Metabolic Networks

Hsa metabolic Network (Compound-based)

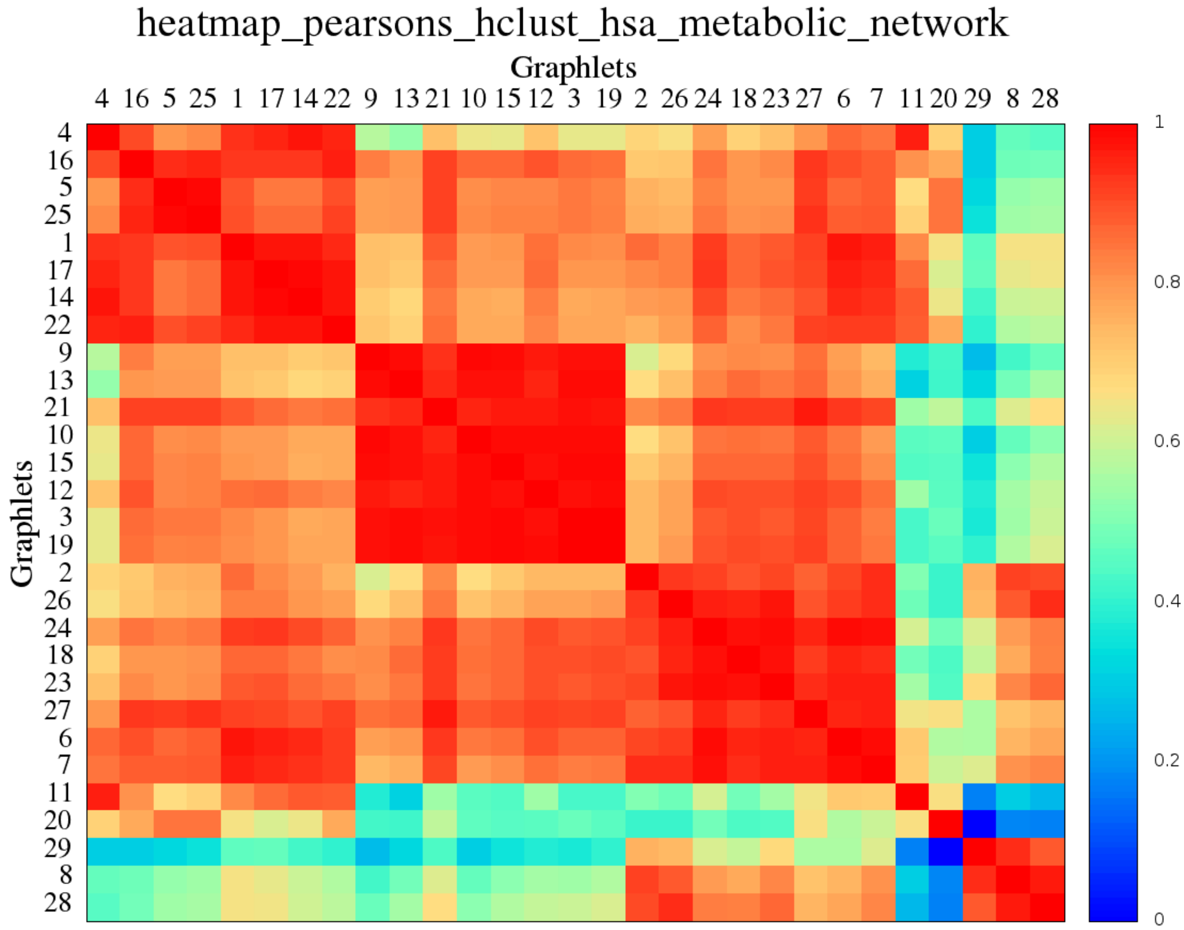


Figure 1

For the HSA metabolic network, we can clearly see several clusters of graphlets that have been formed along the main diagonal. These 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.

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.

CCA - Hsa metabolic Network (Compound-based)

canonical correlation	0.517688900136879
"sig29"	-0.34583422564784
"sig2"	-0.367846409747838
"sig8"	-0.374418098124076
"EC5"	-0.114222050692381
"EC2"	-0.126006303968814
"EC4"	-0.161439048762524
"EC1"	-0.161557005857164
"EC3"	-0.210568893756238
"EC6"	-0.406153663937205

There is some degree of correlation between the Graphlets and the EC numbers (0.517). All the cross-loadings from both the Graphlets and the EC numbers 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" (Wiki: http://en.wikipedia.org/wiki/Ligase). The reason why the magnitude of EC6 is quite high (0.4) compared to the other indicators is because the neighbourhood of the ligase compound is made of the two large molecules that have a lot of interactions and feedback loops between them. These interactions and feedback loops yield a high number of graphlets in the network topology.

Other metabolic networks

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.

Example - CCA result for E.Coli - Enzyme-based Metabolic network

canonical correlation	0.251508972397617
”sig25”	-0.0151119270013791
”sig5”	-0.0197252221535908
”sig20”	-0.0293681175027652
”EC1”	0.159458849722403
”EC6”	0.12180017345131
”EC3”	0.0537111323926065
”EC4”	9.93665948890473e-05
”EC5”	-0.0443838552104944
”EC2”	-0.21807362156073

