On a New Signature that quantifies Topological Structure in Biological and Economic networks

Razvan Valentin Marinescu

Department of Computing
Imperial College London

23 June 2014

Neighbourhood comparison using the clustering coefficient

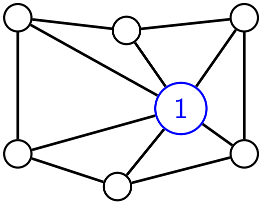


Figure: Graph 1

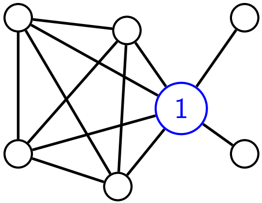


Figure: Graph 2

Neighbourhood comparison using the clustering coefficient

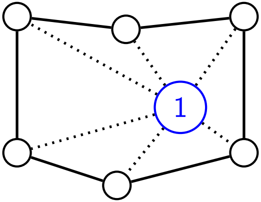


Figure: Clustering coefficient: 0.4

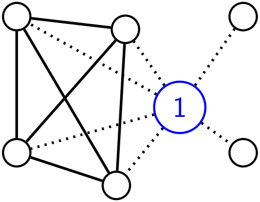
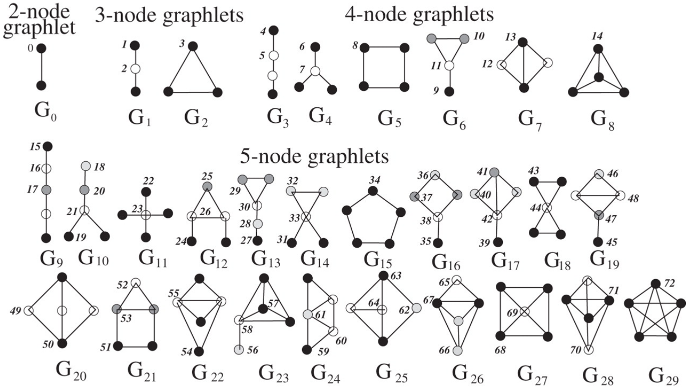


Figure: Clustering coefficient: 0.4

- Problem: the clustering coefficient cannot distinguish between the two graphs . . .
- Solution: generalise the clustering coefficient!

Clustering coefficient generalisation

- **Graphlet Cluster Vector** (GCV) - generalises the clustering coefficient.
- $GCV = \{F_1, F_2, \ldots, F_{29}\}$



Comparison using the GCV

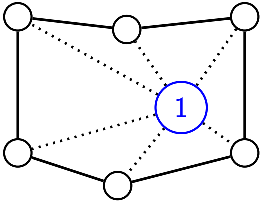


Figure:
$GCV = \{6, 0, 6, 0, 0, 0, 0, 0, \ldots\}$

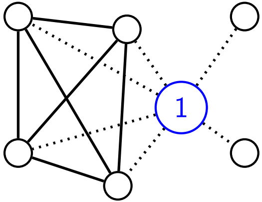


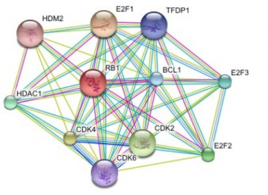
Figure:
$GCV = \{0, 4, 0, 0, 0, 0, 0, 1, \ldots\}$

Presentation outline

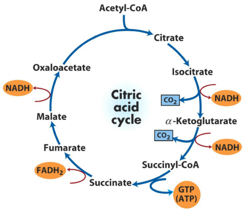
- GCV Implementation
- Applications

World Trade networks

Protein Interaction networks



Metabolic networks



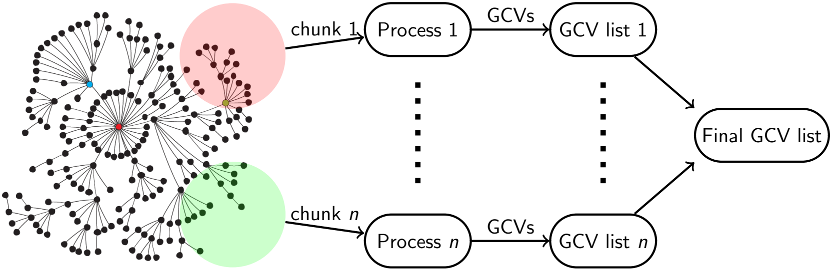
- Evaluation on random graphs clustering

