A Weighted Spectrum Metric for Comparison of Internet Topologies

Dr. Damien Fay
Computer Laboratory
University of Cambridge

Andrew W. Moore
Computer Laboratory
University of Cambridge

Hamed Haddadi
University College London
London, United Kingdom

Steve Uhlig
Delft University of Technology
Delft, The Netherlands

Richard Mortier
Vipadia Ltd
Cambridge, United Kingdom

UNIVERSITY OF CAMBRIDGE

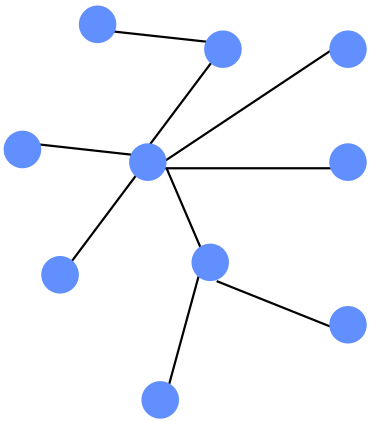
Presentation Outline

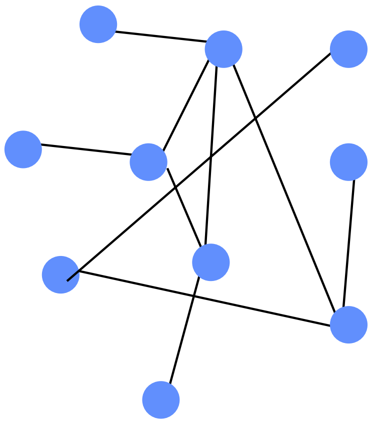
- Problem description
- Overview: the graph spectrum.
- Metric development.
- Results.
  - Parameter grids.
  - Optimised parameters.
- Conclusions.

UNIVERSITY OF CAMBRIDGE

Overall goal

- **How 'similar' are these two graphs/topologies?**
  - **How do we define similarity:**
    - **Link count,**
    - **Node degree distribution,**
    - **Centrality,**
    - **Edit distance.**

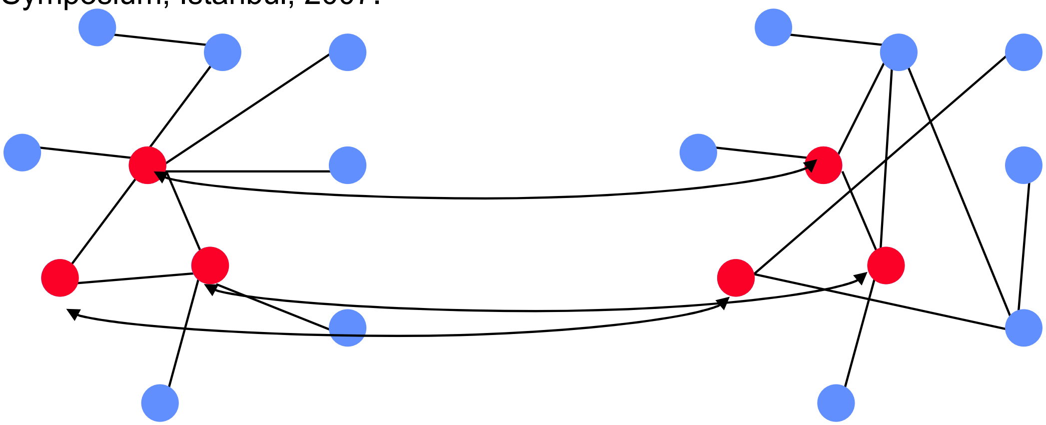




UNIVERSITY OF CAMBRIDGE

Related work

- **Cluster analysis**
  – A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems14. MIT Press, 2002
- **Graph matching**
  – B. Luo and E. Hancock. Structural graph matching using the em algorithm and singular value decomposition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(10):1120–1136, Oct 2001.
- **Topology tuning.**
  – S. Hanna. Representation and generation of plans using graph spectra. In 6th International Space Syntax Symposium, Istanbul, 2007.



Overall goal

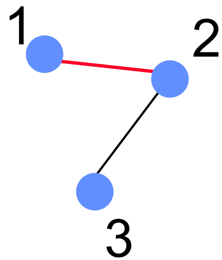
- The graphs we examine are different:
  - we seek to generate 'internet *like*' topologies not internet topologies.
- Determine how the distribution of clusters are different in two graphs.
- The exact location of clusters is not important.
- Given a metric this can then be used to determine optimum parameters of a topology generator.
- Validation: it is difficult to validate a metric as this would require a metric!
  - Does it perform as expected?

UNIVERSITY OF CAMBRIDGE

Graph Spectra.

