CPA project

Edwin Ansari
Maria Sidko

1 Introduction

Real world implementation of the mathematical notion of graph theory has countless applications in modern computer science, which have become crucial in our daily life, from efficient handling of data to pertinent responses of search engines. Throughout this project, we have attempted to understand, implement and optimize some of those unavoidable algorithms, following the guidelines provided by our professor for the subject.

In order to reach that goal, we have decided to implement our code using the Java language, given its reasonable calculation efficiency, in addition to its object oriented paradigm and provided APIs, reaching an optimal middle ground between speed and coding comfort.

2 TME 1 - Handling a large graph

Exercise 1 - Preparation The provided files are comprised of the edges of the graph, so we can easily count the edges by looping through it. The nodes on the other hand should only be counted once, so we used an HashSet for that purpose. The class `Counter` implements this process. Following are the results of the counting on the examples :

Amazon benchmark :	number of vertices	334863
	number of edges	925872
Live journal benchmark :	number of vertices	3997962
	number of edges	34681189
Orkut benchmark :	number of vertices	3072441
	number of edges	117185083

N. B. The last provided benchmark, Friendster, which has 65608366 Nodes and 1806067135 Edges according to the website it was taken from. That gives us $|E| \approx 10^9$ and $|G| \approx 10^9 + 10^7$ Just reading this file and storing it in a simple list structure would take $10^9 \times sizeof(int) = 10^9 \times 4Bytes \approx 4GB$ of RAM and then the whole program that counts it would take on top of it around 5GB of RAM (that is nearly impossible with our basic computers having 8GB of RAM in total). The same reasoning could be applied in terms of time complexity. We were therefore unfortunately unable to use that benchmark throughout the project.

Exercise 2 - Three graph data structures Given that the graphs in the provided files are essentially already stored as a list of edges, the parsing, done by the `GraphLoader Class`, creates automatically a list of edges representation (`GraphList Class`), from which are derived the two other representations, `GraphArray Class` and `GraphMatrix Class`. The number of vertices of the two latter representations is `MAX NODE ID + 1`. In an attempt to improve speed, we decided to replace java Scanner with a buffered `FastReader.java`. Our results were the following :

Amazon :	List of edges	1934ms
	List of edges using FastReader	939ms
	Adjacency array	89ms
	Adjacency matrix	OutOfMemoryError
Live-journal :	List of edges	91363ms
	List of edges (using FastReader)	31400ms
	Adjacency array	OutOfMemoryErrorr
	List of edges (allocating 4500MB of RAM)	32031ms
	Adjacency array (allocating 4500MB of RAM)	17712ms
Orkut :	List of edges (using FastReader)	OutOfMemoryError
	Adjacency array	OutOfMemoryErrorr
	List of edges (allocating 5800MB of RAM)	441201ms
	Adjacency array (allocating 5800MB of RAM)	OutOfMemoryErrors

Space Complexity :
List of Edges : $O(|E|)$
Adjacency Array : $O(|V| + |E|)$
Adjacency Matrix : $O(|V|^2)$

We have the same time complexity for structures reading.

It is noticeable right away that, while the list of edges representation is scaled up reasonably easily, and so is the adjacency array representation, the matrix representation quickly gets out of range, and despite additional memory allocation, reaches the hardware limits of our machines.

Exercise 3 - BFS Diameter results using `GraphDiameter` class :

Amazon	Lower-Bound	47
	Lower-Bound time	2330ms
	Upper-Bound	57
	Upper-Bound time (after Lower Bound executed)	296ms
Live-Journal	Lower-Bound	21
	Lower-Bound time	74739ms
	Upper-Bound	24
	Upper-Bound time (after Lower Bound executed)	10685ms

Exercise 4 - Triangles The intuitive way to calculate the triangles would be for each edge of the graph to compute the intersection of each two nodes. However, due to efficiency issues, some adjustments can be made to that algorithm. In order to avoid registering the same triangle multiple times, as well as to optimize processing time, we can visit the vertices in a non increasing degree order, and only considering neighbor vertices of lower degree. This is implemented by the `TriangleFinder` class

amazon	triangles	667129
	counting time	2840ms

3 TME2

Exercise 1 - PageRank (Directed Graph) Implementing the Pagerank algorithm in the `PageRank` class we could achieve accurate results after about 5 or 6 iterations, using the Wikipedia database. Here are the top 5 listed pages :

```
max 1 : United States
max 2 : United Kingdom
max 3 : 2007
max 4 : Germany
max 5 : 2006
```

and the 5 bottom :

```
min 1 : Schismogenesis
min 2 : Thomas W. Thompson
min 3 : Thomas W. Thompson
min 4 : List of examples in general topology
min 5 : James Dempsey
```

Exercise 2 - Correlations In order to plot we used linear scales.

