Bioinformatics@Data Science A.Y. 2017-2018

Project 2: Chagas Disease (cont.)

MATADEEN Craig[1], SELEK Aydan[1] and VUKOVIC Vlado[1]

[1]Group no. 1

Abstract

As a continuation of the previous project, we went on to gather new properties about Chagas disease. In this project our target was identifying the Chagas disease modules. For this propose, we used three clustering methods and one non-cluster-based method. Firstly, we started by obtaining the global measures of Seed Gene Interactomes (SGI), Intersection Interactome (I), Union Interactome (U). After this, we mainly dealt with the largest connected component (LCC) of the intersection and union interactomes rather than the entire list of interactomes. For our cluster-based method, we applied three kinds of clustering algorithms for I-LCC and U-LCC (simulated annealing, MCL and Louvain). For each clustered partitions, we found modules (so called "putative disease modules") in which seed genes are statistically overrepresented ($p<0.05$) by applying a hypergeometric test. Later, we carried out GO and pathway analysis for those which had more than 20 proteins between putative disease modules. Using the Netcarto tool, we found the roles of nodes only for I-LCC and drew the cartography map. Finally, we also applied a different approach to identifying disease modules, which is DIAMOnD. DIAMOnD evaluates the significance of the connections instead of the density. We used only two reference interactomes for this method and we finalised the project with GO and pathway analysis for the result of DIAMOnD.

1 Global measures of SGI, I, U, I-LCC, U-LCC

* SGI has less than 20 nodes. Our analyses will be for I and U.

* I is a single large connected network. For this reason, I and I-LCC results are same.

TABLE 1.1 Global Measures

	Number of nodes	Number of links	Number of connected components	Number of isolated nodes	Average path length	Average degree	Average clustering coefficient	Network diameter & radius	Centralization
I	130	660	1	0	2.612	8.262	0.269	4-3	0.352
U	4,683	19,840	2	0	3.137	4.410	0.267	4-1	0.353
I-LCC	130	660	-	-	2.612	8.262	0.269	4-3	0.352
U-LCC	4,681	19,828	-	-	3.137	4.411	0.267	4-3	0.353

2 Figures of the SGI and of the I networks

FIGURE 2.1: SGI Network Figure

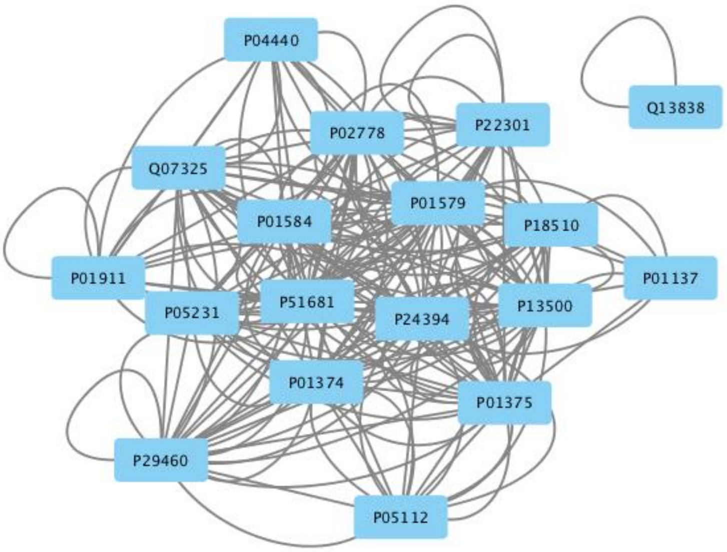
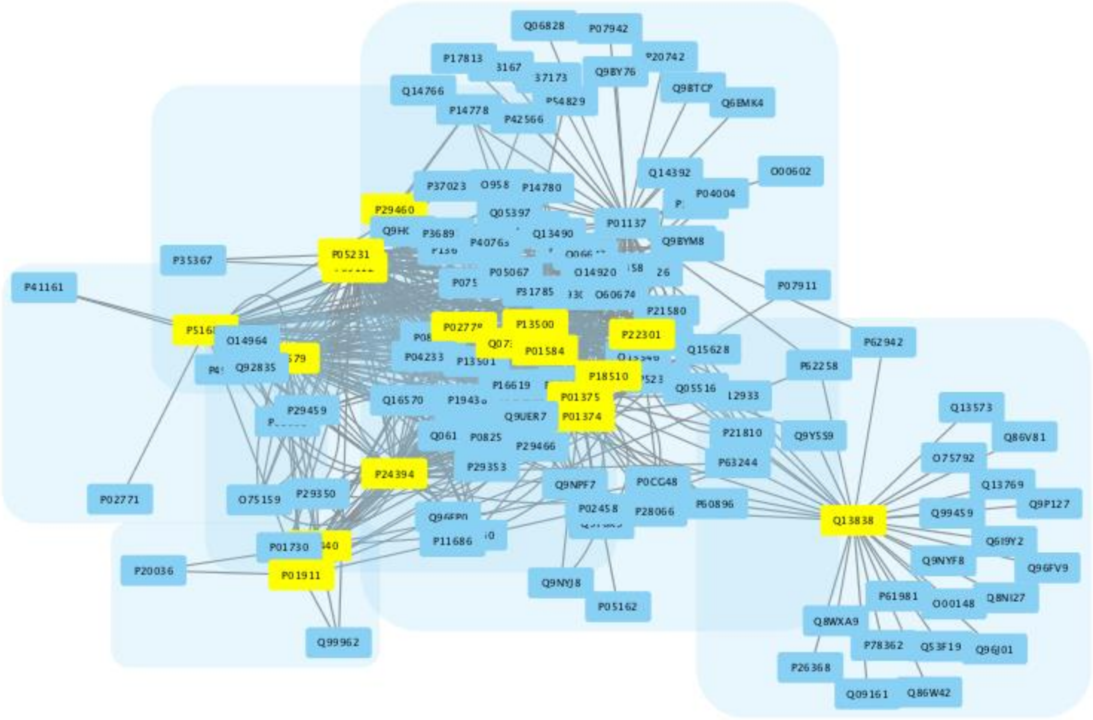


