The discovery of structural form

Charles Kemp & Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA

Algorithms for finding structure in data are important both as tools for scientific discovery (*1*) and as models of human learning (*2, 3*). Most of these algorithms assume that the *form* of the structure is known in advance: clustering algorithms (*2*) assume that the data fall into some number of disjoint groups, hierarchical clustering algorithms assume that the data are tree-structured, and algorithms for dimensionality-reduction (*4–6*) typically assume an underlying Euclidean geometry. Unlike these algorithms, humans can discover structural forms from data, and this ability drives major advances in scientific understanding and cognitive development: Linnaeus and Mendeleev discovered forms for biological species and the chemical elements, and young children may make analogous discoveries as they learn about social networks and category hierarchies. We present a computational model that explains discoveries like these as Bayesian inferences over probabilistic models generated by graph grammars. The model simultaneously discovers the form and the specific structure of that form that best explain the available data.

Philosophers, psychologists, and statisticians have suggested that scientists and children use similar strategies to learn about the structure of the world (*7–13*). In both science and cognitive development, the problem of structure discovery can be addressed on at least two levels. At the

first level, the form of the data is assumed known and the task is to choose the instance of that form that best explains the data (Figure 1A). Biologists, for instance, have long agreed that tree structures are useful for organizing living kinds but still debate which tree is best. Traditional taxonomies group crocodiles with lizards, snakes and turtles, but contemporary phylogenies assert that crocodiles are better grouped with birds (Figure 1A) (**?**). Similar problems arise when chemists attempt to locate a new element in the periodic table, or when children attempt to locate a newly encountered animal or food in an intuitive hierarchy of categories

At the second, deeper level, the problem is to discover the structural form of a domain: to discover, for example, that living kinds are tree structured, or that the chemical elements have a periodic structure (Figure 1B). The problem of form discovery is prominent in the history of biological classification. For centuries, the great chain of being (*14*) was thought to be the natural representation for living kinds (Figure 1B), but this linear form has been replaced by the tree structures introduced by Linnaeus (*15*). Other forms are also logically possible: a ring structure might not seem suitable for the species in Figure 1B, but has recently been suggested as the best model of relationships between microbes (*16*). Form discovery is also a problem for children, who learn, for example, that social networks are often organized into cliques, that the seasons can be arranged into a cycle, that relations like "heavier than" are transitive (*17,18*), and that category labels can be organized into hierarchies (*19*). It may even be solved by members of other species: baboons, for example, may make the genuine discovery that their troops are organized into dominance hierarchies (*20*).

We present a computational framework that addresses structure discovery at both these levels. Form discovery is mostly ignored by existing approaches, but is arguably the more fundamental problem. Solving this problem can have dramatic consequences: structural forms provide powerful constraints on inductive inference, allowing confident predictions about objects that are sparsely observed, or perhaps not observed at all. Discovering the periodic structure

of the elements allowed Mendeleev to predict both the existence and the properties of several novel elements. Similarly, a baboon who has discovered that dominance relations are linearly ordered within his troop should be able to predict the outcome of a confrontation between two animals who may never have interacted previously.

Form discovery, then, is one way of acquiring inductive constraints, a possibility that is rarely considered by theories of human or machine learning. One tradition recognizes the need for structural constraints, but assumes that they are provided as part of the initial specification of the learning algorithm (**?**, *21–25*). Chomsky (*21*), for instance, has claimed 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." Many approaches to unsupervised learning implicitly endorse this claim: they assume the form of the data is known and search for the best instance of that form (*26,27*). A second tradition denies the importance of structural form, proposing models like associative networks (*28*), multilayer perceptrons (*29*) and Bayesian networks (*30*) that can apply to domains with any kind of structure. Without the benefit of domain-specific constraints, generic models like these can require massive quantities of data to achieve human-level performance (*23*) — data that are often unavailable in scientific applications or in real-world human learning. Our framework offers a third approach to structure discovery that combines insights from both of these traditions. We show how structured, domain-specific representations cna be acquired using domain-general statistical inference, and demonstrate that the structural forms of many real-world domains can be discovered from relatively sparse data sets.

