Lab5: Finding community structures

Jakub Šalagovič
Simon Van den Eynde

November 24, 2016

1 Introduction

This report consists of 2 parts. In the first part we will compare some of igraph's community-finding algorithms on specifically chosen graphs, using the metrics: Triangle Partition Ratio (high is best), expansion (low is best), conductance (low is best) and modularity (high is best). In the second part we will analyse if it makes sense to use fast-greedy community detection algorithm of provided Wikipedia network.

2 Task 1 – Comparison of community-finding algorithms

2.1 Results

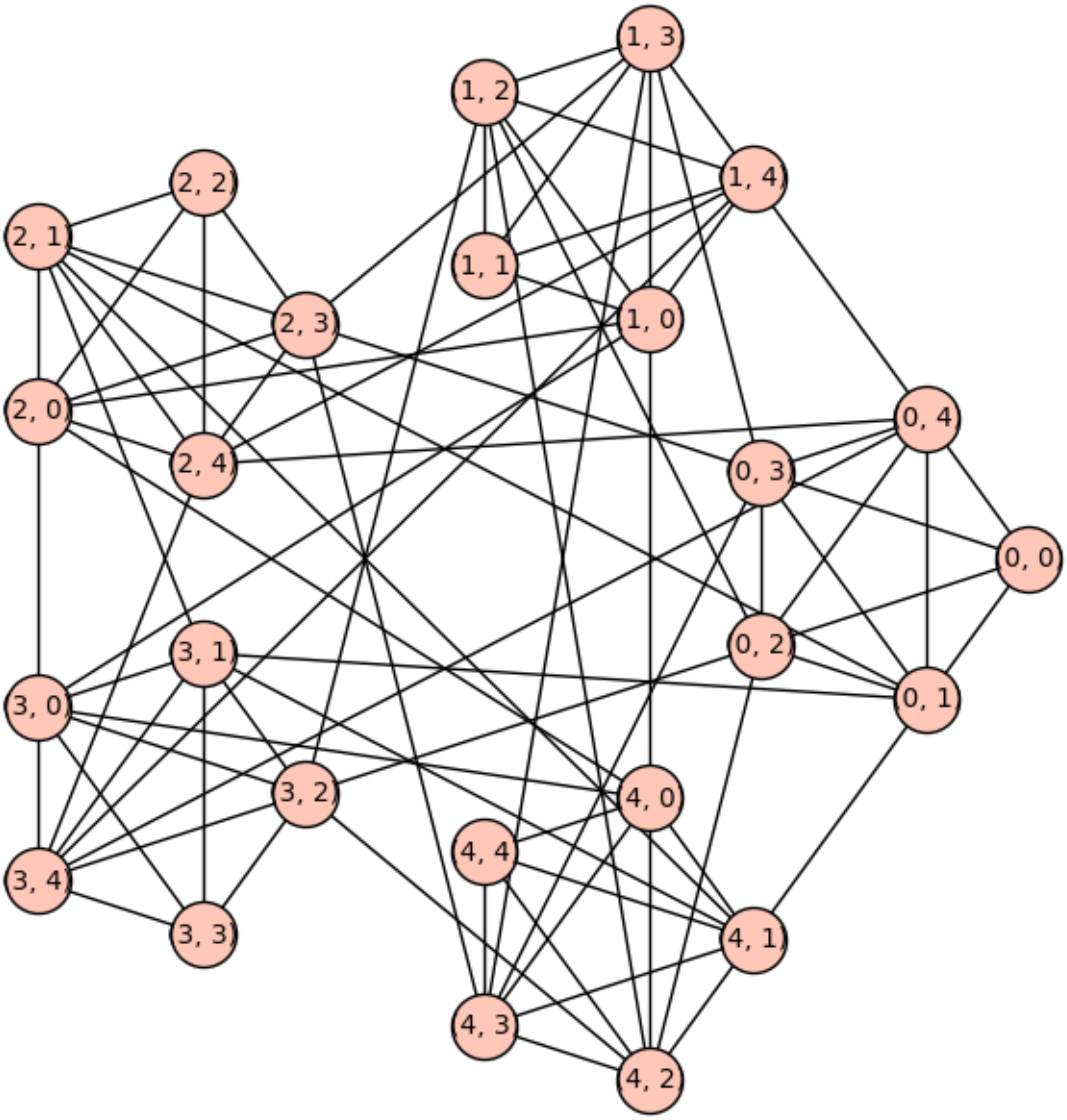

Figure 1: The Hanoi Tower graph with 5 pegs and 2 discs. The labels on the vertices indicate the positions of the two discs ($(1, 3)$ means: peg 1 on disk 1, peg 2 on disc 2)

