"Politehnica" University of Timişoara
Faculty of Automation and Computers
Departament of Computer and Software Engineering
2, Vasile Pârvan Bv., 300223 – Timişoara, Romania
Tel: +40 256 403261, Fax: +40 256 403214
Web: http://www.cs.upt.ro

METABOLIC NETWORK ACTIVITY MODELING

Dissertation Thesis

Larisa MOCU

Supervisors:
Lect. Dr. Eng. Versavia ANCUSA

Timişoara,
2013

Table of Contents

1. Introduction

In the last decade, there has been a significant progress in the research of complex networks and systems, the main reason being the expansive availability of huge network data resources. Albert Laszlo Barabasi and his team popularized one of their most important finding, namely that real networks behave substantially distinct from acknowledged hypothesis of network theory.

Normally, it was assumed that real networks have a majority of nodes of about the same number of connections around an average. Modern network researchers proved that in real networks, the majority of nodes are very poor connected and that, in contrast with that, there exists some nodes of very strong connectivity. These power-law (scale-free) characteristics describe many real systems, from biological ones to social networks.

Plenty of techniques and models were developed by researchers in order to help us understand and predict the behavior of complex systems. In this paper I review developments in this field, including the explanation of concepts such as *small-world* effect, degree distribution, clustering, network correlations, random graph models, models of network growth, dynamical processes taking place on networks.

Networks can be found everywhere. In real world, we deal with social networks, biological networks, information networks, technological networks. In the field of biology and medicine, it was noticed that representing biological systems as complex networks can be very useful. For example, in biology, metabolic pathways can be represented as a network, which illustrates metabolic substrates and products with directed edges joining them if a known metabolic reaction exists that acts on a given substrate and produces a given product. Another example of biological network is the one of mechanistic physical interactions between proteins, which is usually reffered to as protein interaction network. An important class of biological networks is the genetic regulatory network. The genome itself is a switching network with vertices representing the proteins and directed edges representing dependence of protein production on the proteins at other vertices. The expression of genes,

the production by transcription and translation of the protein for which the gene codes, can be controlled by other proteins, activators or inhibitors.

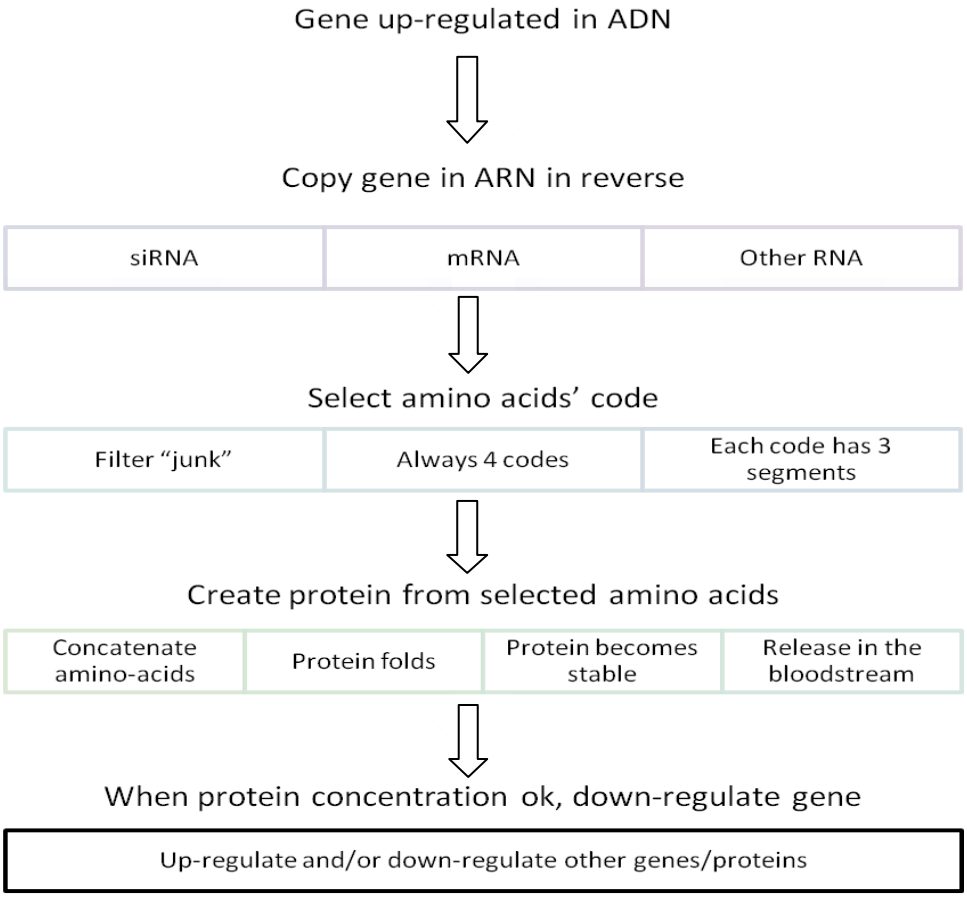
For networks of tens or hundreds of vertices, it is relatively simple to draw a picture and to answer specific questions about network structure by examining its properties. On the other hand, networks of thousands of vertices can be quite difficult to represent. Network analysis tools facilitate qualitative and quantitative analysis of complex networks. They describe features of a network either through numerical or visual representation. Some of the most popular this kind of tools are Gephy, Cytoscape, Graphviz, Pajek, yEd Graph Editor, etc.

On the other hand, architecture and regulation of metabolic networks make the subject of current research in molecular biology. In present, the European Molecular Biology Network is studying the human metabolic network in order to develop a framework for rationally designing clinical intervention strategies and diagnostic for type-2 diabetes. A combination of modeling, bioinformatics and experimental approaches are used to observe metabolic networks and how they are controlled. [1]

Interdisciplinarity between computer science and network medicine became a challenge in research today, and it is widely promoted by Albert Laszlo Barabasi and Paul Smolen. Combining these two fields may answer questions from metabolic networks area that still remain largely elusive.

2. Background

Models for biological networks, from a medical (bio-chemistry) point of view, are very hard to understand and to be transposed into intuitively models. They have the main disadvantage of not being able to represent qualitative/quantitative data. After I did some research about genes expression and metabolic networks, I reached to a simplified model represented below.



Example of medical model for genes

Albert Laszlo Barabasi is best known for his work in the research of network theory. He had important contributions in network medicine and network-biology. Barabasi

discovered the scale-free network property and showed how it emerges in biological systems, namely in metabolic networks and protein-protein interaction networks. He introduced the concept of “diseasome”, or disease network, and showed that diseases link to each other through shared genes.

Barabasi was awarded in 2005 with FEBS Anniversary Prize for Systems Biology. He is member of the Center for Complex Network Research and has many publications in biological networks, complex networks, network medicine, human dynamics and network dynamics areas.

In modeling metabolic networks, Barabasi uses massive simulation and facilitate analysis and correlations which are then used and validated by specialists. The main drawback is that he uses known interactions for genes and only 10% of human genes were studied until now.

Another approach comes from Paul Smolen, a researcher and member of the Neurobiology and Anatomy Department from the University of Texas Medical School at Houston, whose research focuses on differential-equation based modeling to help understand and predict mechanisms of gene regulation underlying circadian rhythmicity. He states that mathematical modeling is essential, for example, to develop a predictive and comprehensive picture of how genes and proteins transduce brief stimuli into long lasting memories.

There are two basic approaches in use for gene network modeling: “the logical network” or “Boolean” method and “the dynamic-systems” method. The dynamic-systems approach uses ordinary differential equations to describe the rates of change of gene products (proteins) concentrations. Terms in these differential equations describe how gene expression rates are modified by changes in the levels of transcription factors or other effector molecules.[2]

According with Smolen, the dynamic system approach is often preffered over the logical-network approach because it is more accurate. However, further development of experimental techniques is expected to facilitate construction of detailed models of gene networks important for controlling key biological events.[2]

3. Theoretical foundations

3.1 Characteristics of Network Science

Network science is distinguished, not only by its subject matter, but also by its methodology. Here are some key characteristics of the approach network science adopted to understand complex systems.

Interdisciplinary nature

Network science can be seen as a language through which different disciplines communicate and interact with each other. For example, in cell biology and computer science fields, researchers must characterize the wiring diagram behind their system and extract information from incomplete and noisy datasets in order to understand the systems' robustness to failures or deliberate attacks. A cross-disciplinary development of tools and methods came as a challenge to explore and conceive the common character that many issues from various fields have. As an example, the concept of "betweeness centrality" that emerged in social network literature in 1970s, today is also used to identify high traffic nodes on the Internet. Computer scientists developed algorithms that can be successfully used in cell biology applications.[3]

Empirical data

Computer science algorithms that represent the base for network science tools have their roots in graph theory, a field of mathematics. The distinction between network science and graph theory is the empirical, data driven nature. Abstract mathematical tools are not enough to describe a network property. Network science tools are tested with real data and their results can be used and validated by specialists.[3]

Computational nature

Definitely, given the size of many of the networks are explored, and the exceptional amount of data behind them, network science offers a series of formidable computational challenges. Hence, the field has a strong computational character, actively borrowing from algorithms, database management and data mining. A series of software tools help practitioners with diverse computational skills analyze networks.[3]

3.2 The Impact of Network Science

The impact a new research filed has can be measured by its intellectual achievements and by the reach and potential of its applications. Although Network Science is a young field, it is of very high impact.

Over the past decade there has been a growing public fascination with the complex "connectedness" of modern society. At the heart of this fascination is the idea of a network — a pattern of interconnections among a set of things — and one finds networks appearing in discussion and commentary on an enormous range of topics. The diversity of contexts in which networks are invoked is in fact so vast that it's worth deferring precise definitions for a moment while one might first recount a few of the more salient examples.

To begin with, the social networks we inhabit—the collections of social ties among friends — have grown steadily in complexity over the course of human history, due to technological advances facilitating distant travel, global communication, and digital interaction. The past half-century has seen these social networks depart even more radically from their geographic underpinnings, an effect that has weakened the traditionally local nature of such structures but enriched them in other dimensions.

The information we consume has a similarly networked structure: these structures too have grown in complexity, as a landscape with a few purveyors of high-quality information (publishers, news organizations, the academy) has become crowded with an array of information sources of wildly varying perspectives, reliabilities, and motivating intentions.

Understanding any one piece of information in this environment depends on understanding the way it is endorsed by and refers to other pieces of information within a large network of links. Our technological and economic systems have also become dependent on networks of enormous complexity. This has made their behavior increasingly difficult to reason about, and increasingly risky to tinker with. It has made them susceptible to disruptions that spread through the underlying network structures, sometimes turning localized breakdowns into cascading failures or financial crises.

The imagery of networks has made its way into many other lines of discussion as well: Global manufacturing operations now have networks of suppliers, Web sites have networks of users, and media companies have networks of advertisers. In such formulations, the emphasis

is often less on the structure of the network itself than on its complexity as a large, diffuse population that reacts in unexpected ways to the actions of central authorities. The terminology of international conflict has come to reflect this as well: for example, the picture of two opposing, state-supported armies gradually morphs, in U.S. Presidential speeches, into images of a nation facing "a broad and adaptive terrorist network", or "at war against a far-reaching network of violence and hatred".

Economic impact

Successful networks like *Google, Facebook, Apple,* have their technology and model business based on networks. Google, the biggest networking mapping operation, relies it's search technology on network characteristics and creates a comprehensive map of the World Wide Web. Since its emergence, Facebook is one of the most used social networking site, with an impressive number of users. The tools developed by network science fuel these sites, aiding everything from friend recommendation to advertising.[3]

Health. Drug design and metabolic engineering.

To fully understand the functionality of human cells and to find the origin of diseases, accurate representations are needed to tell how genes and other cellular components interact with each other. Cellular processes, such as insulin signaling pathway or how the organisms sense environmental changes, can be mapped as molecular networks. The breakdown of these networks lead to most of human diseases. A new subfield of biology, network biology, emerged in order to help better understanding of cellular behavior. Tools have been developed to store databases, explore and analyze patient and genetic data. A parallel movement within medicine, called network medicine, aims to uncover the role of networks in human disease. Networks are particularly important in drug development. The ultimate goal of network pharmacology is to develop drugs that can cure diseases without significant side effects. This goal is pursued at many have made significant investments in networks and systems medicine seeing it as the path towards future drugs.[3]

Security

Terrorism is one of the 21st century main problems. Significant resources are used to combat it worldwide. Network thinking is increasingly present in the arsenal of various law enforcement agencies in charge of limiting terrorist activities. It is used to disrupt the financial network of terrorist organizations, to map terrorist networks, and to uncover the role of their members and their capabilities.. Network concepts have impacted military doctrine as well, leading to the concept of net-war, aimed at fighting low intensity conflicts and crime waged

by terrorist and criminal networks that employ decentralized flexible network structures. One of the first network science programs at the college level was started at West Point, the US Army's military academy. In 2009 the Army Research Lab and the Department of Defense devoted over $300 million to support network science centers across the US.[3]

Epidemics. From forecasting to halting deadly viruses.