PPI networks

Human PPI

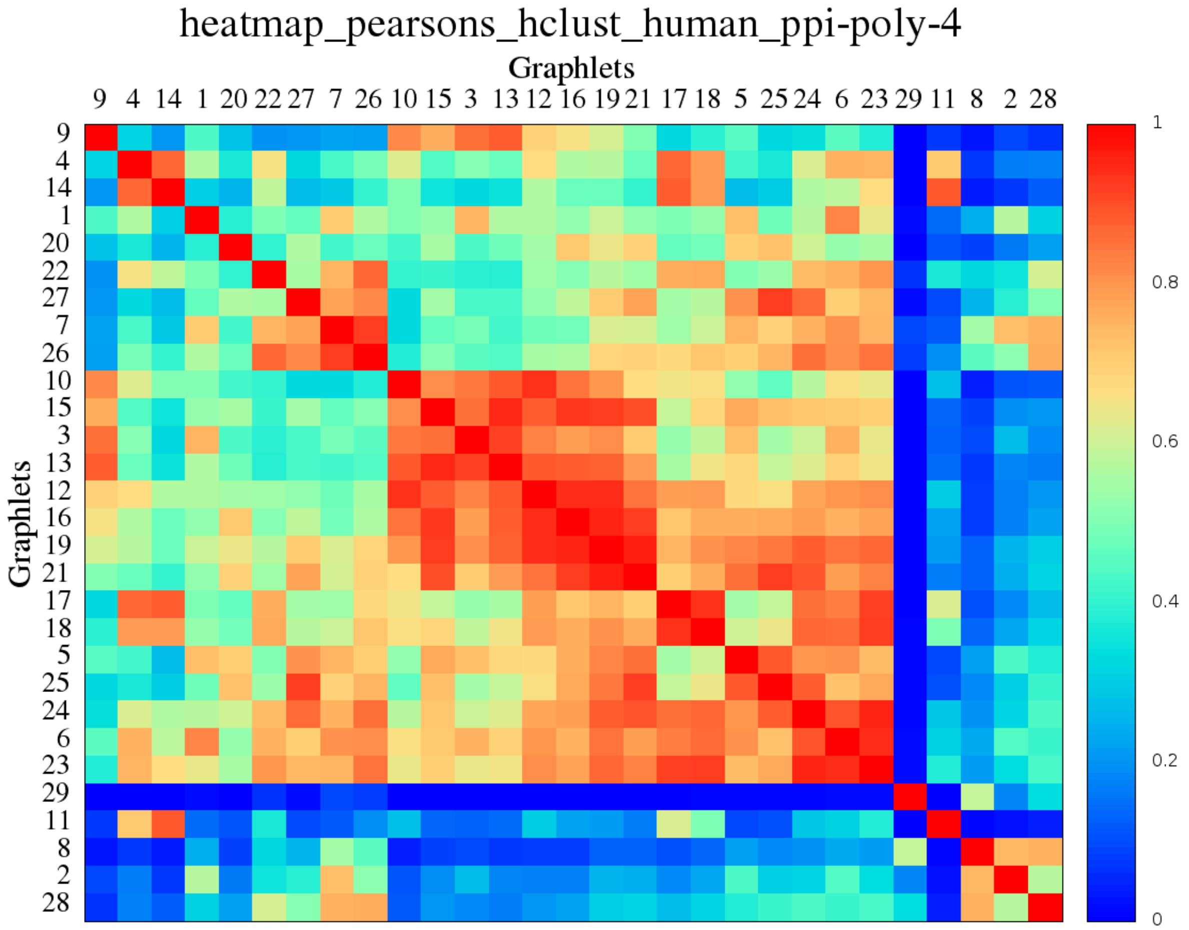


Figure 2

The above figure represents the Pearsons’s correlation heatmap for the Human PPI network. As opposed to the Metabolic network, the clusters obtained are much less noticeable. The cluster that seems most proeminent is made of Graphlets 10,15,3,13,12,16,19 and 21. These graphlets all contain a P4 (path on 4 nodes, graphlet G3). This time, it also seems that the cliques are uncorrelated.

The lack of graphlet correlations in the Human PPI is something that we cannot explain at the current time. Further research needs to be done into this area.

Human PPI CCA

canonical correlation	0.155502366337325
”sig1”	-0.0681076890184487
”sig8”	-0.071701548908936
”sig29”	-0.0721921026141832
”sig2”	-0.0816038161136185
”RNA.metabolism”	-0.065515997126898
”death”	-0.130453937016065

As it can be noticed the correlation is really low, suggesting that the number graphlet signatures doesn’t reflect the annotations of the proteins. However, the CCA results for yeast look more promising (see next subsection)

The 18 experiments

We performed CCA on 6 different human ppi networks with one annotation file and also on 6 yeast networks using 2 annotation files. The networks were as follows:

- Human - one annotation file (14) – 6 experiments in total
  1. human HI 2012 preliminary
  2. human i2d full
  3. human i2d hc
  4. human ppi 56k
  5. biogrid human ppi noUBC full (removed the ubiquitous protein)
  6. biogrid human ppi hc
- Yeast - two annotation files: boone (14) and merin (13) – 6x2 = 12 experiments in total
  1. yeast apms collins
  2. yeast biogrid genetic
  3. yeast lc
  4. yeast y2h union yu ito uetz
  5. biogrid yeast ppi noYLL039C full (removed the ubiquitous protein)
  6. biogrid yeast ppi hc

The best results have been obtained for the following yeast networks, for both annotation files:

1. yeast apms collins
2. yeast biogrid genetic (correlation of 0.35 with boone)
3. biogrid yeast ppi noYLL039C full (removed the ubiquitous protein)
4. biogrid yeast ppi hc

They have an average overall CCA correlation of 0.5 (apart from *yeast biogrid genetic*). 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) (*yeast apms collins* and *biogrid yeast ppi noYLL039C full*), transcription (mainly *biogrid yeast ppi hc*). Therefore, we can state that the proteins that are involved in translation or transcription are more likely to have a neighbourhood rich in graphlets, especially cliques.

The other combinations of networks and annotation files have yielded much poorer correlation results (only aprox 0.2). One of the reason for this might be because of the high noise of the data that is prevalent in PPI networks.