Any algorithm for form discovery must specify, explicitly or otherwise, the space of structural forms it is able to discover. We represent structures using graphs, and use graph grammars (**?**) as a unifying language for expressing a wide range of structural forms (Figure 2). Of the many possible forms, we assume that the most natural are those that can be derived from

simple generative processes (*31*). The first six forms in Figure 2A can be generated using a single context-free production that replaces a parent node with two children, and specifies how to connect the children to each other and to the neighbors of the parent node. Figures 2B and 2C show how two of these productions generate linear structures and orders. More complex forms, including multidimensional spaces and cylinders, can be generated by combining these basic forms.

It is striking that the simple grammars in 2a generate many of the structural forms discussed by psychologists (*26*) and assumed by algorithms for unsupervised learning. Partitions (*2, 32*), chains (*33*), orders (*7, 32, 34*), rings (*35*), trees (*7, 36, 37*), hierarchies (*38, 39*) and grids (*40*) recur again and again in formal models across many different literatures. To highlight just one example, Inhelder and Piaget (*7*) suggest that the elementary logical operations in children's thinking are founded on two forms: a classification structure that can be modelled as a tree, and a seriation structure that can be modelled as an order. The popularity of the forms in Figure 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 observed data D about a finite set of entities, we wish to find the form F and the structure S of that form that best capture the relationships between these entities. We take a Bayesian approach, defining a hierarchical generative model (Figure 1A) and searching for the structure S and form F that maximize the posterior probability:

$$P(S, F|D) \propto P(D|S)P(S|F)P(F). \tag{1}$$

$P(F)$ is a uniform distribution over the forms under consideration. Structure S is a *cluster graph*: an instance of one of the forms in Figure 2, where the nodes represent clusters of entities. Working with clusters allows the model to learn representations that are only as complex as the data require. The prior $P(S|F)$ favors graphs with small numbers of nodes: for any structure

S that is compatible with F, $P(S|F) \propto \theta(1-\theta)^k$, where θ is a parameter and k is the number of nodes in S; $P(S|F) = 0$ if S is incompatible with F (see Supporting Online Material for the definition of compatibility). The normalizing constant for $P(S|F)$ depends on the number of structures allowed by a given form, and ensures that simpler forms are preferred whenever possible. For example, any chain S_{chain} is a special case of a grid, but $P(S_{chain}|F_{chain}) > P(S_{chain}|F_{grid})$ since there are more possible grids than chains given a fixed number of entities. The remaining term in Equation 1, the likelihood $P(D|S)$, depends on the nature of the data D. We consider three kinds of data: feature data, similarity data, and relational data.

Suppose first that D is an entity-feature matrix, where the (i, j) entry in the matrix indicates the value of entity i for feature j. We represent the structure of the data set using undirected *entity graphs*. Cluster graphs are converted to entity graphs by adding a node for each entity, connecting each entity to the cluster node that contains it, and replacing each directed edge with an undirected link. Given an entity graph, we expect nearby entities in the graph to have similar features. Formally, this expectation can be captured by assuming that the features are independently generated by a Gaussian process over the graph (*41*). Under this generative model, a graph accounts well for the data D if the features vary smoothly over the graph.

We generated synthetic data to test this model on cases where the true structure was known. Figure 3 shows graphs used to generate five data sets, and the structures found by fitting five different forms to the data. For each data set, we ran a greedy search for each of the candidate forms. The search begins with all the entities at a single cluster, and splits a node at each iteration using a production from Figure 2. After each split, the algorithm attempts to improve the current score using several additional proposals, including proposals that move an entity from one cluster to another (see Supporting Online Material). The final column in Figure 3 compares the scores for the five forms, and we see that the true form is correctly recovered in each case.

Next we applied the model to several real-world data sets, in each case considering all eight forms in Figure 2. The first data set is a matrix of human judgments about the features of animal species. The tree is the form that scores best under Equation 1, which is consistent with the finding that cultures all over the world organize living kinds into tree-structured representations (*22*). The best tree (Figure 4A) organizes the animals into mammals, birds, fish, insects and reptiles, and includes subtrees that correspond to intuitively plausible subcategories (e.g., primates, cetaceans, flying insects, rodents, flightless birds). The second data set is a matrix of Supreme Court votes (*42*). Others (*43*) have argued that a unidimensional structure accounts for most of the variance in Supreme Court data, and we find that the chain is the form with the highest score. The best chain (Figure 4B) organizes the thirteen judges from liberal (Marshall and Brennan) to conservative (Thomas and Scalia).