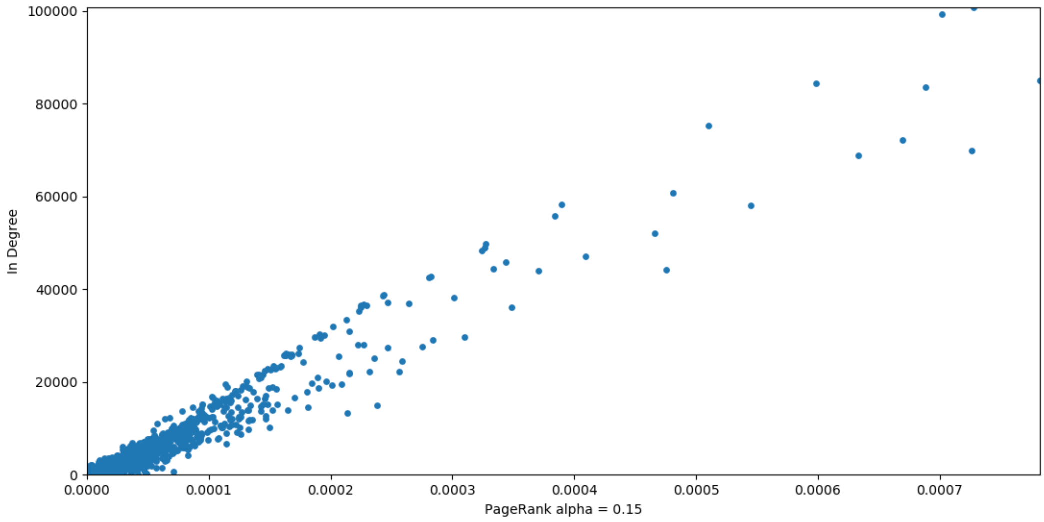

FIGURE 1 – Scatter plot using the Wkikpedia dataset

Intuitively, the more references(In Degree) a node has, the higher its rank is. We can notice that on the scatter plot, as it is somehow linearly correlated to the Pagerank results, which means that it is linked to the ranking, but not identical, given that other factors enter in the Pagerank calculation.