Implementation

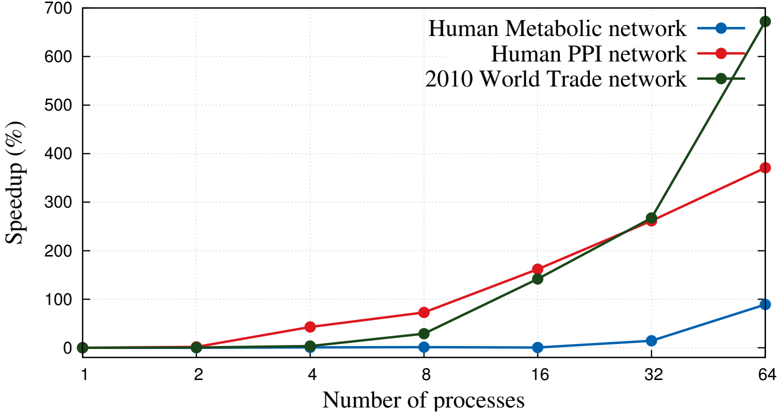
- programming language: `C++`
- we leveraged code (`ncount.cpp`) that was computing a similar signature, called the GDV
- graph was represented by a complex data structure containing both:
  - an adjacency matrix
  - an adjacency list

Parallelisation

- allowed us to run the GCV computation on large networks such as the PPI networks (11,000 nodes)



Parallelisation – speedup

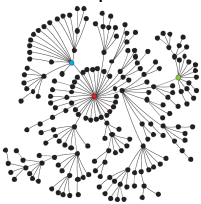


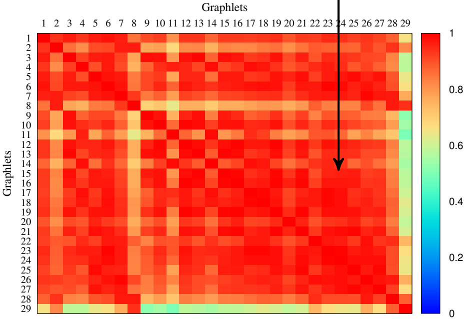
Pearson's GCV correlation matrices - computation

Graphlets	GCV (node 1)	GCV (node 2)	GCV (node 3)	GCV (node 4)	...
G1	2	3	6	9	...
G2	1	10	23	0	...
G3	0	3	5	14	...
G4	4	9	6	2	...
...	...	...	...	...	...
G29	1	14	6	0	...

$\rho(i,j)$ = Pearson's correlation between row vectors i and j

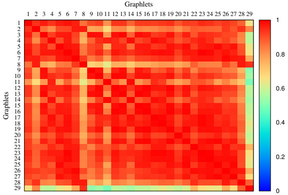
Graphlet Cluster Vectors (GCVs)



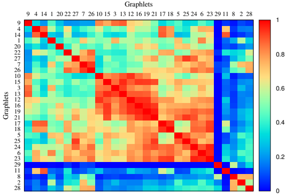


Correlation Matrix Normalisation Process

Initial matrix

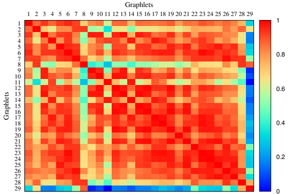


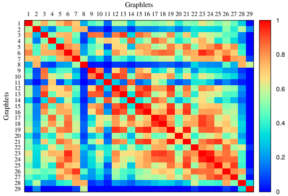
Final matrix



feature scaling:
$x' = \frac{x - min}{max - min}$

hierarchical clustering
(complete linkage)