Fundamental advantages were made to understand the role of networks in the spread of viruses. Before 2000 epidemic modeling was dominated by compartment models, assuming that everyone can infect everyone else. The emergence of a network-based framework has fundamentally changed this, offering a new level of predictability in epidemic phenomena.

Today epidemic prediction is one of the most active applications of network science. It is the source several fundamental results, covered in this book, that are used to predict the spread of both biological and electronic viruses. The impact of these advances are felt beyond biological viruses. In January 2010 network science tools have predicted the conditions necessary for the emergence of viruses spreading through mobile phones. The first major mobile epidemic outbreak that started in the fall of 2010 in China, infecting over 300,000 phones each day, closely followed the predicted scenario.[3]

Mapping neural network

The lack of maps telling the connections between neurons make the human brain, consisting of hundreds of billions of interlinked neurons, one of the least understood networks from network science perspective. Brain research could become the most prolific application area of network science. Driven by the potential impact of such maps, in 2010 the National Institutes of Health has initiated the *Connectome* project, aimed at developing the technologies that could provide an accurate neuron-level map of mammalian brains.[3]

Management: Uncovering the internal structure of an organization

While traditionally management uses the official chain of command to understand the inner structure of an organization, it is increasingly evident that the informal network, capturing who really communicates with whom, matters even more for the success of a company. Accurate maps of this network can expose lack of communication between key units, can identify individuals who play an outsize role in bringing different departments and products together, and help higher management diagnose diverse organizational issues.

Furthermore, there is increasing evidence in the management literature that the position of an employee within this network correlates with his/her productivity.

Therefore, several dozen consulting companies have emerged with expertise to map out the true structure of an organization. Established consulting firms, from *IBM* to *SAP*, have added social networking capabilities to their consulting business. These companies offer a host of services from identifying opinion leaders to preventing employee churn and from identifying optimal groups for a task to modeling product diffusion. Hence lately network science tools are increasingly indispensable in management and business, enhancing productivity and boosting innovation within an organization.

Network science can therefore offer a microscope for higher management, helping them improve the company's effectiveness by uncovering the true network behind any organization.[3]

3.3 Real Network characteristics

The main goal of network science is to build models than can accurately reproduce real networks properties.

Most networks do not have the comforting regularity of a crystal lattice or the predictable radial architecture of a spider web. Rather, at first inspection most real networks look as if they were spun randomly. Random network theory embraces this apparent randomness by constructing networks that are *truly random*. [4]

A random network consists of N labeled nodes where each node pair is connected with the same probability *p*.

A network can be viewed as a simple object, from a modeling perspective, consisting only of nodes and links. The real challenge is to place the links between the nodes in a way to reproduce the complexity and apparent randomness of real networks. [4]

Real networks are supercritical. Two predictions of random network theory are of special interest for real networks.

- Once the average degree exceeds $\langle k \rangle = 1$, a giant component emerges that contains a finite fraction of all nodes. Hence only for $\langle k \rangle > 1$ the nodes organize themselves into a recognizable network.

- For $\langle k \rangle > lnN$ all components are absorbed by the giant component, resulting in a single connected network.

Measurements indicate that real networks extravagantly exceed the $\langle k \rangle = 1$ threshold. Indeed, sociologists estimate that an average person has around *1,000* acquaintances; a typical neuron is connected to dozens of other neurons, some to thousands; in human cells, each molecule takes part in several chemical reactions, some, like water, in hundreds. [4]

Most networks are in the supercritical regime, meaning that these networks have a giant component, but it coexists with many disconnected components and nodes. This is true only if real networks are random.

The *small world phenomena*, also known as *six degrees of separation*, has long fascinated the general public. It states that if any two individuals are choosen anywhere on earth, a path of at most six acquitances will be found between them.[4]

The fact that individuals who live in the same city are only a few handshakes from each other is by no means surprising. The small world concepts goes even further, stating that even individuals who are on the opposite side of the globe axis are six or fewer hand-shakes from each other.[4]

In the language of network science, small world phenomena implies that the distance between two randomly chosen nodes in a network can be surprisingly short.

Small-world networks tend to contain cliques, and near-cliques, meaning sub-networks which have connections between almost any two nodes within them. This follows from the defining property of a high clustering coefficient. Secondly, most pairs of nodes will be connected by at least one short path. This follows from the defining property that the mean-shortest path length be small. Several other properties are often associated with small-world networks. Typically there is an over-abundance of *hubs* - nodes in the network with a high number of connections (known as high degree nodes). These hubs serve as the common connections mediating the short path lengths between other edges. By analogy, the small-world network of airline flights has a small mean-path length (i.e. between any two cities you are likely to have to take three or fewer flights) because many flights are routed through hub cities.[4]

This property is often analyzed by considering the fraction of nodes in the network that have a particular number of connections going into them (the degree distribution of the network). Networks with a greater than expected number of hubs will have a greater fraction of nodes with high degree, and consequently the degree distribution will be enriched at high degree values. This is known colloquially as a fat-tailed distribution. Specifically, if a network has a degree-distribution which can be fit with a power law distribution, it is taken as a sign that the network is small-world. Networks with power law degree distribution are also known as scale-free networks. Graphs of very different topology qualify as small-world networks as long as they satisfy the two definitional requirements above.[4]

3.4 Complex Network Theory - Definition and properties.

Networks are all around us, and we ourselves, as individuals, are the units of a network of social relationships of all kinds, and as biological systems, the result of a network of biochemical reactions.

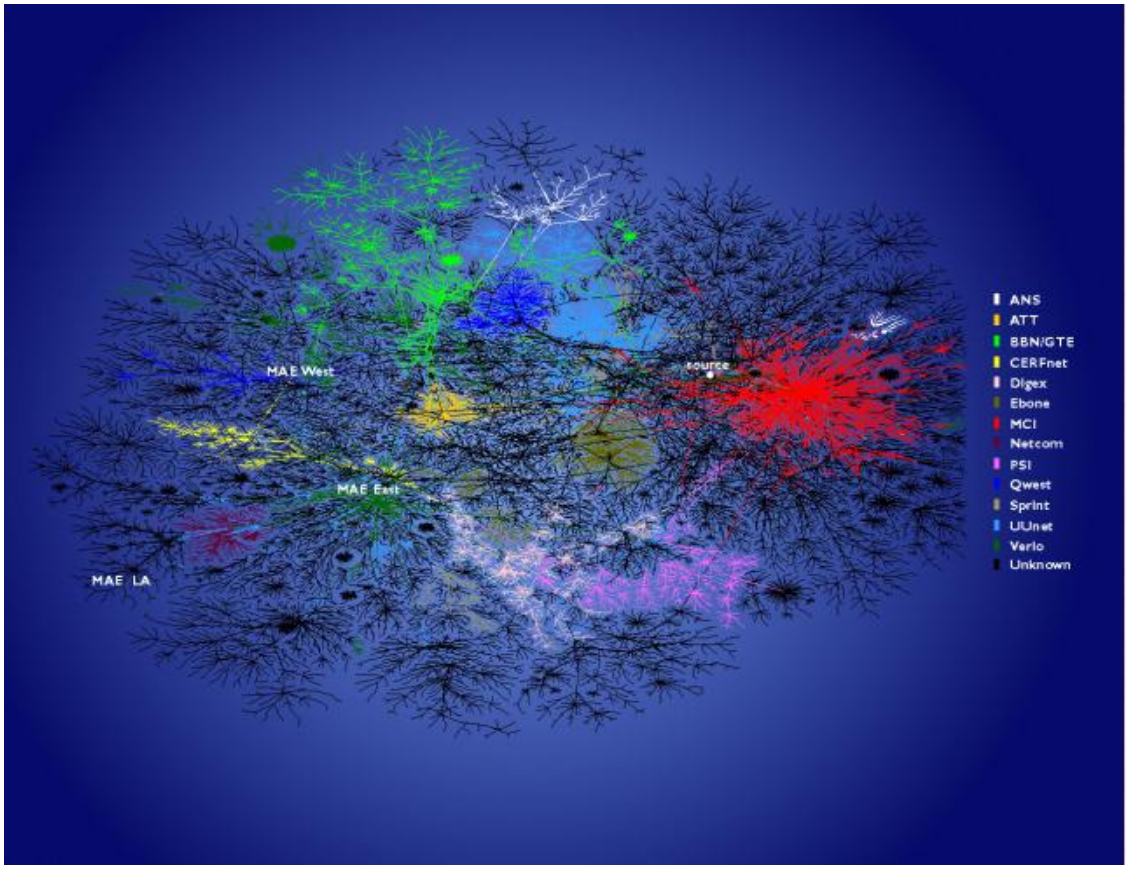
At the beginning, the study of networks was mainly the domain of a branch of discrete mathematics, knows as *graph theory*. In 1736, the Swiss mathematician Leonhard Euler published the solution to the Königsberg bridge problem which consisted in finding one way that traversed each of the bridges of the prussian city of Königsberg exactly once. Since then, the graph theory has experienced rapid and stimulating developments. It has provided answers to a serial of practical issues, such as: what is the maximum flow per unit time from source to sink in a network of pipes, how to color the regions of a map using the minimum number of colors so that neighboring regions receive different colors, or how to fill n jobs with n people, having maximum total utility. [5]

The study of networks has seen important achievements in some specific contexts, as for instance in the social sciences. To begin with, the social networks we inhabit—the collections of social ties among friends — have grown steadily in complexity over the course of human history, due to technological advances facilitating distant travel, global communication, and digital interaction. The past half-century has seen these social networks depart even more radically from their geographic underpinnings, an effect that has weakened the traditionally local nature of such structures but enriched them in other dimensions.

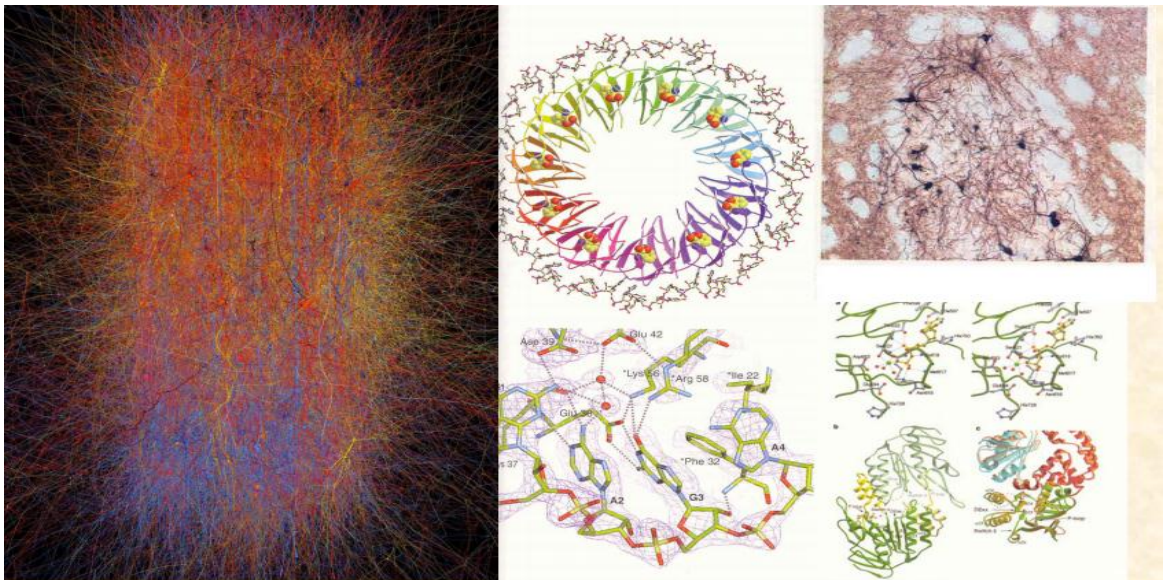
In the last decade, a deep interest was developed to the study of networks with irregular structure, networks that are complex and dynamically evolve in time. The focus moved from the analysis of small networks to that of systems with thousands or millions of vertices.

