University of Palermo
Department of Engineering
PhD in Information and Communication Technologies
XXXV Cicle

Ph.D. Thesis

Knowledge Extraction from Biological and Social Graphs

Mariella Bonomo

Supervisor
Prof. Simona Ester Rombo

Academic year 2021-2022

Author's Address:
Mariella Bonomo
Department of Engineering
University of Palermo
email: *mariella.bonomo@community.unipa.it*

I dedicate this thesis
to myself, my dreams, my force,
because the price of success is hard work.

Abstract

Many problems from real life deal with the generation of enormous, varied, dynamic, and interconnected datasets coming from different and heterogeneous sources. Analysing large volumes of data makes it possible to generate new knowledge useful for making more informed decisions, in business and beyond. From personalising customer communication to streamlining production processes, via flow and emergency management, Big Data Analytics has an impact on all processes.
The potential uses of Big Data go much further: two of the largest sources of data are including individual traders' purchasing history, the use of Biological Networks for disease prediction or the reduction and study of Biological Networks. From a computer science point of view, the networks are graphs with various characteristics specific to the application domain. This PhD Thesis focuses on the proposal of novel knowledge extraction techniques from large graphs, mainly based on Big Data methodologies.
Two application contexts are considered and three specific problems have been solved: *Social data*, for the optimization of advertising campaigns, the comparison of user profiles, and neighborhood analysis. *Biological and Medical data*, with the final aim of identifying biomarkers for diagnosis, treatment, prognosis, and prevention of diseases.

Preface

The contents of this Thesis are conveniently divided in three parts, proposed approaches and results respectively in 8 Chapter. The first part (Chapter 1) introduces the problem of knowledge extraction on Graphs in two different application contexts: Social Networks and Biological Networks. After the introductory Chapter 1 preliminary biological notions on Graphs are reported in Chapter 2, such as definitions, data structures, examples on different types of Graphs.
The second part describes the step-by-step methods, the algorithms used in two different contexts. Chapter 3 presents the state of the art of three problems considered in this Thesis: Chapter 4 the novel method based on the optimization of Advertising Campaign. Chapter 5 the method based on extracting knowledge encoded in the Biological Networks topology. Chapter 6 analyses the method based on prediction of lncRNA-disease Associations.
In third part the results of problems on the two contexts dealt with are shown in Chapter 7. The Thesis ends with (Chapter 8), conclusion common to both the two contexts, where the possible future developments of the presented work are addressed.

Acknowledgments

Arriving at this point, defending my doctoral thesis, was not obvious. Life has presented me with many obstacles and dark periods that have served as experience, which I managed to overcome on my own and with hindsight. Firstly I'd like to thank myself for the strength and the courage of a lion that never gives up even at the first obstacle. I do not deny that it was hard and there were difficult moments along these years full of disappointment, regret but also joy and satisfaction.

This project would not have been possible without the support of many people: many thanks to my advisor, Simona E. Rombo, who read my several revisions and helped me to make sense of the confusion. She has been an ideal teacher, mentor, and thesis supervisor, offering me advice and encouragement with a perfect blend of insights and humor. If I became the person I am today, I owe it to her who saw potential in me by always pushing and advising me. Not in chronological order, and definitely not in terms of importance, I'd like to thank to my parents and my sister Teresa, whom I think I made them proud of me during my university career, they were my pillars, even in the difficult moments they always support me not only financially, and I'll never stop to thanking them. A special thanks goes to my grandparents, because they never leave me alone, I dedicate this moment to them.

Other thanks go to those people who have been a source of inspiration for me during my PhD: Salvatore Morfea and Simona Panni, members of my Kazaam Lab team, I enjoyed with their wealth of knowledge and valuable tips. Another thanks goes to Professor Susana Vinga,

Monica and Roberta for the support that gave me during my experience in Lisbon as visiting scholar, and for the beautiful friendship established. Thanks to two brilliant professors: Raffaele Giancarlo and Giovanni Pilato, who were great and essential mentors during the course of my studies. They deserve my admiration.

Last but not least, I'd like to thank my colleagues and friends: Antonella, the precious friend that everyone would like to have. I shared many beautiful moments. Chiara, my guide and true friend, who I met in Lisbon. Giusy, the friend of social network: we met for the first time online, and she is a sweet person.
Gabriella, Margherita, Antonella and Flaviana are friends of lifetime, who have always been there.

Mariella Bonomo

Funding

The research presented in this PhD Thesis has been supported by different research projects. In particular, the PhD scholarship has been funded in part by the PRIN research project "Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond", grant n. 2017WR7SHH (MIUR). Moreover, participation to international conferences for the presentation of some of the obtained results has been possible also thanks to the funds by the GNCS research projects "Algorithms, methods and software tools for knowledge discovery in the context of Precision Medicine" (2020) and "Big knowledge graphs modelling and analysis for problem solving in the web and biological contexts" (2022, CUP E55F22000270001), provided by INDAM GNCS. All mentioned projects have also allowed to join very stimulating research meetings and events which have pushed toward interaction with other researchers working on related topics.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

Abstract

This Chapter introduces the motivations which have pushed for the study of problems and the proposal of solutions described in this PhD Thesis. The main focus of this Thesis is on knowledge extraction from graphs, specialized in two different application contexts: Social Networks and Biological Networks.

Many problems of real life can be modeled as graphs, able to take into account important relationships between interacting "actors". On the other hand, we are daily drowned in a very large amount of data, coming from different sources, that are complex in contents, heterogeneous in formats and order of terabytes in size.
These "Big Data" provide unprecedented opportunities to work on exciting problems, but also raise many new challenges for data mining and analysis. Indeed, most of the current analytical tools become obsolete as they fail to scale with data, especially when graphs seem to be the most suitable models to be adopted. Moreover, data are usually obtained from different information sources, and they need to be suitably integrated on the cloud. Therefore, performant technologies are required for data integration and data-intensive analysis, and algorithms need to be designed in order to be efficient and effective in this scenario. As sketched in [14], approaches proposed in this PhD Thesis are focused on the proposal of novel methodologies based on knowledge extraction in the context of "Big Data" modeled as graphs. Graphs are powerful

models to represent Networks of real world entities such as: events, objects, situation, concepts, by illustrating the relationship between them. These graphs usually involve nodes associated with the main players of the process under analysis, and edges represent the relationships between these players. As an example, nodes may represent users in the case of Social Networks, or proteins and/or other cellular components in the case of Biological Networks. What is important to analyse, in addition to the topology of the network, is the semantic encoded in nodes and edges, as well as the algorithmic techniques in the networks context. Here we analise two main different application contexts:

- **Social Networks**, where users' data are often analyzed in order to learn more about their interests and connect them with contents and advertising relevant to their preferences. Many individuals, teams and organizations are part of a number of Networks that give access to knowledge, markets, technology, reputation or influence;
- **Biological Networks**, where an important source of Big Data is given by the biological high-throughput techniques, and the representation of interacting elements, such as cellular components, genotypic-phenotypic associations, etc., is particularly relevant in order to take into account important information which would be missed by looking at each element singularly.

It is worth pointing out that, although the proposed methodologies are often general enough to be applied also in other application contexts, part of the contribution of this PhD project consists on providing satisfying solutions which may be used in practice in the social and medical scenarios.

1.1 Social Networks

Automatic systems able to suggest a set of target users for advertising campaigns provide three main benefits:

1. Minimization of costs for the dissemination of the advertising campaign through social media, which is often very expensive;

2. Improvement of the user experience in Online Social Networks (OSNs), since only the possibly interested customers are contacted with advertisements which could be useful for them;

3. Avoid the spread useless information through OSNs.

The presented research consists of the proposal of novel recommendation approaches based on the comparison between the OSNs profiles associated with users (possible customers) and advertisers (brands), according to the considered campaign. Profile matching is then applied relying on such a graph representation, and suitable similarity measures are considered for each category. When categories involve textual documents containing information on interests and preferences (e.g., posts and comments), the document is represented as a bag of words and the Term Frequency-Inverse Document Frequency (TF-IDF) is used to weight the importance of the words inside the text.
Social advertising allows for quick and efficient social engagement on Social Networks.

1.2 Biological Networks

One of the most important challenges of this century is the proposal of precision therapies, that is, medical therapies adaptive with respect to specific categories of individuals, presenting well targeted features (e.g., genomic signatures, phenotypes, etc.). The recent advances in sequencing technologies have led to an exponential growth of biological data, allowing for high throughput profiling of biological systems in a cost-efficient manner. Molecules such as genes, proteins and RNA together contribute to cellular life, and it is commonly accepted that they have to be analyzed as interacting elements when they take part in common biological processes [70].
More recently, great attention is turning towards the possible associations between cellular components and macroscopic disorders or complex diseases. In this context, we have studied two main problems: (i) the importance of centrality measures in extracting functional knowledge from Biological Networks, and (ii) the prediction of long non coding RNA

(lncRNA)-diseases associations (LDA).
The first problem includes comparative analysis of nine outstanding topological measures, based on compact views obtained from the rank they induce on a given input biological network. The goal is to understand their ability in correctly positioning nodes/edges in the rank, according to the functional knowledge implicitly encoded in biological networks. To this aim, both internal and external (gold standard) validation criteria are taken into account, and six networks involving three different organisms (yeast, worm and human) are included in the comparison.
The second problem is based on an approach for the prediction of lncRNA-disease associations based on neighborhood analysis performed on a tripartite graph. The idea is to discover hidden relationships between lncRNAs and diseases through the exploration of their interactions with intermediate molecules (e.g., miRNAs).

1.3 Summary of the main results

In the two analysed contexts, the research behind this PhD project has led to the following main results.
The technique focused on Social Networks is applied to *brand-affinity matching* has been presented in [15]. In particular, the profile-matching technique (presented in Chapter 4) is based on tree-representation of user profiles and applied it on Facebook ego-Networks. The approach presented extends those results, showing that a suitable combination of profile-matching and neighborhood analysis is more successful in identifying the best k users for advertisements distribution.
Neighborhood analysis performs better than other techniques presented in the literature and it is not based on known lncRNA-disease associations (as described in Chapter 3). Approaches based on integrative networks, typically combining networks from different studies that investigate the same or similar research questions have indeed shown to reach better performance. The techniques in the biological context based on extracting knowledge encoded in the Biological Networks topology shows that a distinct handful of best performing measures can be identified for

each of the considered organisms, independently from the reference gold standard. Moreover, it seems that the proposed paradigm works better on denser networks, possibly due to the fact that the encoded information is larger than for sparse networks. The techniques based on prediction of lncRNA-disease associations identify novel LDA by analyzing the behaviour of neighbor lncRNAs, showing that the consideration of indirect relationships between lncRNAs and diseases through neighborhood analysis is more effective, and performs better than other techniques and not based on known lncRNA-disease associations. Centrality method and collaborative filtering are based better than other techniques in the literature in terms of accuracy.

Chapter 2

Background

Abstract

This Chapter introduces basic definitions and examples on different types of graphs, which are important for the full understanding of the main topics accounted for in this Thesis. In particular, we focus on two different contexts: Social Networks and Biological Networks. The approaches proposed in this Thesis have been implemented relying on Big Data technologies, using Apache Spark, therefore the chapter recalls some fundamentals on these technologies.

2.1 Basics on Graphs

Definition 1 (Graph) *A Graph is defined as a pair $\mathcal{G} = (\mathcal{V}, E)$ on two sets $\mathcal{V}$ and E, such that the elements in $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ are the vertices (or nodes), and the elements in $E = \{e_1, e_2, ...e_m\}$ are the edges, that is, the connections between the vertices.*

Each edge $e \in$ E is said to join two vertices, which are its *endpoints*. If e join $u, v \in \mathcal{V}$, we write $e = \langle u, v \rangle$. In addition a *self-loop* is an edge that joins a single endpoint to itself. In Figure 2.1 an example of the representation of a graph is shown.

Definition 2 (Same edge) *In an undirected graph $\mathcal{G}$, if (u, v) is an*

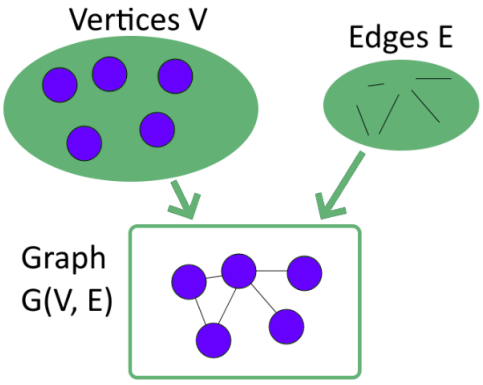


Figure 2.1: The representation of a graph with $\mathcal{V}$ as the set of vertices and E as the set of edges.

edge in E we assume: $(u, v) = (v, u)$*, whose* (v, u) *is the same edge of* (u, v)*.*

Definition 3 (Directed Graph) *$\mathcal{G}$ is a directed graph or digraph or oriented if the edges are ordered pairs (u, v) of vertices; while $\mathcal{G}$ is an undirected Graph if the edges are unordered pairs of distinct vertices.*

In Figure 2.2 an example of the representation of a graph directed and undirected is shown. The following definitions refer to analogous ones as introduced in the literature [12, 31].

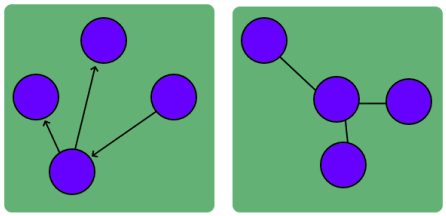


Figure 2.2: The difference between directed graph (graph on the left) and undirected graph (graph on the right).

Definition 4 (Adjacency) *For a given graph $\mathcal{G}$, the vertices u and v*

are adjacent if $e_1 = uv \in \mathcal{G}$. *Two edges* $e1 = uv$ *and* $e_2 = uw$ *having a common end, are adjacent with each other.*

Definition 5 (Neighbor) *For any graph* $\mathcal{G}$ *and vertex* $v \in \mathcal{V}(\mathcal{G})$, *the neighbor set* $N(v)$ *of* v *is the set of vertices (other than* v *adjacent to* v *that is*

$$N(v) = \{w \in \mathcal{V}(\mathcal{G}) | v \neq w, \exists e \in E(\mathcal{G}) : e = (u, v)\} \tag{2.1}$$

Definition 6 (Degree) *The number of edges incident with a vertex* v *in a graph* $\mathcal{G}$ *is the degree of* v, *denoted as* $deg(v)$. *Loops are counted twice. A vertex* v *of degree 0 is an isolated vertex while a vertex* v *of degree 1 is denoted end-vertex.*

Nodes with "high" degree are *hubs*, since they are connected to many neighbors (different criteria may be used in order to define the minimum degree characterizing a hub, often depending on the application context under analysis).

Definition 7 (Subgraph) *A graph* H *is a subgraph of* $\mathcal{G}$ *if* $V(H) \subseteq V(\mathcal{G})$ *and* $E(H) \subseteq E(\mathcal{G})$ *such that for all* $e \in E(H)$ *with* $e = (u.v)$, *we have that* $u, v \in V(H)$. *When* H *is a subgraph of* $\mathcal{G}$, *we write* $H \subseteq G$.

Definition 8 (Induced Subgraph) *For a given graph* $\mathcal{G}$, *the subgraph induced on a vertex subset* U *of* V_G, *denoted by* $\mathcal{G}(U)$, *is the subgraph of* $\mathcal{G}$ *whose vertex-set is* U *and whose edge-set consists of all edges in* $\mathcal{G}$ *that have both endpoints in* U. *If* v *is a vertex of a graph* $\mathcal{G}$, *then the vertex deletion subgraph* $\mathcal{G} - v$ *is the subgraph induced by the vertex set* $V_G - v$.

Suppose a closed walk in the connected graph that visits every vertex of the graph exactly once (except starting vertex) without repeating the edges. A *trail* is a walk in which all edges are distinct; a *path* [4] is an alternating sequence of vertices and edges (see Figure 2.3). This is represented by the sequence $\{v_1, v_2, ..., v_n\}$ of vertices on the path, and there are no repeated edges or vertices (except possibly the initial and final vertices). A path is *simple* if all edges and all vertices on the path are distinct. The *length* of a path is the number of edges of which it is

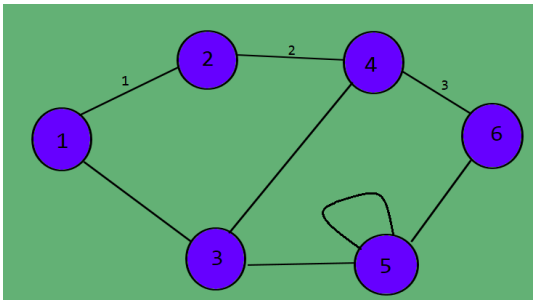


Figure 2.3: In this example, we have a graph $\mathcal{G}$ with six vertices and seven edges. Example of a path of a graph is defined as the finite sequence of edges which joins a sequence of vertices (1,2,4,6) which by most definitions, are all distinct in a graph $\mathcal{G}$.

made up. A *Weighted Graph* $\mathcal{G}$ is a graph for which each edge e has an associated real-valued number $w(e)$ that is weight. The weight w_{ij} of the edge between nodes i and j represents the relevance of the connection (e.g., sequence similarity network) in most cases. A *shortest path* in a graph $\mathcal{G}$ between two vertices u and v is a path with the minimum number of edges. If the graph is weighted, the shortest path is a path such that the sum of edge weights has the minimum value.

Definition 9 (Cyclic Graph) *A graph $\mathcal{G} = (\mathcal{V}, E)$ is cyclic if it has at least one cycle.*

Definition 10 (Acyclic Graph) *A graph $\mathcal{G} = (\mathcal{V}, E)$ is acyclic if zero cycles are present, and an acyclic graph is the complete opposite of a cyclic graph.*

Definition 11 (Connected Graph) *A graph $\mathcal{G} = (\mathcal{V}, E)$ is connected if for every pair of vertices u and v in $\mathcal{V}$, there is a path from u to v.*

A digraph is strongly connected if there exists a directed path between every pair of distinct vertices from D. A digraph is weakly connected if its underlying graph is connected.

Definition 12 (Disconnected Graph) *A graph $\mathcal{G} = (\mathcal{V}, E)$ is disconnected if there are at least two vertices separated from one another.*

Definition 13 (Isomorphic Graphs) *Two graphs* $\mathcal{G}1 = (\mathcal{V}, E)$ *and* $\mathcal{G}2 = (\mathcal{V}^*, E^*)$ *are isomorphic if there is a one-to-one mapping* $\phi : \mathcal{V} \rightarrow \mathcal{V}^*$ *such that for every such that for every edge* $e \in E$ *with* $e = \langle u, v \rangle$*, there is a unique edge* $e* \in E*$ *with* $e^* = \langle \phi(u), \phi(v) \rangle$*.*

Stated differently, two graphs $\mathcal{G}1$ and $\mathcal{G}2$ are isomorphic [94] if we can uniquely map the vertices and edges of $\mathcal{G}1$ to those of $\mathcal{G}2$ such that if two vertices were joined in $\mathcal{G}1$ by a number of edges, their counterparts in $\mathcal{G}2$ will be joined by the same number of edges. The difference between labelled and unlabelled graphs becomes more apparent when we try to count them.

2.2 Data Structures

There are different ways to represent graphs. A graph can be represented using a **adjacency matrix**. An *adjacency matrix* is a binary square matrix $\mathcal{M}$ of order $n = |\mathcal{V}|$ defined by:

$$\mathcal{M}_{i,j} = 1, \text{ if } (v_i, v_j) \subseteq \mathcal{E}(\mathcal{G})$$
$$\mathcal{M}_{i,j} = 0 \text{ otherwise.}$$

It is immediately evident from the representation by the adjacency matrix that:

- An adjacency matrix is symmetric, that is for all $i, j, A[i, j] = A[j, i]$. This property reflects the fat that an edge is represented as an unordered pair of vertices $e = (v_i, v_j) = (v_j, v_i)$;
- A graph $\mathcal{G}$ is simple if and only if for all $i, j, A[i, j] \leq 1$ and A[i,j]=0;
- The sum of values in row i is equal to the degree of vertex v_i, that is, $deg(v_i) = \sum$ A[i,j].

An alternative representation is an incidence matrix. An incidence matrix $\mathcal{M}$ of graph $\mathcal{G}$ consists of n rows and m columns such that $\mathcal{M}[i, j]$ counts the number of times that edge e_j is incident with vertex v_i. The following properties are easy to verify:

- A graph $\mathcal{G}$ has no loops if and only if for all i, j, $\mathcal{M}[i, j]$ $leq1$;

- The sum of all values in row i is equal to the degree of vertex v_i. In mathematical terms, this is expressed as $\forall i : deg(v_i) = \sum \mathcal{M}[i, j;$
- Because each edge has exactly two, not necessarily distinct end points, we know that for all $j, \Sigma\mathcal{M}[i, j{=}2$.

$$M = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Figure 2.4: The adjacency matrix for the subgraph in Figure 2.3.