4^{th} degree polynomial scaling:
$x' = x^4$

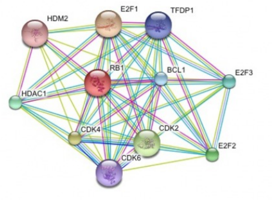
Presentation outline

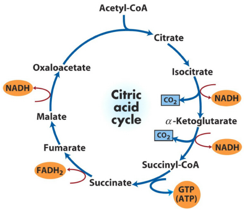
- GCV Implementation
- Applications

World Trade networks

Protein Interaction networks

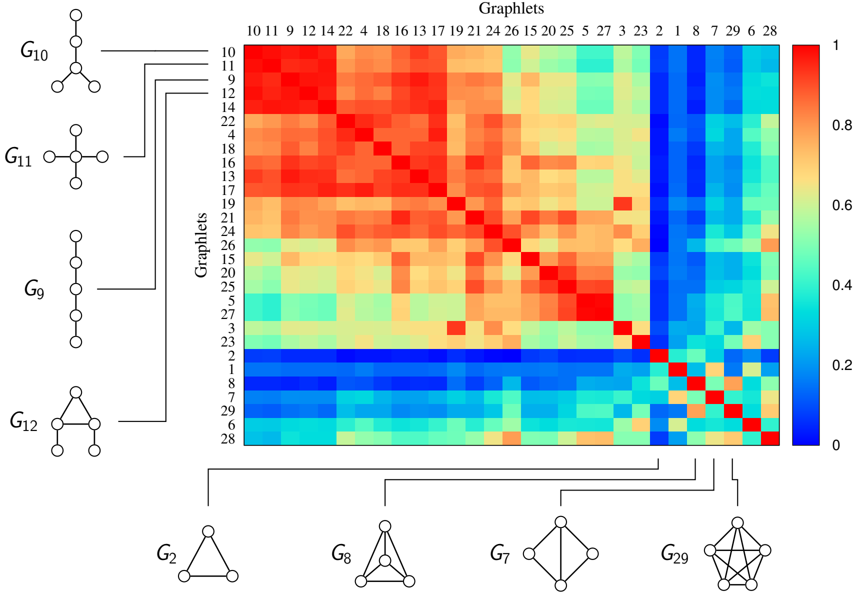
Metabolic networks



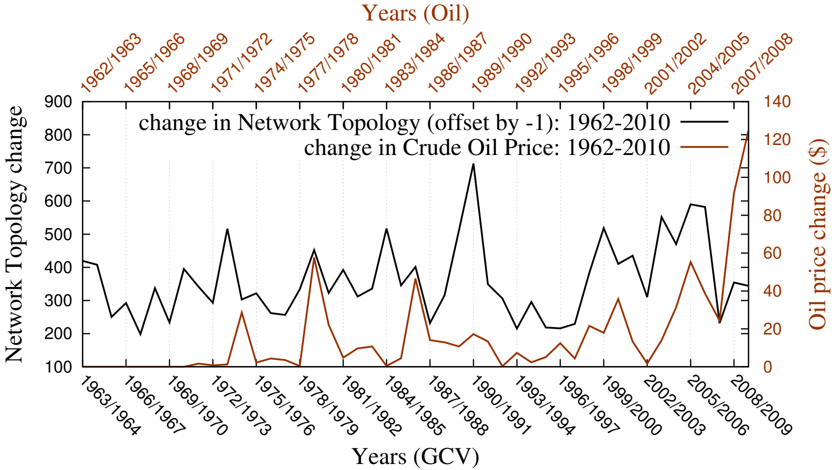


- Evaluation on random graphs clustering

Pearson's GCV correlation matrix - World Trade network

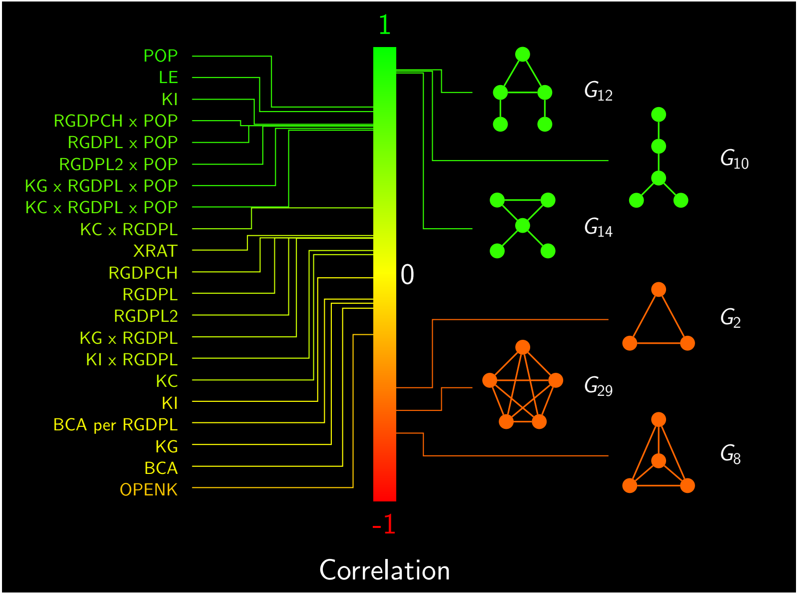


Correlation matrix change during 1962–2010



