Structure detection methods
Complex Networks, CSYS/MATH 303, Spring, 2010

Prof. Peter Dodds

Department of Mathematics & Statistics
Center for Complex Systems
Vermont Advanced Computing Center
University of Vermont

Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.
Outline

Overview

Methods

Hierarchy by aggregation
Hierarchy by division
Hierarchy by shuffling
Spectral methods
Hierarchies & Missing Links
General structure detection

Final words

References
Structure detection

▶ Zachary’s karate club [10, 7]

► Possible substructures:
  hierarchies, cliques, rings, …

► Plus:
  All combinations of substructures.

► Much focus on hierarchies…

The issue:
how do we elucidate the internal structure of large networks across many scales?
Hierarchy by division

Bottom up:

- **Idea:** Extract hierarchical classification scheme for $N$ objects by an agglomeration process.
- Need a measure of distance between all pairs of objects.
- **Note:** evidently works for non-networked data.
- **Procedure:**
  1. Order pair-based distances.
  2. Sequentially add links between nodes based on closeness.
  3. Use additional criteria to determine when clusters are meaningful.
- Clusters gradually emerge, likely with clusters inside of clusters.
- Call above property **Modularity**.
Hierarchy by division

Bottom up problems:

- Tend to plainly not work on data sets with known modular structures.
- Good at finding cores of well-connected (or similar) nodes... but fail to cope well with peripheral, in-between nodes.
Hierarchy by division

Top down:

- **Idea:** Identify *global structure first* and recursively uncover more detailed structure.

- **Basic objective:** find dominant components that have significantly more links within than without, as compared to randomized version.


- See also
  

Hierarchy by division

- **Idea:**
  Edges that connect communities have higher betweenness than edges within communities.
Hierarchy by division

One class of structure-detection algorithms:

1. Compute edge betweenness for whole network.
2. Remove edge with highest betweenness.
3. Recompute edge betweenness
4. Repeat steps 2 and 3 until all edges are removed.

5. Record when components appear as a function of # edges removed.
6. Generate dendogram revealing hierarchical structure.

Red line indicates appearance of four (4) components at a certain level.
Hierarchy by division

Key element:

- Recomputing betweenness.
- **Reason:** Possible to have a low betweenness in links that connect large communities if other links carry majority of shortest paths.

When to stop?:

- How do we know which divisions are meaningful?
- **Modularity measure:** difference in fraction of within component nodes to that expected for randomized version:

\[
Q = \sum_i [e_{ii} - (\sum_j e_{ij})^2] = \text{Tr}E - \|E^2\|_1,
\]

where \(e_{ij}\) is the fraction of edges between identified communities \(i\) and \(j\).
Hierarchy by division

Test case:

- Generate random community-based networks.
- $N = 128$ with four communities of size 32.
- Add edges randomly within and across communities.
- Example:
  \[
  \langle k \rangle_{\text{in}} = 6 \text{ and } \langle k \rangle_{\text{out}} = 2.
  \]
Hierarchy by division

- Maximum modularity $Q \sim 0.5$ obtained when four communities are uncovered.
- Further ‘discovery’ of internal structure is somewhat meaningless, as any communities arise accidentally.
B. Zachary's karate club network

We now turn to applications of our methods to real-world networks. Figure 10 shows the collaborations network between physicists who conduct research in condensed matter physics. The network as presented in Fig. 10 consists of researchers from mostly Boston University physicists and scientists from other institutions, including those in the rest of the U.S., Japan, and Europe. Figure 10 was constructed by adding to the network as vertices, and edges between them indicate coauthorship of one or more papers appearing in the Physics archive at arxiv.org, specifically the condensed-matter section. Thus the collaborative ties represented in the figure instance, are among the nodes in this network.

The network of collaborations going on between the groups on the lower right of the picture consists of researchers in the U.S. and Europe, and the large group in the middle which consists of researchers mostly in southern Europe. In the last panel of Fig. 9, we show the dendrogram and recalculation.

Hierarchy by division

Fig. 9. In the leftmost two panels, we show the dendrograms generated by the shortest-path and random-walk versions of our algorithm. In this example, the network is split into two groups. The administrator after the fission of the club, open circles those who joined after the club split up.

The network as presented in Fig. 10 consists of researchers from mostly Boston University physicists and scientists from other institutions, including those in the rest of the U.S., Japan, and Europe. Figure 10 was constructed by adding to the network as vertices, and edges between them indicate coauthorship of one or more papers appearing in the Physics archive at arxiv.org, specifically the condensed-matter section. Thus the collaborative ties represented in the figure instance, are among the nodes in this network.

