2. Polynomial scaling

By feature scaling we denote a uniform scaling that would make the data fit on the (0,1) range:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.3}$$

Polynomial scaling would apply a polynomial function to the input value:

$$x' = x^n \tag{3.4}$$

For each matrix, feature scaling was first applied followed by polynomial scaling. Because the input to the polynomial scaling was in the (0,1) range, the output was also in the same range, regardless of the parameter n that has been used.

3.3.6 Hierarchical clustering

Hierarchical clustering is a method that clusters data points according to how similar they are. In our case the data points were the 29 GCV correlation vectors, and the similarity measure used was given by the Euclidean distance. See section 2.5.4 for more information on hierarchical clustering. The reason for clustering them was because we needed to find out which graphlets are similar and which ones are different with each other. The graphlets that are similar will probably have some common properties that be would be able to identify and interpret according to the dataset they are applied to.

After normalisation (both feature scaling and polynomial functions) and hierarchical clustering have been applied, the heatmaps obtained are much easier to interpret. The subsequent three sections describe the results obtained for the three main classes of networks that I have applied them on:

- Protein-protein interaction networks in section 3.5
- Metabolic networks in section 3.6
- Trade networks in section 3.7

3.4 Canonical Correlation Analysis (CCA)

In the previous sections we have shown how the Graphlets correlate with each other in a variety of experiments. However, they only give us a measure of topology in the network. In order to assign them some real meaning we have to perform Canonical Correlation Analysis (CCA), which will associate each graphlet with some of the annotations specific to the network being analysed. This will allow us to draw conclusions and insights from the data. For further information on CCA and full derivations of the eigenvectors and cross-loadings see section 2.6. The subsequent sections present the results of the CCA on our networks.

3.5 PPI networks

We shall now present the Pearson's correlation matrix a Human PPI network and Canonical Correlation Analysis results for six different Human and Yeast PPI networks. In short, the heatmap of the Pearson's GCV correlation matrix was not helpful in giving us any insights, since graphlets formed faint clusters. However, the CCA results have helped us get some interesting insights into the interactions of the proteins present in these networks.

3.5.1 Analysis of Pearson's GCV Correlation Matrix

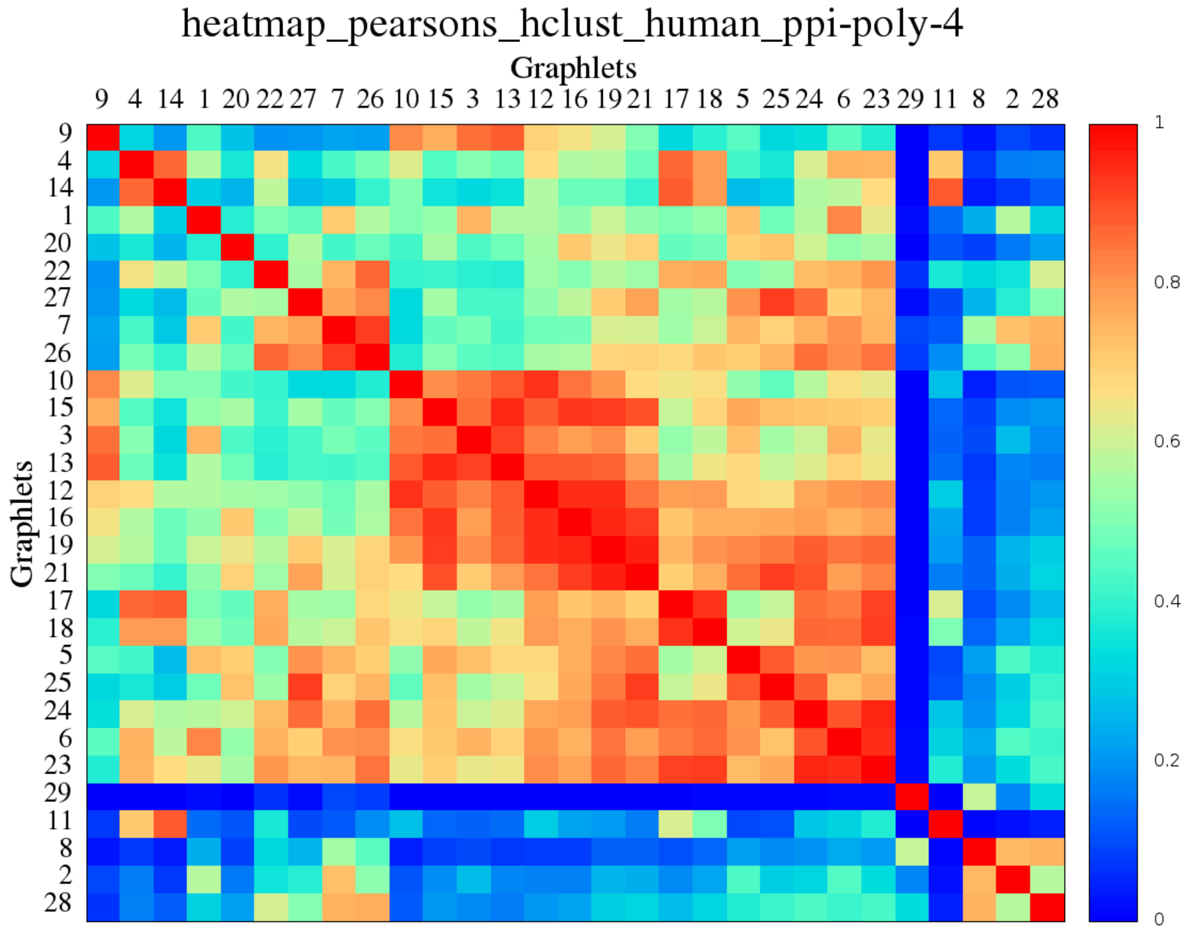


Figure 3.8: Heatmap for the Pearson's GCV correlation matrix of the Human PPI network. The heatmap has been first normalised with feature scaling and a 4^{th} degree polynomial and then hierarchically clustered.

The heatmap from figure 3.8 represents the Pearsons's correlation heatmap for the Human PPI network. It has been first normalised with a simple feature scaling and then with a 4^{th} degree polynomial, because the original correlation matrix yielded correlations that were too strong.[1] There are several clusters formed on the diagonal:

- {10,15,3,13,12,16,19 and 21}. These graphlets all contain a P4 (path on 4 nodes, graphlet G_3).
- {7,26} contain a G_7
- {4,14} contain a G_4
- {17,18} contain 2 G_2's (triangles)
- {5,25} contain a G_5
- {24,6,23} contain a G_6

[1] Having all correlations close to 1 made the identification of clusters impossible

The lack of clear graphlet clusters in the Human PPI is something that we cannot explain at the current time. Because of this, it has not been possible for us to get any insights from the Human PPI correlation matrix. Similar Human and Yeast PPI networks have yielded similar results. Further research needs to be done into this area.

3.5.2 Canonical Correlation Analysis

The next step after the Pearson's GCV correlation matrix was to run CCA on the network. We have set the X vector to be the GCV and the Y vector to be the values of Boone's annotation. For setting up the Y vector, we have labeled each protein with a vector of binary entries, with the i^{th} entry as follows:

$$Y_i = \begin{cases} 1, & \text{if the protein is annotated with the } i^{th} \text{ annotation} \\ 0, & \text{otherwise} \end{cases}$$

Since each protein had only one annotation, each sample Y_i only contained non-null entry. Unfortunately, the results of the CCA on this network didn't were not good, since the correlation was really low and the p-value is above 0.05. A table with the full results of the CCA experiment can be found in the Appendix in section 1. In the next section we will explain the subsequent experiments that have been performed on the PPI networks.

3.5.3 Results for other PPI networks

The 18 experiments

Since the CCA results with the Human PPI network didn't give us much meaningful information, we thought of exhaustively running it on several types of Human and Yeast PPI networks. We ran the same process on 5 other Human PPI networks with Boone's annotation file and on 6 Yeast networks using the two different annotation files: von Mering's and Boone's (see section 2.7.1). For these experiments we have also used high-confidence networks, which contain only protein interactions that have been confirmed by two independent sources. The networks analysed are as follows:

- Human
  1. A high-quality Human PPI network determined by Stitch-seq protocol[49], CCA results in fig. B.2
  2. Two networks from I2D, a database of PPI networks maintained by Jurisca lab[50] at Ontario Cancer Institute
     - A high-confidence version, CCA results in fig. B.10
     - Full version, CCA results in fig. B.11
  3. Two networks from BioGRID:
     - Full version, CCA results in fig. B.13
     - High-confidence version, CCA results in fig. B.14
- Yeast
  1. A network obtained through affinity-purification mass spectroscopy (AP-MS) by Collin's et al[51] - Co-complex membership associations, CCA results in fig. B.15, B.5
  2. A genetic network from BioGRID, CCA results in fig. B.16, B.6
  3. Literature-curated interactions[52], CCA results in fig. B.17, B.7

Figure 3.9

Canonical Correlation		0.15550	
p-value		0.00001	
X variate		Y variate	
G11	-0.02678	other.metabolic.processes	0.05697
G14	-0.02889	transport	0.04117
G17	-0.03063	protein.metabolism	0.02476
G21	-0.03148	cell.cell.signaling	0.01750
G25	-0.03170	cell.organization.and.biogenesis	0.01459
G27	-0.03207	developmental.processes	0.01438
G24	-0.03211	Gnal.transduction	0.01196
G12	-0.03279	cell.adhesion	0.00820
G16	-0.03324	stress.response	0.00620
G19	-0.03382	other.biological.processes	0.00192
G20	-0.03450	DNA.metabolism	-0.00567
G18	-0.03475	cell.cycle.and.proliferation	-0.01553
G22	-0.03501	RNA.metabolism	-0.06552
G10	-0.03561	death	-0.13045
G15	-0.03600		
G23	-0.03671		
G26	-0.03805		
G13	-0.03941		
G9	-0.04260		
G4	-0.04307		
G28	-0.04353		
G5	-0.04618		
G6	-0.04621		
G3	-0.04884		
G7	-0.04994		
G1	-0.06811		
G8	-0.07170		
G29	-0.07219		
G2	-0.08160		

4. y2h union yu ito uetz, CCA results in fig. **??**, **??**
5. Two PPI networks from BioGRID
   - Full version, CCA results in fig. B.3, B.8
   - High-confidence version, CCA results in fig. B.4, B.9

The best results have been obtained for the following yeast networks, for both von Mering's and Boone's annotation files:

1. Collin's AP-MS network
2. BioGRID Full
3. BioGRID High-confidence.

On the graphlet side, the highest correlations are usually with cliques 2,8 and 29 (correlation values 0.45-0.5). On the annotation side, the highest correlations are with translation (corr. value 0.5) (*Collin's Yeast AP-MS* and *Yeast BioGRID PPI - Full version*), transcription (mainly *Yeast BioGRID PPI - High-confidence version*). Therefore, we can state that the proteins involved in translation or transcription are more likely to have a neighbourhood rich in graphlets, especially cliques.

All the other combinations of networks and annotation files have yielded much poorer correlation results (only aprox 0.2) and high p-values above 0.5. One of the reason for this might be because of the high noise of the data that is prevalent in PPI networks. In the next subsections we will present these key results and provide biological interpretations for the observed fenomena. The other CCA results for all the 18 experiments are shown in the Appendix section 2.