Selected CCA Results

6: Yeast Colling apms - boone

canonical correlation	0.530126834561433
”sig9”	-0.212189480521127
”sig10”	-0.23169736846479
...	...
”sig7”	-0.454313584962548
”sig29”	-0.46430615776707
”sig8”	-0.479142491460243
”sig2”	-0.499532936369011
...	...
”RNA.processing”	-0.0475609645756522
”Ribosome.translation”	-0.508987082686986

10: Biogrid Yeast PPI noYLL039C full - boone

canonical correlation	0.45880430741723
”sig11”	-0.0335008490033594
”sig14”	-0.0391537348322753
”sig10”	-0.0409595646198318
...	...
”sig7”	-0.275508344217967
”sig28”	-0.289739702115225
”sig29”	-0.299368501765292
”sig8”	-0.324296804477194
”sig2”	-0.345230081760038
...	...
”Chromatin.transcription”	-0.0283716381202604
”RNA.processing”	-0.240026232630838
”Ribosome.translation”	-0.352812642444517

17: Biogrid Yeast ppi hc - merin

canonical correlation	0.424493771565584
”sig11”	0.00615470207634573
”sig4”	0.00504623777854404
”sig14”	0.00442874775308286
...	...
”sig28”	-0.10302660841484
”sig29”	-0.11273512738831
”sig7”	-0.139579432212895
”sig8”	-0.165033816607569
”sig2”	-0.223518564656672
”X.M.”	0.0999257313876536
”X.A.”	0.0719525568889254
...	...
”X.P.”	0.0036170242868089
”X.B.”	-0.080911133899033
”X.T.” (transcription)	-0.407239925520341

Von merin encoding:

- E - energy production
- G - amino acid metabolism
- M - other metabolism
- P - translation
- T - transcription
- B - transcriptional control
- F - protein fate
- O - cellular organization
- A - transport and sensing
- R - stress and defense
- D - genome maintenance
- C - cellular fate / organization
- U - uncharacterized

Trade network

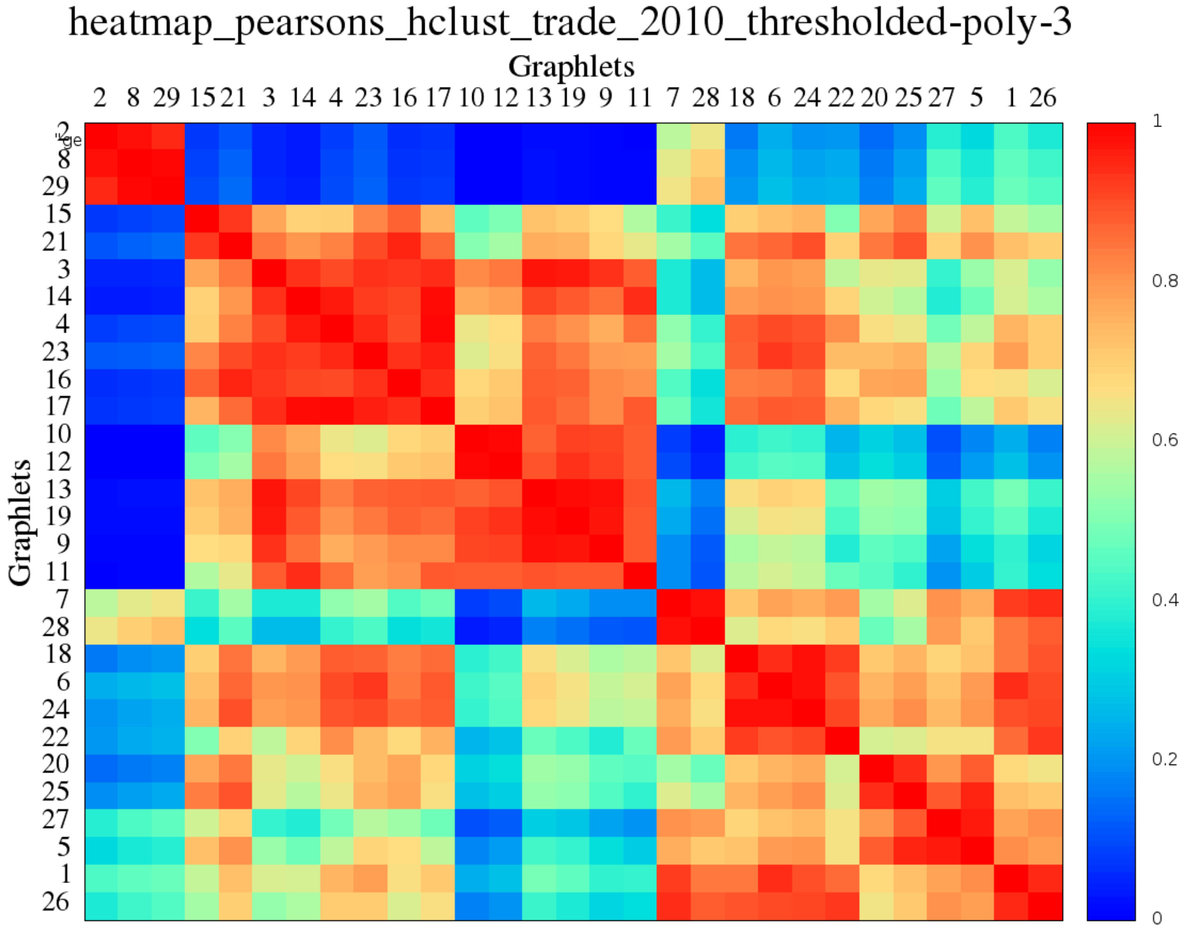


Figure 3

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

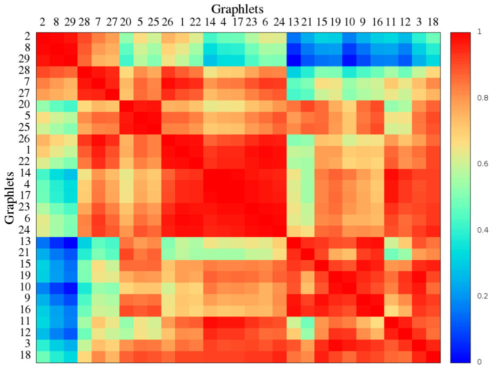
- Cliques cluster made of graphlets 2,8,29. If a country has a lot of cliques in its neighbourhood, then it is part of a densely connected group of countries.
- A cluster that is made of graphlets 15,21,3,14,4,23,16,17,10,12,13,19,9 and 11 which can be split into 2 further sub-clusters:
  - P4 cluster 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), which meant that our metric might not be suitable for analysing this type of networks.