2.3 Types of Graphs

Let $\mathcal{G} = (\mathcal{V}, \mathrm{E})$ be a graph. The following definitions hold.

Definition 14 (Null Graph) *If $E = \emptyset$, then $\mathcal{G}$ is a null graph.*

A null graph $\mathcal{G}$ is a graph with no edges. It may have one or more vertices.

Definition 15 (Trivial Graph) *If the size of $\mathcal{V}$ is equal to one, then $\mathcal{G}$ is trivial graph.*

The trivial graph is the smallest possible graph that can be created with the minimum value of vertices that is one vertex only.

Definition 16 (Complete Graph) *A graph $\mathcal{G}$ is complete if E contains an edge between all possible pairs of vertices in $\mathcal{V}$*

Definition 17 (Multigraph) *A multigraph $\mathcal{G}$ is a graph without loops, multiple edges having the same end vertices.*

Definition 18 (Regular Graph) *If each vertex in $\mathcal{V}$ has the same degree then $\mathcal{G}$ is a regular graph. If each vertex in $\mathcal{V}$ has degree r, then $\mathcal{G}$ is regular of degree r or r-regular.*

Definition 19 (Finite Graph) *If the number of vertices in $\mathcal{V}$ and the number of edges E are finite in number, then $\mathcal{G}$ is a finite graph.*

Definition 20 (Bipartite Graph) *If $\mathcal{V}$ can be partitioned into two disjoint subsets V_1 and V_2 such that $(u, v) \in E$ implies either $u \in V_1$ and $v \in V_2$ OR $v \in V_1$ and $u \in V_2$, then $(V_1, V2)$ is a bipartition of $\mathcal{G}$, and $\mathcal{G}$ is a bipartite graph.*

The following definitions refer to analogous ones as introduced in the literature [92].

Definition 21 (Complete Bipartite Graph) *A bipartite graph $\mathcal{G}$ is complete (m,k)-bipartite, if $|X| = m, |Y| = k$, and $uv \in \mathcal{G}$ for all $u \in X$ and $v \in Y$.*

Definition 22 (Complete Multipartite Graph) *A set of graph vertices $\mathcal{V}$ decomposed into k disjoint sets: $v_1, v_2..., v_k$ is a complete k-partite graph. A graph $\mathcal{G}$ that is complete k-partite for some k is a complete multipartite graph [30].*

Definition 23 (Euler Graph) *A connected graph $\mathcal{G}$ is eulerian, if it has a closed trail containing every edge of $\mathcal{G}$. Such a trail is an Euler tour.*

Definition 24 (Hamiltonian Graph) *A path P of a graph $\mathcal{G}$ is a Hamilton path, if P visits every vertex of $\mathcal{G}$ once. A cycle C is a Hamilton cycle if it visists each vertex once. A graph $\mathcal{G}$ is hamiltonian if it has a Hamilton cycle.*

2.4 Social Networks

A Social Networks (Figure 2.5) is constructed from relational data [90] and it is defined as a set of social entities, such as people, groups, and organizations, with some pattern of relationships or interactions between them. From a technological point of view, in large giants such as Facebook or Twitter Online Social Networks (OSN), the key task has increasingly involved the association of possible customers to brands

Figure 2.5: Example of a Social Network.

[11, 56, 84, 99] based on the comparison of profiles [47] in the same OSN. Social Network profiles represent relationships between categories and subcategories and the objective is to associate the appropriate categories and subcategories to both user and brand profiles.

Definition 25 (Social Network) *A Social Network is represented by an undirected graph $\mathcal{G} = (\mathcal{V}, E)$, where nodes in $\mathcal{V}$ are associated to the users, and two nodes are linked in $\mathcal{G}$ if a social relationship (e.g., friendship, common interests, etc.) occurs between the users.*

A network $\mathcal{G}$ (Figure 2.6) consists of a non empty set $\mathcal{V}$ (vertices) that are associated to users (e.g. customers, brands, users) and a set of edges E, represent relationships between them (e.g friendship). Social information contents published by OSNs users are usually associated with textual data (e.g posts and comments). Each profile includes descriptors such as age, location, interests, multimedia content; the visibility of a profile varies by site and according to the privacy of the user.

2.5 Biological Networks

Networks are widely used in many branches of biology [52, 85] as a convenient representation of patterns of interaction between appropriate

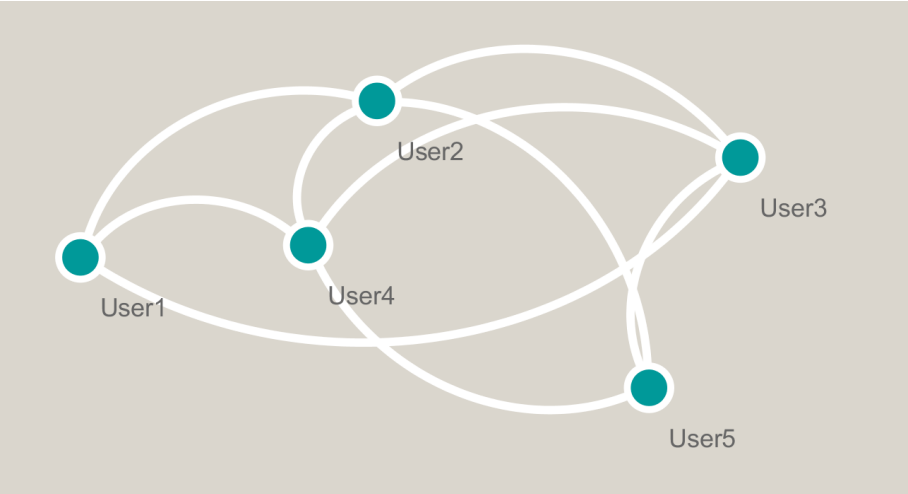


Figure 2.6: Example of Social Network.

biological elements. These Biological Networks include molecular networks, biochemical networks, neural networks, and ecological networks.

Definition 26 (Biological Network) *A Biological Network is an undirected graph $\mathcal{G}$ in which the nodes ($\mathcal{V}$) represent the cellular components, such as genes and proteins, and the edges (E) between two nodes are the physical interactions between such components or other types of association.*

$$\mathcal{G} = (\mathcal{V}, E)$$

In Figure 2.7 an example of a Biological Network. Different types of components can exist in the same network: cellular components such as genes and proteins, which contribute to cellular life and take part in biological processes [21, 35, 39, 73]; the associations may be related to physical, functional or phenotypic interactions [42, 57]. Networks composed of functional links, although different, may have different links depending on what the edge represents; for example, two genes may share an edge if there is at least one disease involving mutations in both of them. The identification of biological pathways [76] will lead us to distinguish whether the removal of a node interrupts communication

between pairs of nodes in the network. Underlying the interactions of networks are various types of biological graphs that we will analyse in the next paragraphs: Protein-Protein interaction Networks, Interactome Networks, Co-expression Networks, Diseasoma, lncRNA-miRNA-disease Networks.

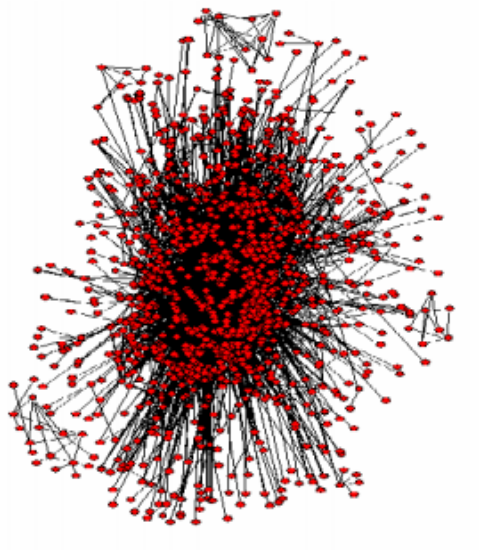


Figure 2.7: Example of a Biological Network.

2.5.1 Protein-Protein interaction Networks

Proteins are long-chain molecules formed by the concatenation of a series of basic units defined amino-acids [36]. Once created, a protein does not stay in a loose chain-like form, but folds on itself, whose shape depends on the amino acid sequence. The folded form dictates the physical interaction it can have with other molecules. Hence, the primary mode of protein-protein interaction is physical rather than chemical, their complicated folded shapes interlocking to create so-called protein complexes but without the exchange of particles that defines chemical reactions. The protein-protein interactions (PPI) [60, 83] of a given organism are modelled by a network which highlights the reciprocal in-

teractions between proteins; they play a large role in every biological functions [35, 76]. In a protein-protein interaction network [37, 61, 82] the vertices are proteins and two vertices are connected by an undirected edge if the corresponding protein physically interacts. However, this representation omits useful information. Interactions that involve three or more proteins are represented by multiple edges, and there is no way to understand from the network itself that such edges represent aspects of the same interaction. The development of new technologies has improved experimental techniques [95] or the detection of PPIs. This problem could be addressed by adopting a bipartite representation, with proteins and interactions as different types of vertices, and undirected edges connecting proteins to the interactions participated by them.
Many computational approaches [86] use information from different sources:

- Primary sources store annotations, both produced manually and by computational prediction;
- Secondary sources involve the integration of PPI from different primary sources that provide the weights between the edges and different organisms.

The graph represented by protein-protein interaction is generated by two different approaches with different global properties such as the relationships between the number of interacting proteins and the essential gene.
Since the interacting proteins are usually in the same subcellular compartment one uses functional similarity to predict the interacting proteins. In the past, experimental methods and computational approaches have been useful on various organisms. The methods used in the prediction of PPI requires reliable data on positive and negative samples, the sets of negative samples have been randomly created from paired proteins or by selecting pairs of proteins that do not share the same cell compartment, but nevertheless these samples were created on the basis of cell position, by means of semantic similarity. There are four parameters that must be estimated to provide an accurate map of network interaction, allowing comparison with other maps:

- Completeness: a measure guaranteeing the physical number of protein pairs actually tested in a given search space;

- The sensitivity assay: measures which interactions can be detected with a sensitive tool;

- The sensitivity sample: the fraction of all detectable interactions found by a single implementation;

- The precision: the proportion of biophysical interactors.

However, with careful comparison with other network maps, network interactions have become more complex, due to the organisation of the cell, which has changed from a simple *baggage of enzymes* to a complex of macromolecular interactions.

2.5.2 Interactome Networks

The range of macromolecular interactions constitutes the *Interactome Networks*. Interactome Networks serve as an information bridge to extract properties of the local or global graph. Several approaches are used to capture these types of networks that differ in the possible interpretations of the network map, some of which occur:

- Through the compilation of existing data available in the literature, obtaining physical or biochemical interactions;

- Through computational predictions based irrespective of physical or biochemical such as sequence similarity, and through the presence or absence of genes in sequences.

In general, the functioning of macromolecular structures known as proteomes and transcriptomes in Interactome networks is represented through arrays comprising all the genes of an organism. These types of networks have been applied to detect genes potentially involved in cancer.

2.5.3 Co-expression Networks

In *Co-expression Networks*, the graph representing the network is characterised by nodes representing genes and edges associated with gene links, which show a co-expression above a set threshold [22]. This threshold value is calculated by the topology of the network, usually a high value is chosen above which gene interactions are considered relevant. There is An overlap between the edges of interaction in Co-expression Networks and the edges of Interactome Networks, which give useful information for the global estimation of the biological significance. At the same time, many correlations can be significant on sets of data sets such as protein protein interactions that correspond to pairs of genes whose expression can be correlated.

2.5.4 Diseasome

The *Human Diseasome* is represented by a bipartite graph consisting of two disjoint sets of nodes. Nodes in the first set represent diseases while nodes in the second sets represent genes. A disorder and a gene are then connected by a link if mutations in that gene are implicated in that disorder. From the bipartite diseasome (Figure 2.8, middle), one can construct the human disease network, the network of human diseases connected by sharing common genetic components (Figure 2.8, left). One can also construct the human disease gene network, the network of human genes, connected by implicating common human disorders (Figure 2.8, right). The first version of the diseasome was created based on the list of human disorders, disease genes and associations between them obtained from the OMIM database as of December 2005 [46].

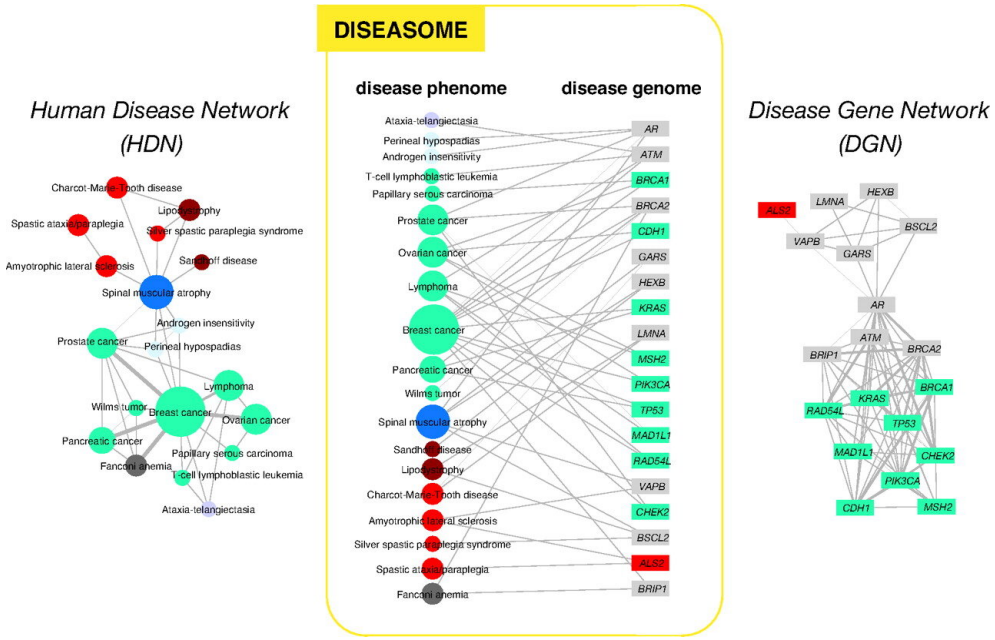


Figure 2.8: Schematic illustration of the Human Diseasome. Adapted from [46]. Copyright (2007) National Academy of Sciences, USA.

2.5.5 lncRNA-miRNA-disease Networks

Most of the genome in Figure 2.9 is not encoded; that is, the information it contains is not used for protein synthesis. For a long while, such non-coding regions of the genome have been considered "junk DNA". However, it is now well recognized that DNA sequences that do not give rise to proteins, also known as non-coding RNA, may be important for specific cell functions. Non-coding RNAs differ in the length of nucleotides. There are two types of ncRNAs:

- Long-non-coding RNAs (**lncRNAs**) are molecules emerging as key regulators of various critical biological processes, and their alterations and dysregulations have been associated with many important complex diseases [24, 63];

- MicroRNA (**miRNA**) small molecules characterized by approximately 20 to 22 nucleotides that are particularly active in the regulation of gene expression at transcriptional and post - transcriptional levels.

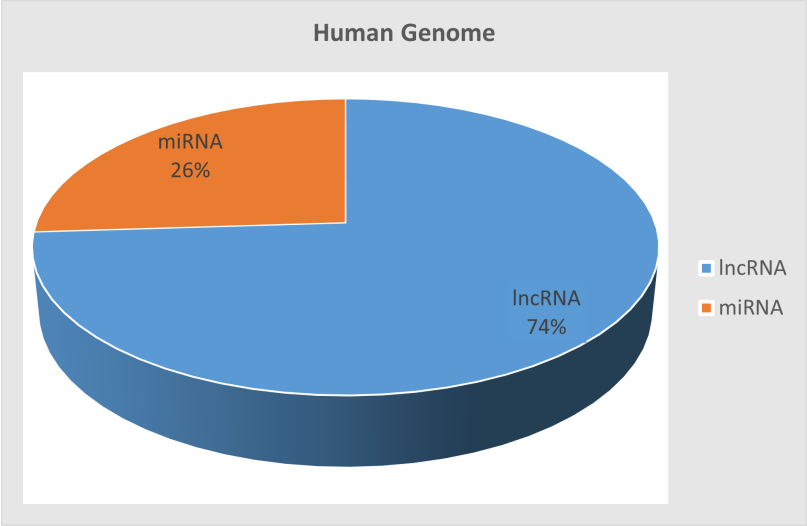


Figure 2.9: Distribution of miRNA and lncRNA in the Human Genome.

Let $\mathcal{L} = \{l_1, l_2, \ldots, l_h\}$ be a set of lncRNAs and $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$ be a set of diseases, and $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$ be a set of miRNA. Let T_{LMD} be a tripartite graph defined on the three sets of disjoint vertexes L, M and D, (in Figure 2.10) which can also be represented as $T_{LMD} = \langle (l, m), (m, d) \rangle$, where (l, m) are edges between vertexes in L and M, (m, d) are edges between vertexes in M and D, respectively. In such a context, edges of the type (l, m) represent molecular interactions between lncRNAs and miRNAs, edges of the type (m, d) correspond to known associations between miRNAs and diseases.

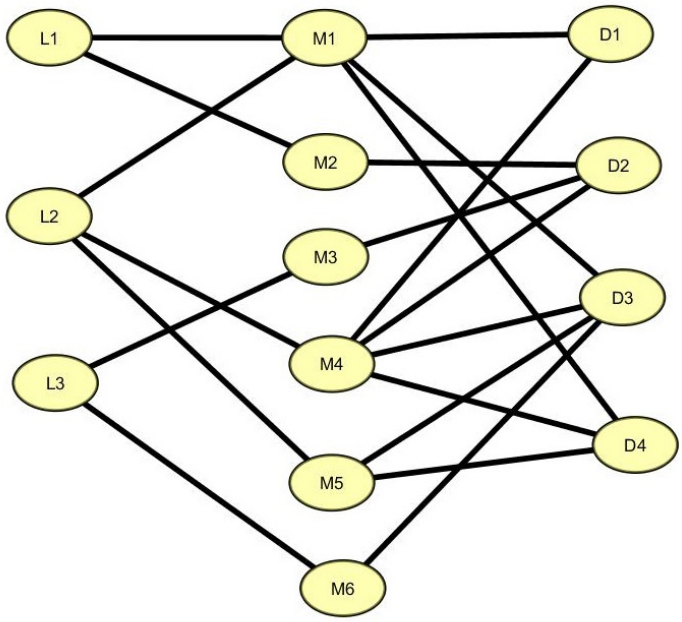


Figure 2.10: Tripartite graph which nodes are represented from lncRNAs, miRNAs and diseases.

2.5.6 Network clustering

A number of approaches rely on traditional hierarchical clustering methods [55], other ones are based on graph partitioning algorithms. Most methods require the number of clusters to be known in advance. However, this information is not always available, thus some algorithms are executed with different cluster numbers and results satisfying a quality criteria are considered to be the most reliable. Obviously, the necessity of running an algorithm different times may cause losses in efficiency. The principal problem that arises (for example in PPI networks) [80] is the choice of the metric adopted to measure the distance between two proteins. In this kind of graphs, due to the structure of the interactions, it has been found that the distances among many nodes are often identical. In such a case the adopted clustering method fails in finding good solutions, due to the presence of ties that have to be solved arbitrarily. For this reason many approaches prefer the analysis of the topological measures.

2.6 The adopted Big Data technologies

Big Data is a combination of structured, semistructured and unstructured data collected [32, 78] by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications. Systems that process and store big data have become a common component of data management architectures in organizations, combined with tools that support big data analytics. Big data is often characterized by the three V's:

- the large **volume** of data in many environments;
- the wide **variety** of data types frequently stored in big data systems;
- the **velocity** at which much of the data is generated, collected and processed.

These characteristics were first identified in 2001 by Doug Laney.

2.6.1 The Map Reduce paradigm

One of the best-known frameworks for turning raw data into useful information is known as *MapReduce*. MapReduce [27, 29] is a method for taking a large data set and performing computations on it across multiple computers, in parallel. It is a programming paradigm proposed by Google researchers in 2004, and the term MapReduce is often used to refer to the implementation of the corresponding model. Basically, MapReduce [59] consists of two primitives:

- The *Map function* performs the tasks of sorting and filtering, taking data and placing it inside of categories, so that it can be analyzed. This function takes data structured in $\langle key, value \rangle$ pairs;

- The *Reduce function* analyzes data returned by the Map in order to produce the results of the MapReduce program.

