Final report: drug-drug interaction network

Assignment This homework aims at analysing the structure of a drugs interactions network.

To study the network I extract the following analytics, dividing the work in two parts:

- Part I:
  - a) average degree, higher moments;
  - b) average distance, diameter;
  - c) CCDF and ML fitting;
  - d) clustering coefficient and its distribution;
  - e) assortativity;
  - f) robustness to random failures and attacks;
- Part II:
  - a) PageRank;
  - b) HITS;
  - c) ranking comparison: degree, betweenness, PageRank and HITS;
  - d) communities detection;
  - e) link-prediction.

At the end of each part I report some considerations about the results; the paper final consideration are at the end of the analysis.

In the Appendix I put some informations about the drugs mentioned during the ranking comparison (Part II.c), which are not essential for this work, but it helps to contextualize it.

Introduction In pharmacology a Drug-drug interactomes (DDI) is a complex network with drugs as nodes and interactions between them as edges. DDIs are used to predict potential interaction, even unknown ones, to be sure some interactions will be avoided and to study links between the pharmaceutical properties and drug interactions. The research for candidates for drug repositioning can also be accelerated thanks to DDI studies, and as 20% of the new drugs brought to marked in 2013 were repositioning [1], it would bring economic benefits too.

I decided to use the most famous public database of drugs *DrugBank* [2] and in particular the network of *U.S. Food and Drug Administration* approved drugs, available on-line thanks to the *Stanford's BioSNAP project* [3].

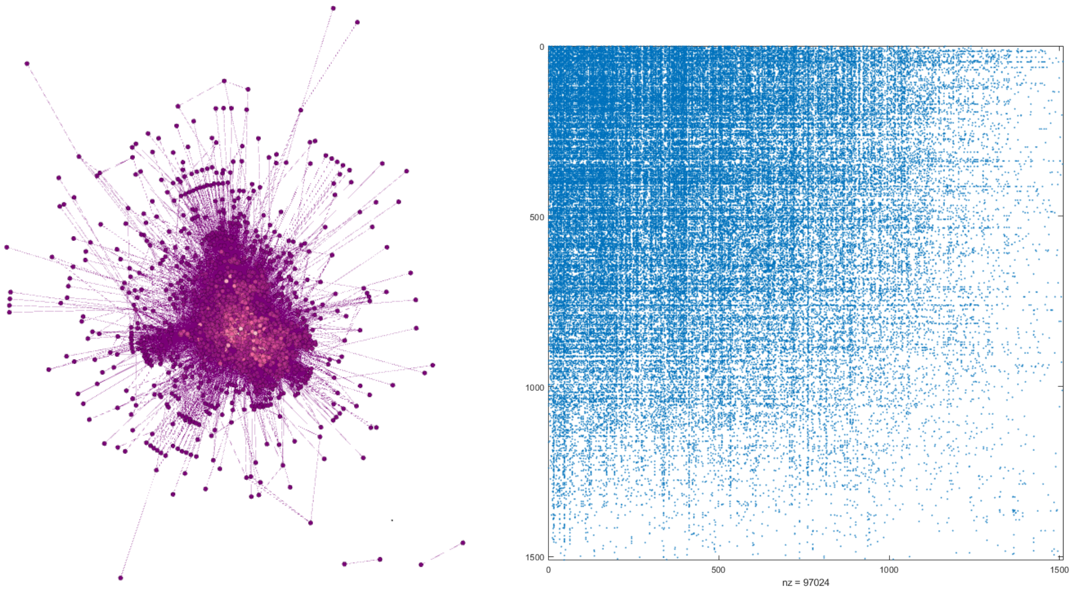
The Dataset The original file has to be processed before being properly used in MATLAB as it's just two columns of drugs codes representing edges between nodes. So from this raw data, I use a script created in Python to generate two files:

1. *drugs_nodes.txt*: it's a table ID - Drug names tag, it's useful to have actual drug names references since they get converted in numbers;

2. *drugs_edges.txt*: it's a Source-Target edges table, all the names have been converted in numbers so no other preprocessing is needed in MATLAB.

The project has been developed and tested with the following tools:
Python: version 3.7.5, using Pandas external library; IDE: Spyder 4.0.1.
MATLAB: version R2019b, service update 5.
Gephi: version 0.9.2.

Pre-processing By looking at the edges with *Gephi* in Figure 1a, it is clear that there is a very large connected component plus some other nodes. So after loading the data in MATLAB I find the largest connected component and develop my analysis on this network.
I obtain the upper triangular adjacency matrix B and with a simple operation I create the symmetric adjacency matrix A as B + B′, Figure 1b.
At the end I obtain a graph of 1510 nodes and 48512 undirected edges, they originally were 1514 and 48514 respectively. There are no loops in the connections.
The network is very dense, so it has a very high number of edges compared to the nodes number.