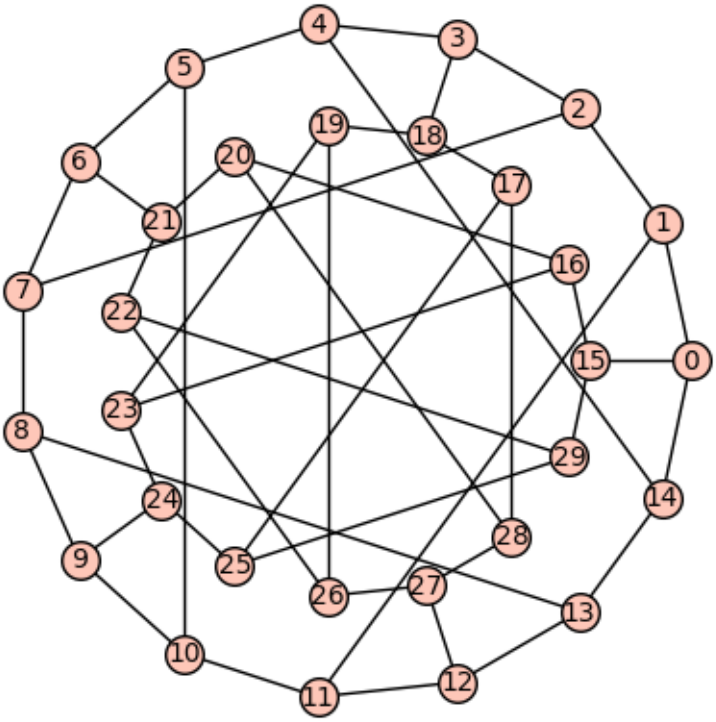

Figure 2: The double star snark

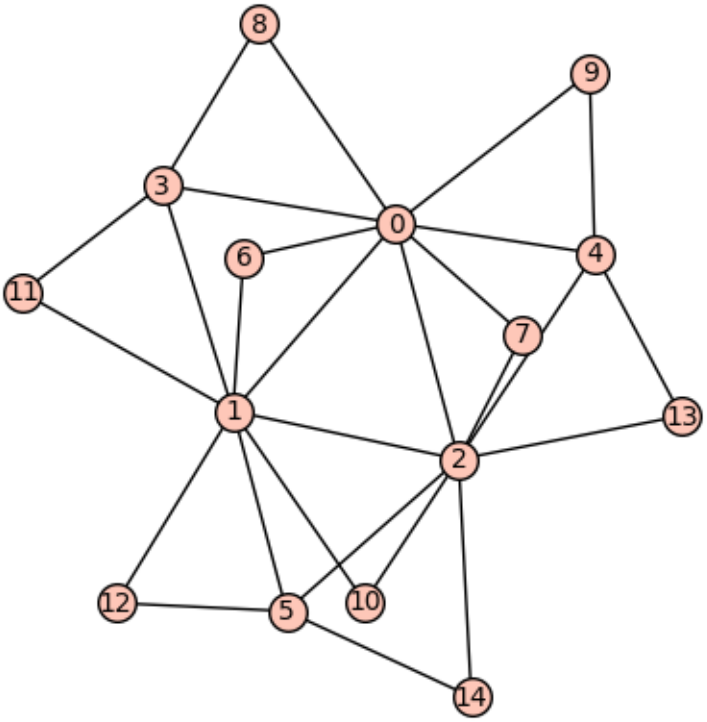

Figure 3: The Dorovstev-Goltsev-Mendes graph after 3 iterations

	TPT	expansion	conductance	modularity
edge.betweenness	1.000	4.400	0.524	0.425
fastgreedy	1.000	4.400	0.524	0.425
label.propagation	1.000	4.160	0.482	0.340
leading.eigenvector	1.000	3.200	0.333	0.000
multilevel	1.000	4.400	0.524	0.425
optimal	1.000	4.400	0.524	0.425
spinglass	1.000	4.400	0.524	0.425
walktrap	1.000	4.400	0.524	0.425
infomap	1.000	4.400	0.524	0.425

