Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Semantic clustering and community detection in biological networks

by

Hassan Mahmoud Mohamed Ramadan Mohamed

Theses Series **DIBRIS-TH-2015-04**

DIBRIS, Università di Genova
Via Opera Pia, 13 16145 Genova, Italy http://www.dibris.unige.it/

Università degli Studi di Genova

Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Dottorato di Ricerca in Informatica

Ph.D. Thesis in Computer Science

Semantic clustering and community detection in biological networks

by

Hassan Mahmoud Mohamed Ramadan Mohamed

May, 2015

Dottorato di Ricerca in Informatica
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università degli Studi di Genova

DIBRIS, Univ. di Genova
Via Opera Pia, 13
I-16145 Genova, Italy
http://www.dibris.unige.it/

Ph.D. Thesis in Computer Science (S.S.D. INF/01)

Submitted by Hassan Mahmoud Mohamed Ramadan Mohamed
DIBRIS, Univ. di Genova, Italy
hassan.mahmoud@unige.it

Date of submission: February 2015

Title: Semantic clustering and community detection in biological networks

Advisor: Francesco Masulli
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova, Italy
francesco.masulli@unige.it

Ext. Reviewers:
Roberto Tagliaferri
Dipartimento di Informatica Campus di Fisciano
Università di Salerno, Italy
robtag@unisa.it

Pietro Lio'
Computer Laboratory
Cambridge University, Cambridge, UK
pl219@cam.ac.uk

Abstract

The recent advances of software systems and devices such as Facebook in social networks and Next Generation Sequencing in the biological domain are leading to an increasing accumulation of tremendous masses of data with complex interactions. Often, such data entail graph structures where interaction strengths based on specific degree (e.g., of friendship or of Protein-Protein Interactions *(PPIs)*) can be represented as ties linking network entities. However, vagueness and uncertainty can affect the relationships between linked data due to variations in strength of interaction and network evolution. The art of inferring the modular structure of networks is called *community detection* which refers to finding large subgraphs characterized by high connection densities within nodes belonging to it, and low interaction with entities residing in other communities. It is worth to note that a community can provide a significant way to identify functionally relevant or strictly related groups of nodes rather than analyzing data independently. Community discovery is a complex task since communities are usually hidden behind complicated relationships. Moreover, it is unclear how to extract coverage in real world networks. Several attempts were devoted in the recent literature to characterize overlapping communities, however none of them can efficiently always infer the hidden structure of the network.

The main contributions of this thesis are summarized as follows:

- A survey of state of the art of fuzzy, kernel, and modularity based methods used for identifying disjoint and overlapping communities such as spectral approaches. To this aim a classification of these methods showing their pros and cons is presented.
- A comparison of centrality measures used for characterizing hubs, bridges, and outlier nodes in the analyzed networks, and on how to use them in case of overlapping networks.
- The proposal of a Fuzzy Spectral Modularity *(FSM)* community detection approach and its possibilistic version *(PSM)*. The *FSM* employs graph based and spectral approaches to infer possible fuzzy communities. A comparison of *PSM* and *FSM* methods with state of the art methods is presented.
- The proposal of an approach based on an ensemble community detection method of *PSM* and semantic information for identifying biological communities that improves the quality of the discovered overlapping aggregations.

- The evaluation of the quality of *PSM* and the other related approaches in detecting communities in benchmark networks such as Zachary and Dolphin. To this aim a fuzzy version of Rand index, and a fuzzy modularity measure are employed.

- The application of the proposed approach for inferring significant communities in *PPIs* networks to homo sapiens *HIV1* network, and to yeast saccharomyces cerevisiae *PPIs*. The experimental results show that the inclusion of the semantic information for enriching and annotating community discovery allow us to identify many significant pathways consistent with the literature.

- The proposal of a correlation-based version of *FSM* for analyzing networks obtained by time series data and its application to the Italian financial stock network. This approach can be extended to community discovery in dynamical networks.

Publications

1. Mahmoud H., Masulli F., Rovetta S., Russo G.: Detecting overlapping protein communities in disease networks. Lecture Notes in Computer Science (LNCS), Computational Intelligence Methods for Bioinformatics and Biostatistics, (in press).

2. Mahmoud H., Masulli F., Marina R., Rovetta S., Abdulatif A.: Hubs and Communities Identification in Dynamical Financial Networks. Neural Nets and Surroundings Smart Innovation, Systems and Technologies, 24th Italian Workshop on Neural Networks, WIRN2014, Vietri sul Mare, Salerno, Italy (in press).

3. Mahmoud H., Masulli F., Rovetta S., Russo G.: Exploiting Quantitative and Semantic Information in Protein-Protein Interactions Networks Analysis. Computational Intelligence Methods for Bioinformatics and Biostatistics-11th International Meeting (CIBB 2014), Cambridge, UK, June 26-28, (2014).

4. Mahmoud H., Masulli F., Rovetta S., Russo G.: Finding Fuzzy Biological and Ecological Aggregations in Spectral Space. Poster presentation, Dagli Atomi al Cervello le Scienze di Base per la Comprensione delle Funzioni del Cervello, Università degli Studi di Milano, Italy, pp 79, Jan 27, (2014). http://www.cnism.it/web/it/events/dagli_atomi_al_cervello.

5. Mahmoud H., Masulli F., Rovetta S., Russo G.: Characterizing evolving protein communities. Poster presentation, Bioinformatics Italian Society Eleventh Annual Meeting (BITS 2014), Rome, Italy, pp 80, Feb 26-28, (2014). http://bits2014.uniroma2.it/.

6. Mahmoud H., Masulli F., Rovetta S., Russo G.: Identifying overlapping Interactome communities using spectral and semantic approaches. In Quarta Giornata Ligure di Bioinformatica (GLIB2014). Congress Centre CBA IRCCS AOU San Martino IST, Genoa, Italy, Dec 19, (2014).

7. Mahmoud H., Masulli F., Rovetta S., Russo G.: Community Detection in Protein-Protein Interaction Networks Using Spectral and Graph Approaches. Lecture Notes in Computer Science (LNCS), Computational Intelligence Methods for Bioinformatics and Biostatistics, Springer, ISBN/ISSN: 978-3-319-09041-2, pp 62–75, (2014), *doi* : 10.1007/978 – 3 – 319 – 09042 – 9_5, http://link.springer.com/chapter/10.1007/978-3-319-09042-9_5#page-1.

8. Rosasco R., Mahmoud H., Rovetta S., Masulli F.:A quality-driven ensemble approach to automatic model selection in clustering. Recent Advances of Neural Network Models and Applications, Proceedings of the 23rd Workshop of the Italian Neural Networks Society (SIREN), WIRN2013, May 23-25, Vietri sul Mare, Salerno, Italy, Springer, ISBN/ISSN:

978-3-319-04128-5, vol. 26, pp 53–61, *doi* : 10.1007/978 – 3 – 319 – 04129 – 2_6, http://link.springer.com/chapter/10.1007%2F978-3-319-04129-2_6.

9. Mahmoud H., Masulli F., Rovetta, S. Russo G.: Community detection in Protein-protein interaction networks, Abstract, The 9 th Conference of Italian Researchers in the World, Houston, Texas, USA, pp 62, Dec 14, (2013). http://houston.comites-it.org/conferenza2013.pdf.

10. Mahmoud H., Masulli F., Rovetta S.: Finding fuzzy biological and ecological aggregations in spectral space, Poster, Giornata Ligure di Bioinformatica 2013 (GLIB2013), Genova, Italy, Nov, (2013).

11. Rovetta S., Masulli F., Mahmoud H.: Neighbor-based similarities. Lecture Notes in Computer Science (LNCS), Tenth International Workshop on Fuzzy Logic and Applications (WILF 2013), Genoa, Italy, Springer, ISBN/ISSN: 978-3-319-03199-6, vol. 8256, pp 161–170, (2013), *doi* : 10.1007/978 – 3 – 319 – 03200 – 9_17, http://link.springer.com/chapter/10.1007/978-3-319-03200-9_17.

12. Bulotta S., Mahmoud H., Masulli F., Palummeri E., Rovetta S.: Fall Detection using an Ensemble of Learning Machines. Neural Nets and Surroundings Smart Innovation, Systems and Technologies, 22nd Italian Workshop on Neural Nets, WIRN 2012, May 17-19, Vietri sul Mare, Salerno, Italy, Springer Heidelberg New York Dordrecht London, ISBN/ISSN: 978-3-642-35466-3, vol. 19, pp 81–90, (2013), *doi* : 10.1007/978 – 3 – 642 – 35467 – 0_9, http://link.springer.com/chapter/10.1007%2F978-3-642-35467-0_9.

13. Mahmoud H., Masulli F., Rovetta S., Russo, G.:Community detection in Protein-Protein Interaction networks. Computational Intelligence Methods for Bioinformatics and Biostatistics-10th International Meeting (CIBB 2013), Nice, France, Jun 19-22, (2013).

14. Mahmoud H., Masulli F., Rovetta S.: Feature-Based Medical Image Registration Using a Fuzzy Clustering Segmentation Approach. Lecture Notes in Computer Science, Computational Intelligence Methods for Bioinformatics and Biostatistics, Springer Berlin Heidelberg, ISBN/ISSN: 978-3-642-38341-0, vol. 7845, pp 37–47, (2013), *doi* : 10.1007/978 – 3 – 642 – 38342 – 7_4, http://link.springer.com/chapter/10.1007/978-3-642-38342-7_4.

15. Mahmoud H., Masulli F., Rovetta S.: A Fuzzy Clustering Segmentation Approach for Feature-Based Medical Image Registration. Ninth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB2012), In Proceedings of CIBB, ISBN/ISSN: 978-88-906437-1-2, Houston, Texas, USA, Jul 12-14, (2012).

16. Mahmoud H., ElMessiry H., ElBahnasy K., Khalifa M.E.: Mutual Information threshold guided approach for image registration using resolution pyramid. Egyptian Computer Science Journal (ECS) Vol. 35(1), ISSN:1110-2586, Jan 2011 http://www.informatik.uni-trier.de/~ley/db/journals/ecs/ecs35.html.

17. Khalifa M.E., ElMessiry H., ElBahnasy K., Mahmoud H.:Medical Image Registration Using Mutual Information Similarity Measure , 13th International Conference on Biomedical Engineering (ICBME 2008), Singapore, Springer, ISBN/ISSN: 978-3-540-92840-9, vol. 23, PP 151–155, (2008), *doi* : 10.1007/978 – 3 – 540 – 92841 – 6_37, http://www.springerlink.com/content/x7r0l36n730u2107/.

18. Khalifa M.E., ElMessiry H., ElBahnasy K., Mahmoud H.: Medical Image Registration Techniques. Fourth International Conference on Intelligent Computing and Information Systems (ICICIS 2009), Cairo, Egypt, pp 473–479, March 19-22, (2009).

To my family

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.

(Clifford Stoll)

Acknowledgements

There are many people to thank, these few lines may not be enough. I would like to express my deep thanks for Prof. Francesco Masulli (DIBRIS-Genoa university), and Prof. Stefano Rovetta (*DIBRIS*-Genoa university) for their support and guidance all over my PhD. Their scientific discussions, facilities provided, and continual support to publish my work and participate in international and national events and schools improved my knowledge and understanding of the field. Moreover, during my PhD they involved me in scientific workshops organization (e.g., I was member of the scientific secretariat of the Tenth International Workshop on Fuzzy Logic and Applications (*WILF 2013*), and Clustering high dimensional data workshop (*CHDD2012*), and Genoa Bioinformatics workshop). In addition to peer reviewing (I am reviewer for the Journal of Biomedical informatics "*JBI–Elsiever*", *IEEE* Transactions on Human-Machine Systems Journal, Journal of Computer Science and information technology (ISSN: 2331-6071), Ajoy Journal of Soft Computing and Applications (*AJSCA*), International Conference on Artificial Neural Networks "*ICANN2012*", Image Analysis and Processing *ICIAP2013*, and member of the editorial committee of International conference of bioinformatics and genetics (*ICBG*) and the American society of Science and Engineering). They introduced me to different scientific groups (I become a member of of *INdAM*: Italian National Institute of Advanced Mathematics, and *GNCS*: National Group of Scientific Computing).

I would like to thank Prof. Giuseppe Russo (Temple University, Sbarro Institute of Cancer Research and Molecular Medicine, Philadephia, USA) for our discussions on biological issues that resulted in different joint publications, and Prof. Marina Resta (Economy Department-Genoa University) for collaborating and publishing together.

I thank Prof. Roberto Tagliaferri (Salerno University), and Prof. Pietro Lio' (Cambridge University) for reviewing this thesis, and I thank the members of the Bioinformatics Italian Society (*BITS*) (among them Prof. Paolo Romano, Prof. Annalisa Barla, and Prof. Angelo Facchiano) for the interesting events they organized in Proteomics and Next Generation Sequencing, and for the discussions we had that opened new ideas in my research.

I thank all the Professors of DIBRIS for organizing many interesting scientific seminars (I attended more than 60 seminars mainly in machine learning and bioinformatics and 14 PhD schools during my PhD), and I thank my friends and colleagues in *DIBRIS* (among them Andrea La Camera, Alessandro Solimando, Anna, and Federico Iuricich) for having a friendly and amazing working environment. Finally, I dedicate this work to my father, my mother, my brother (Dr. Mohamed), and my sisters (Dr. Fatma, and Hasnaa).

