[ P(S, F | D) \propto P(D | S) P(S | F) P(F). \]  

\[ P(F) \] is a uniform distribution over the forms under consideration
(Fig. 2). Structure S is a cluster graph, an instance of one
of the forms in Fig. 2, where the nodes represent clusters of entities (Fig. 4A shows a cluster graph with the form of an order).
The prior \( P(S | F) \) favors graphs where \( k \), the number of clusters,
is small: \( P(S | F) \propto \theta^k \) if \( S \) is compatible with \( F \), and \( P(S | F) = 0 \)
otherwise [see supporting information (SI) Appendix for
the definition of compatibility]. The parameter \( \theta \) determines the

\[ P(S, F | D) \propto P(D | S) P(S | F) P(F). \]  

\[ P(F) \] is a uniform distribution over the forms under consideration
(Fig. 2). Structure S is a cluster graph, an instance of one
of the forms in Fig. 2, where the nodes represent clusters of entities (Fig. 4A shows a cluster graph with the form of an order).
The prior \( P(S | F) \) favors graphs where \( k \), the number of clusters,
is small: \( P(S | F) \propto \theta^k \) if \( S \) is compatible with \( F \), and \( P(S | F) = 0 \)
otherwise [see supporting information (SI) Appendix for
the definition of compatibility]. The parameter \( \theta \) determines the
extent to which graphs with many clusters are penalized, and is fixed for all of our experiments. The normalizing constant for $P(S/F)$ depends on the number of structures compatible with a given form, and ensures that simpler forms are preferred whenever possible. For example, any chain $S_c$ is a special case of a grid, but $P(S_c|F = \text{chain}) > P(S_c|F = \text{grid})$ because there are more possible grids than chains given a fixed number of entities. It follows that $P(S_c, F = \text{chain}|D) > P(S_c, F = \text{grid}|D)$ for any

Fig. 3. Structures learned from biological features ($A$), Supreme Court votes ($B$), judgments of the similarity between pure color wavelengths ($C$), Euclidean distances between faces represented as pixel vectors ($D$), and distances between world cities ($E$). For $A$–$C$, the edge lengths represent maximum a posteriori edge lengths under our generative model.

Fig. 4. Structures learned from relational data ($Upper$) and the raw data organized according to these structures ($Lower$). ($A$) Dominance relationships among a troop of sooty mangabeys. The sorted data matrix has most of its entries above the diagonal, indicating that animals tend to dominate only the animals below them in the order. ($B$) A hierarchy representing interactions between members of the Bush administration. ($C$) Social cliques representing friendship relations between prisoners. The sorted matrix has most of its entries along the diagonal, indicating that prisoners tend only to be friends with prisoners in the same cluster. ($D$) The Kula ring representing armshell trade between New Guinea communities. The relative positions of the communities correspond approximately to their geographic locations.
dataset $D$, and that the grid form will only be chosen if the best grid is substantially better than the best chain.

The remaining term in Eq. 1, $P(D|S)$, measures how well the structure $S$ accounts for the data $D$. Suppose that $D$ is a feature matrix like the matrix in Fig. 1. $P(D|S)$ will be high if the features in $D$ vary smoothly over the graph $S$, that is, if entities nearby in $S$ tend to have similar feature values. For instance, feature $F^i$ is smooth over the tree in Fig. 1B, but $F^{(i)}$ is not. Even though Fig. 1 shows binary features, we treat all features as continuous and capture the expectation of smoothness by assuming that these features are independently generated from a multivariate Gaussian distribution with a dimension for each node in graph $S$. As described in SI Appendix, the covariance of this distribution is defined in a way that encourages nearby nodes in graph $S$ to have similar feature values, and the term $P(D|S)$ favors graphs that meet this condition.

In principle, our approach can be used to identify the form $F$ that maximizes $P(F|D)$, but we are also interested in discovering the structure $S$ that best accounts for the data. We therefore search for the structure $S$ and form $F$ that jointly maximize the scoring function $P(S, F|D)$ (Eq. 1). To identify these elements, we run a separate greedy search for each candidate form. Each search begins with all entities assigned to a single cluster, and the algorithm splits a cluster at each iteration, using the production for the current form (Fig. 2). After each split, the algorithm attempts to improve the score, using several proposals, including proposals that move an entity from one cluster to another and proposals that swap two clusters. The search concludes once the score can no longer be improved. A more detailed description of the search algorithm is provided in SI Appendix.

We generated synthetic data to test this algorithm on cases where the true structure was known. The SI Appendix shows graphs used to generate five datasets, and the structures found by fitting five different forms to the data. In each case, the model recovers the true underlying form of the data.

Next, we applied the model to several real-world datasets, in each case considering all forms in Fig. 2. The first dataset is a matrix of animal species and their biological and ecological properties. It consists of human judgments about 33 species and 106 features and amounts to a larger and noisier version of the matrix in Fig. 1. The best scoring form for this dataset is the tree, and the best tree (Fig. 3A) includes subtrees that correspond to categories at several levels of resolution (e.g., mammals, primates, rodents, birds, insects, and flying insects). The second dataset is a matrix of votes from the United States Supreme Court, including 13 judges and their votes on 1,596 cases. Some political scientists (35) have argued that a unidimensional structure best accounts for variation in Supreme Court data and in political beliefs more generally, although other structural forms [including higher-dimensional spaces (36) and sets of clusters (37)] have also been proposed. Consistent with the unidimensional hypothesis, our model identifies the chain as the best-scoring form for the Supreme Court data. The best chain (Fig. 3B) organizes the 13 judges from liberal (Marshall and Brennan) to conservative (Thomas and Scalia).