Table 1: Metrics for HanoiTower(5,2) (HT)

	TPT	expansion	conductance	modularity
edge.betweenness	0.000	1.933	0.476	0.451
fastgreedy	0.000	2.000	0.501	0.411
label.propagation	0.000	1.800	0.430	0.420
leading.eigenvector	0.000	1.667	0.385	0.389
multilevel	0.000	1.933	0.476	0.451
optimal	0.000	1.933	0.476	0.451
spinglass	0.000	2.067	0.527	0.442
walktrap	0.000	1.933	0.476	0.451
infomap	0.000	1.867	0.456	0.416

Table 2: Metrics for Double Star Snark (DSS)

2.2 Discussion

For more information on the chosen graphs, see section 2.3.

- In general we notice that the modularity sometimes becomes zero, this happens only if the entire network is one community. In this case we can discard these findings.
- In graphs without triangles (BA, DSS) or with many triangles (HT) we find that that the Triangle Partition Ratio does not convey any useful information.
- We see that for if there are clear communities as in HT, almost every community-finding method can identify them correctly.
- When looking at the metric values for the DSS network, we see that the leading.eigenvalue gives rise a special community structure. Its expansion and conductance are a lot lower than for other graphs, which is good. But its modularity is also a lot lower, which is a bad sign. When comparing network sizes (not included) we note it is the only community structure which consists of only 2 communities. So

	TPT	expansion	conductance	modularity
edge.betweenness	0.600	2.467	0.529	0.228
fastgreedy	0.400	2.733	0.609	0.217
label.propagation	1.000	1.800	0.333	0.000
leading.eigenvector	1.000	1.800	0.333	0.000
multilevel	0.600	2.600	0.565	0.222
optimal	0.600	2.467	0.530	0.239
spinglass	0.400	2.600	0.577	0.230
walktrap	0.400	2.600	0.575	0.181
infomap	1.000	1.800	0.333	0.000

Table 3: Metrics for the Dorovtsev-Goltsev-Mendes(3) Graph (DGM)

	TPT	expansion	conductance	modularity
edge.betweenness	0.000	1.075	0.381	0.680
fastgreedy	0.000	1.100	0.393	0.678
label.propagation	0.000	1.200	0.451	0.614
leading.eigenvector	0.000	1.075	0.381	0.680
multilevel	0.000	1.100	0.393	0.678
optimal	0.000	1.075	0.381	0.680
spinglass	0.000	1.100	0.393	0.678
walktrap	0.000	1.125	0.406	0.652
infomap	0.000	1.125	0.406	0.670

Table 4: Metrics for Barabasi-Albert (BA)

depending on how many communities you want, you might consider using different metrics our different community-finding methods.

- The values for the conductance and expansion of the BA network are lower than that of other networks. Also the modularity is higher. This means that this power-law delivers stronger communities than many other graphs.

2.3 Methods

2.3.1 Metrics

Here we will discuss why we want a high value for some metrics, and a low value for others.

Triangle partition ratio This metric measures the number of nodes in a community belonging to a triangle. The higher this number is, the more vertices belong to triangles, and generally the more connected a community is. So for good communities we want this high.

	TPT	expansion	conductance	modularity
edge.betweenness	0.794	3.000	0.497	0.401
fastgreedy	0.824	2.853	0.454	0.381
label.propagation	0.882	2.706	0.427	0.338
leading.eigenvector	0.735	3.059	0.502	0.393
multilevel	0.794	2.912	0.468	0.419
optimal	0.794	2.912	0.468	0.420
spinglass	0.794	2.912	0.468	0.420
walktrap	0.588	3.235	0.545	0.353
infomap	0.882	2.706	0.419	0.402

Table 5: Metrics for Zachary's Karate network (ZK)

Expansion Measures the number of edges leaving the cluster per node. So if this number is high, there are a lot of connections from this community to other communities. However we prefer if the communities are very clear and thus that there are almost no edges going out. So we want this number low.