Summary of the CCA Results from the 18 experiments

Figure 3.10 shows the CCA results for Collin's AP-MS[2] PPI network. The results mainly show that Ribosome Translation is correlated with all the graphlets, since their weights have the same sign. The spectrum of graphlets runs from the least dense graphlets {9,10,13,11,12} on top, having the lowest correlation magnitude of around 0.2 to the most dense {29,8,2} on the bottom, having the highest correlation magnitude of aproximately 0.46. This property of the graphlet vector also holds for the other relevant PPI experiments. Therefore, the observation we can make is the following: proteins involved in Ribosome translation generally interact with clusters of other proteins. This is a surprising result, since we would have expected the proteins involved in Ribosome translation to be interacting with long amino-acid chains, that are best represented by graphlets that are made of long paths such as G_9 and G_3. This result is also confirmed by the same experiment that was run using Von Mering's annotation, with *Translation* also correlating positively with all the graphlets (see figure B.5). One possibility is that these clusters are found in the Ribosome complex, a molecular machine that serves as the site for protein synthesis. It is usually made up of dozens of distinct proteins that interact with each other.

The same argument also applies to RNA processing, which converts primary transcript RNA into mature RNA[53]. Since the RNA is a single-stranded chain of nucleotides, we would have expected the proteins involved in its processing to interact sequentially with its nucleotides, a behaviour that is also best represented by graphlets such as G_9 and G_3. However, the other experiments (see fig. B.3 and B.3) have actually yielded a higher-magnitude cross-loading of around -0.2, which means that the correlation cannot be attributed to chance or noise. If we try to understand the RNA processing a bit further, we find out that there are three main tasks that occur in the cell nucleus before the RNA is translated:[54]

[2] affinity-purification mass spectroscopy

- 5' capping
- 3' polyadenylation
- RNA splicing

Polyadenylation is a process in which a segment of the newly made pre-mRNA is first cleaved off by a *set of proteins.* This protein complex then synthesize the poly(A) tail at the RNA's 3' end. We believe that this protein complex might be one of the reasons why cliques correlate highly with proteins involved in the polyadenylation step of RNA processing. The third step of the RNA processing, refered to as RNA splicing, is a process in which regions of the RNA that do not code for protein (i.e. introns) are removed and the remaining nucleotide sequence (i.e. exon) is re-connected to form a single continuous molecule. This splicing reaction is also catalyzed by a large protein complex called the *spliceosome* that is assembled from several smaller protein complexes and small nuclear RNA molecules.[55]

On the other end of the Y variate we have Golgi Endosome vacuole sorting with a weight of 0.11. Golgi endosome vacuole sorting is an environment where material is sorted before it reaches the degradative state. CCA analysis shows that proteins involved in the Golgi endosome have a sparse environment, since all the graphlets correlate negatively with the Golgi endosome index[3]. The explanation for this is that proteins involved in the Golgi endoscope mainly interact with the proteins that need to be sorted, but these don't interact with each other. This result is also confirmed by experiments 10[4] and 11[5](see figures B.3 and B.3 in the appendix).

Figure 3.10 also shows that the Metabolism/mitochondria index is negatively correlated with all the graphlets. This suggests that the proteins present in mitochondria interact with other proteins that interact much less with each other. This could be explained by the fact that the proteins present in mitochondria each have a variety of different functions and therefore their partner proteins are more unlikely to interact with each other because they have different functions. The main functions of the proteins found in mitochondria are related to:

- Energy production and cellular metabolism - the main function of a large number of mitochondria proteins is the production of Adenosine triphosphate (ATP), commonly reffered to as the energy currency of the cell[56]
- Pyruvate and the citric acid cycle[56]
- Electron transport chain[56]
- Heat production[56]
- Storage of calcium ions[57]
- Signaling through mitochondrial reactive oxygen species[58]
- Regulation of the membrane potential[56]
- Apoptosis-programmed cell death[59]
- Calcium signaling (including calcium-evoked apoptosis)[60]
- Regulation of cellular metabolism[61]
- Certain heme synthesis reactions[62]

[3] they correlate negatively since their weights have different signs: Golgi endosome has a weight of 0.11, while all the graphlets have negative weights

[4] Experiment 10 was run using the Yeast BioGRID PPI - Full version network + Boone's annotation

[5] Experiment 11 was run using the Yeast BioGRID PPI - High confidence version network + Boone's annotation

- Steroid synthesis[63]

We can illustrate our last argument using a small, simple example. *Cytochrome c* is a small protein found in the inner membrane of the mitochondrion. It is an essential protein in the Electron transport chain, where it carries one electron. Apart from electron transportation, it is also invoved in the initiation of Apoptosis, that is the programmed cell death. However, the interacting partners of *Cytochrome c* are much less likely to interact with each other, since they are split in two different functional groups. This process can be extended for the other of its subfunctions. Now, from a topological point of view, that is why the network of partners of *Cytochrome c* will be more likely to form sparser graphlets such as {9,10,13,11 or 12} as opposed to dense graphlets such as {29 or 28}.

Figure 3.10

Canonical Correlation		0.53013	
p-value		0.00000	
X variate		Y variate	
G9	-0.21219	Golgi.endosome.vacuole.sorting	0.11166
G10	-0.23170	Metabolism.mitochondria	0.09505
G13	-0.23282	DNA.replication...repair.HR.cohesion	0.09497
G11	-0.27316	Chromatin.transcription	0.08089
G12	-0.29269	Cell.polarity.morphogenesis	0.07997
G15	-0.29388	Signaling.stress.response	0.07165
G18	-0.29605	Chrom..seg..kinetoch..spindle.microtub.	0.06670
G14	-0.30226	Protein.folding...glycosylation.cell.wall	0.05669
G16	-0.30491	ER.Golgi.traffic	0.05629
G3	-0.32983	Nuclear.cytoplasmic.transport	0.04242
G19	-0.32988	Cell.cycle.progression.meiosis	0.01010
G21	-0.34306	Protein.degredation.proteosome	0.00767
G17	-0.35499	RNA.processing	-0.04756
G20	-0.36336	Ribosome.translation	-0.50899
G4	-0.36451		
G23	-0.37679		
G25	-0.39331		
G24	-0.39659		
G22	-0.40344		
G6	-0.40955		
G5	-0.41656		
G27	-0.42197		
G26	-0.42617		
G28	-0.44430		
G1	-0.45305		
G7	-0.45431		
G29	-0.46431		
G8	-0.47914		
G2	-0.49953		

Ribosome translation, RNA processing, Golgi endoscope sorting and Metabolism/mitochondria are the annotations that have consistently shown up with strong correlations in all our relevant[6] experiments. The other annotations have varied in weights, so we will consider that our GCV

[6] that is experiments with the Biogrid and Collin's yeast networks, since they have a p-value below 0.05 and relatively high canonical correlations

signature coupled with Canonical Correlation Analysis cannot capture any patterns in proteins that are part of those processes.

3.6 Human Metabolic network

3.6.1 Analysis of Pearson's Correlation Matrix

Figure **??** illustrates the final Pearson's GCV correaltion matrix for the Human metabolic network. We can clearly distinguish several clusters of graphlets that have been formed along the main diagonal. Section 2.1 describes the graphlet terminology in detail. The main clusters are as follows:

1. **Claw** cluster made of Graphlets 4,16,5,25,1,17,14,22. These graphlets all have a C4 (claw - graphlet G4) as a subgraph.

2. **Paths** cluster made of Graphlets 9,13,21,10,15,12,3,19. These graphlets all have a P4 (path - graphlet G3) as a subgraph.

3. **Triangles** cluster made of Graphlets 2,26,24,18,23,27,6,7. These graphlets all have triangles (graphlet G2) as subgraphs

4. **Cliques** cluster made of Graphlets 29,8,28. These graphlets are cliques, with the exception of 28, which is almost a clique as it is missing as edge. Note that the 3 node clique is missing because being a triangle, it is correlated more with the triangle group above.

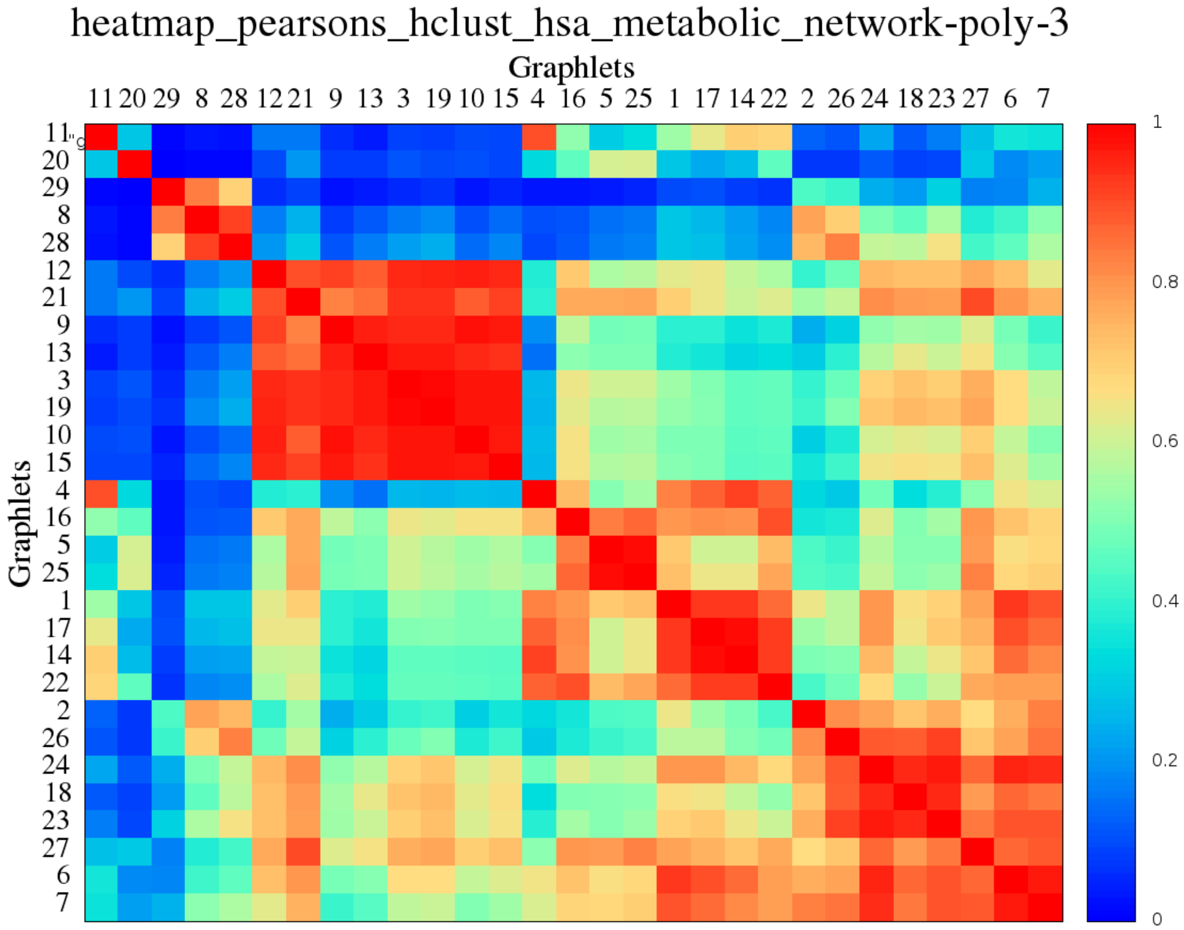


Figure 3.11: Heatmap for the Pearson's GCV correlation matrix of the compound-based Human Metabolic network. The heatmap has been normalised with a 3^{rd} degree polynomial and hierarchically clustered.

Furthermore, we can also notice that graphlets from clusters 1,2,3 also have a high degree of inter-correlation, since they might contain claws, paths and triangles at the same time. This is not the case for group 4, which is made of cliques. The cliques only bear some correlation with the third cluster made of triangle-like graphlets, which is not surprising for the following reasons:

- Cliques contian a lot of triangles
- Cliques do not contain claws C4 or paths P4, which miss several edges.