- **The adjacency matrix:**

$$A(G) = \begin{cases} 1 & u, v \text{ are connected} \\ 0 & u, v \text{ not connected} \end{cases}$$



$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

  - **Symmetric. (bi-directional)**
  - **Fully connected.**

Graph Spectra.

- **Associated matrices:**

- **Computational Laplacian**

$$L(G) = D - A$$

- **Normalised Laplacian**

$$\Im(G) = D^{-1/2}\, L(G)\, D^{-1/2}$$

- **Walk Laplacian**

$$\Im_{rw}(G) = D^{-1} L(G)$$

Example

***A*(*G*)**

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

D

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

L*(*G*) = *D* - *A

$$\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

$$\mathfrak{I}(G) = D^{-1/2} L(G)\, D^{-1/2}$$

$$\begin{bmatrix} 1 & -1 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & 1 & 1 \end{bmatrix}$$

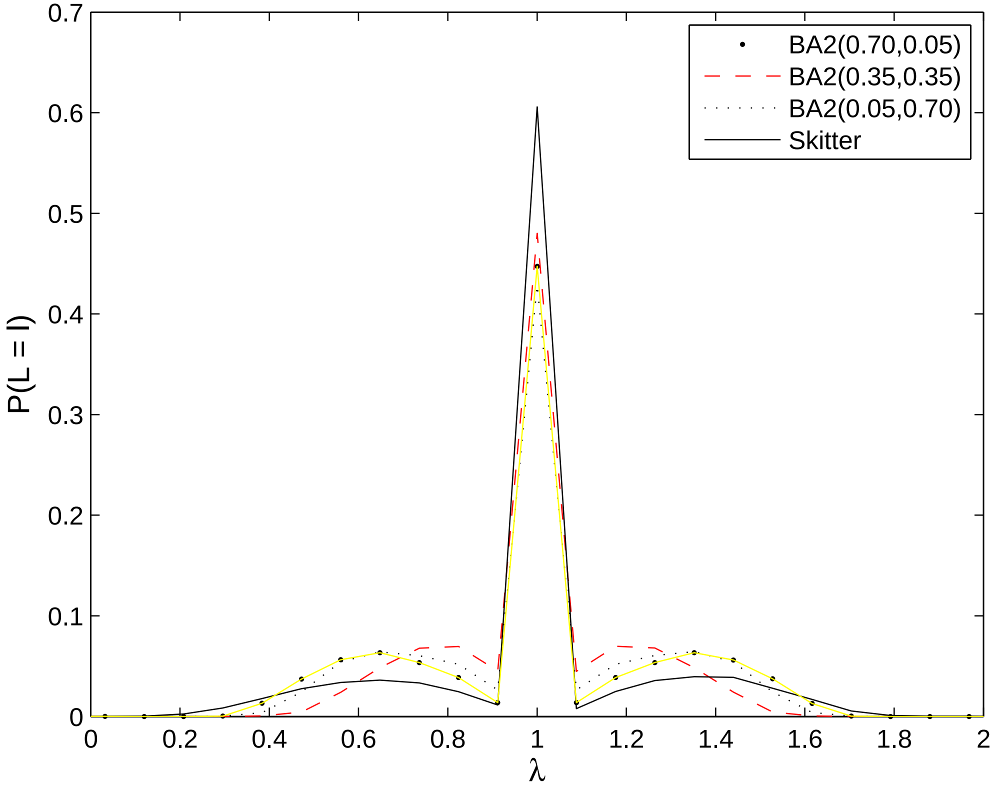
UNIVERSITY OF CAMBRIDGE

Graph spectrum definition.

- **The graph spectrum is simply the eigenpairs of the normalised laplacian.**
  - $\{\lambda_1, \lambda_2, \ldots, \lambda_{n\text{-}1}\}$ $\{\phi_1, \phi_2, \ldots, \phi_{n\text{-}1}\}$
  - **For a connected graph these are symmetrical around 1:**

$$\lambda_i = 1 - \lambda_{n-i}$$

  - **The set of eigenpairs forms an orthonormal basis for $\mathfrak{I}(G)$**
  - $\{\lambda_1, \lambda_2, \ldots, \lambda_{n\text{-}1}\} \in (0,2)$



0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(L = l)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
λ
BA2(0.70,0.05)
BA2(0.35,0.35)
BA2(0.05,0.70)
Skitter

Graph spectrum-interesting features.