(a) (b)

Figure 1: *(a) Visualization of the original network using Gephi. On the bottom-right corner is possible to see 4 nodes not attached to the main component; (b) Adjacency matrix A of the network.*

Part I

Degree
From the adjacency matrix A, I compute the average degree $\langle k \rangle$, second moment degree $\langle k^2 \rangle$ and third moment degree $\langle k^3 \rangle$:

$$\langle k \rangle = 64.2543;$$
$$\langle k^2 \rangle = 8\,800.6490;$$
$$\langle k^3 \rangle = 1\,670\,104.1550.$$

The $\langle k \rangle$ is very high, but it was expected as the graph is dense.
Another expected result is the *connected regime* of the network, as by definition

$$\langle \mathrm{k} \rangle > ln(N) = 7.3199,$$

and by construction, the graph's nodes are connected together.

Distance
The DDI network has an *average distance* $\langle \mathrm{d} \rangle$ and a diameter d_{max} as follows:

$$\langle d \rangle = 2.5149;$$
$$d_{max} = 7.$$

Another two different estimation for the average distance are the one valid for a random network and power-law based network:

$$\langle d \rangle_{random\ net} = \frac{ln(N)}{ln(\langle k \rangle)} \simeq 1.7584$$

$$\langle d \rangle_{powerlaw\ net} = ln(ln(N)) \simeq 1.9906$$

But it's clear that these $\langle d \rangle$ do not well approximate the real value.

CCDF and ML fitting
From the adjacency matrix A, I compute the *Complementary Cumulative Density Function* CCDF and attempt to fit the curve with a power law $p_k = C \cdot k^{-\gamma}$ where γ is estimated with a *Maximum Likelihood* ML criterion:

$$\gamma = 1 + \frac{N}{\sum_i ln(\frac{k_i}{k_{min}})}.$$

The estimated γ is 5.8145, which according to the theory is characteristic for a random network regime. In Figure 2 the plot of the CCDF and ML fit.

Clustering coefficient
The clustering coefficient C is used to denote how strongly a node is connected to its neighbours.
In this DDI network $\langle C \rangle$ is equal to 0.30477, which is a low value considering C lives in the interval [0,1].
In Figure 3 the plot of $C_{i|k_i=k}$ and $\langle C \rangle$.

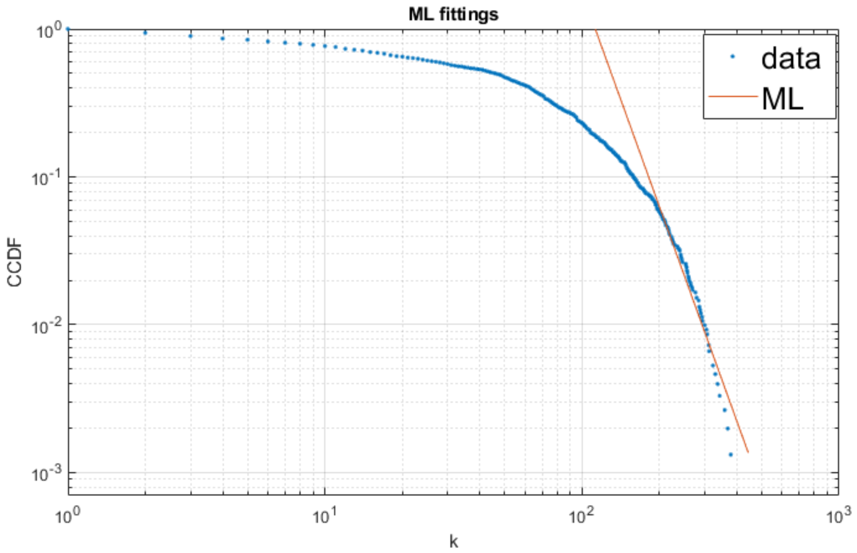


Figure 2: *Degree distribution CCDF and ML fit.*

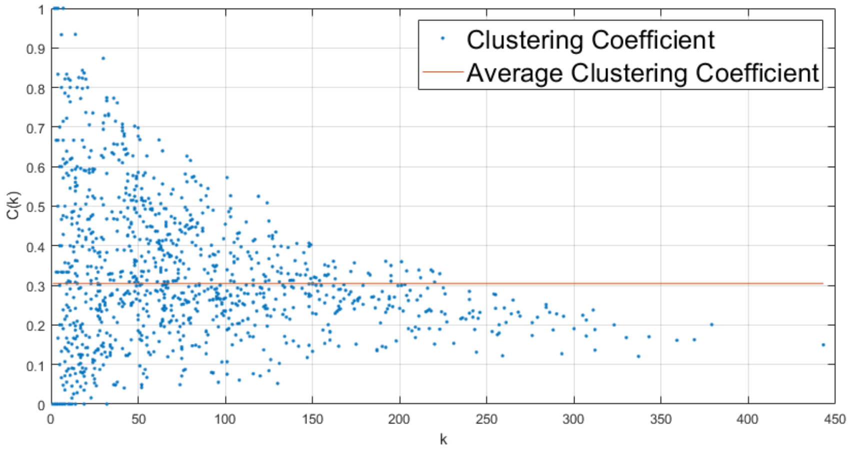


Figure 3: *The clustering coefficients* $C_{i|k_i=k}$ *in blue,* $\langle C \rangle$ *in red.*