It should also be noted that graphlets 11 and 20 have been left outside, as they don't strongly correlate with any of the other groups. The cluster closest to these 2 grahlets is the claw cluster.

To sum up, we can see how the graphlets cluster together according to what basic shapes they contain.

3.6.2 Canonical Correlation Analysis

In order to run Canonical Correlation Analysis on the metabolic networks we used *Enzyme Commission* (EC) numbers as annotations for the network nodes. See section 2.7.2 for more information on EC numbers.

There is some degree of correlation between the Graphlets and the EC numbers ($\rho = 0.517$), with a p-value < 0.05. All the cross-loadings from both the Graphlets and the EC numbers

Canonical Correlation		0.51769	
p-value		0.00000	
X variate		Y variate	
G20	-0.13839	EC5	-0.11422
G11	-0.17646	EC2	-0.12601
G9	-0.17733	EC4	-0.16144
G10	-0.18420	EC1	-0.16156
G13	-0.18757	EC3	-0.21057
G16	-0.20243	EC6	-0.40615
G12	-0.20399		
G4	-0.20521		
G15	-0.20553		
G3	-0.20671		
G19	-0.20956		
G22	-0.21329		
G5	-0.21897		
G25	-0.22033		
G14	-0.22256		
G17	-0.23125		
G21	-0.23144		
G18	-0.24860		
G1	-0.25408		
G27	-0.25641		
G6	-0.26110		
G24	-0.26365		
G23	-0.26813		
G7	-0.28239		
G26	-0.28362		
G28	-0.32754		
G29	-0.34583		
G2	-0.36785		
G8	-0.37442		

Figure 3.12: CCA analysis on the compound-based Human Metabolic network

have the same sign, which suggests that they are positively correlated. Cliques 8, 2 and 29 have the highest magnitude in their weights, while EC6 (ligands) have the highest magnitude in the EC vector.

EC6 refers to ligases, which are enzymes that can catalyze the joining of two large molecules by forming a new chemical bond.[64] The reason why the magnitude of EC6 is quite high (0.4) compared to the other indicators is because a ligase normally catalyzes two large molecules by forming a new chemical bond. The two large molecules would be represented in the metabolic network by cliques or dense clusters because they have a lot of interactions and feedback loops between them. This is why cliques {8,2,29} or dense graphlets such as G_{28} have the weights with the highest magnitude. This however doesn't exclude other sparser clusters to be part of the two molecules catalyzed by the ligase, since graphlets such as {9,10,11 or 12} also correlate positively with EC6.

3.6.3 CCA Results for other model organisms

We have analysed other metabolic networks that belong to the following organisms: C. elegans, D.melanogaster, E.coli, M.musculus, S.cerevisiae. For all these organisms, we have analysed both compound-based networks and also enzyme networks.

They results for the other compund-based networks confirm the correlation heatmaps and CCA results that were obtained for Homo Sapiens. Average CCA correlation is around 0.5, EC6 has the highest magnitude at around 0.4 and cliques 2,8,29 are usually the most correlated with it (0.35).

However, the enzyme networks display a much lower CCA correlation (around 0.25). This is the case for all the organisms, including humans. The Graphlet signatures have very low signatures, while EC numbers don't have magnitudes above 0.22.

3.6.4 CCA on the Kegg categories

We have tried to use the Kegg categories as annotations for the enzymes in the metabolic network. We have initially annotated the enzymes with the following high-level categories:

- Metabolism
- Genetic Information Processing
- Environmental Information Processing
- Cellular Processes
- Organismal Systems
- Human Diseases

The CCA correlation obtained was only around 0.6, so we tried running CCA on the lower-level category. That is, for each of those 6 high-level categories, we ran CCA on its subcategories. The best results, presented here, were obtained for the Human Diseases, Cellular Processes and Organismal Systems sub-categories.

Figure 3.13 shows the CCA for Cellular Processes. We can clearly see that Graphlet G_9 correlates positively with Transport and catabolism. The reason for this is because in Catabolism, large molecules such as polysaccharides, lipids and nucleic acids are broken down into smaller units such as monosaccharides, fatty acids or nucleotides. Since molecules such as polysaccharides are made up of long chains of small monomer units, Graphlets that are made of long paths such as G_9 will be overly represented in these processes. Similarly enzymes involved in transport are transporting nutrients from one chemical to another, so their interactions will be characterised by long "transportation" paths that are best represented by graphlet G_9. At the other end of the spectrum, Cell growth and death and Cell communication are correlated with graphlets {1,2,7 and 8}. The reason for this is because in Cell Communication, if a cell is stimulated, it's needs to send signals to it's neighbours through the use of molecules. First of all, in order to ensure that a signal is successfully transmitted, several molecules carrying the same message could be transmitted and there has to be a lot of different possible paths to reach the destination. This is why a graphlet like G_9 would correlate negatively with these, because G_9 is made of a long path of 5 nodes and if one of the nodes fails, then the whole signal is lost. Graphlets like {2,7 and 8} correlate positively because these are highly connected ({2,8} being cliques) or because they contain several alternative paths for message transmission (G_7). However, the reason why graphlet G_1 correlates with Cell Communication is still a matter or research.

Figure 3.13: Cellular Processes CCA

Canonical Correlation		0.98633	
p-value		0.00000	
X variate		Y variate	
G9	0.04828	Transport.and.catabolism	0.52121
G21	0.01960	Cell.motility	0.20502
G25	0.01441	Cell.communication	-0.40751
G5	0.01434	Cell.growth.and.death	-0.69712
G16	0.00969		
G13	0.00199		
G12	-0.00048		
G27	-0.00134		
G20	-0.00256		
G3	-0.00412		
G24	-0.01287		
G19	-0.01528		
G10	-0.01623		
G18	-0.01681		
G14	-0.02579		
G11	-0.02667		
G23	-0.02851		
G15	-0.03092		
G17	-0.03201		
G29	-0.04271		
G6	-0.04386		
G28	-0.04750		
G4	-0.05059		
G26	-0.05235		
G22	-0.05877		
G8	-0.05881		
G7	-0.07069		
G2	-0.07388		
G1	-0.07463		

Figure 3.14 shows the CCA for Organismal Systems. These results show that enzymes involved in Environmental Adaptation and Excretory systems are usually rich in interactions and their neighbours are also highly clustered, since all the graphlets correlate positively with these. On the other hand, Circulatory and Digestive metabolic pathways are sparse and would ideally contain few graphlets. One explanation for this is because in these systems enzymes,

proteins and metabolites have to circulate throughout the whole body and interact with distant enzymes, which don't cluster together. Enzymes at the other end of the spectum (Environmental Adaptation and Excretory system), the enzymes and proteins are much more localised in the body. For instance, excretory system enzymes are mainly active in the kidney or liver. Moreover, the enzymes in the Circulatory and Digestive systems will probably have much less interactions compared to their counterparts in the Environmental Adaptation and Excretory systems.

Figure 3.14: Organismal Systems CCA

Canonical Correlation		0.96925	
p-value		0.00000	
X variate		Y variate	
G26	0.30978	Environmental.adaptation	0.20426
G24	0.29818	Excretory.system	0.19729
G23	0.29308	Development	0.07461
G18	0.28901	Endocrine.system	0.04192
G6	0.27857	Nervous.system	-0.01315
G12	0.26520	Sensory.system	-0.06276
G19	0.25419	Immune.system	-0.15192
G3	0.24988	Digestive.system	-0.23211
G14	0.23274	Circulatory.system	-0.37659
G13	0.23155		
G1	0.22611		
G17	0.22546		
G7	0.21225		
G27	0.20440		
G10	0.19976		
G9	0.19547		
G25	0.19422		
G16	0.19135		
G28	0.18874		
G4	0.18421		
G5	0.18381		
G20	0.17359		
G11	0.15537		
G21	0.14453		
G29	0.11208		
G8	0.10681		
G2	0.10369		
G22	0.08192		
G15	0.01276		

Figure 3.14 shows the CCA for Human Diseases such as Cancers, Immune, Neurodegenerative and Cardiovascular diseases. The result that is most striking here is that Cardiovascular diseases and Substance dependence correlate negatively with almost all the graphlets (apart from {2 and 8}). That implies the enzymes and proteins involved in these Human Diseases have a low number of interactions and when they do have interactions, their neighbourhood only contains small clusters of 3-4 nodes maximum. The explanation for this might be the same as for the Organismal Systems: the enzymes involved in Cardiovascular diseases and substance dependence travel through long pathways throughout the body and end up interacting with distant chemicals that do not interact with each other because of their distant location in the body. However, the small clusters of interactions might occur because of locality, that is all

chemicals involved will be in the same area and therefore inevitably interact with each other. In the case of the Cardiovascular diseases, the enzymes travel long distances because they are transported through the blood vessels, while the ones involved in Substance dependence are again transported through the blood vessels or other channels.

Figure 3.15: Human Diseases CCA

Canonical Correlation		0.99479	
p-value		0.00000	
X variate		Y variate	
G2	0.01681	Cardiovascular.diseases	0.99171
G8	0.00462	Substance.dependence	0.57989
G29	-0.00737	Infectious.diseases	0.21844
G7	-0.00812	Neurodegenerative.diseases	0.00366
G1	-0.00989	Endocrine.and.metabolic.diseases	-0.02545
G26	-0.01077	Immune.diseases	-0.03029
G24	-0.01274	Cancers	-0.08682
G6	-0.01321		
G28	-0.01321		
G15	-0.01342		
G23	-0.01369		
G22	-0.01420		
G21	-0.01438		
G14	-0.01449		
G12	-0.01465		
G17	-0.01516		
G16	-0.01521		
G18	-0.01526		
G13	-0.01528		
G19	-0.01562		
G9	-0.01565		
G10	-0.01641		
G4	-0.01727		
G3	-0.01742		
G11	-0.01781		
G20	-0.01936		
G25	-0.02017		
G5	-0.02438		
G27	-0.02731		