Complex networks are the skeleton of complex systems in the real-world. They are known as networks with non-trivial topology and dynamics. It is very useful, maybe mandatory, to have an image of the complex network we want to study, in order to be able to explore and manipulate its properties, to filter, navigate and cluster the network data.

To be able to understand what a complex network means, I will give a few examples of networks we find in different fields.

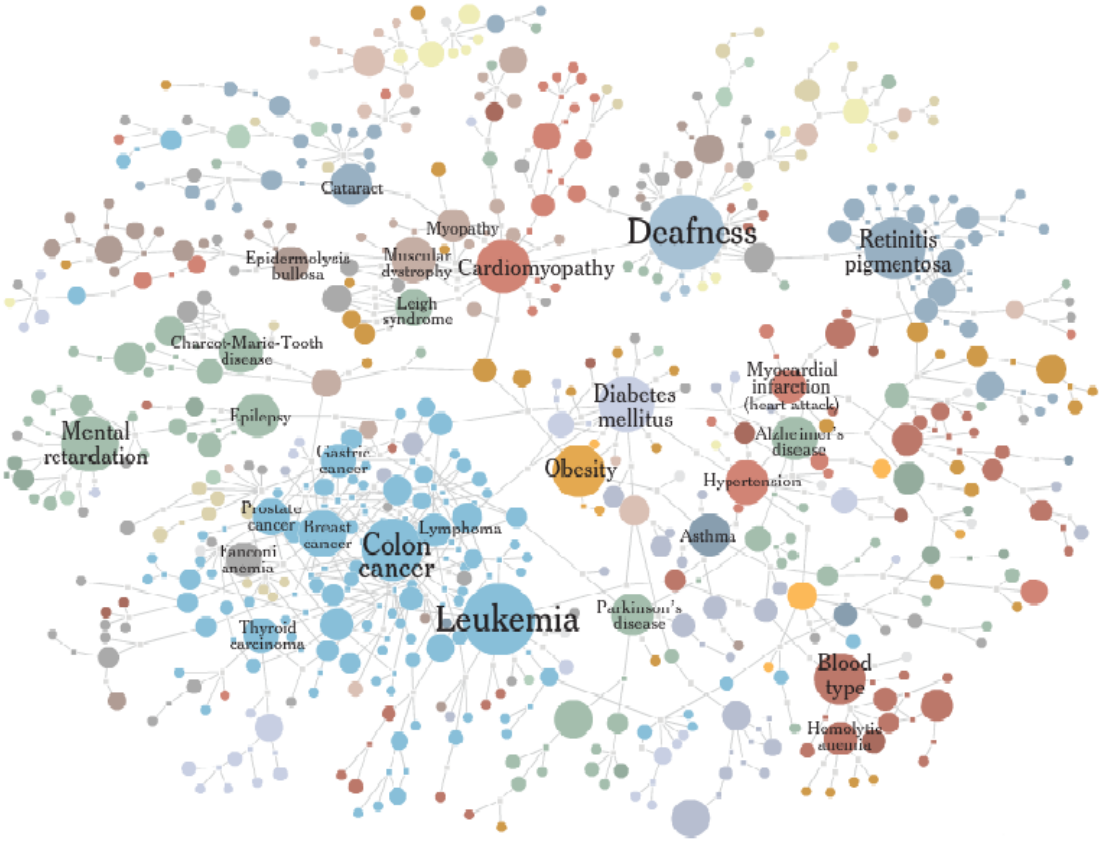


Complex Network: Internet



Biological Network

Twitter Social Network



Human Disease Network

There are two basic parameters which describe a network: *the number of nodes* and *the number of links.* The number of nodes is often denoted with N and represents the number of components in the system. N is also called the size of the network. The number of links, L, counts the total number of the connections between the components of the system. Links can be directed or undirected. An example of a network with directed links is the World Wide Web network, from one document to another. On the other hand, transmission lines on an electrical circuit are undirected links (the electric current flows in both directions).

A directed network has all its links directed, while an undirected one has all of its links undirected.

Other key properties of nodes which are helpful in studying networks, are the degree, the average degree and the degree distribution. The degree of a node is the number of links it has to other nodes. For an undirected network, the number of nodes can be represented like this:[6]

$$L = \frac{1}{2}\sum_{i=1}^{n} k_i \text{ , where } k_i \text{ is the degree of node } i$$

Two other important properties of a network are the density of the network and its average degree.

The density of the network measures how many edges are compared to the maximum possible number of edges between vertices. The average degree is another measure of how many edges are compared to the number of vertices. For an undirected network, it can be represented like this:[6]

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

In directed networks there is a distinction between the incoming degree, representing the number of links that point a node, and the out coming degree, representing the number of links that point from a node to other nodes.

The degree distribution provides the probability that a node randomly selected from the network has degree k. The degree distribution has taken an important role in network theory following the discovery of scale-free networks. Another reason for its importance is that the calculation of most networks requires to know the p_k. For example the average degree can be also known as:

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

A full description of a network requires to keep track of its links. For mathematical purposes, a network is usually represented by its adjacency matrix. The adjacency matrix has N rows and N columns, its elements being: A_{ij} = 1, if there is a link between node i and node j, Aij = 0, if node i is not connected to node j. [6]

The adjacency matrix for an undirected network has two entries for each link. The degree k_i of a node i can be directly determined from the elements of the adjacency matrix. For undirected networks a node's degree is a sum over either the rows or the columns of the matrix.[6]

$$k_i = \sum_{i=1}^{N} A_{ij} = \sum_{j=1}^{N} A_{ij}$$

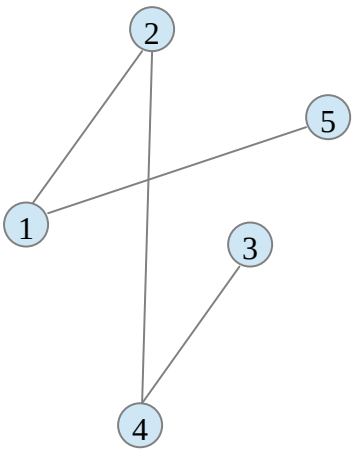
For directed networks the sums over the adjacency matrix' rows and columns provide the incoming and outgoing degrees:

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \qquad k_i^{out} = \sum_{i=1}^{N} A_{ij}$$

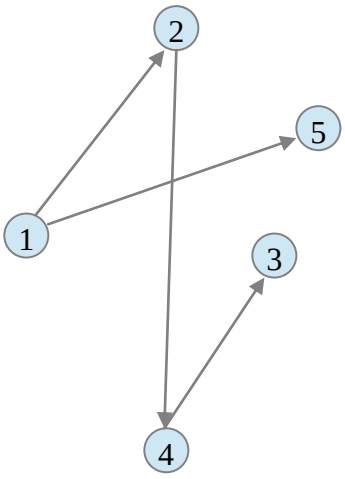
Given the fact that in an undirected network the number of outgoing links equals the number of incoming links, results that:

$$2L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j}^{N} A_{ij}$$

Undirected network



Directed network



Networks can be weighted or un-weighted. For an un-weighted network, all links have the same weight, i.e A_{ij} =1. Yet, in many applications, it is useful to have weighted networks, where each link (i,j) has a unique weight w_{ij}. [6]

For weighted networks the elements of the adjacency matrix carry the weight of the link, $A_{ij} = w_{ij}$.

Most networks of scientific interest are weighted, but the appropriate weights can not be always measured.

In physical systems the components are characterized by obvious distances. In networks, the physical distances are replaced by the path length. A path is a route that runs along the links of the network, its length representing the number of links it contains.[6]

A path can intersect itself and pass through the same link repeatedly. In network science, paths play an important role.

Shortest Path (also called geodesic path) between nodes i and j is the path with fewest number of links. The shortest path is often called the distance between nodes i and j, and is denoted simply with *d*.[6]

The shortest path never contains loops or intersect itself.

In an undirected network $d_{ij} = d_{ji}$, i.e. the distance between node *i* and *j* is the same as the distance between node *j* and *i*. In a directed network often d_{ij} differs from d_{ji}. Furthermore, in a directed network the existence of a path from node *i* to node *j* does not guarantee the existence of a path from *j* to *i*.[6]

In real networks, frequently it is needed to determine the distance between two nodes. If we deal with a small network, this is an easy task. In contrast, for networks of millions of nodes, finding the shortest path between two nodes can be really difficult and time consuming. The length of the shortest path and the number of such paths can be formally obtained from the networks adjacency matrix. In practice, it is often used the Breadth First Search (BDF) algorithm to measure the distance between two nodes. This algorithm works like this: starting from a randomly chosen node labeled "0", are identified all its neighbors, labeling them "1". Then, the unlabeled neighbors off all labeled "1" nodes are labeled with "2", and so on, in each iteration increasing the labels, until no node is left unlabeled. The

length of the shortest path or the distance between node "0" and some other node *i* in the network is given by the label of node *i*. [6]

Properties of Complex Networks

- **Components and Connectedness**

Most networks are built to ensure connectedness, and this is an important utility. For an undirected network two nodes i and j are connected if there is a path between them on the graph. They are disconnected if such a path does not exist, in which case we have $d_{ij} = \infty$.[6]

A network is called connected if all pairs of nodes in the network are connected. It is disconnected if there is at least one pair with $d_{ij} = \infty$.

A component is a subset of nodes in a network, so that there is a path between any two nodes that belong to the component, but one cannot add any more nodes to it that would have the same property. If a network consists of two components, a properly placed single link can connect them, making the network connected. Such a link is called a bridge. In general a bridge is any link that, if cut, disconnects the graph.[6]

A small network can be easily visually inspected and one can decide if it is connected or disconnected. For a large network, consisting of millions of nodes, connectedness is a challenging question. There are several mathematical tools that can be used to identify the connected components of a network. For a disconnected network, its adjacency matrix can be rearranged into a block diagonal form, so that all nonzero elements in the matrix are contained in square blocks along the matrix' diagonal and all other elements are zero. Each square block corresponds to a component. Tools of linear algebra can be used to decide if the adjacency matrix is block diagonal, helping to identify the connected components. In practice, for large networks, the Breadth First Algorithm is more efficient to identify the components.[6]

- **Clustering Coefficient**

The local clustering coefficient expresses the degree to which the neighbors of a given node link to each other. For a node *i* with degree k_i , the local clustering coefficient is defined as:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

where L_i represents the number of links between the k_i neighbors of node *i*.[6]

C_i is between 0 and 1. $C_i = 0$ if none of the neighbors of node i link to each other; $C_i = 1$ if the neighbors of node i form a complete graph, i.e. they all link to each other. In general C_i is the probability that two neighbors of a node link to each other: C = 0.5 implies that there is a 50% chance that two neighbors of a node are linked. In summary, C_i measures the network's local density: the more densely interconnected the neighborhood of node i, the higher is C_i.[6]

The degree of clustering of a whole network is captured by the average clustering coefficient, $\langle C \rangle$, representing the average of C_i over all nodes i = 1, ...,N.

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

In line with the probabilistic interpretation $\langle C \rangle$ is the probability that two neighbors of a randomly selected node link to each other.[6]

The clustering coefficient can be generalized to directed and weighted networks as well.

In the network literature one also often encounters the *global clustering coefficient*. It measures the total number of closed triangles in a network. Indeed, *Li* is the number of triangles that node *i* participates in, as each link between two neighbors of node *i* closes a triangle. Hence the degree of a network's global clustering is captured by the global clustering coefficient, defined as

$$C = \frac{3xNumberOfTriangles}{NumberOfConnectedTriples}$$

where a connected triplet consists of three nodes that connected by two (open triplet) or three (closed triplet) undirected links. For example, an A, B, C triangle is made of three triples, ABC, BCA and CAB. In contrast a chain of connected nodes A, B, C, in which B connects to A and C but A does not link to C forms a single open triplet. The roots of the global clustering coefficient go back to the social network literature of the 1940s, hence C is often called the number of transitive triplets.[6]

The average clustering coefficient and the global clustering coefficient are not equivalent.