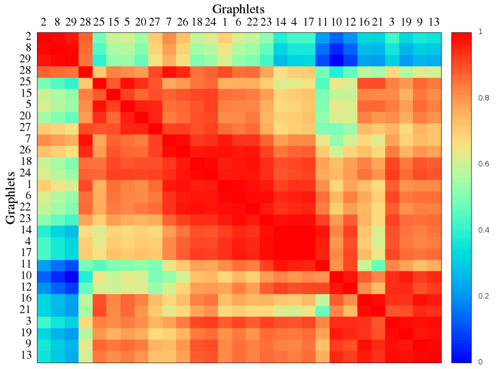
Trade network over the years

1962 to 1972

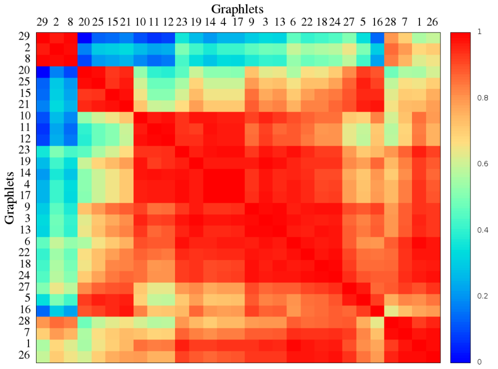
heatmap_pearsons_hclust_1962



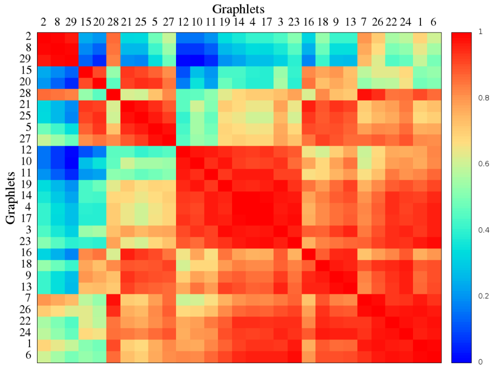
heatmap_pearsons_hclust_1964



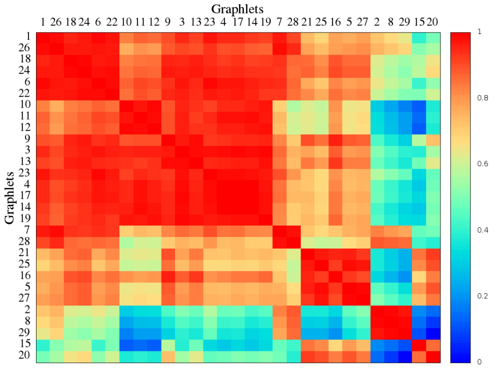
heatmap_pearsons_hclust_1966



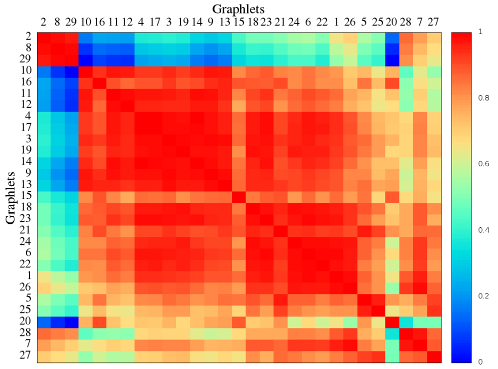
heatmap_pearsons_hclust_1968



heatmap_pearsons_hclust_1970



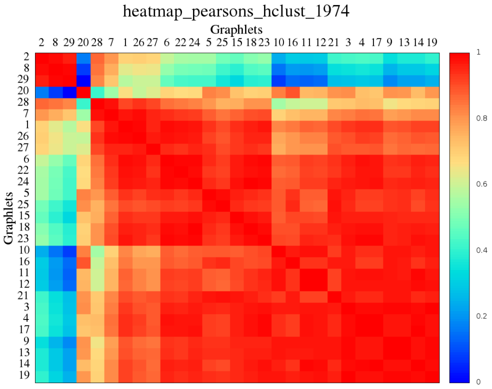
heatmap_pearsons_hclust_1972

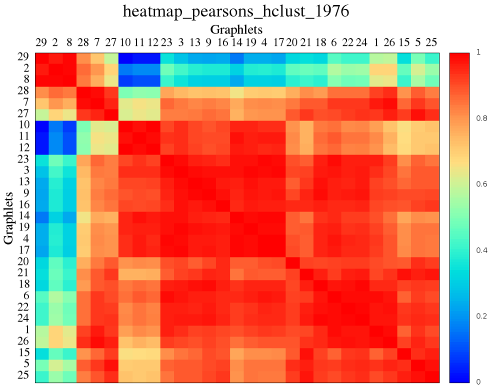


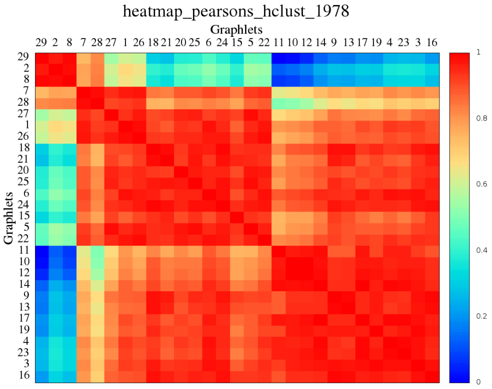
The first thing we notice is that the cliques 2,8 and 29 cluster together in each of the years analysed. In most of the years (apart from 1972) we also observe a cluster containing Graphlets that are made of cycles of length 4 (graphlet G5), with proeminent graphlets including 5, 20, 21 and 25.