FIGURE 2.2: I Network Figure – Seed Genes Highlighted



3 The first 20 highest ranking genes for betweenness for I- and U-LCCs

TABLE 3.1 I-LCC (Local Measures with 20 highest ranking for betweenness)

Name	Degree	Betweenness centrality Normalized	Eigenvector Centrality Normalized	Closeness centrality Normalized	Node ratio betweenness / degree
Q13838	31	1.00000	0.11972	0.44725	0.03226
P01137	49	0.89380	0.59142	0.86273	0.01824
P01375	86	0.65012	1.00000	1.00000	0.00756
P24394	69	0.33600	0.87182	0.87889	0.00487
P05112	56	0.28388	0.65790	0.74249	0.00507
P51681	43	0.25504	0.79361	0.75670	0.00593
O60674	22	0.22567	0.53975	0.85475	0.01026
P01579	54	0.18371	0.88968	0.87078	0.00340
P05231	38	0.17991	0.72592	0.68779	0.00473
P01374	26	0.17320	0.51222	0.64256	0.00666
P13500	57	0.16272	0.90415	0.76388	0.00285
P01584	41	0.15700	0.78423	0.70791	0.00383
Q07325	27	0.13577	0.56598	0.60560	0.00503
P23458	16	0.13357	0.39344	0.77843	0.00835
P52333	14	0.12182	0.34774	0.77113	0.00870
Q13546	12	0.10911	0.30461	0.72158	0.00909
P22301	50	0.09249	0.71202	0.68779	0.00185
P0CG48	6	0.06997	0.13396	0.68118	0.01166
P02778	30	0.06946	0.62396	0.58183	0.00232
P01911	18	0.06533	0.27715	0.45221	0.00363

TABLE 3.2 U-LCC (Local Measures with 20 highest ranking for betweenness)

Name	Degree	Betweenness centrality Normalized	Eigenvector Centrality Normalized	Closeness centrality Normalized	Node ratio betweenness / degree
Q13838	1,799	1.00000	0.36715	0.73940	0.00056
P01137	2,307	0.54047	0.68170	0.74691	0.00023
P01375	2,208	0.43026	1.00000	0.91344	0.00019
P51681	1,748	0.25527	0.95390	0.78191	0.00015
P05112	1,316	0.23454	0.50858	0.65333	0.00018
P05231	1,201	0.13146	0.74173	0.64229	0.00011
P13500	1,214	0.11913	0.79430	0.71493	0.00010
P01579	918	0.11441	0.61197	0.78311	0.00012
P01584	1,175	0.09777	0.87689	0.70579	0.00008
P01911	715	0.09568	0.33501	0.52086	0.00013
P02778	1,186	0.09217	0.84458	0.69655	0.00008
P24394	722	0.07859	0.46142	0.73175	0.00011
Q07325	1,034	0.07360	0.74485	0.68995	0.00007
P04440	559	0.06896	0.27976	0.51126	0.00012
P18510	474	0.05098	0.44091	0.69159	0.00011
P22301	638	0.03578	0.47731	0.59669	0.00006
P01374	427	0.01989	0.35720	0.61345	0.00005
P29460	373	0.01816	0.32813	0.63986	0.00005
P16871	29	0.01178	0.14261	0.95648	0.00041
P06239	27	0.01119	0.14507	0.99489	0.00041

