Clustering signatures classify directed networks

NetSci 08

June 26, 2008

Sebastian Ahnert

University of Cambridge

In collaboration with Thomas Fink, Institut Curie

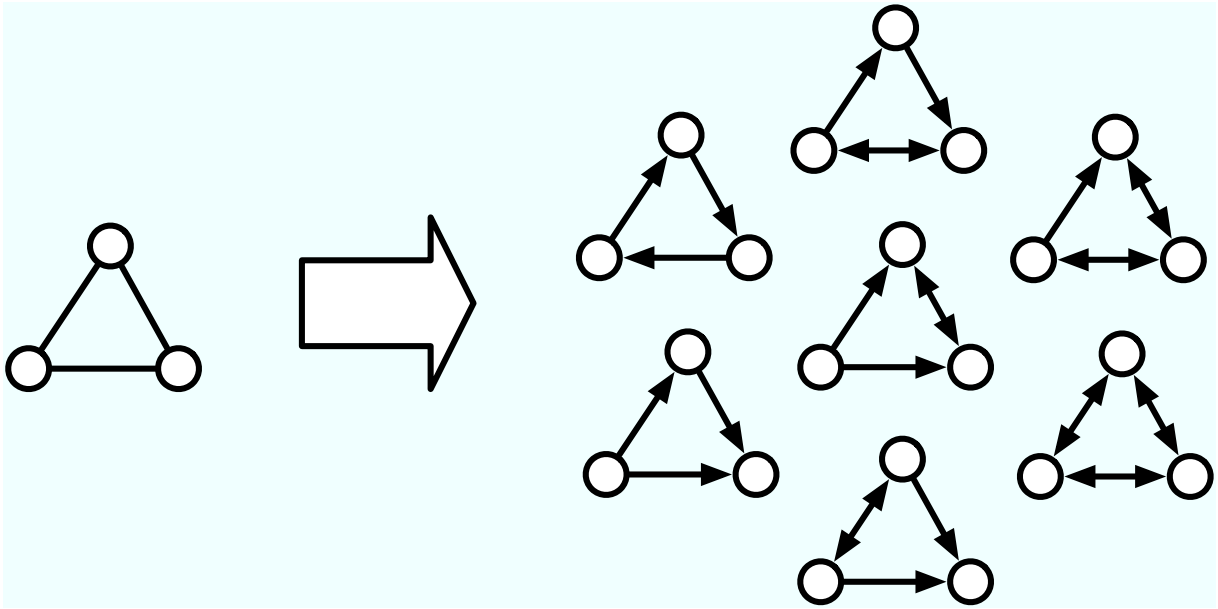
Directed networks

Majority of research done on undirected networks. Often direction is present but dropped.

Why?

Directed networks are much more complex.

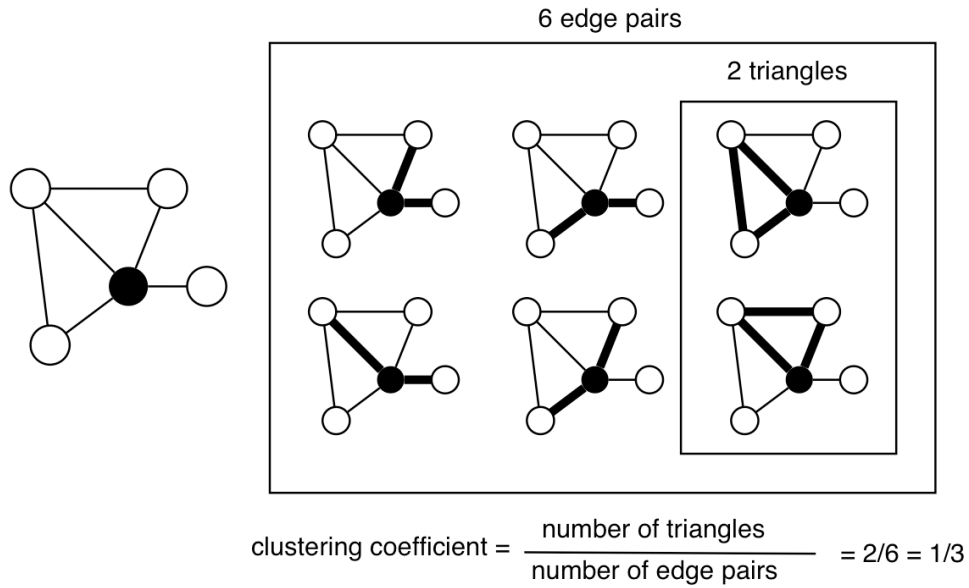
Directed networks



One type of triangle vs. seven types of triangle.