Perhaps the most influential and established tool for analyzing big data is known as *Apache Hadoop*. Apache Hadoop [101] is a framework for storing and processing data at a large scale, and it is completely open source. Hadoop can run on commodity hardware, making it easy to use with an existing data center, or even to conduct analysis in the cloud. Another tool is Apache Spark, which we find in more detail in section 2.6.2, used for the development of the methods in this thesis, it is a unified analysis engine for large-scale data processing with integrated modules for SQL, data flows, machine learning and graph processing.

2.6.2 Apache Spark

Apache Spark has emerged as a unified engine for large-scale data analysis across a variety of workloads. It has introduced a new approach for data science and engineering where a wide range of data problems can be solved using a single processing engine with general-purpose languages. Apache Spark [91, 1, 48, 67, 109] has been adopted as a fast and scalable framework, it provides an interface for programming entirely using clusters, using parallelism and implementing fault tolerance. The Spark ecosystem comprises five components:

- Spark Core [109] is the foundation of Apache Spark and is the base of the whole project. It provides distributed task dispatching, scheduling, and basic I/O functionalities as *Resilient Distributed Datasets* (RDD);
- Spark SQL [7] is a Spark module for structured data processing;
- Spark Streaming [111] is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams;
- MLlib [67] is Spark's machine learning library. Its goal is to make practical machine learning scalable;
- GraphX [48, 104] is a component for graphs and graph-parallel computation.

At a high level, every Spark application consists of a driver program that runs the user's main function and executes various parallel operations on a cluster. The core abstraction of Spark is called *Resilient Distributed Datasets (RDD)* [110] (Figure 2.11), which is a distributed collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. The key performance driver of Spark is that an RDD can be cached in memory of the Spark cluster compute nodes and thus can be reused by many iterative tasks. The basic unit of parallelism in an RDD is called partition. Each partition is one logical division of data which is immutable and created through some transformation on existing partitions. Immutability helps to achieve consistency in computations. RDDs achieve fault tolerance through a notion of lineage: if the partition of an RDD is lost, the RDD has enough information about how it was derived from other RDDs to be able to rebuild just that partition. Physically an RDD is a Scala object and it is created by starting with a file in the Hadoop file system (or any other Hadoop-supported file system), or an existing Scala collection in the driver program, and transforming it. RDDs can be created through deterministic operations:

- *Transformations* are deterministic, but lazy, operations which define a new RDD without immediately computing it; transformations return pointers to new RDDs. Example of transformations:

  - **map***(func)* that passes each dataset element through a function and returns a new RDD representing the results;
  - **filter***(func)* return a new dataset formed by selecting those elements of the source on which func returns true;
  - **distinct**[*numPartition*] return a new dataset that contains the distinct elements of the source dataset.

- *Actions* return values or results to the driver program. Example of actions:
  - **reduce***(func)* that aggregates all the elements of the RDD using some function and returns the final result to the driver program;
  - **collect()** return all the elements of the dataset as an array at the driver program. This is usually useful after a filter or other operation that returns a sufficiently small subset of the data.
  - **foreach***(func)* passes each element through a user provided function.

An application on Apache Spark is usually split into a driver and several executors. The driver takes care of the execution of the spark application, managing the resources to be allocated and the tasks to be performed by each executor running in the cluster, while the driver may be running on the client. The instance of the SparkContext object in the Spark (driver) program is responsible for requesting the resources needed to execute the executors. A Spark application is formed by jobs, one for each action, each job is composed of a set of stages that depend on each other executed sequentially, each of them is executed by many tasks, performed parallel by the executors, an example is shown in Figure 2.12. Executors perform the tasks assigned by the driver, they have an assigned amount of memory allocated and cannot communicate with each other unless they first save the data to disk. The executors have a lifetime equal to the lifetime of the application.
Apache Spark has introduced several improvements for its data abstraction which yield a better computation model as well. One of these

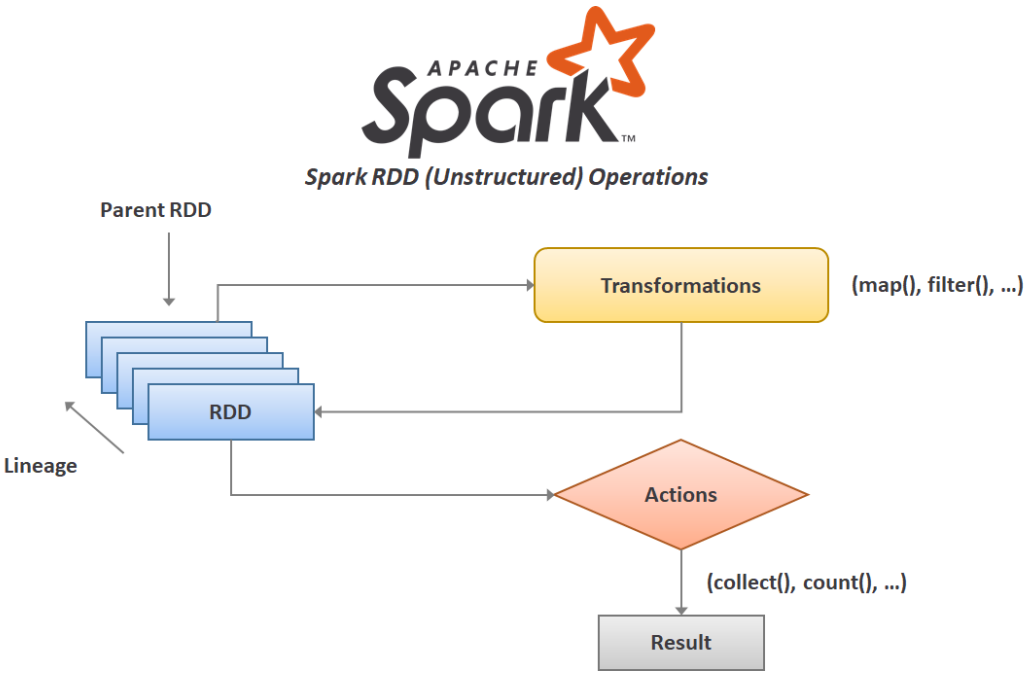


Figure 2.11: RDD (Resilient Distributed Dataset). Adapted from spark.apache.org.

improvements is the DataFrame API which is part of Spark SQL [7]. A DataFrame is conceptually equivalent to a table in a relational database. It is a distributed collection of data, like RDD, but organized into named columns. Another improvement is the Dataset API which is a new experimental interface added in Spark 1.6. It is an extension of the DataFrame API that provides a type-safe, object-oriented programming interface. A Dataset is a strongly typed, immutable collection of objects that are mapped to a relational schema [7].

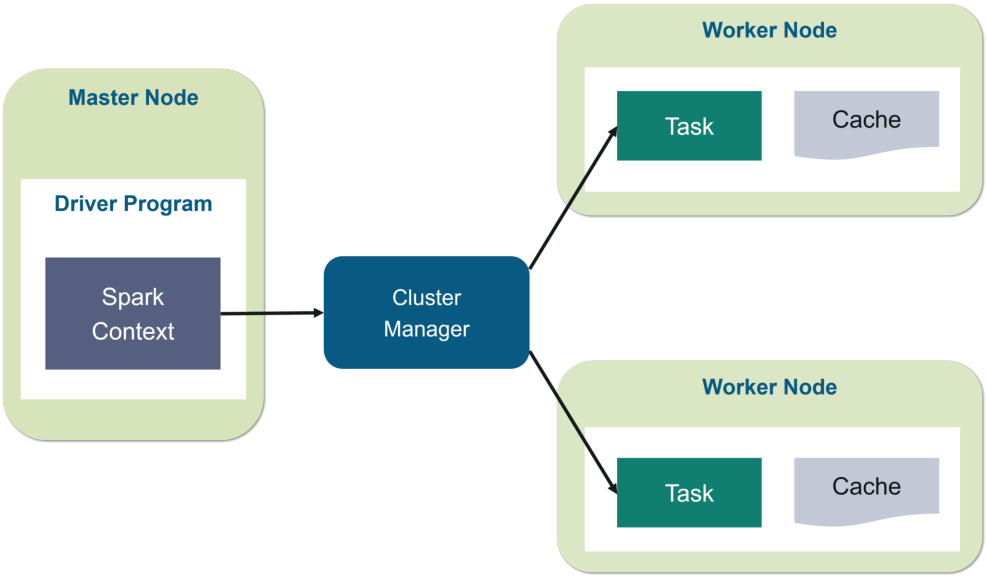


Figure 2.12: Apache Spark Architecture. Adapted from spark.apache.org.

Chapter 3

Problems and state of the art

Abstract

This Chapter presents the detailed analysis and study of the literature focusing on knowledge extraction from two different contexts: Social and Biological Networks. The section 3.1 is concerned with the presentation of literature in the context of semantic and action-based approaches for the optimization of Advertising Campaigns. The section 3.2 is concerned with the literature for the identification of significant descriptors of a Biological Network and the prediction of new association on methods based on two different categories already known/unknown associations and matrix factorization.

3.1 Social Networks

This paragraph shows the study of the literature and the detailed analysis of the Social Networks context, for the optimization of Advertising Campaigns.

3.1.1 Optimization of Advertising Campaigns

Modeling the user profiles from social media raw data is usually a challenging task. The extensive spread of the Internet around the world

has created an effective medium for instant communication at low or no cost, and online communication has become an important platform for consumers to express their opinions about experiences concerning products [28]. Consumers are motivated to use social networking sites for a variety of reasons. Chenyan Xu in [105] discussed several different theoretical reasons and concluded that gratification, utilitarian motivation and social presence provided explanations for the user's motivations to use social networking sites. Spreading product relation information through Facebook can be evaluated from the diffusion of innovation perspective where the motivations can be studied. Facebook [51] is by far one of the largest social networking sites with over 800 million users who frequently log-on to the site every day. Facebook creates a platform for users to talk about their favorite interests and hobbies with friends. Therefore, suggestions and recommendations from friends may be considered to be a more influential source of product information. Consumers might be more receptive to such information and, perhaps, are more likely to try a product because of the recommendation by their friends. The approaches proposed in the literature to this aim may be roughly classified in two main categories:

- The first category includes approaches based on the analysis of user generated contents (here referred to as *semantic approaches*);
- The second category of approaches characterize individuals by "actions", e.g., visited web pages (*action-based approaches*).

3.1.2 Semantic approaches

The authors of [93] use Differential Language Analysis (DLA) in order to find language features across millions of Facebook messages that distinguish demographic and psychological attributes. They show that their approach can yield additional insights (correlations between personality and behavior as manifest through language) and more information (as measured through predictive accuracy) than traditional apriori word-category approaches. The framework proposed in [64] relies on a semi-supervised topic model to construct a representation of an app's version as a set of latent topics from version metadata and textual descriptions.

The authors discriminate the topics based on genre information and weight them on a per-user basis, in order to generate a version-sensitive ranked list of apps for a target user. In [62] the authors propose a dynamic user and word embedding algorithm that can jointly and dynamically model user and word representations in the same semantic space. They consider the context of streams of documents in Twitter, and propose a scalable black-box variational inference algorithm to infer the dynamic embeddings of both users and words in streams. They also propose a streaming keyword diversification model to diversify top-K keywords for characterizing users' profiles over time.
The first technique applied to brand-affinity matching that is not an action-based approach has been presented in [15]. In this work [19] the author proposes a general framework for the recommendation of possible customers (users) to advertisers (e.g., brands) based on the comparison between OSN profiles. This approach belongs to the first category, discussed above, to the best of our knowledge the only techniques applied to brand-affinity matching are action-based approaches. In particular, the method associates suitable categories and subcategories to both user and brand profiles in the considered OSN. When categories involve posts and comments, the comparison is based on word embedding, and this allows to take into account the similarity between the topics of particular interest for a brand and the user preferences. Furthermore, user personal information, such as age, job or genre, are used for targeting specific advertising campaigns.

3.1.3 Action-based approaches

In [84] individuals are associated with each other due to some actions they share (e.g., they have visited the same web pages). The proximity between individuals on networks built upon such relationships is informative about their profile matching. In particular, brand-affinity audiences are built by selecting the social-network neighbors of existing brand actors, identified via co-visitation of social-networking pages. This is achieved without saving any information about the identities of the browsers or content of the Social Network pages, thus allowing for user anonymization. In [2] compact and effective user profiles are generated

from the history of user actions, i.e., a mixture of user interests over a period of time. The authors propose a streaming, distributed inference algorithm which is able to handle tens of millions of users. They show that their model contributes towards improved behavioral targeting of display advertising relative to baseline models that do not incorporate topical and/or temporal dependencies. In [53] a computer user behavior is represented as the sequence of the commands she/he types during her/his work. This sequence is transformed into a distribution of relevant subsequences of commands in order to find out a profile that defines its behavior. Also, because a user profile is not necessarily fixed but rather it evolves/changes, the authors propose an evolving method to keep up to date the created profiles using an Evolving Systems approach. The observation that behavior of users is highly influenced by the behavior of their neighbors or community members is used in [103] to enrich user profiles, based on latent user communities in collaborative tagging.

3.2 Biological Networks

In this paragraph there is the study of the literature of two problems in the Biological Network context. First, the paradigm for the identification of significant "global" descriptors of a Biological Network, relying on the characterization of the relevance of nodes and edges across the network structure; second, the prediction of new associations of lncRNAs-diseases.

3.2.1 Topological measures

Biological Networks topology yields important insights into biological function, occurrence of diseases and drug design. In the last few years, different types of topological measures have been introduced and applied to infer the biological relevance of network components-interactions, according to their position within the network structure.
Two Biological Networks $\mathcal{N} = \langle V, E \rangle$ and $\mathcal{N}' = \langle V', E' \rangle$ are isomorphic ($\mathcal{N} \simeq \mathcal{N}'$) if there exists a bijection $\phi : V \to V'$ such that $(u, v) \in E$ if and only if $(\phi(u), \phi(v)) \in E'$. Similarly to the definition in [20], we provide the following.

Definition 27 (Topological measure) *Let $\mathcal{N} = \langle V, E\rangle$ and $\mathcal{N}' = \langle V', E'\rangle$ be two isomorphic biological networks. Let X be V or E. A real-valued function $w : X \to \mathbb{R}$ is a topological measure if and only if: $\forall x \in X$, $\mathcal{N} \simeq \mathcal{N}' \implies w_{\mathcal{N}}(x) = w_{\mathcal{N}'}(\phi(x))$, where $w_{\mathcal{N}}(x)$ denotes the value $w(x)$ in $\mathcal{N}$.*

Definition 28 (Topological Overlap Measures (TOM)) *Given an edge (i, j), the Topological Overlap Measure (TOM) and its Generalized version (GTOMm) score the degree of overlap among the neighbors of those two nodes. The larger the overlap, the higher the weight assigned to (i, j). TOM considers only the immediate neighbors, whereas* GTOMm *includes all the neighbors at distance $\leq m$, as follows:*

$$w_{ij} = \begin{cases} \dfrac{|N_m(i) \cap N_m(j)| + a_{ij}}{min\{|N_m(i)|, |N_m(j)|\} + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{otherwise} \end{cases} \tag{3.1}$$

In Equation 3.1, $N_m(i)$ denotes the set of i's neighbors reachable with a shortest path of length at most m from i, and $a_{ij} = 1$ if and only if there exists an edge connecting vertices i and j, otherwise $a_{ij} = 0$. The following definition refers to analogous ones as introduced in literature [88, 106].

Definition 29 (Edge Clustering Value) *Similarly to TOM, Edge Clustering Value (ECV) quantifies how much the i's and j's neighborhood overlap:*

$$w_{ij} = \frac{|N_1(i) \cap N_1(j)|^2}{|N_1(i)| \cdot |N_1(j)|}. \tag{3.2}$$

Unlike TOM, which normalizes the size of common neighborhood over the smallest between i and j neighborhoods (see Equation 3.1), ECV is equal to 1 if and only if i and j have the same *exact* neighbors. It is worth noting that both TOM and ECV can be interpreted as a biological, neighborhood-normalized version of Granovetter's *embeddeness* measure [66], historically used to characterize tie-strength in social networks. The following definition refers to analogous ones as introduced in the literature[97].

For a given edge (i, j), the Dispersion measure [9] extends Granovetter's

tie-strength measure [66], taking into account both the size and the *connectivity* of i,j's common neighborhood. Intuitively, it quantifies how *"not well"*-connected is the i,j's common neighborhood within G_i, the subgraph induced by i and its neighbors. More formally, let G_i be the subgraph induced in G by $\{i\} \cup N_1(i)$. In G_i, let j be a neighbor of i (i.e. $j \in N_1(i)$), and denote with $C_{ij}^{(i)} = N_1(i) \cap N_1(j)$ the set of common neighbors of i and j *within* the induced subgraph G_i.

Definition 30 (Dispersion) *The absolute dispersion is defined as follows:*

$$disp(i,j) = \sum_{s,t \in C_{ij}^{(i)}} d_v(s,t) \tag{3.3}$$

where d_v is a boolean function such that $d_v = 1$ if and only if s and t are not connected by a path of length ≤ 2 in G_i. The following definition refers to analogous ones as introduced in the literature [9].

The authors define two enhanced versions of dispersion: *parametric* dispersion (Equation 3.4) and *recursive* dispersion (Equation 3.5):

$$param(i,j,\alpha,\beta,\gamma) = \frac{(disp(i,j)+\beta)^\alpha}{emb(i,j)+\gamma} \tag{3.4}$$

$$rec(i,j) \leftarrow \frac{\sum_{w \in C_{ij}^{(i)}} x_w^2 + 2\sum_{s,t \in C_{ij}} d_v(s,t) x_s x_t}{emb(i,j)} \tag{3.5}$$

where $emb(i,j)$ is equal $|C_{ij}^{(i)}|$ (see details in [9]). It is easy to show that none of these measures are symmetric. Since in our context we are considering undirected graphs, we unambiguously assign a weight to the edge (i,j) by defining and applying the following three variants:

1. **rec_max** (KB1). Assigns a dispersion weight to edge (i,j) as $w_{ij} = rec\_max(i,j) = MAX\{rec(i,j), rec(j,i)\}$.

2. **rec_min** (KB2). Assigns a dispersion weight to edge (i,j) as $w_{ij} = rec\_min(i,j) = MIN\{rec(i,j), rec(j,i)\}$.

3. **param_sym** (KB3). Computes a *symmetric* parametric variant of $w_{ij} = param(i, j, \alpha, \beta, \gamma)$ with parameters $\alpha = 0.61$, $\beta = 0$, $\gamma = 5.0$ (i.e., parameters used by authors, see details in [9]), considering the common neighborhood C_{ij} in G (and not in the induced subgraphs G_i, G_j, as in the original – non-symmetric– definition).

Definition 31 (Edge Betweenness) *Given an edge e_{ij}, the Edge Betweenness (EB) is the fraction of shortest paths in the network $\mathcal{N}$ containing it:*

$$w_{ij} = \sum_{s,t \in V} \frac{\sigma_{st}(e_{ij})}{\sigma_{st}} \tag{3.6}$$

In Equation 3.6, σ_{st} represents the total number of shortest paths connecting nodes s, t ($s \neq t$) in the network, whereas $\sigma_{st}(e_{ij})$ counts only the shortest paths between the same nodes containing the edge e_{ij}. The higher w_{ij} is, the more likely the edge e_{ij} acts as a *bridge* connecting separate communities, i.e., it represents an *inter*-community edge. The following definition refers to analogous ones as introduced in the literature[45].

Definition 32 (Edge Clustering Coefficient) *Given the edge e_{ij}, the Edge Clustering Coefficient (ECC3) is the number of triangles the edge e_{ij} belongs to, divided by the number of triangles that might potentially include it:*

$$w_{ij} = -\frac{z^3_{i,j} + 1}{s^{(3)}_{i,j}} = -\frac{|N_1(i) \cap N_1(j)| + 1}{min[(|N_1(i)| - 1)(|N_1(j)| - 1)]} \tag{3.7}$$

In Equation 3.7, $z^3_{i,j}$ is the number of triangles built on the edge e_{ij} and $s^{(3)}_{i,j}$ is the maximal possible number of them. The minus sign is explained as follows. Many triangles exist within dense communities. Therefore, the higher $|w_{ij}|$ the more likely e_{ij} lies within a dense community, being an *intra*-community edge. Nevertheless, in order to identify communities [87] use $|w_{ij}|$ in a Girvan-Newman fashion [45]: at each step of the divisive algorithm, edges with the *lowest* $|w_{ij}|$ are removed, eventually splitting the original network into separate connected components.