Lemma 3. *Given a graph G, $\lambda_1 > 0$ if and only if G is connected. More generally, $\lambda_i = 0$ if and only if there are at least $i+1$ connected components of G.*

Lemma 4. *Given a graph G, $\lambda_{n-1} \leq 2$ and $\lambda_{n-1} = 2$ if and only if G has a bipartite component.*

Theorem 6. *Let G be a weighted graph and H a connected subgraph of G with $|V(H)| = t$. If*

$$\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1} \qquad \text{and} \qquad \theta_0 \leq \theta_1 \leq \cdots \leq \theta_{n-1}$$

are the eigenvalues of $\mathcal{L}(G)$ and $\mathcal{L}(G-H)$ respectively, then for $k = 0, 1, \ldots, n-1$ we have

$$\lambda_{k-t+1} \;\leq\; \theta_k \;\leq\; \begin{cases} \lambda_{k+t-1} & H \text{ is loopless and bipartite,} \\ \lambda_{k+t} & \text{otherwise,} \end{cases}$$

where $\lambda_{-t+1} = \cdots = \lambda_{-1} = 0$ and $\lambda_n = \cdots = \lambda_{n+t} = 2$.

Ref: Three lectures on spectral graph theory, Steve Butler: http://www.math.ucsd.edu/~sbutler/spectral/

Technical development.

- **We will examine 3 views of spectra:**
  - **Clustering analysis point of view.**
  - **Eigenvalue decomposition and statistical approximation.**
  - **Graph theoretic.**
- **Construct a metric for our purpose.**

Cluster analysis.

- **Taking the eigenpairs $\{\lambda_1, \lambda_2, \ldots, \lambda_{n-1}\}$ and $\{\phi_1, \phi_2, \ldots, \phi_{n-1}\}$, $\mathfrak{I}(G)$ can be expanded as:**

$$\mathfrak{I}(G) = \sum \lambda_i \phi_i \phi_i^{T}$$

- **In order to identify the clusters in the original topology we can simply thresehold ϕ_1:**

$$\begin{array}{cl} \left[\begin{array}{c} 0.3 \\ 0.2 \\ 0.1 \\ 0.8 \\ 0.1 \\ 0.9 \\ 0.8 \\ 0.6 \end{array}\right] & \begin{array}{l} \text{Cluster 1.} \\ \text{Cluster 1.} \\ \text{Cluster 1.} \\ \text{Cluster 2.} \\ \text{Cluster 1.} \\ \text{Cluster 2.} \\ \text{Cluster 2.} \\ \text{Outlier.} \end{array} \end{array}$$

Cluster analysis.

- **Note that the eigenvectors form a basis for $\mathfrak{I}(G)$ with λ_i as the coefficients of expansion:**

$$\mathfrak{I}(G) = \sum_i \lambda_i \phi_i \phi_i^T$$

- **Clustering can be improved by taking just the first *k* eigenvectors and process using k-means clustering:**

$$\begin{array}{cl} \begin{bmatrix} 0.3 & 0.7 & 0.7 \\ 0.2 & 0.2 & 0.2 \\ 0.1 & 0.2 & 0.2 \\ 0.8 & 0.7 & 0.7 \\ 0.1 & 0.2 & 0.2 \\ 0.9 & 0.7 & 0.2 \\ 0.8 & 0.2 & 0.7 \\ 0.6 & 0.7 & 0.2 \end{bmatrix} & \begin{array}{l} \text{Cluster 1-1.} \\ \text{Cluster 1-2.} \\ \text{Cluster 1-2.} \\ \text{Cluster 2-1.} \\ \text{Cluster 1-1.} \\ \text{Cluster 2-1.} \\ \text{Cluster 2-3.} \\ \text{Cluster 3.} \end{array} \end{array}$$

Eigenvalue decomposition.

- **Note that the eigenvectors form a basis for $\Im(G)$ with λ_i as the coefficients of expansion:**

$$\Im(G) = \sum_i \lambda_i \phi_i \phi_i^T$$

- **We could approximate $\Im(G)$ with:**

$$\hat{\Im}(G) = \lambda_1 \phi_1 \phi_1^T$$

- **The sum squared error in this approximation is:**

$$\sigma^2_{\hat{\Im}(G)} = \frac{C}{(1-\lambda_1)}$$

Conclusions so far.

- **The eigenvectors represent differing levels of clustering in a network.**
- **The eigenvalues represent the strength or *distribution* of those clusters.**

UNIVERSITY OF CAMBRIDGE

Graph theoretic viewpoint.