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the recent years the continuous advances and spread of software systems and devices such as Facebook in social networks and Next Generation Sequencing [139] in biological domain lead to an exponential process of continuous accumulation of tremendous amounts of data characterized by complex interactions.

Such data entail a graph structure where interaction strength based on specific function (e.g., friendship or protein interaction [100]) could be represented by ties linking network entities. In addition, vagueness and uncertainty affect the relationships between linked data due to variations in strength of interaction and network evolution (see Fig.2.1).

1.1 Community detection

The art of inferring the modular structures in networks is called *communities detection*, which refers to finding large subgraphs with high internal connection densities within nodes belonging to it, and sparse interaction between entities residing in other subgraphs or communities (see, e.g., Fig.1.1). It is worth to note that a community can provide a scalable way to identify functionally important or closely related groups of nodes rather than analyzing data independently.

1.2 Challenges of community detection

Usually community discovery is a complex task since communities are hidden behind complicated relationships. Moreover, it is unclear how to extract coverage in real world networks. Several attempts were devoted in the recent literature to characterize overlapping communities [148]. Unfortunately none of them is always able to detect the ground truth (when known) [118,

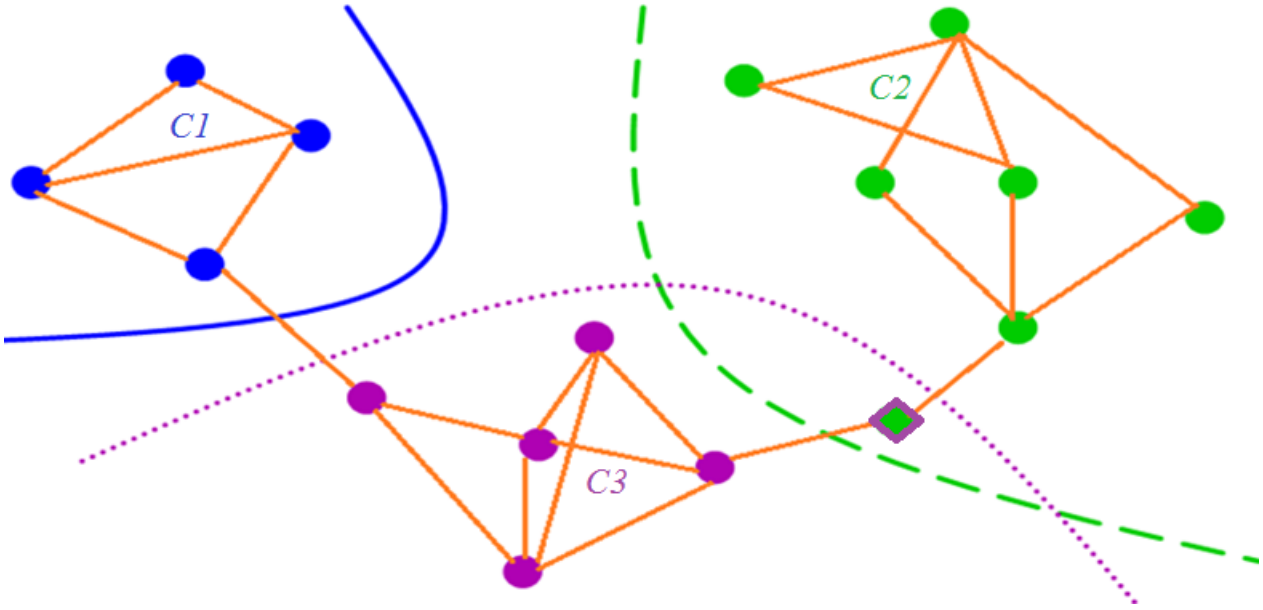


Figure 1.1: *Syntactic example of three overlapping communities detection (C1,C2,C3), the overlapping node is labeled by diamond.*

79, 200], and their performance is usually affected by the network topology [10, 54] and by the dimensionality of data (Curse of dimensionality problem).

The term "Curse of dimensionality" refers to any problem in data analysis that results from a large number of variables (attributes). When the dimensionality is high[1] as data becomes sparse [16, 58]. The analysis of data having large number of features (that is a typical situation in biology) is challenging because:

- The concept of distance becomes less precise due to concentration effects;
- Different clusters might be found in different subspaces;
- Clusters might exist in arbitrarily oriented affine subspaces [1];
- Data could likely include irrelevant features.

[1] when the dimensions is moderate (>10)

1.3 Community detection in biological networks

Biological networks are formed by the complex interactions among genes, proteins and metabolites, which give rise to various complex aggregations [25] such as expression networks, pathway networks, gene regulatory networks, metabolic networks, and Protein-Protein interaction networks *(PPIs)* (see Fig.1.2).

Protein-protein interactions (*PPIs*) occur when two or more proteins bind together in a cell (in vitro or in a living organism) as the interaction interface of proteins is evolved to a specific purpose. The interactions between proteins are connected to biological functions. Not all possible *PPIs* occur in any cell at a given time [41]. In studies of biological networks, such as *Saccharomyces cerevisiae* PPIs [104] networks, community detection techniques are used to extract aggregations showing dense relationships.

Protein interactome can give us the most valuable information about biological functions in cells. Protein Interactions requires an exact fit, as proteins evolve they might change and form new interactions. How ever even a tiny change in the protein structure could make some existing interactions no longer functional and this could cause risks to the entire organism. In most of these biological networks, the interaction map of various proteins are still unknown. In addition, even known interactions are changing due to new discoveries obtained using nowadays biological technologies, such as Next Generation Sequencing (*NGS*).

In biological networks we are interested in aggregating molecular components based on their behavior and interaction such as enzymes, transcriptional module, and signaling pathway. Inferring significant protein communities is fundamental task in many biological researches like drug discovery. The detected communities may help in revealing the relevance of specific macromolecular assemblies or proteins affecting a specific biological process or disease (e.g., cancer) [100]. It is known that proteins involved in the same cellular processes often interact with each other. Therefore, the functions of uncharacterized proteins can be predicted through comparison with the interactions of similar known proteins, and the detection of pertinent communities in *PPIs* networks can be used to predict the function of uncharacterized proteins based on the functions of others they are grouped with.

1.4 Limitations of existing community detection techniques

Adopting an efficient community detection technique in biological networks is an open problem. For example, when we apply a clustering approach to community detection problem we have to face the following problems:

- Initialization criteria (e.g., choosing an initial number of clusters is required in partitional clustering like K-Means [114], but not needed in hierarchical clustering).

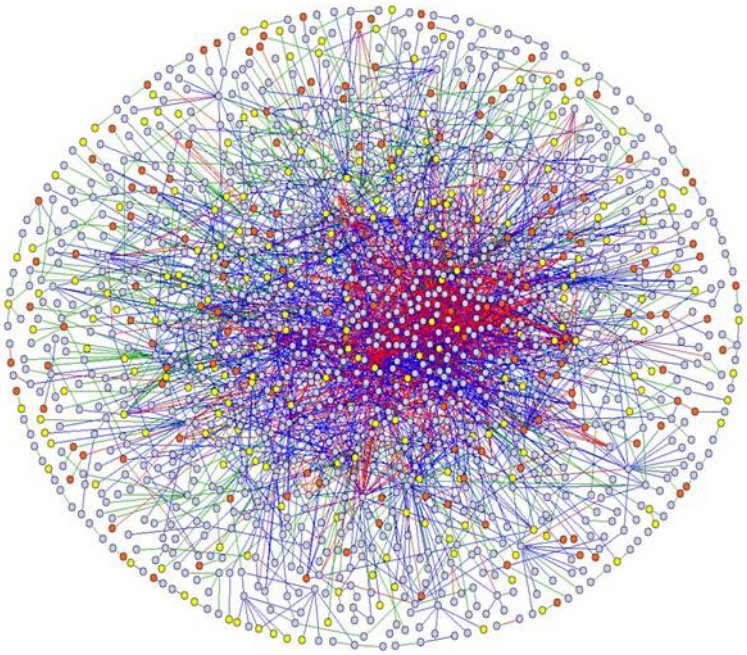


Figure 1.2: *Map of human protein interactions. (From [171])*

- Accuracy (e.g., a main drawback of hierarchical clustering is the possible misclassification of some nodes [143], while removing edges may result in singleton clusters in graph bisection approach).
- Stability (e.g., results may differ depending on the specific similarity measure used and on the random initialization of cluster centers in partitional clustering).
- Complexity (e.g., deciding whether a cut exists is an *NP*-Complete problem even for regular graphs or else for spectral clustering there is a cost concerning the computation of the first k eigenvectors of their Laplacian matrix).
- Noise sensitivity (e.g., hierarchical clustering is very sensitive to noise artifacts).
- Overlap detection (e.g., the algorithm ability to detect possible overlapping between communities) as depicted in Fig.1.1.

Clustering approaches differ in several aspects. The more relevant is the underlying hypothesis on what kind of aggregation should constitute a cluster. This a-priori model selection is often implicit in how the clustering procedure is specified. There are many approaches used for clustering, but only some of them are suitable for community extraction specially in biological data. Moreover, there is no validity criteria for the communities obtained.

1.5 Thesis aims

This thesis explores the challenges of biological network clustering for discovering biological communities. To this aim:

- It proposes some approaches for community detection on high dimensional data.
- It exploits the enrichment framework for identifying significant biological communities.
- It applies the proposed approaches on real biological cases such as Yeast protein-protein interactions, *HIV-1* infection, Leukemia, and dynamical networks.

1.6 Thesis novelty

The main novelty of this thesis can be summarized in the following items:

1. Survey of state of the art of fuzzy, kernel, and modularity crisp and overlapping community detection methods (published in [125], and submitted to a journal).
2. A comparison of centrality measures used for characterizing hubs, bridges, and outlier nodes in networks (published in [119, 125]).
3. A proposal of a possibilistic spectral modularity *(PSM)* and its fuzzy version *(FSM)* for inferring fuzzy communities in networks (published in [126], and [161]).
4. A proposal an ensemble of *FSM* and semantic information *(SE-FSM)* for inferring semantically enriched overlapping (biological) communities (published in [122], and [165]).
5. Evaluating the quality of *PSM* using the state of the art fuzzy similarity measures and proposing a network spreadability measure (submitted to a journal).
6. Application of the proposed approaches for inferring significant communities in *PPIs* networks, such as *S.cerevisiae*, *HIV-1*, and Leukemia (published in [119], [125], [121], and [124]).
7. Proposal of a correlation-based version of *PSM* employed for analyzing dynamical networks and a fuzzy version *(COR-FSM)* (published in [120], [123], and submitted to a journal).

1.7 Thesis structure

This thesis is organized as follows: Chapter 2 contains state of the art related studies; Chapter 3 proposes novel methods for overlapping community detection; Chapter 4 evaluates the quality of the proposed methods compared to the related studies on benchmark networks from the literature; Chapter 5 applies the proposed methods in chapter 3 on real world biological networks in yeast S.cerevasie, and homosapiens (in particular HIV-1, and leukemia protein-protein interaction netwotks); Chapter 6 extend the proposed methods in chapter 3 for dynamical community detection with an application in Italian stock market financial network; while chapter 7 contains the thesis conclusion.

Chapter 2

State of the art

2.1 Graph based measures for community detection

A network can be represented as a weighted graph $G(v;e)$, where v is the set of vertices or nodes and e is the set of edges or links and a weight value is assigned to each edge. The length of a path with endpoint vertices s and t in a graph $G(v;e)$ is the sum of the weights on its edges (see Fig.2.1). Community detection studies [143, 19, 68, 17] devote huge efforts to capture complex relational network structures by attempting to exploit weights between interacting nodes. In a social network weights can be a function of the relationship between linked individuals like co-authorship, duration, or friendship, while in PPI networks weights refer to the biological interactions between nodes (proteins).

The art of identifying nodes having more influence over the network structure than others is referred to as a *node centrality* study. A vertex with high centrality implies that it lies on considerable fractions of shortest paths connecting vertices. Various *centrality measures* are used in network analysis such as centrality degree, closeness, betweenness, and modularity [166, 83, 174, 69, 20, 99].

In the following of this section, the implemented graph based measures for community detection are presented:

- **The Vertex degree** (or *Centrality degree*) $C_D(v)$ of a node v indicates the risk of catching the information flow and is defined as the number of links incident upon v, however $C_D(v)$ may be deceiving due to its locality:

$$C_D(v) = \frac{deg(v)}{n-1}, \tag{2.1}$$

  where n is number of nodes.