Indeed, for a network that is a double star consisting of N nodes, where nodes 1 and 2 are joined to each other and to all other vertices, and there are no other links. Then the local clustering coefficient C_i is 1 for i ≥ 3 and 2/(N − 1) for $i = 1, 2$. It follows that the average clustering coefficient of the network is $\langle C \rangle = 1 - O(1)$, while the global clustering coefficient gives $C \sim 2/N$. In less extreme networks the definitions will give more comparable values, but they will still differ from each other. [6]

- **Centrality**

The centrality of a node measures how important that node is for the network. The more "central" it is, the more it influences the whole network. The idea of centrality comes from social networks. There are different types of centrality:

1. *Degree Centrality*. It reflects how many ties a node has to other nodes. Nodes who have more ties may have multiple alternative ways and resources to reach goals. The degree centrality for an undirected network is straightforward. If *node1* is connected to *node2*, then by definition *node2* is connected to *node1*. For a directed network we can define the *indegree centrality* and the *outdegree centrality.* Indegree centrality is a count of the number of edges directed to the node and outdegree centrality refers to the number of ties that the node directs to other nodes.

2. *Betweenness Centrality* is based on shortest paths in a network. It is a measure of the extent to which a node is connected to other nodes that are not connected to each other. Betweenness centrality can also be viewed as a measure of network resilience – it tells how many geodesic paths will get longer when a vertex is removed from the network. Betweenness appears to follow a power law for many networks and propose a classification of networks into two kinds based on the exponent of this power law.

3. *Closeness Centrality* can be calculated as a measure of inequality in the distribution of distances across the objects. These measures rely on the sum of the geodesic distances from each node to all others. The idea of closeness centrality is that nodes are more central if they can reach other nodes "easily".

4. *Eigenvector Centrality* is one method of computing the "centrality", or approximate importance, of each node in a graph. The assumption is that each node's centrality is the sum of the centrality values of the nodes that it is connected to. The eigenvector approach to

measure closeness uses a factor analytic procedure to discount closeness to small local subnetworks.

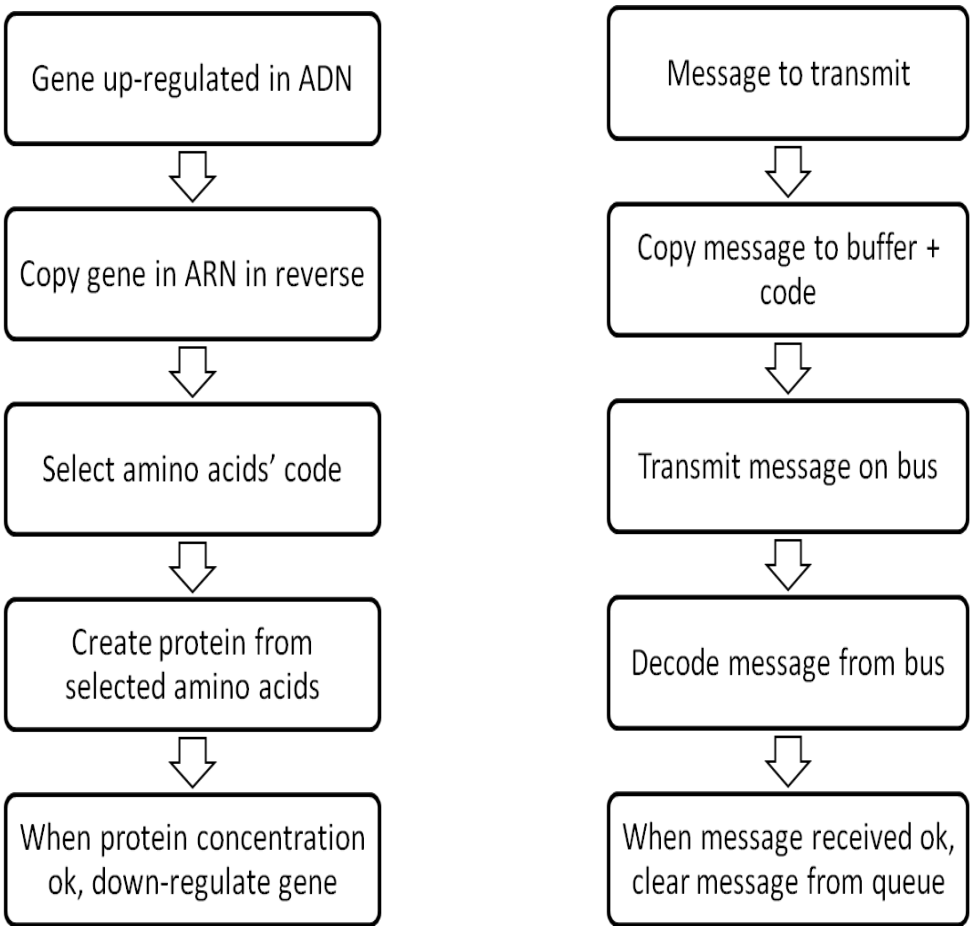
- **Modularity** is a measure for evaluation of community decomposition. It

measures the quality of a division of a network into groups or communities. The idea behind the definition of modularity is simple. A set of nodes form a community in the sense of modularity if the fraction of links inside the community is higher than expected in a network considered as reference. Modularity is defined only for undirected graphs. A node can be placed in just one community at a time, while nodes in real networks usually belong to many communities.[7]

4. Application proposed

The overall aim of this thesis is to provide a solution in modeling and analyzing different metabolic networks.

After I did some research about gene expression, I reached to a simplified model of this process and found certain similarities with processes from computer science.



First of all, I did some research about gene expression process, which is very complex and depends on many factors, both internal and external. I reached to a simplified model that describes it, and found similarities with processes from computer science. For example, a gene being up-regulated in DNA is equivalent to a message we want to transmit on a bus, copying the gene in ARN in reverse means copying the message to a buffer. The process of selecting amino acids' code and creating protein (genes products) from the selected amino acids, is similar to the process of a transmitting a message on a bus and decoding it from the bus. Finally, when a protein reaches a specific threshold of concentration and the gene is down-regulated, means that the message has been received ok and so it can be removed from the queue.

The tool I used to explore the networks is called Gephi, an interactive open-source platform which helps data representation and analysis, manipulating structures and intuitively discover patterns. It is a complementary tool to traditional statistics.

The application starts with a program generated in Visual Basic 2007, which parses a database and generates a spreadsheet .csv file ready to be imported in Gephi. Afterwards, using layout algorithms and suitable metrics and filters Gephi provides, a complex network with an intuitive representation is generated.

The first network studied and modeled is the clock gene network. More than a dozen clock genes were identified by researchers and a biochemical feedback mechanism was defined. Recent models consider the clock a biochemical and cellular oscillator, and also a genetic network.

The clock gene network has complex regulatory architecture and perturbation, also known as knockdown, of one clock component can lead to changes in the levels of others. Knockdown of a clock gene may up-regulate, down-regulate or have no effect on the expression of other genes.

Up-regulation is a metabolic process within a gene triggered by a signal, which results in increased expression of that gene and the proteins it encodes. Down-regulation is the inverse process, resulting in decreased gene and corresponding protein expression.

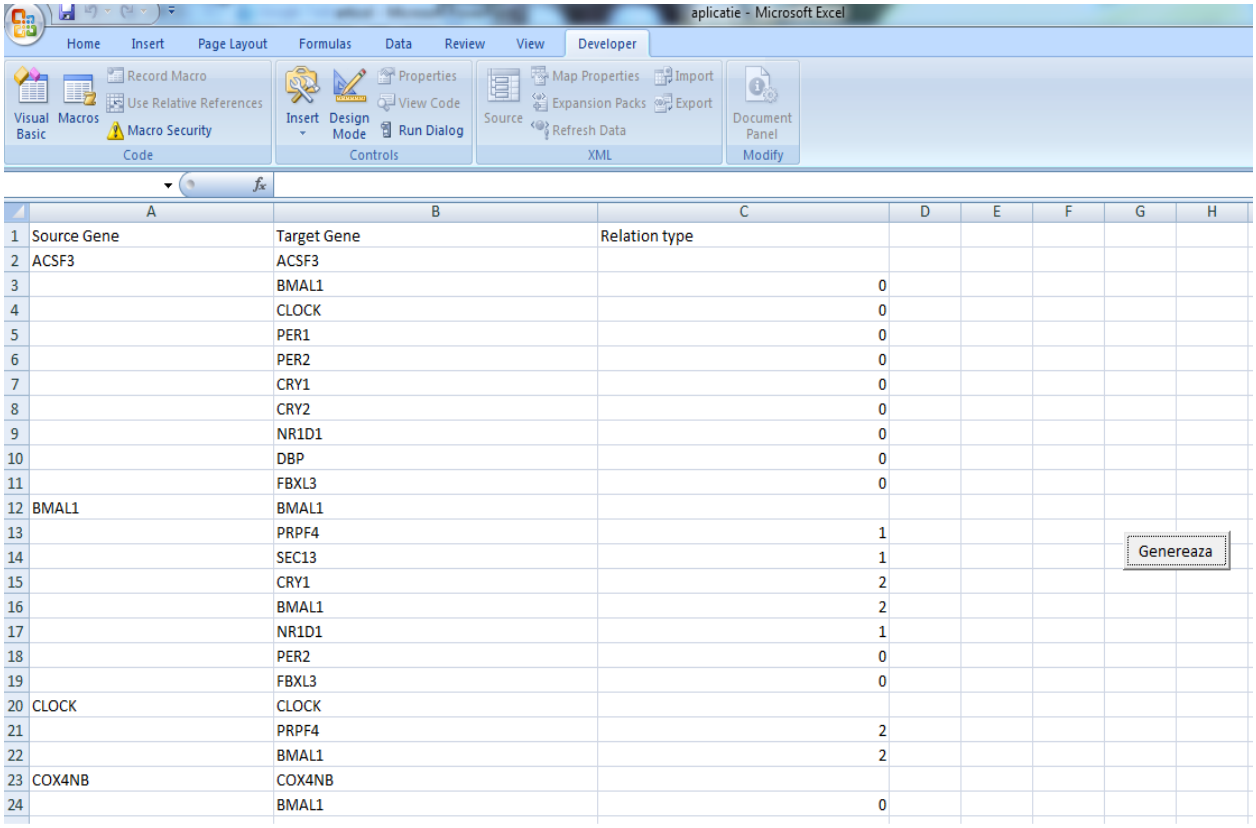
It is assumed that there are additional clock genes and modifiers of the circadian rhythm in mammals. Biochemical experiments were conducted and researchers found a subset of genes with a dosage-dependent effect on circadian oscillator function. Based on their results, I first created a database which reflects the dose-dependent effects (up-regulation, down-regulation, no effect) on known clock gene expression.

	BMAL1	CLOCK	PER1	PER2	CRY1	CRY2	NR1D1	DBP	FBXL3
CRY2						down			
FBLXL3				down	down		down	down	down
POLR3F							down	down	
PRPF4	up	down					down	down	
SEC13	up				up			down	
CRY1	down			up	down	up			
ACSF3									
MPG			down				down	down	
BMAL1	down	down	up		up	up	down	down	
NR1D1	up					up	down		
COX4NB				down			down	down	

Table 1. Summary of dose-dependent effects on known clock gene expressions [7]

Information contained by the table was introduced into a spreadsheet which afterwards was processed and resulted into a .csv file containing data that Gephi needs to generate a complex network.

The “target” genes are those whose knockdown determines changes on the expression of other genes, also known as “source” genes. “Relation type” encodes the effect a target gene has upon a source gene: 0 – no effect, 1 – up-regulation, 2- down-regulation.