The network of collaborations going on between the groups on the lower right of the picture consists of researchers in the U.S. and Europe, and the large group in the middle which consists of researchers mostly in southern Europe. In the last panel of Fig. 9, we show the dendrogram and recalculation.

Factions in Zachary’s karate club network. [10]
Betweenness for electrons:

- Unit resistors on each edge.
- For every pair of nodes $s$ (source) and $t$ (sink), set up unit currents in at $s$ and out at $t$.
- Measure absolute current along each edge $\ell$, $|I_{\ell,st}|$.

- Sum $|I_{\ell,st}|$ over all pairs of nodes to obtain electronic betweenness for edge $\ell$.
- (Equivalent to random walk betweenness.)
- Electronic betweenness for edge between nodes $i$ and $j$:

\[ B_{ij}^{\text{elec}} = a_{ij} |V_i - V_j|. \]
Electronic betweenness

- Define some arbitrary voltage reference.
- Kirchoff’s laws: current flowing out of node $i$ must balance:

$$\sum_{j=1}^{N} \frac{1}{R_{ij}} (V_j - V_i) = \delta_{is} - \delta_{it}.$$  

- Between connected nodes, $R_{ij} = 1 = a_{ij} = 1/a_{ij}$.
- Between unconnected nodes, $R_{ij} = \infty = 1/a_{ij}$.
- We can therefore write:

$$\sum_{j=1}^{N} a_{ij} (V_i - V_j) = \delta_{is} - \delta_{it}.$$  

- Some gentle jiggery pokery on the left hand side:

$$\sum_j a_{ij} (V_i - V_j) = V_i \sum_j a_{ij} - \sum_j a_{ij} V_j$$
$$= V_i k_i - \sum_j a_{ij} V_j = k_i \delta_{ij} V_j - \sum_j a_{ij} V_j = [(K - A) \tilde{V}]_i.$$
Electronic betweenness

- Write right hand side as \([I^{\text{ext}}]_i = \delta_{is} - \delta_{it}\), where \(I^{\text{ext}}\) holds external source and sink currents.
- Matrixingly then:

\[(K - A) \vec{V} = I^{\text{ext}}.\]

- \(L = K - A\) is a beast of some utility—known as the Laplacian.
- Solve for voltage vector \(\vec{V}\) by **LU decomposition** (Gaussian elimination).
- Do not compute an inverse!
- **Note:** voltage offset is arbitrary so no unique solution.
- Presuming network has one component, null space of \(K - A\) is one dimensional.
- In fact, \(\mathcal{N}(K - A) = \{c\vec{1}, c \in R\}\) since \((K - A)\vec{1} = \vec{0}\). 

Electronic betweenness

- Write right hand side as \([I^{\text{ext}}]_i = \delta_{is} - \delta_{it}\), where \(I^{\text{ext}}\) holds external source and sink currents.
- Matrixingly then:

\[(K - A) \vec{V} = I^{\text{ext}}.\]

- \(L = K - A\) is a beast of some utility—known as the Laplacian.
- Solve for voltage vector \(\vec{V}\) by **LU decomposition** (Gaussian elimination).
- Do not compute an inverse!
- **Note:** voltage offset is arbitrary so no unique solution.
- Presuming network has one component, null space of \(K - A\) is one dimensional.
- In fact, \(\mathcal{N}(K - A) = \{c\vec{1}, c \in R\}\) since \((K - A)\vec{1} = \vec{0}\). 

Electronic betweenness

- Write right hand side as \([I^{\text{ext}}]_i = \delta_{is} - \delta_{it}\), where \(I^{\text{ext}}\) holds external source and sink currents.
- Matrixingly then:

\[(K - A) \vec{V} = I^{\text{ext}}.\]

- \(L = K - A\) is a beast of some utility—known as the Laplacian.
- Solve for voltage vector \(\vec{V}\) by **LU decomposition** (Gaussian elimination).
- Do not compute an inverse!
- **Note:** voltage offset is arbitrary so no unique solution.
- Presuming network has one component, null space of \(K - A\) is one dimensional.
- In fact, \(\mathcal{N}(K - A) = \{c\vec{1}, c \in R\}\) since \((K - A)\vec{1} = \vec{0}\).
Alternate betweenness measures:

Random walk betweenness:

- **Asking too much:** Need full knowledge of network to travel along shortest paths.
- **One of many alternatives:** Consider all random walks between pairs of nodes $i$ and $j$.
- **Walks start at node $i$, traverses the network randomly, ending as soon as it reaches $j$.**
- **Record the number of times an edge is followed by a walk.**
- **Consider all pairs of nodes.**
- **Random walk betweenness of an edge = absolute difference in probability a random walk travels one way versus the other along the edge.**
- **Equivalent to electronic betweenness.**
Hierarchy by division

- Third column shows what happens if we don’t recompute betweenness after each edge removal.
Scientists working on networks

(a)

FIG. 10. Illustration of the use of the community-structure algorithm to make sense of a complex network.

The initial network is a network of coauthorships between physicists who have published on topics related to networks. The figure shows only the largest component of the network, which contains 145 scientists. There are 90 more scientists in smaller components, which are not shown.

Application of the shortest-path betweenness version of the community-structure algorithm produces the communities indicated by the shades of the vertices.

A coarse-graining of the network in which each community is represented by a single node, with edges representing collaborations between communities. The thickness of the edges is proportional to the number of pairs of collaborators between communities. Clearly panel (c) reveals much that is not easily seen in the original network of panel (a).
Scientists working on networks

FIG. 10. Illustration of the use of the community-structure algorithm to make sense of a complex network.

The initial network is a network of coauthorships between physicists who have published on topics related to networks. The figure shows only the largest component of the network, which contains 145 scientists. There are 90 more scientists in smaller components, which are not shown.

Application of the shortest-path betweenness version of the community-structure algorithm produces the communities indicated by the shades of the vertices.

A coarse-graining of the network in which each community is represented by a single node, with edges representing collaborations between communities. The thickness of the edges is proportional to the number of pairs of collaborators between communities. Clearly panel (c) reveals much that is not easily seen in the original network of panel (a).
Scientists working on networks

FIG. 10. Illustration of the use of the community-structure algorithm to make sense of a complex network.

(a) The initial network is a network of coauthorships between physicists who have published on topics related to networks. The figure shows only the largest component of the network, which contains 145 scientists. There are 90 more scientists in smaller components, which are not shown.

(b) Application of the shortest-path betweenness version of the community-structure algorithm produces the communities indicated by the shades of the vertices.

(c) A coarse-graining of the network in which each community is represented by a single node, with edges representing collaborations between communities. The thickness of the edges is proportional to the number of pairs of collaborators between communities. Clearly panel (c) reveals much that is not easily seen in the original network of panel (a).
Dolphins!

The split into two groups appears to correspond to a known division of the dolphin community. Lusseau reports that for a period of about two years during observation of the dolphins they separated into two groups along the lines found by our analysis, apparently because of the disappearance of individuals on the boundary between the groups. When some of these individuals later reappeared, the two halves of the network joined together once more. As Lusseau points out, developments of this kind illustrate that the dolphin network is not merely a scientific curiosity but, like human social networks, is closely tied to the evolution of the community. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals: the largest subgroup consists almost of entirely of females and the others almost entirely of males, and it is conjectured that the split between the male groups is governed by matrilineage.
Les Miserables

The split into two groups appears to correspond to a known division of the dolphin community. Lusseau reports that for a period of about two years during observation of the dolphins they separated into two groups along the lines found by our analysis, apparently because of the disappearance of individuals on the boundary between the groups. When some of these individuals later reappeared, the two halves of the network joined together once more. As Lusseau points out, developments of this kind illustrate that the dolphin network is not merely a scientific curiosity but, like human social networks, is closely tied to the evolution of the community. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals: the largest subgroup consists almost of entirely of females and the others almost entirely of males, and it is conjectured that the split between the male groups is governed by matrilineage.

Figure 12 shows the community structure of the network of interactions between major characters in Victor Hugo’s sprawling novel of crime and redemption in post-restoration.
Shuffling for structure

- “Extracting the hierarchical organization of complex systems”
  Sales-Pardo et al., PNAS (2007) [8, 9]
- Consider all partitions of networks into $m$ groups
- As for Newman and Girvan approach, aim is to find partitions with maximum modularity:

$$Q = \sum_i [e_{ii} - (\sum_j e_{ij})^2] = \text{Tr}E - \|E^2\|_1.$$
Shuffling for structure