- Spearman's correlation: 0.34
- p-value: 0.01
- offset: -1 $\Longrightarrow$ the network topology causes the changes in oil price!

Canonical Correlation Analysis (CCA)



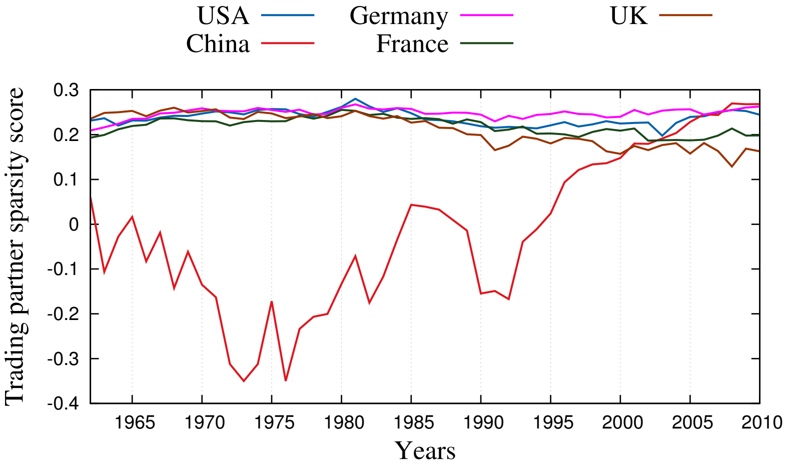
Trading partners sparsity index

- Measures how sparse the neighbourhood of a node is.

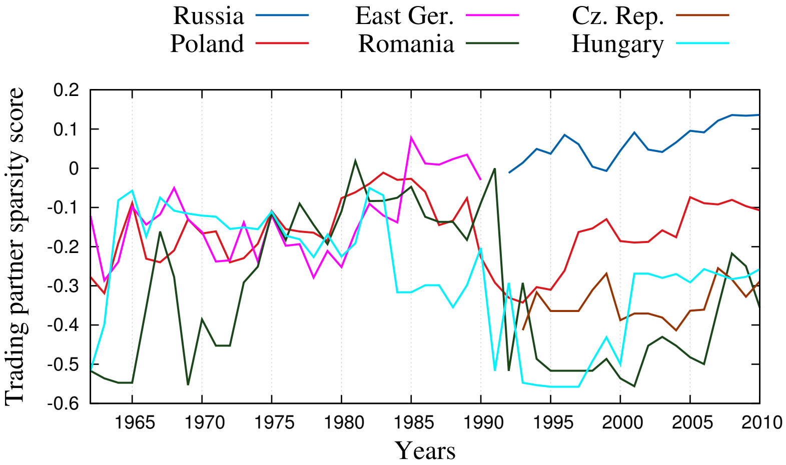
- $T = \underbrace{w_{12} F_{12} + w_{10} F_{10} + w_{14} F_{14}}_{\text{sparse graphlets}} + \underbrace{w_8 F_8 + w_{29} F_{29} + w_2 F_2}_{\text{dense graphlets}}$

