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

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.
Finding and evaluating community structure in networks
M. E. J. Newman1,2 and M. Girvan2,3
1Department of Physics and... ©2004 The American Physical Society

Tend to plainly not work on data sets with known edges... but fail to cope well with peripheral, in-between nodes.

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.
- We’ll first work through “Finding and evaluating community structure in networks” by Newman and Girvan (PRE, 2004).
- See also

Hierarchical clustering methods:
- Overview
- Methods
- Hierarchy by aggregation
- Hierarchy by division
- Hierarchy by shuffling
- Spectral methods
- Missing links
- General structure detection
- Final words
- References

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.

Idea:
Edges that connect communities have higher betweenness than edges within communities.
Tests on computer-generated networks

First, as a controlled test of how well our algorithms perform, we have...

Maximum modularity $Q \approx 0.5$ obtained when four communities are uncovered.

Further ‘discovery’ of internal structure is somewhat meaningless, as any communities arise accidentally.

Test case:

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

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 $\ell$, $|\ell_{st}|$.
- Sum $|\ell_{st}|$ 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

- Write right hand side as $[l_{\text{ext}}]_i = \delta_{is} - \delta_{it}$, where $l_{\text{ext}}$ holds external source and sink currents.
- Matrixingly then:
  \[
  (K - A) \vec{V} = l_{\text{ext}}.
  \]
- $L = K - A$ 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, $N(K - A) = \{c \vec{1}, c \in \mathbb{R}\}$ since $(K - A) \vec{1} = \vec{0}$.

Electronic betweenness

- Define some arbitrary voltage reference.
- Kirchoff’s laws: current flowing out of node $i$ must balance:
  \[
  \sum_{j=1}^{N} R_{ij} (V_j - V_i) = \delta_{is} - \delta_{it}.
  \]
- Between connected nodes, $R_{ij} = 1 = a_{ij} = 1/a_{ij}$.
- Between unconnected nodes, $R_{ij} = \infty = 1/a_{ij}$.
- We can therefore write:
  \[
  \sum_{j=1}^{N} a_{ij} (V_j - V_i) = \delta_{is} - \delta_{it}.
  \]
- Some gentle jiggery pokery on the left hand side:
  \[
  \sum_{j} a_{ij} (V_i - V_j) = V_i \sum_{j} a_{ij} - \sum_{j} a_{ij} V_j
  = V_i k_i - \sum_{j} a_{ij} V_j = k_i \delta_{ij} V_j - \sum_{j} a_{ij} V_j = [(K - A) \vec{V}]_i
  \]

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 starts 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 $= \text{absolute difference in probability a random walk travels one way versus the other along the edge}$.
- 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)

(b)

(c)

Scientists working on networks
Dolphins!

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:

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

Les Miserables

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 $A_{ij}$.
- $A_{ij} = \text{Pr}$ 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...

- $N = 640$, $\langle k \rangle = 16$, 3 tiered hierarchy.

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, ...

**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><em>Escherichia coli</em> KEGG</td>
<td>739</td>
<td>1,369</td>
<td>39</td>
<td>13</td>
</tr>
<tr>
<td><em>E. coli</em> UCSD</td>
<td>507</td>
<td>947</td>
<td>28</td>
<td>17</td>
</tr>
</tbody>
</table>
General structure detection

- “Detecting communities in large networks” Capocci et al. (2005) [1]
- Consider normal matrix $K^{-1}A$, random walk matrix $A^TK^{-1}$, 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.
Hierarchies and missing links

Clauset et al., Nature (2008)

- 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...).

General structure detection

- Second eigenvector’s components:

\[ x_i \]

0 0.2 0.4
-0.2 -0.4
0 5 10 15 20

Hierarchies and missing links

- Network of word associations for 10616 words.
- Average in-degree of 7.
- Using 2nd to 11th vectors of a modified version of \( AA^T \):

<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>Synopsus</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

- Consensus dendrogram for grassland species.
- Copes with disassortative and assortative communities.

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.

General structure detection


Example learned structures:

- Biological features: Supreme Court votes; perceived color differences; face differences; & distances between cities.
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 II

The discovery of structural form.
pdf

Scientific collaboration networks. II. Shortest paths,
weighted networks, and centrality.

Erratum: Scientific collaboration networks. II.
Shortest paths, weighted networks, and centrality
[Phys. Rev. E 64, 016132 (2001)].

References III

Finding and evaluating community structure in
networks.

[8] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and
L. A. N. Amaral.
Extracting the hierarchical organization of complex
systems.

[9] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and
L. A. N. Amaral.
Extracting the hierarchical organization of complex
systems: Correction.

References IV

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