Definition 33 (Edge Centrality Proximity Distance) *Edge Centrality Proximity Distance (ECPd) is based on the notion of Edge Centrality:*

$$L^{\kappa}(e) = \sum_{s \in V} \frac{\sigma_s^{\kappa}(e)}{\sigma_s^{\kappa}} \tag{3.8}$$

$L^{\kappa}(e_{ik})$ represents the fraction of times a random walker traverses the edge e_{ik} running through a random sample path of length at most κ. The following definition refers to analogous ones as introduced in the literature [68, 69].

Definition 34 (Edge Centrality Proximity Distance) *(ECPd) is defined as:*

$$w_{ij} = 1 - \sqrt{\sum_{k=1}^{n} \frac{(L^{\kappa}(e_{ik}) - L^{\kappa}(e_{kj}))^2}{d(k)}} \tag{3.9}$$

Equation 3.9 represents a *distance* between nodes i and j: the higher w_{ij} the more likely the nodes i and j belong to different communities, and e_{ij} being an inter-community edge. In particular, $(L^{\kappa}(e_{ik}) - L^{\kappa}(e_{kj}))^2$ expresses a *proximity* between nodes i and j: the probability that a message propagated to node i reaches also node j, with node k being a common neighbor of those the two.

Definition 35 (Node Clustering Coefficient) *Given a node i, Node Clustering Coefficient (NCC) expresses how densely connected is the i's neighborhood. Let k_i be the number of neighbors of i, and let n_i be the number of edges connecting such neighbors:*

$$x_i = \frac{2n_i}{k_i(k_i - 1)}. \tag{3.10}$$

In Equation 3.10, the denominator is equal to the maximum value for k_i (recall $\mathcal{N}$ is an undirected graph). As a result, the greater the value of x_i, the closer the i's neighborhood to be a clique. The following definition refers to analogous ones as introduced in the literature [100].

Definition 36 (Eigenvector Centrality) *Given a node i, the Eigenvector centrality (EGC) is a measure of topological "importance" for the*

node i in the network $\mathcal{N}$. Specifically, node i can acquire high centrality either by having a high degree or by being connected to other highly-important.

$$x_i = \frac{1}{\lambda} \sum_{j \in G} A_{ij}\, x_j \tag{3.11}$$

The following definitions refer to analogous ones as introduced in the literature [13, 20]. In Equation 3.11, λ is the largest positive eigenvalue of the adjacency matrix A, satisfying the equation $A\,\boldsymbol{x} = \lambda\,\boldsymbol{x}$, with $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ being the vector of node centralities. Notably, Google's PageRank [75] is a variant of Eigenvector Centrality.

Definition 37 (Betweenness Centrality) *Given a node i, Node Betweenness (BC) quantifies the extent to which node i lies on geodesic (shortest) paths between other pairs of vertices:*

$$x_i = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3.12}$$

The following definition refers to analogous ones as introduced in the literature [38]. In Equation 3.12, $\sigma_{st}(v)$ represents the number of shortest paths from node s to node t that pass through i, whereas σ_{st} is the total number of shortest paths between the same nodes. Intuitively, the higher x_i the more likely i lies on a path between nodes in different communities.

Definition 38 (Subgraph Centrality) *Given a node i, Subgraph Centrality (SGC) quantifies the centrality of node i based on the number of subgraphs it belongs to:*

$$x_i = \sum_{k=0}^{\infty} \frac{(A^k)_{ii}}{k!} \tag{3.13}$$

In Equation 3.13, $(A^k)_{ii}$ is the number of closed paths of length k, starting and ending on node i. Closed walks are weighted such that smaller walks are given higher weights (see [33]). Therefore, smaller subgraphs are given higher weights than larger ones, which makes SGC able to

quantify the extent to which node i participates in network motifs within real-world networks [33, 73, 74].

Definition 39 (κ-Path Centrality) *Given a node i, the κ-Path Centrality (KPC) is defined as the sum, over all possible source nodes s, of the probability that a message originating in s goes through i, assuming the message runs along random simple paths of length at most κ:*

$$x_i = \sum_{s \neq i} \frac{\sigma_s^\kappa(i)}{\sigma_s^\kappa} \tag{3.14}$$

The following definition refers to analogous ones as introduced in the literature [5]. In Equation 3.14, $\sigma_s^\kappa(i)$ is the number of messages originating at node s passing through node i, whereas σ_s^κ is the total number of messages originated from node s. Despite a similar formulation, KPC differs substantially from NB. In particular, KPC does not assume information flows necessarily across shortest paths, as NB does.
This type of measures, which originated initially in the context of Social Networks Analysis, find practical application in numerous other types of networks such as that of the internet, urban networks, networks representing the spread and contagion of a disease and, of course, also in Biological Networks.

3.2.2 Prediction of lncRNAs-diseases Associations

The approaches may be divided in two different categories:

- those that do not use already known lncRNAs-diseases associations;
- those that do use known lncRNAs-diseases associations, which usually rely on matrix factorization.

Methods based on already known/unknown associations The method presented in [24] (**HGLDA**) is not based on already known lncRNA-disease associations. The use of an HyperGeometric distribution for LDAs inference is proposed in order to predict LDAs by

integrating miRNA-disease associations and lncRNA-miRNA interactions. HGLDA has been successfully applied to predict Breast Cancer, Lung Cancer and Colorectal Cancer-related lncRNAs. To quantify the functional similarity of lncRNAs on a large scale, the model LncRNA Functional Similarity Calculation based on the information of miRNAs (LFSCM) used, which integrates semantic similarities, miRNA-disease associations and lncRNA-miRNA interactions.

The **ncPred** method uses interactions experimentally verified. This approach provides a recommendation technique to find novel lncRNA-disease associations proposed in [6]. ncPred is based on a resource propagation methodology, which uses a tripartite network to guide the inference process of novel ncRNA–disease associations. The tripartite network permits exploiting the interactions between ncRNA–target and target–disease. The method uses two datasets containing experimentally verified interactions between ncRNAs, targets, and diseases. Interactions in the considered network associate each ncRNA with a disease through its targets. The algorithm is based on a multilevel resource transfer technique, which computes the weights between each ncRNA–disease pair and, at each step, considers the resource transferred from the previous step.

An integrative framework, **IntNetLncSim**, is presented in [26] to infer lncRNA functional similarity by modeling the information flow in an integrated network that comprises both lncRNA related transcriptional and posttranscriptional information. An approach that relies on the analysis of lncRNAs related information stored in public databases, as well as their interactions with other types of molecules is described in [17, 18]. In particular, large amounts of lncRNA–miRNA interactions (LMIs) have been collected in public databases, and plenty of experimentally confirmed MDAs are available as well. Therefore, the prediction of LDAs may be based on known LMIs, and MDAs. In the considered approach, the problem of LDA prediction is modeled as a neighborhood analysis performed on tripartite graphs (described in Chapter 2), in which the three sets of vertices represent lncRNAs, miRNAs, and diseases, and the vertices are linked according to LMIs and MDAs.

Methods based on matrix factorization Methods in the literature based on recommendation systems and the matrix of preferences are described in this section. *Collaborative filtering* is a principal problem in recommendation research. In the more abstract sense, collaborative filtering is the problem of weighting missing edges in a bipartite graph. Typically, this bipartite graph is represented by its adjacency matrix, which is called the *preference matrix*. In the literature on recommender systems in general and collaborative filtering specifically, two dominant perspectives have emerged: the model based perspective and the memory-based perspective [89].

The method **MFLDA** [96] is based on a different technique: the matrix of recommendation. In general these matrix factorization-based models (see Figure 3.1) show great potential in recovering latent associations between various biological molecules. They implicitly assume that each data source has equal relevance towards the target prediction task, and do not differentiate among the quality of different data sources. Therefore, their performance might be seriously compromised by noisy (irrelevant or low quality) data sources. MFLDA first encoded directly (or indirectly) relevant data sources related to lncRNAs or diseases in individual relational data matrices and presets weights for these matrices. Next, it simultaneously optimizes the weights and low-rank matrix tri-factorization of each relational data matrix.

Another method [108] based on Collaborative Filtering model called **CFNBC** for inferring potential lncRNA-disease associations is proposed based on Naïve Bayesian Classifier. In CFNBC, an original lncRNA-miRNA-disease tripartite network is constructed first by integrating known miRNA-lncRNA associations, miRNA-disease associations and lncRNA-disease associations, and then, an updated lncRNA-miRNA-disease tripartite network is further constructed through applying the item-based collaborative filtering algorithm on the original tripartite network. These case studies of glioma, colorectal cancer and gastric cancer demonstrate the excellent prediction performance of CFNBC as well. A computational approach using graph regularized non-negative matrix factorization (**LDGRNMF**) [98], which considers disease associated lncRNAs identification as recommendation system problem. This method calculates the disease similarity matrix based on known lncR-

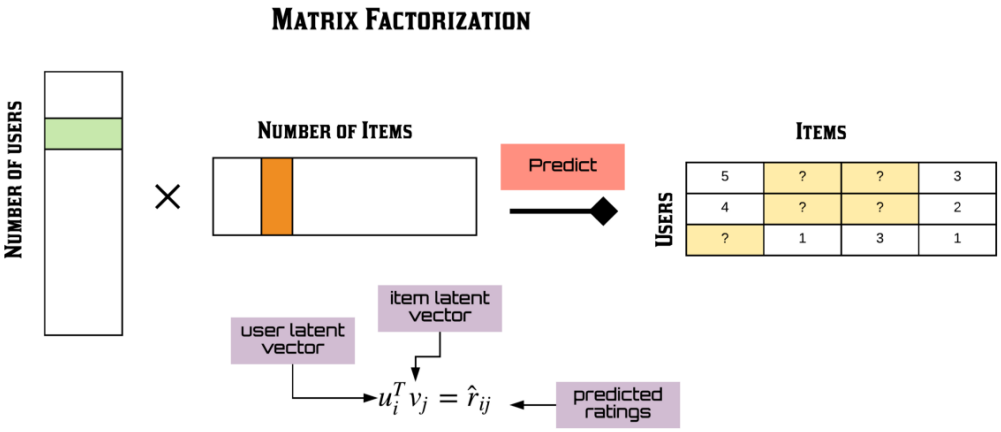


Figure 3.1: Representation of Matrix Factorization: the goal of a recommendation system is to predict the blanks in the utility matrix. Copyright (Recommendation System series part 4: the 7 variants of matrix factorization for Collaborative Filtering) (https://towardsdatascience.com).

NAs -diseases associations, disease semantic information and the similarity matrix of lncRNAs based on known lncRNAs -diseases associations. In this study, the semantic similarities between different diseases can be computed using directed acyclic graphs (DAGs) based on the Mesh database, where nodes represent diseases and edges represent the association between diseases.

Part II

Proposed Approaches

Chapter 4

Optimization of Advertising Campaign

Abstract

This Chapter presents an approach based on the use of automated system that founds the k best customers for Advertisement Campaigns via OSN. The proposed approach is based on two main aspects: the comparison between Online Social Network profiles and neighborhoods analysis on the Online Social Network. Profile matching between users and brands is considered based on bag of words representation of textual contents coming from the social media, and measures such as the Term Frequency-Inverse Document Frequency are used in order to characterize the importance of words in the comparison.

4.1 Introduction

In the last few years, with the exponential growth in the use of social media (reviews, forums, discussions, blogs and Social Networks), people and companies are increasingly using information (opinions and preferences) published in these media for their decision-making process. However, monitoring and searching for opinions on the Web by one or more companies is a very difficult problem due to the proliferation of thousands of sites; in addition, each site contains a huge volume of text

that cannot always be optimally decipherable (e.g., the long messages of forums and blogs). The use of automated systems is therefore necessary.

4.2 A novel approach for the optimization of an Advertising Campaign

The main goal of the proposed approach [15, 19] is to identify the most suitable k possible *buyers* to whom distributing a given advertisement campaign. To this aim, two important aspects have to be taken into account:

- Ideally, users to whom distributing the campaign should have interests compatible with the specific features of the advertiser (i.e., the *brand*);

- It would be better if the chosen possible buyers would know other users whose interests are close to those expected for the campaign success as well.

User profiles complement network topology information. In particular, each node in the network points to data associated with a user and retrieved from the considered social media. An important aspect for this research is the textual information available about user general interests and activities, coming for example from private communications, posts, comments, short text messages [15]. Therefore, the user profile of u is represented here by a text T_u, characterizing u with references to the considered OSN.
Also a brand profile is represented by a text, that can be for example easily extracted from the web-page describing brand activities or from other textual documents containing information on the advertisement campaign. In the following, we refer indistinctly to brand profile and advertisement campaign, since both may be described by textual documents and then handled in the same way in the context of the proposed approach.

Profile matching. Profile matching between users and brands is considered, based on bag-of-words representation of textual contents coming from the social media. Let u be a node in an input OSN $\mathcal{N}$ and T_u be its user profile. Moreover, let T_b, the text associated with the brand profile. The first step of our approach is to understand how much T_u and T_b are "similar", i.e., to which extent they *match* each other. To this aim, we consider Term Frequency Inverse Document Frequency (TF-IDF) and cosine similarity measures in order to understand if and how much textual contents associated to u and to the brand are semantically related. This is sketched in the following, for the specific case under consideration (i.e., only two textual documents, T_u and T_b). Let w_{ij} be a word occurring in the text T_j $(j = 1, \ldots, m)$.

Definition 40 (Term Frequency Inverse Document Frequency) *We introduce the measure of Term Frequency Inverse Document Frequency. Let $\{D_1, ...D_m\}$ be a set of textual documents and w_{ij} be a word occuring in the document $D_j (j = 1, ...m)$. This function for w_{ij} is defined as:*

$$TF\text{-}IDF(w_{ij}) = TF(w_{ij}) * IDF(w_{ij}) \tag{4.1}$$

such that:

$$TF(w_{ij}) = \frac{|w_{ij}|}{|T_j|} \tag{4.2}$$

where $|w_{ij}|$ is the frequency of the term w_{ij} in the text T_j and $|T_j|$ is the number of words in T_j, and: $IDF(w_{ij}) = \log \frac{m}{h}$, where $h \leq m$ is the number of texts where w_{ij} occurs.
The following definition used in [15, 19] is needed in order to provide the notion of match between the social profiles considered here. The TF-IDF is used to weigh the importance of the words inside the text.

Definition 41 (Cosine Similarity) *Let V_u and V_b be two arrays of k real values. The cosine similarity between V_u and V_b is defined as:*

$$\mathcal{CS}(V_u, V_b) = \frac{\sum_{i=1}^{k} V_u[i] * V_b[i]}{\sqrt{\sum_{i=1}^{k} V_u[i]^2} * \sqrt{\sum_{i=1}^{k} V_b[i]^2}} \tag{4.3}$$

The numerator is the inner product of the two arrays and the denominator is the product of the norms of the two arrays.

The cosine similarity is computed in order to measure the match between customer and brand profiles.
Three measures were calculated for each node: affinity, centrality and utility.

Definition 42 (Affinity) *The value of affinity between the profiles associated to an user and a brand is then computed as the cosine similarity between arrays containing the TF-IDF values of the words occurring in T_u and T_b:*

$$\mathcal{A}(T_u, T_b) = \mathcal{CS}(V_u, V_b) \tag{4.4}$$

In order to make more effective the advertisement campaign, for each node u in V, it is important not only to measure to what extent its profile matches with the brand profile, but also how many nodes in the neighborhood of u could be possibly interested in that campaign as well. That is, the best targets are those nodes whose profile matches with the brand, and that are surrounded by other nodes with this same feature. Let u be a node in the set of vertices V and T_u and T_b be the user and brand profiles, respectively. Moreover, let N_u be the set of nodes linked to u by at least one edge in the set of edges E of $\mathcal{N}$. In Figure 4.1 an example of a small OSN, with the measure is shown.

Definition 43 (Centrality) *The* centrality *of u for the given considered brand (or advertisement campaign) is defined as:*

$$\mathcal{C}(u, b) = \frac{\sum_{v \in N_u} \mathcal{A}(T_v, T_b)}{|N_u|} \tag{4.5}$$

It is worth pointing out that, in order to focus the advertising campaign on those interested users only, a threshold value can be chosen on the affinity values according to which filtering only nodes in the network scoring affinity values larger than that threshold. As already explained, the final aim of our approach is to identify the best k nodes to which distribute advertisements according to their profile matching with the

brand (or campaign, respectively). On the other hand, in order to maximize the gain, we are also interested in detecting nodes whose neighbors in the OSN may be interested in the same advertisements.

Definition 44 (Utility) *The* utility *of a node for a specific brand/campaign is defined as follows:*

$$\mathcal{U}(u,b) = \alpha \cdot \mathcal{A}(T_u, T_b) + (1-\alpha) \cdot \mathcal{C}(u,b) \tag{4.6}$$

where α is a real value in $[0,1]$ used to balance two different contributions, i.e., the match between user and brand, and the match between user neighbors and brand.

Example 1 Figure 4.1 depicts a small OSN and, for each node, the corresponding affinity value that is supposed to be computed with respect to a given brand is also shown. Suppose that the brand is interested to send its advertising campaign to 5 nodes on this network (i.e., $k = 5$).

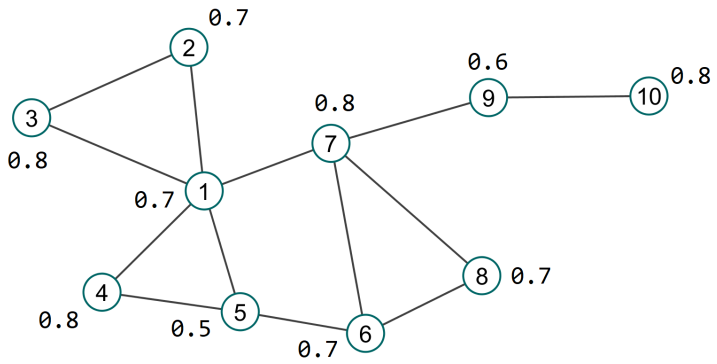


Figure 4.1: A small OSN. For each node, the corresponding affinity value is also shown.

As an example, for Node 1:

$$\mathcal{C}(1,b) = \frac{0.7 + 0.8 + 0.8 + 0.5 + 0.8}{5} = 0.72$$

and, for $\alpha = 0.4$:

$$\mathcal{U}(1, b) = 0.4 \cdot 0.7 + (1 - 0.4) \cdot 0.72 = 0.71$$

while for $\alpha = 0.6$:

$$\mathcal{U}(1, b) = 0.6 \cdot 0.7 + (1 - 0.6) \cdot 0.72 = 0.7$$

Usually, to the best of our knowledge, the available data on OSNs consist only on the graph topology, no information about user interests and profiles are publicly available. Web scraping has been used here in order to collect and extract useful contents for user profiles characterization. In particular, we have avoided to associate randomly the information obtained by web scraping to nodes in the considered OSN graph, due to the fact that a random association would have altered the natural mechanism according to which users in the same neighbors have similar interests.
We have considered the web-pages associated with four brands. We consider TF-IDF and cosine similarity measures in order to understand if and how much textual contents associated with the user and to the brand are semantically related. The networks are built upon the following information: each node corresponds to a web page that could be associated with a brand (see Figure 4.2). This technique is based on measures which aim at detecting neighbor nodes with similar interests. Firstly this approach selects some nodes from the `twitter-2010` OSN and some web-pages focused on different topics (cooking, fashion, cars, etc.), but the method avoids to associate randomly the information obtained by web scraping. User profile may be obtained by scraping the contents of a web-page on a specific topic for example for Amarelli Brand (as shown in Figure 4.3). Then, a visit in depth of the OSN has been performed starting from each of the seeds and stopping when the entire network was visited. For each new node to be visited, a new web-page has been visited as well, following the cross-page links on the considered web-pages. Secondly, the values of affinity, centrality and affinity are

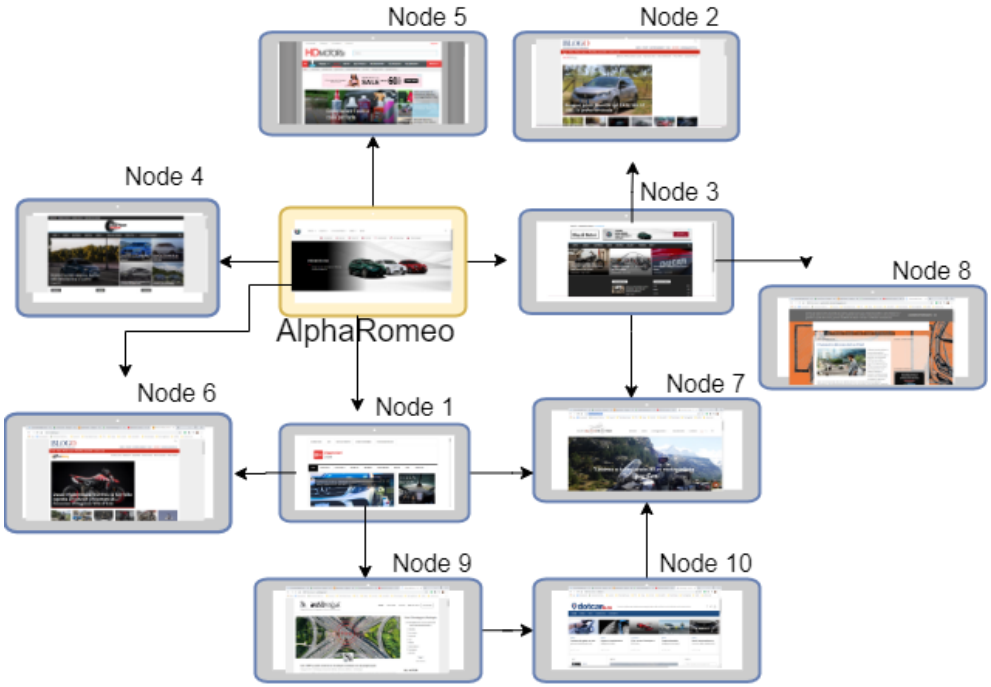


Figure 4.2: Links among the first 10 target nodes for Alfa Romeo Brand.

computed for each brand (node of the network). Then these values are ranked in descending order, according to each of these measures. The results have been compared with a random choice of the k nodes to which distribute the advertisement (as shown in Chapter 7).