Assortativity

To see how the nodes are connected together, I check the assortativity of the network with $k_{nn}(k)$ and $k_{nn}^{R-S}(k)$ (*Degree Preserving Randomization with Simple Links* technique). The *assortative factor* is -0.03, the results plot in Figure 4 confirm that the DDI network is neutral and that it's a structural property of the graph.

Robustness

To check the robustness of the network, I simulate a random node deletion attack and then an adversary attack. In Figure 5 it is possible to see the plot of the robustness when the attack consists in a simple random node removal (in green), which is very linear; a more specific attack, removing the highest degree nodes, reveals a f_c relatively large, about 0.73, probably because the DDI is a dense network. In the plot I highlight the 0.1 threshold to show how faster the attack methods is compared to random failures.

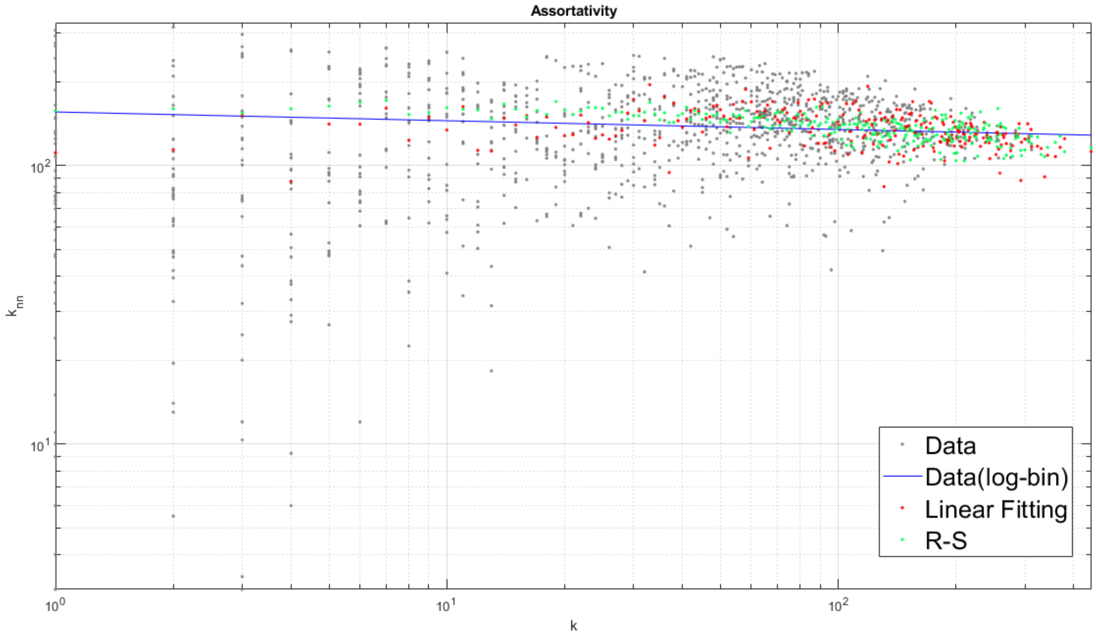


Figure 4: *Assortativity of the network.*

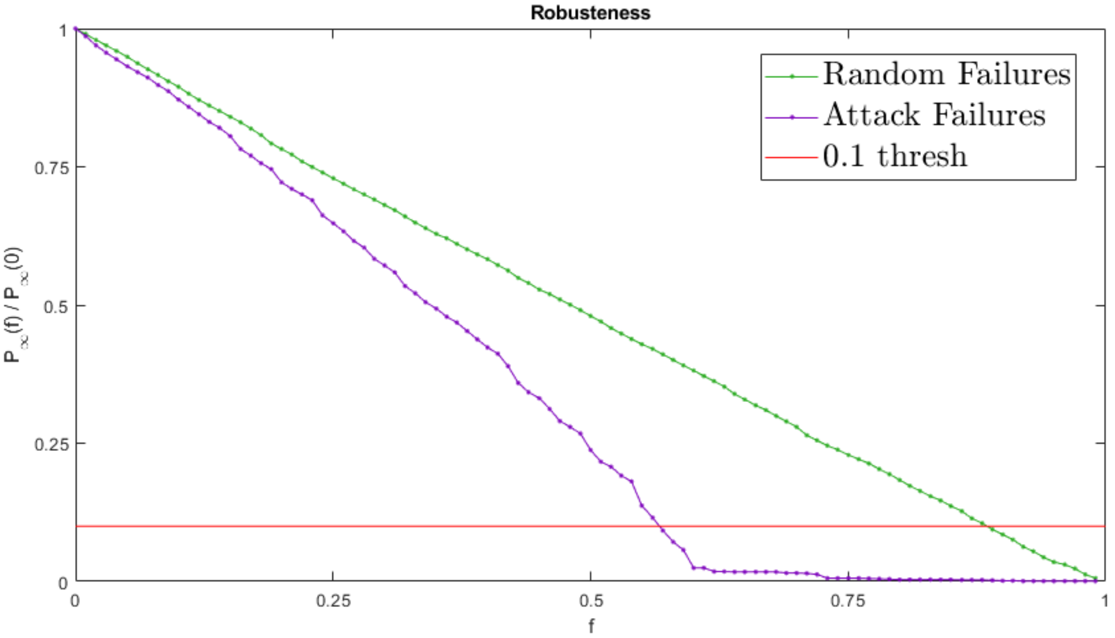


Figure 5: *Random and adversary attack on the network.*

Consideration on part I

The network is neutral, random, which means a physician must be careful to prescribe new drugs to a patient while he's taking something already as it could interfere with the treatment even if totally unrelated.

For what concerns the robustness, an f_c of 0.73 could also mean that the drugs tend to have many interaction between them, so it's hard for a physician to balance the patient's cure. On the other hand should be easier repositioning a drug to work with other drugs and have some possibly positive effects for the patient.