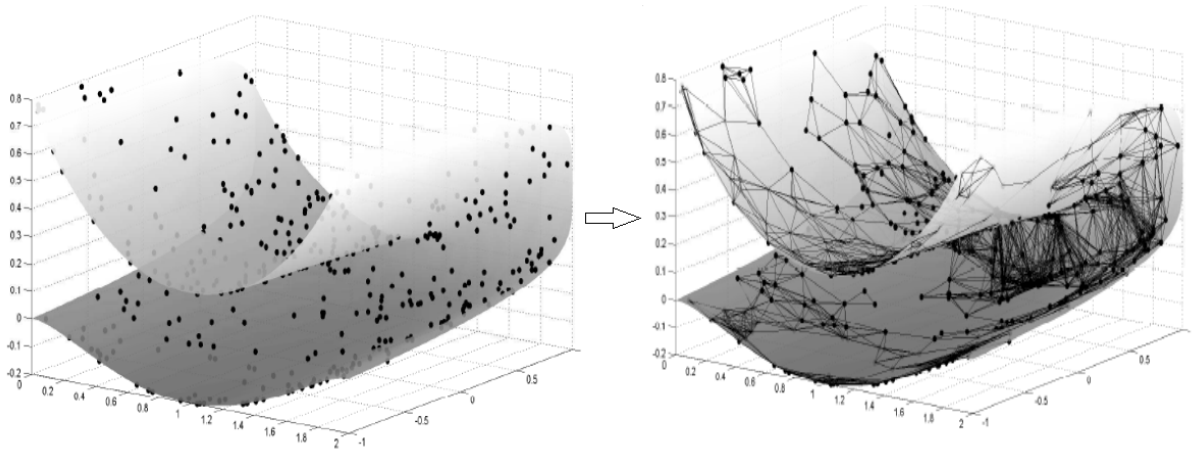


Figure 2.1: *Networks construction (right) using (neighborhood) similarities (left). (From [88, 135])*

- **The Node Closeness** $C_C(v)$ is the inverse of "farness" or distance from other vertices such that $d_{G(v,t)}$ is the shortest distance between nodes v and t [166]:

$$C_C(v) = 1 \Big/ \sum_{t \in V} d_{G(v,t)} \,. \tag{2.2}$$

- **The Betweenness** $C_B(e)$ of an edge e is measured by the ratio between shortest paths linking each vertex pairs s and t that pass through e referred as $\sigma_{st}(e)$ and all shortest paths between these pairs σ_{st} [69, 20]:

$$C_B(e) = \sum_{s,t \in V, s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}, \tag{2.3}$$

  Girvan and Newman [142] proposed a divisive method based on progressive removal of edges (see Tab. 2.1). Edges to be eliminated are chosen on the basis of the updated evaluation of betweenness scores after each edge removal. Betweenness can be computed for all vertices in time $O(mn)$ and requires $O(n+m)$ space for a network with m edges and n vertices [20]. In addition to the complexity of this approach that makes it unfeasible in application to large networks, another disadvantage is that there is no quantitative evaluation of the resultant communities.

- **The Clustering coefficient** indicates network tendency to cluster. It was introduced by Watts and Strogatz [191]. The idea is based on measuring the ratio between number of cycles incident on edge and the largest possible cycles may reside on it. Clustering coefficient

Table 2.1: The Newman's edge betweenness community detection method [142].

1. Calculate the betweenness of all existing edges in the network.
2. The edge with the highest betweenness is removed.
3. The betweenness of all edges affected by the removal is recalculated.
4. Repeat steps 2 and 3 are until no edges remain.

is given by:

$$C_{CC}(v) = \frac{2m_v}{k_v(k_v-1)}, \tag{2.4}$$

where m_v denotes the number of links connecting the k_v neighbors of node v to each other. Hence, the average clustering coefficient is defined as:

$$\langle c \rangle = \frac{1}{n}\sum_{v=1}^{n} C_{CC}(v). \tag{2.5}$$

The network contains small linked clusters or homogenous if the clustering coefficient is independent of k, while in case of hierarchical networks the clustering coefficient $\backsim k^{-1}$, indicating that sparsely connected nodes are part of highly clustered areas.

- **The Average Path Length** *(APL)* of a network is the average number of steps along the shortest paths for all possible pairs of network nodes, and for nodes x and $x^{'}$ having distance $d(x,x^{'})$ in a network with n nodes is given by:

$$APL = \frac{2}{n(n-1)}\sum d(x,x^{'}), \text{ s.t.,} x \neq x^{'}. \tag{2.6}$$

Moreover, we notice that the *APL* is considered as a measure of the efficiency of information transport on the network.

- **The Modularity** (Q) of a network [143] is defined as:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i,c_j), \tag{2.7}$$

where A_{ij} is the weight of edge linking vertices i and j, $k_i = \sum_j A_{ij}$ is the degree of vertex i, c_i is the community to which node i is assigned, $m = \frac{1}{2}\sum_{ij} A_{ij}$, and $\delta(c_i,c_j)$ function is 1 if c_i is the same as c_j and 0 otherwise.

Network modularity is used for measuring the strength of community structure in networks and also as an objective function to maximize with suitable optimization methods. Q is a scalar value ranging between -1 and 1. Networks with high modularity implies the existence of dense connections within communities and of sparse links between them. Although modularity suffers a resolution limit specially in case of detecting small communities, it has the advantages of not requiring prior knowledge about the number or sizes of communities, and it is capable of discovering network partitions composed of communities having different sizes.

It is worth noting that both Newman's betweenness [142] and Newman & Girvan modularity [143] approaches can not support overlapping communities detection. Several overlapping trials of Newman & Girvan modularity were proposed [146, 141, 203, 195, 87, 113].

2.2 Fuzzy community definitions

After applying a fuzzy clustering algorithm such as Fuzzy C-Means to a network, the fuzzy communities obtained can be characterized using the following definitions:

- **Crisp nodes** are the network nodes having a membership in only one community.
- **Fuzzy nodes** are the network nodes having significant membership in more than one community.
- **Hub nodes** are nodes having highest degree in a community.
- **Bridgeness** quantifies the degree to which a given vertex (or node) is shared among different clusters. It is given by: $b(s) = 1 - \sqrt{k\sigma^2(U_{1..k}(s))}$. If a vertex s belongs to all the clusters in the graph with equal probabilities, then the variance evaluates to zero, which in turn gives a bridgeness score of 1. This implies that ideal bridges in the network will belong to multiple communities with equal probabilities. Moreover, vertices with low degree and high bridgeness usually correspond to **outliers** (a node that does not have a dominant community) [141].
- **Bridge nodes** are fuzzy nodes linking the communities having highest bridgeness [141].
- **Node spreadability** is a novel measure will be proposed in this thesis (see Sect.3.4) that measures each node capability to distribute the information among the network communities.

2.3 Neighbor-based similarities

Recently a shift from feature-based data representation to similarity-based representation can be noticed. This shift has prompted a renewed interest in methods based on similarities which are not evaluated as distances in some suitable space, but given as inputs obtained from some complex, costly, or unobservable source. This section examines some proximity measures derived from the analysis of the ordered list of neighbors to data items [161], and proposes some fuzzy generalizations that allows to use these criteria as real-valued similarity measures.

As opposed to geometric distance, these criteria are applicable even when data are not Euclidean. We are interested in fuzzy data, since they are more realistic. For the sake of concreteness, fuzziness is assumed to derive from measurement uncertainty in an Euclidean setting, similarly to the cases studied by Yager [198]. The methods studied can be applied in other cases.

These measures may be interesting even when a primary similarity is available. For popular methods such as kernel classifiers [37] and spectral clustering [186] a suitably sparse proximity (similarity) matrix has definite computational advantages. In addition, while nearest neighbor classification criteria are asymmetric, similarities based on *shared* neighbor lists are symmetric (and positive semidefinite), i.e., they are possible metrics.

The similarity can be considered a fuzzy generalization of identity, so it is possible to unify binary, discrete-valued, and continuous-valued similarity measures under a general fuzzy framework.

2.3.1 Fuzzy equality and inequality

It assumes that each data item in X and Y is a fuzzy vector (a vector whose components are fuzzy numbers), with Gaussian membership of identical variance.

Crisp equality may be expressed by the following indicator function:

$$\delta(x,y) = \begin{cases} 1 & \text{if } x-y=0, \\ 0 & \text{otherwise,} \end{cases} \tag{2.8}$$

so that $\delta(x,y) = 1 \iff x = y$. A fuzzy generalization ($\mho$) of (2.8) can be interpreted as the indicator of the fuzzy number "about 0" computed for the value $x-y$. This is a fuzzy equivalence relation [201], and several choices are possible. For instance, in [158] interval values were used, and (in)equality tests compared the overlap between intervals. The Gaussian membership connects fuzzy knowledge and probabilistic uncertainty [161, 198] (a very common source of imprecise knowledge). Equality may be defined as follows:

$$\mho(a,b) = e^{-(a-b)^2/\gamma^2} \quad \text{for } a, b \text{ fuzzy number centroids,} \tag{2.9}$$

where the fuzziness parameter γ should be tuned to the range of the data under study.

We also need to "overload" the indicator $\mho$ for vector arguments: a natural choice is based on the logical conjunction of equality for the N individual components,

$$\mho(\mathbf{x},\mathbf{y}) = \left(\prod_{p=1}^{N} \mho(\mathbf{x}_p,\mathbf{y}_p)\right)^{1/N} \qquad \text{for } \mathbf{x},\,\mathbf{y} \text{ fuzzy } N\text{-vectors.} \tag{2.10}$$

The geometric mean provides a normalization with respect to N, whereas the simple product would yield a smaller value for larger N. For scalars ($N = 1$), (2.10) reduces to (2.9).

To model *fuzzy inequality*, Yager [198] employs the probability distribution of the difference of two values. In [163] a similar approach was followed, but it is based on a pair of equal variance Gaussian distributions which yields a sigmoid:

$$\lambda(a,b) = \frac{1}{1+e^{(a-b)/\gamma}}\,, \tag{2.11}$$

a natural generalization of the Heaviside indicator with fuzzification parameter γ. In the limit for $\gamma \to 0$ the inequality tends to become crisp. The value of $\lambda(a,b)$ represents the degree of truth of the statement "a is larger than b", with value 0 when $a \ll b$, 1 when $a \gg b$, and a value approaching 0.5 when $a \approx b$. We remark that similarly to Yager's approach, the fuzzy indicator $\lambda()$ can be used to directly define fuzzy rankings [163].

2.3.2 A taxonomy of neighbor-based similarity measures

To design neighbor-based similarity values, several choices are possible, although not all combinations make sense. Here we present these choices along with the naming conventions. Measures will be given a symbol indicating its supervised / unsupervised and crisp / fuzzy nature, and further choices will be specified by a subscript, a sequence of *key* characters chained in the same order as they are presented in the following. Some examples are given at the end of this section.

- **Supervised vs unsupervised measures**; Measures based on neighbors assume the existence of a primary pairwise similarity information, and use it to define a new similarity. In the supervised case this primary information is given by a class labeling for Y. Classification in the same class is a kind of similarity measure, binary only in the crisp case, used for instance in correlation clustering [9].

  A base name will be given to each measure as follows: t is a supervised measure, s an unsupervised measure.

- **Crisp vs fuzzy measures**; Depending on the availability of fuzzy information, measures can be either crisp or fuzzy. Symbols for crisp measures are : t, s; while for fuzzy measures are : $\hat{t}$, $\hat{s}$

- **Number k of neighbors**; The number k of neighbors considered can range from 1 to $|Y|$. In general we only distinguish the case $k = 1$ from $k > 1$. A subscript starting with `1` has $k = 1$; a subscript starting with `k` has $k \neq 1$, to be specified.

- **Near vs far neighbors**; The neighbors considered for measuring similarities are usually the nearest to the items considered. However, some approaches use the *farthest* neighbors.

  A subscript that has `1` or `k` followed by `n` indicates that neighbors are the nearest; if followed by `f`, neighbors are farthest. Note that all remaining choices are meaningful only for $k > 1$, so subscripts starting with `1` end with either `n` or `f`.

- **List vs set of neighbors**; Neighbors can conceptually be arranged in a list or in a set. Comparing two lists requires that all neighbors considered appear with the same rank in both lists. Comparing two sets does not take ranks into account, only the existence of points. The key `l` indicates list, while `s` stands for set. This choice does not apply for $k = 1$.

- **Measuring strategy**; There are several ways to evaluate similarity between two data items. Here we consider three approaches: The number of coincident neighbors or of neighbors sharing the same class label can be used directly (this is indicated by the key `c`). Or, this number can be offset by a threshold ("ε-insensitive" count), so that if the shared items are less than this threshold the similarity is zero, and only the excess is counted (key `e`). Finally, a binary similarity is obtained by simply considering counts that are/are not above the threshold (key `t`).

  Note that in the fuzzy case the concepts of "coincident items" and "larger than a threshold" must be suitably defined, as we did in Section 2.3.1, and give rise to a fuzzy truth degree rather than a binary value. This choice does not apply for $k = 1$.

For instance, the measure $\hat{t}_{1\mathrm{n}}$ is a supervised, fuzzy similarity based on the nearest neighbor only; s_{knst}, a crisp unsupervised similarity, is 1 if the count of the shared near neighbors among the nearest k (taken in any order) is above threshold t, where k and t must be specified; and $\hat{s}_{\mathrm{kflc}}$, a fuzzy unsupervised similarity, is the count of fuzzy shared farthest neighbors taken with their ranks, where k must be specified.

2.3.3 Supervised measures

- **k nearest neighbor classification**; The well-known nearest neighbor classification rule [63, 36] states that a point is attributed to the same class as its closest reference point:

$$c(x_i) = c\left(y_{I_{i1}}\right) \ . \tag{2.12}$$

Therefore, the nearest neighbor similarity between x_i and x_j is

$$t_{1\mathrm{n}}(x_i, x_j) = \begin{cases} 1 & \text{if } c(y_{I_{i1}}) = c(y_{I_{j1}}) \ , \\ 0 & \text{otherwise.} \end{cases} \tag{2.13}$$

Nearest neighbor rules can be stated as a Bayes decision criterion working on a crude estimation of class-conditional data densities [63, 48]. Using a set of k neighbors makes the nearest neighbor rule less sensitive to local variation in data distribution. A point is in the class that is most represented among the k nearest neighbors, or:

$$c(x_i) = \arg\max_c \left| \{ y_j \,:\, c(y_j) = c,\, R_{ij} \leq k \} \right| \ , \tag{2.14}$$

where $|S|$ = cardinality of S. Once a majority class is established, the k nearest neighbor similarity is again given by $s_{1\mathrm{n}}$ (2.13).