Under our generative model for features, the matrix D influences the distribution $P(D|S)$ only through the number of features m and the covariance matrix $\frac{1}{m}DD^{\mathsf{T}}$. As long as these components are provided, our approach to structure discovery can be used even if none of the features is observed. Assuming that similarity is a measure of covariance, we used this idea to discover structure in similarity data. First we analyzed similarity ratings between 14 pure-wavelength color hues (*44*). The ring in Figure 4C is the best structure for these data, and corresponds to the color circle described by Newton. Next we analyzed a similarity data set where the entities are faces that vary along two dimensions of masculinity and race. The model chooses a grid structure that recovers these dimensions (Figure 4D). Finally, we applied the model to a data set of distances between 35 world cities. Our model chooses a cylinder where the chain component corresponds roughly to latitude, and the ring component corresponds roughly to longitude.

Consider now a distribution $P(D|S)$ that can be used with Equation 1 to analyze data about relationships between entities. Suppose that D is a square frequency matrix where $D(i, j)$

indicates the number of times a certain relation has been observed between i and j. We define a model where $P(D|S)$ is high if if the large entries in D correspond to edges in the cluster graph S. 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 Figure 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 data set representing interactions among a troop of sooty mangabeys. The model discovers that the order is the most appropriate form, and the best order found (Figure 5A) is consistent with the dominance hierarchy inferred by primatologists studying this troop (*45*). We then applied the model to a data representing interactions between 13 high-ranking members of Bush's first-term administration. The best form is an undirected hierarchy, and the best hierarchy found (Figure 5B) closely matches an organizational chart built by connecting individuals to their immediate superiors. Next we analyzed social preference data (*46*) that represent friendships between prison inmates. Clique structures are often claimed to be characteristic of social networks (**?**), 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 (*47*). The model discovers the Kula ring, an exchange structure first described by Malinowski (*48*).

We have presented a framework for structure discovery that subsumes many popular approaches to unsupervised learning, discovers the structural form of a data set, and suggests how human learners or other primates might do the same. Our hypothesis space of forms (Figure 2) includes some of the most commonly encountered forms, but does not exhaust the set of cognitively natural or scientifically important forms. Ultimately, psychologists should aim to develop a "Universal Structure Grammar" (cf (*49*)) 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 spectulate about how these principles might look, but one starting place is a meta-grammar for generating graph grammars. For instance, all of the grammars shown in Figure 2a can be be generated by a simple meta-grammar described in the Supporting Online Material.

Our framework may be most readily useful as a tool for data analysis and scientific discovery, but we hope that it will also be explored and tested as a model of conceptual development. As they learn about the structure of different domains, children make discoveries as impressive as those of Linnaeus and Mendeleev, and approaches like ours may ultimately explain how these discoveries are possible. As our model encounters more data, it can show qualitative transitions from a simple form to a more complex form that more faithfully represents the structure of the domain (see Supplementary Online Material). These transitions resemble the conceptual leaps of children (*8*) or the paradigm shifts of scientists (*10*), and deciding how deep this resemblance goes is a major question for future research.