- For graph G, with subsets X_i and X_j, we have:

$$\min_{i \neq j} dist(X_i, X_j) = \max_{i \neq j} \left\lceil \frac{\ln \sqrt{\dfrac{vol\bar{X}_i vol\bar{X}_j}{volX_i volX_j}}}{\ln \dfrac{\lambda_{n-1} + \lambda_k}{\lambda_{n-1} - \lambda_k}} \right\rceil$$

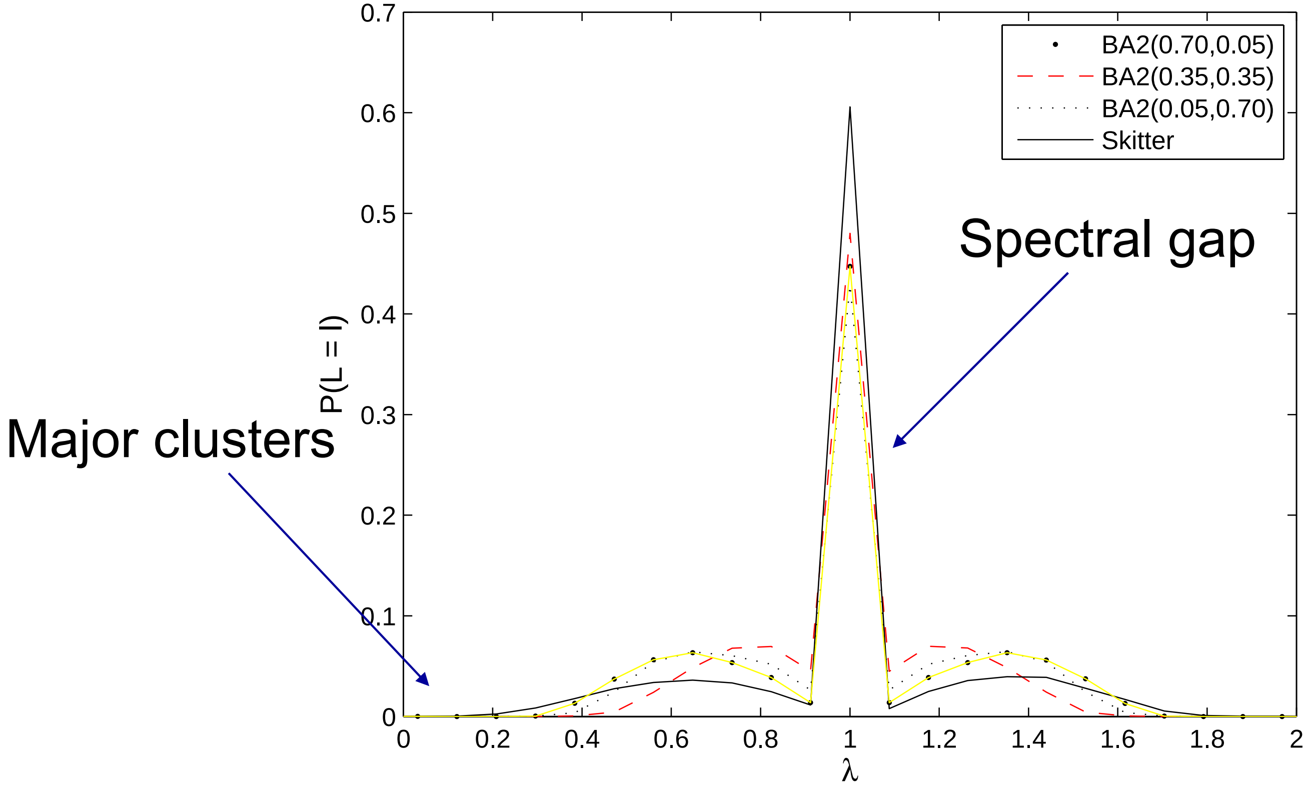
- where vol(X) is the total number of edges in X and dist(Xi,Xj) is the distance between subset i and j

- Equation (1) may be interpreted as representing the distance between subsets for *k* subsets *k* =1,...,n-1.

- In other words the eigenvalues may be used to estimate the number of subsets in a network without forcing the distances to be too short [11].

- **The *Distribution* of eigenvalues is central.**

UNIVERSITY OF CAMBRIDGE



UNIVERSITY OF CAMBRIDGE

Metric construction.

- Lets define the distribution of the eigenvalues as : f_λ
- Then a metric may be constructed based on f_λ in the normal way as:

$$J(G_1, G_2) = \int_\lambda \mu(\lambda)\left(f_{\lambda_1} - f_{\lambda_2}\right)^p d\lambda$$

- 3 Questions now arise:
  - What is the best value of p?
  - What is a suitable weighting function, $\mu(\lambda)$?
  - How do we calculate f_λ?

UNIVERSITY OF CAMBRIDGE

Metric construction.

- 3 Questions now arise:
  - What is the best value of p?
  - What is a suitable weighting function, $\mu(\lambda)$?
  - How do we calculate f_λ?