- Consider **partition network**, i.e., the network of all possible partitions.
- **Defn:** Two partitions are connected if they differ only by the reassignment of a single node.
- Look for local maxima in partition network.
- Construct an **affinity matrix** with entries $A_{ij}$.
- $A_{ij} = \Pr$ random walker on modularity network ends up at a partition with $i$ and $j$ in the same group.
- C.f. **topological overlap** between $i$ and $j = \#$ matching neighbors for $i$ and $j$ divided by maximum of $k_i$ and $k_j$. 
Shuffling for structure

- **A**: Base network; **B**: Partition network; **C**: Coclassification matrix; **D**: Comparison to random networks (all the same!); **E**: Ordered coclassification matrix; Conclusion: no structure...
Shuffling for structure

- Method obtains a distribution of classification hierarchies.
- Note: the hierarchy with the highest modularity score isn’t chosen.
- Idea is to weight possible hierarchies according to their basin of attraction’s size in the partition network.
- **Next step:** Given affinities, now need to sort nodes into modules, submodules, and so on.
- **Idea:** permute nodes to minimize following cost

\[
C = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} |i - j|.
\]

- Use simulated annealing (slow).
- **Observation:** should achieve same results for more general cost function:  
  \[
  C = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} f(|i - j|)
  \]
  where \(f\) is a strictly monotonically increasing function of 0, 1, 2, ...
Let $\tilde{P}_{\text{max}}$ be the set of partitions for which the modularity $M$ is a local maxima, that is, partitions for which neither $\tilde{P}$ nor $\tilde{P}'$ are other local maxima. Algorithms are not.

We propose that the affinity $\mathbf{P}_{ij}$ of a pair of nodes $(i,j)$ to $\tilde{P}_{\text{max}}$ is the probability that starting from partition $\tilde{P}_{\text{max}}$, two single node transforms one partition into the other. This landscape has node $\tilde{a}$ and for simplicity, we connect two partitions if the change of a single

node $\tilde{b}$.

In the hierarchical clustering tree, the vertical axis shows clustering and the “box clustering” method we propose. In the hierarchical clustering tree, the vertical axis shows clustering and the “box clustering” method we propose.

We generate a model network comprised of 640 nodes with average degree 16 and with a three-level hierarchical organization of the network. (see Fig. 1B), the levels are very cohesive, for high values of $\rho$. The most straightforward way to determine the hierarchical structure (see

with average degree 16 and with a three-level hierarchical organization of the network. (see

for two different measures: (see

modular structure beyond a locally dense connectivity pattern. In contrast, the coclassification measure clearly shades of the color used for the modules in the level lower level, we color the nodes in the submodules with different colors indicate different modules at that level. To better identify which are the submodules at a hierarchy.

Consider again the landscape in Fig. 1. The most straightforward way to determine the hierarchical structure (see

$P_{\text{max}}$) is the probability that when local maxima partition $\tilde{P}_{\text{max}}$, the probability that when local maxima partition $\tilde{P}_{\text{max}}$, the probability that when local maxima partition $\tilde{P}_{\text{max}}$.

$N = 640$, $\langle k \rangle = 16$, 3 tiered hierarchy.
Shuffling for structure

Table 1. Top-level structure of real-world networks

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>Modules</th>
<th>Main modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air transportation</td>
<td>3,618</td>
<td>28,284</td>
<td>57</td>
<td>8</td>
</tr>
<tr>
<td>E-mail</td>
<td>1,133</td>
<td>10,902</td>
<td>41</td>
<td>8</td>
</tr>
<tr>
<td>Electronic circuit</td>
<td>516</td>
<td>686</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>Escherichia coli KEGG</td>
<td>739</td>
<td>1,369</td>
<td>39</td>
<td>13</td>
</tr>
<tr>
<td>E. coli UCSD</td>
<td>507</td>
<td>947</td>
<td>28</td>
<td>17</td>
</tr>
</tbody>
</table>
Shuffling for structure

- Modules found match up with geopolitical units.
**Shuffling for structure**

![Graphical representation](image)

- **Modularity structure for metabolic network of E. coli (UCSD reconstruction).**
General structure detection

▶ “Detecting communities in large networks” Capocci et al. (2005) [1]
▶ Consider normal matrix $K^{-1}A$, random walk matrix $A^T K^{-1}$, Laplacian $K - A$, and $AA^T$.
▶ Basic observation is that eigenvectors associated with secondary eigenvalues reveal evidence of structure.
▶ Build on Kleinberg’s HITS algorithm.
General structure detection

- Example network:

![Network Diagram](image)

Thus, solving the eigenproblem is equivalent to minimizing function (1) with constraint (2), where the $x_i$'s are eigenvectors components. The absolute minimum corresponds to the trivial eigenvector, which is constant. The other stationary points correspond to eigenvectors where components associated to well connected nodes assume similar values.