- can have both positive or negative values

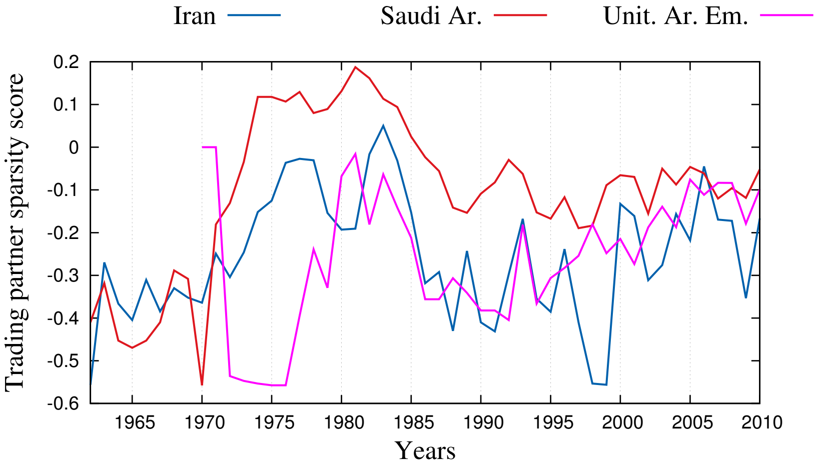
Trading partners sparsity index – G20



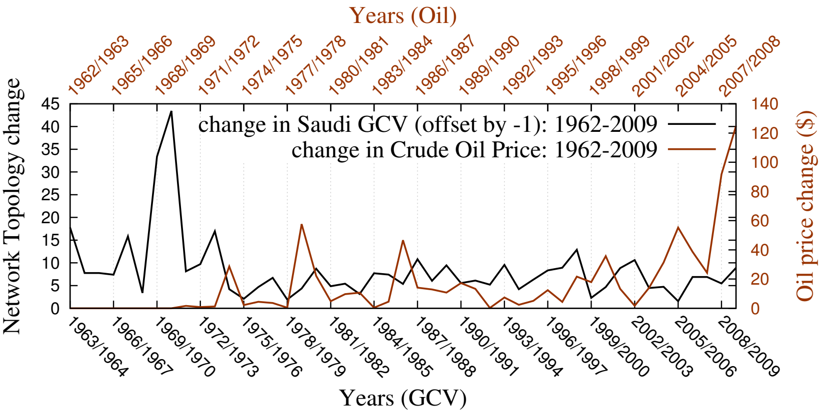
Trading partners sparsity index – Eastern Europe



Trading partners sparsity index – OPEC



Case study – Saudi Arabia



- Spearman's correlation: -0.32
- p-value: 0.026
- offset: -1

Results for other networks

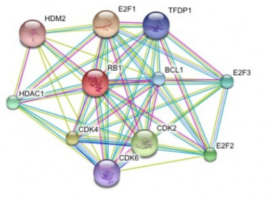
- PPI networks – key protein functions:
    - Ribosome translation
    - RNA processing
    - Metabolism
    - Golgi endosome vacuole sorting
- Metabolic networks – main results in:
    - Cellular Processes
    - Organismal Systems
    - Human diseases

Presentation outline

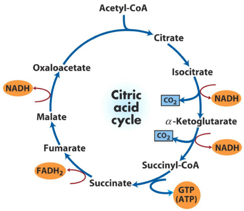
- GCV Implementation
- Applications

World Trade networks

Protein Interaction networks



Metabolic networks



- Evaluation on random graphs clustering

Evaluation on random graphs clustering

We evaluated the GCV signature at classifying 5 random graphs:

- Erdős-Rényi (ER)
- Erdős-Rényi (with preserved degree distribution) (ER-DD)
- Geometric networks (GEO)
- Scale-free Barabási-Albert – preferential attachment (SF)
- Stickiness index-based (STICKY)

Other signatures evaluated

We compared the performance of GCV against 5 other signatures:

- Degree Distribution
- Average clustering coefficient
- Spectral distribution
- Graphlet Frequency Vector (GFV) – RGFD distance
- Graphlet Distribution Vector (GDV) – GCD73 distance

Multi-dimensional scaling

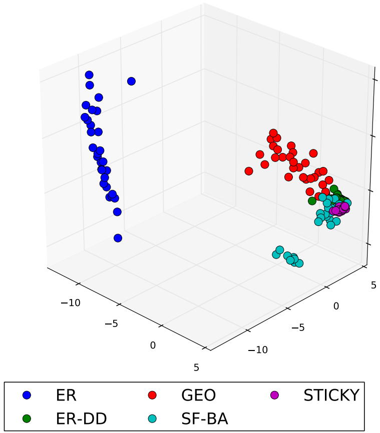


Figure: GCV MDS

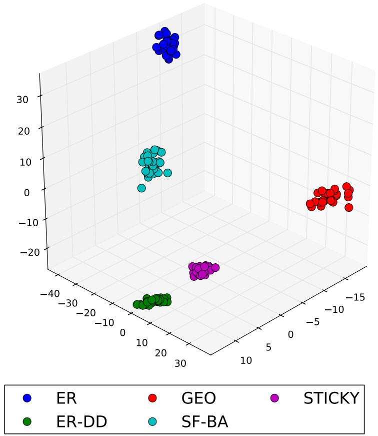
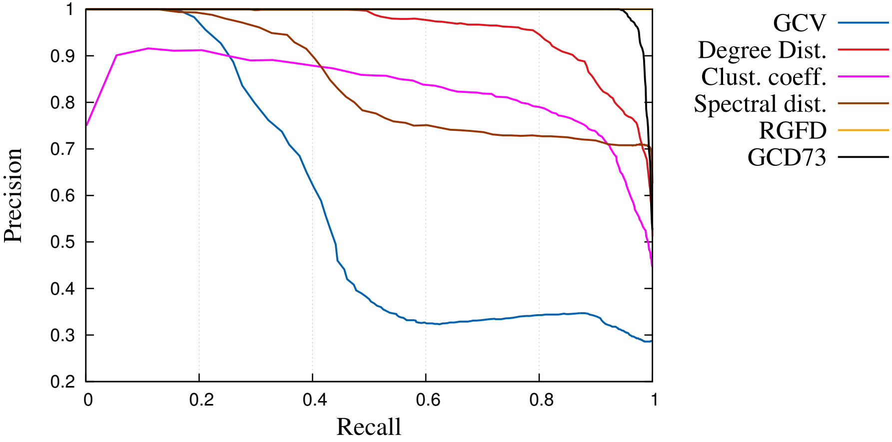
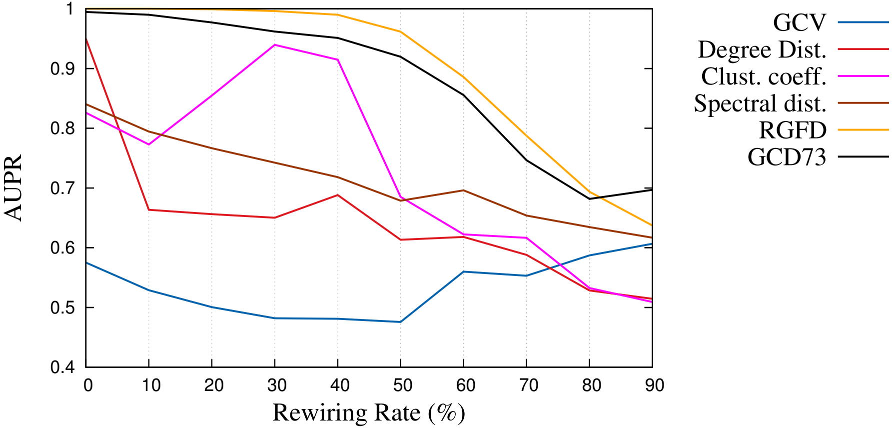


