Optimal supply & Structure detection
Complex Networks, SFI Summer School, June, 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

Single Source

Distributed Sources
- Facility location
- Size-density law
- A reasonable derivation
- Global redistribution networks

Structure Detection
- Hierarchy by division
- Hierarchy by shuffling
- Spectral methods
- Hierarchies & Missing Links
- General structure detection

Final words

References
What’s the best way to distribute stuff?

- Stuff = medical services, energy, nutrients, people, ...
- Some fundamental network problems:
  1. Distribute stuff from single source to many sinks
  2. Collect stuff coming from many sources at a single sink
  3. Distribute stuff from many sources to many sinks
  4. Redistribute stuff between many nodes
- Q: How do optimal solutions scale with system size?
Basic Q for distribution/supply networks:

- How does flow behave given cost:
  \[ C = \sum_j I_j^\gamma Z_j \]
  where
  \( I_j \) = current on link \( j \)
  and
  \( Z_j \) = link \( j \)'s impedance?

- Example: \( \gamma = 2 \) for electrical networks.
Single source optimal supply

(a) $\gamma > 1$: Braided (bulk) flow
(b) $\gamma < 1$: Local minimum: Branching flow
(c) $\gamma < 1$: Global minimum: Branching flow

From Bohn and Magnasco\textsuperscript{[3]}
See also Banavar et al.\textsuperscript{[1]}
Single source optimal supply

Optimal paths related to transport (Monge) problems:

Xia (2003) [24]
Growing networks:

FIGURE 1. $\alpha = 0.6, \beta = 0.5$

α=0.6, β=0.5, ε=2

α=0.6, β=0.5, ε=3

α=0.6, β=0.5, ε=4

α=0.6, β=0.5, ε=5

Xia (2007) [23]
Growing networks:

**Figure 3. A maple leaf**

Xia (2007) [23]
Single source optimal supply

An immensely controversial issue...

- The form of river networks and blood networks: optimal or not? \([22, 2, 7]\)

Two observations:

- Self-similar networks appear everywhere in nature for single source supply/single sink collection.
- Real networks differ in details of scaling but reasonably agree in scaling relations.
Stream Ordering:

- Label all source streams as order $\omega = 1$.
- Follow all labelled streams downstream
- Whenever two streams of the same order ($\omega$) meet, the resulting stream has order incremented by 1 ($\omega + 1$).

- If streams of different orders $\omega_1$ and $\omega_2$ meet, then the resultant stream has order equal to the largest of the two.
- Simple rule: 
  \[
  \omega_3 = \max(\omega_1, \omega_2) + \delta_{\omega_1, \omega_2}
  \]

  where $\delta$ is the Kronecker delta.
Horton’s laws in the real world:

The Mississippi

The Nile

The Amazon

Single Source
Distributed Sources
Facility location
Size-density law
A reasonable derivation
Global redistribution networks
Structure Detection
Hierarchy by division
Hierarchy by shuffling
Spectral methods
Hierarchies & Missing Links
General structure detection
Final words
References
Many scaling laws, many connections

<table>
<thead>
<tr>
<th>relation:</th>
<th>scaling relation/parameter: [6]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell \sim L^d$</td>
<td>$d$</td>
</tr>
<tr>
<td>$T_k = T_1(R_T)^{k-1}$</td>
<td>$T_1 = R_n - R_s - 2 + 2R_s/R_n$</td>
</tr>
<tr>
<td>$n_\omega/n_{\omega+1} = R_n$</td>
<td>$R_n$</td>
</tr>
<tr>
<td>$\bar{a}<em>{\omega+1}/\bar{a}</em>\omega = R_a$</td>
<td>$R_a = R_n$</td>
</tr>
<tr>
<td>$\bar{\ell}<em>{\omega+1}/\bar{\ell}</em>\omega = R_\ell$</td>
<td>$R_\ell = R_s$</td>
</tr>
<tr>
<td>$\ell \sim a^h$</td>
<td>$h = \log R_s / \log R_n$</td>
</tr>
<tr>
<td>$a \sim L^D$</td>
<td>$D = d/h$</td>
</tr>
<tr>
<td>$L_\perp \sim L^H$</td>
<td>$H = d/h - 1$</td>
</tr>
<tr>
<td>$P(a) \sim a^{-\tau}$</td>
<td>$\tau = 2 - h$</td>
</tr>
<tr>
<td>$P(\ell) \sim \ell^{-\gamma}$</td>
<td>$\gamma = 1/h$</td>
</tr>
<tr>
<td>$\Lambda \sim a^\beta$</td>
<td>$\beta = 1 + h$</td>
</tr>
<tr>
<td>$\lambda \sim L^\varphi$</td>
<td>$\varphi = d$</td>
</tr>
</tbody>
</table>