Part II

PageRank

With a random walk over the network the PageRank methods estimates the importance of each node in the network. This algorithm gives a (stochastic) vector p_∞, containing the probabilities to end on a node with a random surf over the net.
p_∞ is an eigenvector of the matrix

$$M_1 = cM + (1-c)q1^T, \text{with } M = A \cdot diag^{-1}(d).$$

With the *Power iteration* method, $p \simeq p_\infty$ is extracted; the convergence of the algorithm is fast and, from the theory, is proportional to the second largest eigenvalue of M_1, λ_2:

$$||p_t - p_\infty|| \sim [c \mid \lambda_2 \mid]^t.$$

In Figure 6 on the left the *Power iteration* convergence, on the right the eigenvalues λ_i of the matrix M_1 inside the unit circle.

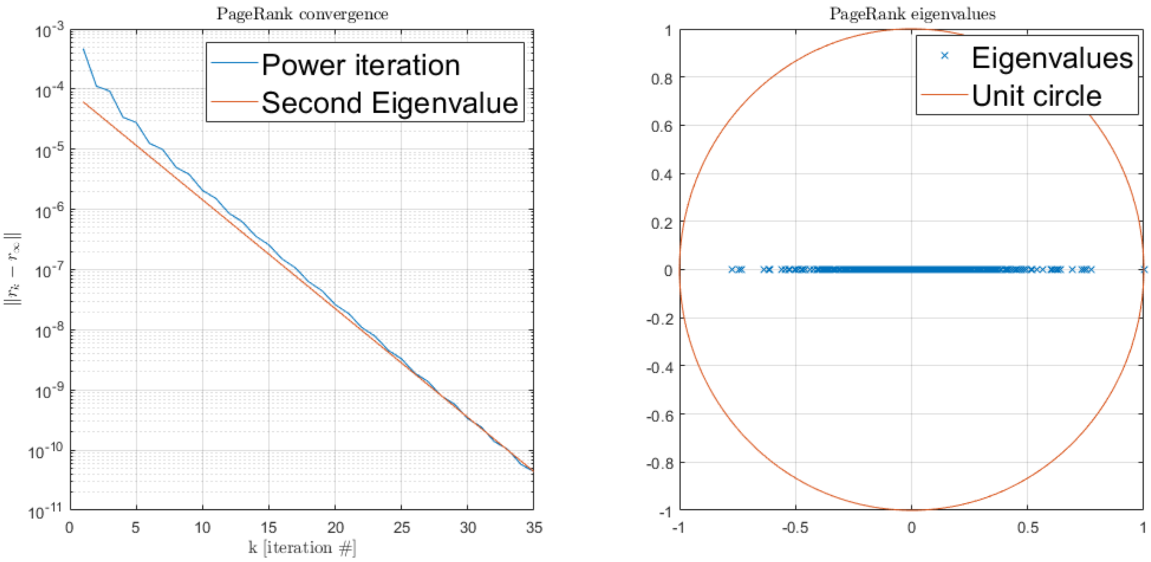


Figure 6: *On the left the PageRank convergence, on the right the eigenvalues λ_i of the matrix M_1 inside the unit circle*

Hiperlink induced topic search - HITS

The DDI network is undirected so the score for the authority and hub is the same.
Once again the *Power iteration* method can be used to solve the HITS algorithm as we are interested in the principal eigenvector of the matrix $M = A^T A$.
From the theory, this time the convergence of the *Power iteration* solution is proportional to λ_1 and λ_2:

$$||a_t - a_\infty|| \sim \left(\frac{\lambda_2}{\lambda_1}\right)^t, \text{with } \lambda_1 \text{ and } \lambda_2 \text{ biggest eigenvalues of } M.$$

In Figure 7 the convergence plot of the algorithm.