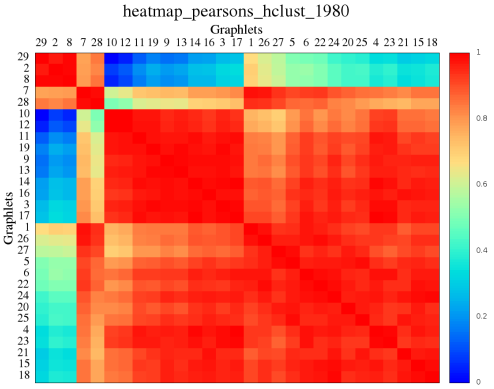
Another trend we notice is that the graphlets become more and more correlated (this will be even more obvious in the second batch of years 1974-1984). This might be an effect of globalization, as countries become more and more connected and the diameter of the trade network gets smaller. This in turn causes the countries to have share a higher proportion of the neighbourhood, which yields a higher graphlet correlation.

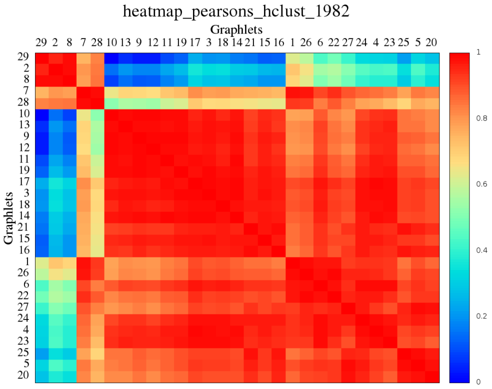
1974 to 1984

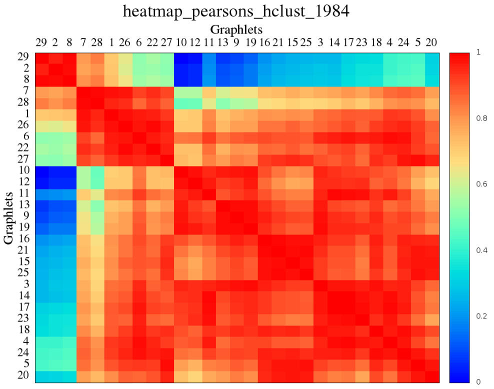












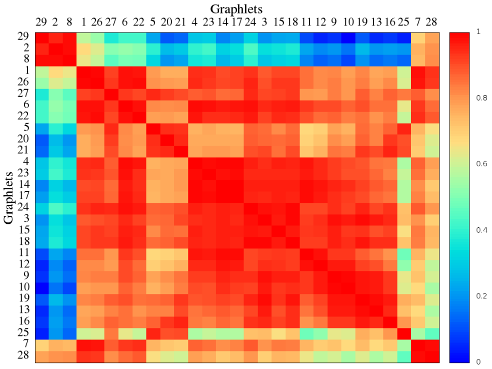
We also observe that the cliques 2,8 and 29 cluster together in every year that has been analysed in this period. In year 1980 we also observe a cluster that is made of graphlets 10,12,11,19,9,13,14,16,3 and 17. All of these contain P4's (paths of length 4, graphlet G3). This cluster on P4 can also be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

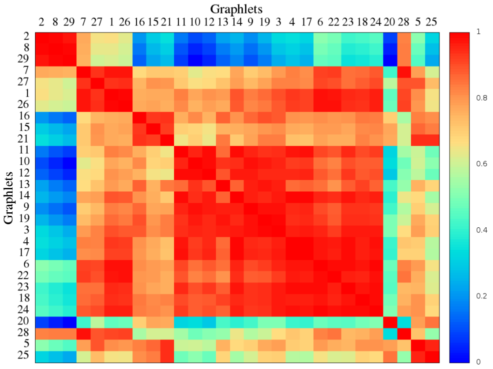
On broad terms, we also notice that the graphlets become much more correlated in this period, as a result of globalisation.

