A taxonomy of networks

J-P Onnela, D J Fenn, S Reid, M A Porter, P J Mucha, M Fricker, N S Jones

Jukka-Pekka “JP” Onnela

Harvard University

Lisbon, September 15, 2010
Introduction

- Role of classification in science (periodic table, phylogenetic trees, etc.)
- Aim to classify networks based on their “overall” structural properties
- Why is structural classification or categorization useful?
  - Can define distance between networks in a given category
  - Can identify potential approaches to empirical analysis across fields
  - Can identify potential approaches to theoretical modeling across fields
Introduction

• What is the appropriate structural level for comparing networks?
  • Microscopic properties (e.g. degree)
  • Macroscopic properties (e.g. diameter)
• Many networks have fat-tailed degree (strength) distributions and possess the small-world property
• Consequently, microscopic and macroscopic characteristics do not differentiate effectively between networks
Community detection

- Community? A set of densely connected nodes


Mesoscopic response function

• Choose an infinite range Potts model, which generalizes modularity methods by incorporating a resolution parameter $\lambda$ (excluding $i=j$ terms):

$$H(\lambda) = - \sum_{i,j} \left[ A_{ij} - \lambda \frac{k_i k_j}{2m} \right] \delta(\sigma_i, \sigma_j)$$

• Find the values $\Lambda_{\text{min}}$ and $\Lambda_{\text{max}}$ and sweep $\Lambda_{\text{min}} \leq \lambda \leq \Lambda_{\text{max}}$

• Number of partitions (communities) increases from 1 to $N$ (number of nodes)

$$H(\lambda) = - \sum_{i \neq j} J_{ij} \delta(C_i, C_j) = - \sum_{i \neq j} (A_{ij} - \lambda P_{ij}) \delta(C_i, C_j), \quad J_{ij} = A_{ij} - \Lambda_{ij} P_{ij} = 0$$

$$\Lambda_{\text{min}} = \max_{i,j} \{ \Lambda_{ij} | \eta(\lambda) = 1 \}$$

$$\Lambda_{\text{max}} = \max_{i,j} \{ \Lambda_{ij} \} + \epsilon$$

$$\Lambda^* = \min_{i,j} \{ \Lambda_{ij} | A_{ij} > 0 \}$$
Mesoscopic response function

- **Effective energy**

\[
H_{\text{eff}}(\lambda) = \frac{H(\lambda) - H_{\min}}{H_{\max} - H_{\min}}
\]

- **Effective partition entropy**

\[
S_{\text{eff}}(\lambda) = \frac{S(\lambda) - S_{\min}}{S_{\max} - S_{\min}}
\]

- **Effective number of communities**

\[
\eta_{\text{eff}}(\lambda) = \frac{\eta(\lambda) - \eta_{\min}}{\eta_{\max} - \eta_{\min}}
\]

\[
p_k = \frac{n_k}{N}
\]

\[
S(\lambda) = -\sum_{k=1}^{\eta(\lambda)} p_k \log p_k
\]
Mesoscopic response function

- **Effective energy**

  \[ H_{\text{eff}}(\lambda) = \frac{H(\lambda) - H_{\text{min}}}{H_{\text{max}} - H_{\text{min}}} \]

- **Effective partition entropy**

  \[ S_{\text{eff}}(\lambda) = \frac{S(\lambda) - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \]

- **Effective number of communities**

  \[ \eta_{\text{eff}}(\lambda) = \frac{\eta(\lambda) - \eta_{\text{min}}}{\eta_{\text{max}} - \eta_{\text{min}}} \]

  \[ p_k = \frac{n_k}{N} \]

  \[ S(\lambda) = -\sum_{k=1}^{\eta(\lambda)} p_k \log p_k \]
Mesoscopic response function

Figure 1:
Mesoscopic response function

• Two challenges remain:
  • The range $\Lambda_{\text{min}} \leq \lambda \leq \Lambda_{\text{max}}$ is different for different networks
  • The quantities are strongly affected by very large lambda values

• Formulate in terms of the (effective) fraction of antiferromagnetic links

$$\xi = \xi(\lambda) = \frac{\ell^A(\lambda) - \ell^A(\Lambda_{\text{min}})}{\ell^A(\Lambda_{\text{max}}) - \ell^A(\Lambda_{\text{min}})}$$

• Here $\ell^A(\lambda)$ is the number of antiferromagnetic interactions in the network

• Now all quantities are normalized between 0 and 1

• Note that $\xi$ is a monotonically increasing function of $\lambda$

• Sweep $\xi$ from 0 to 1: $\eta(\xi = 0) = 1; \eta(\xi = 1) = N$
Plotting effective energy, partition entropy, and number of communities against $\xi$ produces a three-fold signature we call the mesoscopic response function.

