Homework #5: Network of Diseases

Anže Pečar (63060257)

January 11, 2013

1 Introduction

In homework #5 we build a network of diseases from the OMIM database. We use a network clustering algorithm by Raghavan et al and the python networkx library.

2 Data

We get the list of Diseases and their genes from the OMIM morbidmap file. Because some records are very similar we use some simple heuristic to group them.

We constructed the graph following this simple principle: if two diseases share a gene their corresponding graph nodes should be connected. As seen in the Figure 2 the graph is not scale free. The diameter of the largest connected component is 12.

3 Methods

To cluster similar diseases we used the network clustering algorithm with label propagation by Raghavean et al. The algorithm is actually quite simple. You choose a random node in the graph, and find it's neighbours. Once you have all the neighbours you rename your original node with the most common label from the neighbours. If there is more than one label tied to be the most common you choose a random one.

4 Results

In both Figure 1 and Figure 2 we could not use $pylab.scale(''log'')$ as the graph would be drawn in a weird way. Our solution is not perfect as the y axis values can be misleading.

As seen in Figure 3, the networkx library didn't do the best job with the spring layout (even though we tried using as much as 3000 iterations). Nevertheless it can be seen how at least some of the similar disease nodes tend to stick together.

Figures 4 5 and 6 show all the extracted clusters "Breast cancer", "Deafness", and "Diabetes mellitus". I am not an expert, but it seems to me that the clustering algorithm did a decent job. We can see that the algorithm grouped some other cancers together with Breast cancer which seems to show that clustering has indeed worked.

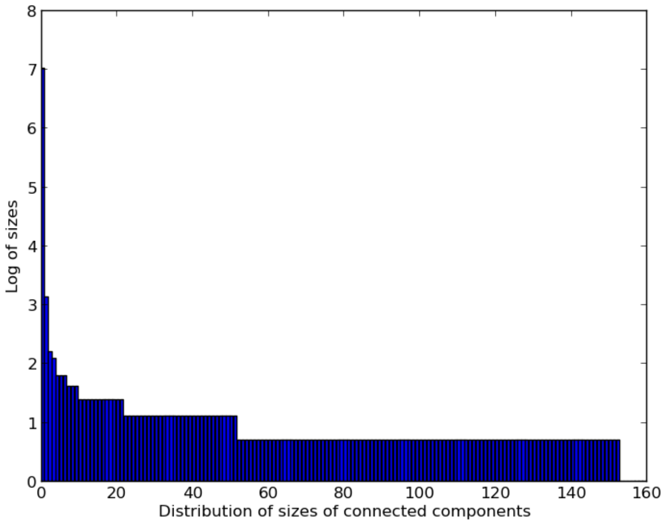

Figure 1: Distribution of sizes of connected components

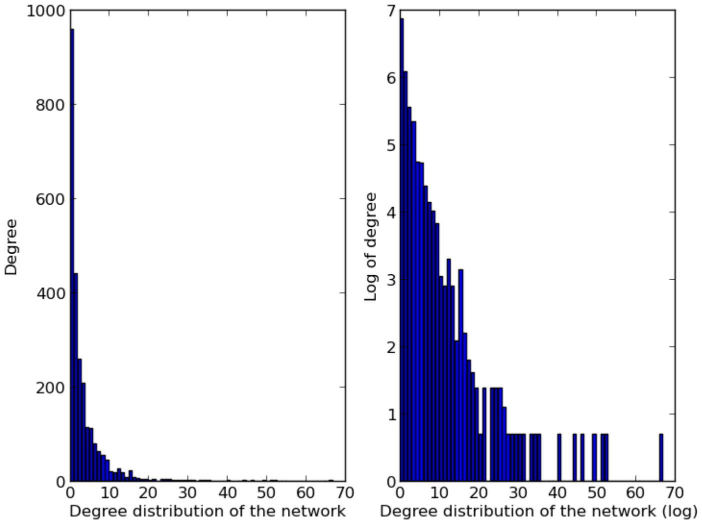

Figure 2: Degree distribution of the network with and without log

Honor Code

My answers to homework are my own work. I did not make solutions or code available to anyone else. I did not engage in any other activities that will dishonestly improve my results or dishonestly improve/hurt the results of others.