4 Apply clustering methods for disease modules discovery

We did not find any putative disease modules for I-LCC using MCL, Louvain or simulated annealing algorithms. We only found one putative disease module from U-LCC using the MCL algorithm.

You can find the information (including p-values) regarding all modules that were found, under the “Notes and Comments” section of this paper.

TABLE 4.1 U-LCC Putative Disease Module Information

Clustering Algorithm	Module ID	Number of seed genes in the module	Total n. of genes in each module	Ratio number of seed genes/ total genes in module	p-value
MCL	mcl_12	1	8	0.125	0.030298

5 The role cartography map

If we are to extract the significant information from the topology of a large, complex network, knowledge of the role of each node is of crucial importance. A cartographic analogy is helpful to illustrate this point. Guimera and Amaral demonstrated that they could find functional modules in complex networks, and classify nodes into universal roles according to their pattern of intra and inter module connections. The method thus yields a ‘cartographic representation’ of complex networks. The first step in the method is to identify the functional modules in the network, then to classify the nodes in the network into a small number of system-independent ‘universal roles’. For this, the simulated annealing method is used, because it maximizes the network’s modularity. Simulated annealing enables us to perform an exhaustive search and to minimize the problem of finding sub-optimal partitions.

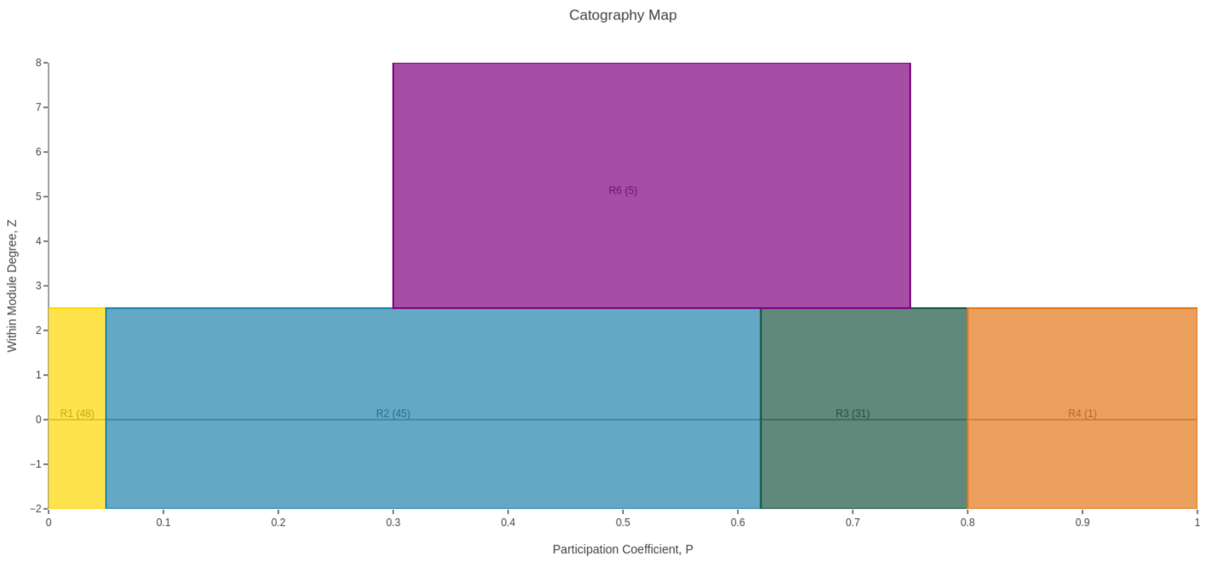
This approach predicts that the role of a node can be determined, to a great extent, by its within-module degree and its participation coefficient, which defines how the node is positioned within its own module and with respect to other modules. These two properties are easily computed once the modules of a network are known.