In order to compute cluster sizes and distribution, methods such as bisection or edge-betweenness based ones are very poor in detecting the end of the recursive splitting. Our approach, instead, immediately detects the number of clear clusters from the eigenvectors profile.

As an illustrative example, we show in Fig. 2 the profile of the second eigenvectors of $D/C_0 W$ corresponding to the simple graph shown in Fig. 1 with $S = 19$ nodes, where random weights between 1 and 10 were assigned to the links (Figs. 1 and 2).

The components of the eigenvectors assume approximately constant values on nodes belonging to the same community. Thus, the number of communities emerges naturally and it is not needed as input.

However, as aforementioned, when dealing with large networks with no clear partitioning, the precise value of the eigenvector components is of little use. In such situations, the typical eigenvector profile is not step-like, but resembles a continuous curve. Nevertheless, our method can still be applied, and efficiently detects sets of well connected nodes. In fact, components corresponding to nodes belonging to the same communities are still strongly correlated taking, in each eigenvector, similar values among themselves.

Thus, a natural way to identify communities in an automatic manner, is to measure the correlation $r_{ij} = \frac{\langle x_i x_j \rangle}{\sqrt{\langle x_i^2 \rangle \langle x_j^2 \rangle}}$, (4), where the average $\langle \cdot \rangle$ is over the first few nontrivial eigenvectors. The quantity $r_{ij}$ measures the community closeness between nodes $i$ and $j$.

Though the performance may be improved by averaging over more and more eigenvectors, with increased computational effort, we find that indeed a small number of eigenvectors suffices to identify the community to which nodes belong, even in large networks.
When dealing with a directed network, links do not correspond to any equivalence relation. Rather, pointing to common neighbors is a significant relation, as suggested in the sociologists’ literature where this quantity measures the so-called structural equivalence of nodes. Accordingly, in a directed network, clusters should be composed by nodes pointing to a high number of common neighbors, no matter their direct linkage. For directed networks, we thus modify our method in the streamline of the HITS algorithm. The HITS algorithm was proposed on empirical bases to find the main communities in large oriented networks. It assumes that the largest components (in the absolute value) of eigenvectors of the matrices $AA^T$ and $A^TA$ correspond to highly clustered nodes belonging to a single community. Such algorithm efficiently detects the main communities, even when these are not sharply defined. However, it becomes computationally heavy when one is interested in minor communities, which correspond to smaller eigenvalues. As explained in the undirected case, we tackle this issue by combining information from the first few eigenvectors of the normal matrix and extracting the community structure from correlations between the same components in different eigenvectors. To detect the community structure in a directed network, we therefore replace, in the previous analysis, the matrix $W$ with a matrix $Y = WW^T$: This corresponds to replacing the directed network with an undirected weighted network, where nodes pointing to common neighbors are connected by a link, whose intensity is proportional to the total sum of the weights of the links pointing from the two original nodes to the common neighbors. Then, one performs the analysis on the undirected network as described previously. Thus, the function to minimize in this case is

$$y(x) = \sum_{ij} S_{ijl} (x_i/C_0 x_j)^2 w_{il} w_{jl}. \quad (5)$$
General structure detection

- Network of word associations for 10616 words.
- Average in-degree of 7.
- Using 2nd to 11th eigenvectors of a modified version of $\mathbf{A}\mathbf{A}^T$.

Table 1
Words most correlated to science, literature and piano in the eigenvectors of $Q^{-1}\mathbf{W}\mathbf{W}^T$