References

1. N. Friedman, *Science* **303**, 799 (2004).

2. J. R. Anderson, *Psychological Review* **98**, 409 (1991).

3. T. K. Landauer, S. T. Dumais, *Psychological Review* **104**, 211 (1997).

4. K. Pearson, *Philosophical Magazine* **2**, 559 (1901).

5. C. E. Spearman, *American Journal of Psychology* **5**, 201 (1904).

6. W. S. Torgeson, *Psychometrika* **30**, 379 (1965).

7. B. Inhelder, J. Piaget, *The early growth of logic in the child* (Routledge & Kegan Paul, 1964).

8. S. Carey, *Conceptual change in childhood* (MIT Press, Cambridge, MA, 1985).

9. A. Gopnik, A. N. Meltzoff, *Words, thoughts, and theories* (MIT Press, Cambridge, MA, 1997).

10. T. S. Kuhn, *The structure of scientific revolutions* (University of Chicago Press, Chicago, 1970), second edn.

11. W. Whewell, *The philosophy of the inductive sciences, founded upon their history* (1840).

12. E. T. Jaynes, *Probability theory: The logic of science* (Cambridge University Press, Cambridge, 2003).

13. P. Spirtes, C. Glymour, R. Schienes, *Causation prediction and search* (Springer-Verlag, New York, 1993).

14. A. O. Lovejoy, *The great chain of being* (Harvard University Press, 1970).

15. C. Linnaeus, *Systema Naturae* (1766).

16. M. C. Rivera, J. A. Lake, *Nature* **431**, 152 (2004).

17. J. Piaget, *The child's conception of number* (Norton, New York, 1965).

18. D. Mareschal, T. R. Shultz, *Connection Science* **11** (1999).

19. E. Rosch, *Cognition and categorization*, E. Rosch, Lloyd, eds. (1978), pp. 27–48.

20. D. L. Cheney, R. Seyfarth, *Cognition* **37**, 167 (1990).

21. N. Chomsky, *Rules and Representations* (Basil Blackwell, Oxford, 1980).

22. S. Atran, *Behavioral and Brain Sciences* **21**, 547 (1998).

23. S. Geman, E. Bienenstock, R. Doursat, *Neural Computation* **4**, 1 (1992).

24. T. M. Mitchell, *Machine learning* (McGraw Hill, New York, 1997).

25. I. Kant, *Critique of pure reason* (1993).

26. R. N. Shepard, *Science* **210**, 390 (1980).

27. R. O. Duda, P. E. Hart, D. G. Stork, *Pattern classification* (Wiley, New York, 2000).

28. R. A. Rescorla, A. R. Wagner, *Classical conditioning II: Current research and theory*, A. H. Black, W. F. Prokasy, eds. (Appleton-Century-Crofts, New York, 1972), pp. 64–99.

29. T. T. Rogers, J. L. McClelland, *Semantic cognition: a Parallel Distributed Processing approach* (MIT Press, 2004).

30. D. Heckerman, *Learning in Graphical Models*, M. I. Jordan, ed. (MIT Press, Cambridge, MA, 1998), pp. 301–354.

31. M. Leyton, *Symmetry, causality, mind* (MIT Press, Cambridge, MA, 1992).

32. A. P. Fiske, *Psychological Review* **99**, 689 (1992).

33. L. Guttman, *American Sociological Review* **9**, 139 (1944).

34. R. A. Bradley, M. E. Terry, *Biometrika* **39**, 324 (1952).

35. L. Guttman, *Mathematical thinking in the social sciences*, P. F. Lazarsfeld, ed. (1954), pp. 258–348.

36. P. H. Sneath, R. R. Sokal, *Numerical Taxonomy – the principles and practice of numerical classification* (1973).

37. J. P. Huelsenbeck, F. Ronquist, *Bioinformatics* **17**, 754 (2001).

38. A. Collins, M. R. Quillian **8**, 240 (1969).

39. J. D. Carroll, *Psychometrika* **41**, 439 (1976).

40. T. Kohonen, *Self-Organizing Maps* (Springer-Verlag, New York, 1997).

41. X. Zhu, J. Lafferty, Z. Ghahramani, Semi-supervised learning: from Gaussian fields to Gaussian processes, *Tech. Rep. CMU-CS-03-175*, Carnegie-Mellon University (2003).

42. H. J. Spaeth (2005).

43. B. Grofman, T. Brazill, *Public Choice* **112**, 55 (2002).

44. G. Ekman, *J. Psychol* **38** (1954).

45. F. Range, R. Noë, *American Journal of Primatology* **56**, 137 (2002).

46. J. MacRae, *Sociometry* **22**, 360 (1960).

47. P. Hage, F. Harary, *Exchange in Oceania: A graph theoretic analysis* (Oxford University Press, 1991).

48. B. Malinowski, *Argonauts of the Western Pacific: An account of native enterprise and adventure in the archipelagoes of Melanesian New Guinea* (1922).

49. N. Chomsky, *Aspects of the theory of syntax* (MIT Press, Cambridge, MA, 1965).

50. C. White, *An account of the regular gradation in man* (1799).

51. Supported by the William Asbjornsen Albert memorial fellowship (CK) and the Paul E. Newton career development chair (JBT). We thank P. Gunkel, E. Newport, A. Perfors, and W. Richards for valuable discussions.