The within-module degree Z_i measures how ‘well-connected’ node i is to other nodes in the module. High values of Z_i indicate high within-module degrees and vice versa. The participation coefficient P_i measures how ‘well-distributed’ the links of node i are among different modules. The participation coefficient P_i is close to 1 if its links are uniformly distributed among all the modules, and 0 if all its links are within its own module. According to the within-module degree, we classify nodes with $Z>=2.5$ as module hubs and nodes with $Z<2.5$ as non-hubs.

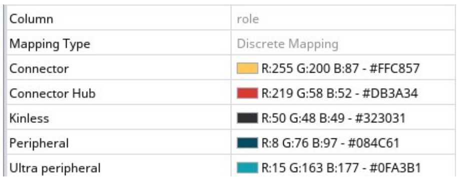
TABLE 5.1

R1	ultra-peripheral nodes (nodes with all their links within their module (P<=0.05, Z<2.5))
R2	peripheral nodes (nodes with most links within their module (0.05<P<=0.62, Z<2.5))
R3	non-hub connector nodes (nodes with many links to other modules (0.62<P<=0.80, Z<2.5))
R4	non-hub kinless nodes (nodes with links homogeneously distributed among all modules (P>0.80, Z<2.5))
R5	provincial hubs (hub nodes with the vast majority of links within their module (P<=0.30, Z>=2.5))
R6	connector hubs (hubs with many links to most of the other modules (0.30<P<=0.75, Z>=2.5))
R7	kinless hubs (hubs with links homogeneously distributed among all modules (P>0.75, Z>=2.5))

FIGURE 5.1. Cartograpy Map - Guimera and Amaral



Note: (n) in the above represents the number of nodes with respective role.



Column	role
Mapping Type	Discrete Mapping
Connector	🟨 R:255 G:200 B:87 - #FFC857
Connector Hub	🟥 R:219 G:58 B:52 - #DB3A34
Kinless	⬛ R:50 G:48 B:49 - #323031
Peripheral	🟦 R:8 G:76 B:97 - #084C61
Ultra peripheral	🟦 R:15 G:163 B:177 - #0FA3B1