Only 3 parameters are independent... [6]
Reported parameter values: [6]

<table>
<thead>
<tr>
<th>Parameter:</th>
<th>Real networks:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_n$</td>
<td>3.0–5.0</td>
</tr>
<tr>
<td>$R_a$</td>
<td>3.0–6.0</td>
</tr>
<tr>
<td>$R_\ell = R_T$</td>
<td>1.5–3.0</td>
</tr>
<tr>
<td>$T_1$</td>
<td>1.0–1.5</td>
</tr>
<tr>
<td>$d$</td>
<td>1.1 ± 0.01</td>
</tr>
<tr>
<td>$D$</td>
<td>1.8 ± 0.1</td>
</tr>
<tr>
<td>$h$</td>
<td>0.50–0.70</td>
</tr>
<tr>
<td>$\tau$</td>
<td>1.43 ± 0.05</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>1.8 ± 0.1</td>
</tr>
<tr>
<td>$H$</td>
<td>0.75–0.80</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.50–0.70</td>
</tr>
<tr>
<td>$\varphi$</td>
<td>1.05 ± 0.05</td>
</tr>
</tbody>
</table>
Data from real blood networks

<table>
<thead>
<tr>
<th>Network</th>
<th>$R_n$</th>
<th>$R_r^{-1}$</th>
<th>$R_l^{-1}$</th>
<th>$-\frac{\ln R_r}{\ln R_n}$</th>
<th>$-\frac{\ln R_l}{\ln R_n}$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>West et al.</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.5</td>
<td>0.33</td>
<td>0.75</td>
</tr>
<tr>
<td>rat (PAT)</td>
<td>2.76</td>
<td>1.58</td>
<td>1.60</td>
<td>0.45</td>
<td>0.46</td>
<td>0.73</td>
</tr>
<tr>
<td>cat (PAT)</td>
<td>3.67</td>
<td>1.71</td>
<td>1.78</td>
<td>0.41</td>
<td>0.44</td>
<td>0.79</td>
</tr>
<tr>
<td>(Turcotte et al. [21])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog (PAT)</td>
<td>3.69</td>
<td>1.67</td>
<td>1.52</td>
<td>0.39</td>
<td>0.32</td>
<td>0.90</td>
</tr>
<tr>
<td>pig (LCX)</td>
<td>3.57</td>
<td>1.89</td>
<td>2.20</td>
<td>0.50</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>pig (RCA)</td>
<td>3.50</td>
<td>1.81</td>
<td>2.12</td>
<td>0.47</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>pig (LAD)</td>
<td>3.51</td>
<td>1.84</td>
<td>2.02</td>
<td>0.49</td>
<td>0.56</td>
<td>0.65</td>
</tr>
<tr>
<td>human (PAT)</td>
<td>3.03</td>
<td>1.60</td>
<td>1.49</td>
<td>0.42</td>
<td>0.36</td>
<td>0.83</td>
</tr>
<tr>
<td>human (PAT)</td>
<td>3.36</td>
<td>1.56</td>
<td>1.49</td>
<td>0.37</td>
<td>0.33</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Animal power

Fundamental biological and ecological constraint:

\[ P = c M^\alpha \]

\( P \) = basal metabolic rate

\( M \) = organismal body mass
1964: Troon, Scotland:
3rd symposium on energy metabolism.
\( \alpha = \frac{3}{4} \) made official . . .

. . . 29 to zip.
Some data on metabolic rates

Heusner’s data (1991)\textsuperscript{[11]}

- 391 Mammals
- blue line: 2/3
- red line: 3/4.
- \(B = P\)
Some regressions from the ground up...