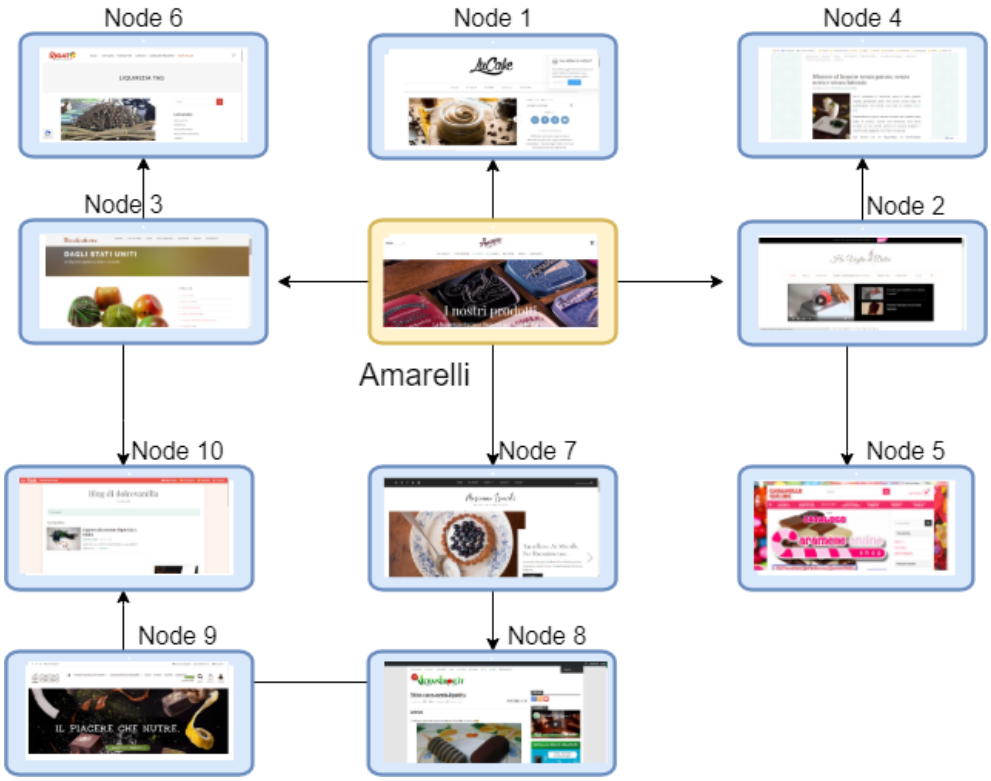


Figure 4.3: Links among the first 10 target nodes for Amarelli Brand.

Chapter 5

Extracting knowledge encoded in the Biological Networks topology

Abstract

Here a paradigm is described for the identification of significant global descriptors of a biological network, relying on the characterization of the relevance of nodes and edges across the network structure. To this aim, the main building boxes of our methodological framework are topological measures, which assign a real weight to nodes or edges based not only on the network topology, but also of novel, hidden, functional information.

5.1 Introduction

The main goal of the approach described here is to provide an overview for the identification of significant "global" descriptors of a Biological Network, based on the characterization of the relevance of (nodes) edges across the network structure. The paradigm for the identification of significant "global" descriptors of a Biological Network relies on the characterization of the relevance of nodes and edges across the network structure. To this aim, the main building boxes of our methodological

framework are topological measures [45], which assign a real weight to nodes or edges based only on the network topology. In particular, it has emerged recently that systematic explorations of different topological measures, with a comprehensive understanding of their dependence on the specific network type and application context, may contribute to accelerate the solution of difficult problems (e.g., drug repositioning) and their clinical translation [10]. It is worth pointing out that the proposed methodology is general enough to apply also in contexts different from biological systems.

5.2 A novel approach to infer hidden knowledge via topological rank

This approach builds the compact hierarchical views from the topological ranks of nodes and edges induced by such measures, in two different assets, *static* and *dynamic*. While in the former case the network is considered in its entirety, in the latter one weights and rankings are assigned dynamically during the relevance discovery process. We propose a methodology for the evaluation of topological ranks obtained from different measures, that relies on two different criteria:

- *statistical significance*, via Montecarlo Hypothesis Test;
- *biological relevance*, quantified by comparing topological ranks against those obtained from external knowledge (e.g., gold standards).

The former is a sort of *internal* criteria, which allows us to discriminate the most significant ranks independently from the specific application context. The latter criteria aims to measure to what extent hidden information may be retrieved from a Biological Network taken as a whole, and intuitively this depends also on the specific type of information one is looking for.

Biological Networks. The Biological Networks considered in this study may be roughly distinguished in two main categories:

- *genotype-phenotype associations* networks;
- *physical-interaction* networks.

Six Biological Networks involving three organisms (C.elegans, S. cerevisiae, H. sapiens) and seven different gold standards have been included in the experimental analysis.

Methodology. For the purpose of this study, among the measures defined in the Chapter 2, two classes are of interest: *incremental* and *decremental* (see in Table 5.1 and 5.2). We consider a *incremental view* to display the subgraphs induced by incrementally considering sets in the edge rank, according to the priority, i.e. 'relevance', of the edges given by the rank. And a *decremental view* to remove edges according to the edge rank, such that the priority of the rank indicates irrelevance.

We propose to build compact hierarchical views from the *topological*

	Incremental	
TOM	Topological Overlap Measure	[88, 106]
GTOMm	Generalized Topological Overlap Measure	[88, 106]
ECV	Edge Clustering Value	[97]
KB1	Variant of Dispersion	[9]
KB2	Variant of Dispersion	[9]
KB3	Variant of Dispersion	[9]
	Decremental	
EB	Edge Betweenness	[45]
ECC3	Edge Clustering Coefficient	[87]
ECP	Edge Centrality Proximity Distance	[68]

Table 5.1: Edge topological measures.

ranks of nodes and edges induced by such measures, in two different assets, *static* and *dynamic*, shown in Tables 5.2 and 5.1. In the first asset the value of the input topological measure is computed for each of the edges of the input graph. In the second asset, at each step, edges with the highest score are appended to the partial solution and deleted from the graph. This allows to intercept edges with an important role in

Incremental		
NCC	Node Clustering Coefficient	[100]
EGC	Eigenvector centrality	[13]
Decremental		
BC	Node Betweenness	[38]
SGC	Subgraph Centrality	[33]
KPC	κ-Path Centrality	[5]

Table 5.2: Node topological measures.

their topological context, yet hidden by other edges scoring much higher values. We provide explicit definitions for edge (node, resp.) ranks and their associated views, together with procedures for generating them with the use of topological measures.

Definition 45 (Edge rank) *An* edge rank *of $\mathcal{N}$ is an ordered list $\mathcal{E} = (E_1, E_2, \cdots, E_k)$ of subsets of E such that they are a partition of E.*

Intuitively, by displaying the subgraphs induced by incrementally considering, in the order given, the sets in $\mathcal{E}$, one can get incremental views of $\mathcal{N}$, according to the priority, i.e, "relevance", of the edges given by the ranking. A decremental view can be obtained analogously by removing edges according to $\mathcal{E}$. In this latter case, the priority of the ranking indicates irrelevance. Formally, one can define a sequence of *views* of $\mathcal{N}$, based on $\mathcal{E}$, as follows.

Definition 46 (i-th incremental (decremental) view) *Given an integer $1 \leq i \leq k$, the* i-th incremental view *of $\mathcal{N}$* w.r.t. *$\mathcal{E}$ is the subgraph $\mathcal{N}_i$ of $\mathcal{N}$ induced by the set $S_i = \bigcup_{j=1}^{i} E_j$. The* i-th decremental view *is defined analogously, except that $S_i = E \setminus (\bigcup_{j=1}^{i} E_j)$.*

Definition 47 $i\%$ percentage incremental (decremental) view *Let p be the largest integer such that the cardinality of $S_{i\%} = \cup_{j=1}^{p} E_j$ is at most $i\%$ of the edges in E. The $i\%$ incremental view of $\mathcal{N}$ is defined as the subgraph $\mathcal{N}_{i\%}$ of $\mathcal{N}$ induced by the set $S_{i\%}$. The* $i\%$ percentage decremental view *is defined analogously, except that $S_{i\%} = E \setminus (\bigcup_{j=1}^{p} E)$.*

The difference between views in Definitions 46 and 47, is that the former is focused on the partitions built on the rank (e.g., they may be associated to fixed value intervals of the considered measure), whereas the latter uses *at most* a specified number of edges (nodes) in $\mathcal{N}$, therefore it is focused on the "size" of the view one wants to generate.
As for nodes, the definitions of rank and views are analogous to the ones given above for edges and therefore omitted. It is worth noting that, in terms of views, the one corresponding to a node rank $\mathcal{V}$ reduces to an edge rank $\mathcal{E}^*$, that we refer to as *equivalent edge rank*. Indeed, informally, given $\mathcal{V}$, one can construct $\mathcal{E}^*$ by progressively growing the sets of edges in $\mathcal{E}^*$ as they are inserted/removed in the view corresponding to $\mathcal{V}$. Formal details are omitted for brevity.
Suppose that a rank view induced by a specific topological measure is given. The study presented here aims to understand how much it is representative not only of the biological knowledge directly encoded by the network topology, but also of novel, hidden, functional information. As usual in both supervised and unsupervised classification contexts [19, 25, 36, 37, 44, 77, 80, 82], the "performance" of the rank view in discovering hidden functional knowledge may be evaluated by using *external* criteria. Here, an external criterion relies on the existence of a ranking associated to some gold standard, obtained via information not dependent on the topology of the input network. The rank induced by a topological function can thus be compared against the gold standard rank. Once that quantification is available, it is also important to assess how statistically significant it is. To this end, one can resort to a Montecarlo Hypothesis Test (see [43] for analogous applications of this test in the biological domain), where the Null Hypothesis H_0 is that the mentioned quantification is due to chance. That is, its value is no better than the one obtained by a random ranking.

Global comparison. Assume that each edge is numbered with an integer in $[1, |E|]$. Let $\mathcal{E}_w = (E_{1,w}, E_{2,w} \cdots, E_{k,w})$ and
$\mathcal{E}_g = (E_{1,g}, E_{2,g} \cdots, E_{p,g})$ be the rankings coming out of w and g, respectively. If the sequence of those ranks were a permutation of the edge numbers, then we could easily compare them via standard methods such as Kendall rank index [34]. Unfortunately, since there may be ties, i.e.,

more than one edge may be associated to the same integer representing its ranking, we cannot use the mentioned index directly, as well as many others (see discussion in [34]). A rank with ties is referred to usually as partial. Among the many possibilities, we have chosen to use K_{haus}, a distance function on partial ranks defined by [34] and that belongs to a class of distances specifically designed for partial ranks. Given two partial ranks, K_{haus} counts the number of inversions in the ranks, excluding ties. It is normalized so that it has value in $[0, 1]$, where zero indicated identity. In order to assess how close $\mathcal{E}_w$ is to $\mathcal{E}_g$, we use K_{haus}. The lower its value, the better the performance of w with respect to the gold standard g. Consider $\mathcal{E}_w = (E_{1,w}, E_{2,w} \cdots, E_{k,w})$, $\mathcal{E}_g = (E_{1,g}, E_{2,g} \cdots, E_{p,g})$ and K_{haus}. We use two Null models. The first is referred to as *total random* and denoted by TR, in which a random permutation of the edges of the network is generated. The second, referred to as *equal classes* and denoted by EC, in which each class of $\mathcal{E}_w$ is assigned the same number of edges it has, but this time chosen randomly, without replacement, from the set of edges of the network. We perform a MonteCarlo simulation consisting of 100 iterations for both models. In each iteration and for each model, we compute K_{haus} between $\mathcal{E}_g$ and the randomly generated permutation. Then we set the significance level at 1% as a measure of relevance.

Chapter 6

Prediction of lncRNA-disease Associations

Abstract

This Chapter presents a novel approach for the prediction of lncRNA-disease Associations. The main idea here is to discover hidden relationships between lncRNAs and diseases through the exploration of their interactions with intermediate molecules (e.g., miRNAs) in the tripartite graph, based on the consideration that while a few of lncRNA-disease Associations are still known, plenty of interactions between lncRNAs and other molecules, as well as associations of the latters with diseases, are available.

6.1 Introduction

The main goal here is to provide a computational method able to predict novel LDA candidate for experimental validation in laboratory, given further external information on both molecular interactions and genotype-phenotype associations, but without relying on the knowledge of existing validated LDA. The main idea presented in [17] is to discover hidden relationships between lncRNAs and diseases through the

exploration of their interactions with intermediate molecules (e.g., miRNAs) in a tripartite graph in Figure 6.1, where the three sets of vertices represent lncRNAs, miRNAs, and diseases, respectively, and vertices are linked according to lncRNA-miRNA interactions (LMI) and miRNA disease associations (MDA).

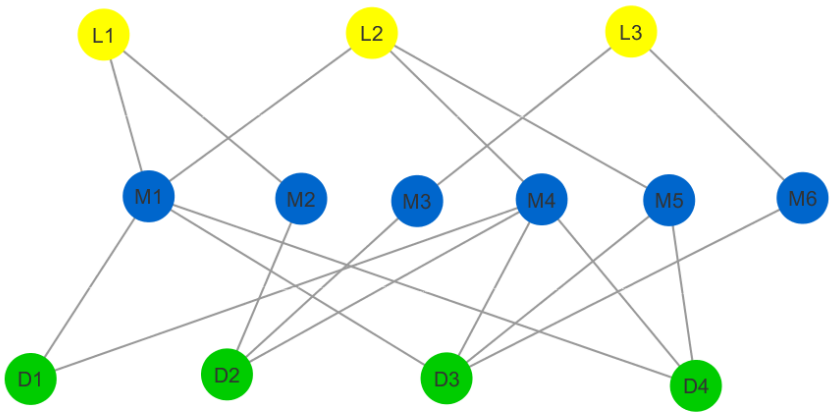


Figure 6.1: Large amounts of lncRNA-miRNA interactions and miRNA-disease associations have been collected in public databases.

6.2 A novel approach to predict new lncRNA-disease associations

The idea of not including any information on existing LDAs in the approach is based on the consideration that only a restricted number of validated LDAs is yet available, therefore a not exhaustive variability of real associations would be possible, affecting this way the correctness of the produced predictions. On the other hand, larger amounts

of interactions between lncRNAs and other molecules (e.g., miRNAs, genes, proteins), as well as associations between those molecules and diseases are known, and we have focused our approach on the use of such datasets. In particular, we have considered only miRNAs as intermediate molecules, however the approach is general enough to allow the inclusion of other molecules in the future.

Problem Statement. Let $\mathcal{L} = \{l_1, l_2, \ldots, l_h\}$ be a set of lncRNAs and $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$ be a set of diseases. The goal is to return a set $\mathcal{P} = \{(l_x, d_y)\}$ of predicted LDAs. Let T_{LMD} be a tripartite graph defined on the three sets of disjoint vertices L, M and D, which can also be represented as $T_{LMD} = \langle (l, m), (m, d) \rangle$, where (l, m) are edges between vertices in L and M, (m, d) are edges between vertices in M and D, respectively. In the proposed approach, L is a set of lncRNAs, M is a set of miRNAs and D is a set of diseases. In such a context, edges of the type (l, m) represent molecular interactions between lncRNAs and miRNAs, experimentally validated in laboratory; edges of the type (m, d) correspond to known associations between miRNAs and diseases, according to the existing literature. In both cases, we refer to interactions and associations suitably annotated and stored in public databases. A commonly recognized assumption is that lncRNAs with similar behaviour in terms of their molecular interactions with other molecules, may also reflect this similarity in their involvement in the occurrence and progress of disorders and diseases [65]. This is even more effective if the correlation with diseases is "mediated" exactly by the molecules they interact with, i.e., miRNAs.
Based on the assumption that similar lncRNAs interact with similar diseases [65], we aim to identify novel LDA by analyzing the behaviour of *neighbor lncRNAs*, in terms of their intermediate relationships with miRNAs. A score is assigned to each LDA (l, d) by considering both their respective interactions with common miRNAs, and the interactions with miRNAs shared by the considered disease d and other lncRNAs in the neighborhood of l. We define the *prediction-score* $S(l_i, d_j)$ for the LDA

(l_i, d_j) as:

$$S(l_i, d_j) = \alpha \cdot \frac{|M_{l_i} \bigcap M_{d_j}|}{|M_{l_i} \bigcup M_{d_j}|} + (1 - \alpha) \cdot \frac{|\bigcup_x (M_{l_x} \bigcap M_{d_j})|}{|\bigcup_x M_{l_x} \bigcup M_{d_j}|} \tag{6.1}$$

where M_{l_i} is the set of miRNA associated to l_i, M_{d_j} is the set of miRNA associated to d_j, α is a real value in [0,1] used to balance the two terms of the formula, M_{l_x} are all miRNA of those lncRNAs sharing at least one miRNA with l_i.

Example 2

$$S(l_i, d_j) = 0.5 \cdot \frac{1}{2} + (1 - 0.5) \cdot \frac{(1 \times 2) + (1 \times 2)}{(2 \times 3)(3 \times 3)} = 0.25 + 0.13 = 0.38$$

Prediction score with Hypothesis Test. Given a set $\mathcal{A}$ of LDAs scored according to the prediction-score computed as described above, it is necessary to select the only associations which are statistically significant, for producing the output predictions. To establish the statistical significance of the considered LDAs, we perform a Hypothesis Test via a Montecarlo simulation [43, 49]. The Null Hypothesis is that lncRNAs and diseases have been associated by chance. It is important to focus on the importance that the intermediate miRNAs have in the prediction-score computation and, more in general, in the measure of how similar is the behaviour of different lncRNAs with respect to the occurrence of diseases. In particular, in the adopted model interactions with miRNAs are the key factors determining the association between a lncRNA and a disease. Let then $(\hat{l}, \hat{m})$ be the pairs in $\mathcal{A}$ and shuffle them for 100 times by producing 100 new sets of pairs $\mathcal{A}_i$. The meaning is to interchange the associations between lncRNAs and miRNAs, still maintaining the same number of interactions. The test to reject the Null Hypothesis consists on comparing the prediction-score $S(l, d)$ of an association (l, d) in $\mathcal{A}$ with the maximum value of prediction-score $\hat{S}(l, d)$ obtained by the same pair in the 100 $\mathcal{A}_i$. The Null Hypothesis is rejected if $S(l, d) > \hat{S}(l, d)$.

Neighborhood based approach. The method consists of the following steps: let T_{LMD} a tripartite graph, we define a prediction score LDA,

based on neighborhood analysis with Apache Spark. The first step of the system consists of parsing the input data from the considered databases, extracting all the different miRNAs, lncRNAs, and diseases. In the second step, following the graph generation, there is the score computation step over all the associations generated in the previous step. We performed a statistical test to establish significant predictions and then we tested several methods and chose an appropriate test based on recent experimental literature : *False Discovery Rate* (FDR). The predicted LDAs are ranked according to their corrected score. Each verified LDA is left out in turn as a test sample; when the rank of this test sample exceeds a given threshold, the model provides a successful prediction. At the varying of the threshold, we compute: *true positive rate* (TPR) that it represents the *sensitivity*: the test samples whose ranking is higher than the given threshold and *false positive rate* (FPR), that it represents the *specificity*: the test samples that are below the threshold. The last step is the result analysis, performed through ROC analysis, which will be analysed in chapter (7). Apache Spark includes MLLib, a library with a set of functionalities to calculate the ROC metrics. In order to calculate the AUC, it is necessary to convert the data into a format (score, prediction), in which the score represents the value calculated and the prediction is 1 or 0, depending on the results (respectively True or False).