The number of majority class representatives can be used to introduce a degree of classification confidence [157], allowing to implement a rejection option in classification tasks. In the k-nearest neighbor distance, confidence can be used to grade the distance, making it non-binary. In [157], t_{knlc} is also used.

- **Fuzzy k nearest neighbors**; Fuzzy supervised nearest neighbor criteria may be based on metrics $\hat{t}_{1\mathrm{n}}$ and $\hat{t}_{\mathrm{knst}}$. However Keller *et al.* proposed a fuzzy k nearest neighbor classification [97], based on the idea of applying the fuzzy c-means membership function to Y. The fuzzy k nearest neighbor classification is a C-vector of memberships $u(x_i) = [u_1(x_i), \ldots, u_C(x_i)]$, so that

$$u_h(x_i) = \frac{\sum_{j=1}^{k} c_h(y_{I_{i,1}}) \left(1/D_{ij}\right)^{2/(m-1)}}{\sum_{j=1}^{k} \left(1/D_{ij}\right)^{2/(m-1)}} \ , \qquad h : 1 \ldots C \tag{2.15}$$

and the nearest neighbor similarity between x_i and x_j can be computed as the degree of similarity between two modified membership vectors

$$\hat{s}'_{\mathrm{knlc}}(x_i, x_j) = \frac{\sqrt{u'(x_i) \cdot u'(x_j)}}{k} \ , \tag{2.16}$$

a variation of $\hat{s}_{\mathrm{knlc}}$ where

$$u'(x) \begin{cases} u(x) & \text{if } u(x) \in \text{top } k \text{ memberships,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

Note that according to this model, the nearest neighbor ($k = 1$) criterion is always crisp, unless the target class indicators $c_h(y_j)$ are fuzzy to start with. This requirement is not usually satisfied.

When the y_j are the centroids of meaningful groupings in data, e.g., class centroids or centroids of convex components of classes, these criteria are called "nearest centroid". Usually, however, they are prototypes [50] or landmarks [39] in the data space, in a more generic sense. A similar representation, with $k \equiv m$, has been used in [57].

2.3.4 Unsupervised measures

- **Shared near neighbors**; A binary similarity measure was proposed by Jarvis and Patrick [94]. The measure is inherently unsupervised. It assigns two points to the same cluster whenever, among the k nearest neighbors of each point, at least t are common to both ("shared"). In other words, it uses the s_{knst} measure: If $n_n(k) = \left|\{y_p : R_{ip} \leq k,\, R_{jp} \leq k\}\right|$ is the number of shared near neighbors among the nearest k,

$$s_{\text{knst}}(x_i, x_j) = \begin{cases} 1 & \text{if} \quad n_n \geq t\ , \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

- **Shared farthest neighbors**; Another binary, unsupervised similarity measure was presented in [162]. This measure assigns two points to the same cluster whenever the *farthest* neighbor of the two points is the same:

$$s_{\text{1f}}(x_i, x_j) = \begin{cases} 1 & \text{if} \quad I_{in} = I_{jn}\ , \\ 0 & \text{otherwise.} \end{cases} \tag{2.19}$$

  This measure can be applied recursively up to depth k, using s_{kflt} for a sequence of thresholds t. Similarly to the k nearest neigbors rule, we can apply this criterion to the set of k farthest neighbors.

- **Fuzzy shared near and far neighbors**; Using the definitions of fuzzy equality and inequality given in Section 2.3.1, fuzzy neighbor-based similarities can be readily defined. These similarities measure the degree of overlap between the k nearest or farthest neighbors. If the lists of neighbors include similar, albeit not identical, points, these give some (reduced) contribution to the measure.

  The following is the definition of measures, for both nearest and farthest neighbors, using the "list" representation, i.e., keeping ranks into account:

$$\hat{s}_{\text{knlc}}(x_i, x_j) = \sum_{p=1}^{k} \text{eq}\left(y_{I_{ip}}, y_{I_{jp}}\right)\ ; \qquad \hat{s}_{\text{kflc}}(x_i, x_j) = \sum_{p=m-k+1}^{m} \text{eq}\left(y_{I_{ip}}, y_{I_{jp}}\right)\ . \tag{2.20}$$

  In measures using the "set" representation, data items are required to be included in the two lists of neighbors, but not necessarily with the same rank in both.

$$\hat{s}_{\text{knsc}}(x_i, x_j) = \sum \text{top}(k) \left\{\text{eq}\left(y_{I_{ip}}, y_{I_{jq}}\right),\ i = 1 \ldots k,\ j = i \ldots k\right\}\ ; \tag{2.21}$$

$$\hat{s}_{\text{kfsc}}(x_i, x_j) = \sum \text{top}(k) \left\{\text{eq}\left(y_{I_{ip}}, y_{I_{jq}}\right),\ i = m-k+1 \ldots m,\ j = i \ldots m\right\}\ . \tag{2.22}$$

  Here $\text{top}(k)$ is the top-k query operator, applied to the fuzzy equality level $\text{eq}()$ of $k(k+1)/2$ pairs of neighbors (either nearest or farthest).

  Finally, threshold-based nearest and farthest neighbor measures, according to the ideas of Jarvis and Patrick, are defined as follows:

$$\hat{s}_{\text{knst}}(x_i, x_j) = \lambda\left(\hat{s}_{\text{knsc}}, t\right)\ ; \qquad \hat{s}_{\text{kfst}}(x_i, x_j) = \lambda\left(\hat{s}_{\text{kfsc}}, t\right)\ . \tag{2.23}$$

2.4 Crisp and overlapping community detection approaches

Several approaches have been proposed in recent literature to infer community structures [142, 143, 65, 107, 153, 15]. They can be categorized with different criteria, either aim at characterizing how connectivity of a (candidate) community is measured, i.e., using the objective function (e.g., modularity) [143], or how it is optimized, i.e., the algorithm (e.g., random walk or topological structure analysis). The most used community detection algorithms can be categorized into graph based partitioning [43], hierarchical clustering [100], partitional clustering [86, 127], spectral clustering [59, 137], edge removal [142], random walk, modularity based, an overlapping methods [143]. We list here some state of the art methods for a graph with m edges, and n nodes, each method representing one of those groups, we select for comparative analysis we will report in our experimental analysis.

2.4.1 Greedy divisive based approaches

- **Betweenness based**; This approaches depend on shortest path analysis. Newman and Girvan [142] proposed a divisive method based on progressive removal of edges having high betweenness scores after each edge removal. The complexity of this approach is $O(m^2n)$ that makes it unfeasible in application to large networks, another disadvantage is that there is no quantitative evaluation of the resultant communities.

- **Link clustering**; A hierarchical approach reused Girvan and Newman divisive objective, but edges are removed based on their clustering coefficient[153]. The computation time is faster than betweenness approach $O(\frac{m^4}{n^2})$ and $O(n^2)$ on sparse graphs due to its locality, moreover the stopping criteria depends on intrinsic community characteristics rather than using a quality criteria like modularity.

2.4.2 Simulated annealing based approaches

These approaches benefit of simulated annealing technique for optimizing the objective function used in aggregations discovery e.g modularity [80]. The cost function or temperature is optimized until it reaches the global optima, hence global moves reduce the risk of getting trapped in local minima. The optimization process makes a local random movement of a node to a different random community, this movement is accepted if it increases modularity Q for instance, otherwise it is accepted with probability $exp(-\beta\Delta Q)$, where β is a global time-varying parameter, Other attempts experimented moving a series of nodes simultaneously. In comparison, these algorithms are slower than most of the divisive paradigms but yield higher accuracies.

2.4.3 Modularity based approaches

- **Heuristic maximization bases approaches**; Network modularity is used for measuring the strength of community structure in networks and also as an objective function to maximize with suitable optimization methods [143]. Although modularity suffers a resolution limit specially in case of detecting small communities, it has the advantages of not requiring prior knowledge about the number or density of communities. This approach requires $O((m+n)n)$ complexity or $O(n^2)$ on a sparse graph.

- **Blondel method**; This method [15] is a greedy model implementing an iterative process terminating when a maximum modularity (Q) is obtained. This method repeatedly attempts to optimize the *network modularity* Q [143] in a small local community scale, then, after a partition is identified, communities are replaced by their super-nodes. This leads to hierarchical decomposition of the network. The computational time is $O(m)$, which makes it more scalable than other greedy approaches, but, anyway, it may stuck in a local minima and may not find a good optima.

- **Multiresolution based greedy approaches**; Due to resolution limit for Newman-Girvan modularity, multiresolution approaches attempt to use tunable parameters to control the size of communities to be detected. Such methods aim at identifying the right communities by exploring all possible scales. Moreover, theses approaches showed the capability of discovering communities in graphs characterized by hierarchical structures in which clusters resides inside each others. This principle was extended in [107] by adding a tunable multiresolution parameter α having $O(n^2 log(n))$ complexity.

- **Eigenvector based approaches**; The *eigenvector approach* [144] is aimed to maximizing the modularity using eigenvalues and eigenvectors derived from the full adjacency matrix of the original graph. This method gives good results in case of bisection graphs, while it is less accurate with more than two communities. The algorithm converges in $O(n^2 log(n))$ time.

2.4.4 Random walk based approaches

An example of this family is Markov clustering based approach [46]. It is a repetitive procedure, starting with an expansion step that generates a stochastic diffusion matrix from the original affinity matrix. Then it estimates its integer power as an indicator for random walking probability. An inflation hypothesis is then applied by raising the matrix elements to power α targeting trapping the random walker within a community. Expansion and inflation are then repeated until obtaining a forest adjacency matrix. The components of the resultant disconnected tree are the communities. The computational complexity is $O(nK^2)$, where k refers to the remaining matrix largest elements after each inflation step.

2.4.5 Partitional based approaches

Centroid-based clustering techniques [92] use a prototypical point in the data space, usually, but not exclusively, defined as the barycenter of points to represent each cluster. Membership of a point in a cluster is usually calculated using a distance-based criterion. This results in clusters which, in crisp cases, are the tiles of a Voronoi tessellation, and are therefore convex; in fuzzy clustering, tile boundaries are fuzzy, but their shape is still convex. Moreover, not all of clustering algorithms are effective for community detection. The most popular approaches of this family are crisp k-means [114, 86], fuzzy c-means [14].

- ***K*-Means** (See Sect. 3.2).
- **Fuzzy *C*-Means** (See Sect. 3.3).
- **Possiblistic *C*-Means** (See Sect. 3.5).

2.4.6 Spectral based approaches

Spectral clustering-based approach, known also as *affinity-based approach*, where clustering is performed on the pairwise distance or similarity (affinity) matrix [59, 186]. This approach performs clustering in the data spectral space, grouping in the same cluster instances featuring a mutual intracluster similarity that is larger than the similarity to instances in other clusters . As for other clustering-based approaches to community detection we assume that each obtained cluster corresponds to a different community, but, as spectral clustering exploits the local connectivity, it may yield to very complex overall shapes modeling with precision the communities as shown in Fig. 2.3.

Starting from the adjacency matrix that describes the network topology, several derivations built upon it to characterize network properties such as Laplacian matrix, the normalized Laplacian matrix, and the correlation matrix [30]. Spectral clustering exploits such matrices to infer possible complex network community structures which may be difficult to be obtained using traditional partitional clustering approaches for instance.

The hypothesis behind these approaches is that the eigenvector components corresponding to nodes in the same community should be similar [45]. Spectral clustering exploits such matrices to infer possible complex network community structures which may be difficult to be obtained using partitional clustering approaches for instance. In Chapter 3, three Spectral clustering-based methods for community detection: one based on the algorithm proposed by Ng et al. [145] and two novel fuzzy versions of it will be presented in detail.

Fig. 2.2 shows clustering data using partitional K-Means vs. spectral space. K-Means is biased to dense spherical clusters and performs poorly, while in the affinity/laplacian space given by two

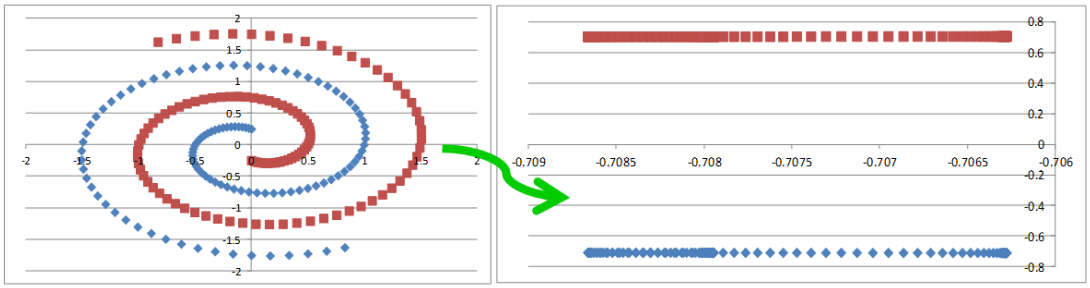


Figure 2.2: Clustering in data partitional K-means (left) vs. spectral space (right).