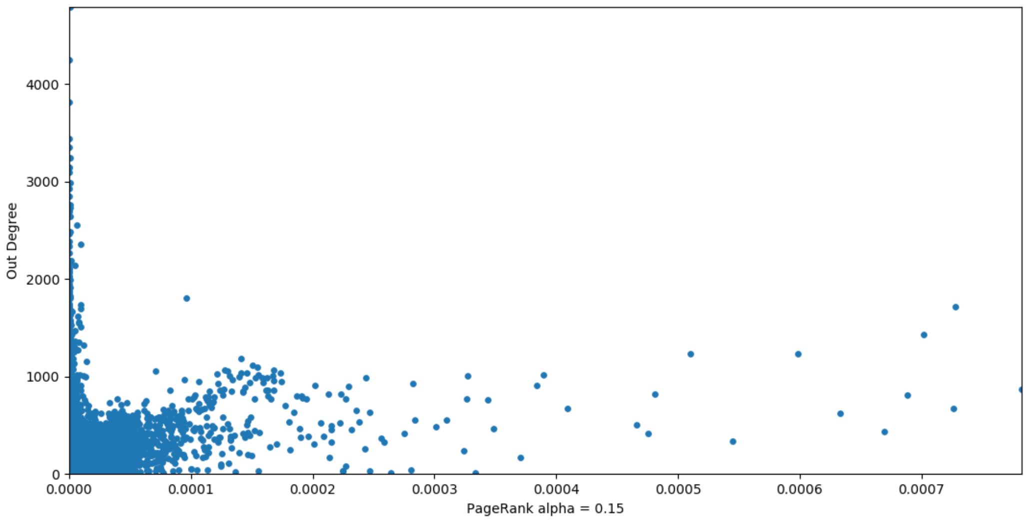

FIGURE 2 – Scatter plot using the Wkikpedia dataset

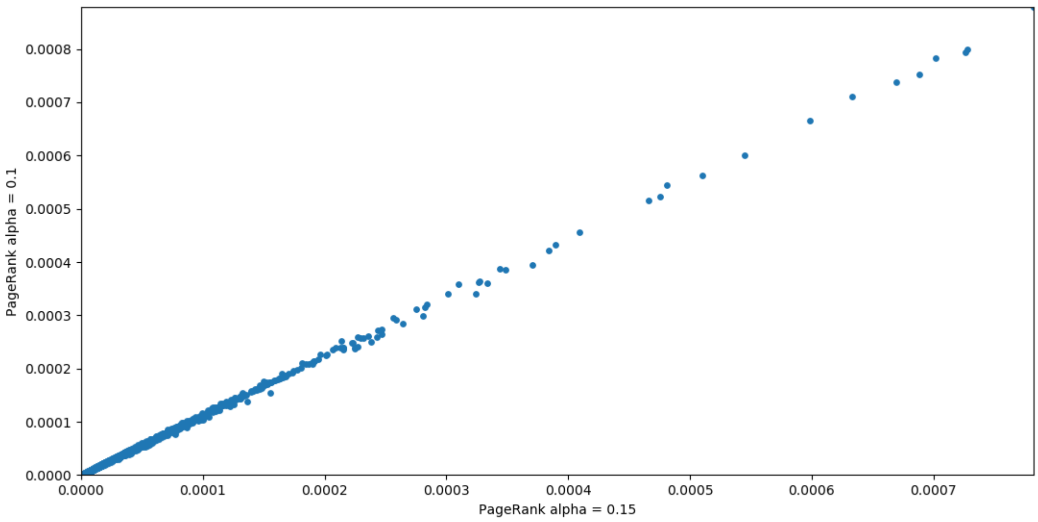

FIGURE 3 – Scatter plot using the Wkikpedia dataset

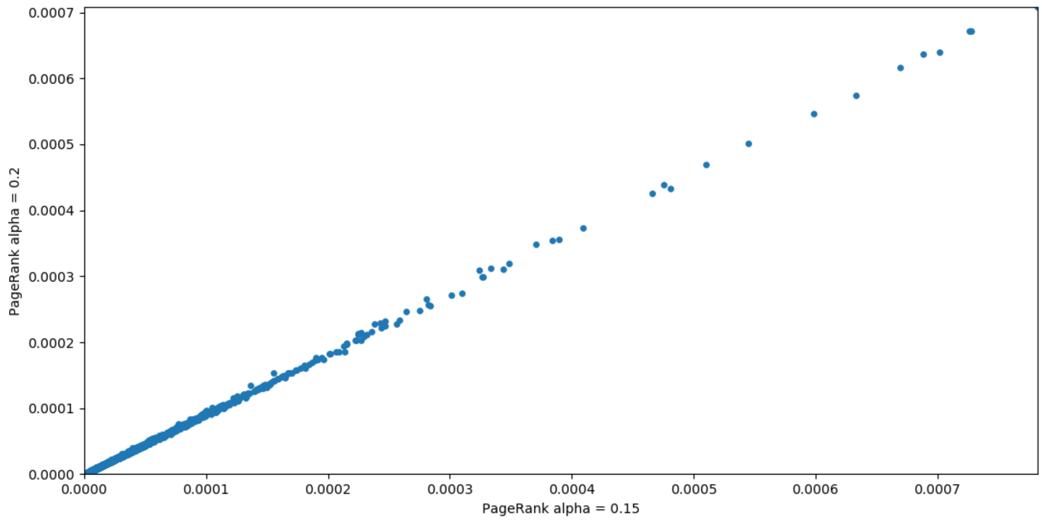

FIGURE 4 – Scatter plot using the Wkikpedia dataset

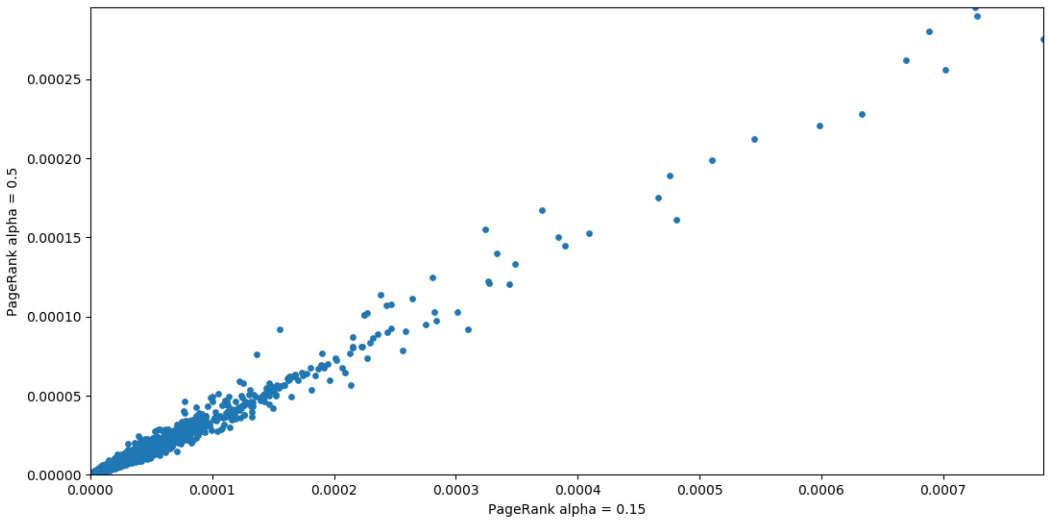

FIGURE 5 – Scatter plot using the Wkikpedia dataset

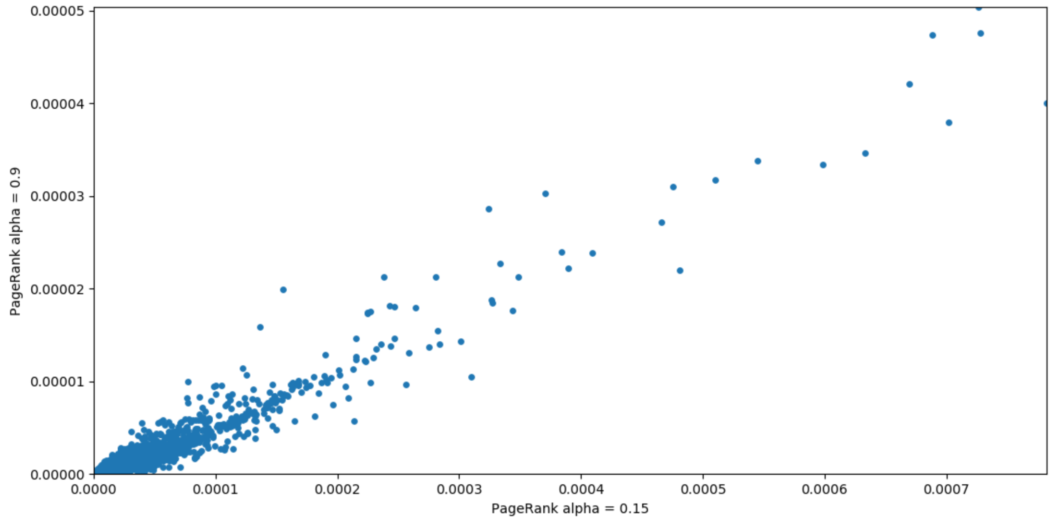

FIGURE 6 – Scatter plot using the Wkikpedia dataset

We can notice that the α value of the PageRank algorithm has an influence on its results, the closer they are the more linear the plot is, meaning that the PageRank results of the two calls are similar. When the values become more distant, the rankings change, leaving the scatter plot more sparse.

Exercise 3 - Personalized Page Rank

Magnus Carlsen : To find very close results to node(Magnus Carlsen) we first found its Page Id : 442682 and then run the *rootedPageRank* using an initial P_0 vector such that :
$P_0(\texttt{magnus\_id}) = 1$
$P_0(i) = 0\ \forall i \neq \texttt{magnus\_id}$

By computing personnalized PageRank using *Magnus Carlsen* as the root node, we obtained the following top and bottom pages :

```
max 1 : Magnus Carlsen
max 2 : Chess
max 3 : Grandmaster (chess)
max 4 : Federation Internationale des echecs
max 5 : Elo rating system

min 1 : Transmission and infection of H5N1
min 2 : Ital
min 3 : (Sittin' on) the Dock of the Bay
min 4 : Ministry of New and Renewable Energy (India)
min 5 : Superclub
```

Chess Boxing :
In order to have a relevant restart vector using a set of subjects (Chess and Boxing in this case) we offer two methods.