Clustering coefficient

Clustering coefficient measures density of connections in neighbourhood of node *i*.



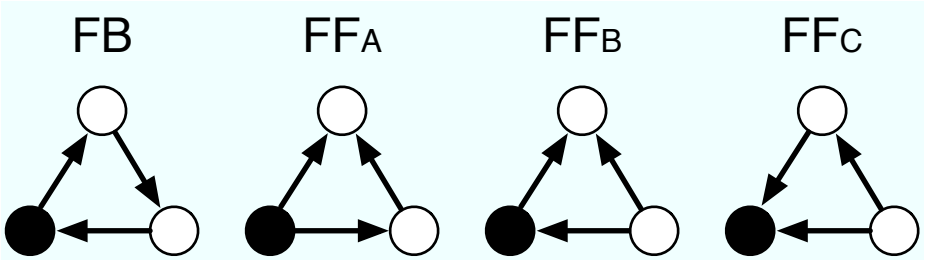
Clustering coefficient

In terms of the entries of the adjacency matrix we write:

$$c_i = \frac{\sum_{j,k} a_{ij} a_{jk} a_{ik}}{d_i(d_i - 1)/2} = \frac{\sum_{j,k} a_{ij} a_{jk} a_{ik}}{\sum_{j,k} a_{ij} a_{ik}}$$

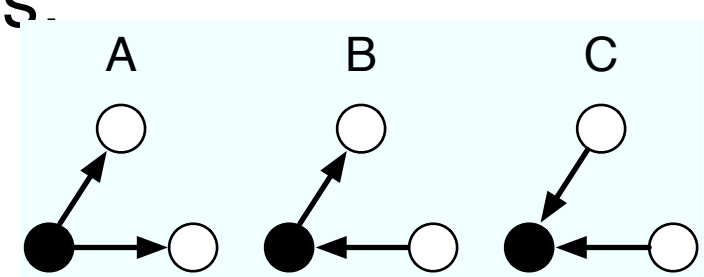
Directed clustering coefficient

In a directed network we can distinguish four basic scenarios for a given node to be located in a triangle:



Directed clustering coefficient

In a directed network we can distinguish three basic edge pair scenarios:



Directed clustering coefficient

Hence we can construct four clustering coefficients[1]:

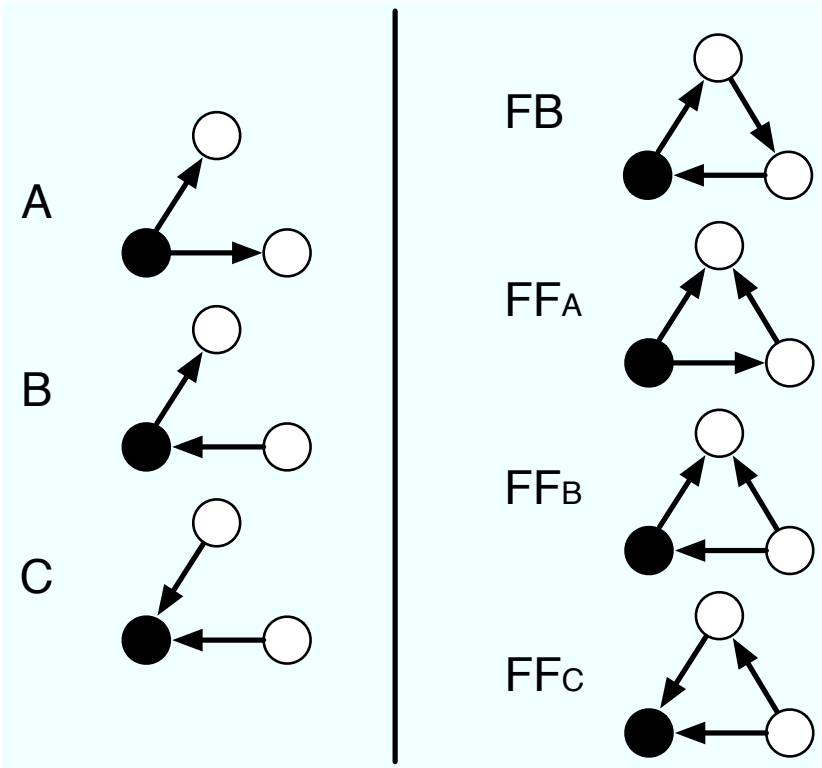
$$c_i^{FB} = N_{FB} / M_B$$

$$c_i^{FF_A} = N_{FF_A} / M_A$$

$$c_i^{FF_B} = N_{FF_B} / M_B$$

$$c_i^{FF_C} = N_{FF_C} / M_C$$

These we call the *clustering signature*.



1) see also Fagiolo, PRE 76, 026107 (2007).

Clustering signature

$$\mathbf{C}^{(i)} = \left(\frac{N_{\mathrm{FB}}^{(i)}}{M_{\mathrm{B}}^{(i)}}, \frac{N_{\mathrm{FF_A}}^{(i)}}{M_{\mathrm{A}}^{(i)}}, \frac{N_{\mathrm{FF_B}}^{(i)}}{M_{\mathrm{B}}^{(i)}}, \frac{N_{\mathrm{FF_C}}^{(i)}}{M_{\mathrm{C}}^{(i)}} \right)$$

where

$$M_{\mathrm{A}}^{(i)} = \sum_{j,k} a_{ij} a_{ik} \quad ; \; M_{\mathrm{B}}^{(i)} = \sum_{j,k} a_{ij} a_{ki} \quad ; \; M_{\mathrm{C}}^{(i)} = \sum_{j,k} a_{ji} a_{ki}$$

$$\begin{aligned} N_{\mathrm{FB}}^{(i)} &= \sum_{j,k} a_{ij} a_{jk} a_{ki} \qquad N_{\mathrm{FF_A}}^{(i)} = \sum_{j,k} a_{ij} a_{kj} a_{ik} \\ N_{\mathrm{FF_B}}^{(i)} &= \sum_{j,k} a_{ij} a_{kj} a_{ki} \qquad N_{\mathrm{FF_C}}^{(i)} = \sum_{j,k} a_{ji} a_{kj} a_{ki} \end{aligned} \tag{1}$$

Clustering signature

To visualize this quantity we normalize the signature such that:

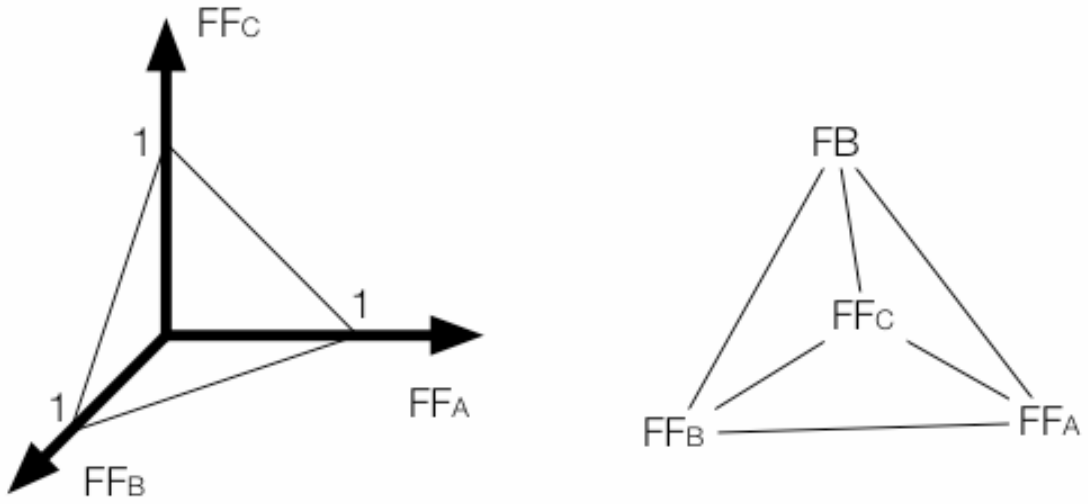
$$\tilde{\mathbf{C}}^{(i)} = \frac{\mathbf{C}^{(i)}}{T^{(i)}}$$