3.7 Trade networks

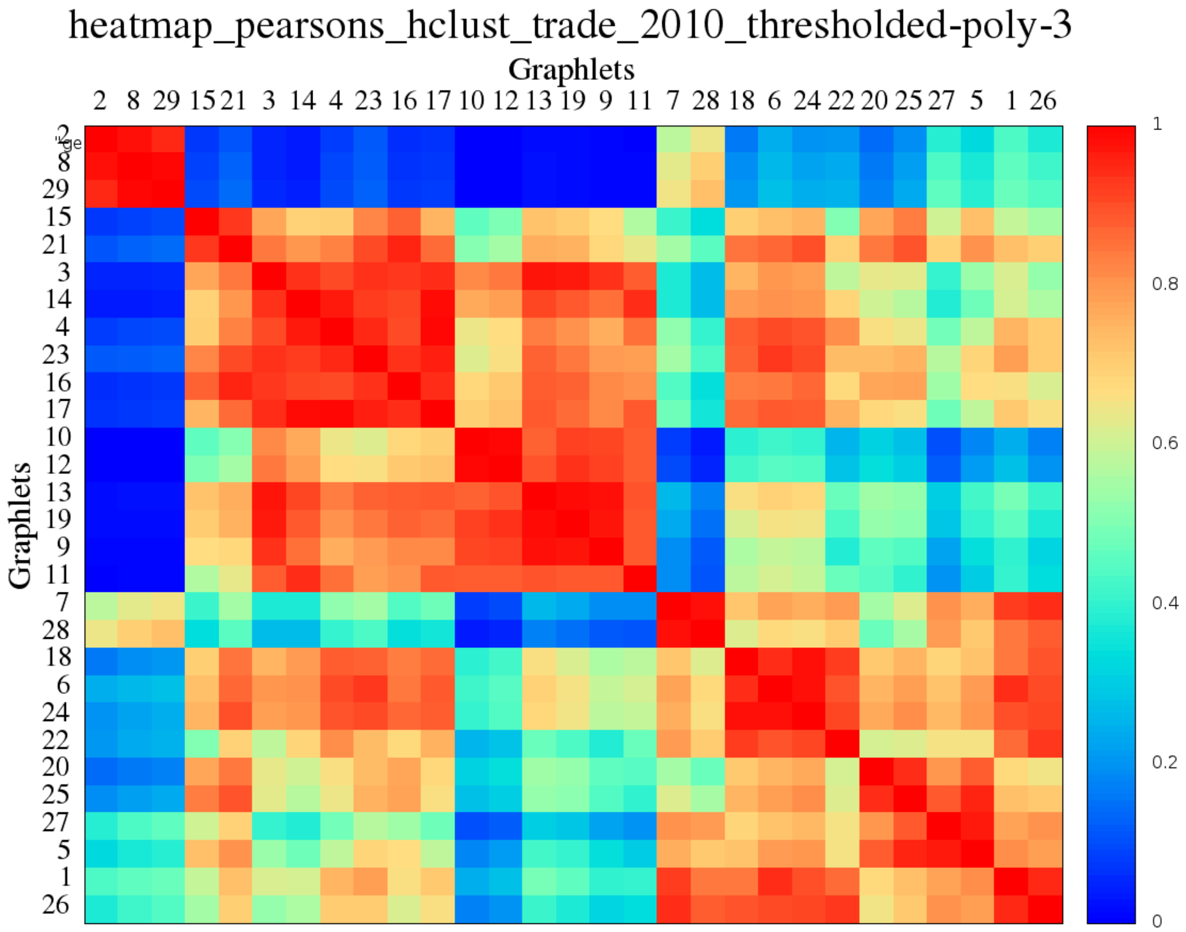


Figure 3.16

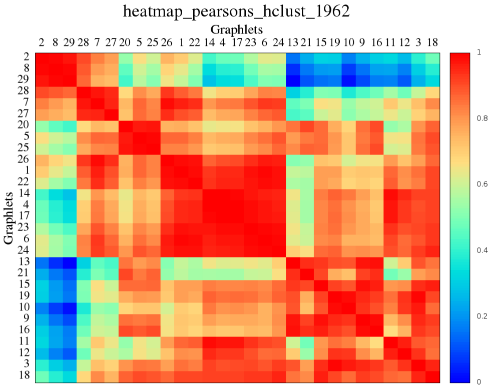
In the trade network, we can observe several clusters of graphlets that ahave been formed along the diagonal:

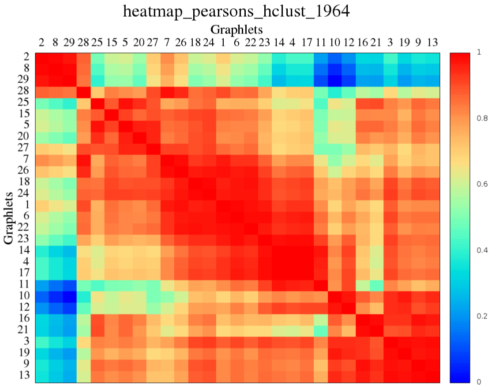
- Cliques cluster made of graphlets 2,8,29. If a country has a lot of cliques in its neighbourhood, then it is part of a densely connected group of countries.
- A cluster that is made of graphlets 15,21,3,14,4,23,16,17,10,12,13,19,9 and 11 which can be split into 2 further sub-clusters:
  - P4 cluster made of graphlets 15,21,3,14,4,23,16,17. These are all graphlets that contain a P4 (path on 4 nodes, graphlet G3).
  - Claw cluster made of 10,12,13,19,9,11. These graphlets all contain C3 (claw on 3 nodes, graphlet G4)

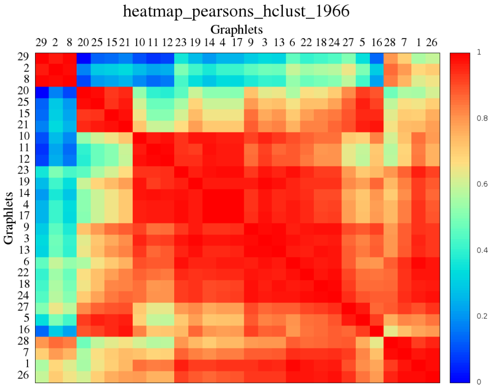
It should be noted that the diameter of the trade network is really small (aprox 5). This means that nodes will share a large proportion of their neighbourhood, especially hub nodes. In order to fix this issue, we have tried to threshold the economic networks at a level lower than 85% in order to remove some of the edges and thus yield a higher diameter. However, this has not resulted in a lower diameter (it stayed constant at around 5), most probably because of the scale-free properties of the network.[7] We have therefore decided to continue our analysis with the initial 85% thresholding.

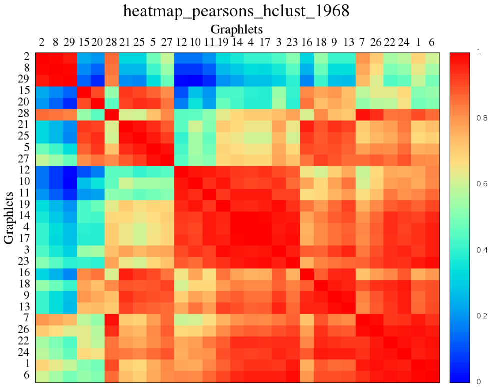
3.7.1 Analysis of trade networks during 1962-2010

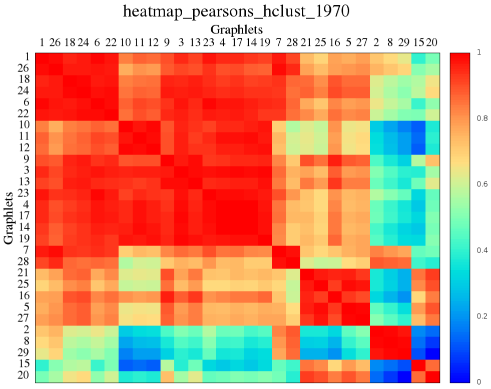
1962 to 1972

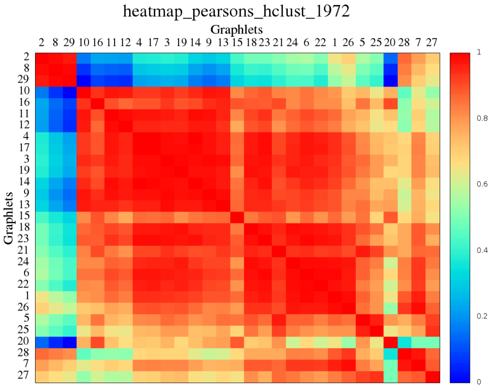












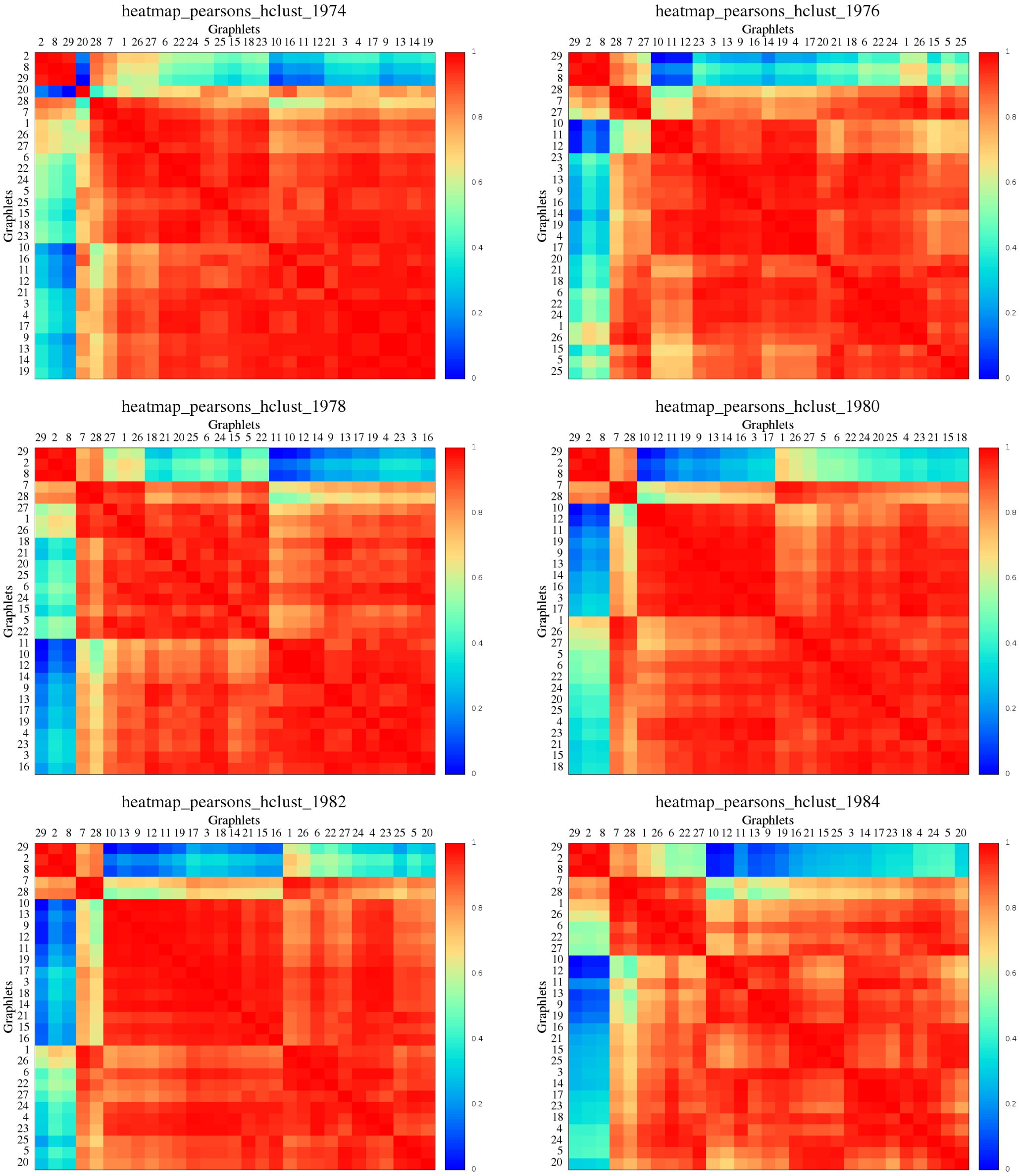
The first thing we notice is that the cliques 2,8 and 29 cluster together in each of the years analysed. In most of the years (apart from 1972) we also observe a cluster containing Graphlets that are made of cycles of length 4 (graphlet G5), with proeminent graphlets including 5, 20, 21 and 25.

Another trend we notice is that the graphlets become more and more correlated (this will be even more obvious in the second batch of years 1974-1984). This might be an effect of

[7] We have tried 9 different thresholding levels in increments of 10% each (i.e. 90%, 80%, 70% ...), but all of the resulting networks had the same network diameter(aprox. 5). As the thresholding decreased and the network was becoming smaller and smaller, only the hub nodes were kept (i.e. the big or wealthy countries such as USA, China, Japan, etc ..)

globalization, as countries become more and more connected and the diameter of the trade network gets smaller. This in turn causes the countries to have share a higher proportion of the neighbourhood, which yields a higher graphlet correlation.