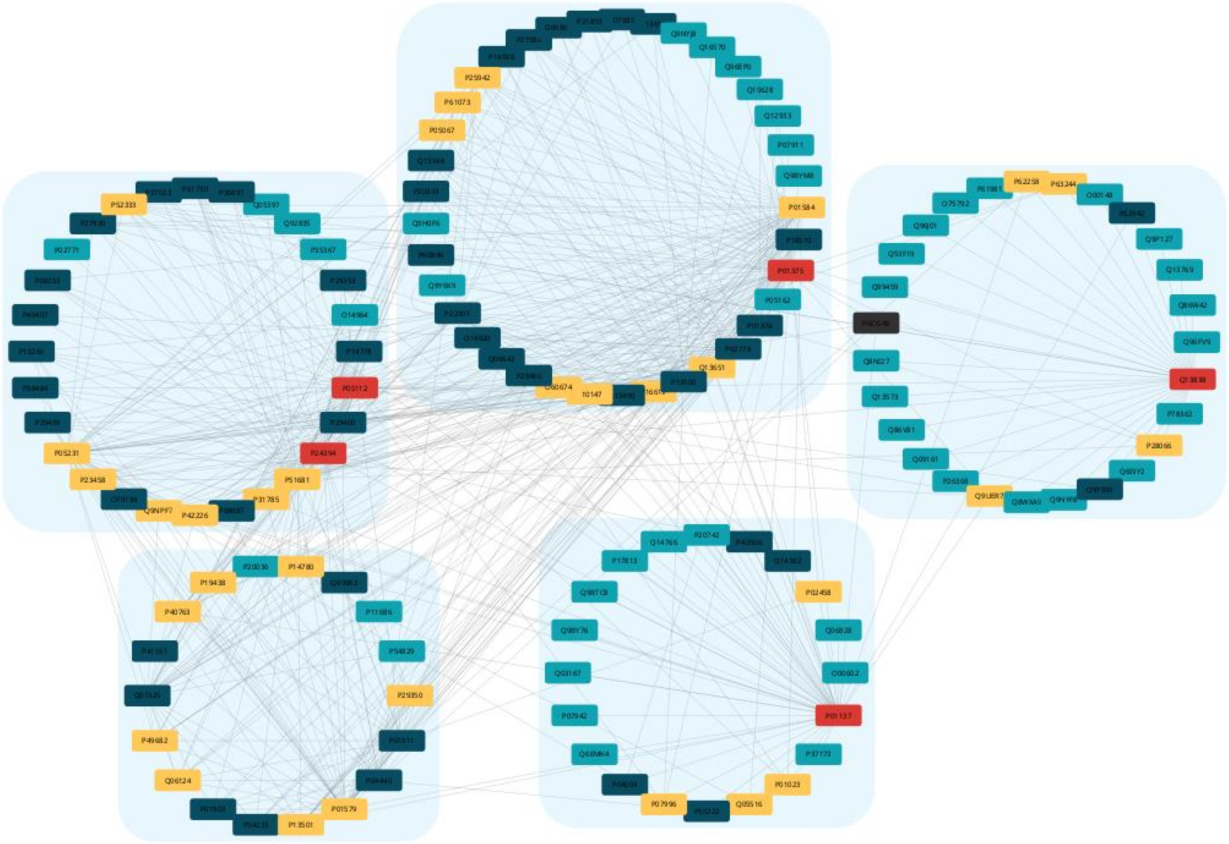


FIGURE 5.2. Cartography Map via Cytoscape

6 DIAMOnD

TABLE 6.1 The first 40 genes coming from the DIAMOnD tool for each of the two reference interactomes, and the intersection of the two putative disease proteins list.

APID	BIOGRID	INTERSECTION
P51677	P13501	P20849
P10147	P49682	P09486
P13236	P51677	P55268
O00590	P16619	Q13438
P16619	P21810	P02462
P32246	P13611	Q7Z4W2
P80075	P02776	Q8NBL1
Q16627	P48061	P53708
P13501	P32246	P47992
P80098	P80098	P20916
Q2F862	O00590	Q8IVL6
Q16570	P80075	P12110
Q9NPB9	P07585	Q92791
Q99616	P15502	P25067
P41597	P35555	P16619
P51671	P98095	Q13683
P49682	P55001	P01023
P27487	Q96GW7	P0C862
P08254	P08253	P08185
O15467	P20742	Q9BZ76
P32302	P01023	P08253
O14625	P61812	P06756
P13611	O00602	P21810

P98066	P14778	P49682
P02776	P02458	Q15262
P10145	P07996	P39060
P48061	P14780	Q16635
O00585	P35442	Q9BXJ5
O43927	P20908	Q02809
P34015	P02462	Q9Y215
Q99731	P04085	P51677
P46092	P12109	Q86YD3
O15444	P20916	P14780
Q9NRJ3	P20849	P13501
P24766	P02461	Q02388
Q9Y4X3	P01127	Q05707
O57300	P02452	Q8IYK4
Q61581	P09486	P54802
P78556	Q02809	P07585
P47992	P08572	Q6P9A2

TABLE 6.2 Overrepresented GO categories for the intersection list joined with the seed genes list.

Pathway Name	Pathway ID	Source Name	Gene Count	Genes for this entity	Pathway p-value	Pathway p-value (corrected)
ATP binding	GO:0005524	molecular function	1	1,493	0.99959	1.00000
DNA binding	GO:0003677	molecular function	1	2,108	0.99999	1.00000
cytosol	GO:0005829	cellular component	3	2,642	0.99992	1.00000
nucleic acid binding	GO:0003676	molecular function	1	1,162	0.99758	1.00000
regulation of transcription, DNA-templated	GO:0006355	biological process	1	1,898	0.99996	1.00000
transcription, DNA-templated	GO:0006351	biological process	1	1,938	0.99997	1.00000
nucleus	GO:0005634	cellular component	6	5,730	1.00000	1.00000
Golgi apparatus	GO:0005794	cellular component	1	765	0.98019	0.98543
endoplasmic reticulum membrane	GO:0005789	cellular component	1	727	0.97583	0.98255
mitochondrion	GO:0005739	cellular component	3	1,411	0.97603	0.98200

TABLE 6.3 Overrepresented Pathways for the intersection list joined with the seed genes list.

Pathway Name	**Pathway ID**	**Source Name**	**Gene Count**	**Genes for this entity**	**Pathway p-value**	**Pathway p-value (corrected)**
Pathway Name	Pathway Id	Source Name	Pathway uploaded gene count	Genes in InnateDB for this entity	Pathway p-value	Pathway p-value (corrected)
EGFR1	15908	NETPATH	2	472	0.90522	0.90522
Innate Immune System	17476	REACTOME	3	563	0.84864	0.85205
Metabolism	19429	REACTOME	10	1,535	0.82957	0.83626
Metabolism of lipids and lipoproteins	16920	REACTOME	4	554	0.67386	0.68204
Class I MHC mediated antigen processing & presentation	19282	REACTOME	2	256	0.62537	0.63554
Adaptive Immune System	18371	REACTOME	5	604	0.55773	0.56911
TGF_beta_Receptor	15911	NETPATH	2	220	0.54231	0.55565
Endocytosis	4386	KEGG	2	214	0.52731	0.54250
Gastrin-CREB signalling pathway via PKC and MAPK	13219	REACTOME	2	212	0.52224	0.53951