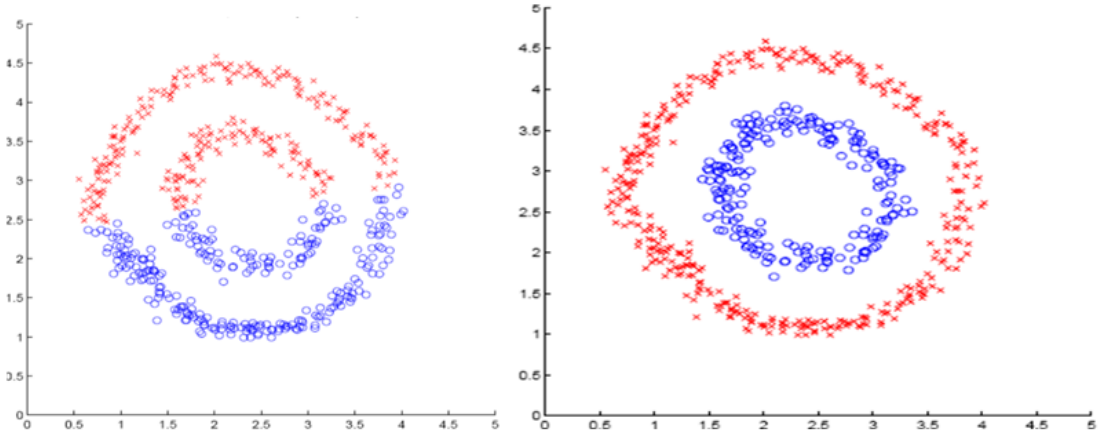


Figure 2.3: Clustering 2 Circles using: K-Means (left), Normalized Spectral clustering (right).

leading eigenvectors, clusters are trivial to separate. The clustering result on 2 circles is shown in Fig. 2.3. It is worth noting that the eigengap of the Laplacian matrix can be used to infer the number of clusters (this will be discussed in eigengap index illustrated in Sect. 2.5).

2.4.7 Overlapping community detection approaches

Several mentioned methodologies such as Newman's betweenness [142] and Newman & Girvan modularity [143] approaches can not support overlapping communities detection. Many overlapping community detection methods exist in the literature. These methods can be classified into the following categories 1) Divisive approaches based on splitting nodes based by maximizing a quality criteria [79, 77, 75], 2) Link partitioning approaches based on splitting links instead of nodes [65] 2) Clique percolation approaches extract communities based on the connected subgraphs (cliques) [148, 65, 195], 3) Seed based approaches expand the communities starting from set of seed nodes [107, 195], 4) Membership based approaches analyze the fuzzy member-

ship of each node to detect the dominant (significant) community to which each node belongs [119, 122, 125, 141, 203], 5) Label propagation approaches assume that nodes with same label that form a community are allowed to have multiple labels. [76, 79, 65].

This section will highlight the following methods used in the experiments:

- **Cluster Overlap Newman- Girvan Algorithm** (CONGA) is a divisive based approach that expands Newman algorithm to support overlapping detection [79]. It splits vertices linkages among clusters, if their betweenness exceeds the maximum edge betweenness. It has a complexity of $O(m^3)$ or $O(n^3)$ in case of sparse graphs.

- **Percolation approaches** such as Clique Percolation Method (*CPM*) that is a local search algorithm looks for communities that may overlap[148]. This approach for inferring communities consists in extracting the largest possible subgraphs that can be explored by rolling k-cliques across the network, where a k-clique rolls by rotating about any of its component $(k-1)$ cliques. *CPM* detects overlapping communities of various networks especially in case of low k values in a reasonable manner, and it is robust against removal or insertion of a single link.

  The method depends on network topology and requires existence of significant number of cliques. Moreover, the computational time needed to find all k-cliques of a graph is an exponentially growing with graph size $O(exp(n))$.

- **Simulated annealing based approaches** such as Nepusz approach [141] that detects overlapping communities based on optimizing a fuzzy version of Newman modularity and simulated annealing, then detect node belonging using a measure called bridgeness that will be introduced presented in Chapter 3.

- **Spectral clustering based fuzzy approach** such as Zhang approach [203] that depends on having the number of communities in advance then use it to calculate the eigenvectors.

- **Nonnegative matrix factorization approaches** (NMF) assume that membership is based on factorizing the network Laplacian matrix by using a diffusion kernel for instance [204].

- **Label propagation methods** [76, 79, 65] are iterative methods assume that each node determines its community based on the labels of its neighbors and usually join the community to which the maximum number of its neighbors belong to, such that the occurring ties are broken uniformly at random. The iterative process terminates when the network nodes no longer change their labels.

2.5 Evaluating the detected communities

The main goal of data clustering is to obtain low inter-cluster similarity (separation) among members of different clusters, and high intra-cluster similarity (cohesion) among members of different clusters. Clustering validity (quality) can be estimated by exploiting different aspects of the data distribution among clusters, such as well separateness of data, optimum number of clusters, stability of the community detection method, significance, reproducibility of results even in case of noise, and ability to detect anomalies. Most methods cluster noisy data either in one of the existing clusters or create an additional cluster (outliers).

Validity indices [92, 84, 85] are statistical measures used to judge the quality of a clustering by evaluating some various aspects of clusters, such as monotonicity, noise, densities, sub-clusters and skewed distributions. There is a huge literature devoted to the subject. Some indices are specific of a given clustering method, while others are general, and work on the membership matrix only.

The quality of a clustering may be measured on the basis of point distribution only, as in the case of *internal* quality measures, or with the aid of external knowledge attached to the available data, such as class labels, as in the case of *external* measures. Finally, we can mention stability-based approaches, that relate cluster quality to how much they change as data or algorithm parameters are perturbed.

In fuzzy environments, they can be classified into:

1. Validity indices based only on the clustering membership values.
2. Validity indices consider dataset and membership values.
3. Hybrid approaches involving other measures to overcome limitation of depending only on distance between cluster centroids.

It is worth noting that clustering validity can be used to infer the correct number of clusters. Moreover, from experimental studies validity indices vary in their stability, even some of them could not be adapted to some clustering methods. Here we are mainly interested in well-established measures that can be applied to each of the cases of our interest, i.e., crisp centroid-based clusters, fuzzy centroid-based clusters, and clusters obtained by spectral partitioning. Some relevant validity indices we have implemented in [165, 120] are illustrated in the the following of this section.

These measures can be used to evaluate the detected partition obtained using a community detection method on a dataset (or network) having X data points (obtained from the similarity matrix in a graph), n points (nodes), c clusters (communities in a graph), U membership matrix, and v cluster centers:

- **Davies Bouldin index** (V_{DB}) is based on measuring the average similarity between each cluster and its most similar one, hence a lower V_{DB} indicates better clustering compactness and separation [38, 51]. It is defined as:

$$V_{DB} = \frac{1}{c}\sum_{i=1}^{c} R_i, \tag{2.24}$$

where $R_i = \underset{i=1,\ldots,c, i\neq j}{max} R_{ij}, i = 1,\ldots,c$, such that $R_{ij} = \frac{s_i+s_j}{d_{ij}}$ is the similarity measure of clusters based on the cluster dissimilarity measure between centers $d_{ij} = d(v_i, v_j)$, and the dispersion measure of a cluster obtained by $s_i = \frac{1}{\|c_i\|}\sum_{x\in c_i} d(x, v_i)$, having $\|c_i\|$ nodes.

- **Beni-Xie index** (V_{BX}) [194] was explicitly proposed for fuzzy clustering. This index is a function of the data set and the centroids of the clusters based on measuring the ratio of the total variation of the partition and the centroids (U,V) and the separation of the centroids vectors. The minimum values of this index under comparison indicates the best partitions. It is defined as:

$$V_{BX} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n} u_{ij}^2 ||x_i - v_j||^2}{n(min_{j\neq l}||v_j - v_l||^2)} \tag{2.25}$$

- **Partition Entropy index** (V_{PE}) is a fuzzy validity measure based on membership matrix. It modified the partition coefficient index by considering logarithmic values of membership matrix. It is defined as:

$$V_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} log(u_{ij}), \tag{2.26}$$

where estimating the optimal number of clusters c^* is an optimization problem obtained by $\underset{k<n}{min}(V_{PE}(k))$.

- **Dunn Index** (V_D) is given by:

$$V_D = \underset{i=1,\ldots,c}{min}\left\{\underset{j=i+1,\ldots,c}{min}\left(\frac{d(c_i,c_j)}{\underset{k=1,\ldots,c}{max}(diam(c_k)}\right)\right\}, \tag{2.27}$$

where $d(c_i,c_j) = \underset{x\in c_i, y\in c_j}{min}\{d(x,y)\}$ and $diam(c_i) = \underset{x,y\in c_i}{max}\{d(x,y)\}$. For c clusters obtained. A larger value of Dunn index is obtained if clusters are well-separated, consequently large distances reside between different clusters and the diameters of each cluster is expected to be small. This index was proposed for crisp clustering validation. The maximization of Dunn Index leads to inferring optimal number of clusters c^*, its main limitations are:

1. Time complexity with increasing c and n.

2. Noise sensitivity as cluster diameter may be large in a noisy environment. Several Dunn-like index derivations exploit this index using different definition for cluster distance and cluster diameter.

- **Alternative Dunn Index** (ADI) modified Dunn's index methodology in measuring cluster pairs dissimilarity by considering cluster centers v:

$$ADI(c) = \min_{i \in c} \left\{ \min_{j \in c, i \neq j} \left(\frac{\min\limits_{x \in c_i, y \in c_j} \left| d(y, v_j) - d(x, v_j) \right|}{\max\limits_{k \in c} \left\{ \max\limits_{x,y \in c} d(x,y) \right\}} \right) \right\}. \tag{2.28}$$

- **The Rand index and its fuzzy variations** (RI) [154] is a commonly adopted measure for evaluating an obtained partition compared to a reference partition defined as:

$$RI = \frac{a+d}{a+b+c+d} \tag{2.29}$$

where,

  - a, is the number of object pairs whose elements are in the same cluster in partition P_1 and also in P_2;
  - b, is the number of object pairs whose elements are in the same cluster in P_1 but are in different clusters in P_2;
  - c, is the number of object pairs whose elements are in the same cluster in P_2 but are in different clusters in P_1;
  - d, is the number of object pairs whose elements are in different clusters in P_1 and are also in different clusters in P_2.

  The Rand index lies in $[0,1]$ with zero indicating that the two partitions do not agree on any pair of elements and one indicating that the two partitions are identical.

  Several extensions have been proposed for fuzzification of the Rand index, such as measures based on the comparison of objects using dot product association of memberships [70, 26, 160], measures based on building a bonding matrix containing the cosine association of partitions then evaluate a, b, c, and d [22], and measures which depend on the contingency matrix of the two partitions [2], however when considering the fuzzy Rand index as a distance matrix these measures lack some properties such as reflexivity (i.e., $d(P_1, P_1) = 0$), moreover some measures such as [2] fail to guarantee symmetry property as well [89].

  The Rand index is viewed as distance function $D_{RI} = 1 - RI$ [89].

$$a = (1 - |u - v|) \times u \times v; \tag{2.30}$$

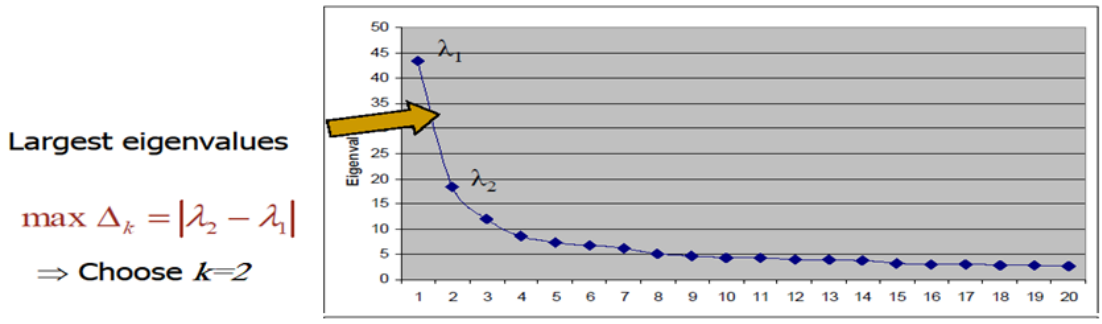


Figure 2.4: Eigengap analysis, choosing number of clusters (k) that maximize eigengap Δ_k.

$$b = (1 - |u - v|) \times (1 - u \times v); \tag{2.31}$$

$$c = max((u - v), 0); \tag{2.32}$$

$$d = max((v - u), 0). \tag{2.33}$$

where $u = E_{P_1}(x,x^{'})$, $v = E_{P_2}(x,x^{'})$ are a fuzzy equivalence relation on X in terms of a similarity measure on the associated membership vectors $P(X) = P_1(X), P_2(X), ...P_k(X) \in [0,1]^k$ given by:

$$E_{P_1}(x,x^{'}) = 1 - \parallel P(x) - P(x^{'}) \parallel, \tag{2.34}$$

where $\parallel . \parallel$ is a proper distance on $[0,1]^k$

The distance measure on fuzzy partitions is then defined as the normalized sum of degrees of discordance:

$$D(P_1,P_2) = \frac{\sum_{(x,x^{'}) \in C} |E_{P_1}(x,x^{'}) - E_{P_2}(x,x^{'})|}{n(n-1)/2}) \tag{2.35}$$

Hence, the fuzzy Rand index is given by:

$$RI_f = 1 - D(P_1,P_2) \tag{2.36}$$

- **Eigengap index** aims at finding the gap between the eigenvalues in spectral clustering with c clusters, an usual criterion for binary partitioning quality is algebraic connectivity [56] (the second eigenvalue). Since this is not a valid criterion for completely connected, weighted graphs as those obtained from Euclidean data, we use the following related index: $q_{\text{GAP}} = 1 - \lambda_{c-1}/\lambda_c$, valid for multi-way clustering as well, where λ_c is the c-th eigenvalue of the graph Laplacian. This index is positively related to quality and has 1 as its maximum value (see Fig. 2.4).