Conductance This metric measures the fraction of total edge volume that points outside the cluster. If more edges go out, this number will increase and if there are less internal edges this number will increase as well. So we want the conductance to be low.

Modularity Measures the difference between the number of internal edges in a community and a community with the same degree sequence. The higher this number, the more internal edges it has, compared to a random model. So we want modularity to be high.

2.3.2 Graphs

We chose 5 different graphs. As a real network we chose Zachary's karate (ZK) network. We also analysed a Barabási-Albert (BA) graph on 40 vertices. Furthermore we considered 3 special graphs:

HanoiTower(5,2) (HT) This graph has as nodes the game states of Hanoi Tower game with 5 pegs and 2 disks, there is an edge if you can go from one game state to another in one move. This graph basically consists of 5 copies of K_5 which are sparsely connected. See figure 1.

Double Star Snark (DSS) This is graph is snark, which means it doesn't contain any triangle (this graph does only contain cycles with length > 5), it is also cubic and has no bridges. It consists of 30 vertices. See figure 2.

Dorovstev-Goltsev-Mendes graph (DGM) This is a graph which can be constructed, starting with K_2, as follows: for every edge a triangle (add a vertex and 2 edges). If

we do this 3 times we get a graph with 15 vertices, 27 edges and a lot of triangles. See figure 3.

3 Task 2 - Wikipedia

For the task we have chosen to use the fast-greedy algorithm for community detection of this given network. Then we computed the metrics discussed in the previous part, the results are as follows: modularity = 0.745, expansion = 2.9, conductance = 0.368, TPR = 0.46 (all values have been rounded to three decimal places). We can argue that modularity in this case is relatively high, which points to good division into communities. On the other hand, we have 2.9 edges per node leaving community which can be either a high or a normal value for good division. This depends on a lot of factors, for example average number of edges per node and so on. Conductance, which is fraction of total edge volume that points outside community, with value 0.368 can be higher than one expects for good division into communities, but it still does not mean the division into communities is not good. The same is valid for the triangle partition ratio.

Therefore we had to look if the dividing into communities made sense in different ways. We have used three different approaches, each of them is presented below.

3.1 Titles of each community

The first approach is to look at titles (vertices) in communities and decide if it makes sense to have them in one community (if we can find some common topic of other connection by meaning by titles in one community). Note that for this type of analysis one needs to have some background knowledge in the analysed terms. Therefore we skipped cases with terms we were not familiar with and, for example stated in this text, we have selected the simple ones. Also it is impossible to manually analyse all communities, so we decided to look just for biggest and analyse the first n titles. The result for communities containing at least 50 titles can be found in the file *WikiTitleCommunities.txt* attached to this document.

For better visualisation and getting general view we decided to plot these results as text clouds. Every text cloud contains 50 selected titles from communities that contain at least 50 titles. The size of each title depends on the number of vertices the corresponding vertex is connected to. The text cloud for a sample of 50 titles of the biggest community is displayed in the Fig. 4. We can say that almost all titles seem to be more or less related. Generally speaking, it looks like for each of the plotted text clouds we can find a common topic or category to which they belong. All word clouds can be found in file *WikiTextClouds.pdf* attached to this document.