where

$$T^{(i)} = \frac{N_{\mathrm{FB}}^{(i)}}{M_{\mathrm{B}}^{(i)}} + \frac{N_{\mathrm{FF_A}}^{(i)}}{M_{\mathrm{A}}^{(i)}} + \frac{N_{\mathrm{FF_B}}^{(i)}}{M_{\mathrm{B}}^{(i)}} + \frac{N_{\mathrm{FF_C}}^{(i)}}{M_{\mathrm{C}}^{(i)}}$$

Clustering signature tetrahedron

The normalization means that we can plot the clustering signatures in a tetrahedron by using FF_A, FF_B and FF_C, with FB given implicitly:



Average clustering signature

We can average the normalized clustering signature over all nodes in the network:

$$\tilde{\mathbf{C}} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{C}}^{(i)}$$

This tells us about the average local connectivity properties of a directed network.

Application to real networks

We calculate the average normalized clustering signatures for 16 directed real-world networks:

3 food webs

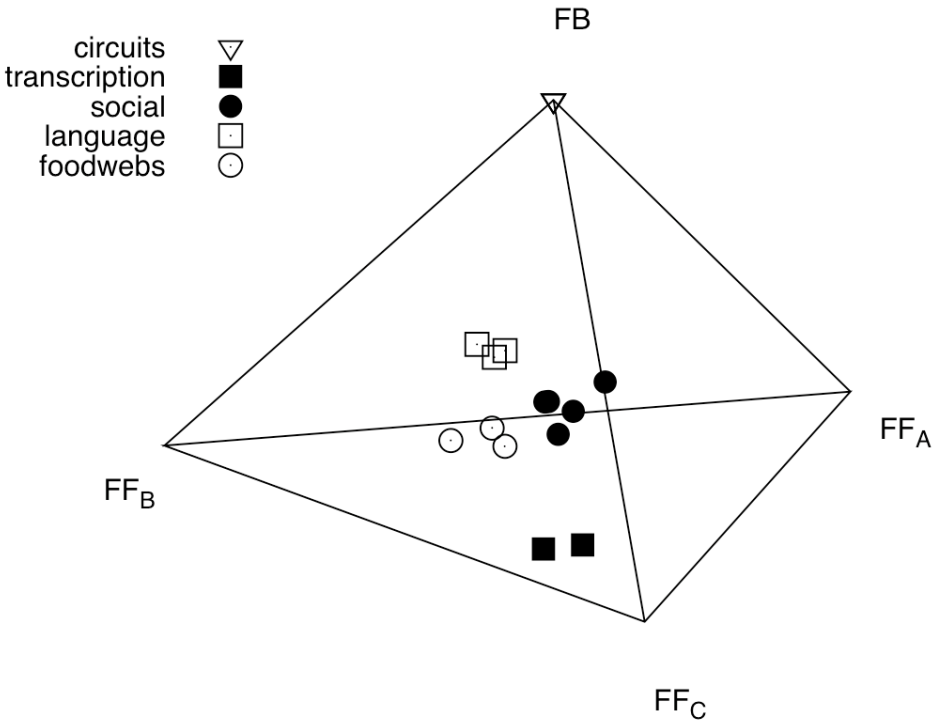
2 transcription networks

3 language networks

5 social networks

3 electrical circuits

Application to real networks

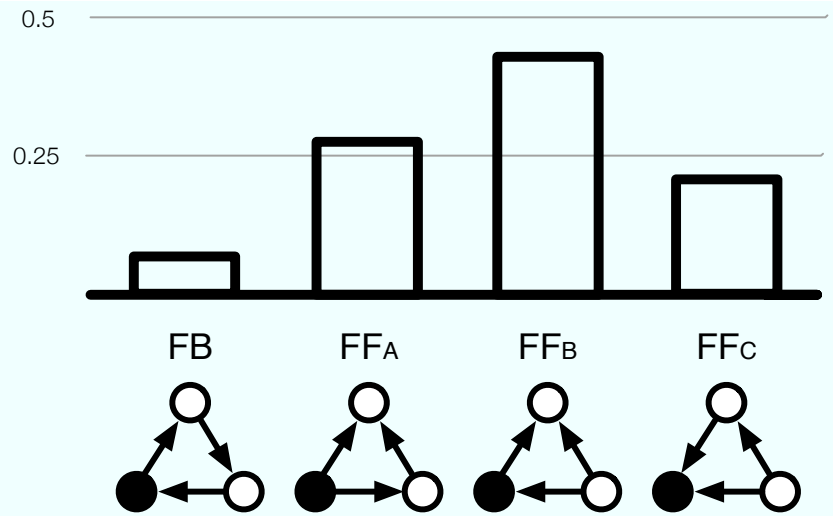


K-means clustering:

Correct classification with 89% of variance explained.

Food webs

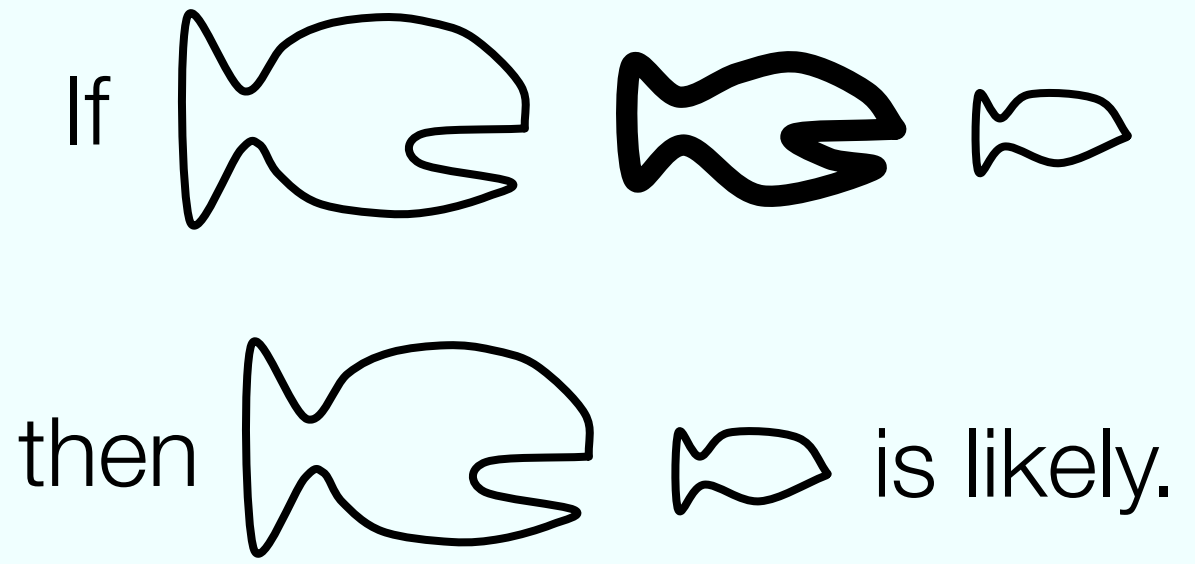
Three food webs: N = 39 to 128, E = 177 to 2106



FF_B is dominant component. What does this mean?