<table>
<thead>
<tr>
<th>range of $M$</th>
<th>$N$</th>
<th>$\hat{\alpha}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 0.1$ kg</td>
<td>167</td>
<td>$0.678 \pm 0.038$</td>
</tr>
<tr>
<td>$\leq 1$ kg</td>
<td>276</td>
<td>$0.662 \pm 0.032$</td>
</tr>
<tr>
<td>$\leq 10$ kg</td>
<td>357</td>
<td>$0.668 \pm 0.019$</td>
</tr>
<tr>
<td>$\leq 25$ kg</td>
<td>366</td>
<td>$0.669 \pm 0.018$</td>
</tr>
<tr>
<td>$\leq 35$ kg</td>
<td>371</td>
<td>$0.675 \pm 0.018$</td>
</tr>
<tr>
<td>$\leq 350$ kg</td>
<td>389</td>
<td>$0.706 \pm 0.016$</td>
</tr>
<tr>
<td>$\leq 3670$ kg</td>
<td>391</td>
<td>$0.710 \pm 0.021$</td>
</tr>
</tbody>
</table>
Analysis of residuals—p-values—mammals:

(a) $M < 3.2$ kg
(b) $M < 10$ kg
(c) $M < 32$ kg
(d) all mammals.

For a-d, $ho_{2/3} > 0.05$ and $ho_{3/4} \ll 10^{-4}$. 
Analysis of residuals—p-values—birds:

- (a) $M < 0.1$ kg
- (b) $M < 1$ kg
- (c) $M < 10$ kg
- (d) all birds.

For a-d,

\[ p_{2/3} > 0.05 \] and
\[ p_{3/4} \ll 10^{-4}. \]
Many sources, many sinks

How do we distribute sources?

- Focus on 2-d (results generalize to higher dimensions)
- Sources = hospitals, post offices, pubs, ...
- **Key problem:** How do we cope with uneven population densities?
- Obvious: if density is uniform then sources are best distributed uniformly.
- Which lattice is optimal? The **hexagonal lattice**
  - **Q1:** How big should the hexagons be?
  - **Q2:** Given population density is uneven, what do we do?
Optimal source allocation

Solidifying the basic problem

- Given a region with some population distribution $\rho$, most likely uneven.
- Given resources to build and maintain $N$ facilities.
- **Q:** How do we locate these $N$ facilities so as to minimize the average distance between an individual’s residence and the nearest facility?
- Problem of interested and studied by geographers, sociologists, computer scientists, mathematicians, ...
- See work by Stephan $^{[19, 20]}$ and by Gastner and Newman (2006) $^{[8]}$ and work cited by them.
Optimal source allocation


- Approximately optimal location of 5000 facilities.
- Based on 2000 Census data.
- Simulated annealing + Voronoi tessellation.

From

Optimal supply & Structure detection

Single Source
Distributed Sources
Facility location
Size-density law
A reasonable derivation
Global redistribution networks
Structure Detection
Hierarchy by division
Hierarchy by shuffling
Spectral methods
Hierarchies & Missing Links
General structure detection
Final words
References

Frame 24/78
Optimal supply & Structure detection

Single Source
Distributed Sources
Facility location
Size-density law
A reasonable derivation
Global redistribution networks

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

Final words
References

Frame 25/78

Optimal source allocation

From Gastner and Newman (2006) [8]

- Optimal facility density $D$ vs. population density $\rho$.
- Fit is $D \propto \rho^{0.66}$ with $r^2 = 0.94$.
- Looking good for a 2/3 power...
Optimal source allocation

Size-density law:

$D \propto \rho^{2/3}$

In $d$ dimensions:

$D \propto \rho^{d/(d+1)}$

Why?

Very different story to branching networks where there is either one source or one sink.

Now sources & sinks are distributed throughout region...
Optimal source allocation

- Zipf-like approach: invokes principle of minimal effort.
- Also known as the Homer principle.
Deriving the optimal source distribution:

- **Basic idea:** Minimize the average distance from a random individual to the nearest facility.
- Assume given a fixed population density $\rho$ defined on a spatial region $\Omega$.
- Formally, we want to find the locations of $n$ sources $\{\hat{x}_1, \ldots, \hat{x}_n\}$ that minimizes the cost function

$$F(\{\hat{x}_1, \ldots, \hat{x}_n\}) = \int_{\Omega} \rho(\hat{x}) \min_i ||\hat{x} - \hat{x}_i|| d\hat{x}.$$ 

- Also known as the $p$-median problem.
- Not easy... in fact this one is an NP-hard problem. [8]
Size-density law

Can (roughly) turn into a Lagrange multiplier story:

- By varying \( \{\vec{x}_1, \ldots, \vec{x}_n\} \), minimize

\[
G(A) = c \int_\Omega \rho(\vec{x}) A(\vec{x})^{1/2} d\vec{x} - \lambda \left( n - \int_\Omega [A(\vec{x})]^{-1} d\vec{x} \right)
\]

- Involves estimating typical distance from \( \vec{x} \) to the nearest source (say \( i \)) as \( c_i A(\vec{x})^{1/2} \) where \( c_i \) is a shape factor for the \( i \)th Voronoi cell.