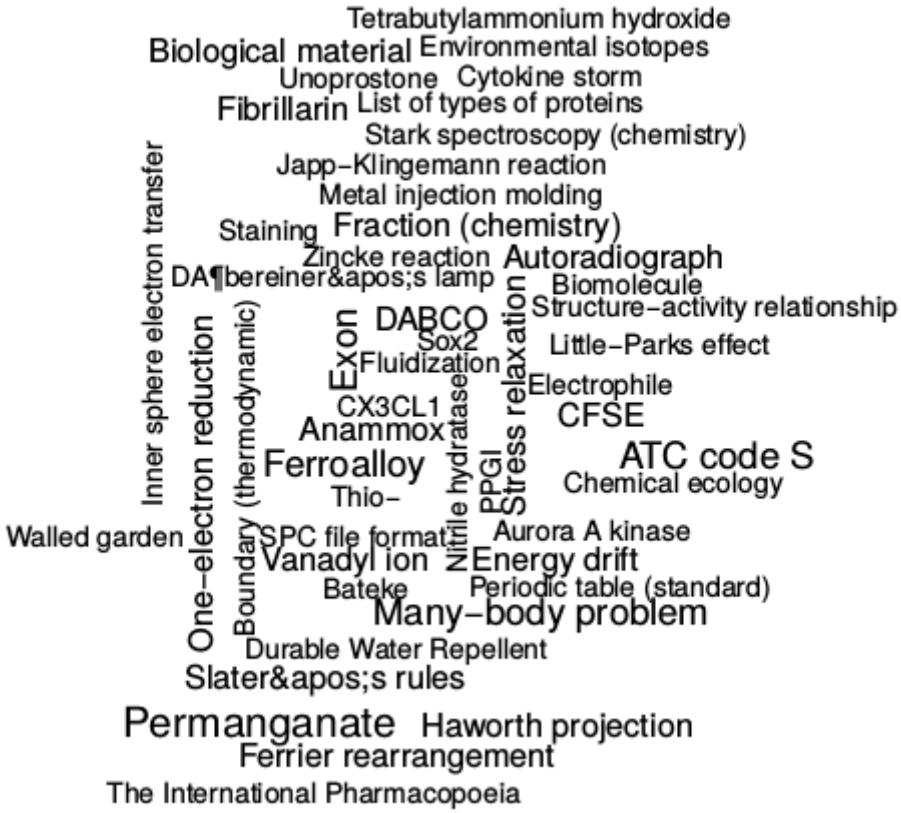

Figure 4: Text cloud of sample of 50 titles of Community 1

It is normal that there are cases when some titles looks completely unrelated to the community they were assigned to. For example in Fig. 5 we can see that titles seem to be number related. But for example titles *French punk* and *Statue of Liberty Bike* clearly don't belong to this category. Therefore we explored this whole community and have found that apart of the *number-related* articles there are smaller groups of *motorbike* and *punk* related articles. It is normal that we do not get perfect communities, but the results in general make some sense. Aside of the algorithm one can also question connections made by authors and editors of articles.