Food webs

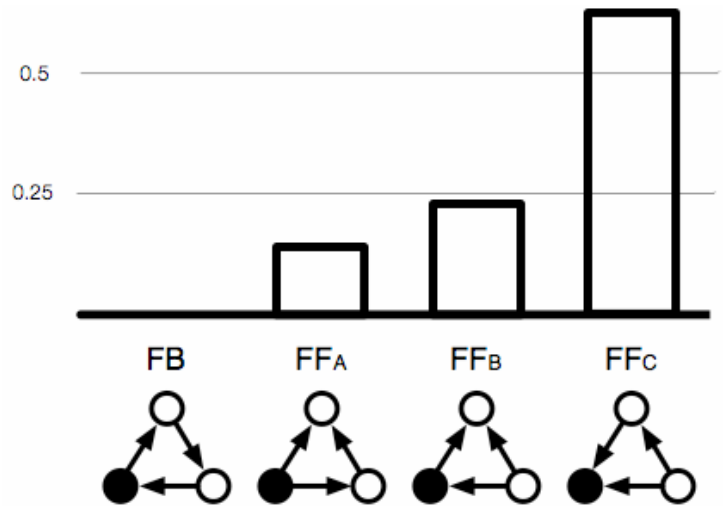
It means:



Transcription networks

Two transcription networks:

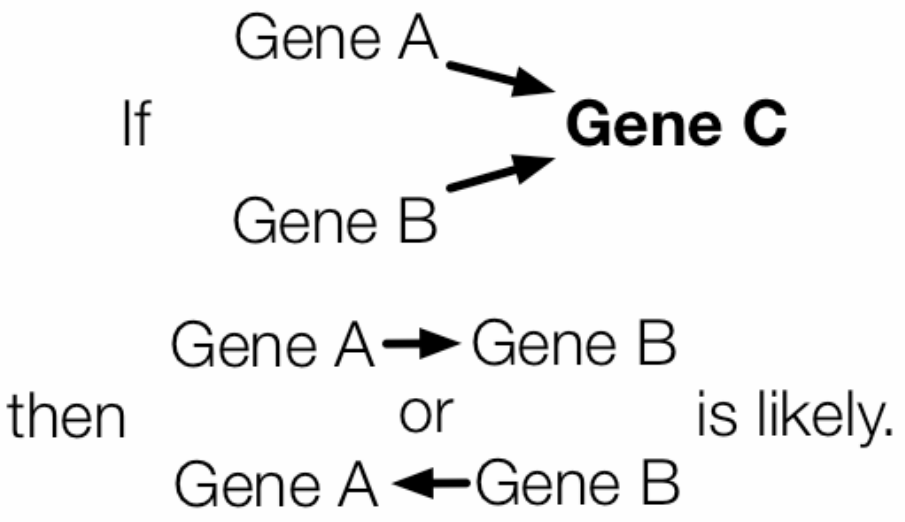
N = 423, 688 and E = 519, 1079



Strongly dominant FF_C component.

Transcription networks

Transcription networks: Feed-forward loops, small in-degrees, large out-degrees.

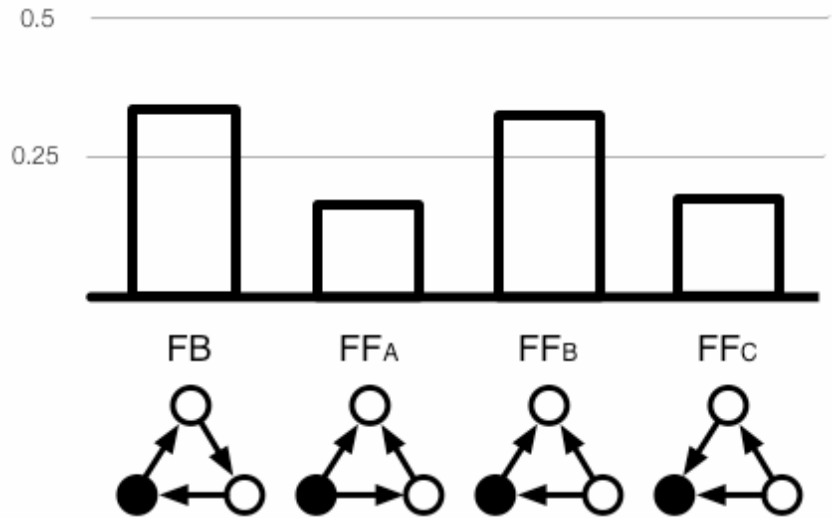


Language networks

English, French and Japanese word adjacency.

