Structure detection methods
Complex Networks, Course 303A, Spring, 2009

Prof. Peter Dodds
Department of Mathematics & Statistics
University of Vermont

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

Structure detection

- Zachary’s karate club
- 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?

Outline

Overview

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

References

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

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.

![Diagram showing community structure]

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

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.
A. Tests on computer-generated networks

First, as a controlled test of how well our algorithms perform, we have generated networks with known community structure. This allows us to assess the accuracy of our algorithms in identifying communities. The results show that our algorithms can correctly identify the communities with a mean number of connections to vertices outside their community as it out all the way to around 0.5, which is typical.

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 \( j \).

- 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.

B. Zachary’s karate club network

We now turn to applications of our methods to real-world networks. One such example is the Zachary karate club network, which consists of researchers and their interactions over a period of two years in the early 1970s. Wayne Zachary observed social interactions between the club's members and recorded ties between them. The network as presented in Fig. 10 has a strong hierarchy, and it is clear that the group divisions are well-identified by our algorithms. The peak of modularity is obtained when four communities are uncovered.

In the last panel of Fig. 9, we show the dendrogram and the network as a whole. The results suggest that the network has a hierarchical structure, and our algorithms are able to detect these layers. Furthermore, the modularity does not reach nearly 1, indicating that there are still some misclassifications.

In summary, our algorithms perform noticeably better than the random-walk version, especially in cases where the network is not too dense. We recommend the latter for most applications. However, when dealing with sparse networks, our algorithms are more effective in capturing community structure.
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:
  \[ B^\text{elec}_{ij} = a_{ij}|V_i - V_j|. \]

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.
- Matrixing 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, \(N(K - A) = \{c\vec{1}, c \in R\}\) 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 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 = absolute difference in probability a random walk travels one way versus the other along the edge.
- Equivalent to electronic betweenness.
Third column shows what happens if we don’t recompute betweenness after each edge removal.

Scientists working on networks

(b)

(c)
As for Newman and Girvan approach, aim is to find structures in networks. Consider all partitions of networks into $m$ groups, and the circles are subdivided further into smaller groups as shown. The modularity for the split is $Q = \sum_i [e_{ii} - \left( \sum_j e_{ij} \right)^2] = \text{Tr}E - \|E^2\|_1$, extracted using the shortest-path version of our algorithm. The squares and circles denote the primary split of the network into two groups, and the circles are subdivided further into four smaller groups as shown. The modularity for the split is $Q = 0.52$. The network has been drawn with longer edges between vertices in different communities than between those in the same community. The subgroupings within the larger half of the network also seem to correspond to real divisions among the dolphins they separated into two groups along the known division of the dolphin community.

Conjectured that the split between the male groups is governed by matrilineage, and conjectured that the split between the females and the others almost entirely of males, and it is consistent with the human social networks, is closely tied to the evolution of the animal social systems. The greatest modularity obtained in the shortest-path version of our algorithm is $Q = 0.54$ and corresponds to the 11 communities shown.

Points out, developments of this kind illustrate that the dolphins separated into two groups along the known division of the dolphin community conjectured that the split between the male groups is governed by matrilineage, and conjectured that the split between the females and the others almost entirely of males, and it is consistent with the human social networks, is closely tied to the evolution of the animal social systems. The greatest modularity obtained in the shortest-path version of our algorithm is $Q = 0.54$ and corresponds to the 11 communities shown.

When some of these individuals later reappeared, the two lines found by our analysis, apparently because of the disappearance of individuals on the boundary between the groups. The greatest modularity obtained in the shortest-path version of our algorithm is $Q = 0.54$ and corresponds to the 11 communities shown.

Consider partition network, i.e., the network of all possible partitions. Define: 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. Consider partition network, i.e., the network of all possible partitions. Define: 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 $A_{ij}$.
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, ...

Shuffling for structure

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

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.

Modularity structure for metabolic network of E. coli (UCSD reconstruction).

Example network:
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

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

Hierarchies and missing links

- 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 dendogram 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.

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

References I

Detecting communities in large networks.

Hierarchical structure and the prediction of missing links in networks.

Community structure in social and biological networks.

References II

The discovery of structural form.

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)].