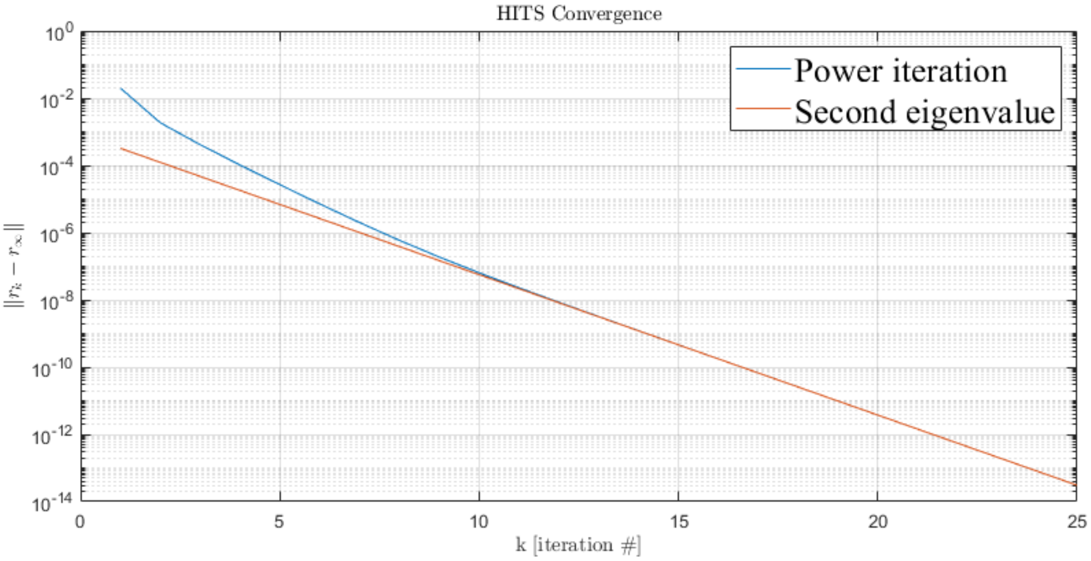


Figure 7: *HITS convergence.*

Ranking comparison

From the plot of the convergences, in Figure 6 (left) and 7, HITS seems faster as it needs less interactions to reach the same precision. As it is possible to see on Figure 8, the rankings are similar, but not the same.

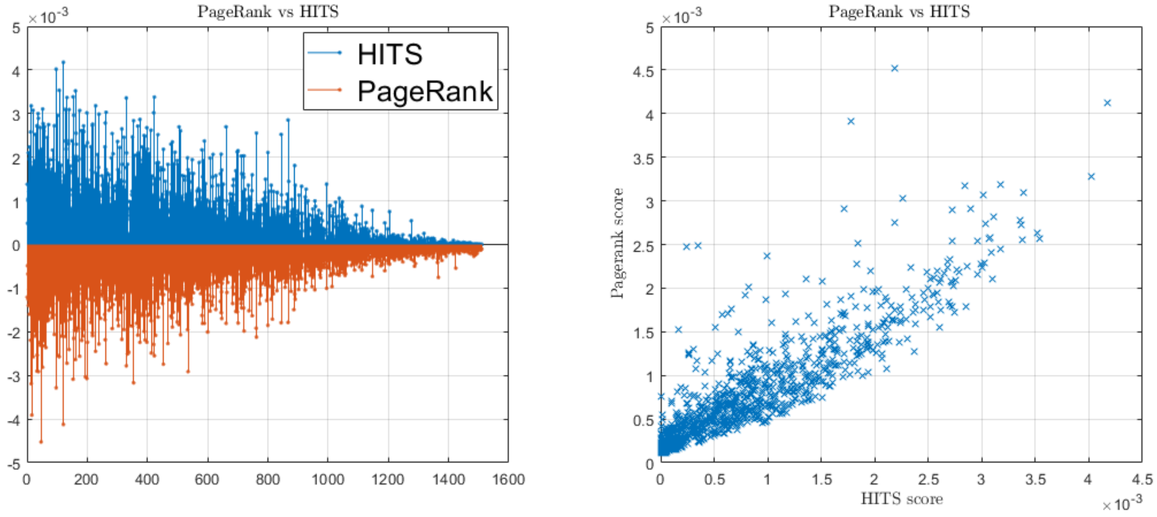


Figure 8: *PageRank and HITS comparison plot.*

Finally, in Table 1 a more pragmatic vision of the nodes classification based on their degree, betweenness and the aforementioned PageRank and HITS.