If similarity is assumed to be a measure of covariance, our model can also discover structure in similarity data. Under our generative model for features, the expression for $P(D|S)$ includes only two components that depend on $D$: the number of features observed and the covariance of the data. As long as both components are provided, Eq. 1 can be used even if none of the features is directly observed. We applied the model to a matrix containing human judgments of the similarity between all pairs of 14 pure-wavelength hues (38). The ring in Fig. 3C is the best structure for these data and corresponds to the color circle described by Newton. Next, we analyzed a similarity dataset where the entities are faces that vary along two dimensions: masculinity and race. The model chooses a grid structure that recovers these dimensions (Fig. 3D). Finally, we applied the model to a dataset of distances between 35 world cities. Our model chooses a cylinder where the chain component corresponds approximately to latitude, and the ring component corresponds approximately to longitude.

The same algorithm can be used to discover structure in relational data, but we must modify the distribution $P(D|S)$. Suppose that $D$ is a square frequency matrix, where $D(i,j)$ indicates the number of times a certain relation has been observed between entities $i$ and $j$ (Fig. 4). We define a model where $P(D|S)$ is high if the large entries in $D$ correspond to edges in the graph $S$. A similar model can be defined if $D$ is a binary relation rather than a frequency matrix. Given a relation $D$, it is important to discover whether the relation tends to hold between elements in the same cluster or only between different clusters, and whether the relation is directed or not. The forms in Fig. 2A all have directed edges and nodes without self-links, and we
expanded this collection to include forms with self-links, forms with undirected edges, and forms with both of these properties.

First, we applied the model to a matrix of interactions among a troop of sooty mangabeys. The model discovers that the order is the most appropriate form, and the best order found (Fig. 4A) is consistent with the dominance hierarchy inferred by primatologists studying this troop (39). Hierarchical structure is also characteristic of human organizations, although tree-structured hierarchies are perhaps more common than full linear orders. We applied the model to a matrix of interactions between 13 members of George W. Bush’s first-term administration (40). The best form is an undirected hierarchy, and the best hierarchy found (Fig. 4B) closely matches an organizational chart built by connecting individuals to their immediate superiors. Next, we analyzed social preference data (41) that represent friendships between prison inmates. Clique structures are often claimed to be characteristic of social networks (42), and the model discovers that a partition (a set of cliques) gives the best account of the data. Finally, we analyzed trade relations between 20 communities in New Guinea (43). The model discovers the Kula ring, an exchange structure first described by Malinowski (44).

We have presented an approach to structure discovery that provides a unifying description of many structural forms, discovers qualitatively different representations for a diverse range of datasets, and can handle multiple kinds of data, including feature data, relational data, and measures of similarity. Our hypothesis space of forms (Fig. 2) includes some of the most common forms, but does not exhaust the set of cognitively natural or scientifically important forms. Ultimately, psychologists should aim to develop a “Universal Structure Grammar” (compare with ref. 45) that characterizes more fully the representational resources available to human learners. This universal grammar might consist of a set of simple principles that generate all and only the cognitively natural forms. We can only speculate about how these principles might look, but one starting place is a metagrammar (46) for generating graph grammars. For instance, all of the forms shown in Fig. 2A can be generated by a metagrammar shown in SI Appendix.

Our framework may be most readily useful as a tool for data analysis and scientific discovery, but should also be explored as a model of human learning. Our model helps to explain how adults discover structural forms in controlled behavioral experiments (40), and is consistent with previous demonstrations that adults can choose the most appropriate representation for a given problem (47). Our model may also help to explain how children learn about the structure of their world. The model shows developmental shifts as more data are encountered, and often moves from a simple form to a more complex form that more faithfully represents the structure of the domain (Fig. 5 and SI Appendix). Identifying the best form for a domain provides powerful constraints on inductive inference, constraints that may help to explain how children learn new word meanings, concepts, and relations so quickly and from such sparse data (48–51). Discovering the clique structure of social networks can allow a child to predict the outcome of interactions between individuals who may never have interacted previously. Discovering the hierarchical structure of category labels allows a child to predict that a creature called a “chihuahua” might also be a dog and an animal, but cannot be both a dog and a cat.

As examples like these suggest, form discovery provides a way of acquiring domain-specific constraints on the structure of mental representations, a possibility that departs from two prominent views of cognition. A typical nativist view recognizes that inductive inference relies on domain-specific constraints but assumes that these constraints are innately provided (52–54). Chomsky (52), for instance, has suggested that “the belief that various systems of mind are organized along quite different principles leads to the natural conclusion that these systems are intrinsically determined, not simply the result of common mechanisms of learning or growth.” A typical empiricist view emphasizes learning but assumes no domain-specific representational structure. Standard methods for learning associative networks (55) and neural networks (56) use the same generic class of representations for every task, instead of attempting to identify the distinctive kinds of structures that characterize individual domains. Without these constraints, empiricist methods can require unrealistically large quantities of training data to learn even very simple concepts (57). Our framework offers a third view that combines insights from both these approaches and shows how domain-specific structural constraints can be acquired by using domain-general probabilistic inference. As children learn about the structure of different domains, they make discoveries as impressive as those of Linnaeus and Mendeleev, and approaches like ours may help to explain how these discoveries are possible.

ACKNOWLEDGMENTS. We thank P. Gunkel, E. Newport, A. Perfors, and W. Richards for valuable discussions and D. Casasanto, M. Frank, N. Goodman, V. Mankinsgka, R. Saxe, J. M. Tenenbaum, D. Tymoczko, the editor, and two anonymous reviewers for helpful suggestions. This work was supported in part by the William Asbjornsen Albert memorial fellowship (to C.K.), the Paul E. Newton career development chair (J.B.T.), the James S. McDonnell Foundation Causal Learning Research Collaborative, Air Force Office of Scientific Research Grant FA9550-07-1-0075, and the NTT Communication Sciences Laboratory.