1974 to 1984

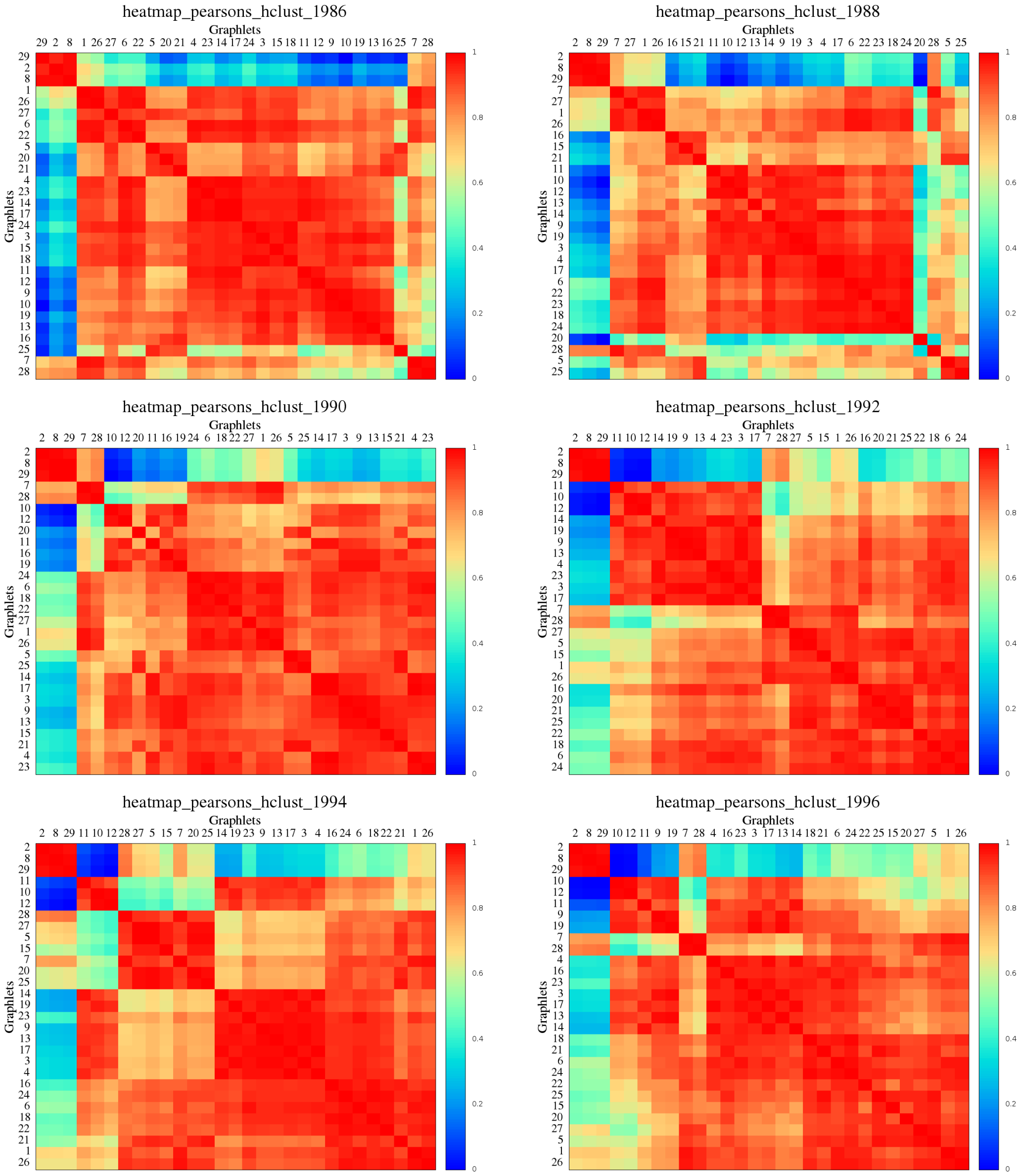


We also observe that the cliques 2,8 and 29 cluster together in every year that has been analysed in this period. In year 1980 we also observe a cluster that is made of graphlets 10,12,11,19,9,13,14,16,3 and 17. All of these contain P4's (paths of length 4, graphlet G3). This cluster on P4 can also be observed in years 1976 and 1982, but the order in which graphlets appear is different, due to the clustering algorithm.

In year 1984, we can actually notice several small clusters on the diagonal. One cluster is made of graphlets 7, 28, 1, 26, 6, 22 and 27, which all have a P3 (path on 3 nodes, grahplet G1). However, cluster 10,12,11,13,9,19 is made of graphlets that don't seem to have a lot in common apart from P2's. Cluster 16,21,15 and 25 is mostly made of graphlets that have cycles of length 4, apart from graphlet G15 which has a cycle of length 5.

On broad terms, we also notice that the graphlets become much more correlated in this period, as a result of globalisation.

1986 to 1996



3.7.2 Correlation matrix change during 1962-2010

After calculating the Pearson's GCV correlation matrix for all the years between 1962-2010, we can now calculate the change in the correlation matrix during the respective timeframe. In order to calculate the change in correlation matrix between year Y and $Y + 1$, we simply subtract in a pairwise manner the two matrices and then return the sum of squares of all the elements in the matrix. For the exact formula used see equation 2.8 from section 2.5.3.

In this section we are trying to find whether there is a connection between network topology and Crude Oil price. If one of these attributes changes, it might be possible that the other reacts with a certain number of years delay. In order to account for this, we have shifted the vector of GCV correlation change by [-2,-1,0,1,2] years. For each of these 5 cases, we have calculated the Spearman's rank correlation coefficient and the respective p-value.

The best correlation has been obtained when the vector of GCV correlation has been shifted by -2 years. This scenario is plotted in fig. 3.17. Suprisingly, the oil change in inversely correlated to the change in network topology: the Spearman's rank correlation coefficient is -0.49, having a p-value of 0.0004. The explanation for this is as follows: high oil prices generally have a large negative impact on the global economic growth. Slower growth would lead to diminished investment-related activity in the countries affected, which would in turn deter the creation of new trading partners that would keep the network structure unchanged. No major changes in the network structure would result in a low GCV correlation change.

However, there are several other major economic events for which we do not have a big change in the topology of our network, such as the 2007 sub-prime mortgage crisis or the 1997 Asian financial crisis. Similar results that use a normalised version of the GCV for the change are better correlated with global economic and social events(see section 3.11.2).

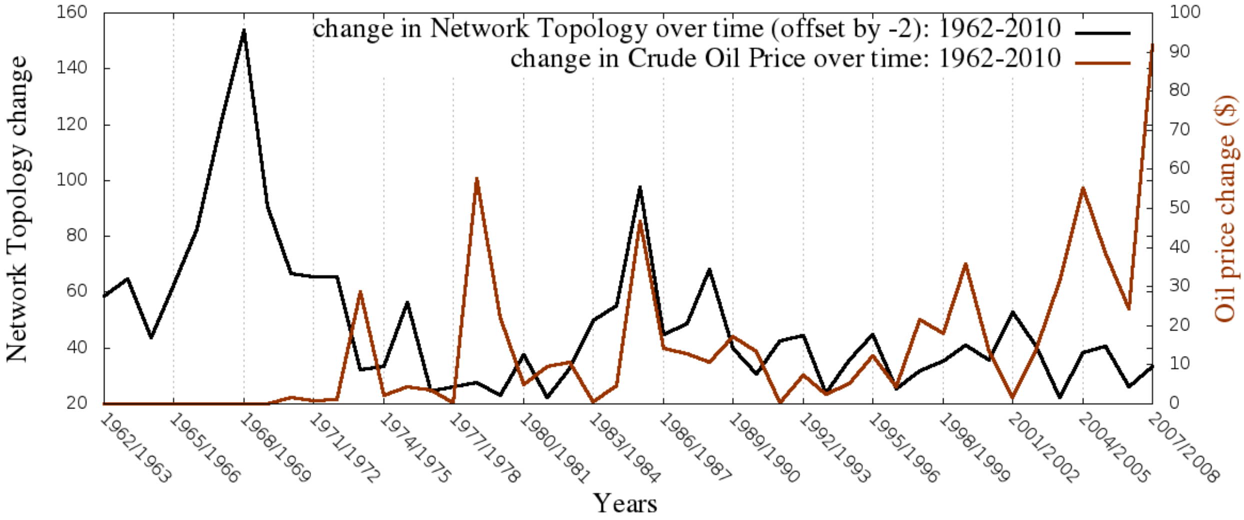


Figure 3.17: Evolution of trade network structure during 1962-2010. Plotted in black is the change in GCV correlation that has been offset by -2 years, while the change in Crude Petroleum Price is plotted in brown. Spearman's rank coefficient between oil price change and change in network topology is -0.49 with a pvalue of 0.0004. This suggests that when the change in GCV correlation between countries changes, then the oil price stays the same.

3.8 CCA - 1980-2010 Trade networks

Figure 3.18

Canonical Correlation		0.89595	
p-value		0.00000	
X variate		Y variate	
OPENK	0.24745	G20	-0.44454
BCA	0.20019	G15	-0.57546
KG	0.17422	G16	-0.62203
BCAperRGDPL	0.14775	G25	-0.63754
KI	0.05999	G5	-0.66142
KC	-0.09780	G21	-0.67142
XRAT	-0.11999	G29	-0.68727
KIxRGDPL	-0.17869	G27	-0.70899
KGxRGDPL	-0.21884	G9	-0.72424
RGDPL2	-0.26616	G10	-0.72915
RGDPL	-0.26629	G11	-0.73369
RGDPCH	-0.26634	G12	-0.74053
KCxRGDPL	-0.42077	G18	-0.74056
KGxRGDPLxPOP	-0.69344	G28	-0.74525
KCxRGDPLxPOP	-0.69962	G19	-0.74581
RGDPL2xPOP	-0.71683	G13	-0.74857
RGDPLxPOP	-0.71877	G8	-0.75020
RGDPCHxPOP	-0.71889	G26	-0.75734
KIxRGDPLxPOP	-0.72552	G14	-0.76293
LE	-0.75818	G24	-0.76458
POP	-0.76667	G23	-0.76656
		G22	-0.76781
		G17	-0.77813
		G4	-0.78270
		G3	-0.78361
		G7	-0.79365
		G6	-0.80315
		G2	-0.80569
		G1	-0.82332

CCA results clearly show that big and rich countries that have a high population and GDP per capita have a neighbourhood rich in Graphlets, while small and poor countries with account deficits have a sparser neighbourhood. The population of the country seems to be quite an important factor for determining whether it will have a neighbourhood rich in graphlets because of the following two reasons:

- In the indicators vector, population has the weight with the highest magnitude: 0.766
- Most of the other indicators that have a high weight are obtained by mutiplying population with other indicators. It should be noted that this is the case also in Omer et al's paper "Revealing the hidden language" ...

3.8.1 Canonical correlation analysis on Economic Integration

I have tried to analayse whether the level of *Economic Integration* of a country is positively correlated with dense graphlets and negatively correlated with sparse graphlets. This is something to be expected, since when a country is part of a strong trading bloc, then it's neighbours have a higher probability of doing heavy trade with one another. This is because there is incentive for the country to trade more with the partners from the same bloc, that are already trading a lot with each other. This would in turn result in denser graphlets in the neighbourhood of that country.