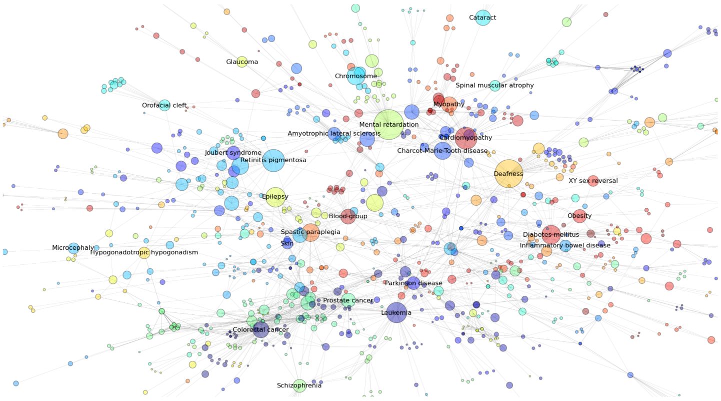

Figure 3: Largest connected component of your network

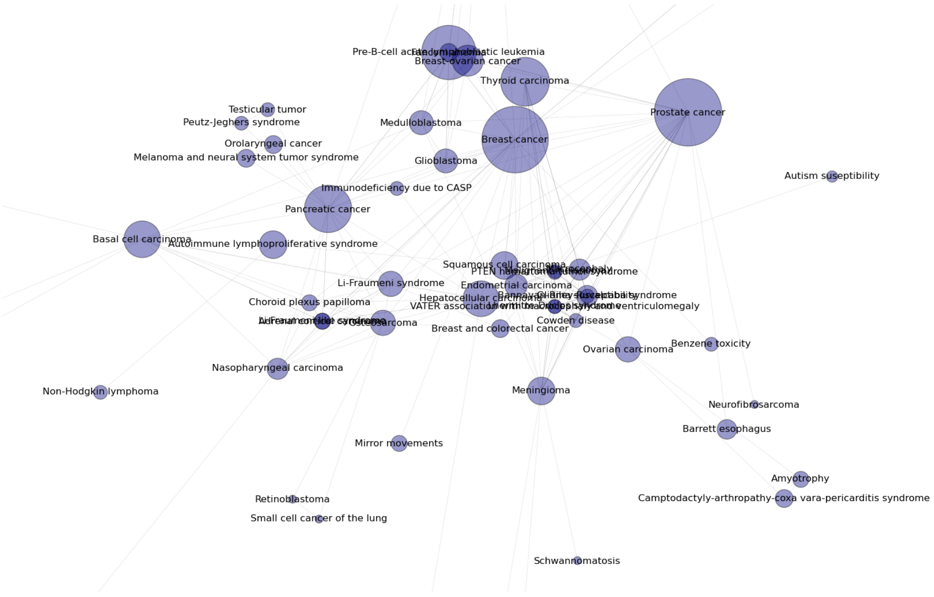

Figure 4: Breast cancer cluster

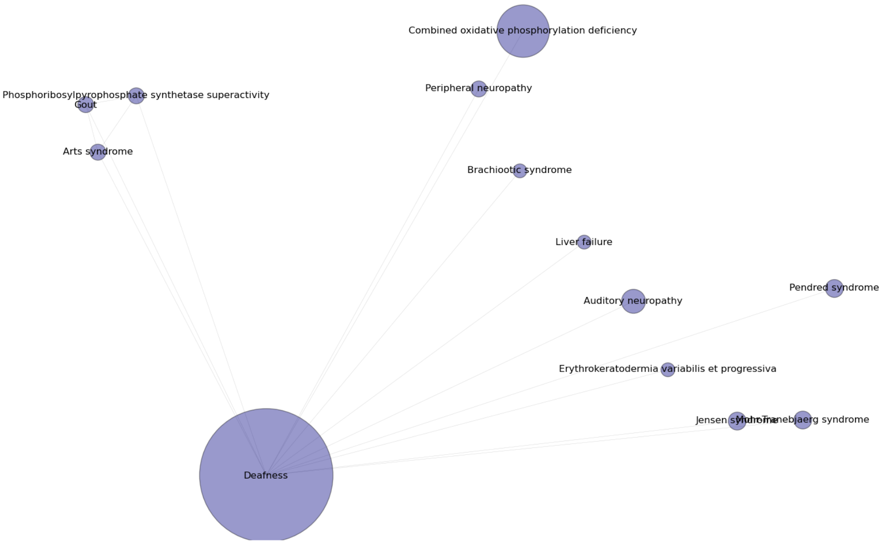
Combined oxidative phosphorylation deficiency
Phosphoribosylpyrophosphate synthetase superactivity
Gout
Peripheral neuropathy
Arts syndrome
Brachiootic syndrome
Liver failure
Pendred syndrome
Auditory neuropathy
Erythrokeratodermia variabilis et progressiva
Jensen syndrome
Mohr-Tranebjaerg syndrome
Deafness

Figure 5: Deafness cluster

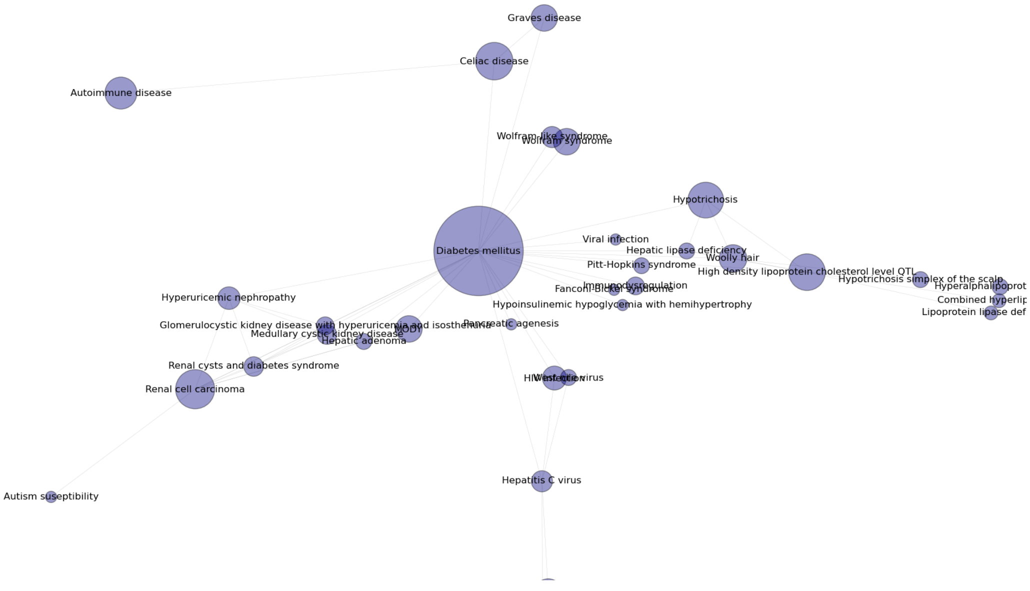

Figure 6: Diabetes mellitus cluster