Phenytoin is drug to pay most attention with, as it's on the Top 3 ranking for all the systems, followed by Mifepristone, not present on the Top 5 only for its betweenness (11^{th} position).

In the Appendix more informations about the drugs in the table.

In Figure 9 and 10 the visual representation of the ranking for all the four methods.

Rank	Degree	Betweenness	PageRank	HITS
#1	Phenytoin	Warfarin	Warfarin	Phenytoin
#2	Mifepristone	Acenocoumarol	Phenytoin	Mifepristone
#3	Paroxetine	Phenytoin	Acenocoumarol	Hydrocodone
#4	Tranylcypromine	Clozapine	Mifepristone	Aripiprazole
#5	Phenelzine	Tacrolimus	Deferasirox	Paroxetine

Table 1: *Comparison rank table based on nodes degree, betweenness, PageRank and HITS score. I highlight with the same colours the same drugs among different indexes.*

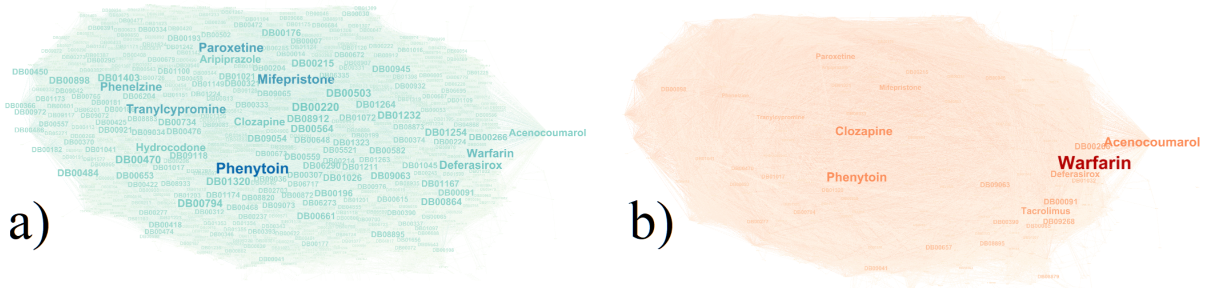


Figure 9: *Visual representation of the a) degree and b) betweenness ranking score (Gephi).*

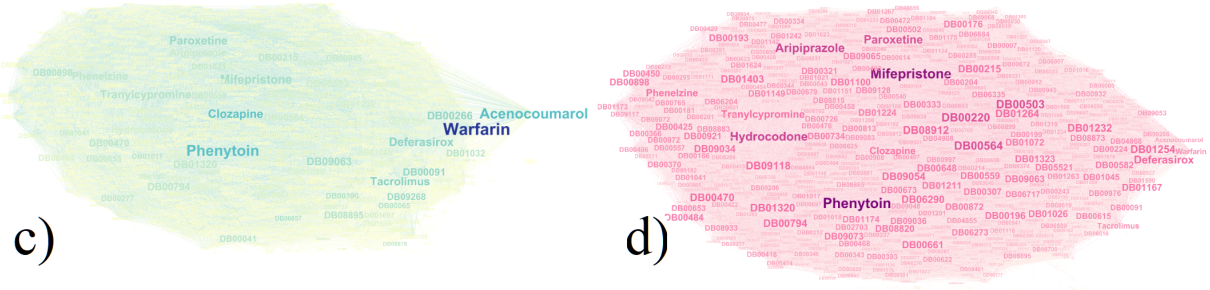


Figure 10: *Visual representation of the c) PageRank and d) HITS ranking score (Gephi).*

Communities

Using the spectral approach, exploiting the *Normalized Lapacian* matrix

$$L = I - D^{-\frac{1}{2}} \cdot A \cdot D^{-\frac{1}{2}}\ , \text{with } D = diag(d),$$

and in particular the eigenvectors of the smallest K eigenvalues, I identify the communities inside the network. In Figure 11 the plot of the eigenvalues and the eigengap. After sorting the nodes based on Fiedler's eigenvector v_{N-1} and computing the conductance $\Phi(A_i)$, looking at Figure 12 I identify six local minima, so the network is divided in seven communities as reported in following Table 2:

Community	#1	#2	#3	#4	#5	#6	#7
Size (number of nodes)	135	36	108	282	369	494	86

Table 2: *Sizes of the communities*

In Figure 13 the 2D representation using normalized eigenvectors to see the division in communities.