I have therefore annotated each country with a number (1-6) that measures the degree of economic integration using a map from a Wikipedia article[65]:

- 0 - no economic integration
- 1 - Multilateral Free Trade Area (AFTA, CEFTA, CISFTA)
- 2 - Customs union (EAC, EUCU, MERCOSUR)
- 3 - Common market (EEA, EFTA)
- 4 - Customs and Monetary Union (CEMAC/franc, UEMOA/franc)
- 5 - Economic union (CSME, EU)
- 6 - Economic and monetary union (CSME + EC dollar, EU + euro)

Canonical Correlation		0.61882	
p-value		0.01280	
X variate		Y variate	
G29	0.28704	Integration	0.61882
G8	0.28349		
G2	0.27862		
G22	0.26820		
G28	0.26806		
G7	0.26021		
G26	0.26011		
G18	0.24661		
G1	0.23837		
G24	0.23133		
G6	0.22969		
G4	0.22021		
G27	0.22013		
G17	0.21021		
G14	0.20631		
G23	0.20420		
G5	0.20149		
G20	0.19391		
G25	0.19122		
G21	0.19076		
G16	0.19026		
G11	0.18330		
G3	0.17730		
G15	0.16814		
G13	0.16764		
G19	0.16051		
G9	0.15087		
G12	0.14886		
G10	0.14638		

Figure 3.19

The results confirmed our initial expectations, with dense graphlets correlating most with the integration index, while the sparse graphlets correlating least. However, since the data has been manually annotated using some potentially unreliable source, we have searched for some official index that quantifies political integration for each country around the world. Although we haven't found something that quantifies countries according to the 6-level scale that we previously mentioned, we have found on the WTO[8] website some indices that measure the number of *Regional Trade Agreements* of a country[66]. These *Regional Trade Agreements* are defined as trade agreements that are concluded between countries that are normally geographically close to each other. They facilitate trade on a regional basis and can be of several types:

- A Free Trade Agreement (FTA)
- A Customs Union (CU)
- Economic Integration Agreement (EIA)

[8] World Trade Organisation

Canonical Correlation		0.81460	
p-value		0.00000	
X variate		Y variate	
G10	0.04910	Services EIAs	0.00187
G11	0.04673	Physical RTAs	-0.14733
G9	0.04035	Goods RTAs	-0.15447
G12	0.03890		
G20	0.03736		
G14	0.03384		
G13	0.03298		
G16	0.03295		
G15	0.02992		
G19	0.02690		
G17	0.01933		
G18	0.01594		
G21	0.01429		
G3	0.01405		
G4	0.01362		
G25	0.00856		
G22	0.00484		
G24	0.00225		
G23	-0.00702		
G27	-0.01388		
G5	-0.01887		
G6	-0.02347		
G26	-0.03082		
G1	-0.06278		
G7	-0.07024		
G28	-0.07331		
G29	-0.14967		
G8	-0.15671		
G2	-0.16970		

Figure 3.20: Canonical Correlation Analysis on Trade Integration using the number of Regional Trade Agreements as an indicator of trade integration.

- Partial Scope Agreement[9] (PS)

The results of the Canonical Correlation Analysis are shown in figure 3.20. The results with this dataset are even better than the ones with the manual Wikipedia annotations. As we expected the Goods and Physical RTAs are correlating positively with dense graphlets such as cliques {2,8 and 29} and negatively with sparse graphlets such as {10,11,9 and 12}. This suggests that once a country is acceeding to a trading block, its entire trade shifts towards its partners within the block, which will trade mainly with each other, hence the dense graphlets in the neighbourhood structure. Suprisingly, the services EIAs are not showing this correlation, having a small but positive weight of 0.00187. This imply that when a country negotiates EIAs, that doesn't result in the total trade getting redirected towards the signatories of the EIAs. Further research needs to be done in order to explain why this is the case.

[9] covers only certain types of products

3.9 Trade network - Revision of GCV - normalisation

3.10 CCA - 1980-2010 Trade networks with the normalised GCV

Figure 3.21

Canonical Correlation		0.94594	
p-value		0.00000	
X variate		Y variate	
POP	0.73628	G12	0.90456
LE	0.71650	G10	0.89337
KIxRGDPLxPOP	0.66038	G14	0.87536
RGDPCHxPOP	0.65383	G17	0.85955
RGDPLxPOP	0.65376	G9	0.84692
RGDPL2xPOP	0.65226	G11	0.83708
KGxRGDPLxPOP	0.64238	G19	0.70966
KCxRGDPLxPOP	0.63303	G4	0.69198
KCxRGDPL	0.29252	G16	0.67490
XRAT	0.17083	G3	0.63019
RGDPCH	0.16079	G18	0.60564
RGDPL	0.16071	G24	0.59760
RGDPL2	0.16038	G13	0.59247
KGxRGDPL	0.15848	G22	0.54531
KIxRGDPL	0.10411	G23	0.47154
KC	0.08634	G15	0.43876
KI	-0.01620	G21	0.32221
BCAperRGDPL	-0.10953	G20	0.28966
KG	-0.12868	G26	0.27068
BCA	-0.14935	G6	0.23057
OPENK	-0.26502	G27	0.15386
		G25	0.14823
		G5	0.11232
		G28	-0.15016
		G1	-0.16367
		G7	-0.21277
		G2	-0.48656
		G29	-0.52462
		G8	-0.63741

Economic interpretation

CCA shows that the good indicators such as RGDPL, POP, LE are positively correlated with the graphlets {12,10,14,17,9,11,19,4,16, etc ..}. On the other hand, the bad indicators correlate positively with graphlets 8,29,2,7,1,28. We now wonder what do graphlets from these two sets have in common. We first notice that graphlets 8,29,2,7,1,28 represent cliques 8,29,2 or almost cliques 7,1,28. Since these graphlets are very densely connected, this suggests that the trading partners of small and poor countries are trading heavily with each other or form highly connected clusters. This suggests that **the majority of the trading partners of small and poor countries are the big and rich countries that are always trading heavily with each other**.

This theory seems to be confirmed by taking a few small and poor countries and looking at their trading partners. Note that since the network has only 119 nodes, most of the poorest countries from Africa of South Asia have already been filtered out. Therefore, I have selected Morocco as a small and poor country relative to the others, although in reality it considered to have a medium level of development. Morocco's main trading partners are: Saudi Arabia, China, France ,USA , Spain , Germany, Italy. These countries are big and rich and every single pair of them clearly trade with each other. Similar results have been observed for other countries such as Bulgaria or Uruguay. Moreover, all of Morocco's trading partners are part of G20, a club of big and wealthy countries that collaborate with each other on economic matters. This leads us to a second theory: **since the trading partners of a small and poor country form highly connected clusters, these clusters represent big and rich economic groups such as G8, G20, OECD, Paris-club, BRIC**. Again, this is validated by selecting a few countries and looking at their neighbours. For example, the trading partners of Tunisia are Germany, France and Italy, all part of G8, G20, OECD and Paris-club.

Regarding the first group of graphlets (that is {12,10,14,17,9,11,19,4,16, etc ..}), we wonder what they all have in common? We can find the answer quite easily by temporarily looking only at the 4-node graphlets, that is {3,4,5,6,7,8}. We find that the sparse graphlets {3,4} have a high positive weight, medium-dense graphlets {5,6} have a low positive weight (0.11 and 0.23), while dense graphlets {7,8} have a negative weight. This suggests that the graphlets are ordered according to how dense they are, from sparsely connected graphlets (coloured in blue) having a high positive weight to densely connected graphlets (coloured in red) having a negative weight.

Now that we now know to interpret the positively weighted part of the graphlet vector as sparse graphlets, canonical correlation tells us that the trading partners of big and wealthy countries have a lot of sparse graphlets in their neighbourhood. The economic reason for this is because **big and rich countries like USA, China, Russia are trading with a lot of small, isolated countries which in turn do not trade with each other**. This theory is supported by a closer analysis with Cytoscape. Using this software I have found that the clustering coefficient of a country is inversely correlated with the wealth and size of that country, suggesting that big and rich countries have a relatively much sparser neighbourhood.

The population of the country seems to be quite an important factor for determining whether it will have a neighbourhood rich in sparse graphs because of the following two reasons:

- In the indicators vector, population has the weight with the highest magnitude: 0.73
- Most of the other indicators that have a high positive weight are obtained by mutiplying population with other indicators (KIxRGDPLxPOP, RGDPCHxPOP). It should be noted that this is also the case in Omer et al's paper "Revealing the hidden language" ...

3.11 Pearson's GCV correlation matrix

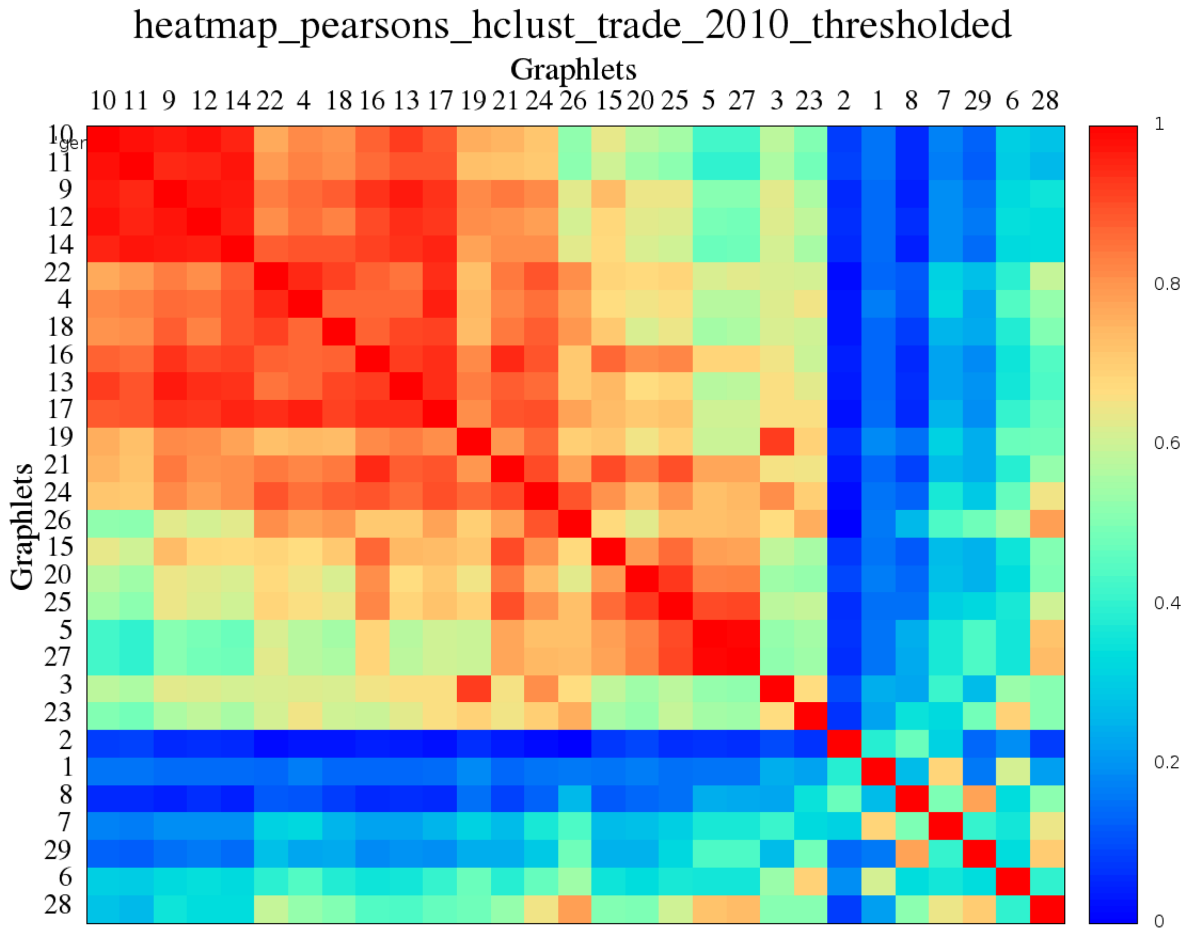


Figure 3.22

In the trade network, we can observe several clusters of graphlets that ahave been formed along the diagonal:

- **A**: Cluster made of graphlets {10,11,9,12,14}. These are all sparse graphlets, and CCA analysis shows that these graphlets are correlated with big and rich countries. Note that graphlet 9 does not always correlate in previous years. All apart from G9 contain a C4.
- **B**: A slightly similar cluster that is also correlated with the one above is {22,4,18,16,13,17}. These graphlets all contain a C4 as a subgraph, so perhaps that is the reason for being correlated together.
- **C**: Another cluster is formed by graphlets {5,25,27,20}, with graphlet 20 being added because it is highly correlated especially in other years. These graphlets all contain a cycle of length 4 (S4).