<table>
<thead>
<tr>
<th>Science</th>
<th>1</th>
<th>Literature</th>
<th>1</th>
<th>Piano</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific</td>
<td>0.994</td>
<td>Dictionary</td>
<td>0.994</td>
<td>Cello</td>
<td>0.993</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.990</td>
<td>Editorial</td>
<td>0.990</td>
<td>Fiddle</td>
<td>0.992</td>
</tr>
<tr>
<td>Physics</td>
<td>0.988</td>
<td>Synopsis</td>
<td>0.988</td>
<td>Viola</td>
<td>0.990</td>
</tr>
<tr>
<td>Concentrate</td>
<td>0.973</td>
<td>Words</td>
<td>0.987</td>
<td>Banjo</td>
<td>0.988</td>
</tr>
<tr>
<td>Thinking</td>
<td>0.973</td>
<td>Grammar</td>
<td>0.986</td>
<td>Saxophone</td>
<td>0.985</td>
</tr>
<tr>
<td>Test</td>
<td>0.973</td>
<td>Adjective</td>
<td>0.983</td>
<td>Director</td>
<td>0.984</td>
</tr>
<tr>
<td>Lab</td>
<td>0.969</td>
<td>Chapter</td>
<td>0.982</td>
<td>Violin</td>
<td>0.983</td>
</tr>
<tr>
<td>Brain</td>
<td>0.965</td>
<td>Prose</td>
<td>0.979</td>
<td>Clarinet</td>
<td>0.983</td>
</tr>
<tr>
<td>Equation</td>
<td>0.963</td>
<td>Topic</td>
<td>0.976</td>
<td>Oboe</td>
<td>0.983</td>
</tr>
<tr>
<td>Examine</td>
<td>0.962</td>
<td>English</td>
<td>0.975</td>
<td>Theater</td>
<td>0.982</td>
</tr>
</tbody>
</table>

Values indicate the correlation.
Hierarchies and missing links

- Idea: Shades indicate probability that nodes in left and right subtrees of dendogram are connected.
- Handle: Hierarchical random graph models.
- Plan: Infer consensus dendogram for a given real network.
- Obtain probability that links are missing (big problem...).
Model also predicts reasonably well

1. average degree,
2. clustering,
3. and average shortest path length.

Table 1 | Comparison of original and resampled networks

<table>
<thead>
<tr>
<th>Network</th>
<th>$\langle k \rangle_{\text{real}}$</th>
<th>$\langle k \rangle_{\text{samp}}$</th>
<th>$C_{\text{real}}$</th>
<th>$C_{\text{samp}}$</th>
<th>$d_{\text{real}}$</th>
<th>$d_{\text{samp}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>T. pallidum</em></td>
<td>4.8</td>
<td>3.7(1)</td>
<td>0.0625</td>
<td>0.0444(2)</td>
<td>3.690</td>
<td>3.940(6)</td>
</tr>
<tr>
<td>Terrorists</td>
<td>4.9</td>
<td>5.1(2)</td>
<td>0.361</td>
<td>0.352(1)</td>
<td>2.575</td>
<td>2.794(7)</td>
</tr>
<tr>
<td>Grassland</td>
<td>3.0</td>
<td>2.9(1)</td>
<td>0.174</td>
<td>0.168(1)</td>
<td>3.29</td>
<td>3.69(2)</td>
</tr>
</tbody>
</table>

Statistics are shown for the three example networks studied and for new networks generated by resampling from our hierarchical model. The generated networks closely match the average degree $\langle k \rangle$, clustering coefficient $C$ and average vertex–vertex distance $d$ in each case, suggesting that they capture much of the structure of the real networks. Parenthetical values indicate standard errors on the final digits.
Hierarchies and missing links

- **Consensus dendrogram** for grassland species.
- Copes with disassortative and assortative communities.
General structure detection

“The discovery of structural form”
General structure detection

- Top down description of form.
- Node replacement graph grammar: parent node becomes two child nodes.
- B-D: Growing chains, orders, and trees.

Diagram:

- **A** Structural Form
  - Partition
  - Chain
  - Order
  - Ring
  - Hierarchy
  - Tree
  - Grid
  - Cylinder

- **B** Generative process
  - Partition
  - Chain
  - Order
  - Ring
  - Hierarchy
  - Tree
  - Grid
  - Cylinder

- **C**
  - Generation rules

- **D**
  - Chain II Chain
  - Chain II Ring
Example learned structures:

- Biological features; Supreme Court votes; perceived color differences; face differences; & distances between cities.
General structure detection

Effect of adding features on detected form.

Straight partition

↓

simple tree

↓

complex tree
General structure detection

Performance for test networks.

<table>
<thead>
<tr>
<th>True</th>
<th>Partition</th>
<th>Chain</th>
<th>Ring</th>
<th>Tree</th>
<th>Grid</th>
</tr>
</thead>
</table>

- For each structure, the bar chart indicates the log posterior probability.

---

References
Final words:

Modern science in three steps:

1. Find interesting/meaningful/important phenomena involving spectacular amounts of data.
2. Describe what you see.
3. Explain it.
References I


References II


References III


An information flow model for conflict and fission in small groups.