- We choose p=2 (sum squared error, other values might be of interest?)
- Again a graph theory result gives us the answer:

$$dev(G) \leq \sum_i (1-\lambda)^4 - 20\sqrt{irr G}$$

$$\mu(\lambda) = (1-\lambda)^4$$

- In this paper the distribution of eigenvalues is estimated by using pivoting and Sylveters law of inertia to compute the number of eigenvalues that fall in a given interval.

Finally the Metric.

$$J(G_1, G_2) = \sum_{k=1}^{M} (1-k)^4 \left(f_{\lambda_1}(k) - f_{\lambda_2}(k) \right)^2$$

- *M* is the number of bins used.
- The integral becomes a summation.
- $(1\text{-}k)^4$ weights eigenvalues furthest from the spectral gap most.

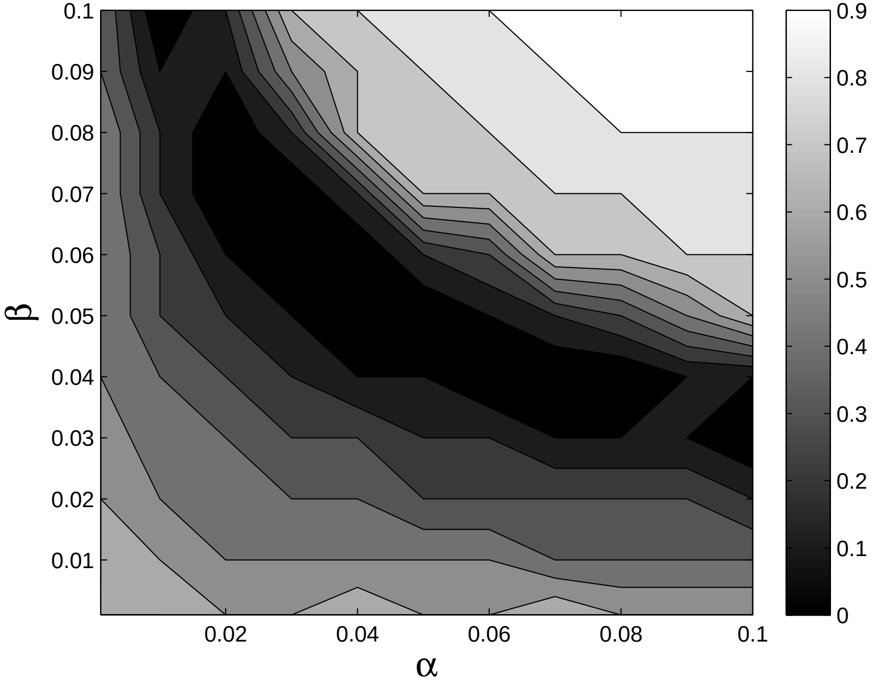
UNIVERSITY OF CAMBRIDGE

An application: Internet graphs.

- **We compare 5 topology generators:**
  - The waxman model
  - The 2nd Barabasi and Albert Model (BA2)
  - The Generalised Linear Preference model (GLP)
  - The INET model
  - Posative Feedback Preference model (PFP)
- **To one data set for the internet at AS level:**
  - The skitter data set.

UNIVERSITY OF CAMBRIDGE

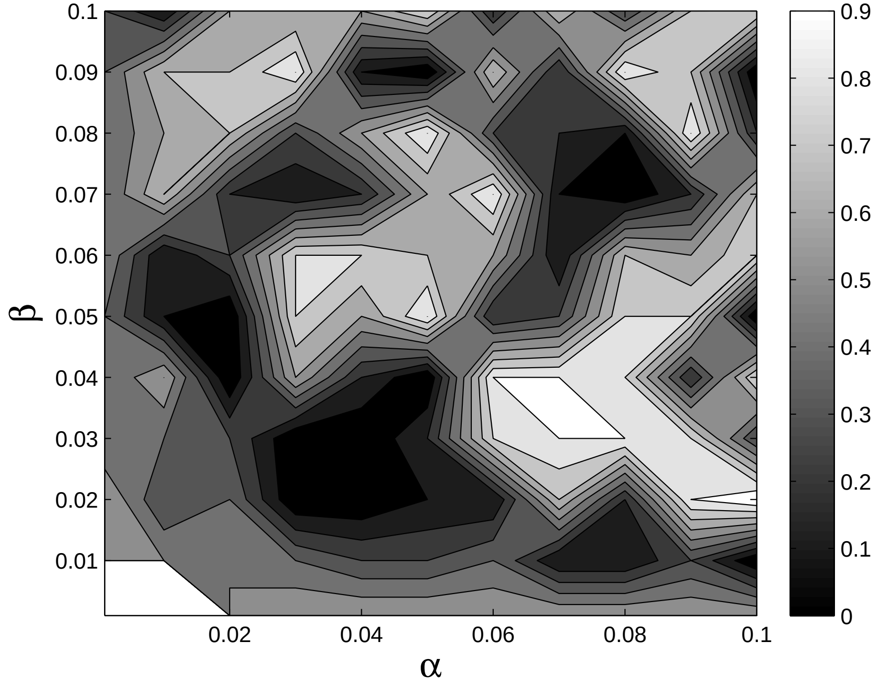
Parameter estimates:



Waxman model – distance with respect to links.