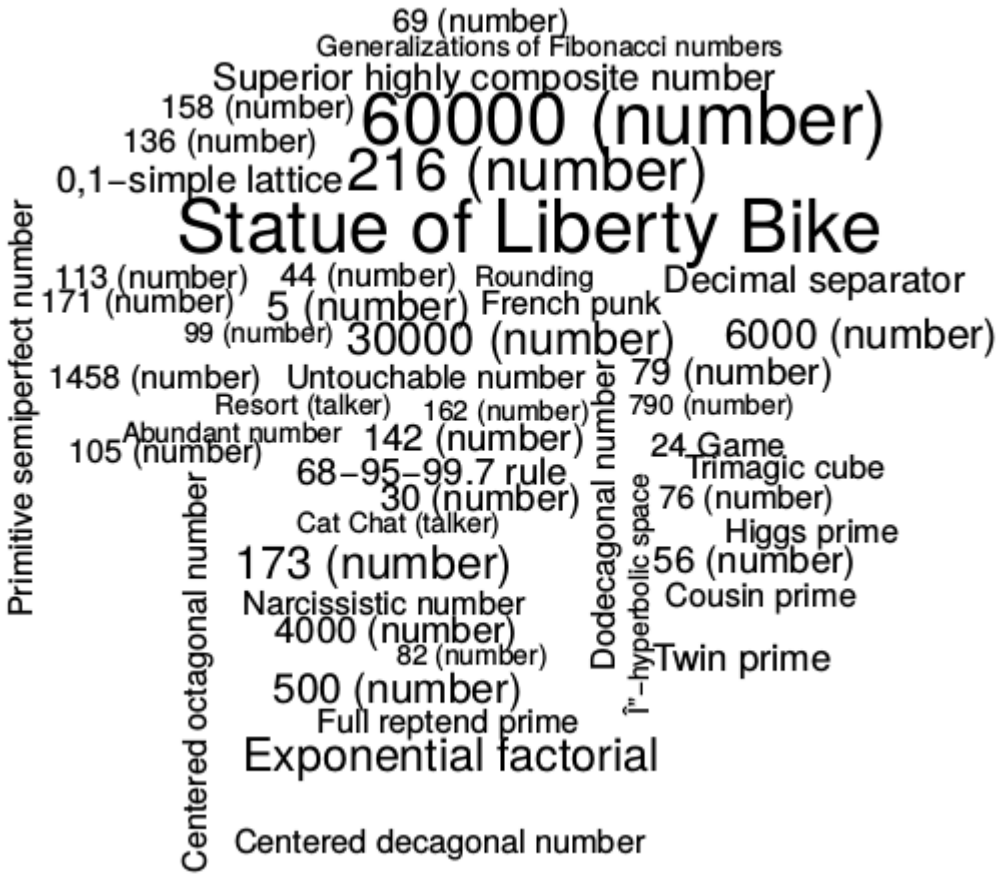

Figure 5: Text cloud of sample of 50 titles of Community 8

3.2 Membership of neighbours

The next approach is to look at all neighbours of a selected vertex and analyse if all of them are in a community that makes sense regarding to the selected vertex and his community. For each vertex we plotted this subgraph displaying corresponding communities in one figure. To make it easy to analyse we decided to use a star layout with the corresponding vertex in the middle and its neighbours around it. For the reasons mentioned earlier it is impossible to analyse results for all vertices, so we decided to plot just the vertices that have 20 neighbours. The 20 neighbours is a value large enough to find some general behaviour, but it is still easy to visualise and analyse. All graphs for vertices with number of neighbours equal to 20 can be found in file *WikiNeighboursOf20.pdf* attached to this document.

In the Fig. 6 we have a central node with title *Strictly non palindromic number*. Most of its neighbour are numbers, assigned to same community as central vertex. As we can

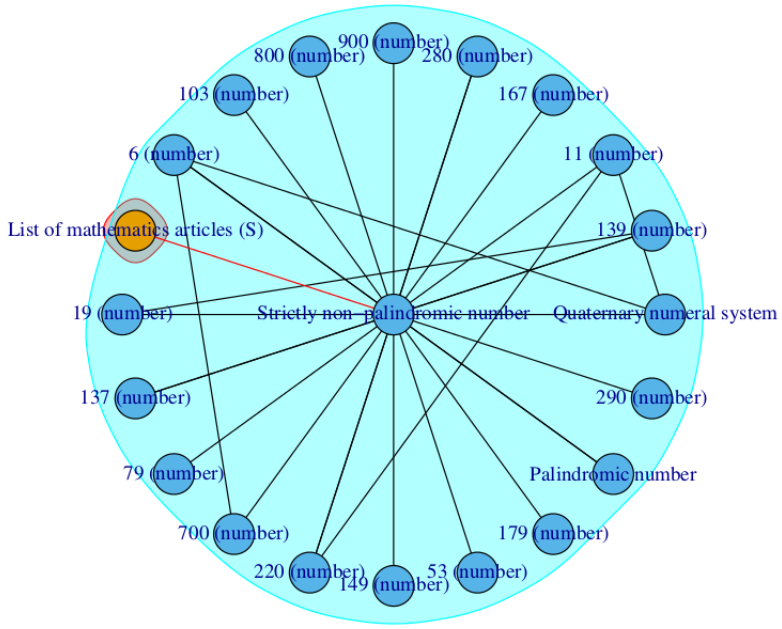

Figure 6: Graph of neighbours of selected vertex and their community membership

see, only the title *List of mathematics articles (S)* is in another community as a central vertex, which in this case makes sense. Also the membership of other titles looks correct, the only questionable thing is membership of title *Quaternary numeral system*.

Fig. 7 perfectly represents the situation when a title can be related to more fields. In this example the central vertex *Conjugation* has several meanings in fields such as linguistics, mathematics, biology, chemistry...[1] Most of them are also present in our plot and their division into communities looks reasonable. The green node belongs to linguistics, blue ones to mathematics, and the orange ones to chemistry/biology. This particular case will increase the value of the *expansion*-metric, but the division into communities seems logical. This probably happens with all titles that have more than one meaning or are not field specific and therefore could have a connection to a wide spectrum of other titles. For example, almost in every article on Wikipedia we can find some number and therefore make a connection to it. The question is wether this makes sense for a community-structure.