2.6 Measuring community stability

2.6.1 Fuzzy communities

Communities identified can have crisp or fuzzy memberships, although a crisp community is really a special case of a fuzzy community. In a fuzzy community the elements can belong to the subsets in a community at various degrees whose values lie in $[0,1]$. The subsets are fuzzy while the full set that is partitioned is crisp. A fuzzy community induces a crisp community if the maximum membership value for each node over the various communities is replaced by one and all other values are replaced by zeros (defuzzification). Even if the objective is to obtain a crisp community it is still useful to use a fuzzy clustering process as an improved way of handling noise.

2.6.2 Stable communities

A community detection method is stable if the communities do not differ too much, in the presence of small variations of a model parameter or of the data, i.e., if, given a measure of similarity between communities:

$$Q(P_1,P_2) \in [0,1], \tag{2.37}$$

if $P_1 = P(\nu_1)$ and $P_2 = P(\nu_2)$ i.e., if they are obtained by the community detection method for two values of the model parameter ν such that $\nu_1 - \nu_2 = \varepsilon$, with ε sufficiently small, then $P\{Q(P(\nu_1,P(\nu_2)) < \delta\} > p$ with $\delta = \delta(\varepsilon)$ and p a probability which is close to 1 for δ small enough (p, as usual, indicates a prescribed confidence level) [160].

2.7 Semantic similarity measures

Proteins encoded by the genes associated with a common disorder interact together, participate in similar pathways, and share Gene Ontology (*GO*) terms [4]. Drug discovery for certain disease may arise from a hypothesis that genes contributing to a common disorder have an increased tendency for their products to be linked at various functional levels, this may be induced from experimental studies of protein-protein interactions, co-regulation, co-expression, and annotated semantic information (e.g., those stored in gene ontology).

GO provides a structured controlled vocabulary of gene and protein biological roles, which can be applied to different species [4] . GO has three different aspects : molecular function, biological process and cellular component.

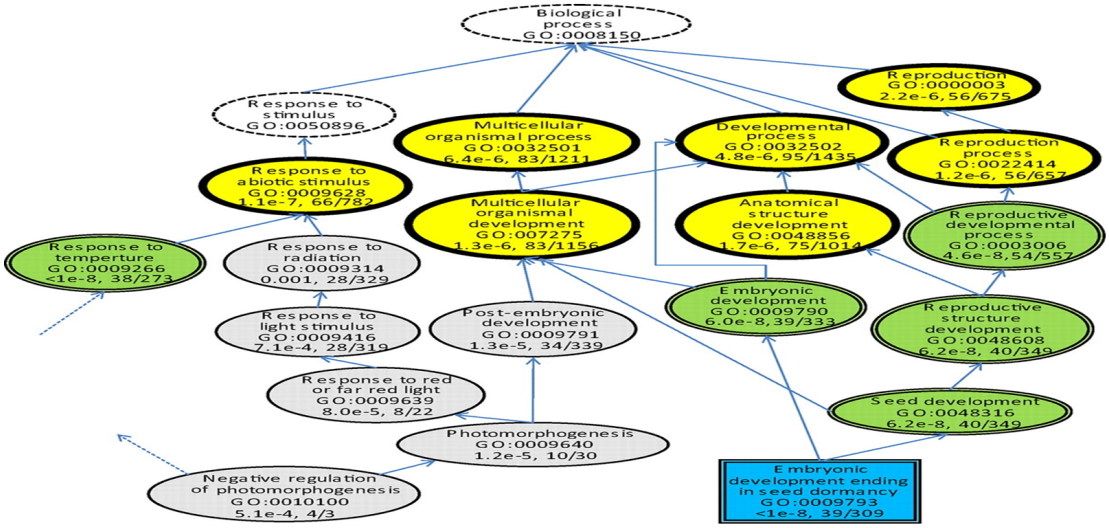


Figure 2.5: Directed acyclic sub graph of Gene Ontology. (From [112])

Each protein has elementary molecular functions that normally are independent of the environment, such as catalytic or binding activities. Sets of proteins interact and are involved in cellular processes, such as metabolism, signal transduction or RNA processing. Proteins can act in different cellular localizations, such as nucleus or membrane [33].

GO organizes the terms in three directed acyclic graphs (DAG), one for each aspect. Each node of the graph represents a concept, and the edges represent the links between the concepts (see Fig. 2.5). Links can represent two relationship types: is-a and part-of.

A main aim of this thesis is to improve the quality of aggregation discovery in dense biological interactions by incorporating such information embedded in biological repositories and mapping them in the feature space.

Several biological available repositories for instance protein data banks [128] and gene ontology contain annotated information and biological knowledge either extracted with the help of experts or from papers or other sources. For improving the effectiveness of community detection many similarity measures entailed in those frameworks can be proposed as shown in the rest of this section.

2.7.1 Feature based measures

Similarity between terms can be calculated by applying a function to their structures, or properties [164].

$$sim_{re}(a,b) = w.S_{synsets}(a,b) + u.S_{features}(a,b) + v.S_{neghiborhoods}(a,b), \tag{2.38}$$

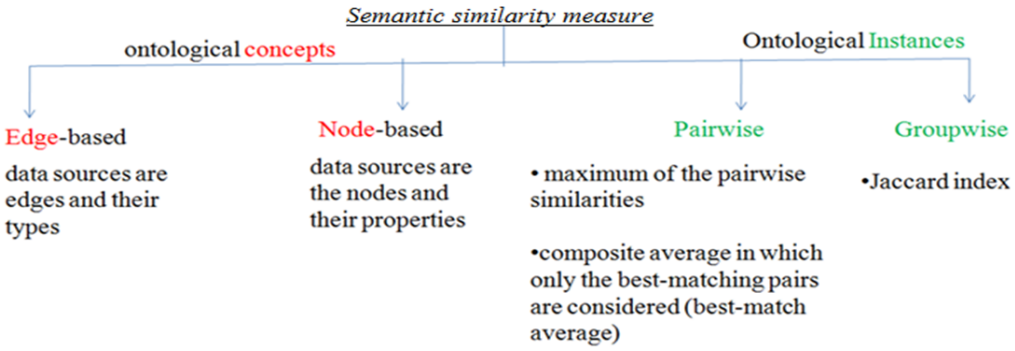


Figure 2.6: Classification of Semantic similarity measures

where $S(a,b) = \frac{|A \cap B|}{|A \cap B| + \gamma(a,b)|A \setminus B| + (1-\gamma(a,b))|B \setminus A|}$. Feature-based measures depend on mutual features and properties between concepts. The lack of this information of them limits their applicability; another problem is their dependency on the weighting parameters that balance the contribution of each feature.

2.7.2 Edge based measures

Similarity between terms can be estimated by measuring the path length linking them in ontology [152]:

$$dis_{rad}(a,b) = min_{\forall i} |path(a,b)|. \tag{2.39}$$

Another edge based approach considers relative depth (N_3) or Least Common Ancestor (LCA) between ontological terms (N_1, N_2) [193]:

$$dis_{w\&p}(a,b) = \frac{2N_3}{N_1 + N_2 + 2N_3}. \tag{2.40}$$

A different measure was given based on both the number of nodes N_p separating the ontological nodes corresponding to terms a and b, including themselves, and the depth D of the taxonomy [108]:

$$dis_{l\&c}(a,b) = -log\left(\frac{N_p}{2D}\right). \tag{2.41}$$

Edge based approaches show a low computational cost compared to other approaches [167]. However, they are not scalable to wide and detailed ontologies because they depend on shortest path between concepts, while other factors are not considered such as frequency and distribution of ancestors.

2.7.3 Information Content based measures

These measures assume that the semantic similarity between two terms or nodes a and b can be evaluated from the amount of mutual information they share [168].

Information Content based measures use the following notations:

- A term a is an **ancestor** of a term (t) when there is at least one path from a to t, this is given by:
$$Ancestors(t) = \{a|Paths(a,t) \neq \emptyset\}. \tag{2.42}$$
- The **Frequency** $(Freq(t))$ of a term (t), can be defined as the number of times that t and all its descendants occur. It is given by:
$$Freq(t) = \sum \{occur(t_i)|t \in Ancestors(t_i)\}. \tag{2.43}$$
- The **Likelihood** $L(t)$ (or probability) of observing an instance of a term (t) is given by:
$$L(t) = \frac{Freq(t)}{maxFreq}. \tag{2.44}$$
where $maxFreq$ is the maximum frequency of all terms. The $maxFreq$ of each aspect of GO is equal to the frequency of the DAG root.
- The **Information Content** (IC) of a term (t) is given by:
$$IC(t) = -log(L(t)). \tag{2.45}$$
Note, the Information Content of a term (t) is inversely proportional to its frequency (see Eq. 2.43) in the ontology.
- The **Least Common Ancestor)** $(LCA$ of terms a and b is a term in a lexical taxonomy (e.g. WordNet or gene ontology) that has the shortest distance from the two terms being compared, where $LCA(a,b) \in Ancestors(a) \cap Ancestors(b)$.
- The **c** between two terms a and b in the ontology can be defined as:
$$Mutual(a,b) = IC(LCA(a,b)). \tag{2.46}$$
However, there exist possible variations of $Mutual(a,b)$ in the literature such as those will be discussed in Sect. 3.8.

Interesting examples for estimating the Information Content based semantic similarity measures between two terms a and b in an ontology are:

- **Resnik semantic similarity measure**; [156] that only considers the Mutual semantic information between a and b:

$$sim_{resnik}(a,b) = Mutual(a,b). \tag{2.47}$$

- **Lin semantic similarity measure**; [111] that enhances Resnik similarity by considering the information content of *LCA* and the two compared concepts:

$$sim_{lin}(a,b) = \frac{2sim_{resnik}(a,b)}{IC(a)+IC(b)}. \tag{2.48}$$

- **Jiang and Conrath semantic similarity measure**; [96] that is similar to Lin's one. If we define a distance $dis_{j\&c}(a,b)$ between a and b as:

$$dis_{j\&c}(a,b) = IC(a)+IC(b)-(2sim_{resnik}(a,b)), \tag{2.49}$$

the Jiang and Conrath semantic similarity ($sim_{j\&c}$) can be estimated as:

$$sim_{j\&c}(a,b) = \frac{1}{dis_{j\&c}(a,b)+1}. \tag{2.50}$$

2.7.4 Topological and other approaches

These approaches combine several structural characteristics (such as path length, depth and local density), then they assign weights to balance the contribution of each component in the final similarity value [110].

2.7.5 Groupwise and pairwise semantic similarities

A protein may participate in several biological processes or carry out different molecular functions hence it may be annotated to several terms in Gene ontology. To obtain the semantic similarity between two interacting proteins, we can combine semantic enrichment measures that can be categorized as:

- **Groupwise** Term Semantic Similarity measures that can be directly extended to measure protein similarity, simply considering as input the two sets of *GO* terms annotating the proteins.
- **Pairwise** Term Semantic Similarity measures that evaluate similarity of pairs of terms and therefore, are not directly applicable to genes and proteins.

Pairwise term similarity measures can be combined by applying a **mixing strategy**. The following mixing strategies are reported in literature for calculating protein **functional similarity** scores:

1. Based on GO term Semantic similarity:

   - Average (*Avg*). The average of all term pairwise similarities [116];
   - Maximum (*Max*). the maximum of all term pairwise similarities [172];
   - Averaging all the Best Matches (*ABM*) for two annotated proteins is the mean of best matches of *GO* terms of each protein against the other [132], it is given by:

$$ABM(p,q) = \frac{1}{n+m}\left(\sum_{t\in T_p^X}\max_{s\in T_q^X} S(s,t) + \sum_{t\in T_q^X}\max_{s\in T_p^X} S(s,t)\right), \tag{2.51}$$

     where in Eqs. 2.51 and 2.52, $S(s,t)$ is the semantic similarity score between terms s and t, T_r^X is a set of *GO* terms in X representing the molecular function (MF), biological process (*BP*) or cellular component (*CC*) ontology annotating a given protein r and $n = |T_p^X|$ and $m = |T_q^X|$ are the number of *GO* terms in these sets. These two approaches produce different scores and they are equal only when $n = m$, which is not often the case in a set of annotated genes or proteins.

   - Best Match Average (*BMA*). the average of similarity between best matching terms [7]; For two annotated proteins p and q it is the mean of the following two values: average of best matches of *GO* terms annotated to protein p against those annotated to protein q, and average of best matches of GO terms annotated to protein q against those annotated to protein p, it is given by:

$$BMA(p,q) = \frac{1}{2}\left(\frac{1}{n}\sum_{t\in T_p^X}\max_{s\in T_q^X} S(s,t) + \frac{1}{m}\sum_{t\in T_q^X}\max_{s\in T_p^X} S(s,t)\right) \tag{2.52}$$

     The difference between *ABM* and *BMA* approaches is subtle in their conception and scores produced by these two approaches differ.

   - funSim. first protein semantic similarities in MF and BP ontologies are determined using *Max*, *Avg* or *BMA* mixing strategies and then they are combined together in a non-linear way [170];
   - Information Theory-based Semantic Similarity. best matching pairs are filtered on the basis of their similarities, then the average is calculated [179];
   - FuSSiMeG. similar to max strategy and the maximum of all term pairwise similarities weighted by the *IC* of the terms is selected [35]. A well known issue with all these

statistical measures of closeness is that they are sensitive to scores that lie at abnormal distances from the majority of scores, or outliers.