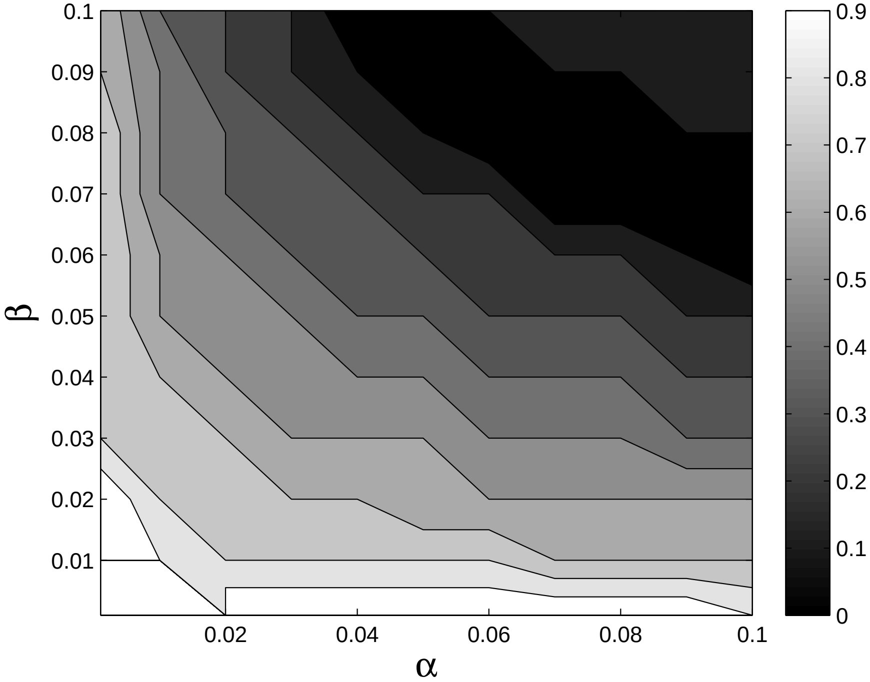
UNIVERSITY OF CAMBRIDGE

Parameter estimates:



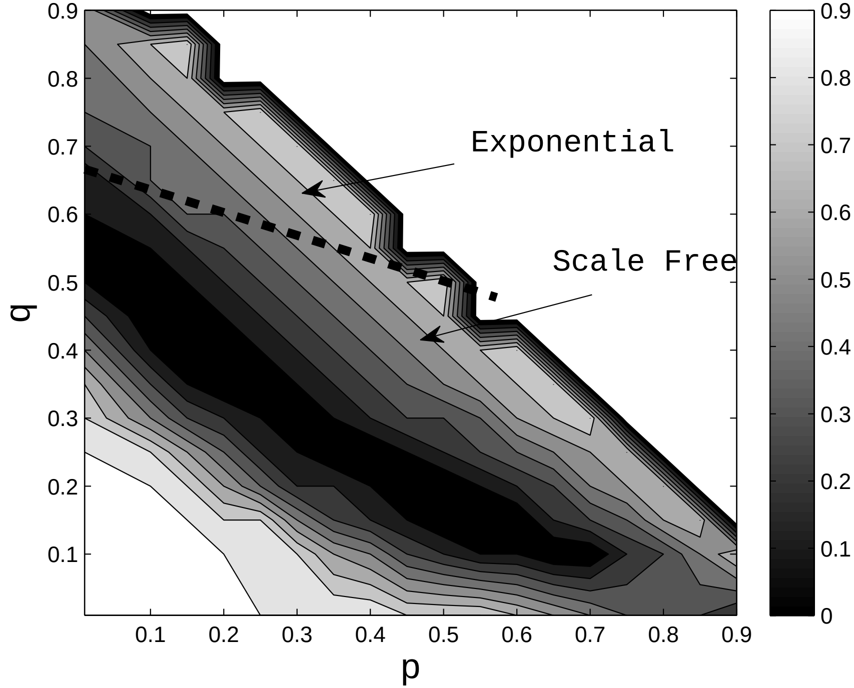
Waxman model – distance with respect to raw spectrum.