- **Sneakiness:** set \( c_i = c \).

- Compute \( \delta G/\delta A \), the functional derivative.

- Solve and substitute \( D = 1/A \), we find

\[
D(\vec{x}) = \left( \frac{c}{2\lambda \rho} \right)^{2/3}.
\]
Global redistribution networks

One more thing:

- How do we supply these facilities?
- How do we best redistribute mail? People?
- How do we get beer to the pubs?
- Gaster and Newman model: cost is a function of basic maintenance and travel time:

\[ C_{\text{maint}} + \gamma C_{\text{travel}}. \]

- Travel time is more complicated: Take ‘distance’ between nodes to be a composite of shortest path distance \( \ell_{ij} \) and number of legs to journey:

\[ (1 - \delta)\ell_{ij} + \delta(\#\text{hops}). \]

- When \( \delta = 1 \), only number of hops matters.
Global redistribution networks

From Gastner and Newman (2006) [8]
Structure detection

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

Zachary’s karate club [25, 16]

Possible substructures:
  - hierarchies, cliques, rings, ...

Plus:
  - All combinations of substructures.
  - Much focus on hierarchies...

Possible substructures:
- hierarchies
- cliques
- rings
...
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:

\[
\langle k \rangle_{\text{in}} = 6 \text{ and } \langle k \rangle_{\text{out}} = 2.
\]
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.
Hierarchy by division

Factions in Zachary’s karate club network. [25]
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 $j$:
  \[ B_{ij}^{\text{elec}} = a_{ij} |V_i - V_j| \]
- Upshot: specific measure of betweenness not too important.
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 reveals much that is not easily seen in the original network of panel.
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).
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.

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.
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 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) [17, 18]
- Consider all partitions of networks into $m$ groups
- As for Newman and Girvan approach, aim is to find partitions with maximum modularity:

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

- \( N = 640, \)
- \( \langle k \rangle = 16, \)
- 3 tiered hierarchy.
Air transportation:

- Modules found match up with geopolitical units.
General structure detection

▶ 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. [13]
General structure detection

- Example network:

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 $D/C_0W$ 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 $r_{ij} = \frac{\langle x_i x_j \rangle}{\langle x_i^2 \rangle^{1/2} \langle x_j^2 \rangle^{1/2}}$, (4), where the average $\langle x_i \rangle$ 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 $y(x) = \sum_{i,j,l} S_{ijkl} - C_0 S_{ij} x_i x_j w_{il} w_{jl}$. (5)

![Graph showing the second eigenvector's components](image)
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...).
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>T. pallidum</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.

---

NATURE | Vol 453 | 1 May 2008 LETTERS

Nature Publishing Group © 2008

---

Consensus dendogram

Copes with disassortative and assortative communities.

---

T. pallidum

---

Network
General structure detection

“The discovery of structural form”
Kemp and Tenenbaum, PNAS (2008) [12]
General structure detection

A Structural Form  Generative process

Partition

Chain

Order

Ring

Hierarchy

Tree

Grid

Cylinder

B C D

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

Performance for test networks.

- True
- Partition
- Chain
- Ring
- Tree
- Grid

Fig. 1: The algorithm used for each of these searches is related to the frequency model, we integrate out the parameters of a data set to the best structures found, and the best scoring structure is marked with an asterisk. The di-...
Example learned structures:

- Biological features; Supreme Court votes; perceived color differences; face differences; & distances between cities.
Effect of adding features on detected form.

Straight partition

simple tree

complex tree
Final words:

Science in three steps:

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

A plea/warning

Beware your assumptions—don’t use tools/models because they’re there, or because everyone else does...
More final words:

A real theory of everything:

1. Is not just about the small stuff...
2. It’s about the increase of complexity

Symmetry breaking/ vs. Universality
Accidents of history

How probable is a certain level of complexity?
References I


References II

Detecting communities in large networks.

Hierarchical structure and the prediction of missing links in networks.

Unified view of scaling laws for river networks.
References III


References IV


References V

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

Finding and evaluating community structure in networks.

Extracting the hierarchical organization of complex systems.
References VI


References VII


[23] Q. Xia. The formation of a tree leaf. Submitted. pdf (↗)