1. In this first method we simply set the restarting vector P_0 such that :
   $P_0(\texttt{id}) = \frac{1}{N}$ if $id \in S$
   $P_0(\texttt{id}) = 0\ \forall id \notin \texttt{S}$
   Where S is the set of pages that are linked to any of the related categories and N is their quantity.
   The results using this method are :

```
max 1 : Chess
max 2 : Grandmaster (chess)
max 3 : Federation Internationale des echecs
max 4 : Chess Olympiad
max 5 : United States
max 6 : Boxing
max 7 : International Master
max 8 : Elo rating system
max 9 : World Chess Championship
max 10 : Chess opening

min 1 : Villeray-Saint-Michel-Parc-Extension
min 2 : Jessica Cutler
min 3 : Electrical connection
min 4 : Bass player
min 5 : Caribou-Targhee National Forest
min 6 : Above the Rim
min 7 : Charlesbourg, Quebec
min 8 : Gaius Calpurnius Piso
min 9 : Lisa Opie
min 10 : Persistence
```

2. We fetch, using all the provided files, the number of categories related to the requested subjects each page is linked to, alongside with searching whether the actual names of the pages contain those subjects. We use this information to weight the restarting vector P_0 such that : $P_0(i) = \frac{r+10}{t}$ if $ct = 0$
   $P_0(i) = \frac{r+ct*k}{t}$ if $ct > 0$
   $P_0(i) = 0 \ \forall i \notin \mathtt{S}$

   Where r is the number of appearances of the related subjects in the page categories, ct is number of subjects that are a sub-string of the page name, k is a factor that can be customized in order to reach convergence (we chose 20 in our tests). and $t = \sum_{i=0}^{n} P_0(i)$
   Using this second method we get the following top and bottom results :

```
max 1 : Chess
max 2 : Grandmaster (chess)
max 3 : Boxing
max 4 : United States
max 5 : Federation Internationale des    checs
max 6 : Chess variant
max 7 : Chess Olympiad
max 8 : Chess problem
max 9 : Chess boxing
max 10 : World Chess Championship

min 1 : Pioneer High School
min 2 : Umea IK
min 3 : Guadalupe River (California)
min 4 : Brian Ashton (rugby player)
min 5 : Grammy Award for Best Latin Rock/Alternative Album
min 6 : Brampton Thunder
min 7 : Pump Up the Volume (song)
min 8 : Sigurd Evensmo
min 9 : Ethel Snowden
min 10 : St Edward's School (Oxford)
```

4 TME 3

Exercise 1 - Simple Benchmark Following are some of the generated random graphs :