Parameter estimates:



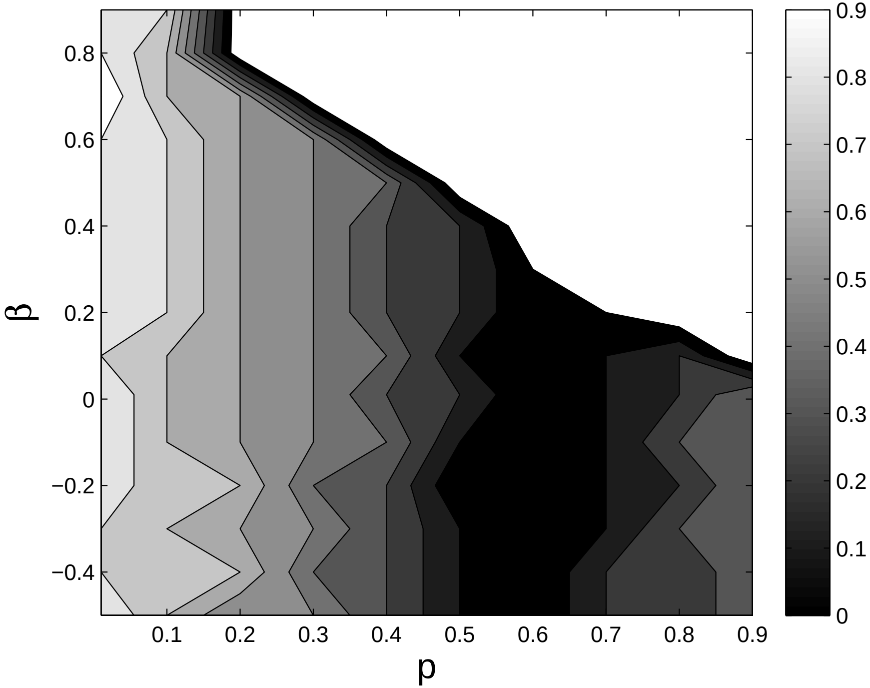
Waxman model – distance with respect to weighted spectrum.

Parameter estimates:



BA model – distance with respect to weighted spectrum.

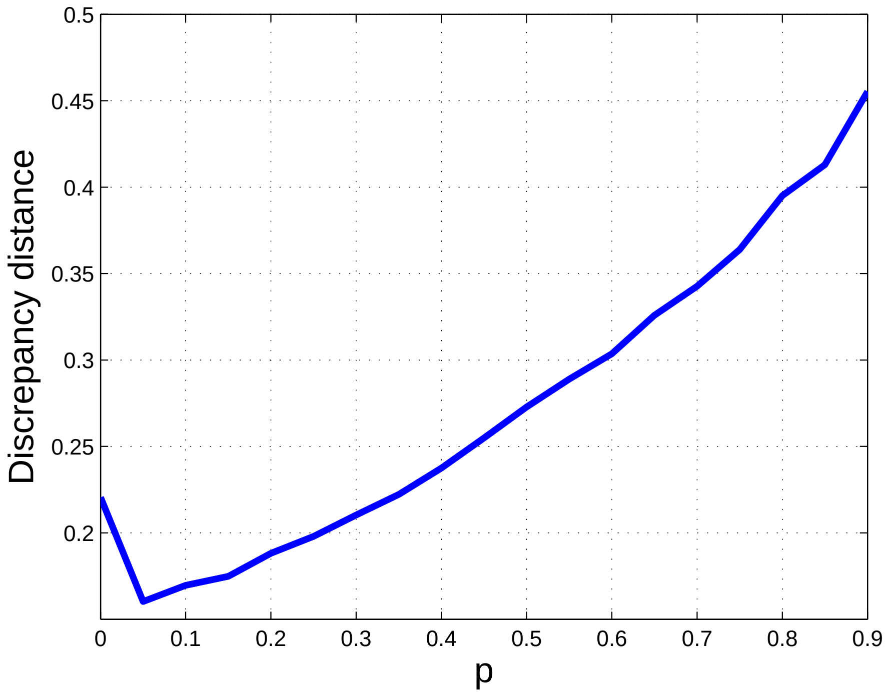
Parameter estimates:



GLP model – distance with respect to weighted spectrum.