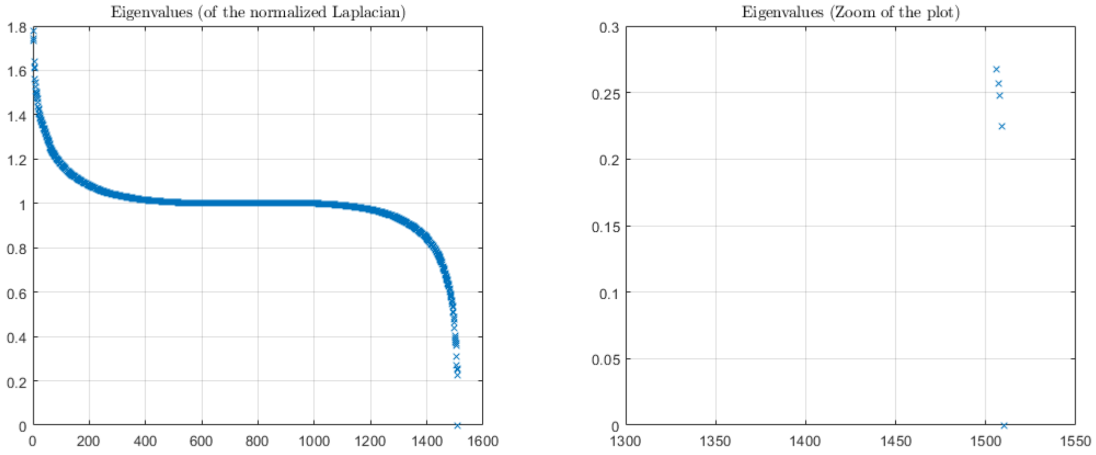


Figure 11: *Eigenvalues of the normalized Lapacian matrix L on the left; smallest eigenvalues of L on the right.*

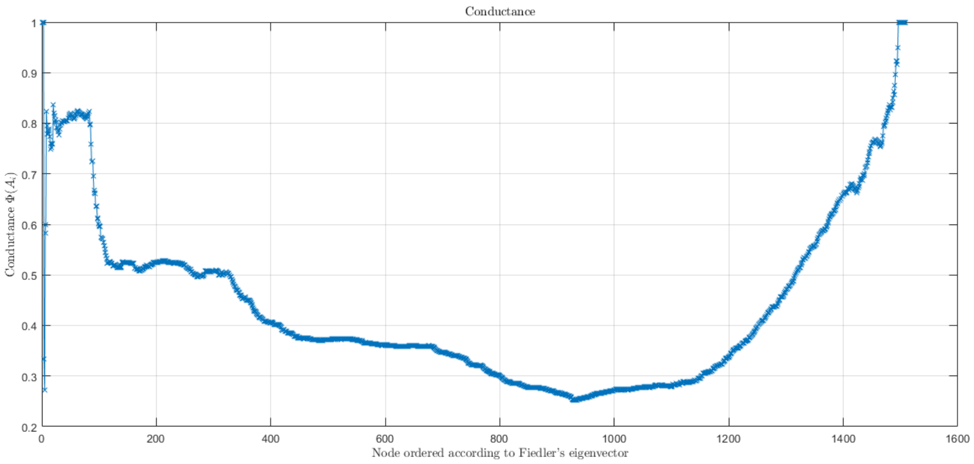


Figure 12: *Conductance of the network.*

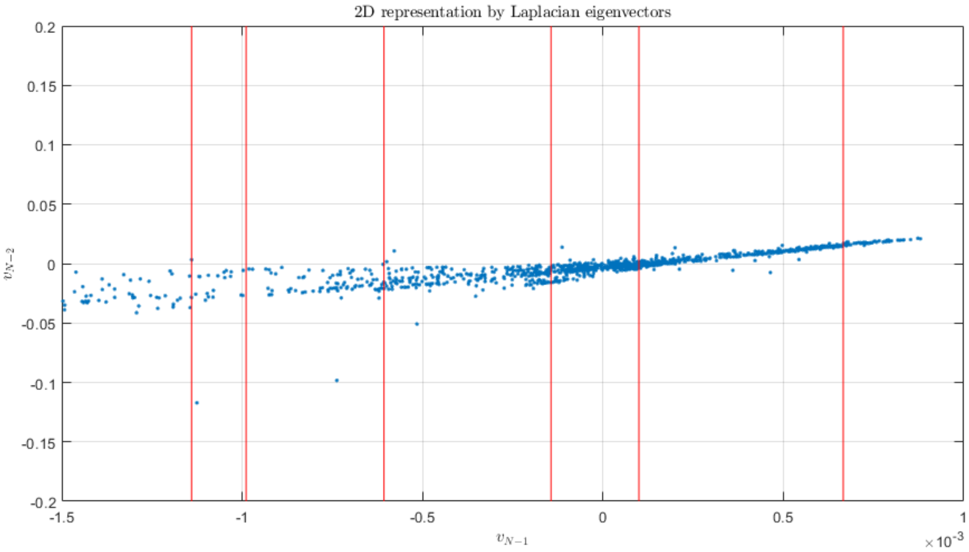


Figure 13: *Communities of the network divided by red lines.*

To explain this division I had to look into the literature and I discovered the seven communities identifies the pharmacological properties of the drugs as follows [1]:

1. drugs acting on the central and peripheral nervous system;
2. inhibitors and inducers of CYP enzymes drugs;
3. drugs related to hemostasis, anti-convulsant and epileptogenic;
4. sympathetic nervous acting drugs;
5. cancer, auto-immune disorders and the musculo-skeletal system targeting drugs;
6. platelet activity and plasma potassium levels interfering drugs;
7. chelatin agents.

Link prediction

I try different methods to analyse the capability to predict future links inside the network, this could be used to say if some drugs could have some interaction yet to be discovered.

The performance of the prediction is measured with the *AUC* and *Precision* (with L = 10% of the total nodes) index; the results with a test set of 80% of edges are reported in Table 3.

	Technique	AUC	Precision
Neighbour based	Common Neighbours - CM	0.8921	0.0397
	Adamic Adar - AA	0.7925	0.1258
	Resource Allocation - RA	0.9008	0.0530
Path Based	Local Path - LP	0.9411	0.6556
	Kats [4]	0.5042	0.0530
Random walk based	Random Walk with Restart - RWR	0.8566	0.1854

Table 3: *AUC and Precision (with L = 151) for CM, AA, RA, LP, Kats and PWR similarity.*

Overall the *Local Path* LP has the best performances and the LP similarity is defined as

$$S_{LP} = A^2 + \beta A^3 \text{ , with } \beta = 0.5.$$

Consideration on part II

This part is important to rank the nodes, so it can be a practical aid for a physician when prescribing multiple drugs to a patient or for when a patient is already on drugs and need to take some others.

The communities division confirmed what I found in the literature.

Final conclusion and most significant end results

In this paper I take a raw DDI dataset and with some initial pre-processing I load the data in MATLAB.

In *Part I* I find the proposed DDI network to be very dense, neutral, robust to random and adversary attacks. I conclude that a physician should be careful when prescribing multiple drugs to a patient as there are many interactions and some could be harmful.

In *Part II* I rank the nodes with some basic (degree and betweenness) and more advanced (PageRank and HITS) metrics to find out the highest ranked nodes, which are the ones corresponding to Warfarin and Phenytoin. The communities discovery confirmed the division found in the literature.

Appendix

The mentioned drugs in this paper with their DrugBank ID [2], main usage and commercial name (when available), all info available on the DrugBank webpage [1]:

- Warfarin (DB00682) - anticoagulant, commercial name: Coumadin;
- Phenytoin (DB00252) - one of the most used anticonvulsants;
- Acenocoumarol (DB01418) - anticoagulant that functions as a vitamin K antagonist (like warfarin);
- Mifepristone (DB00834) - progestational and glucocorticoid hormone antagonist, commercial name: RU-486;
- Deferasirox (DB01609) - iron chelator;
- Hydrocodone (DB00956) - synthetic opioid derivative of codeine, commercial name: with paracetamol is Vicodin;
- Aripiprazole (DB01238) - atypical antipsychotic (treatment of schizophrenia), commercial name: Abilify;
- Paroxetine (DB00715) - serotonin reuptake inhibitor (SSRI), commercial name: Paxil;
- Clozapine (DB00363) - atypical antipsychotic agent;
- Tacrolimus (DB00864) - immunosuppressive, commercial name: Advagraf;
- Tranylcypromine (DB00752) - monoamine oxidase inhibitor (effective in the treatment of major depression), commercial name: Parnate;
- Phenelzine (DB00780) - monoamine oxidase inhibiting antidepressant, commercial name: Nardil.

References

[1] Udrescu, L., Sbârcea, L., Topîrceanu, A. et al. *Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing.* Sci Rep 6, 32745 (2016). `https://doi.org/10.1038/srep32745`

[2] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, Michael Wilson, *DrugBank 5.0: a major update to the DrugBank database for 2018*, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D1074–D1082, `https://doi.org/10.1093/nar/gkx1037`

[3] Marinka Zitnik, Rok Sosič, Sagar Maheshwari, and Jure Leskovec, *BioSNAP Datasets: Stanford Biomedical Network Dataset Collection*, `http://snap.stanford.edu/biodata`, Aug 2018.

[4] Zhang, W., Chen, Y., Liu, F. et al. *Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data.* BMC Bioinformatics 18, 18 (2017). `https://doi.org/10.1186/s12859-016-1415-9`

[1] `https://www.drugbank.ca/`, last access 10 April 2020.