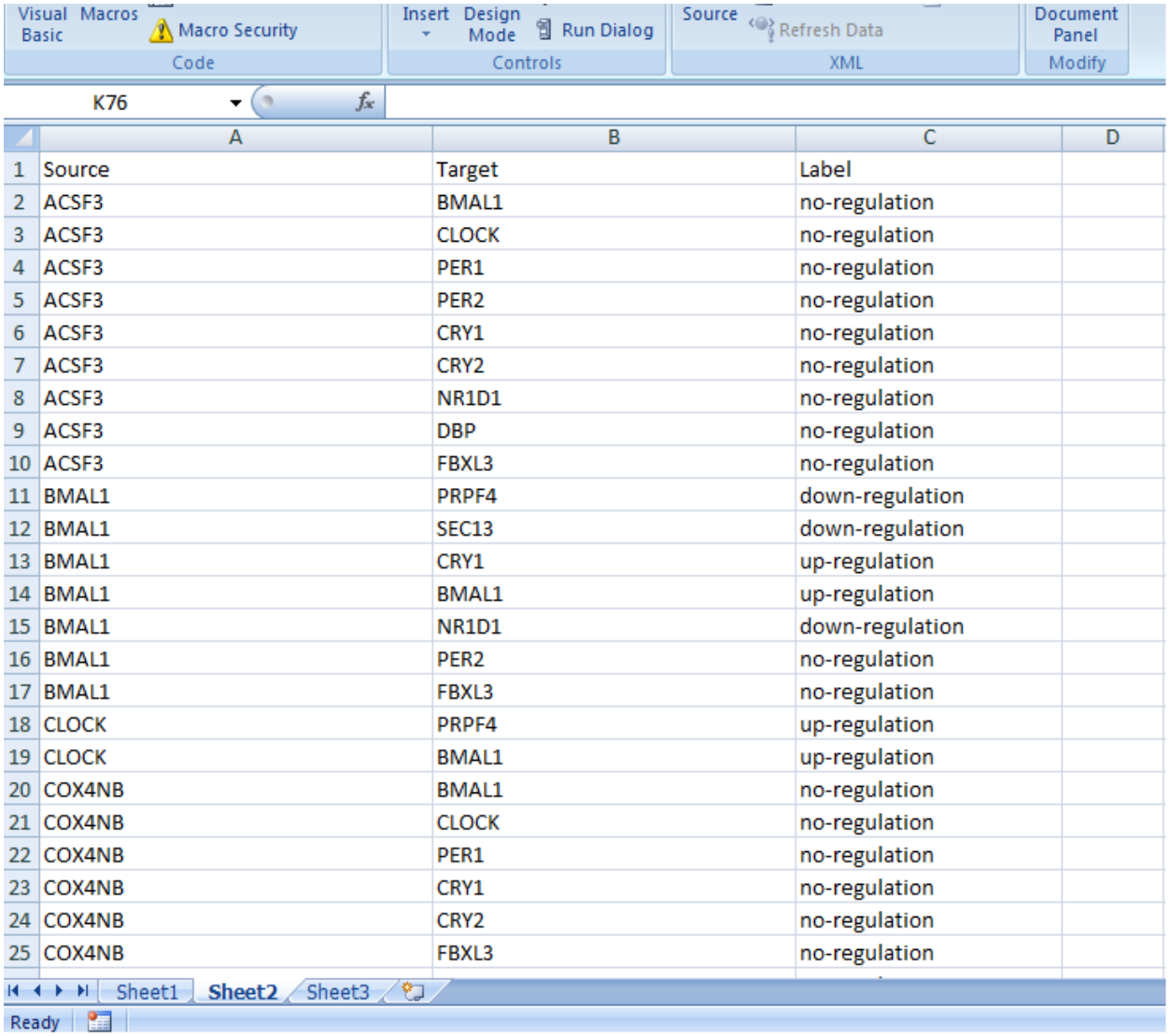


	A	B	C	D	E	F	G	H
1	Source Gene	Target Gene	Relation type					
2	ACSF3	ACSF3						
3		BMAL1	0					
4		CLOCK	0					
5		PER1	0					
6		PER2	0					
7		CRY1	0					
8		CRY2	0					
9		NR1D1	0					
10		DBP	0					
11		FBXL3	0					
12	BMAL1	BMAL1						
13		PRPF4	1					
14		SEC13	1					
15		CRY1	2					
16		BMAL1	2					
17		NR1D1	1					
18		PER2	0					
19		FBXL3	0					
20	CLOCK	CLOCK						
21		PRPF4	2					
22		BMAL1	2					
23	COX4NB	COX4NB						
24		BMAL1	0					

The button “Genereaza” parses the database and generates the .csv spreadsheet to be imported in Gephi.

```
Private Sub CommandButton1_Click()
pozitie = 2
sheetInput = 1
numeGena = Worksheets(sheetInput).Range("A2").Value
genaSursa = Worksheets(sheetInput).Range("B2").Value
Label = Worksheets(sheetInput).Range("C2").Value

contor = 3
listaFinala = 0
While (listaFinala = 0)
    genaTarget = Worksheets(sheetInput).Range("B" & contor).Value
    numeGena = Worksheets(sheetInput).Range("A" & contor).Value
    Label = Worksheets(sheetInput).Range("C" & contor).Value
    contor = contor + 1
    If (genaTarget = "") Then
        listaFinala = 1
        Else
        If (numeGena = "") Then
                Worksheets(2).Range("A" & pozitie).Value = genaSursa
                Worksheets(2).Range("B" & pozitie).Value = genaTarget
                If (Label = 0) Then
                    Worksheets(2).Range("C" & pozitie).Value = "no-
regulation"
                End If

                If (Label = 1) Then
                    Worksheets(2).Range("C" & pozitie).Value = "down-
regulation"
                End If

                If (Label = 2) Then
```

```
                        Worksheets(2).Range("C" & pozitie).Value = "up-
regulation"
                    End If
                    Worksheets(2).Range("D" & pozitie).Value = "Directed"
                pozitie = pozitie + 1
                Else
                genaSursa = genaTarget
                End If
            End If

        Wend

        End Sub
```



	A	B	C	D
1	Source	Target	Label	
2	ACSF3	BMAL1	no-regulation	
3	ACSF3	CLOCK	no-regulation	
4	ACSF3	PER1	no-regulation	
5	ACSF3	PER2	no-regulation	
6	ACSF3	CRY1	no-regulation	
7	ACSF3	CRY2	no-regulation	
8	ACSF3	NR1D1	no-regulation	
9	ACSF3	DBP	no-regulation	
10	ACSF3	FBXL3	no-regulation	
11	BMAL1	PRPF4	down-regulation	
12	BMAL1	SEC13	down-regulation	
13	BMAL1	CRY1	up-regulation	
14	BMAL1	BMAL1	up-regulation	
15	BMAL1	NR1D1	down-regulation	
16	BMAL1	PER2	no-regulation	
17	BMAL1	FBXL3	no-regulation	
18	CLOCK	PRPF4	up-regulation	
19	CLOCK	BMAL1	up-regulation	
20	COX4NB	BMAL1	no-regulation	
21	COX4NB	CLOCK	no-regulation	
22	COX4NB	PER1	no-regulation	
23	COX4NB	CRY1	no-regulation	
24	COX4NB	CRY2	no-regulation	
25	COX4NB	FBXL3	no-regulation	

Fig. 1 Resulted .csv file

When you first import a dataset to Gephi, the nodes' position is random. Gephi provides layout algorithms that can be used to set the graph's shape.

Force Atlas is a commonly used layout algorithm with an easy principle: linked nodes attract each other while non-linked nodes are pushed apart.