N = 3177 to 9424

E = 8300 to 46281



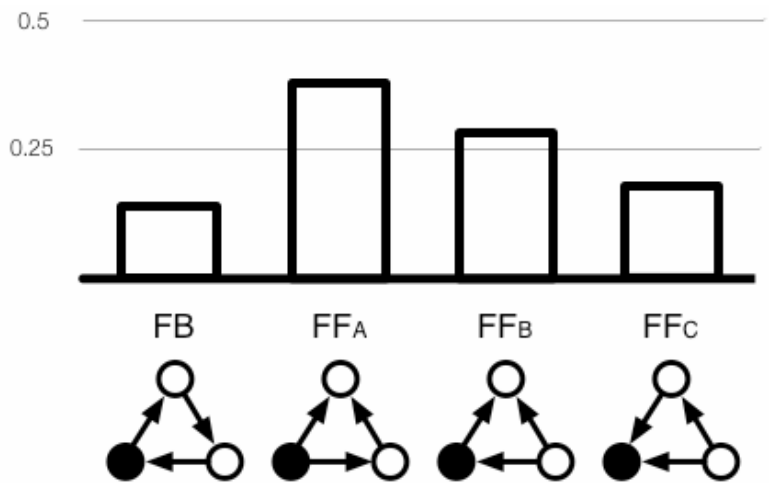
FF_A and FF_C suppressed, as words fall into categories (nouns, verbs, adjectives) which are unlikely to be adjacent.

Social networks

Five social networks: prisoners, team partner selection, factory workers (2), political weblogs.

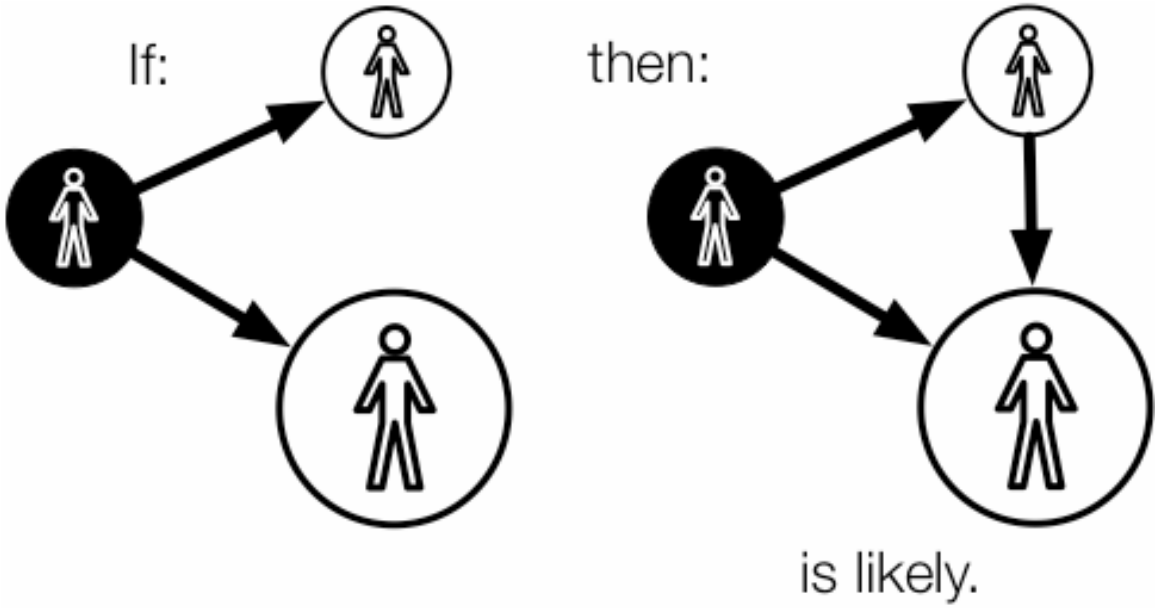
N = 32 to 1491

E = 96 to 19090



Social networks

Enhanced FF_A component, because:



Social networks

Note: Reverse all edges = swap FF_A and FF_C

We have to define directions of edges carefully.