1986 to 1996

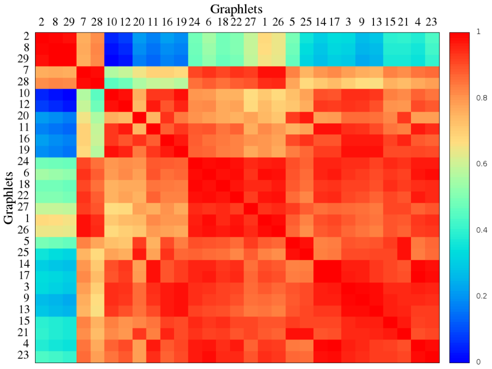
heatmap_pearsons_hclust_1986



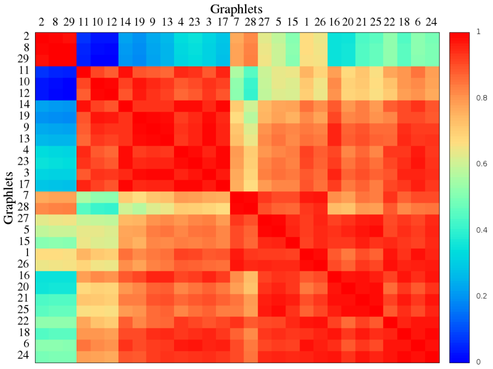
heatmap_pearsons_hclust_1988



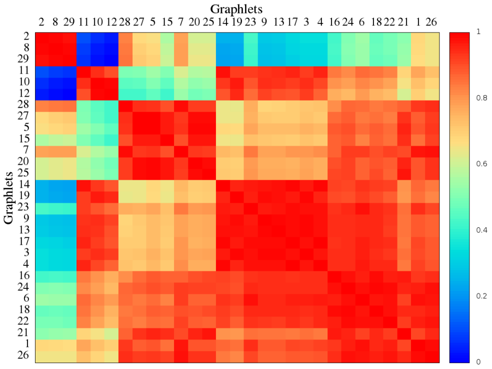
heatmap_pearsons_hclust_1990



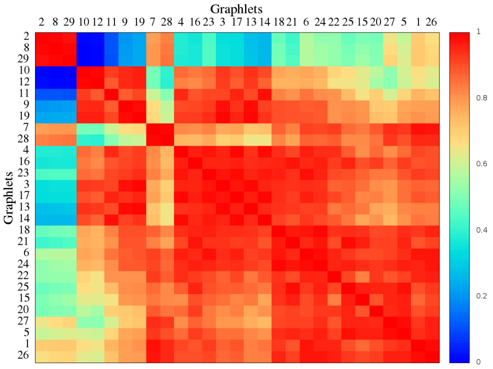
heatmap_pearsons_hclust_1992



heatmap_pearsons_hclust_1994



heatmap_pearsons_hclust_1996



Trade 1980 - 2010 - CCA results

canonical correlation	0.895952852603906
”OPENK”	0.247451850129221
”BCA”	0.200192036020745
”KG”	0.174218020898056
”BCAperRGDPL”	0.147747854581897
”KI”	0.0599886767878585
”KC”	-0.0978003022187681
”XRAT”	-0.119986658349081
”KIxRGDPL”	-0.178691495415784
...	...
”KIxRGDPLxPOP”	-0.725521757848721
”LE”	-0.758181971837348
”POP”	-0.766671415171843
”sig20”	-0.444543442451426
”sig15”	-0.575462152591753
”sig16”	-0.622031342349607
”sig25”	-0.637543533431093
...	...
”sig3”	-0.783606882407559
”sig7”	-0.793649835514637
”sig6”	-0.803147167225324
”sig2”	-0.805688869837944
”sig1”	-0.823315620830077

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” ...

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. 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:

- 0 - no economic integration
- 1 - Multilateral Free Trade Area (AFTA, CEFTA, CISFTA, COMESA, GAFTA, GCC
- 2 - Customs union (CAN, CUBKR, EAC, EUCU, MERCOSUR, SACU)

- 3 - Common market (EEA, EFTA, CES)
- 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	0.618823452163313
p-value (asymptotic Wilks):	0.0128012383082338
”Integration”	0.618823452346073
”sig29”	0.287035191288073
”sig8”	0.283489421651237
”sig2”	0.278615840865378
”sig22”	0.26819674644438
”sig28”	0.268059123234175
”sig7”	0.260207030053052
”sig26”	0.260112720674704
”sig18”	0.246614831141233
”sig1”	0.238368565474503
”sig24”	0.231333551025451
”sig6”	0.229690387810875
”sig4”	0.220206855982917
”sig27”	0.220129084450554
”sig17”	0.210208580507573
”sig14”	0.206306944610344
”sig23”	0.204204021795578
”sig5”	0.201486106029494
”sig20”	0.193906989388751
”sig25”	0.191223668928417
”sig21”	0.190762668949242
”sig16”	0.190257968164896
”sig11”	0.183299329835195
”sig3”	0.177301380839247
”sig15”	0.168139476991408
”sig13”	0.16763975453569
”sig19”	0.160509773821264
”sig9”	0.150872182634174
”sig12”	0.148861465107402
”sig10”	0.146375591601729