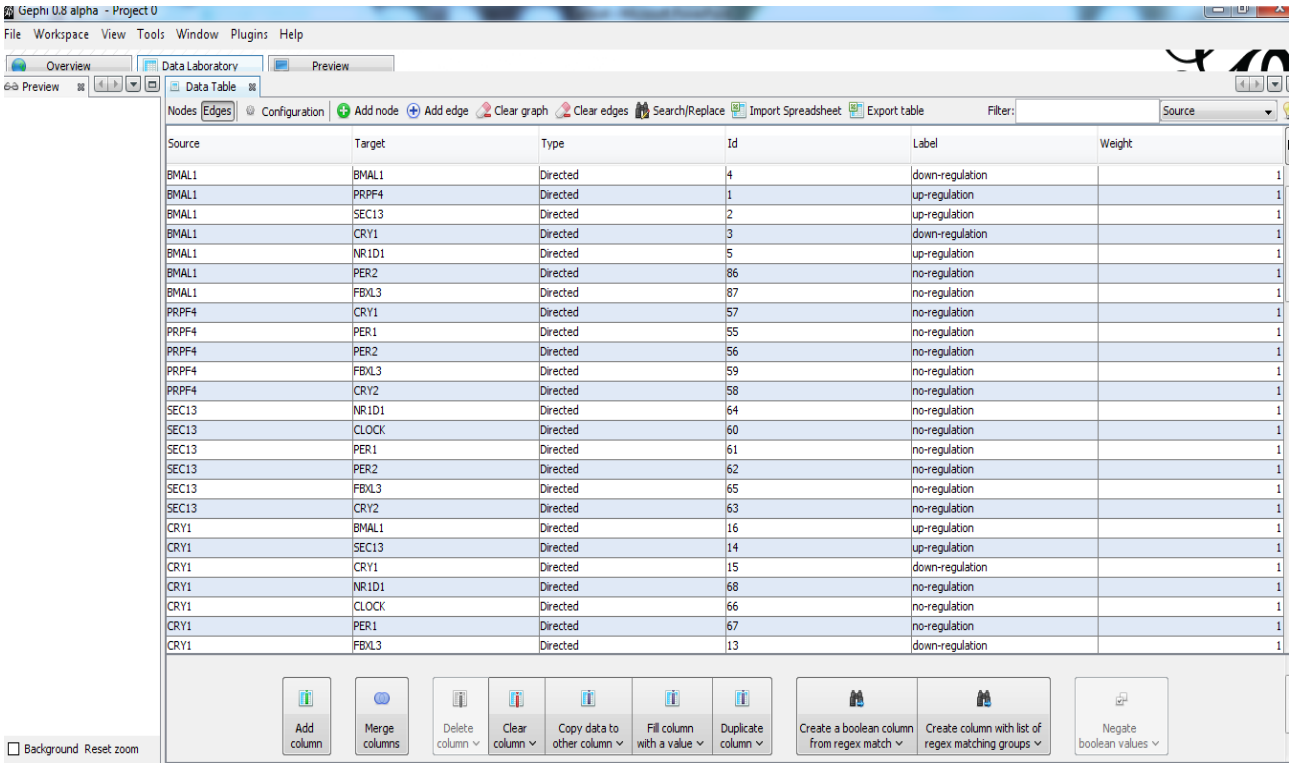


Fig. 2 Importing spreadsheet in Gephi

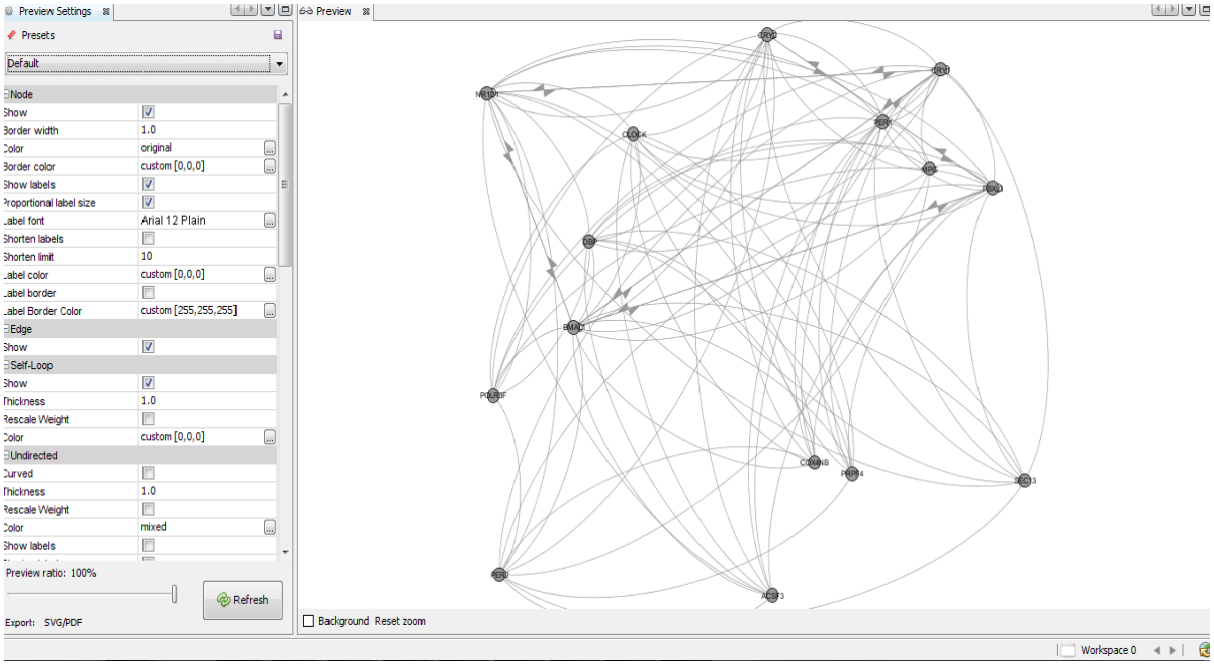


Fig. 3 First view on network in Gephi

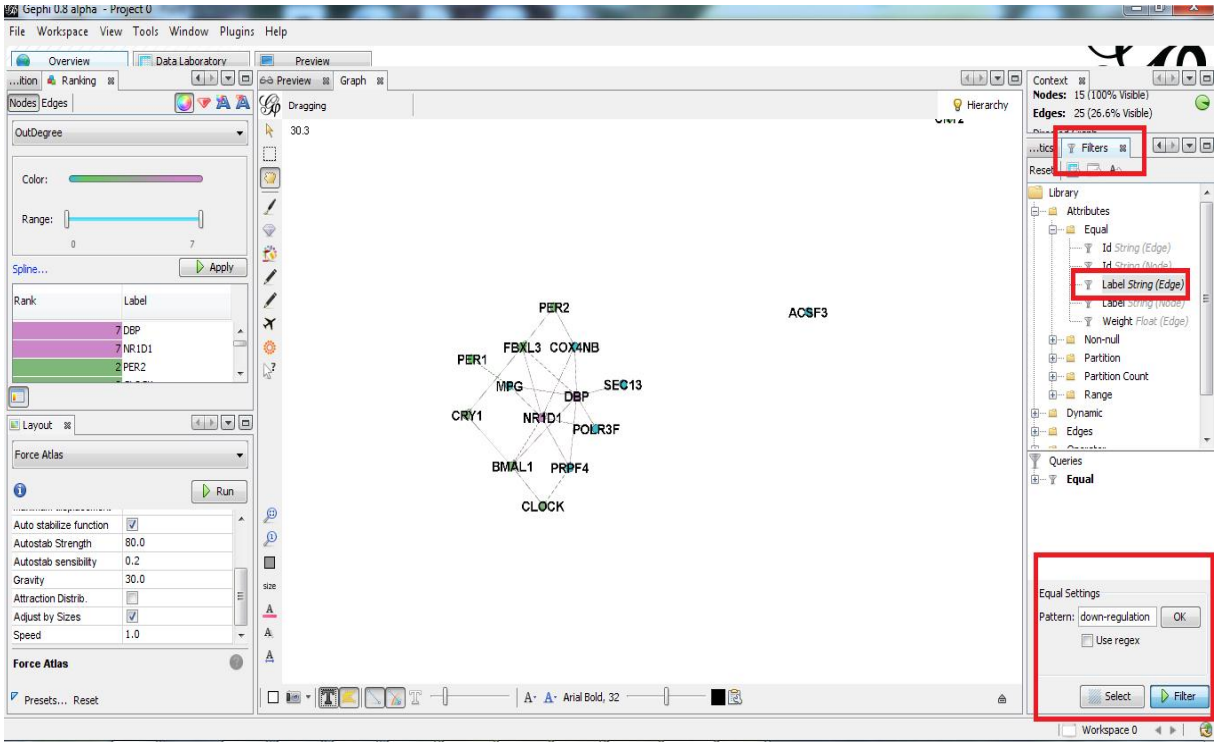
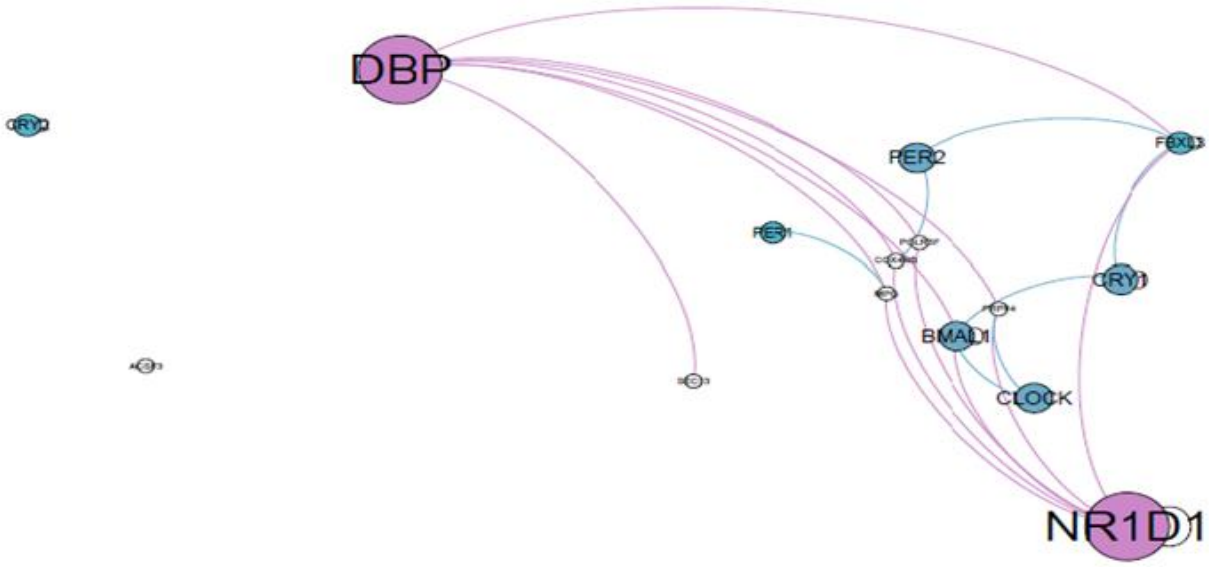


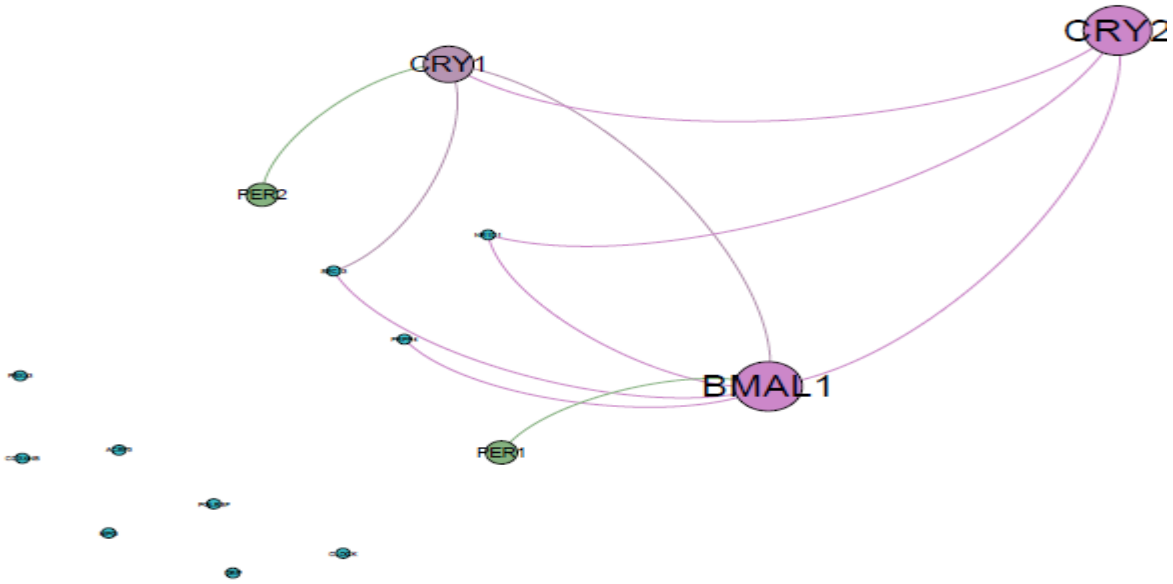
Fig 4. Force Atlas algorithm applied. Network filtered by edge label.

The first network I generated was filtered in order to obtain the most "down-regulated" genes. Gephi provides the possibility to export the resulted network as a SVG or PDF file.



From the resulted network, it can easily be determined that the most "down-regulated" genes are DBP and NR1D1.

The same steps were followed in order to obtain the network of most "up-regulated" genes: CRY2, CRY1, BMAL1.



Researchers have also identified an extensive set of proteins that interacted with gene products or known clock components. The resulted table (Table 2) was transposed into an expanded static clock gene interaction network (Figure 6). The main drawback of the representation is that nodes of basic importance to the network, structure communities or significant interactions are difficult to be observed from it. I proposed another view of the network, manipulating using Gephi the same database. The resulted complex network is more intuitive.

Clock Ref	Clock Symbol	Interactant Ref	Interactant Symbol	Hit Ref	Hit Symbol
NM_001178	ARNTL	XM_374491	PPP1R9A	NM_018067	MAP7D1
NM_001178	ARNTL	NM_021009	UBC	NM_006142	SFN
NM_001178	ARNTL	NM_003884	PCAF	NM_000059	BRCA2
NM_001178	ARNTL	NM_003884	PCAF	NM_003496	TRRAP
NM_001178	ARNTL	NM_002957	RXRA	NM_002434	MPG
NM_001178	ARNTL	NM_001530	HIF1A	NM_003968	UBE1C
NM_001178	ARNTL	NM_001530	HIF1A	NM_017902	HIF1AN
NM_001178	ARNTL	NM_001530	HIF1A	NM_015179	RRP12
NM_001178	ARNTL	NM_005348	HSP90AA1	NM_177559	CSNK2A1
NM_001178	ARNTL	NM_005348	HSP90AA1	NM_005334	HCFC1
NM_004898	CLOCK	NM_015641	TES	NM_005697	SCAMP2
NM_004898	CLOCK	NM_153260	FLJ36812	NM_020772	NUFIP2
NM_004898	CLOCK	XM_377778	LOC402110	NM_005334	HCFC1
NM_004898	CLOCK	NM_002835	PTPN12	NM_001424	EMP2
NM_004898	CLOCK	NM_020183	ARNTL2	NM_001430	EPAS1
NM_002518	NPAS2	NM_003884	PCAF	NM_000059	BRCA2
NM_002518	NPAS2	NM_003884	PCAF	NM_003496	TRRAP
NM_002518	NPAS2	NM_002957	RXRA	NM_002434	MPG
NM_002518	NPAS2	NM_181659	NCOA3	NM_001556	IKBKB
NM_002616	PER1	XM_290629	C14ORF78	NM_014258	SYCP2
NM_002616	PER1	NM_002468	MYD88	NM_016166	PIAS1
NM_002616	PER1	NM_014676	PUM1	NM_018067	MAP7D1
NM_002616	PER1	NM_013333	EPN1	NM_005826	HNRPR
NM_002616	PER1	XM_290629	C14ORF78	NM_001281	TBCB
NM_002616	PER1	NM_021906	USP9X	NM_004697	PRPF4
NM_002616	PER1	NM_006346	C13ORF24	NM_006609	MAP3K2
NM_002616	PER1	NM_015638	TRPC4AP	NM_005781	TNK2
NM_002616	PER1	NM_000249	MLH1	NM_001430	EPAS1
NM_002616	PER1	NM_021906	USP9X	NM_002896	RAB3A
NM_002616	PER1	XM_290629	C14ORF78	NM_017838	NOLA2
NM_002616	PER1	NM_014676	PUM1	NM_002896	RAB3A
NM_002616	PER1	NM_006346	C13ORF24	NM_006210	PEG3
NM_002616	PER1	NM_015927	TGFB1I1	NM_013390	TMEM2
NM_002616	PER1	NM_021906	USP9X	NM_133370	YTED1
NM_022817	PER2	NM_001895	CSNK2A1	NM_005953	MT2A
NM_022817	PER2	NM_001895	CSNK2A1	NM_004327	BCR
NM_022817	PER2	NM_178552	MGC35206	NM_018449	UBAP2
NM_022817	PER2	NM_004572	PKP2	NM_014516	CNOT2
NM_022817	PER2	NM_003906	MCM3AP	NM_005082	KALPHA1
NM_016831	PER3	NM_002616	PER1	NM_013314	BLNK
NM_016831	PER3	NM_002616	PER1	NM_015179	RRP12
NM_016831	PER3	NM_020765	RBAF600	NM_002532	NUP88
NM_004075	CRY1	NM_021138	TRAF2	NM_002460	IRF4
NM_004075	CRY1	NM_021105	PLSCR1	NM_005157	ABL1
NM_004075	CRY1	NM_021138	TRAF2	NM_001556	IKBKB
NM_021117	CRY2	NM_006247	PPP5C	NM_001316	CSE1L
NM_005126	NR1D2	NM_177995	EPB41L1	NM_017920	URG4
NM_021724	NR1D1	NM_005087	FXR1	NM_018449	UBAP2
NM_021724	NR1D1	NM_005087	FXR1	NM_017761	PNRC2
NM_005126	NR1D2	NM_001222	CAMK2G	NM_005781	TNK2
NM_005126	NR1D2	NM_005791	MPHOSPH10	NM_018360	CXORF16
NM_006914	RORB	NM_014071	NCOA6	NM_003259	ICAM5
NM_006914	RORB	NM_003388	CYLN2	NM_001894	CSNK1E
NM_001893	CSNK1D	XM_290629	C14ORF78	NM_014258	SYCP2
NM_001893	CSNK1D	XM_290629	C14ORF78	NM_001281	TBCB
NM_001893	CSNK1D	NM_181870	DVL1	NM_001895	CSNK2A1
NM_001893	CSNK1D	NM_004423	DVL3	NM_001895	CSNK2A1
NM_001893	CSNK1D	XM_290629	C14ORF78	NM_017838	NOLA2
NM_001894	CSNK1E	NM_181523	PIK3R1	NM_013314	BLNK
NM_001894	CSNK1E	NM_000546	TP53	NM_000059	BRCA2
NM_001894	CSNK1E	NM_000546	TP53	NM_006142	SFN
NM_001894	CSNK1E	NM_000546	TP53	NM_005157	ABL1
NM_001894	CSNK1E	NM_000546	TP53	NM_004327	BCR
NM_001894	CSNK1E	NM_000546	TP53	NM_000791	DHFR
NM_001894	CSNK1E	NM_000546	TP53	NM_001261	CDK9
NM_001894	CSNK1E	NM_000546	TP53	NM_139046	MAPK8
NM_001894	CSNK1E	NM_000546	TP53	NM_005381	NCL
NM_001894	CSNK1E	NM_003502	AXIN1	NM_001892	CSNK1A1
NM_001894	CSNK1E	NM_032421	CYLN2	NM_014117	PRO0149
NM_001894	CSNK1E	NM_181870	DVL1	NM_001895	CSNK2A1
NM_001894	CSNK1E	NM_004422	DVL2	NM_001895	CSNK2A1