Matrix factorization. We introduce another method based on recommendation systems and matrix of preferences (see Table 6.1). There are three classes of entities: lncRNAs, miRNAs and diseases respectively. We used this method in two different modes:

- using prediction score of the association of previous method;
- without using a prediction score.

The idea of not including any information on existing LDAs is based on the consideration that only a restricted number of validated LDAs is yet available, therefore a not exhaustive variability of real associations would be possible, affecting this way the correctness of the produced predictions. On the other hand, larger amounts of interactions between

lncRNAs and other molecules (e.g., miRNAs, genes, proteins), as well as associations between those molecules and diseases are known.

The standard approach to matrix factorization based collaborative filtering treats the entries in the lncRNA-disease matrix as explicit common association with miRNA calculated in the first moment with the measure of centrality score. Collaborative filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a lncRNA-disease association matrix. *spark.ml* currently supports model-based collaborative filtering, in which lncRNA and disease are described by a small set of latent factors that can be used to predict missing entries. spark.ml uses the *alternating least squares (ALS)* algorithm to learn these latent factors.
This approach generalized by the algorithm summarized in the following steps:

1 Assign a score to all lncRNA-disease associations with respect to miRNA in common between lncRNA and disease;

2 Select α association that have the highest similarity commonly called the neighborhood;

3 Compute a prediction from a weighted combination of the selected neighbors' ratings.

In step (1) we calculate the score between lncRNA and disease using the measure described in the Formula 6.1, in the paragraph 6.2.

Example 3 We assume that the matrix is represented by 0 and 1, depending on whether lncRNA have miRNA in common with the disease. The sub-matrix factorization (see Table 6.1) includes lncRNAs and diseases respectively in the rows and in the columns. The cells (i, j) identified with an boolean value: 0 or 1. If a lncRNA interacts with a given miRNA and the same miRNA interacts with a given disease there will be 1, otherwise 0. The task is to predict the missing rating between lncRNA e disease.

	Abortion	Immuno Deficiency Syndrome	Adenoma	Acute Coronary Syndrome
HSA-MIR-155	0	0	0	1
HSA-MIR-15A	0	0	1	0
HSA-MIR-15B	0	0	0	0
HSA-MIR-16	0	0	1	0

Table 6.1: Example of the representation of a sub-matrix factorization.

Part III

Results

Chapter 7

Results

Abstract

This Chapter shows the results of the research presented in Chapters 4, 5 and 6 for the Social Networks and Biological Networks contexts. The chapter is divided into three sections, each corresponding to three proposed approaches in previous chapters. In particular, most results of this PhD Thesis have been already pubblished in conferences proceedings [14, 15, 17, 19], book chapters [18] and international journals [16].
The approaches described in the previous chapters have been implemented in Apache Spark (see Chapter 2.6.2).
The considered Apache Spark libraries are libraries are: spark-core and spark-graphX.

7.1 Optimization of Advertising Campaigns

Our experimental analysis has been devoted to understanding to what extent our approach is effective, in order to identify the k most convenient nodes in the input OSN to which distribute the advertisement. The main aim is to optimize two different aspects when identifying the best targets, that is, the fact that interests of considered users are related to the campaign contents, and the fact that they have "friends" on the OSN potentially interested in the distributed advertisements. The proposed approach has been implemented in Java under Apache Spark

1.6. To this respect, the use of Big Data Technologies allows the exploitation of the software tool also on very large OSNs.
We have considered the web-pages associated with four brands, listed in Table 7.1.

Brand	**Web-page**
AlfaRomeo	`www.alfaromeo.it`
Amarelli	`www.amarelli.it`
Carpisa	`www.carpisa.it`
KikoCosmetic	`www.kikocosmetics.com`

Table 7.1: The considered brands and their associated web-pages.

OSN graphs are available for example from Stanford website (`https://snap.stanford.edu/data/`).
We have considered the `twitter-2010` OSN from that repository, having $90,908$ vertices and $443,399$ edges. Unfortunately, the available OSNs consist only of the Graph topology, no information about user interests and profiles are publicly available.
As already introduced in Chapter 4 Web scraping has been used here in order to collect and extract useful contents for user profiles characterization. In particular, we have avoided associating randomly the information obtained by web scraping to nodes in the considered OSN Graph, due to the fact that a random association would have altered the natural mechanism according to which users in the same neighbors have similar interests. In order to mimic such a mechanism, which is important for our approach (indeed the introduced measures aim at detecting neighbor nodes with similar interests), we have proceeded as follows. We have first randomly selected 20 seed nodes from the `twitter-2010` OSN and 20 web-pages focused on different topics (cooking, fashion, cars, etc.). Indeed, with a certain margin of simplification, we have assumed that a user profile may be obtained by scraping the contents of a web-page on a specific topic. Then, a visit in depth of the OSN has been performed starting from each of the seeds and stopping when the entire network was visited. For each new node to be visited, a new web-page has been visited as well, following the cross-page links on the

considered web-pages. We have considered the OSN constructed, and we have computed, for each of the four brands (see Table 7.1), the different values of affinity and utility (with $\alpha = 0,25; 0,5; 0,75$) for all nodes in the network. Then, we have ranked them in descending order, according to each of these measures. We have supposed that the number of target nodes is $k = 100$ and we have fixed to 0.6 the minimum value of affinity between user and brand profiles in order for a user to be considered a possible target. The obtained results have been compared with a random choice of the k nodes to which distribute the advertisement. For 100 different times, 100 nodes have been extracted from the set of vertices V and the affinity between their and brand profiles have been computed at each time. The obtained results for the different brands do not present significant differences, therefore we illustrate only those regarding the brand AlfaRomeo in Table 7.2.

Method	**# of Target Nodes**	**Directly Reached**	**From Neighborhoods**
Affinity	184	100	84
Utility ($\alpha = 0.25$)	152	64	88
Utility ($\alpha = 0.5$)	**192**	99	**93**
Utility ($\alpha = 0.75$)	181	**100**	81
Random	99	13	86

Table 7.2: Total number of nodes (second column) with affinity values larger than the chosen threshold identified by each method (first column), fraction of target nodes directly reached (third column) or instead detected from the neighborhoods (fourth column).

In particular, the considered method is specified in the first column of the table, and for the Random generation we have considered the average of obtained results.
For each method, the number of nodes presenting an affinity value larger than the chosen threshold when the first k nodes in the corresponding ranking is chosen is shown in the third column. It is interesting to observe that, with respect to the random choice, both Affinity and Utility

with a high value of α (0.75) improves by one order of magnitude. Indeed, in these two latter cases, all the considered nodes have affinity values above the threshold. This shows that the profile matching at the basis of our approach is effective in the selection of target users for an advertising campaign. However, the second aspect to take into consideration is related to the number of possible further interested users that can be reached by the advertisement, starting from those k. To this respect, the last column of Table 7.2 shows how many distinct nodes are in the neighborhoods of the first k ones (according to the ranking obtained for each method). The second column of the table shows the total number of nodes with affinity values larger than the threshold that can be reached starting from the first k, for each ranking. It is evident that, again, the worst performance is obtained by the Random method, whereas the best one by Utility with $\alpha = 0.5$ in this case. This confirms what is expected, that is, neighborhood analysis associated with profile matching is the most promising choice.

7.2 Inferring the biological relevance of network components

The approach described in Chapter 5 requires three ingredients: (a) gold standards (b) a measure of agreement between ranks and (c) the specification of H_0 for the statistical significance test. Those points are presented next, focusing only on the edge rank and incremental case, since the equivalent edge rank and decremental cases are analogous.

Networks analysed. Three types of networks are analysed as already described in Chapter 2 (see in Table 7.3): in the first category the Gene Disease Network (***GDN***) is a one-mode projection of the Diseasome bipartite network [46]. Another possible one-mode projection of the Diseasome is the Human Disease Network (***HDN***). In analogy with GDN, the Worm Gene Network (***WGN***) is obtained as a one-mode projection of a bipartite graph obtained for *C. elegans* in [50], by placing in one class 554 essential genes and on the other 94 phenotypic defects. We consider the second category: three different yeast PPI datasets are accounted

for. The first two PPI networks, namely ***YD1*** and ***YD2***, have been built by [112] by filtering two networks, one used by [40] and another containing yeast protein interactions generated by six individual experiments, to delete unreliable interactions. The third PPI network, ***Y2H***, is built upon interactions obtained by high-throughput yeast two-hybrid screening [107], where self-edges have been eliminated according to [3].

Network	Type of nodes	No. Nodes	No. Edges	Link density
GDN	Gene	903	6,760	0.017
HDN	Disease	516	1,118	0.009
WGN	Gene	554	137,918	0.897
$YD1$	Protein	990	4,687	0.010
$YD2$	Protein	1,443	6,993	0.007
$Y2H$	Protein	1,966	2,705	0.001

Table 7.3: Basic structural features of the considered Biological Networks.

For each of the considered Biological Networks, at least one gold standard has been defined, as follows.

Gold Standards G1 and G2. In exploring GDN, it seems natural to expect that one would like first to see edges corresponding to the most strongly correlated gene pairs. Among the many possible weight assignments, we use very simple and intuitive ones which are meant to encode a biological tie-strength proportional of the number of common (1) diseases implied by SNPs (G1) [46], and (2) GO terms (G2) [41] between two genes (with references to the biological process vocabulary only).

Gold Standard G3. For the HDN gold standard ranking, how many SNPs are common to a pair of diseases is considered, according to [46].

Gold Standard G4. For WGN, in analogy with GDN, a gold standard is considered such that the weight of each edge is the number of defects of phenotype that two genes have in common.

Gold Standards G5, G6 and G7. For PPI networks, gold standard rankings have been associated with the number of biological complexes two proteins participate together. To this aim, three reference sets of yeast complexes have been considered here, each specifically selected for the networks in analysis [112]: G5 for D1, G6 D2 and G7 for Y2H, respectively. G5 includes 81 complexes of sizes at least 5, created from MIPS [71]. G6 is made of 162 hand-curated complexes (size no less than 4 proteins) from MIPS [72]. Finally, G7 includes 975 known and curated complexes from `ftp://ftpmips.gsf.de/yeast/catalogues/complexcat`.

Network	Case	Type	F Rank Type	P	R	Fm	View
D1	ER	I	ECV static	0.55	0.69	0.61	50%
D1	ER	D	EB dynamic	**0.89**	**0.77**	**0.82**	10%
D1	ER	D	ECC3 dynamic	0.56	0.72	0.63	40%
D1	EER	D	NB static	0.68	0.33	0.44	10%
D1	EER	D	NB dynamic	0.61	0.27	0.37	30%
D2	ER	I	ECV static	0.62	0.59	0.60	40%
D2	ER	D	EB dynamic	**0.89**	**0.77**	**0.82**	10%
D2	ER	D	ECC3 dynamic	0.56	0.72	0.63	40%
D2	EER	D	NB static	0.68	0.33	0.44	10%
D2	EER	D	NB dynamic	0.61	0.27	0.37	30%
Y2H	ER	I	ECV static	**0.37**	0.18	0.24	30%
Y2H	ER	D	EB dynamic	0.31	**0.33**	**0.32**	30%
Y2H	EER	D	NB dynamic	0.36	0.18	0.24	10%

Table 7.4: Application to PPI networks clustering. The first column shows the considered network; the second one specifies if edge ranking (ER) or edge equivalent rank (EER) is considered; in the third column if incremental (I) or decremental (D) views are considered is reported; the topological measure for which the results are reported on that row is specified in the fourth column; the values of Precision (P), Recall (R) and Fmeasure (Fm) are shown in following three columns, while in the last one the view percentage at which the best performance is reached is reported.

Table 7.4 shows the results obtained by comparing the connected

subgraphs at a given view percentage (between 10% and 50%) against known biological processes, as explained in Chapter 5. The validation is performed through the Precision, Recall and F-measure indices, computed as in [82]. Each row in the table corresponds to the best performance obtained by the only topological measures which pass both the EC and the TR significance tests. Results marked in red indicate a higher value of the corresponding index, with respect to the best results performed by methods in [82]. In bold the best value performed for network and index is highlighted.

Gene Disease Network: gold standard G1. The histogram in Figure 7.1 graphically illustrates the most representative example of the results obtained for this network. In particular, the value of $1 - K_{haus}$ is shown on the vertical axis for the considered topological measures, when the gold standard G1 is considered. The histogram shows that the best compromise between biological relevance and statistical significance is represented by *GTOM2* and *TOM*, in both the static and the dynamic settings (with a slight improvement in the latter case). Also *KB3*, closely followed by *KB2*, has good performance, although both the associated rankings do not pass the EC test.
From Table 5.2 (in Chapters 5) it is evident that only *NCC* (static/dynamic) passes both EC and TR tests, however its performance is not high. *NB* dynamic reaches the best performance, although it passes only the TR test.

Gene Disease Network: gold standard G2. However, the performance of all considered measures is worse, on average, than in the case of G1. Therefore, it seems that the involvement of gene pairs in common biological processes is more difficult to be inferred from GDN, than their influence on common diseases. This is possibly due to the complexity of cell processes, and to the fact that genes whose mutations are involved in the occurrence or progress of the same diseases, may act on different (e.g., complementary) biological processes. The measure *NCC* (static/dynamic) is the only measure passing both significance tests, also for G2. However, in contrast with the case of the gold standard G2, this time the best performance is reached by *NCC* dynamic.

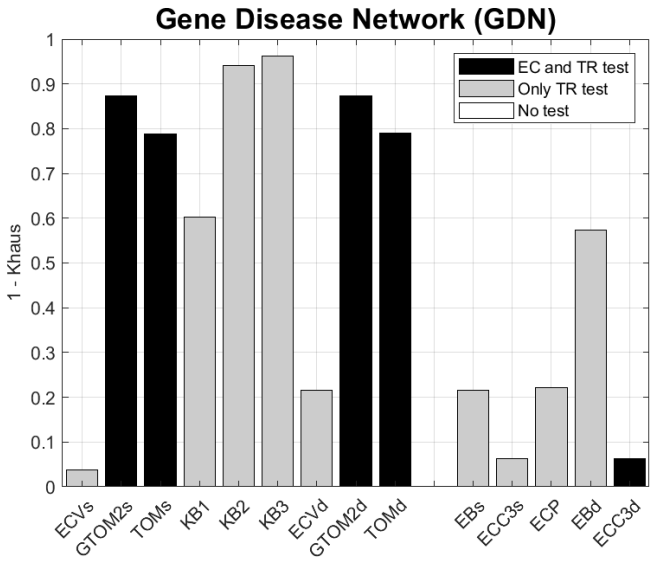


Figure 7.1: Performance and statistical significance for the rankings returned by topological measures for GDN w.r.t. the gold standard G1.

Human Disease Network: gold standard G3. Results (Tables 7.5 and 7.6) are similar to those obtained for GDN on G1, although here *GTOM2* dynamic reaches more markedly the best performance, among those measures passing the significance tests for edge rank. *NCC* dynamic is the only one passing both EC and TR tests, and it also outperforms all other measures, as in the case of GDN with G2. However, the performance of measures is on average slightly worse for HDN than for GDN, possibly due to the fact that the former is sparser than the latter (link density equal to 0.009 and 0.017, respectively).

Worm Gene Network: gold standard G4. The histogram for WGN (see Figure 7.2) shows that decremental measures remarkably outperform incremental ones, in terms of both performance and statistical significance. The best performing measure is *EB*, immediately followed by *ECC3* in the static case. This can be in part explained by the fact that the WGN is a very dense graph (link density equal to 0.897), as opposed to the GDN and HDN variants that are very sparse (0.017 and

View Type	F	Rank Type	K_haus	EC test	TR test
I	ECV	static	0.9616		✓
I	GTOM2	static	0.1267	✓	✓
I	TOM	static	0.2118	✓	✓
I	KB1	static	0.3983		✓
I	KB2	static	0.0590		✓
I	KB3	static	0.0380		✓
I	ECV	dynamic	0.7836		✓
I	GTOM2	dynamic	0.1260	✓	✓
I	TOM	dynamic	0.2108	✓	✓
D	EB	static	0.7841		✓
D	ECC3	static	0.9376		✓
D	ECP	static	0.7787		✓
D	EB	dynamic	0.4266		✓
D	ECC3	dynamic	0.9378	✓	✓

Table 7.5: **Global Comparison for the Gene Disease Network (GDN)** using Edge Ranks and with Golden Standard G1: e_{ij} exists if genes i, j share at least one common disease, and w_{ij} is equal to the number of common diseases according to [46].

0.009 respectively).

Table 7.7 shows that only *NCC* (static) pass both EC and TR tests, although the performance of all measures is worse than in edge rank, in analogy with results obtained for GDN and HDN.

Protein-Protein Interaction Networks: gold standards G5, G6, G7. For the PPI networks the best performance of topological measures is reached on the less sparse network, that is, $D1$ (see Figure 7.3), having link density equal to 0.01 against the 0.007 and 0.001 of $D2$ and $Y2H$, respectively. However, results on the three considered PPI networks are comparable (see Tables 7.8, 7.9, 7.10, 7.11, 7.12 and 7.13), in particular those of $D1$ and $D2$, where both incremental and decremental measures pass the statistical significance tests and the best performing measure is *KB3* for edge rank. As for $Y2H$, the best performance is

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.7739	✓	✓
I	EGC	static	0.4577		✓
I	NCC	dynamic	0.7463	✓	✓
I	EGC	dynamic	0.9329		✓
D	NB	static	0.2431		✓
D	SGC	static	0.9541		✓
D	KPC	static	0.9842		
D	NB	dynamic	0.2238		✓
D	SGC	dynamic	0.9298		✓
D	KPC	dynamic	0.9890		

Table 7.6: **Global Comparison for the Gene Disease Network** (GDN) using Edge Equivalent Ranks and with Golden Standard G1: e_{ij} exists if genes i, j share at least one common disease, and w_{ij} is equal to the number of common diseases according to [46].

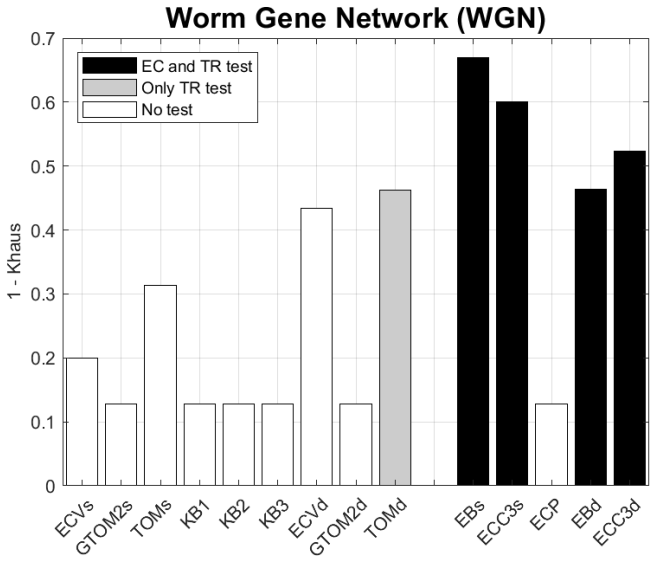


Figure 7.2: Performance and statistical significance for the rankings returned by topological measures for WGN.

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.5296	✓	✓
I	EGC	static	0.4587		✓
I	NCC	dynamic	0.5084	✓	✓
I	EGC	dynamic	0.6357		✓
D	NB	static	0.5997		✓
D	SGC	static	0.6440		✓
D	KPC	static	0.6630		
D	NB	dynamic	0.6110		✓
D	SGC	dynamic	0.6240		✓
D	KPC	dynamic	0.6680		

Table 7.7: **Global Comparison for the Gene Disease Network** (GDN) using Edge Equivalent Ranks and with Golden Standard G2: e_{ij} exists if genes i, j share at least one common disease, and w_{ij} is equal to the total number of shared GO terms.

reached by ECV, although only the decremental measures passes both EC and TR tests. In analogy with all other analyzed networks, edge equivalent rank performs worse than edge rank. EGC is the only incremental measure returning statistically significant results, together with NB and other decremental measures. However, NCC dynamic reaches very good performance on $Y2H$, although it passes only the TR test.