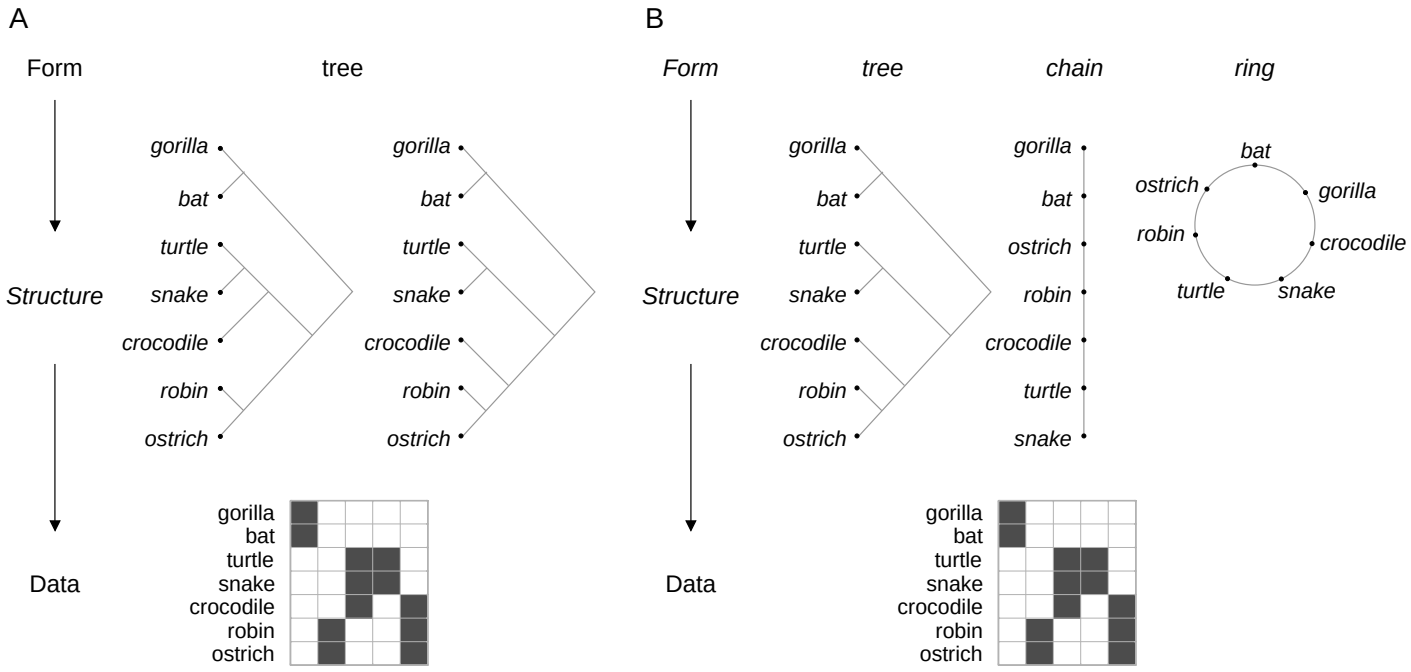

Figure 1: (A) Discovering the structure that best accounts for a set of binary features. The structure is assumed to be a tree: the first candidate is inspired by the Linnaean taxonomy, and the second is a cladogram. (B) Simultaneously discovering the form and the structure that best account for the data. Three possible pairs of forms and structures are shown — the chain is inspired by Bonnet's "scale of natural beings" (*50*).