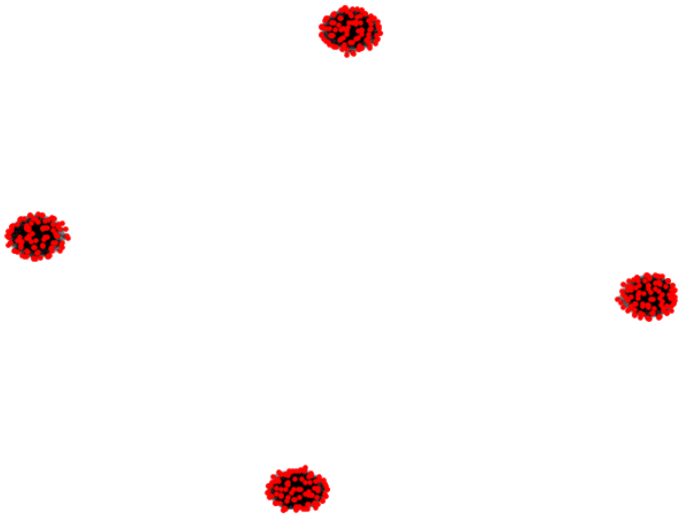

Figure 7 – Random generated graph with p = 1.0 and q = 0.0

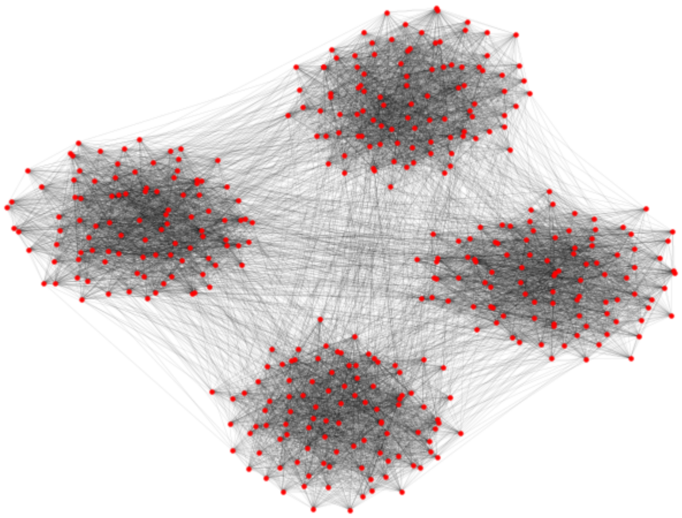

Figure 8 – Random generated graph with p = 0.4 and q = 0.02

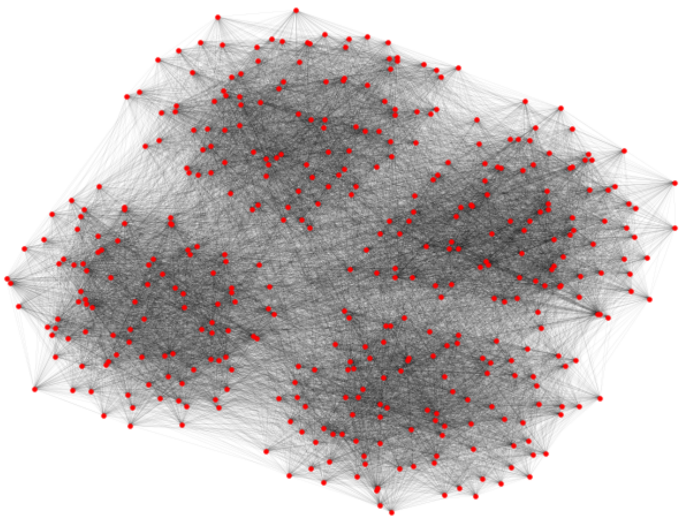

Figure 9 – Random generated graph with p = 0.6 and q = 0.1

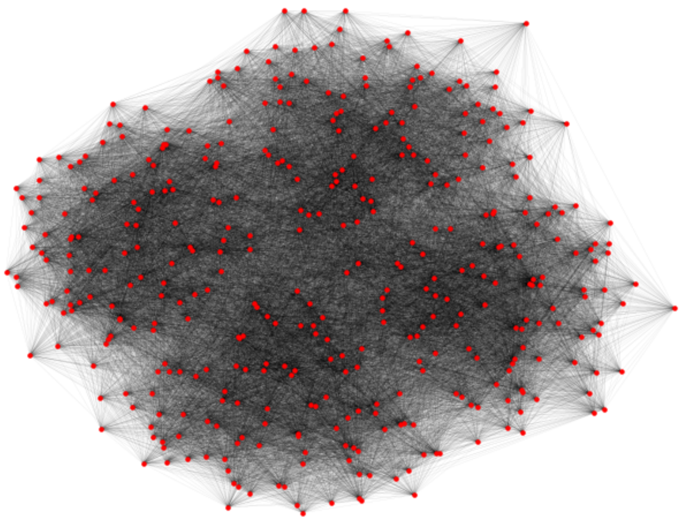

Figure 10 – Random generated graph with p = 0.7 and q = 0.2

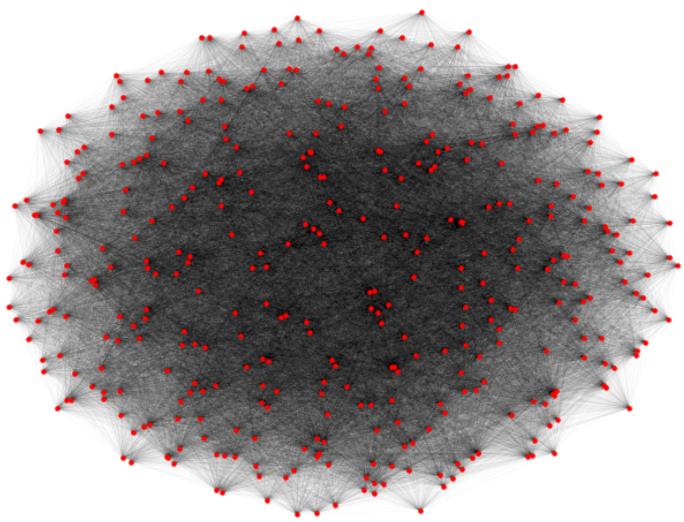

FIGURE 11 – Random generated graph with p = 0.5 and q = 0.5

As we can see on the graphical rendition of the graphs, increasing $\frac{p}{q}$ ratio leads to more noticeable separation in communities, while decreasing it creates a more homogeneous graph with only one big community.