Figure: RGFD MDS

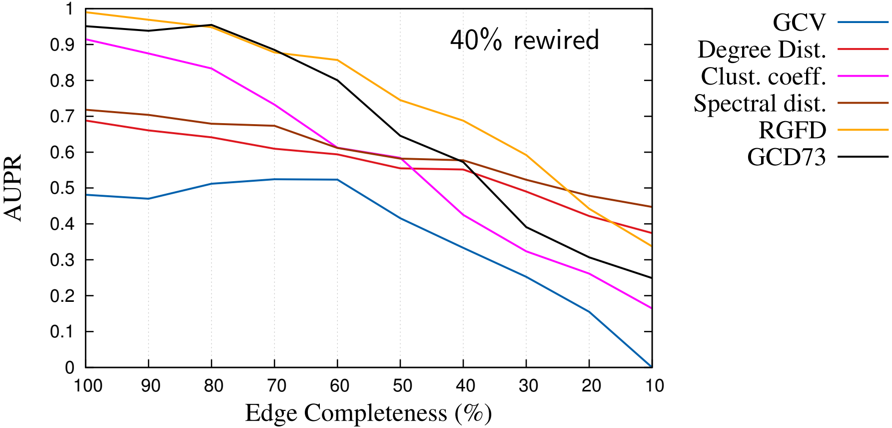
Precision-Recall Curve Analysis



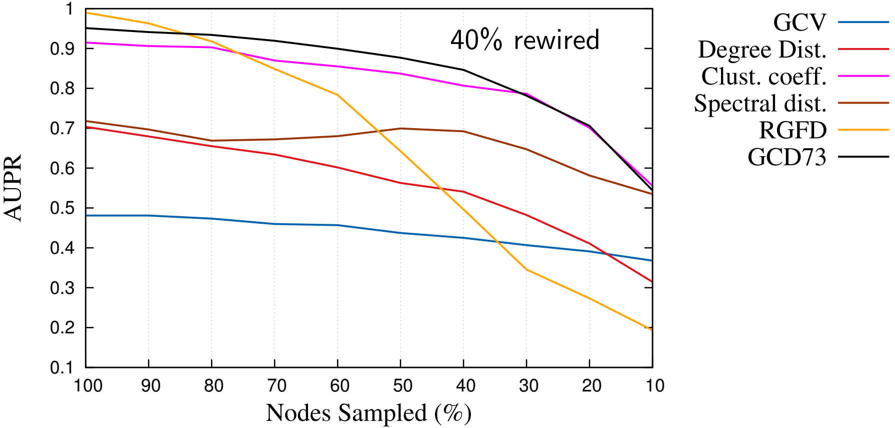
Robustness testing – Noisy data



Robustness testing – Noisy and incomplete data



Signature approximation

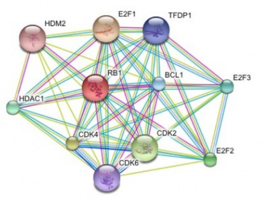


Conclusion

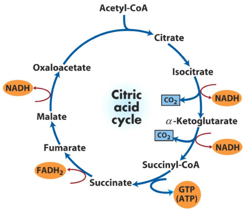
- GCV Implementation
- Applications

World Trade networks

Protein Interaction networks



Metabolic networks



- Evaluation on random graphs clustering

Future work

- Research better normalisation procedures
- Redundancy analysis
- Supporting experiments for our current results
- Apply the signature to other networks
  - Social networks
  - Telecommunication networks
  - Gene Regulatory networks
  - Neuronal networks

Questions?

- Thank you very much!