The graphlets from cluster **A** are all sparse graphs that are associated with big and wealthy countries, according to the CCA analysis (see below). The graphlets that are associated with small and poor countries are the dense graphlets { 2,8,29,1,7,28 }. These dense graphlets are however not correlated with each other. This tells us that the positive topological attributes (sparse graphlets) correlate positively with each other, while negative attributes (dense graphlets) dont't correlate with each other and correlate negatively with the positive ones (this can't be noticed from the plot because the correlation matrix is normalized, but in the original version the dense graphlets correlate negatively with the sparse ones). This in turn means, that if a country has a positive attribute, it will be prone to have more of those positive topological attributes (highly positive correlation), which will in turn lead to make it richer. Similarly, a negative topological attribute will be negatively correlated with a positive attribute, which means that if a country has a negative attribute it will be less like to have a positive attribute at the same time. Therefore, if a country is poor this will somehow stop or delay it from becoming richer. This suggests that the **Rich countries get richer, while poor countries get poorer**. The phrase is generally attributed to a free market system (capitalism) (Wikipedia, Karl Marx - law of increasing poverty).

Most of the graphlets from group A also correlate with one another in other network classes such as Knuth's literature networks. For example, in Anna Karenina graphlets 10,11,12 and 9 also also correlated and form a cluster. Similarly, graphlet from group C are also correlated in literature networks ({5,25,27} in Anna Karenina and {5,27} in David Copperfield). Therefore there are two conclusions I can draw from here:

- The trade networks and the literature networks might have some common property. What I have noticed so far is that these networks have a small number of nodes. This would in turn cause the number of samples for calculating the Pearson's correlation matrix to also be small. Therefore, the reason for the clear clusters might also be because there are less outliers, since only the strong connections are kept. To test this out I compared 2 versions of Anna Karenina. One with a small number of nodes and another with 5x more nodes. The one with a small number of nodes indeed has much clearer clusters compared to the one with a bigger number of nodes.

- The graphlets get clustered because they have a common sub-structure in common (ex: Cluster B contains C4).

Because of the above two points, I do not believe there can be attributed any roles to countries in this scenario. Another aspect is that our GCV metric is much more abstract than the equivalent GDV signature with orbits, since it gives us information about the trading partners of a particular country, but not about the country itself.

3.11.1 Analysis of trade networks during 1962-2010

3.11.2 Correlation matrix change during 1962-2010

The plot in figure 3.23 shows the changes in network structure (quantified by the normalised GCV Pearson's correlation matrix) over the 1962-2010 period and it's connection with the changes in oil price. We find that for the normalised GCV, the correlation is actually positive (0.34) and has a p-value of 0.01. These results are in contrast to the ones obtained using the original GCV signature in section 3.7.2.

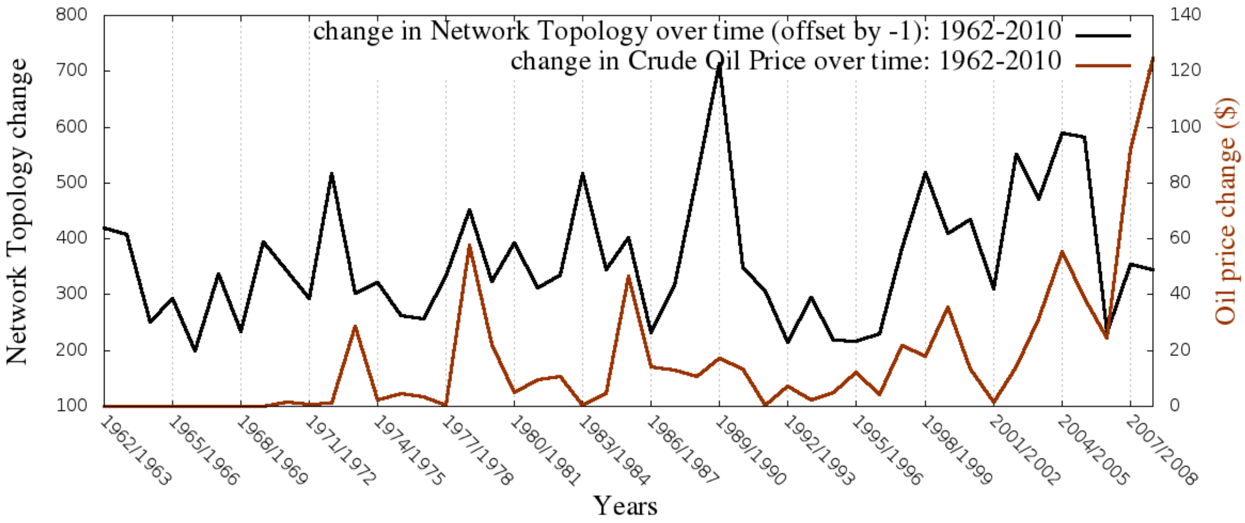


Figure 3.23: Change in overall network topology(as measured by the normalised GCV correlation matrix) versus change in crude oil price. The two plots are possitively correlated, having a Spearman's rank correlation coefficient of 0.34 and p-value of 0.01.

Important economic events that match the graph:

- OPEC oil crisis 1973 (although the peak in our graph occurs 3 years earlier)
- 1990's revolutions in Eastern Europe that mark a starting point of increasing trade between Western Europe and Eastern Europe. Transition to a capitalist-like economy happened thoughout the world.
- 1997 Asian financial crisis

The 1970s were marked by two energy crises (1973 and 1979) that explain the two small peaks seem in both the topology change but also in the oil price change. Afterwards, the 1984 peak in network topology change could be potentially explained by the early 1980 recession, which affected most of the developed world. Revival of neoliberalist economic policies occured in this time which led to reduced government intervention, lower taxes and deregulation. The peak in 1989 might be explained by the fall of communist/socialist governments in Russia, Eastern Europe and around the world accompanied by a fall in heavy industries and increased trade openness. The change in government for some former left-wing or right-wing countries such as Russia, Poland, Chile or South Africa has also led to dramatic changes in their trading partners.

The early 1990s appear as a period of relatively low changes in oil and network topology, which reflects the overall economic stability at that time. However, bigger changes are noticed in the late 1990s, possibly started by the 1997 Asian financial crisis. By the 2000s, even bigger changes can be observed in the network topology plot that could be potentially caused by the commodities boom and rising oil prices and inflation.

However, the reason why these results differ from the results we got withe the original GCV in section 3.7.2 is still an open question. This is one area of further research that can uncover more information about the structure of the trade network.

3.12 Trade partners density index

Using a combination of graphlet frequencies that were part of the GCV, we are now interested to create an index that can will give us a high correlation with the good indicators from section

3.10 such as GDP per Capita (RGDPL) or Level fo Employment (LE). We have therefore used the three graphlets that had the highest correlation with the good indicators {12,10 and 14} and the three that has the lowest correlation with the good indicators {8,29 and 2}. Multiplying each of these by their respective CCA cross-loading would give us a *trading partner density index*. This index can be calculated for every country and for every year and can have both positive and negative values. It gives a measure of the density of the network of the trading partners: the higher the value the sparser the neighbourhood, because the sparse graphlets contribute positively while the dense graphlets contribute negatively. CCA has shown us that a network of trading partners that has sparse graphlets indicates a healthy economy, so we expect the *trading partner density index* to be high for big and wealthy countries and low for small and poor countries.

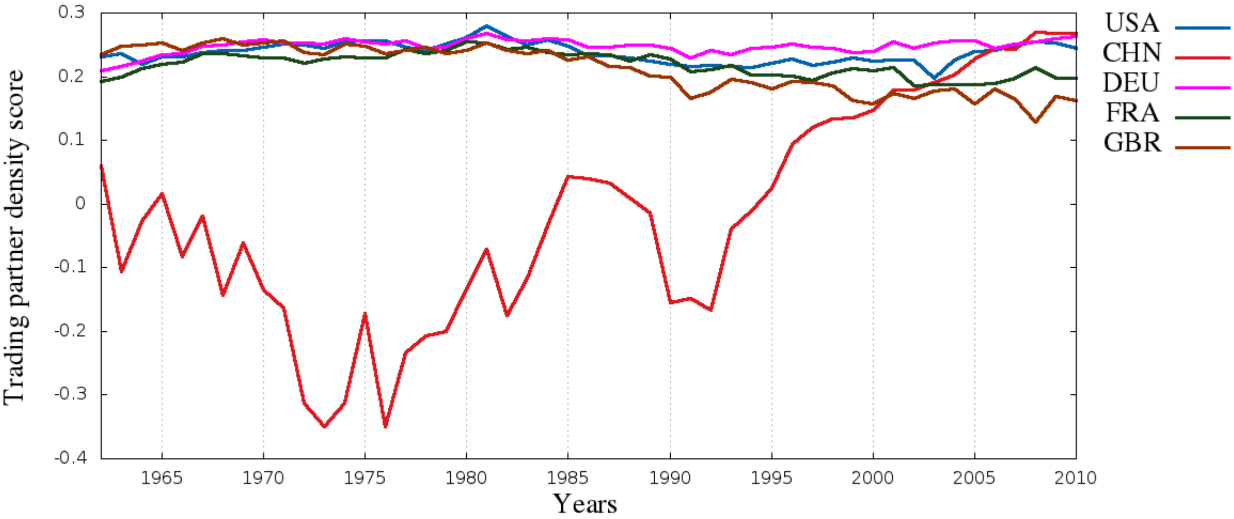


Figure 3.24: Trading partners density index measured for 5 big economies: United States (USA), China (CHN), Germany (DEU), France(FRA) and the United Kingdom (GBR).

Figure 3.24 shows the *trading partners density index* for several big economies of today's world. Throughout the 1965-2010 period, the corresponding index for the United States, Germany, France and the United Kingdom has been approximately even, having a value of 0.2. Some small variation can be seen starting 1990, with Germany and the United States having a slightly bigger index than France and the United Kingdom. Furthermore, for these four countries we don't observe any shocks during economic crises. On the other hand, China suffers a decrease in the trading partners density index during 1965-1976, due to Mao Zedong's Cultural Revolution that resulted in a period of economic decline. However, the index increases again during 1976-1985, probably due to economic reforms that were initiated by Deng Xiaoping which helped revive the economy. Another low point is noticed in 1990-1992 right at the Fall of Communism in USSR and Eastern Europe, global events that resulted in reforms throughout socialist states at that time, including China.