Exercise 2 - Label propagation We have implemented the Label Propagation label in the `LabelPropagator` class. Following are the results of the call of this algorithm of some random generated graphs.

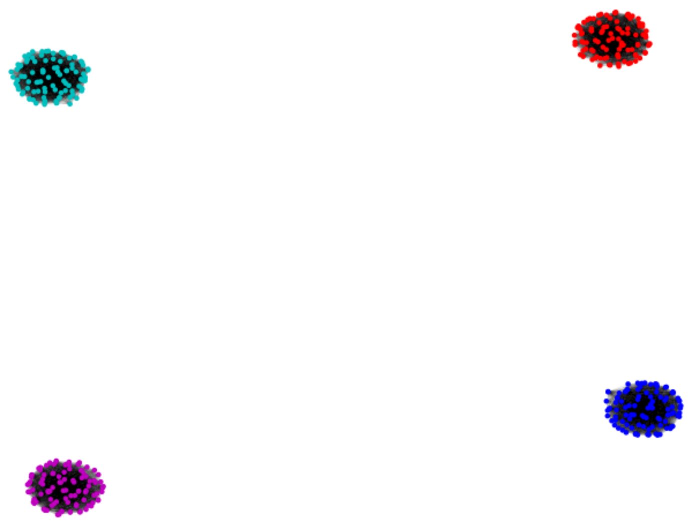

FIGURE 12 – Results of the label propagation algorithm on a random generated graph with p = 1.0 and q = 0.0

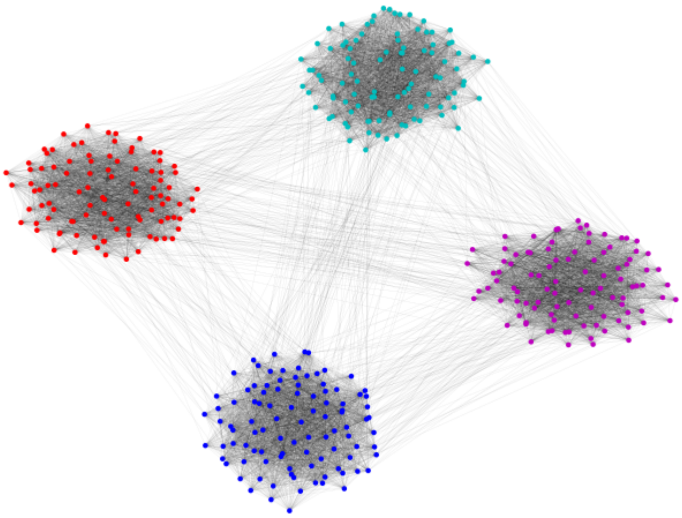

FIGURE 13 – Results of the label propagation algorithm on a random generated graph with p = 0.6 and q = 0.01

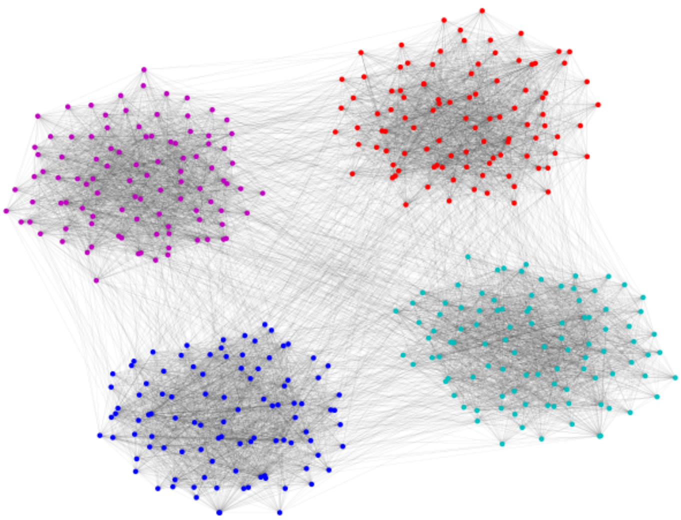

FIGURE 14 – Results of the label propagation algorithm on a random generated graph with p = 0.4 and q = 0.02