In social networks this is more difficult than in most other networks.

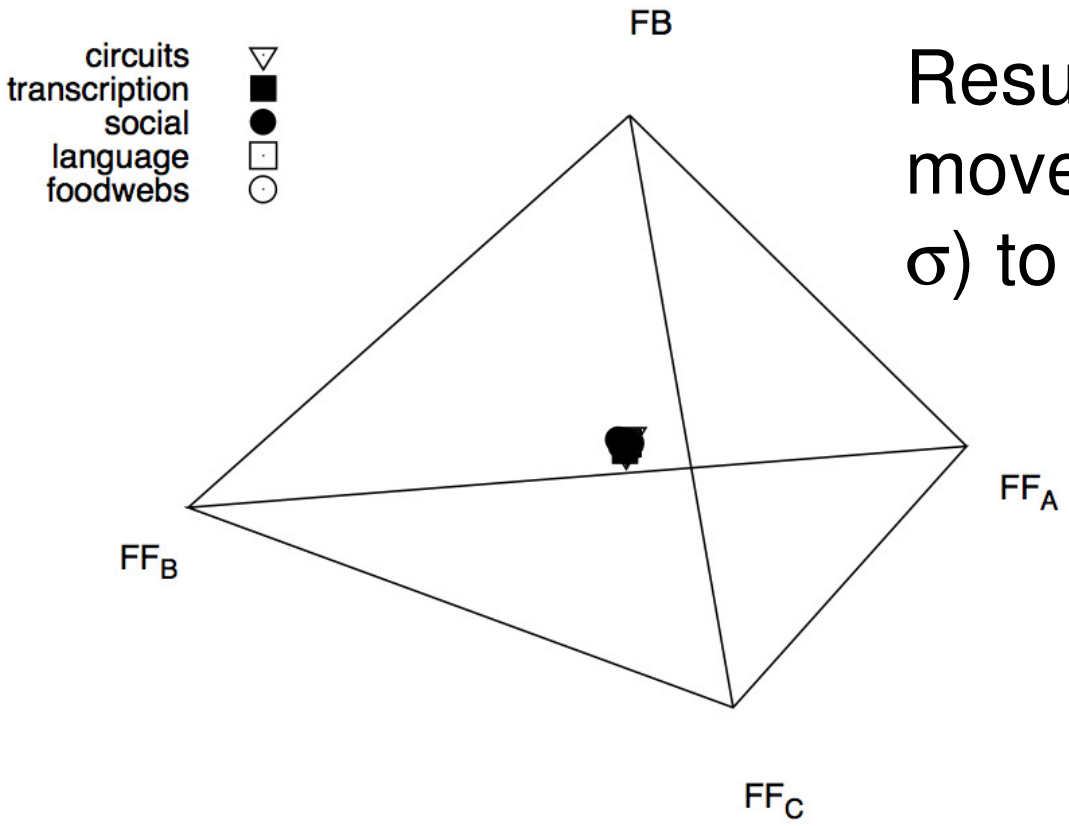
We take direction as an **imbalance** or **transfer** of benefits, e.g. A names B as a friend.

Directionality or density?

Is the clustering signature really measuring local connectivity? Perhaps these networks differ in size or density - could this be responsible for the observed distribution?

Answer: Randomize directions of existing edges and see what happens:

Randomized edge directions



Result: All networks move close (within one σ) to the neutral point.

Weighted networks

This approach can be easily generalized to weighted networks by using the ensemble approach[1], which uses a map from weights to probabilities.

[1] Ahnert et al., Phys. Rev. E 76, 016101 (2007).

Conclusions

Clustering signatures are an effective way of classifying directed networks according to their local connectivity.

The clustering signature tetrahedron offers a compact visualization of this classification.

Acknowledgments

This work was done in collaboration with Thomas Fink of the Institut Curie.

It was supported by:

- The Leverhulme Trust, UK
- Defense Advanced Research Projects Agency (DARPA)