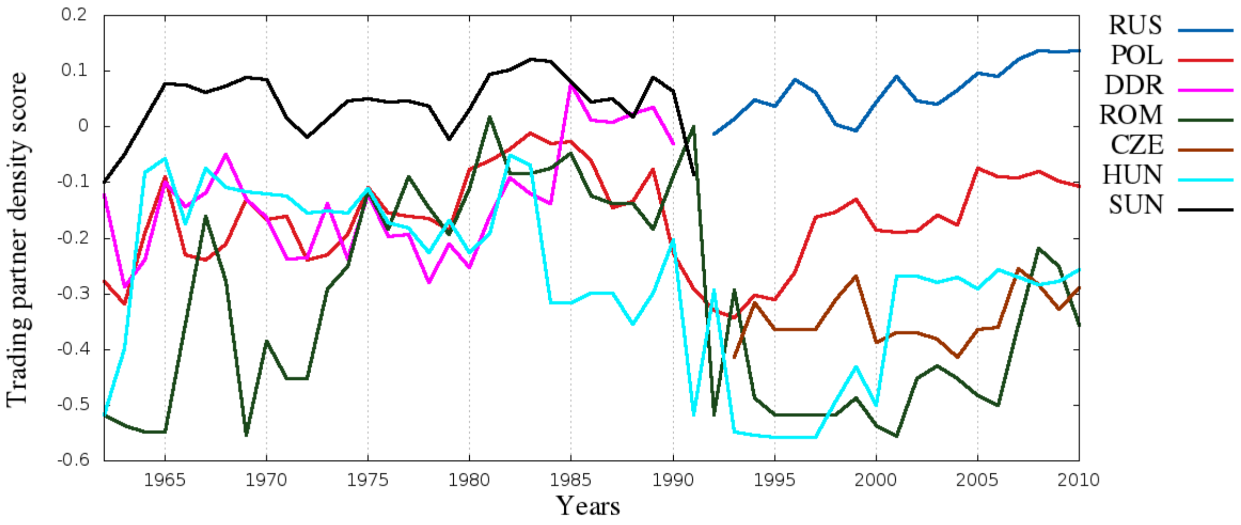


Figure 3.25: Trading partners density index measured for countries from Eastern Europe: Russia (RUS), Poland (POL), East Germany (DDR), Romania (ROM), Czech Republic (CZE), Hungary (HUN) and the USSR (SUN).

Figure 3.25 shows the *trading partners density index* for several countries in Eastern Europe. In the period leading to 1990, the USSR had the highest index since it was a world superpower, while it's sattelite states have a lower index. However, the Revolutions in December 1989 in Eastern Europe have led to a large drop in the *trading partners index*, a fact that is reflected by the economic situation of those countries at that time: unemployment skyrocketed and living standards fell considerably. It took some countries such as Poland of Hungary around aproximately 10-15 years to reach the level in the *trading partners index* that was before 1990.

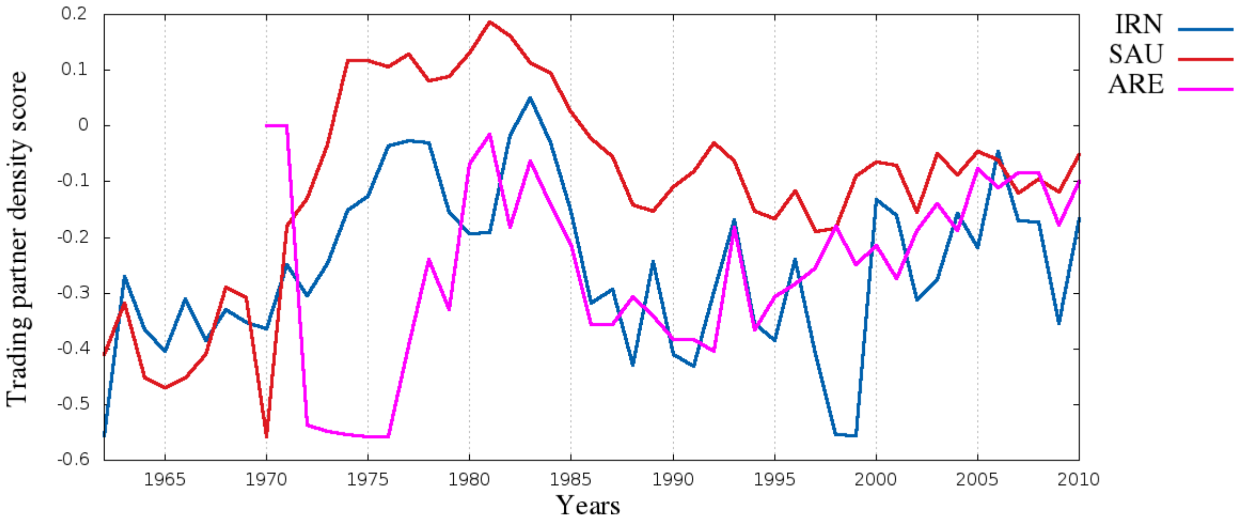


Figure 3.26: Trading partners density index measured for 3 OPEC members: Iran (IRN), Saudi Arabia (SAU) and United Arab Emirates (ARE).

Figure 3.26 shows the *trading partners density index* for three main OPEC members: Iran, Saudi Arabia and United Arab Emirates. For Saudi Arabia and Iran, the rise in petroleum prices in 1970s has led to a surge in it's index. However, during the 1980s the oil glut that was caused by a serious surplus of crude oil and a drop in demand has had detrimental effects on all OPEC members, whose economy is heavily dependent on the price of oil.

3.13 Case study: Saudi Arabia - are the trading partners of Saudi Arabia affected by changes in Crude Oil price?

As we have seen in previous sections, the GCV signature can indeed capture the changes in Crude Oil prices and correlate with key economic and social events around the world. In this section we are trying to apply the same analysis but on a smaller scale, at a country level. We have selected Saudi Arabia as a major oil-exporting country and we are trying to find the answer to the following question: *Are the partners of Saudi Arabia affected by changes in Crude Oil price?*

Saudi Arabia is the world's largest oil-exporting economy and has the largest proven petroleum reserves. It is also a very influential member of OPEC, the *Organisation of the Petroleum Exporting Countries*. It's main export partners are the United States, China and Japan while it's main import partners are China, United States and South Korea. Around 90% of it's exports consist of petroleum and related products.

We have therefore calculated the normalised GCV of Saudi Arabia for each year in the period 1962-2009. Afterwards, the change in GCV between every two years has been calculated using the Euclidean distance between the two vectors. Results of the GCV change along with the Crude Oil price are plotted in figure 3.27. The two plots are negatively correlated, having a Spearman's rank correlation coefficient of -0.32 with a p-value of 0.026, which resembles the results we got for the original GCV change for the overall trade network in section 3.7.2. Fir of all, it must be noted that since Saudi Arabia is an oil-exporting country, it benefits massively from a rise in oil prices. However, high oil prices on the energy markets lead to less demand for petrol and provides other oil-poor countries an incentive for developing alternative sources of energy. The fact that Saudi Arabia benefits from high oil prices might explain why the change in it's trading partner network topology is inversely correlated with oil price: when the price of oil is low, Saudi Arabia always looks for new export markets and thus has a move volatile network of trading partners. On the other hand, when the price of oil is high, it means that the demand is much higher than the supply available, so Saudi Arabian oil companies prefer to export to their old trading partners, since there is no need for extra contracts, negotiations and bureaucracy.

Figure 3.27 shows that big changes in the trading partners of Saudi Arabia occured in 1968/1969 and 1969/1970, which subsumed shortly afterwards. These might be explained as a consequence of the 1967 Oil Crisis, when Saudi Arabia and several Middle Eastern countries limited or completely stopped their oil supplies to Western countries such as the USA, UK and other European states. The consequence was that Saudi Arabia had to look for different export trading partners and that led to a change in the normalised GCV.

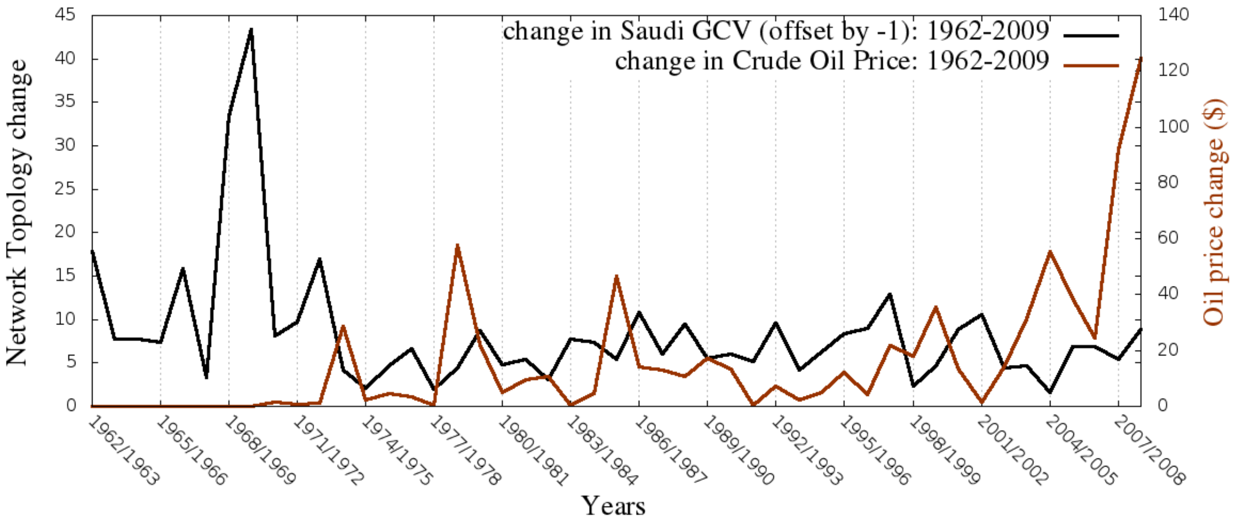


Figure 3.27: The change in the GCV of Saudi Arabia along with the change in Crude Oil price. The two plots are negatively correlated, having a Spearman's rank correlation coefficient of -0.32 with a p-value of 0.026.

In order to find out how each of the individual elements of the GCV vector are influenced by the oil price, we have applied Canonical Correlation on the following two vectors:

- X - short GCV of Saudi Arabia that only contains elements 1-8 of the actual GCV
- Y - a single-element vector containing the Oil price

Results for the CCA analysis are shown in figure. 3.28. It is shown that graphlet G3 correlates positively with the increase in oil price, while graphlets G1, G2 and G8 correlate negatively. One of the things that separates the two ends of the graphlet specturm is their density. Graphlet G3 is a sparse graphlet, while G1, G2 and G8 are dense graphlets having a density of at least 0.66.

Using the results we got earlier from section 3.10, we know that sparse graphlets are correlating with good economic indicators such as GDP per Capita (RGDPL), while dense graphlets are correlating with bad economic indicators such as Balance Current Account (BCA). Using this observation and the fact that sparse graphlets correlate positively with the oil price and dense graphlets vice-versa, we can conclude that for Saudi Arabia the good economic indicators such as GDP per Capita, a result of a healthy economy, must correlate with the Oil price[10]. This is confirmed by the fact that Saudi Arabia is an Oil-exporting economy, and it's GDP per Capita has been shown to strongly correlate with the Oil price[67]. We expect similar behaviour for other oil-exporting economies such as Lybia, Venezuela, Qatar or Russia.

[10] if the correlation of XY is strictly positive and the correlation of YZ is likewise, then the correlation of X and Z is not necessarily strictly positve. This is however the case if the correlations of XY respectively YZ are close to 1.

Canonical Correlation		0.82353	
p-value		0.00000	
X variate		Y variate	
G3	0.49265	Crude Oil price	0.83032
G6	0.09838		
G4	0.05294		
G5	0.03942		
G7	-0.23884		
G8	-0.46603		
G2	-0.50725		
G1	-0.52241		

Figure 3.28: Canonical Correlation Analysis between the short GCV vector and the price of Crude Oil

3.14 Case study: China - does economic growth affect the trading partners of a country?

3.15 Literature networks

Anna Karenina - Knuth literature

Another class of networks for which we have calculated Pearson's GCV correlation coefficient matrices is the literature networks. Fig. 3.29 shows the correlation matrix for the character network of Anna Karenina, a novel by the Russian writer Leo Tolstoy.

Along the diagnonal the following clusters of graphlets are formed:

- {2,8,29}: all these graphlets are cliques
- {28,22,7,26}: all contain as a subgraph graphlet G_7
- {9,12,5,25,13,16,27,10,12,3,19}
- {11,4,14,17}: all contain claws C4 (graphlet G_4)