$\xi$
Mesoscopic response function

• Comparing any two networks $i$ and $j$ amounts to comparing their MRFs:

$$d_{i,j}^H = \int_0^1 |H_{\text{eff}}(\lambda)^i - H_{\text{eff}}(\lambda)^j| d\xi$$

• Analogous measures defined for $d_{i,j}^S$ and $d_{i,j}^\eta$, resulting in $D^H$, $S^S$, $D^\eta$.
Data

- Obtained data for 752 networks including
  - Social (Facebook, collaboration, etc.)
  - Biological (protein interaction, metabolic, etc.)
  - Political (roll-call voting, bill co-sponsorship, etc.)
  - Technological (Internet, WWW, etc.)
  - Financial (NYSE)

- Removed self-edges (loops) and symmetrized directed networks
Data

- Studied networks
  - Come from several scientific fields
  - May represent temporal snapshots of time-dependent systems
  - May represent various observations of the same system using different observational techniques (e.g. in biology)
  - May be directly observable (e.g. WWW)
  - May be derived from suitable similarity measures (e.g. stock returns)
Taxonomy of 192 networks

- First principal component of H-S-eta space
- Similarity networks: Voting & co-sponsorship, House & Senate
- Anomalous networks: Social, language, collaboration, protein interaction

Figure 2:
Taxonomy of network classes

Figure 2:
Case studies

• Case studies (so far):
  • US House of Representatives voting patterns 1789 - 2008
  • US Senate voting patterns 1789 - 2008
  • Fungal growth (across species and time)
  • UN General Assembly voting patterns 1946 - 2008
  • Facebook networks for 100 US universities
  • NYSE return correlation structures 1985 - 2008
Case I: US House

- US Congress is the legislative branch of the US federal government
- House of Representatives is one of the two chambers (House & Senate)
- Analyze roll-call voting for the 1st - 110th Congresses (1789 - 2008)
- Define weighted connections between each pair of Representatives in terms of similarity of their voting dynamics (independently for each two-year Congress)
- Adjacency matrix: \[ A_{ij} = \left( \frac{1}{b_{ij}} \right) \sum_k \alpha_{ijk} \]
  \[ \alpha_{ijk} \] is unity if and only if \( i \) and \( j \) voted the same on bill \( k \)
  \( b_{ij} \) is the total number of bills on which both legislators voted
- Each network gives a snapshot of a two-year period, and we can use the method to study the evolution of voting blocs in time
Case I: US House

- Much research devoted to the extent of partisan polarization (influence of party)
- Popular measure of polarization is the DW-Nominate score (from 46th on)
- Congresses with similar levels of polarization usually appear together (colors)
Case I: US House

- Polarization as a function of time
- Solid line = DW-Nominate score (from 46th Congress onwards)
- Vertical bar is modularity, another measure for polarization, for all Congresses
Case I: US House

High polarization (brown):

- Congresses 5-7: Party politics became more important after the presidency passed from George Washington
- Congress 38: Occurred during the Civil War
Case I: US House

Intermediate polarization (green):

- Congresses 15 - 18: “Era of Good Feeling” (single party politics)
- Congresses 75 - 95: Period of party decline
Case I: US House

Low polarization (blue):

- Congresses 39-40: After the Civil War (reconstruction)
- Congresses 73-75: Beginning of the period of party decline
Case II: UN General Assembly

- The only one of the five principal UN organs in which all member nations have equal presentation
- Since UNGA has no enforcement power, it is considered symbolic by some
- Nevertheless, it is the only forum in which a large number of states meet and vote regularly on international issues

- Analyze roll-call voting for 1st to 63rd sessions, covering period 1946-2008
- Ties between countries determined based on the similarity of their roll-call voting in a single UNGA session (cf. House)
Case II: UN General Assembly

- Red cluster consists of all post Cold War assemblies except 1995
- The red cluster is flat, i.e. short distances between the years
Case III: Facebook

- Examine the LCCs of Facebook networks at 100 US universities
- One-time snapshot in September 2005
- Network sizes range from 762 nodes to 41,536 nodes
- Link density ranges from 0.2% to 6%
- In contrast to previous examples, here we compare multiple realizations of the same type of network in different geographic locations
Case III: Facebook

- Show are number of nodes (top bar) and link density (bottom bar)

- Main two branches are associated with networks with fewer nodes and high link density (left) and networks with more nodes and low link density (right)

- Corresponds to small universities with closely knit social networks and to large universities with sparser social networks, respectively
Case III: Facebook

- Shown are min, average, and max value of MRFs for 100 Facebook networks

- Despite very different sizes of the networks, and the fact that they have evolved independently of one another, all Facebook networks are very similar in terms of their mesoscopic organization
Case IV: NYSE

- New York Stock Exchange (NYSE) is the largest stock exchange in the world as measured by the aggregate US dollar value of the listed securities.
- Each node represents a stock.
- Link weight is proportional to the correlation between the daily logarithmic returns of the given pair of stocks.
- Networks for 100 stocks from the NYSE over the period 1985 - 2009.
- Correlations are calculated for half year time windows.
- Correlation matrices are converted to adjacency matrices according to

\[
A_{ij} = \frac{[\rho(i, j) - \min_{ij} \rho(i, j)]}{[\max_{ij} \rho(i, j) - \min_{ij} \rho(i, j)]} - \delta(i, j)
\]
Case IV: NYSE

• Two clear clusters emerge, where the red one seems to corresponds to periods of market turmoil

• The leafs highlighted in red broadly correspond to “crisis” periods.
  - Includes H2 of 1987 (Black Monday in October)
  - All of 2000-2002 (bursting of the dot-com bubble)
  - H2 of 2007 and all of 2008 (the recent credit crisis)
Case IV: NYSE

NYSE Composite Index measures the performance of all common NYSE stocks

- Shown is the volatility of the Index as a function of time
- Networks assigned to the red cluster correspond to periods of high volatility
Thanks!

For the slides, see jponnella.com