This means that these measures may produce biases which affect protein functional similarity scores [199].

2. Based on the GO term direct information content:

   - The functional similarity approach, *SimGIC* [149, 150], which uses the *IC* of terms directly to compute protein functional similarity from their GO annotations and uses the Jaccard index.
   - SimDic [132] (czekanowski or Lin like measure), is given by:

$$SimDIC(p,q) = \frac{2 \times \sum_{x \in A_p^X \bigcap A_q^X} IC(x)}{\sum_{x \in A_p^X} IC(x) + \sum_{x \in A_q^X} IC(x)} \tag{2.53}$$

   - SimUIC [132] is given by:

$$SimUIC(p,q) = \frac{\sum_{x \in A_p^X \bigcap A_q^X} IC(x)}{\max\left(\sum_{x \in A_p^X} IC(x), \sum_{x \in A_q^X} IC(x)\right)}, \tag{2.54}$$

   where A_r^X is a set of *GO* terms together with their ancestors in X representing the ontology (*MF*, *BP* or *CC*) annotating a given protein r. Note that these two measures are still to be evaluated and compared to the existing functional similarity measures.

2.8 Semantic similarity in biology

Semantic similarity measures *(SS)* are correlated with other biological features and similarity measures as follows:

2.8.1 Semantic similarity and Sequence similarity

Sequence Similarity is the oldest approaches used to establish relations among genes. Proteins with similar sequence likely accomplish similar functions [82].

Many works found strong correlation between sequence and Semantic Similarity at least when considering *MF* ontology, for instance Pearson's linear correlation for a set of protein pairs [151, 116, 134, 109, 93, 115], while Pesquita et al. [149] proposed an assessment based on non-linear regression that aimed at fitting the data with a function that closely follows the behaviour of semantic similarity against sequence similarity.

This evaluation is not as straightforward as it is for Protein-Protein Interactions (PPI) data, different strategies have been used, and not always coherent results are obtained. In general, it seems that BMA mixing strategy should be preferred to Max and Avg approaches. Resnik BMA, simGIC and simIC BMA have often been identified as the best measures [149, 109] (see Sect. 2.7.5).

2.8.2 Semantic similarity and Pfam families

Sequence similarity is not the only kind of structural similarity that can be computed between proteins. Protein family similarity (*Pfam*) [12] is a structural similarity of a higher level than sequence similarity. Each family describes a set of related proteins, which can have identical molecular functions, are involved in the same process, or act in the same cellular location.

Proteins generally comprise one or more functional regions, commonly termed domains. Since proteins sharing the same domains are likely to have some common functional aspects, assessing semantic similarity measures using domain composition information is an appealing alternative to sequence similarity data.

Grouping proteins in families has been a common technique to organize them according to their biological role. For example, the most successful large-scale effort for increasing the coverage of GO annotations within the UniProt database[1] is based on the exploitation of family annotations [27].

Protein family similarity overcomes some of the limitations of sequence similarity as they can represent some evolutionary conserved structure and have implications on the protein's biological role.

Couto et al. [35] showed that, especially when using *MF* ontology, semantic similarity significantly increases when the shared families between two proteins increases. Couto et al. [35] and CESSM [151] evaluate Pearson's linear correlation of semantic similarity and a *Pfam*-based similarity measure, and rank semantic similarity measures according to correlation levels, however Couto concludes that Jang and Conrath measure with *GraSM* (see Sect. 3.8) outperforms *Lin* and *Resnik* measures, while *CESSM* (a tool for semantic similarity evaluation) disagreed.

Benabderrahmane et al. [13] proposed a novel assessment strategy (IntelliGO) and concluded it

[1] http://www.uniprot.org/

is the best. However, *CESSM* disagreed as well.

Although *Pfam* families are good candidates to assess and compare semantic similarity measures, current findings are still not coherent [82].

2.8.3 Semantic similarity and Functional modules

A functional module in a protein interaction network is a set of interacting proteins that share a common biological goal or play a biological role. For instance, a pathway or a protein complex.

A biological pathway is a number of biochemical steps, linked together, that perform a process inside cells. Since proteins within the same pathway are involved in the same biological process, they are likely to be annotated with the same or similar terms in the GO (at least in BP ontology) and therefore having high semantic similarity.

Several studies analyzed the relation between Semantic Similarity measures and **pathways**, see, e.g.:

- Guo et al. [81] that analysed the distribution of Resnik scores when considering pairs of proteins belonging to the same pathway. They showed that all protein pairs within a Kyoto Encyclopedia for Genes and Genomes (KEGG) [2] pathway have significantly higher similarity scores than randomly expected when considering BP ontology. On the other side, semantic similarity on MF and CC ontologies decays exponentially as proteins became farther within the same pathway.
- Wang et al. [190] that compared G- SESAME and Resnik measures (over the MF ontology) using Yeast pathways. They used the scores to hierarchically cluster genes within the same pathway, and by visually inspecting clustering results, they concluded that G-SESAME scores protein pairs consistently with human perception of protein relatedness.
- Benabderrahmane et al. [13] that evaluated the difference between similarity scores between protein pairs within the same pathway and protein pairs from different pathways.

Other studies used the **protein complexes** to assess Semantic Similarity measures, see, e.g.: Li et al. [109] and Wu et al. [192]. To this aim they used the following steps:

- Reconstruct Yeast PPI network relying on Biological Process (BP) and Cellular Component (CC) based semantic similarity scores.
- Map manually annotated Munich Information Center for Protein Sequences (MIPS) [133] complexes on their networks.

[2] http://www.genome.jp/kegg/

- Evaluate how many MIPS complexes were included in their reconstructed network.

Note that Wu's PPI network based on Relative Specificity Similarity (*RSS*) measure (and max mixing strategy) encompassed 120 out of 214 *MIPS* complexes, whereas Li's *PPI* network, based on *simIC* (and max mixing strategy), extended the coverage to 159 complexes. Such analysis revealed the applicability of Semantic Similarity measures for PPI network reconstruction problems and for biological clustering.

2.8.4 Semantic similarity and Expression profiles

Several studies compared Semantic Similarity measures to gene expression profile similarity, see, e.g.:

- *Wang* et al. [189] that found high semantic similarity is significantly associated with strong expression correlation for pairs of genes using Eisen dataset [53], this conclusion agrees with [109, 93].
- Sevilla et al. [172] that concludes that the correlation of gene expression and semantic similarity at low levels of semantic similarity values is negligible, whereas at higher levels of semantic similarity values, they are highly related. This behaviour has been confirmed by [196];

The ranking of the Semantic Similarity measures according to their behaviour compared with expression profile similarity, Li et al. [109] identified simIC as the best measure, followed by Resnik (coupled with max mixing strategy), whereas in [93], *TCSS* is the best measure, followed by Resnik (always using a mixing strategy).

2.8.5 Semantic similarity and Evidence codes

Protein annotations are assigned in many different ways [47]. A big portion of term annotations fall into the electronically inferred category.

Experimentally verified annotations are likely to be correct, but only a small fraction of proteins are annotated through this process. Electronically inferred annotations drastically extend the coverage, but at the expense of introducing a lot of noise and the presence of more generic annotations.

2.9 Semantic terms significance

Terms significance *(p-value)* is based on the hypergeometric test. For instance, the *GO* terms enrichment is obtained as:

$$P(X = x) = \frac{\binom{g}{x}\binom{t-g}{r-x}}{\binom{t}{r}}, \tag{2.55}$$

where t denotes the total number of genes, g refers to the total number of genes belonging to the *GO* category of interest, r the number of differentially expressed *(DE)* genes, and x is the number of *DE* genes belonging to the *GO* category [159].

2.10 Clustering ensemble techniques

Ensemble learning combines the results of different algorithms or the same algorithm with different parameters setting or using resampling methods like bootstrapping, bagging [21] or boosting [169], or representing the data set from different views (e.g., subspaces) to achieve a better result than that of the single learner [90]. Ensemble methods were originally designed for classification [106], but subsequently applied to clustering. It should be noted, unsupervised learning such as clustering is much more difficult to compose the ensemble learner than that of the supervised learning because of the deficiency of the category labels for test samples and prior information among the clustering results and it can be stated as a dual problem.

Clustering ensemble was firstly proposed by Strehl and Ghosh [177] and it can go beyond what is typically achieved by a single clustering algorithm in novelty, stability, robustness, scalability, and parallelization [182, 95]. Different clustering algorithms may produce different partitions because they impose different structure on the data; No single clustering algorithm is optimal. A consensus clustering method (a.k.a. consensus function) constructs an aggregate representation of the ensemble and use it as the basis for extracting a consensus partition. Ensemble techniques designed to circumvent the cluster label correspondence problem [105].

Directly combining partition matrices is possible only in restrictive hypotheses, namely, when the correspondence problem admits a satisfactory solution. This can obviously happen only if the number of clusters is the same for both partitions, or if it is possible to match several clusters from one partition on only one cluster of the other, at least to a reasonable degree. If the above issues can be satisfactorily solved, combining partition matrices involves working with matrices whose size is linear in the number of data items and in the number of clusters.

Mainly, the clustering ensemble task can be divided into the following three main steps:

1. **Producing the partitions**: There are several approaches to produce the partitions (clusterers) such as:

- Using different clustering algorithms (see Sect. 2.4).
- Running the same algorithm many times with different parameters or initializations. (e.g., run K-Means algorithm N times using randomly initialized clusters centers, or different number of clusters, or different dissimilarity measures).
- Using different samples of the data.
- Random projections (feature extraction) (e.g., project the data onto a random subspace).
- Feature selection (e.g., use different subsets of features).

2. **Solving the general correspondence problem**: It implies finding the best-matching partition among several, as opposed to two, partitions (e.g., using Hungarian algorithm [105]). This is a problem of minimizing some average measure of match of the resulting partition with respect to all other partitions in the ensemble, a possibly complex optimization task.

3. **Clustering ensemble** (combining the partitions): According to [185], ensemble clustering techniques can be classified as:

   - Median partition based techniques; It aims at finding a partition P that maximizes the similarity between P and all the N partitions in the ensemble: $P_1, P_2, \ldots, P_N$, and it requires to define the similarity between two partitions such as Normalized mutual information [177], or other indices such as [182, 64], or those discussed in Sect. 2.5.
   - Object co-occurrence based techniques
     - Relabeling/voting based techniques; After finding the corresponding cluster labels among multiple partitions, then obtain the consensus partition through a voting process [42, 6, 49, 60, 183].
     - Co-association matrix based techniques; It computes a co-association matrix based on multiple data partitions, then apply a similarity-based clustering algorithm (e.g., single link and normalized cut) to the co-association matrix to obtain the final partition of the data. For each pair of data points, a co-association matrix indicates whether (or, if fuzzy, how much) they belong to the same cluster.
       This is a powerful technique, adopted for instance in Fred and Jain's "evidence accumulation" technique [67] or in Strehl and Ghosh' method [177]. In this case, the consensus partition is simply obtained by averaging the coassociation matrices: this works even for partitions with different cluster numbers. However, the resulting consensus matrix has a size that grows quadratically with the number of data items. Moreover, once obtained, the consensus matrix must be re-clustered, since it contains only implicit partition information.
     - Hyper-graphs and meta-graphs based techniques; It builds a weighted graph to represent multiple clustering results from the ensemble, then find the optimal partition of data by minimizing the graph cut [55, 177].

Chapter 3

Proposed methods for overlapping community detection

In [125] we proposed the *Fuzzy c-means Spectral Modularity (*FSM*) community detection* method for network analysis. The *FSM* estimates the number of communities k using the maximization of modularity procedure depicted by Newman and Girvan in [143] and then performs data clustering in a subspace spanned by the first k eigenvectors of the graph Laplacian matrix [186]. The method is based on the spectral clustering approach described in [145], with the main difference consisting in using as a technique of clustering the Fuzzy C-Means [14] that makes it possible to identify significant overlapping protein communities.

This chapter is organized as follows: A novel overlapping centrality measure termed *spreadability* is introduced in Sect. 3.4; The proposed Semantically Enriched Fuzzy c-means Spectral Modularity (*SE-FSM*) community detection method is illustrated in Sect. 3.7, while its application to the discovery of communities in the *HIV-1* and Leumekia *PPIs* networks of Homo sapiens is shown in Chapter 5.

3.1 Spectral Modularity community detection methods

There are many approaches to clustering. Among them, the most promising for discovering communities in networks is the spectral graph partitioning, proposed by Donath and Hoffman in 1973 [44].

Spectral clustering refers to methods used to cluster n objects based on the evaluation of the Laplacian matrix obtained from the data similarity matrix (which is symmetric and non negative), and then in application of a clustering technique (such as K-Means) to data in a subspace spanned by the first k eigenvectors of the Laplacian matrix. Several approaches exploit spectral theory

for clustering, such as un-normalized spectral clustering by Shi and Malik [173], normalized spectral clustering by Ng et al. [145], random-walk spectral clustering by Melia and Shi [137].

It is worth to say that most of the previously mentioned spectral clustering approaches differ only in the way they calculate the Laplacian matrix L, and whether they apply a normalization step.