Change over years

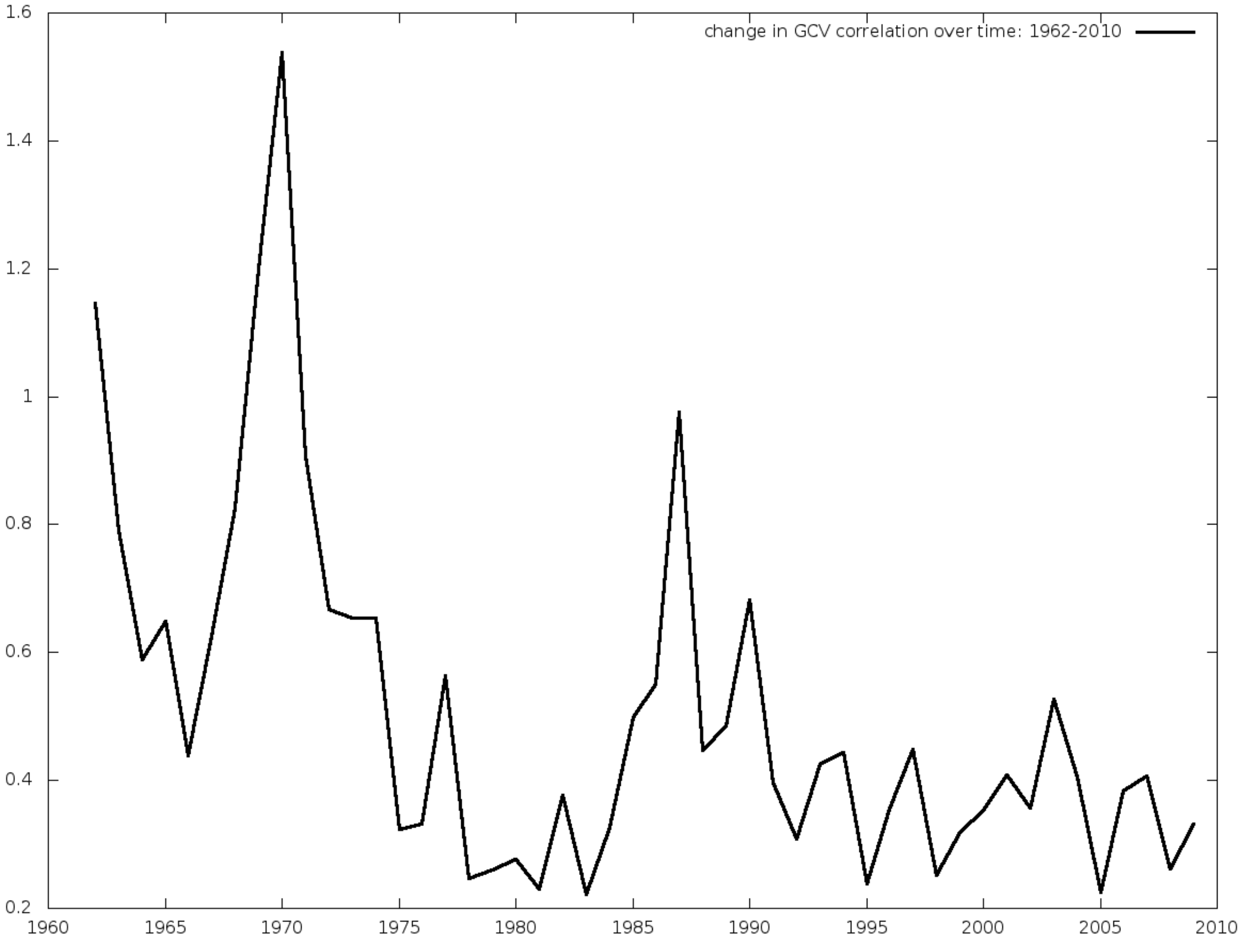


Figure 4

Important economic events that match the graph:

- Black Monday of 1987
- 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
- Arab Spring Revolts

Food and live animals - CCA results

canonical correlation	0.954399648341325
”BCA”	0.547689925367287
”BCAperRGDPL”	0.500823424204372
”OPENK”	0.243606879775848
”KI”	0.209884267951759
...	...
”KCxRGDPL”	-0.420546719870103
”KIxRGDPLxPOP”	-0.871656464696992
”LE”	-0.877434636132884
”POP”	-0.881642665100747
”KGxRGDPLxPOP”	-0.902330822089951
”RGDPLxPOP”	-0.911676654572702
”RGDPCHxPOP”	-0.911684561314991
”RGDPL2xPOP”	-0.911764842878327
”KCxRGDPLxPOP”	-0.914794036209158
”sig29”	-0.671790063754862
”sig8”	-0.713976480166582
...	...
”sig3”	-0.889319524129654
”sig19”	-0.892822123308914
”sig13”	-0.896528616555174

Minerals and fuels - CCA results

Similar results as for food and live animals.

Literature networks

Anna Karenina - Knuth literature

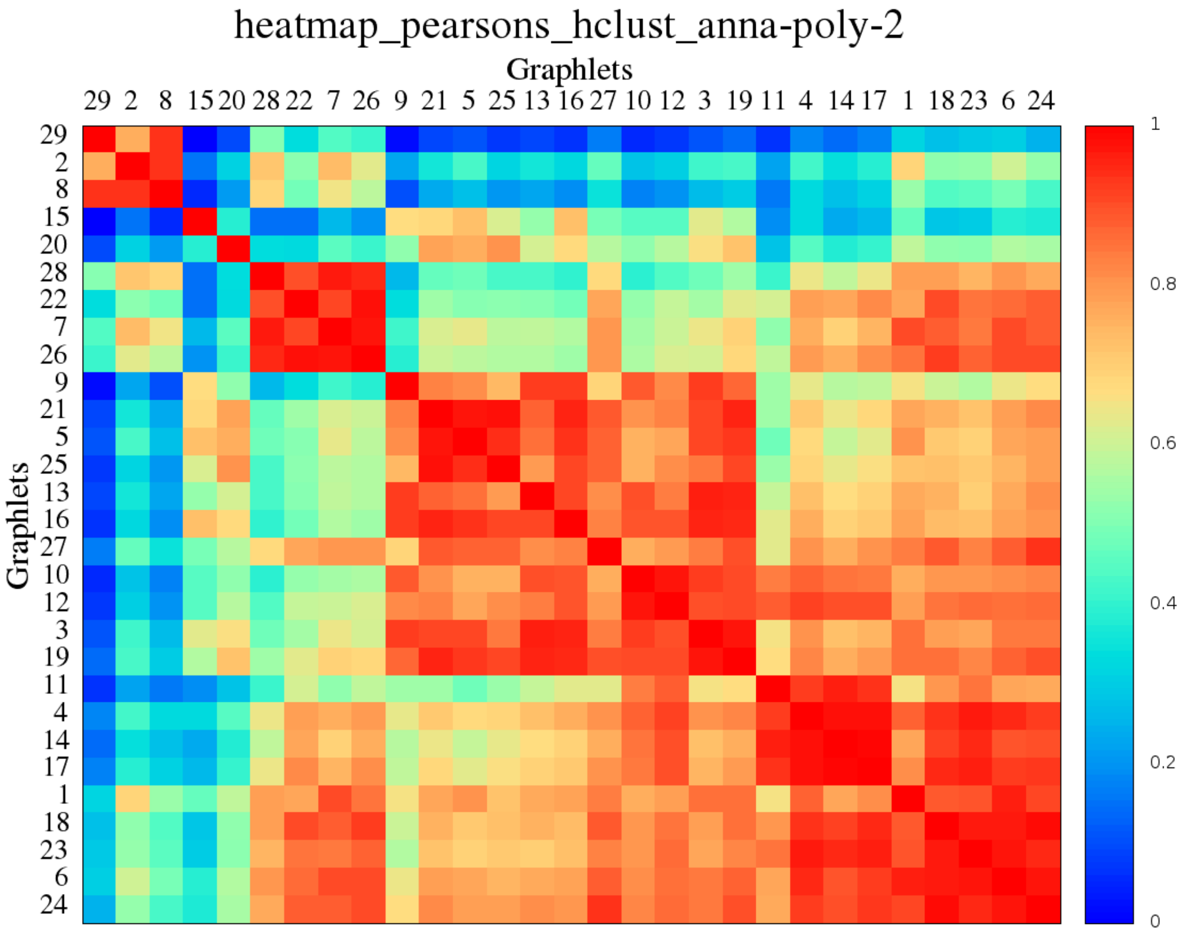


Figure 5

Cliques 2,8 and 29 cluster with each other. Graphlets 28,22,7 and 26 contain as a subgraph graphlet G7. Graphlets 11,4,14 and 17 all contain claws C4 (graphlet G4).