Table 2. List of clock protein interaction[7]

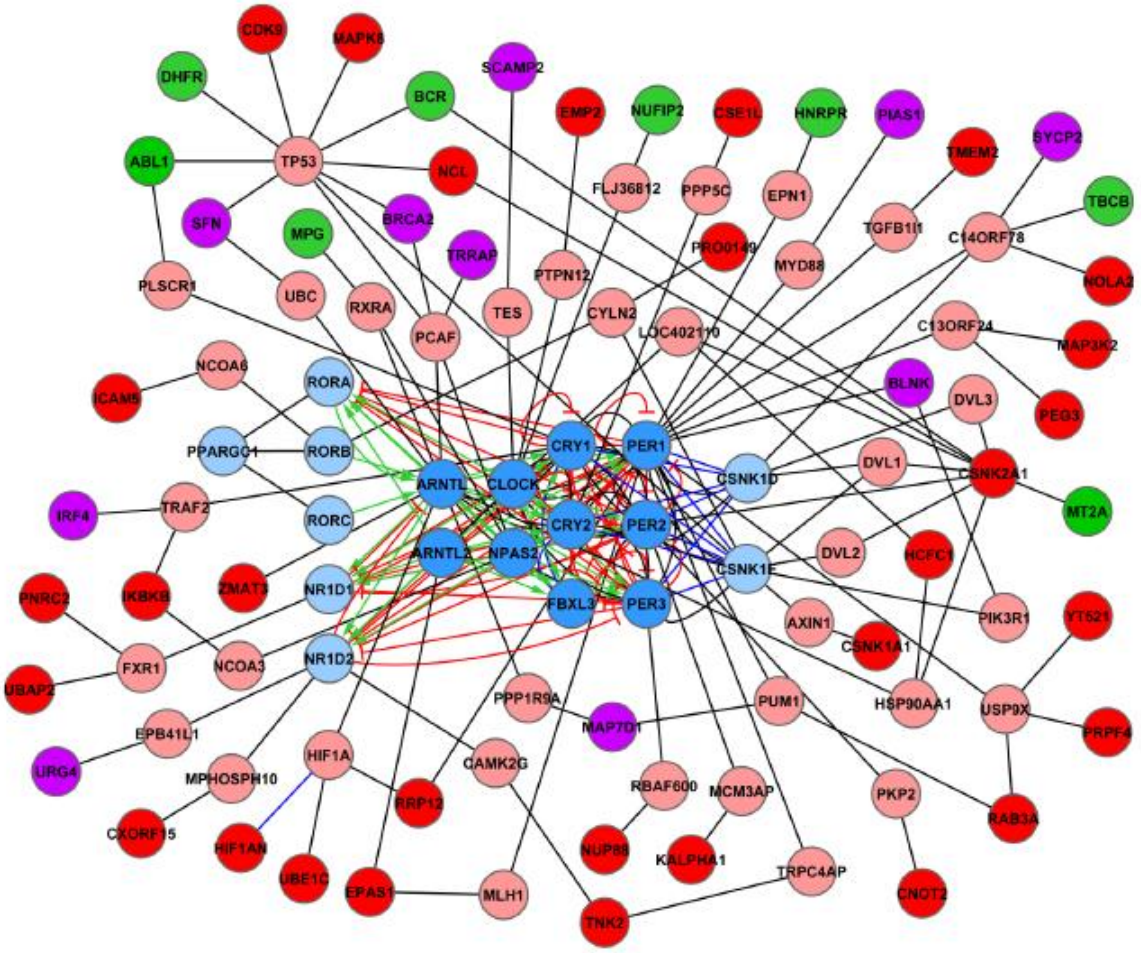


Fig. 5 The expanded clock gene network [7]

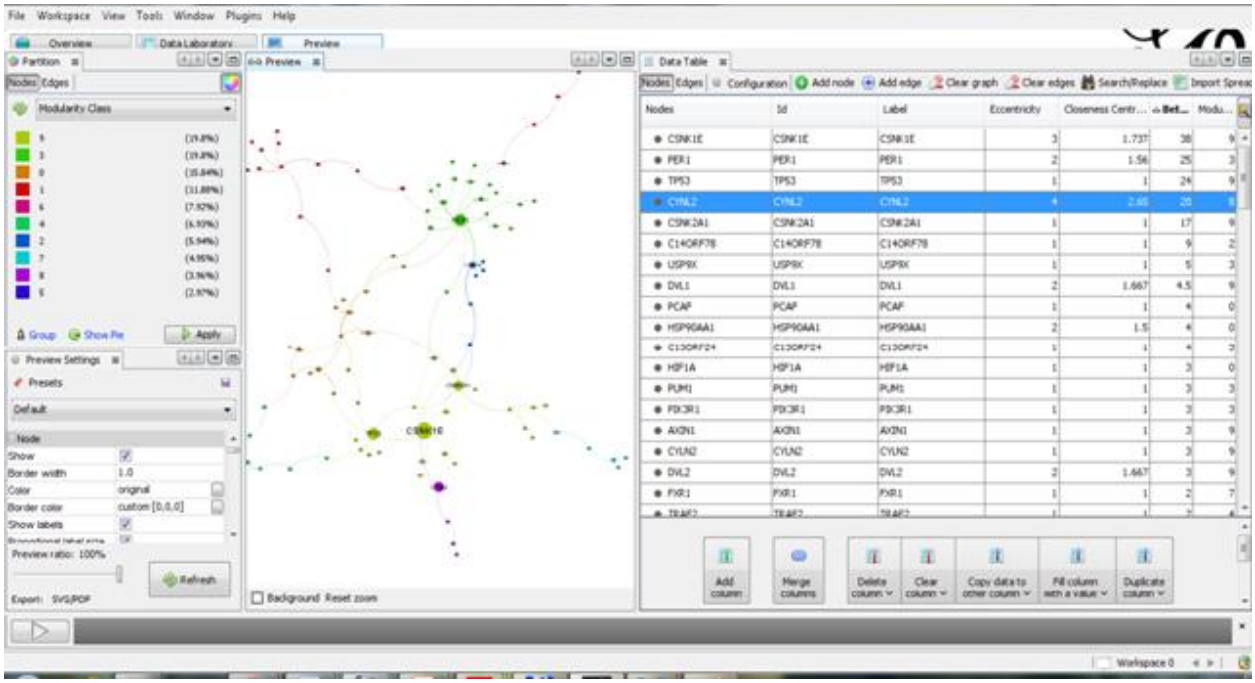


Figure 6. Screenshot from Gephi, showing the network and its statistics.

Gephi gives the possibility to measure important network characteristics that can be interpreted afterwards.