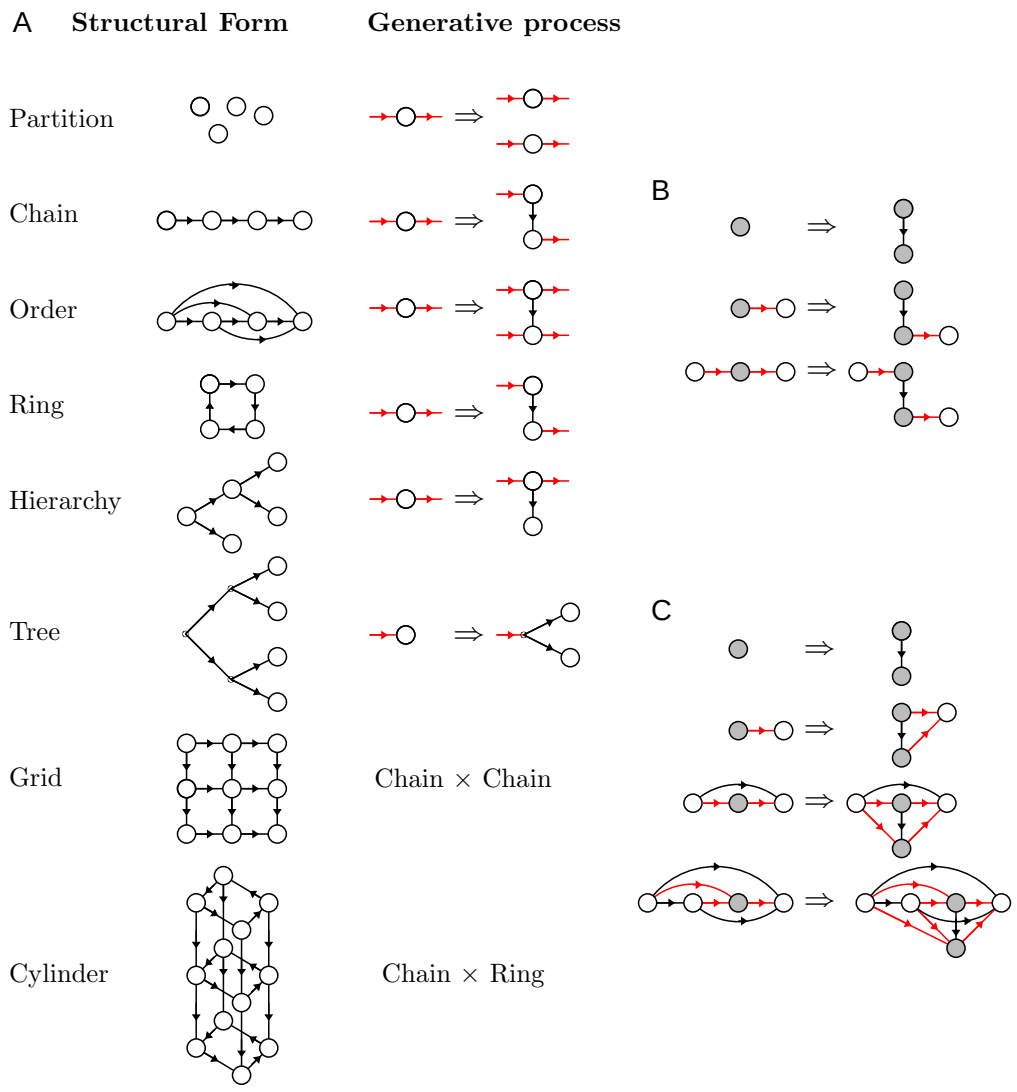

Figure 2: (A) Eight structural forms and the generative processes that produce them. The first six processes are node-replacement graph grammars. Each grammar uses a single production, and each production specifies how to replace a parent node with two children. (B,C) Growing chains and orders. At each step in each derivation, the parent and children nodes are shown in grey. The red arrows in each production represent *all* edges that enter or leave a parent node. When applying the order production, all nodes that sent a link to the parent node now send links to both children.

