Generating artificial complex networks with community structure and realistic topology

James Gilbert

January 20th, 2014

Supervisors:
Andrzej Bargiela, Andrew Wood, Michael Holdsworth.
Aims

Background on complex networks

Background on community detection

Single community model
  - Compare to preferential attachment
  - Compare to empirical data - SeedNet

Multiple community model

Conclusions
Generate artificial networks with topological properties found in real-world biological networks:

- Heterogeneous degree (number of node connections) distribution
- Community or modular structure
- Clustering coefficients
- This will allow better testing of algorithms
Random Graphs

- Fixed probability of forming an edge \( P \)
- At what \( P \) does the graph form a single “giant component”?
- Node degree, approximately, forms a Poisson distribution.

Figure: A random graph 50 nodes, \( P = 0.05 \)
Real Networks aren’t random

**Figure**: EndoNet - Arabadopsis Thaliana Endosperm Co-expression Network [Dekkers et al. 2013]
Watts-Strogatz graphs
[Watts Duncan J. and Strogatz]
- Aims to model “degrees of separation”
- High Clustering coefficients
  - large number of triangles
- Fairly homogeneous degree distribution

Figure: Watts-Strogatz Graph
Barabasi-Albert Graphs
[Barabási and Albert 1999]
- Node degree roughly fits a power law.
- A small number of nodes have most of the connections
- Preferential attachment;
- Networks grow
- Probability for forming edges is proportional to node degree

Figure: A Barabasi-Albert Graph
What are communities?

- We know that real networks have modular structure - clusters not found in random graphs.
- Huge number of algorithms related to detecting community or modular structure in networks.
- No formal definition of what a “community” or “module” is.
- No well defined null models.
Modularity

Modularity Maximisation [Newman and Girvan 2004]

A combinatorial optimisation problem
- Find the set of sub-graphs that have a highest internal density
- Comparing to expected value of randomly connected nodes
- Glassy search space; large number of locally optimal solutions that are near the global optimum
- Resolution limit: often ignores small communities in favour of large ones
- Also gives us a measure \((Q)\) of how modular a network is
Stochastic Block Models

- Each node is assigned a block $K$.
- $K \times K$ Matrix with probabilities for connecting between groups.
- Shown to be a good fit for network community structure.
- However, degree distribution is a fixed condition [Karrer and Newman 2011]; not a result of some stochastic process.

$$M_B = \begin{bmatrix}
\lambda_1 & \lambda_2 & \lambda_3 \\
\lambda_2 & \lambda_4 & \lambda_5 \\
\lambda_3 & \lambda_5 & \lambda_6
\end{bmatrix}.$$  

**Figure**: A $3 \times 3$ block model.
Model Definition

- Fixed number of edges (density, $D$) and nodes ($N$)

$$D = \frac{2|E|}{|N|(|N| - 1)} \quad (1)$$

- Random, non-uniform probability for connecting nodes
- Function $F$ give node $i$ a position about the unit circle, $\theta_i$
- $R(\theta_i) = \alpha_i$ node scores.
  - $F$ and $R$ are wrapped Gaussian density functions.
- Connection probabilities $\beta$,

$$\beta_i = \frac{\alpha_i}{\sum_{j \in N} \alpha_j} \quad (2)$$

- We introduce a rewiring procedure - we want connected graphs.
Generating scale free networks

- $\theta_i$ drawn from $F$ wrapped Gaussian distribution with $\mu = 0, \sigma^2 = 1$
- $R$ wrapped Gaussian function with $\mu = \pi, \sigma^2 = 1$

**Figure**: Gaussian Functions

**Figure**: Observed degree distribution for BA model ($m = 4$)
Fitting Real Data

- Genes that correlate above a given threshold are given an edge
- 8454 nodes, 501522 Edges
- High Average clustering $C = 0.5$ and modularity $Q = 0.56$

**Figure**: Degree distributions - SeedNet (blue), our model (green), our model better fit $F(\sigma^2 = 1.32)$, $R(\sigma^2 = 0.82)$ (red)

- Modularity $Q$, and Clustering coefficient $C$ is far lower in our models ($Q < 0.2, C < 0.1$)
Introduce two new parameters
- $K$, number of communities
- $E_k$, fraction of edges to be assigned between communities

Nodes assigned position $\theta_i$ and rank $\alpha_i$ as before

Each of the $k \in K$, communities is treated independently

Communities are heterogeneous in size and density
With best fit model parameters:

**Figure**: $K = 4$. Low values of K seem to fit SeedNet (black line) quite well for degree and clustering; $E_k = 0.01$ (green), $E_k = 0.3$

**Figure**: $K = 40$. Higher values of $E_k$ are required to fit degree distribution as $K$ increases. $E_k = 0.01$ (blue). $E_k = 0.7$ (yellow)
Community Model - More Results

**Figure** : With fixed $K$, Modularity ($Q$) falls in response to $E_k$. $K = 5$ (blue) $K = 20$ (green), $k = 40$ (red).

**Figure** : With fixed $E_k$, Average Clustering ($C$) increases with $K$. $E_k$ slows the rate of increase. $E_k = 0.01$ (red), $E_k = 0.1$ (green), $E_k = 0.6$ (blue)
Communities can overlap

- It may not always be the case that one node exists in a single community.
- There are many “overlapping” community detection algorithms
  - Link Clustering [Ahn et al.2010]
  - Fuzzy c means clustering [Gregory2011]
- Introduce and overlap parameter $O_f$, fraction of nodes in multiple communities
- We test two models
  - Nodes are selected uniformly at random $M_U$
  - Selection is weighted by $\alpha_i$, $M_H$
Overlapping Communities

With fixed $E_k = 0.15$
$O_f = 0.01$ (Blue), $O_f = 0.1$ (green), $O_f = 0.3$ (red)

**Figure**: $M_U, K = 4$ random selection.

**Figure**: $M_H, K = 4$ Hubs selected. Difference doesn’t appear to be significant at low $K$
Overlapping communities - higher values of $K$

With fixed $E_k = 0.15$
$O_f = 0.01$ (Blue), $O_f = 0.1$ (green), $O_f = 0.3$ (red)

**Figure:** $M_U, K = 40$ random selection. High values of $K$ create problems.

**Figure:** $M_H, K = 40$ Hubs selected. Hub selection makes fitting even harder.
Overlapping Communities - clustering and modularity

Modularity is only a measure; the “correct” partition is meaningless for the overlapping definition.

**Figure**: $K = 40$. $M_U$ (red), $M_H$ (blue). Modularity and clustering (inset) response to $O_f$
Conclusions

- We generate graph topology with scale-free, small world properties
- Simple modification of Gaussian functions to fit degree distributions
- Communities are strongly linked to clustering coefficients
- Overlapping communities strongly effects modularity
- The null model for communities is the graph that can’t be divided further
- We do not attempt to explain how networks form
Future Work

- Fitting procedure for wrapped Gaussian functions
  - Not simple considering communities
  - Graph matching is an extremely difficult problem
- Provides an additional framework for detecting communities.
  - Select an appropriate algorithm for your specific network type.
- Directed and weighted graphs
- Hierarchical Structure
- Integrate into dynamical processes


Stochastic blockmodels and community structure in networks. 

Finding and evaluating community structure in networks. 
*Physical review E*, 69(2):026113.

Collective dynamics of ‘small-world’ networks. 
10.1038/30918.