Figure 7. Screenshot from Gephi, showing the entire network

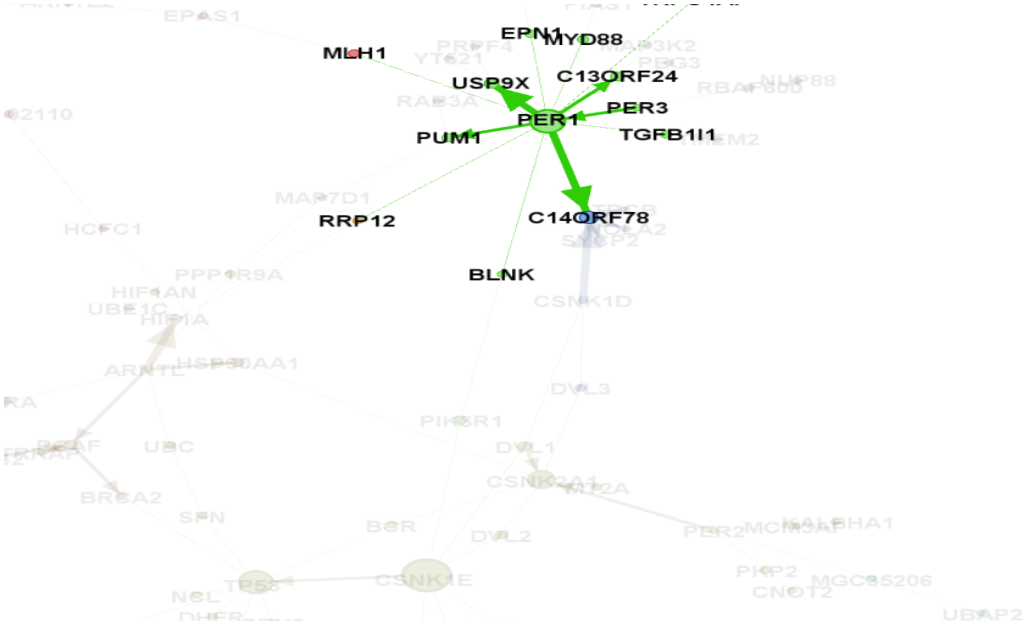
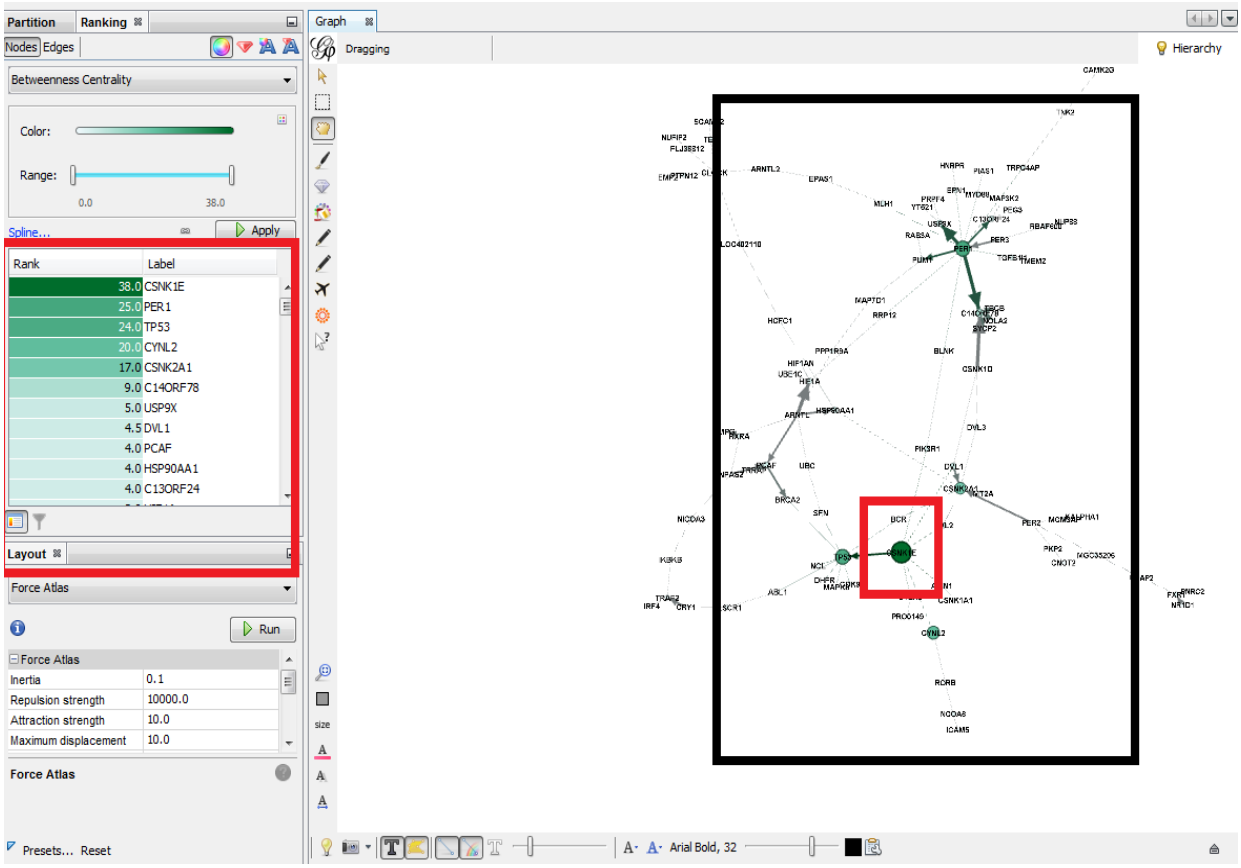
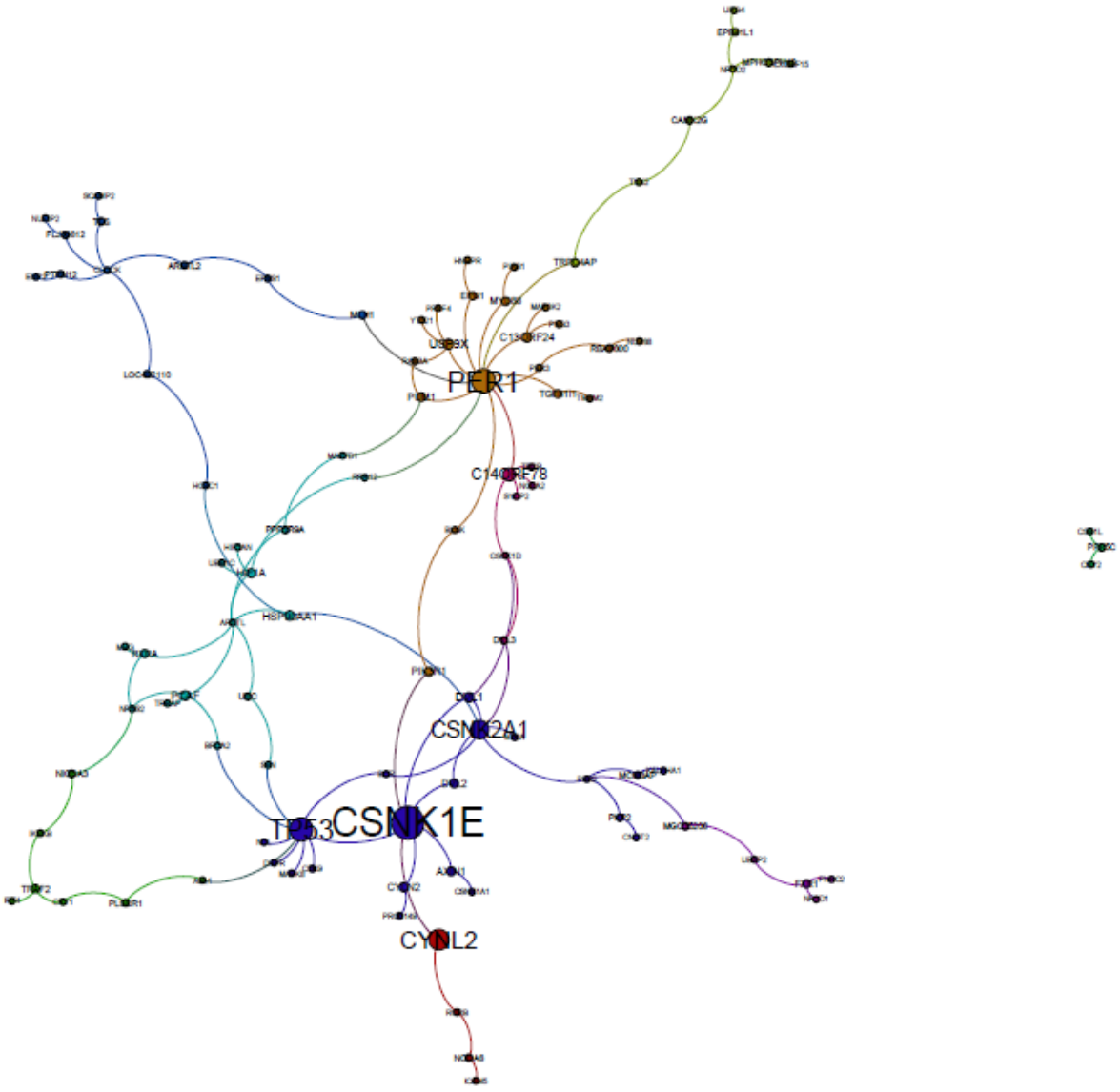


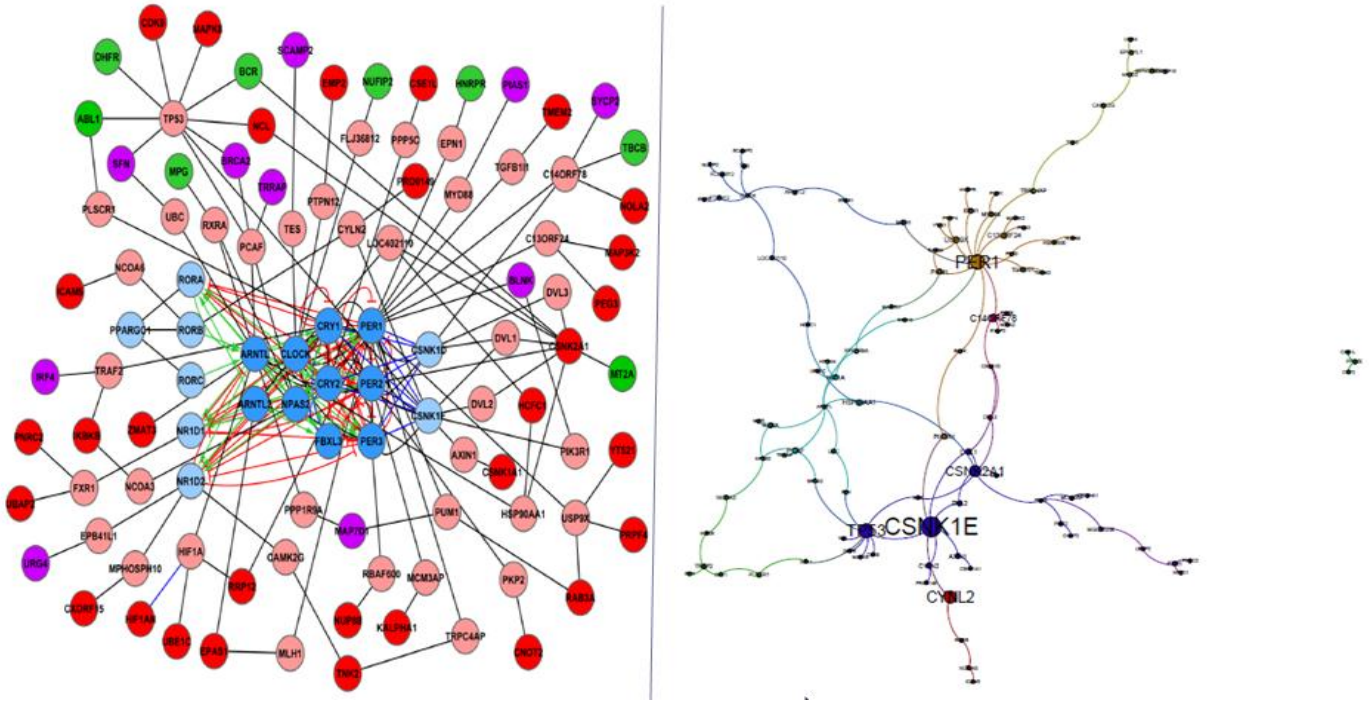
Figure 8. Screenshot from Gephi, with emphasis on one node.

Betweenness centrality measure shows how "central" a node is for a network. A high betweenness centrality suggests that a node connects various different parts of the network together. If we plot size proportional to betweenness centrality we can see that CSNK1E is influential in keeping different parts of the network connected.





The expanded clock gene network modeled in Gephi



Gephi modeling and simulation

5. Conclusions

The presented paper treats interdisciplinarity between complex networks from computer science and molecular biology networks. Cross-disciplinary development of methods is a challenge to explore and conceive the common character that many issues from both fields have. They are novel areas in research and in present an intensive work is done to create frameworks for rationally designing strategies that could give answers to many questions from metabolic network field. Researchers still encounter issues because databases are incomplete (for example, regarding human genome, the exact number of genes is still unknown) and therefore results are inaccurate and of low feasibility.

First of all, I did some research about gene expression process, which is very complex and depends on many factors, both internal and external. I reached to a simplified model that describes it, and found similarities with processes from computer science. For example, a gene being up-regulated in DNA is equivalent to a message we want to transmit on a bus, copying the gene in ARN in reverse means copying the message to a buffer. The process of selecting amino acids' code and creating protein (genes products) from the selected amino acids, is similar to the process of a transmitting a message on a bus and decoding it from the bus. Finally, when a protein reaches a specific threshold of concentration and the gene is down-regulated, means that the message has been received ok and so it can be removed from the queue.

The application I proposed reaches its goal and models the activity of two different metabolic networks. It can be extended and modified in order to be applied to all kind of networks, not only metabolic ones. Its main drawback is that it only makes the static analysis of a gene network. As a future work, a program could be implemented to obtain a dynamic analysis of a network and the resulted output could be examined using Gephi. Dynamicity is hard to achieve, when it comes to gene networks, because not all pre-conditions are known (interactions between genes, total number of genes implied in a metabolic process, genes' exactly behavior, etc).

6. References

[1] Patil, Kiran, Architecture and regulation of metabolic networks, European Molecular Biology Laboratory, Research at a Glance 2013, p. 63

[2] Smolen P., Baxter D.A., and Byrne J.H. (2000) Mathematical modeling of gene networks. *Neuron*, 26: 567-580

[3] Barabasi, Albert-Laszlo, Network Science, Chapter 1, July 2012

[4] Barabasi, Albert-Laszlo, Network Science, Chapter 3, November 2012

[5] Boccalettia S., Latorab V., Morenod Y., Chavez M., Hwanga D.-U. Complex networks: Structure and dynamics, Physics Reports 424 (2006) 175 – 308

[6] Barabasi, Albert-Laszlo, Network Science, Chapter 2, July 2012

[7] Nicosia V., Modularity for community detection: history, perspectives and open issues.

[8] Eric E. Zhang, Andrew C. Liu, Tsuyoshi Hirota, Loren J. Miraglia, Genevieve Welch, Pagkapol Y. Pongsawakul, Xianzhong Liu, Ann Atwood, Jon W. Huss, Jeff Janes, Andrew I. Su, John B. Hogenesch, Steve A. Kay, A Genome-Wide RNAi Screen for Modifiers of the Circadian Clock in Human Cells, 2009, 199-210