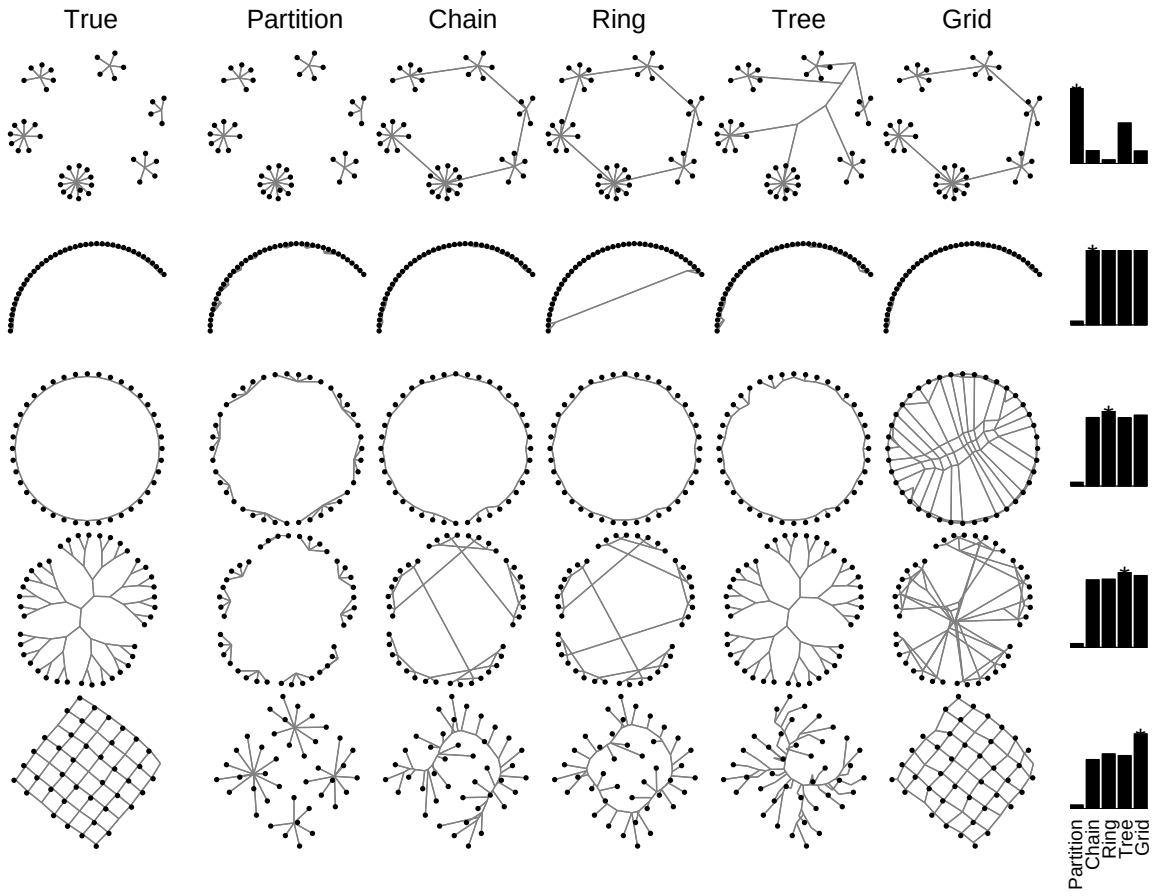

Figure 3: Structure discovery results for synthetic data. Five sets of features were generated over the graphs in the left column, and five forms were fit to each dataset. The structures found are drawn so that entity positions correspond to positions in the picture of the true structure. The final column shows log posteriors $\log(P(S, F|D))$ for the best structures found. Each plot has been scaled so that the worst performing structure receives a score of zero.