For most plots of the selected sample that we were able to analyse by this approach, division into communities done by fast-greedy algorithm looks like it generally makes sense.

3.3 Neighbours in different community

In this approach we have been looking for the neighbours of each vertex that are not in the same category as the selected vertex. We will look at some examples here, all results

[1] https://en.wikipedia.org/wiki/Conjugation

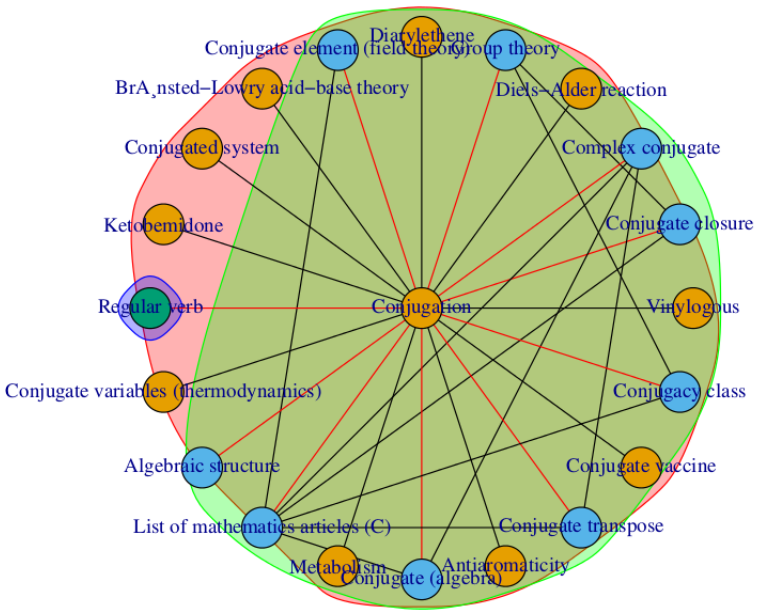

Figure 7: Graph of neighbours of selected vertex and their community membership

can be found in the file *WikiNeighbours.txt* attached to this document.

As with the previous ones, also with this approach we can find situations when division into communities done by fast-greedy algorithm on this data does not get the best results. For example, the title *120 (number)* has the neighbour *119 (number)* which is in a different community. This is probably not the result we had expected. However there may be some criteria, which may not be clear just after doing that simple inspection of results, for which keeping these titles in different communities makes sense (for example because of less edges between odd and even numbers). We also cannot assume that the communities are at the same level of generality or specificity when looking at the titles within one community. For example for one community the thing that have all titles in common can be *biology* without possibility to make it more specific, for another it can be *positive numbers divisible by 6*. A lot also depends on the way the links between Wikipedia articles are being made. This process is not uniform for all articles, just because of the way Wikipedia is being made. The same article edited by different people can result, among other, in different links between articles.

But for most of the articles we picked, the analysed neighbours in different communities make sense. For example titles *64 (number)* and *Lowrider* definitely should not be in one category which fast-greedy algorithm recognised well. A question you could ask is if the reader of the article about lowriders is interested in knowing more about the number 64.

3.4 Conclusion

We have shown that the community detection of this dataset done by fast-greedy algorithm is probably not ideal, but generally makes sense. For example, it could be used when writing new article to give suggestions for categories to which article belongs (categories can be found down at the end of each article).

We have also discussed some problems with this community detection method on this data. One of them is that process of writing and editing articles at Wikipedia is not uniform and depends by the person that write the article and make connections between articles. Currently Wikipedia has approximately 50 policies and uses bots to do automated editing of articles[2]. Therefore there is an effort to make the articles in the similar way.

The next fact is that we do not have information when these networks have been made, but some links between articles currently cannot be found (for example link to number 64 in the article about lowriders mentioned before). It could be interesting to make same network with the current state of Wikipedia and compare the results.

[2]https://en.wikipedia.org/wiki/Wikipedia