Topological views application. The measures best performing in the case of D1 and gold standard G5 have been considered, and the intersection between the complexes intercepted by edges involved in the topological and G5 ranks is computed at different view percentages (see Table 7.14). It is evident from these results that, even if the agreement between edges involved at the same percentage view is not always large, the agreement in terms of captured external knowledge (i.e., complexes) is in some cases highly pronounced. To this respect, KB3 confirms its best performance, being able to capture the 71% of complexes involved in the gold standard already at the 15% view, and although only the 9% of edges are in common between the two ranks at that view. At

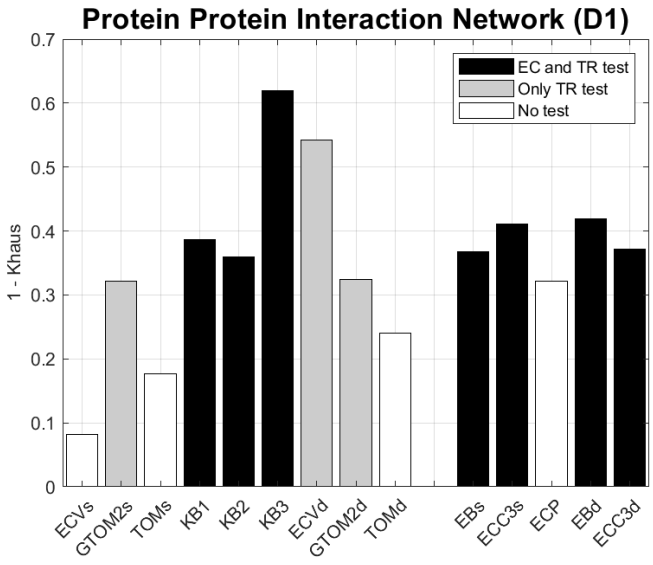


Figure 7.3: Performance and statistical significance for the rankings returned by topological measures for the PPI network D1.

the same view also KB2 reaches good performance (the 50% of common complexes with the 5% of common edges), and all three measures based on dispersion perform very well at the 30% and 60% views. Table 7.15 shows the best performing measures for the three considered organisms (human, worm and yeast), distinguished by those based on clustering coefficient (CC), neighborhoods (N), modularity (M) and dispersion (D). Results show that a distinct handful of best performing measures can be identified for each of the considered organisms, independently from the reference gold standard. Moreover, it seems that the proposed paradigm works better on denser networks, possibly due to the fact that the encoded information is larger than for sparse networks.

View Type	F	Rank Type	K_haus	EC test	TR test
I	ECV	static	0.9213		✓
I	GTOM2	static	0.2631	✓	✓
I	TOM	static	0.3944	✓	✓
I	KB1	static	0.5126		✓
I	KB2	static	0.1579		✓
I	KB3	static	0.1091		✓
I	ECV	dynamic	0.5876		✓
I	GTOM2	dynamic	0.2586	✓	✓
I	TOM	dynamic	0.3878	✓	✓
D	EB	static	0.9043		✓
D	ECC3	static	0.8793		✓
D	ECP	static	0.9023		✓
D	EB	dynamic	0.7337		✓
D	ECC3	dynamic	0.8880		✓

Table 7.8: **Global Comparison for the Human Disease Network** (HDN) using Edge Ranks: e_{ij} exists if diseases i, j share at least one common gene mutated, and w_{ij} is equal to the number of common genes according to [46].

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.5584	✓	✓
I	EGC	static	0.5730		✓
I	NCC	dynamic	0.4324	✓	✓
I	EGC	dynamic	0.6738		✓
D	NB	static	0.5537		✓
D	SGC	static	0.7423		
D	KPC	static	0.7741		
D	NB	dynamic	0.5570		✓
D	SGC	dynamic	0.6646		✓
D	KPC	dynamic	0.8388		

Table 7.9: **Global Comparison for the Human Disease Network** (HDN) using Edge Equivalent Ranks: e_{ij} exists if diseases i, j share at least one common mutated, and w_{ij} is equal to the number of common genes according to [46].

View Type	F	Rank Type	K_haus	EC test	TR test
I	ECV	static	0.7998		
I	GTOM2	static	0.8721		
I	TOM	static	0.6864		
I	KB1	static	0.8721		
I	KB2	static	0.8721		
I	KB3	static	0.8721		
I	ECV	dynamic	0.5657		
I	GTOM2	dynamic	0.8721		
I	TOM	dynamic	0.5379		✓
D	EB	static	0.3301	✓	✓
D	ECC3	static	0.3990	✓	✓
D	ECP	static	0.8721		
D	EB	dynamic	0.5363	✓	✓
D	ECC3	dynamic	0.4771	✓	✓

Table 7.10: **Global Comparison for the Worm Gene Network** (WGN) using Edge Ranks: e_{ij} exists if genes i, j share at least one common observed phenotype following gene knockout, and w_{ij} is equal to the number of common phenotypes according to [50].

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.4862	✓	✓
I	EGC	static	0.7540		
I	NCC	dynamic	0.5383		✓
I	EGC	dynamic	0.7508		
D	NB	static	0.6746		
D	SGC	static	0.7600		
D	KPC	static	0.7459		
D	NB	dynamic	0.6118		
D	SGC	dynamic	0.7524		
D	KPC	dynamic	0.7184		

Table 7.11: **Global Comparison for the Worm Gene Network** (WGN) using Equivalent Edge Ranks: e_{ij} exists if genes i, j share at least one common observed phenotype following gene knockout, and w_{ij} is equal to the number of common phenotypes according to [50].

View Type	F	Rank Type	K_haus	EC test	TR test
I	ECV	static	0.9181		
I	GTOM2	static	0.6778		✓
I	TOM	static	0.8230		
I	KB1	static	0.6139	✓	✓
I	KB2	static	0.4396	✓	✓
I	KB3	static	0.3800	✓	✓
I	ECV	dynamic	0.4583		✓
I	GTOM2	dynamic	0.6760		✓
I	TOM	dynamic	0.7600		
D	EB	static	0.6329	✓	✓
D	ECC3	static	0.5894	✓	✓
D	ECP	static	0.6780		
D	EB	dynamic	0.5814	✓	✓
D	ECC3	dynamic	0.6283	✓	✓

Table 7.12: **Global Comparison for the PPIN D1** using Edge Ranks. Edge e_{ij} exists if proteins i, j interacts physically according to (cite ref for network D1), and edge weight w_{ij} is equal to the number of protein complexes i, j have in common.

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.8950		
I	EGC	static	0.5640	✓	✓
I	NCC	dynamic	0.8369		
I	EGC	dynamic	0.6406		✓
D	NB	static	0.6167	✓	✓
D	SGC	static	0.6860	✓	✓
D	KPC	static	0.7041	✓	✓
D	NB	dynamic	0.5160	✓	✓
D	SGC	dynamic	0.6420		✓
D	KPC	dynamic	0.7173	✓	✓

Table 7.13: **Global Comparison for the PPIN D1** using Equivalent Edge Ranks. Edge e_{ij} exists if proteins i, j interacts physically according to (cite ref for network D1), and edge weight w_{ij} is equal to the number of protein complexes i, j have in common.

View Type	F	Rank Type	K_haus	EC test	TR test
I	ECV	static	0.2707		✓
I	GTOM2	static	0.7537		✓
I	TOM	static	0.7500		✓
I	KB1	static	0.1980		✓
I	KB2	static	0.2003		✓
I	KB3	static	0.2002		✓
I	ECV	dynamic	0.1972		✓
I	GTOM2	dynamic	0.7557		✓
I	TOM	dynamic	0.7606		✓
D	EB	static	0.8090	✓	✓
D	ECC3	static	0.7079	✓	✓
D	ECP	static	0.6780		
D	EB	dynamic	0.8151	✓	✓
D	ECC3	dynamic	0.6950	✓	✓

Table 7.14: **Global Comparison for the PPIN Y2H** using Edge Ranks. Edge e_{ij} exists if proteins i, j interacts physically according to (cite ref for network Y2H), and edge weight w_{ij} is equal to the number of protein complexes i, j have in common.

Organism	**CC**	**N**	**M**	**D**
H. sapiens	ECC3	GTOM2	–	–
		TOM	–	–
C. elegans	ECC3	–	EB	–
S. cerevisiae	ECC3	–	EB	KB1,KB2,KB3

Table 7.15: Best performing measures for edge rank. This table shows the best performing measures for the three considered organisms (human, worm and yeast), distinguished by those based on clustering coefficient (CC), neighborhoods (N), modularity (M) and dispersion (D) (see details in [16]). Results show that a distinct handful of best performing measures can be identified for each of the considered organisms, independently from the reference gold standard. Moreover, it seems that the proposed paradigm works better on denser networks, possibly due to the fact that the encoded information is larger than for sparse networks.

7.3 Prediction of lncRNA-disease Associations

The approach described in Chapter 6 has been applied to the known experimentally verified lncRNA disease associations in the lncRNADisease database [23] according to *Leave-One-Out Cross-Validation* (LOOCV). In particular, each known disease lncRNA association is left out in turn as a test sample. How well this test sample was ranked relative to the candidate samples (all the disease lncRNA pairs without the evidence to confirm their association) with respect to the considered score is evaluated. When the rank of this test sample exceeds the given threshold, this model is considered in order to provide a successful prediction. When the thresholds are varied, true positive rate (TPR, sensitivity) and false positive rate (FPR, specificity) are obtained. Here, sensitivity refers to the percentage of the test samples whose ranking is higher than the given threshold. Specificity refers to the percentage of samples that are below the threshold. Receiver Operating Characteristics (ROC) curve can be drawn by plotting TPR versus FPR at different thresholds. Area under ROC curve (AUC) is further calculated to evaluate the performance of the tested methods. AUC = 1 indicates perfect performance and AUC = 0.5 indicates random performance. We have validated the proposed approach on experimental verified data downloaded from starBase and from HMDD, resulting in 114 lncRNAs, 762 miRNAs, 392 diseases. We have implemented the *p-value* based on a hypergeometric distribution for LDAs inference proposed by [24] and *ncPred* based on recommendation system proposed by [6] and compared our approach against it, with two different dataset *HMDD*. Table 7.16 shows the results with the first dataset *HMDD 2.0* (as shown in Figure 7.5: the proposed neighborhoods-based approach achieved an AUC equal to 0.82, whereas the p-value based approach scored AUC = 0.74, and the ncPred based approach scored AUC = 0.81, showing that the consideration of indirect relationships between lncRNAs and diseases through neighborhood analysis is more effective. As for data extracted from StarBase and HMDD, our approach has produced 7,941 statistically significant LDAs predictions. The results for the second dataset *HMDD 3.0* (as shown in Figure 7.6) presents the centrality method as the best. Results marked in red indicate a higher value of AUC. In

Method	HMDD v.2	HMDD v.3
Centrality	0.82	0.91
PValue	0.74	0.87
ncPred	0.81	0.88
C. Filtering (a)	0.95	0.96
C. Filtering (b)	0.91	0.95

Table 7.16: Table shows the value of AUC for three different methods and two different datasets: Centrality method, Pvalue method and ncPred Method.

the results of the recommendation-system application (see Table 7.16), we calculate the method named Collaborative filtering (a) that uses the matrix without the information on miRNA, this method achieved an AUC equal to 0.95 (with the first dataset), and 0.96 (with the second dataset). In the result of the method named Collaborative filtering (b) we use the information of miRNA and it achieves an AUC of 0.91 (with the first dataset), and 0.95 (with the second dataset). According to simulation results, collaborative filtering models (as shown in Figure 7.4) for lncRNA disease association prediction may be an excellent addition to biomedical research in the future.

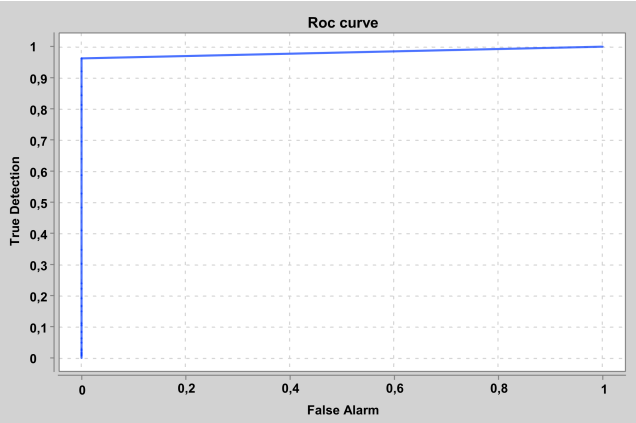
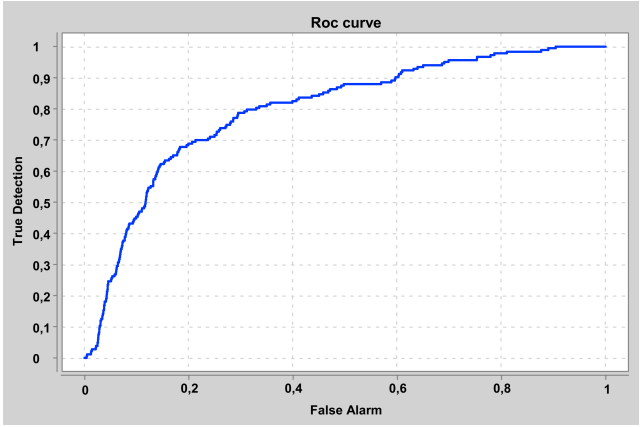
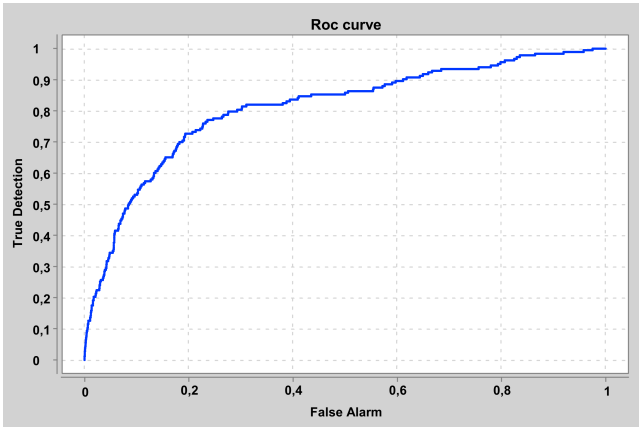


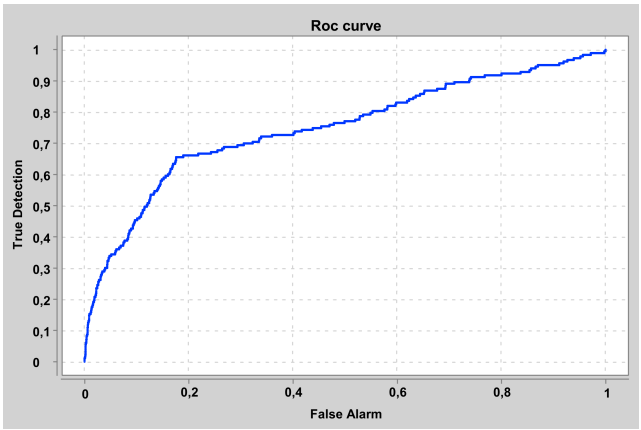
Figure 7.4: Representation of Roc Curve for Collaborative Filtering method.



(a) Centrality method

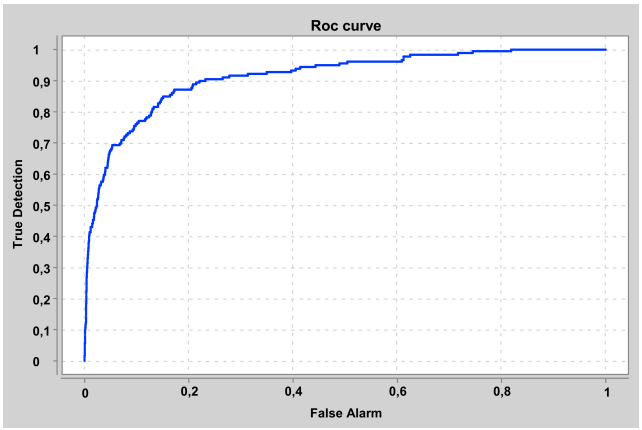


(b) ncPred method

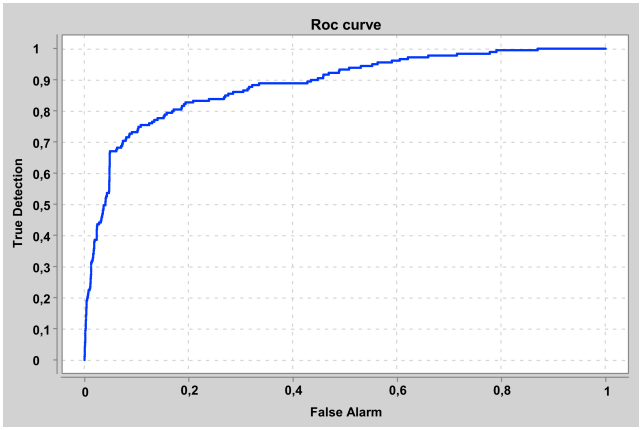


(c) Pvalue method

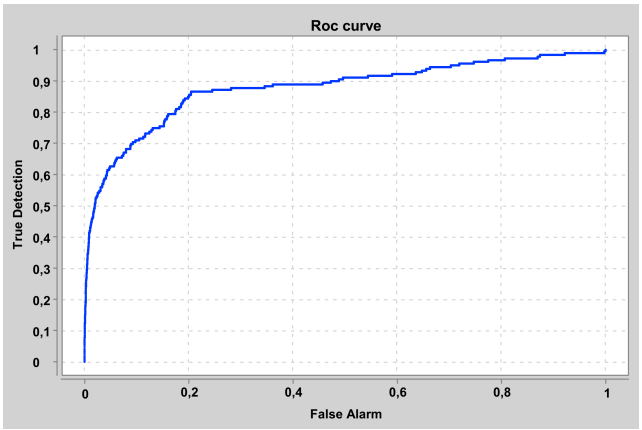
Figure 7.5: Representation of Roc Curve for Centrality method, Pvalue method and ncPred method (using first dataset HMDD).



(a) Centrality method



(b) ncPred method



(c) Pvalue method

Figure 7.6: Representation of Roc Curve for 3 methods (using second dataset HMDD).

Chapter 8

Concluding Remarks

Abstract

This Chapter draws the conclusions of the research presented in the previous Chapters and summarizes the method of optimization of advertising campaign, the approach based on the node/edge centrality measures in the biological context and the prediction of lncRNA-disease Associations.

Here the main contributions presented in this thesis are summarized. In this PhD project, knowledge extraction from large graphs was investigated, with reference to two main application contexts: Social Networks and Biological Networks. Three specific problems have been solved, and for each of them the main preliminary results that have been achieved are succinctly described below, as well as the possible future developments.

8.1 Problem 1: Optimization of Advertising Campaigns

The method in Chapter 4 discusses how the combination of information retrieval measures for profile matching and neighborhood exploration in OSNs may be successful in order to identify a set of target users for the distribution of advertisements. In particular, such users not only have

interests related to the contents of the advertisement, but may also potentially spread the received advertisements to other interested users in the OSN. This allows it to minimize costs for advertising campaigns, improve user experience in OSNs and avoid spreading useless information through OSNs. Results obtained by the measures introduced here on real datasets are promising. However, we are conscious that the proposed approach relies on a naive, although effective, technique for neighborhood exploration. An important problem in the context of OSNs analysis is the absence of publicly available datasets including not only network topology, but also structured information related to the network users, such as interests, general data, actions, etc.. It is worth to point out that the construction of such datasets via web-scraping starting from personal access points on the OSN presents several problems, among which data privacy constraints, the fact that the obtained networks would be mostly ego-networks [8, 58], and the difficulty in building networks that reflect the sizes of real OSNs, often very large [79, 102]. Therefore, providing suitable OSN public datasets which contain both topological and semantic data would be a valuable contribution for the scientific community. We plan to extend in this direction the procedure described here for the construction of big OSNs, and to provide a public repository containing such datasets.

8.2 Problem 2: Extracting functional knowledge from network topology

The method in Chapter 5 is based on a comparative analysis of a set of outstanding topological measures, finalized to show which are the best performing ones in ranking nodes/edges of biological networks, according to their corresponding functional relevance. Although only some of the existing biological network types have been accounted for, the methodology presented here for the comparison of topological measures applies also to other types of biological networks which have not been included in this analysis (e.g., molecular regulatory networks). The provided overview confirms and systematically summarizes previous results of the literature, still leading to novel conclusions. Moreover, it opens

the avenue to further investigations, such as the study of lossy compression in biological networks, based on the succinct global representations induced by the choice of the most relevant topological views, rather than the entire network. Also, the introduced paradigm seems to be successful in boosting important tasks in the context of network analysis, such as network clustering. This could be further explored also for other applications. Another interesting open issue is to study if there are specific network classes for which static and dynamic ranks induce always the same partitions, and other ones for which partitions are always different in the two cases. Moreover, it has been shown that ranks based on edge topological measures outperform those based on node ones, in the proposed comparative analysis. This could be further investigated to understand if there are other problems for which this behaviour changes, e.g., studying which proteins are more relevant in the occurrence and progress of human diseases.