David Copperfield - Knuth literature

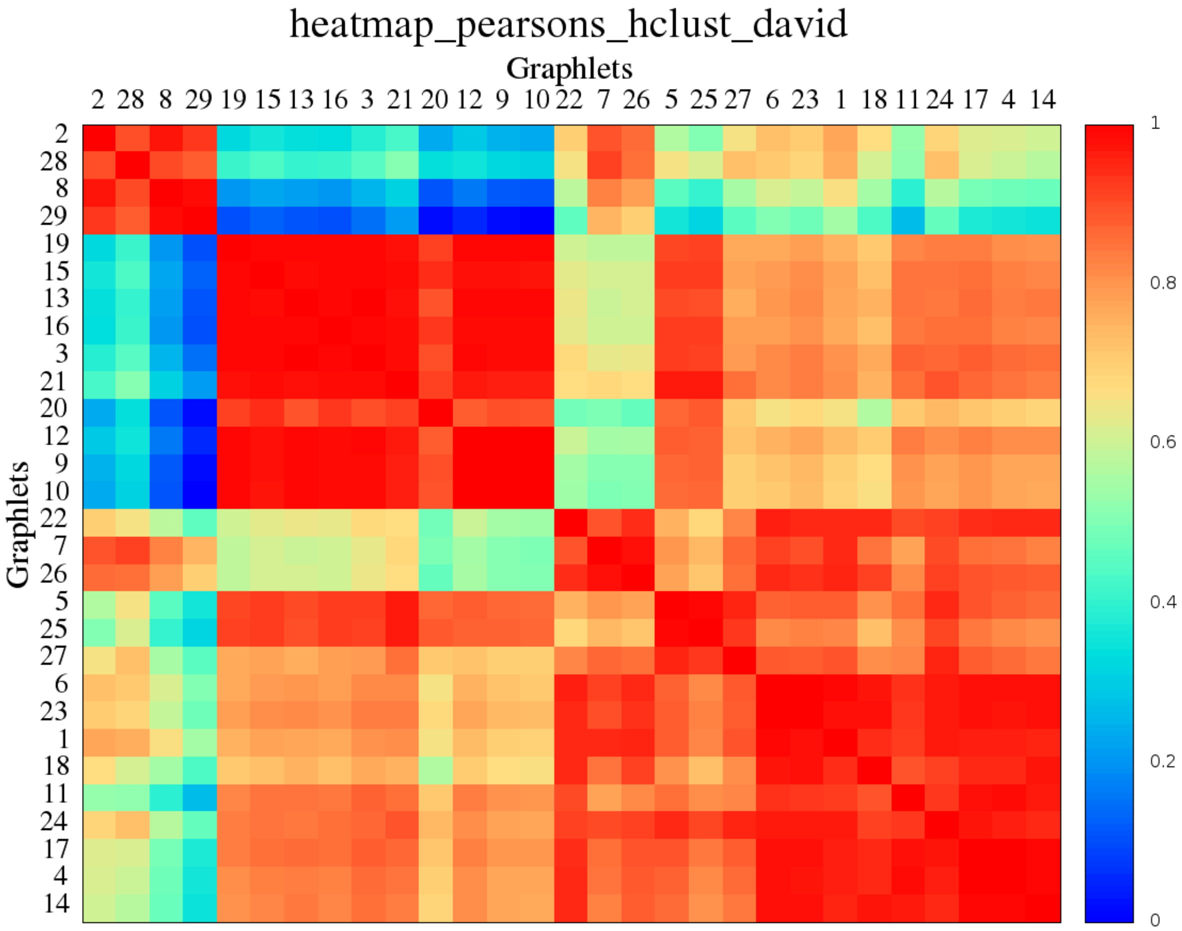


Figure 6

Results for the David Copperfield network are also similar to Anna Karenina: Cliques 2,28,8,29 cluster together. A bigger cluster is formed of Graphlets 19,15,13,16,3,21,20,12,9 and 10 which all contain paths on 4 nodes: P4's (graphlet G3).

Other literature networks for which I have calculated the heatmaps are:

- Knuth Literature:
  - Homer - Iliad or Odyssey
  - Adventures of Huckleberry Finn
  - Jean (??)
- Mreze Literature:
  - Anna Karenina
  - David Copperfield
  - Les Miserables
- Processing under way
  - Bible
  - Old Testament
  - New Testament

Conclusion

To conclude, we can clearly say that the main reason for the graphlets clustering together is because they contain as subgraphs the same smaller graphlets. Moreover, the CCA analysis that has been performed on several metabolic, ppi and trade networks has not been able to clearly separate the graphlets into sets with positive and negative weights. The best CCA correlations have been obtained with the trade network (0.9), followed by some Yeast PPI networks (0.5) and compounds-based metabolic networks (0.5). This means that a high correlation (around 90%) between economic indicators and graphlet GCV signatures can be achieved in the trade network. The reason for the high correlation in the trade network might come as a result of the small size and diameter in this network (119 nodes).

For the yeast ppi networks, a correlation of around 0.5 indicates that there is some connection between the GCV signature and the ppi annotations. The reason for this might be because the GCV signature only captures some information about the neighbourhood of the protein. Moreover, other differences between proteins might not be due to the way they interact with the environment, but due to their chemical and physical properties.