GCV normalisation

- normalised GCV:

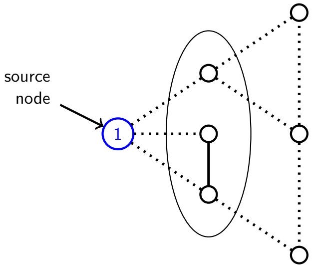
$$GCV(n) = \left(F_n^1, F_n^2, ...F_n^{29}\right)$$

  where

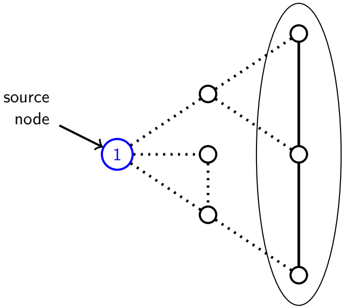
$$F_n^i = \frac{S_n^i}{\sum_{i=1}^n S_n^i}$$

- all the World Trade network results presented here use the normalised GCV

Limitations of the GCV signature



(a) Shell-1 neighbourhood



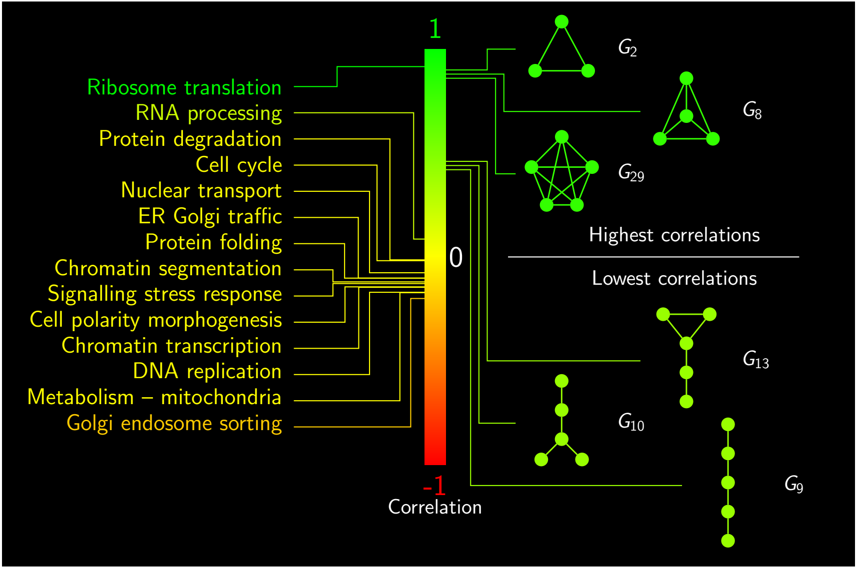
(b) Shell-2 neighbourhood

- the GCV cannot capture information in nodes that are at a distance of 2 or larger away from the source node
- some of the GCV frequencies might be redundant
- computation of the GCV is slow on very dense networks such as the full WTN

Signatures evaluated – definitions

- Spectral distribution: eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ of the Laplacian matrix $L = D - A$ where:
  - D: diagonal degree matrix of G
  - A: adjacency matrix
- Graphlet Frequency vector: $GFV(G) = (F_0(G), F_1(G), ...F_{29}(G))$ where:
  - $F_i(G) = -\log\left(\frac{G_i}{\sum_{i=1}^{n} G_i}\right)$
  - G_i is the total number of graphlets of type i in G
- Graphlet Distribution Vector of node x: a vector $(F_1, F_2, \ldots, F_{72})$, where:
  - f_i measures the number of graphlets that touch node x at automorphism orbit i.

Protein interaction network CCA



Pearson's and Spearman's correlation coefficients

- Pearson's correlation coefficient between X and Y:
  - $\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[X - \mu_X]E[Y - \mu_Y]}{\sigma_X \sigma_Y}$
- Spearman's correlation coefficient between X and Y:
  - each data point X_i and Y_i is converted to their ranks R_i^X and R_i^Y
  - the Pearson's correlation coefficient between R_X and R_Y is computed