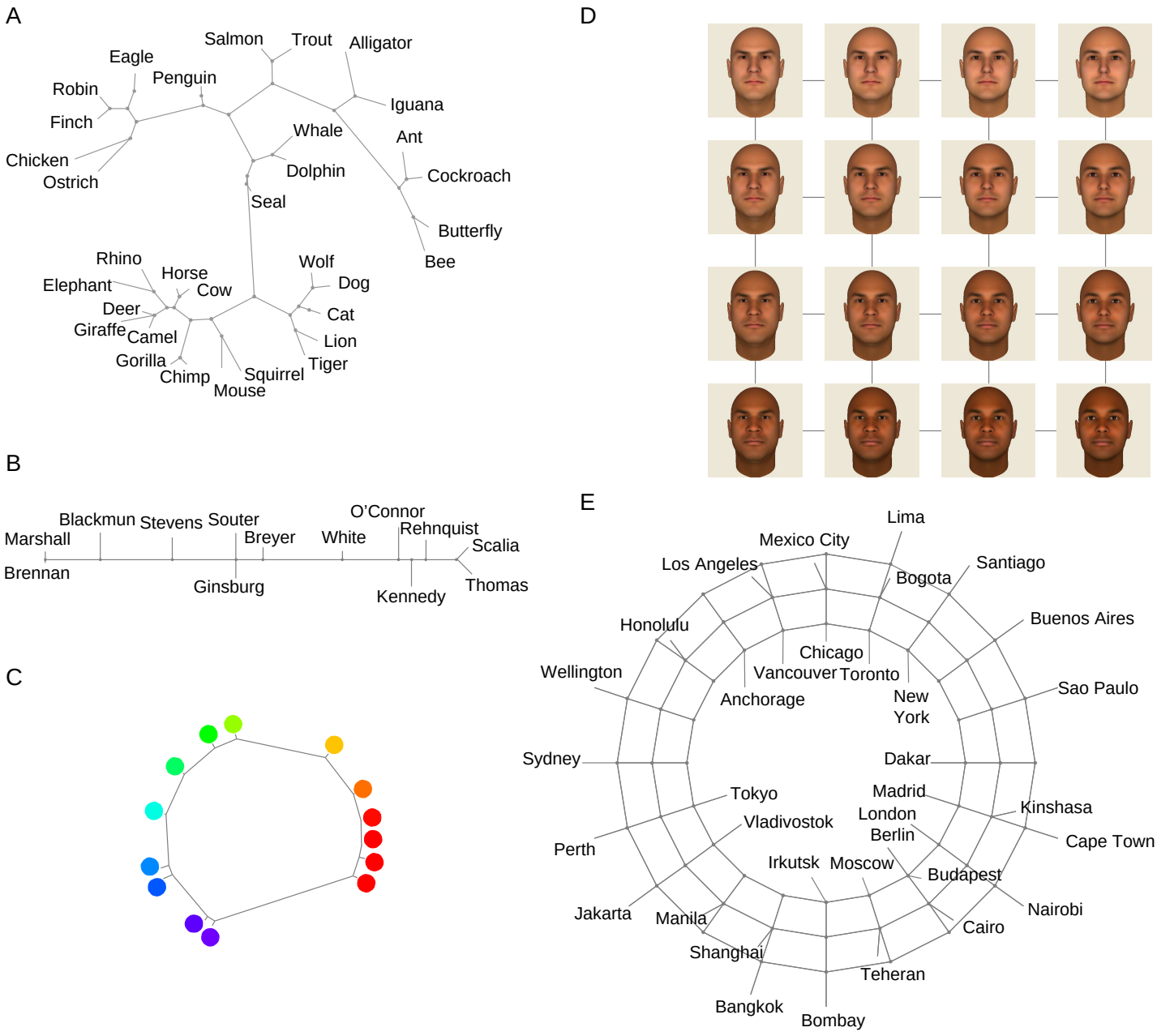

Figure 4: Structures learned from (A) biological features, (B) Supreme Court votes (C) similarity ratings of pure color wavelengths (D) Euclidean distances between faces represented as pixel vectors (E) distances between world cities.

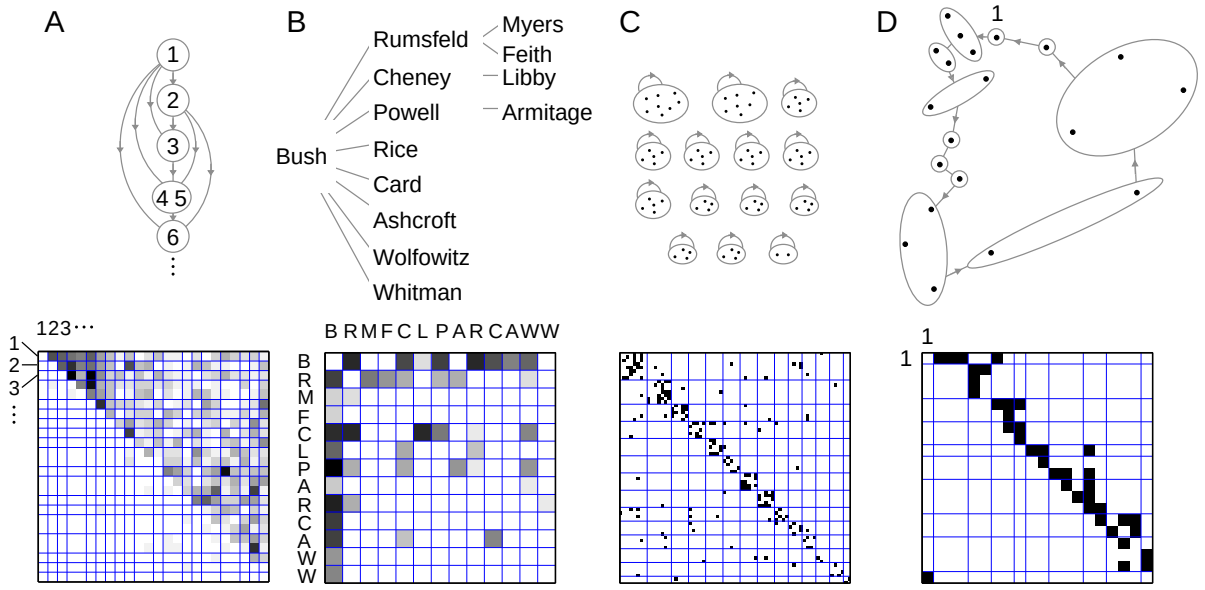

Figure 5: Structures learned from relational data (top row), and the raw data organized according to these structures (bottom row). (A) Primate dominance data. The sorted data matrix has most of its entries above the diagonal, indicating that animals tend to dominate only the animals below them in the order. (B) A hierarchy representing relationships 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 communities are laid out according to their geographic position.