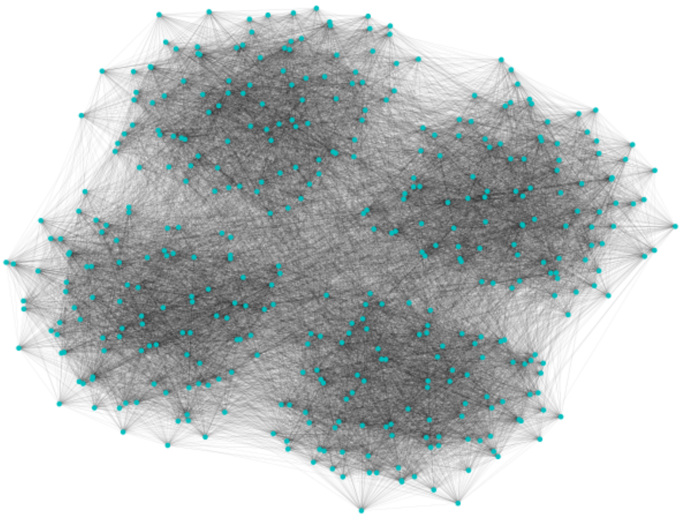

FIGURE 15 – Results of the label propagation algorithm on a random generated graph with p = 0.6 and q = 0.1

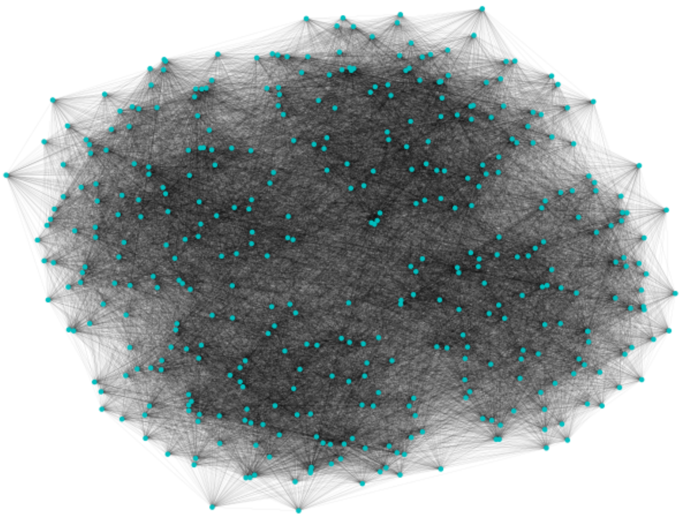

FIGURE 16 – Results of the label propagation algorithm on a random generated graph with p = 0.7 and q = 0.2

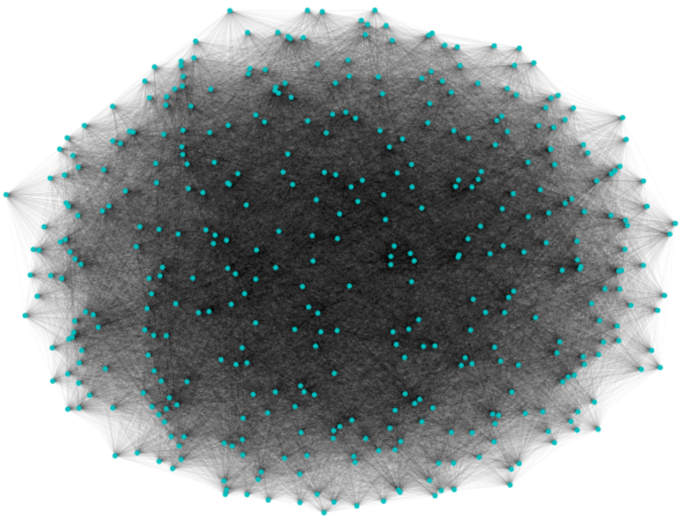

FIGURE 17 – Results of the label propagation algorithm on a random generated graph with p = 0.5 and q = 0.5

While the algorithm is efficient at recognizing communities with a $\frac{p}{q}$ ratio higher than 20, when it decreases it has more trouble and only recognizes one unique community comprising the whole graph.

Exercise 3 - New algorithm For this matter, after searching for different existing algorithms, and in order to experiment on a different approach from label propagation and Louvain algorithms, we decided to turn to divisive approaches, and, after looking more in depth we settled on the Girvman-Newman algorithm. This algorithm computes the communities by successively removing edges from the graph by a certain measure, and then computing the connected components of the graph, which are the communities. The measure, called betweenness is defined as the extent to which that edge lies along shortest paths between all pairs of nodes.The edge with the highest betweenness will be removed at each iteration. We experimented on this algorithm using the `Jung` java API, in the `Clustering` class.