Entropy versus Heterogeneity for Graphs

Lin Han, Edwin R. Hancock and Richard C. Wilson
Department of Computer Science
The University of York
YO10 5DD, UK

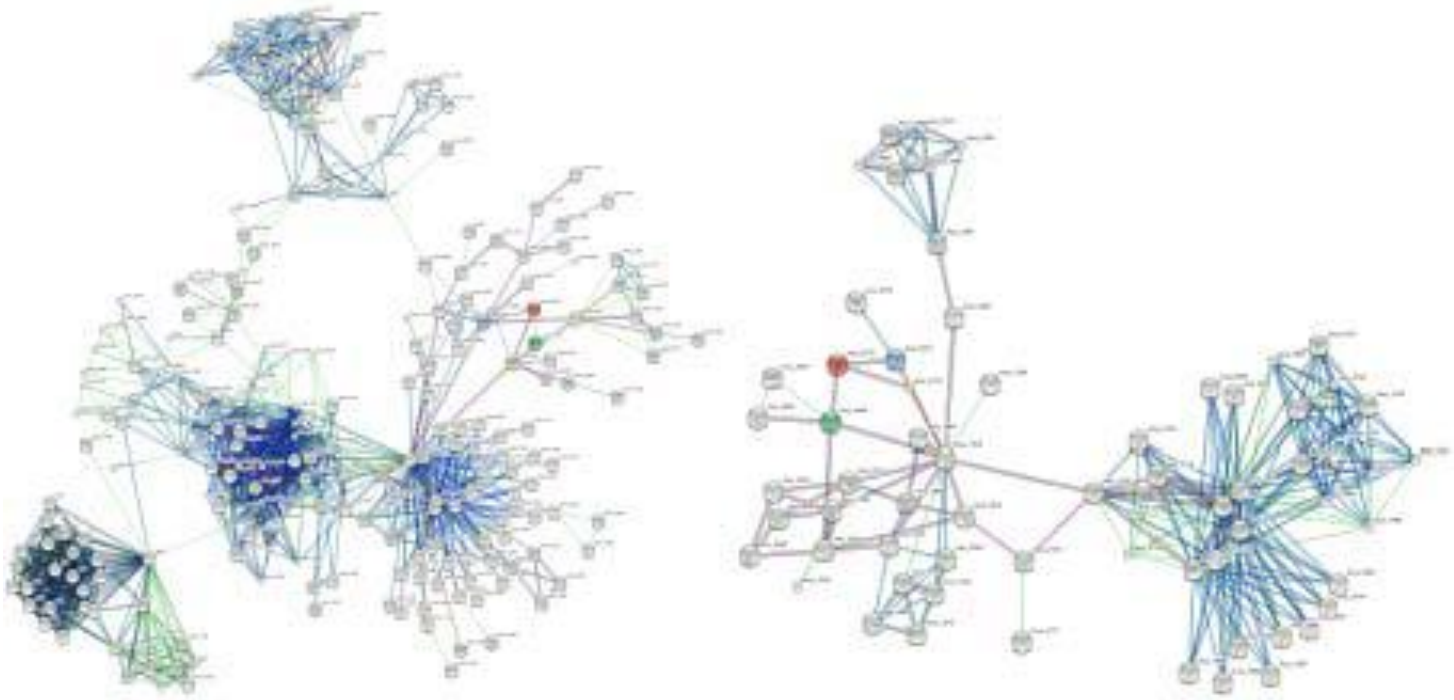

Figure: STRING Protein-Protein Interaction Networks

Characterising graphs

- Topological: e.g. average degree, degree distribution, edge-density, diameter, cycle frequencies etc.
- Spectral or algebraic: use eigenvalues of adjacency matrix or Laplacian, or equivalently the co-efficients of characteristic polynomial.
- Complexity: use information theoretic measures of structure (e.g. Shannon entropy).

Complex systems

- Spatial and topological indices: node degree stats; edge density;
- Communicability: communities, measures of centrality, separation, etc. (Baribasi, Watts and Strogatz, Estrada).
- Processes on graphs: Markov process, Ising models, random walks, searchability (Kleinberg).

Information theory

- Entropic measures of complexity: Shannon , Erdos-Renyi, Von-Neumann.
- Description length: fitting of models to data, entropy (model cost) tensioned against log-likelihood (goodness of fit).
- Kernels: Use entropy to computeJensen-Shannon divergence

Von Neumann Entropy

- Measured by the von Neumann entropy associated with the Laplacian eigenspectrum of graphs (Passerini, Severini, 2008)

$$H_{VN} = -\sum_{i=1}^{|V|} \frac{\hat{\lambda}_i}{2} \ln \frac{\hat{\lambda}_i}{2}$$

$$\hat{L} = D^{-1/2}(D - A)D^{-1/2} = \hat{\Phi}\hat{\Lambda}\hat{\Phi}^T$$

- Comes from quantum mechanics and is entropy associated with density matrix.

Approximation

- Approximate Shannon entropy by quadratic entropy

$$H_{VN} = \frac{1}{4}|V| - \sum_{(u,v)\in E} \frac{1}{4d_u d_v}$$

Homogeneity index

Based on degree statistics

$$\rho(G) = \sum_{(u,v)\in E} (d_u^{-1/2} - d_v^{-1/2})^2$$

$$\rho(G) = \frac{1}{|V| - 2\sqrt{|V| - 1}} \sum_{(u,v)\in E} \left\{ \frac{1}{d_u} + \frac{1}{d_v} - \frac{2}{\sqrt{d_u d_v}} \right\}$$

Homogeneity meaning

Limit of large degree

$$\rho(G) \sim \sum_{(u,v)\in E} \{CT(u,v) - 2A(u,v)\}$$

Largest when commute time differs from 2 due to large number of alternative connecting paths.