7 Notes and comments

All files and scripts are included in the project folder. The following is a description of some of the included files and their function:

- diamond.sh – this is shell file for generating the diamond output.
- rnetcarto.R - this is an rstudio/r file for generating the netcarto output via R.
- bioinformatics_project_2.ipynb is an Ipython notebook to run Louvain via python and for other functions implemented via Python.
- Various CytoScape files for different network analyses.

Part 1:

- SGI has less than 20 nodes. Our analyses will be for I and U.
- I is a single large connected network. For this reason, I and I-LCC results are same.

With regards to question 3, because none of the clustering methods resulted in a module with more than 20 genes and a P value < 0.05 we could not obtain overrepresented GO categories and pathways for the putative disease modules.

Part 4:

TABLE 7.1 I-LCC All Module Information

Clustering Algorithm	Module ID	Number of seed genes in the module	Total n. of genes in each module	Ratio number of seed genes/ total genes in module	p-value
MCL	mcl_1	8	77	0.10390	0.8666
MCL	mcl_2	1	24	0.04167	0.9709
MCL	mcl_3	1	13	0.07692	0.8386
MCL	mcl_4	3	8	0.37500	0.0616
MCL	mcl_5	2	4	0.50000	0.0761
MCL	mcl_6	2	4	0.50000	0.0761
LOUVAIN	L_0	3	17	0.17647	0.4291
LOUVAIN	L_1	7	43	0.16279	0.3766
LOUVAIN	L_2	1	15	0.06667	0.9072
LOUVAIN	L_3	6	33	0.18182	0.2855
LOUVAIN	L_4	1	22	0.04545	0.9728
ANNEALING	SA_0	5	28	0.17857	0.3371
ANNEALING	SA_1	4	18	0.22222	0.2205

ANNEALING	SA_2	7	37	0.18919	0.2159
ANNEALING	SA_3	1	27	0.03704	0.9893
ANNEALING	SA_4	1	20	0.05000	0.9610

TABLE 7.2 U-LCC All Module Information (except putative disease module shown TABLE 4.1)

Clustering Algorithm	Module ID	Number of seed genes in the module	Total n. of genes in each module	Ratio number of seed genes/ total genes in module	p-value
MCL	mcl_1	2	2,344	0.0009	0.9999
MCL	mcl_2	3	721	0.0042	0.5380
MCL	mcl_3	3	636	0.0047	0.4489
MCL	mcl_4	1	301	0.0033	0.6974
MCL	mcl_5	2	235	0.0085	0.2269
MCL	mcl_6	1	155	0.0065	0.4543
MCL	mcl_7	1	81	0.0123	0.2695
MCL	mcl_8	1	71	0.0141	0.2404
MCL	mcl_9	1	49	0.0204	0.1724
MCL	mcl_10	1	43	0.0233	0.1529
MCL	mcl_11	1	37	0.0270	0.1330
LOUVAIN	L_0	3	643	0.0047	0.4582
LOUVAIN	L_1	1	292	0.0034	0.6870
LOUVAIN	L_2	1	1,308	0.0008	0.9973
LOUVAIN	L_3	4	633	0.0063	0.2187
LOUVAIN	L_4	3	446	0.0067	0.2424
LOUVAIN	L_5	2	339	0.0059	0.3786
LOUVAIN	L_6	3	469	0.0064	0.2670

LOUVAIN	L_7	1	551	0.0018	0.8955
ANNEALING	SA_0	1	751	0.0013	0.9573
ANNEALING	SA_1	4	511	0.0078	0.1251
ANNEALING	SA_2	6	865	0.0069	0.0981
ANNEALING	SA_3	1	399	0.0025	0.7995
ANNEALING	SA_4	1	1,394	0.0007	0.9983
ANNEALING	SA_5	5	761	0.0066	0.1554

References

1. Guimera and Amaral (2005) "Functional cartography of complex metabolic networks", NICO and Department of Chemical and Biological Engineering, Northwestern University.

2. Python-Louvain for Louvain clustering: https://github.com/taynaud/python-louvain

3. Simulated Annealing & Netcarto – rnetcarto: https://github.com/cran/rnetcarto