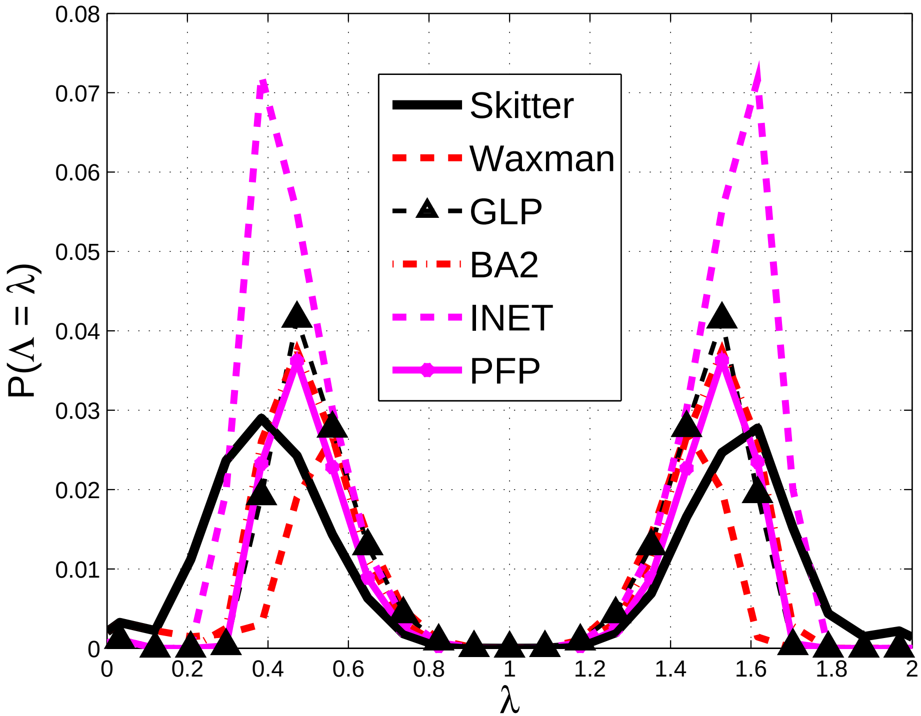
UNIVERSITY OF CAMBRIDGE

Parameter estimates:



INET model – distance with respect to weighted spectrum.

Parameter estimates:

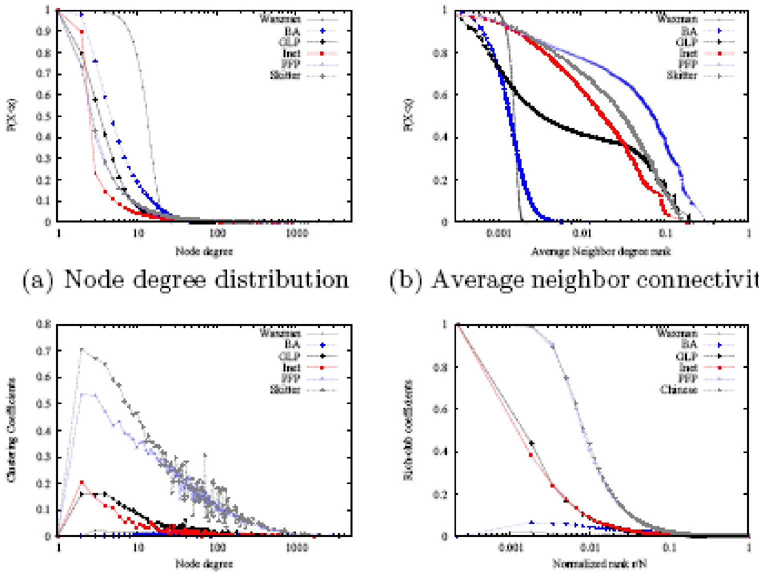


Weighted spectra.

UNIVERSITY OF CAMBRIDGE

Parameters and distances.

Waxman	α=0.08 (default = 0.15)	$\beta = 0.08(default - 0.2)$	$C_3(\theta)$ =0.0026	$\overline{C_3}(\theta)$=0.0797
BA	p=0.2865 (default=0.6)	$q = 0.3145(default = 0.3)$	$C_3(\theta)$ =0.0014	$\overline{C_3}(\theta)$= 0.0300
GLP	p=0.5972 (default = 0.45)	β=0.1004 (default=0.64)	$C_3(\theta)$ =0.0021	$\overline{C_3}(\theta)$=0.0446
Inet	α = 0.1013 (default= 0.3)	–	$C_3(\theta)$ =0.0064	$\overline{C_3}(\theta)$=0.0150
PFP	-	–	$C_3(\theta)$ = 0.0014	$\overline{C_3}(\theta)$=0.0371



(a) Node degree distribution (b) Average neighbor connectivity

(c) Clustering coefficients (d) Rich-Club connectivity

Fig. 7. Comparison of topology generators and Skitter topology

UNIVERSITY OF CAMBRIDGE

Conclusions

- The weighted spectrum appears to be a fair metric for comparing graphs.
- The metric allows identification of where in the structure two graphs differ.
- The metric behaves as expected and allows optimized parameters to be estimated.
- The raw spectrum is not sufficient to explain the structure in a topology.
- Future work: ?.
- Questions?

UNIVERSITY OF CAMBRIDGE