8.3 Problem 3: Prediction of lncRNA-disease Associations

The approach in Chapter 6 for LDAs prediction is based on neighborhood analysis through a tripartite graph built upon lncRNA-miRNA interactions and miRNA-disease associations. An important fact is that the presented approach predicts potential LDAs without relying on the information of known disease-lncRNA associations. Although many previous studies for LDAs prediction use known available LDAs, the latter are still comparatively rare relative to the known lncRNA-miRNA interactions and miRNA-disease associations. Moreover, in the presented research we show that neighborhood analysis performs better than other techniques previously presented in the literature and not based on known LDAs, such as p-value based on HyperGeometric distribution. This is promising and results presented here are to be intended as a first step towards a more complex pipeline, where different types of molecular interactions and associations other than only lncRNA-miRNA will be taken into account (e.g., gene-lncRNA co-expression relationship, lncRNA-protein interactions, etc.). Approaches based on integrative networks

have indeed shown to reach better performance, therefore we plan to combine this strategy with the one proposed here on neighborhood analysis. Moreover, taking inspiration from previous studies on social media [54], we plan also to design suitable co-clustering [80, 81] and network clustering [82] based methods in order to improve the tripartite graph analysis.

Bibliography

[1] S. Agarwal et al. Blinkdb: queries with bounded errors and bounded response times on very large data. In Z. Hanzálek, H. Härtig, M. Castro, and M. F. Kaashoek, editors, *Eighth Eurosys Conference 2013, EuroSys '13, Prague, Czech Republic, April 14-17, 2013*, pages 29–42. ACM, 2013.

[2] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 114–122. ACM, 2011.

[3] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in Networks. *Nature*, 466:761–764, 2010.

[4] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974.

[5] T. Alahakoon, R. Tripathi, N. Kourtellis, et al. K-path Centrality: A New Centrality Measure in Social Networks. In *Proceedings of the 4th Workshop on Soc. Net. Syst.*, SNS '11, pages 1:1–1:6, New York, NY, USA, 2011. ACM.

[6] S. Alaimo, R. Giugno, and A. Pulvirenti. ncpred: ncRNA-disease association prediction through tripartite network-based inference. *Front Bioeng Biot*, 2:71, 2014.

[7] M. Armbrust et al. Spark SQL: relational data processing in spark. In T. K. Sellis, S. B. Davidson, and Z. G. Ives, editors, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 1383–1394. ACM, 2015.

[8] V. Arnaboldi, M. Conti, A. Passarella, and R. I.M. Dunbar. Online social networks and information diffusion: The role of ego networks. *Online Social Networks and Media*, 1:44–55, 2017.

[9] L. Backstrom and J. Kleinberg. Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing*, CSCW '14, pages 831–841, New York, NY, USA, 2014. ACM.

[10] A. Badkas, S. De Landtsheer, and T. Sauter. Topological Network measures for drug repositioning. *Briefings in Bioinformatics*, 22(4), 12 2020.

[11] L. C. Barrett et al. Generation and analysis of large synthetic social contact networks. In Ann Dunkin, Ricki G. Ingalls, Enver Yücesan, Manuel D. Rossetti, Ray Hill, and Björn Johansson, editors, *Proceedings of the 2009 Winter Simulation Conference, WSC 2009, Hilton Austin Hotel, Austin, TX, USA, December 13-16, 2009*, pages 1003–1014. IEEE, 2009.

[12] N. L. Biggs, E. K. Lloyd, and R.J. Wilson. *Graph Theory. 1736-1936*. Clarendon Press, Oxford, 1976.

[13] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2:113–120, 1972.

[14] M. Bonomo. Knowledge extraction from biological and social graphs. In Silvia Chiusano, Tania Cerquitelli, Robert Wrembel, Kjetil Nørvåg, Barbara Catania, Genoveva Vargas-Solar, and

Ester Zumpano, editors, *New Trends in Database and Information Systems - ADBIS 2022 Short Papers, Doctoral Consortium and Workshops: DOING, K-GALS, MADEISD, MegaData, SWODCH, Turin, Italy, September 5-8, 2022, Proceedings*, volume 1652 of *Communications in Computer and Information Science*, pages 648–656. Springer, 2022.

[15] M. Bonomo, G. Ciaccio, A. De Salve, and S. E. Rombo. Customer recommendation based on profile matching and customized campaigns in on-line social networks. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019*, pages 1155–1159. ACM, 2019.

[16] M. Bonomo, R. Giancarlo, D. Greco, and S. E. Rombo. Topological ranks reveal functional knowledge encoded in biological networks: a comparative analysis. *Briefings in Bioinformatics*, 23(3), 2022.

[17] M. Bonomo, A. La Placa, and S. E. Rombo. Prediction of lncrna-disease associations from tripartite graphs. page 205–210, Berlin, Heidelberg, 2020. Springer-Verlag.

[18] M. Bonomo, A. La Placa, and S. E. Rombo. *Prediction of Disease–lncRNA Associations via Machine Learning and Big Data Approaches. In K.P. Mayuri Mehta (eds), Knowledge Modelling and Big Data Analytics in Healthcare Advances and Applications.* CRC Press, 2021.

[19] M. Bonomo, A. La Placa, and S. E. Rombo. Identifying the k best targets for an advertisement campaign via online social networks. In Ana L. N. Fred and Joaquim Filipe, editors, *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2020, Volume 1: KDIR, Budapest, Hungary, November 2-4, 2020*, pages 193–201. SCITEPRESS, 2020.

[20] U. Brandes and T. Erlebach. *Network Analysis: Methodological Foundations (LNCS).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[21] F. Bruno, L. Palopoli, and S. E. Rombo. New trends in graph mining: Structural and node-colored network motifs. *International Journal of Knowledge Discovery in Bioinformatics*, 1(1):81–99, 2010.

[22] B. Chen, W. Fan, J. Liu, and F. Wu. Identifying protein complexes and functional modules - from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics*, 15(2):177–194, 2014.

[23] G. Chen et al. LncRNADisease: a database for long-non-coding rna-associated diseases. *Nucleic Acids Res*, 41:D983–D986, 2013.

[24] X. Chen. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Scientific Reports*, 5:13186, 2015.

[25] X. Chen and G. Yan. Novel human lncrna-disease association inference based on lncrna expression profiles. *Bioinformatics*, 29(20):2617–2624, 2013.

[26] L. Cheng, H. Shi, Z. Wang, Y. Hu, H. Yang, C. Zhou, J. Sun, and M. Zhou. ntnetlncsim: an integrative network analysis method to infer human lncrna functional similarity. *Oncotarget*, 7(30):47864–47874, 2016.

[27] C. Chu et al. Map-reduce for machine learning on multicore. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 281–288. MIT Press, 2006.

[28] A. J. Davis and D. Khazanchi. An exploratory investigation of the development of mutual knowledge in global virtual project teams. In W. Golden, T. Acton, K. Conboy, H. van der Heijden, and V. K. Tuunainen, editors, *16th European Conference on Information Systems, ECIS 2008, Galway, Ireland, 2008*, pages 1801–1813, 2008.

[29] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, jan 2008.

[30] C. Delorme. Eigenvalues of complete multipartite graphs. *Discrete Mathematics*, 312(17):2532–2535, 2012. Proceedings of the 8th French Combinatorial Conference.

[31] C. Demetrescu, U. Ferraro Petrillo, I. Finocchi, and G. F. Italiano. *Progetto di Algoritmi e Strutture Dati in Java.* 2007.

[32] L. Ehrlinger and W. Wolfram. Towards a definition of knowledge graphs. In M. Martin, M. Cuquet, and E. Folmer, editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[33] E. Estrada and J. A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Physical Review E*, 71:056103, May 2005.

[34] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17:134–160, 2003.

[35] N. Ferraro, L. Palopoli, S. Panni, and S. E. Rombo. Asymmetric comparison and querying of biological networks. *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 8(4):876–889, 2011.

[36] V. Fionda, L. Palopoli, S. Panni, and S. E. Rombo. Protein-protein interaction network querying by a "focus and zoom" approach. In Mourad Elloumi, Josef Küng, Michal Linial, Robert F. Murphy, Kristan Schneider, and Cristian Toma, editors, *Proceedings of Bioinformatics Research and Development(BIRD) 2008, Vienna,*

Austria, July 7-9, volume 13 of *Communications in Computer and Information Science*, pages 331–346. Springer, 2008.

[37] V. Fionda, L. Palopoli, S. Panni, and S. E. Rombo. A technique to search for functional similarities in protein-protein interaction networks. *International Journal of Data Mining and Bioinformatics*, 3(4):431–453, 2009.

[38] L. C. Freeman. Centrality in Social Networks conceptual clarification. *Social Networks*, 1(3):1978–1979, 2012.

[39] A. Furfaro, M. C. Groccia, and S. E. Rombo. 2D motif basis applied to the classification of digital images. *The Computer Journal*, 60(7):1096–1109, 2017.

[40] A. C. Gavin et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006.

[41] Gene-Ontology-Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 11 2014.

[42] C. Giallombardo, S. Morfea, and S. E. Rombo. An integrative framework for the construction of big functional networks. In H. J. Zheng, Zoraida Callejas, D. Griol, H. Wang, X. Hu, H. H. H. W. Schmidt, J. Baumbach, J. Dickerson, and L. Zhang, editors, *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Madrid, Spain, December 3-6, 2018*, pages 2088–2093. IEEE Computer Society, 2018.

[43] R. Giancarlo, S. E. Rombo, and F. Utro. Epigenomic k-mer dictionaries: shedding light on how sequence composition influences *in vivo* nucleosome positioning. *Bioinformatics*, 31(18):2939–2946, 2015.

[44] R. Giancarlo and F. Utro. Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis. *Theoretical Computer Science*, 428:58–79, 2012.

[45] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821–7826, 2002.

[46] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[47] K.Y. Goh, C. S. Heng, and Z. Lin. Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *Information Systems Research*, 24(1):88–107, 2013.

[48] J. E. Gonzalez. From graphs to tables the design of scalable systems for graph analytics. In C. W. Chung, A. Z. Broder, K. Shim, and T. Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 1149–1150. ACM, 2014.

[49] A. Gordon. Null models in cluster validation. *Gaul W. Pfeifer D. (eds.), From Data to Knowledge, Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg*, pages 32–44, 1996.

[50] R. A. Green, H.L. Kai, A. Audhya, et al. A High-Resolution C. elegans Essential Gene Network Based on Phenotypic Profiling of a Complex Tissue. *Cell*, 145:470–482, 2011.

[51] M. Hamouda. Understanding social media advertising effect on consumers' responses: An empirical investigation of tourism advertising on facebook. *Journal Enterprise Information Management*, 31(3):426–445, 2018.

[52] W. Huber, V. J. Carey, L. Long, S. Falcon, and R. Gentleman. Graphs in molecular biology. *BMC Bioinformatics*, 8(S-6), 2007.

[53] J. A. Iglesias, P. P. Angelov, A. Ledezma, and A. Sanchis. Creating evolving user behavior profiles automatically. *IEEE Transactions Knowledge and Data Engineering.*, 24(5):854–867, 2012.

[54] K. Ikematsu and T. Murata. A fast method for detecting communities from tripartite networks. *In Proceedings of Social Informatics*, pages 192–205, 2013.

[55] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[56] P. Kazienko. Social network analysis: Selected methods and applications. In J. Pokorný, Václav Snásel, and Karel Richta, editors, *Proceedings of the Dateso 2012 Annual International Workshop on DAtabases, TExts, Specifications and Objects, Zernov, Rovensko pod Troskami, Czech Republic, April 18, 2012*, volume 837 of *CEUR Workshop Proceedings*, page 151. CEUR-WS.org, 2012.

[57] D. Koschützki and F. Schreiber. Comparison of centralities for biological networks. In R. Giegerich and J. Stoye, editors, *Proceedings of the German Conference on Bioinformatics (GCB 2004), Bielefeld, Germany, October 4-6, 2004*, volume P-53 of *LNI*, pages 199–206. GI, 2004.

[58] Y. D. Kwon et al. Effects of ego networks and communities on self-disclosure in an online social network. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019*, pages 17–24, 2019.

[59] I. Lagwankar, A. N. Sankaranarayanan, and S. Kalambur. Impact of map-reduce framework on hadoop and spark MR application performance. In X. Wu, C. Jermaine, L. Xiong, X. Hu, O. Kotevska, S. Lu, W. Xu, S. Aluru, C. Zhai, E. Al-Masri, Z. Chen, and J. Saltz, editors, *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, pages 2763–2772. IEEE, 2020.

[60] X. Lei, J. Tian, L. Ge, and A. Zhang. The clustering model and algorithm of PPI network based on propagating mechanism of artificial bee colony. *Information Sciences*, 247:21–39, 2013.

[61] M. Li, Y. Lu, Z. Niu, and F. Wu. United complex centrality for identification of essential proteins from PPI networks. *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 14(2):370–380, 2017.

[62] S. Liang, X. Zhang, Z. Ren, and E. Kanoulas. Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 1764–1773, 2018.

[63] Q. Liao et al. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co- expression network. *Nucleic Acids Research*, 39:3864–3878, 2011.

[64] J. Lin, K. Sugiyama, M. Kan, and T. Chua. New and improved: Modeling versions to improve app recommendation. In *Proceedings of the 37th International ACM SIGIR Conference on Research; Development in Information Retrieval*, SIGIR '14, pages 647–656. ACM, 2014.

[65] C. Lu et al. Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics*, 34(19):3357–3364, 2018.

[66] P. V. Marsden and K. E. Campbell. Measuring Tie Stength. *Social Forces*, 63:482–501, 1984.

[67] X. Meng et al. Mllib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17:34:1–34:7, 2016.

[68] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Mixing Local and Global Information for Community Detection in Large Networks. *Journal of Computer System Science*, 80(1):72–87, February 2014.

[69] P. De Meo, E. Ferrara, G. Fiumara, and A. Ricciardello. A novel measure of edge centrality in social networks. *Knowledge-Based Systems*, 30:136–150, 2012.

[70] C. V. Mering et al. Comparative assessment of a large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

[71] H. W. Mewes et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, 30(1):31–34, 2002.

[72] H. W. Mewes et al. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Research*, 34(suppl1):D169–D172, 2006.

[73] R. Milo, S. Shen-Orr, S. Itzkovitz, et al. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[74] E. Yeger-Lotem others. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences*, 101(16):5934–5939, 2004.

[75] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[76] S. Panni and S. E. Rombo. Searching for repetitions in biological networks: methods, resources and tools. *Briefings in Bioinformatics*, 16(1):118–136, 2015.

[77] L. Parida, C. Pizzi, and S. E. Rombo. Irredundant tandem motifs. *Theoretical Computer Science*, 525:89–102, 2014.

[78] G. A. Pavlopoulos et al. Using graph theory to analyze biological networks. *BioData Mining*, 4:10, 2011.

[79] S. Peng, G. Wang, and D. Xie. Social influence analysis in social networking big data: Opportunities and challenges. *IEEE Network*, 31(1):11–17, 2017.

[80] C. Pizzuti and S. E. Rombo. *PINCoC*: A co-clustering based approach to analyze protein-protein interaction networks. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007, 8th International Conference, Birmingham, UK, December 16-19,*

2007, Proceedings, volume 4881 of *LNCS*, pages 821–830. Springer, 2007.

[81] C. Pizzuti and S. E. Rombo. A coclustering approach for mining large protein-protein interaction networks. *IEEE/ACM Transanctions on Computational Biology and Bioinformatics*, 9(3):717–730, 2012.

[82] C. Pizzuti and S. E. Rombo. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 2014.

[83] C. Pizzuti and S. E. Rombo. An evolutionary restricted neighborhood search clustering approach for PPI networks. *Neurocomputing*, 145:53–61, 2014.

[84] F. J. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In J. F. Elder IV, F. Fogelman-Soulié, P. A. Flach, and M. J. Zaki, editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 707–716. ACM, 2009.

[85] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 01 2007.

[86] G. Qin and L. Gao. Spectral clustering for detecting protein complexes in protein-protein interaction (PPI) networks. *Mathematical and Computer Modelling*, 52(11-12):2066–2074, 2010.

[87] F. Radicchi et al. Defining and identifying communities in networks. *Proceeding of the National Academy of Sci.*, 101:2658–2663, 2004.

[88] E. Ravasz et al. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297:1551–1555, 2002.

[89] F. Ricci, L. Rokach, and B. Shapira, editors. *Recommender Systems Handbook*. Springer, 2015.

[90] D. M. Romero, B. Uzzi, and J. M. Kleinberg. Social networks under stress: Specialized team roles and their communication structure. *ACM Transactions on the Web*, 13(1):6:1–6:24, 2019.

[91] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang. Big data analytics on apache spark. *International Journal of Data Science and Analytics*, 1(3-4):145–164, 2016.

[92] H. J. Schulz, M. John, A. Unger, and H. Schumann. Visual analysis of bipartite biological networks. In *Proceedings of the Eurographics Workshop on Visual Computing for Biomedicine, VCBM 2008, Delft, The Netherlands, 2008* [92], pages 135–142.

[93] H.A. Schwartz et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 2013.

[94] M. Van Steen. *Graph Theory and Complex Networks: An Introduction*. Maarten van Steen, 2010.

[95] J. M. Urquiza et al. Selecting negative samples for PPI prediction using hierarchical clustering methodology. *Journal of Applied Mathematics*, 2012:897289:1–897289:23, 2012.

[96] F. G. J. Wang, C. Domeniconi, and G. Yu. Matrix factorization-based data fusion for the prediction of lncrna-disease associations. *Bioinformatics*, 34:1529–1537, 2018.

[97] J. Wang, M. Li, J. Chen, and Y. Pan. A Fast Hierarchical Clustering Algorithm for Functional Modules Discovery in Protein Interaction Networks. *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 8(3):607–620, May 2011.

[98] M. N. Wang, Z. H. You, L. Wang, L. P. Li, and K. Zheng. LDGRNMF: lncrna-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing*, 424:236–245, 2021.

[99] S. Wasserman and K. Faust. *Social network analysis - methods and applications*, volume 8 of *Structural analysis in the social sciences*. Cambridge University Press, 2007.

[100] D. J. Watts. *Small worlds*. Princeton University Press, Princeton, 1999.

[101] T. White. *Hadoop - The Definitive Guide: Storage and Analysis at Internet Scale (4. ed., revised & updated)*. O'Reilly, 2015.

[102] Y. Wu et al. An incentive-based protection and recovery strategy for secure big data in social networks. *Information Sciences*, 508:79–91, 2020.

[103] H. Xie et al. Community-aware user profile enrichment in folksonomy. *Neural Networks*, 58:111–121, 2014.

[104] R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica. Graphx: a resilient distributed graph system on spark. In Peter A. Boncz and Thomas Neumann, editors, *First International Workshop on Graph Data Management Experiences and Systems, GRADES 2013, co-located with SIGMOD/PODS 2013, New York, NY, USA, June 24, 2013*, page 2. CWI/ACM, 2013.

[105] C. Xu, S. D. Ryan, V. R. Prybutok, and C. Wen. It is not for fun: An examination of social network site usage. *Information e Management*, 49(5):210–217, 2012.

[106] A. Yip and S. Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8:22, 2007.

[107] H. Yu et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome network. *Science*, 322(5898):104–110, 2008.

[108] J. Yu, Z. Xuan, X. Feng, Q. Zou, and L. Wang. A novel collaborative filtering model for lncrna-disease association prediction based on the naïve bayesian classifier. *BMC Bioinformatics*, 20(1):396:1–396:13, 2019.

[109] R. B. Zadeh et al. Matrix computations and optimization in apache spark. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 31–38. ACM, 2016.

[110] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Steven D. Gribble and Dina Katabi, editors, *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012, San Jose, CA, USA, April 25-27, 2012*, pages 15–28. USENIX Association, 2012.

[111] M. Zaharia et al. Discretized streams: fault-tolerant streaming computation at scale. In M. Kaminsky and M. Dahlin, editors, *ACM SIGOPS 24th Symposium on Operating Systems Principles, SOSP '13, Farmington, PA, USA, November 3-6, 2013*, pages 423–438. ACM, 2013.

[112] N. Zaki, J. Berengueres, and D. Efimov. Prorank: A Method for Detecting Protein Complexes. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*, GECCO'12, pages 209–216, 2012.