This chapter presents in detail three novel Spectral clustering-based methods for community detection exploiting the spectral clustering method proposed by Ng et al. [145] (see Tab. 3.1). This algorithm performs the clustering in the affine subspace spanned by the first k eigenvectors of the symmetric Laplacian matrix defined as [30]:

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}, \tag{3.1}$$

where D is the degree matrix ($D_{ii} = \sum_j W_{ij}$, and $D_{ij} = 0$ for $i \neq j$), W is the adjacency matrix, with $W_{ii} = 0$ and $W_{ij} = e^{-\frac{||x_i - x_j||^2}{\sigma_i \sigma_j}}$ for $i \neq j$ (σ_i and σ_j are scaling parameters). The clustering is obtained with the application of the K-Means algorithm [176, 114, 48] leading to the minimization of the empiric distortion:

$$\sum_{h=1}^{n} \sum_{i=1}^{c} (u_{ih})^m \, \|\mathbf{x}_h - \mathbf{v}_i\|^2, \tag{3.2}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the instances to be clustered, $\mathbf{v}_1, \ldots, \mathbf{v}_c$ are the centroids, and $u_{ih} \in \{0, 1\}$ is an element of the membership matrix U, and the so-called *probabilistic constraint* $\sum_i u_{ih} = 1 \; \forall h$ holds.

We highlight that Ng et al. [145] algorithm, as well as many other popular spectral clustering approaches, when used to detect communities in networks presents the following problems:

1. The solutions are unstable due to the random initialization of number and positions centroids and to the presence of may local minima in the K-Means functional (Eq. 3.2).
2. The temporal complexity of eigenvectors computation increases with the increasing of the number of interacting vertices.
3. The application of the K-Means does not permit to detect possible overlaps between communities, as it is a crisp clustering technique.

We underline that the estimation of the optimal number of communities k is an open problem. Some intra-cluster validity indices (as those proposed by Davies Bouldin, Dunn, and others [92]), or affinity measures based on eigen-gap analysis specific for spectral clustering [186] can be used for estimating the number of clusters, but they are not always reliable when used to estimate the number of communities k inside networks.

Table 3.1: The normalized spectral clustering method by Ng et al. [145].

1. Set the number of clusters k, and the similarity matrix $S \in R^{n \times n}$.
2. Compute the normalized Laplacian L_{sym} given in Eq. 3.1.
3. Obtain the top k eigenvectors $v_1, .., v_k$ of L_{sym}, and calculate $V \in R^{n \times k}$ by reshaping them as columns.
4. Get $U \in R^{n \times k}$ by normalizing the row sum of V to 1, where $u_{ij} = v_{ij} / (\sum_k v_{ik}^2)^{\frac{1}{2}}$.
5. For $i = 1, .., n$, let $y_i \in R^k$ represents the i^{th} row of U.
6. Apply K-Means for clustering $(y_i)_{i=1,..,n}$ instances into k clusters.

3.2 The *K-Means Spectral Clustering Modularity* community detection method

The first method we propose is the ***K-means Spectral Modularity (KSM) - community detection method*** that applies the spectral clustering method by Ng et al. [145], estimating the number of clusters (communities) k using the maximization of modularity procedure depicted by Newman and Girvan in [143]; the estimated number of clusters, say k, is used both for selecting the top eigenvectors of the Laplacian matrix, and to set the number of clusters in the clustering algorithm.

k-means optimizes the empiric distortion:

$$\sum_l \sum_j U_{lj} ||x_l - y_j||^2, \tag{3.3}$$

where U_{lj} is integer and $\sum_j U_{lj} = 1 \ \forall l$.

Due to the usage of the K-Means, the KSM-community detection method is also not capable to detect possible overlaps between communities; therefore we proposed some algorithms using fuzzy clustering instead of K-Means, as shown in the following two sections.

3.3 The *Fuzzy C-Means Spectral Clustering Modularity* community detection method

The ***Fuzzy C-Means Spectral Clustering Modularity*** **(or *FSM*)** community detection method we discuss in this thesis and introduced in [125] applies the following three improvements to the original Ng et al. [145] spectral clustering algorithm, when used to detect communities in networks:

1. First of all, the estimation of the number of clusters is performed using the maximization of modularity procedure depicted by Neuman and Grivan in [143]; the estimated number of clusters, say k, will be used both for selecting the top eigenvectors of the Laplacian matrix, and to set the number of clusters in the clustering algorithm.

2. Then, the clustering in the affine subspace spanned by the first k eigenvectors is obtained with the application of the Fuzzy C-Means (*FCM*) clustering algorithm [14, 52] instead of K-Means (used in our aforementioned crisp *KSM* approach). As *FCM* considers that a point may belong to two or more clusters at the same time, with different membership degrees, this choice supports the detection of overlapping communities and can allow us to understand the role that each protein may play in different communities.

3. Moreover, we automate the parameter tuning in the original *FSM* [125] and introduce a novel calculated spreadability cut measure ϖ discussed in Sec.3.4 (see Eq. 3.6), we apply it after *FCM* to remove nodes with membership to discovered communities below a threshold ϖ. This thresholding allows us to aggregate only proteins having strong memberships, and to handle the noise and possible outliers in communities. In extreme cases it allows us to eliminate insignificant communities including nodes with low membership only. When we have an a-priori knowledge on the number of possible communities.

The Fuzzy C-Means (*FCM*) clustering algorithm [51, 14] performs the minimization of the following distortion:

$$J_m(\mathbf{U},Y) \equiv \sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^m d_{ik} \tag{3.4}$$

where: $X = \{x_1, x_2, \ldots, x_n\}$ is a data set containing n unlabeled sample points; $Y = \{y_1, y_2, \ldots, y_c\}$ is the set of the centers of clusters; $\mathbf{U} = [u_{ik}]$ is the $c \times n$ fuzzy c-partition matrix, containing the membership values of all samples to all prototypes; $m \in (1,\infty)$ is the fuzziness control parameter; d_{ik} is a dissimilarity measure between data point x_i and the center y_k of a specific cluster k. Usually the Euclidean squared distance $d_{ik} \equiv \|x_i - y_k\|^2$ is employed as the dissimilarity measure.

The clustering problem can be formulated as the minimization of J_m with respect to Y, under the probabilistic constraint $\sum_{k=1}^{c} u_{ik} = 1$.

Table 3.2: The FSM community detection method.

1. Detect the number of cluster k using the modularity measurement in (Eq. 2.7).
2. Apply the spectral clustering (e.g., Ng et al. Normalized Spectral Clustering Algorithm [145]) and obtain the spectral space using top k eigen vectors.
3. Cluster the resultant spectral space using Fuzzy C-Means (in Eq.3.4).
4. Assign vertices to clusters having members larger than the threshold ϖ (Eq. 3.6).

The Fuzzy C-Means algorithm usually starts with a random initialization of the fuzzy c-partition matrix $\mathbf{U}$ or of the centroids y_k. The derivatives of the sum of the associated Lagrangian are computed with respect to the u_{ik} and $\mathbf{y}_k$ and are set to zero. This yields the Picard iteration of these equations until convergence, that is usually checked by comparing the change in the position of the centroids or in the cost function with some fixed thresholds:

$$u_{ik} = 1 \Bigg/ \sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{y}_k\|}{\|\mathbf{x}_i - \mathbf{y}_j\|} \right)^{\frac{2}{m-1}} \qquad \mathbf{y}_k = \frac{\sum_{i=1}^{n} (u_{ik})^m \mathbf{x}_i}{\sum_{i=1}^{n} (u_{ik})^m} \,. \tag{3.5}$$

Note that in the limit for $m \to 1$ the Fuzzy C-Means Functional J_m (Eq. (3.4)) becomes the expectation of the *KM* empirical distortion and the *FCM* behaves as the classic *KM* algorithm [176, 114, 48].

After applied the *FCM*, for each node we have the evaluation of its (fuzzy) memberships to all communities. At this point usually a defuzzification step is performed to obtain crisp evaluations of memberships to clusters (i.e. either 0 or 1). A winner takes all (*WTA*) criterion can be applied that assigns (with membership 1) each instance to the cluster where it assumes the higher membership ϖ.

The *FCM* shows more stable solutions than *KM*, but it can still trap in local minima of its functional (Eq. 3.4). Moreover, the *FCM* is subject to the high sensitivity to outliers as it assigns the membership values of instances on the basis of their relative distance from the cluster centers. In particular outliers equidistant from two cluster centroids show very high memberships to those clusters (up to ϖ).

Tab. 3.2 shows the *FSM* community detection method. It is worth noting that the proposed spreadability cut measure can be used for estimate information spreadability of the network nodes as we will show in Sec.3.4.

3.4 The spreadability measure

Spreadability ξ is a novel measure we propose here for estimating the node capability of spreading information among the different communities belonging to a network. A node s has a high ξ if it belongs to more than one community. Such nodes affect the network flow and information broadcasting in different communities.

The spreadability measure depends on the dispersion in node memberships, it is calculated using the following steps:

1. For each node s, having membership $U_{1..k}(s)$ in k communities and standard deviation σ, we measure the spreadability cut given by:
$$\varpi = \sigma(U_{1..k}(s)) - \sigma^2(U_{1..k}(s)). \tag{3.6}$$
2. Assign s to each community c_i having membership $> \varpi$, then estimate the number of belonging communities given by:
$$\lambda_s = \left|I_{/s}\right|, \; I_{/s} = |\{c_i | U_{ci}(s) > \varpi\}| \tag{3.7}$$
.
3. Nodes having $\lambda > 1$ are identified as fuzzy, and the more the $\lambda > 1$ the more the node is spreadable (a.k.a, has significant influence across the network communities), while nodes having $\lambda = 1$ are referred as crisp (a.k.a, located locally in their communities).
4. Spreadability for a fuzzy node s belongs to λ overlapping communities is given by:
$$\xi = \sum_{i=1}^{\lambda} U_i(s), \tag{3.8}$$
s.t, s is member in c_i, while for crisp node is given by:
$$\xi = 1 - max(U_{1..k}(s)). \tag{3.9}$$

The spreadability cut (ϖ) has the following characteristics:

- ϖ has a maximum value .25 if a node has membership equal 1 in one community and zero to the others (crisp membership).
- The maximum of ϖ decreases when increasing the number of communities k, this is reasonable because this means that ϖ automatically adjust itself such that the less the communities we may have the higher the threshold (ϖ).
- The more the fuzziness between the values of $U_{1..k}(s)$ for a node s the lower $\varpi(s)$ obtained.

The spreadability cut (ϖ) experimental analysis (see Sect. 5.4) showed that ϖ is a robust and global measure for identifying node fuzziness.

3.5 The *Possibilistic Spectral Clustering Modularity* community detection method

To overcome the problem of sensibility to outliers, in this subsection we propose the ***Possibilistic c-means Spectral Modularity (*PSM*) - community detection method*** that is a possibilistic version of *FSM* discusses in Sect. 3.3.

The *PSM* methods employes the Possibilistic C-Means (*PCM*) clustering algorithm [103, 102, 181, 129] that relaxes the probabilistic constraint on the memberships of *FCM*, so that the summation of the memberships of each instance to all clusters belongs to the interval [0, 1], and interprets each row of U as a possibility distribution. Hence, in this approach u_{ik} models the typicality of each instance rather than its membership in the cluster.

However, similarly to partitional clustering approaches, *PCM* is sensitive to initialization, and in sometimes it generates coincident clusters. A variation of *PCM* is the *Asymmetric Graded Possibilistic C-Means (AGPCM)* [129] which replaces that assumption of *PCM* with a constraint that forces rows to sum up to *at most* 1; there is also a constraint on the lower value that depends on a parameter α.

The version of *PCM* presented [103] aims to minimize the following functional with respect to the membership matrix U and the codebook $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_c\}$:

$$J(U,Y) = \sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^m \|\mathbf{x}_i - \mathbf{y}_k\|^2 + \sum_{k=1}^{c}\eta_k \sum_{i=1}^{n}(1-u_{ik})^m \ , \tag{3.10}$$

Note that thanks to the penalty term, points with a high degree of typicality have high membership values to clusters, and points not very representative have low membership values in all the clusters.

The minimization of Eq. 3.10 leads to the following equations:

$$u_{ik} = \left[1 + \left(\frac{\|\mathbf{x}_i - \mathbf{y}_k\|^2}{\eta_k}\right)^{\frac{1}{m-1}}\right]^{-1} \qquad \mathbf{y}_k = \frac{\sum_{i=1}^{n}(u_{ik})^m \mathbf{x}_i}{\sum_{i=1}^{n}(u_{ik})^m} \ . \tag{3.11}$$

Finally, after the possibilistic clustering step using *PCM* in the spectral space, *PSM* finds the significant overlapping communities by assigning vertices to clusters in which they have a membership larger than the threshold ϖ (see Eq. 3.6) as we did before in *FSM* (see Sect.3.3).

3.6 The proposal of an ensemble of fuzzy spectral-possibilistic clustering paradigm

This section will propose a clustering ensemble paradigm and highlight its benefit to improve the clustering result. We use spectral and partitional clustering families in our proposal, however other clustering types are possible. The experimental evaluation of this proposal is discussed in Sect. 5.1.

This thesis employs ensemble clustering paradigms (see Sects. 3.7 and 3.8) for community detection. In [165] we proposed a clustering ensemble framework that adopted spectral clustering and several partitional clusterers such as K-Means, Fuzzy C-means, and the Asymmetric Graded Possibilistic C-Means [129]. Moreover, we highlight that the ensemble concept can be used to boost the classification tasks as we did in [24].

Clustering techniques differ in several respects, but one especially crucial area of diversity is the underlying hypothesis on what kind of aggregation should constitute a cluster. This a-priori model selection is in most cases implied in how the clustering procedure is specified.

Centroid-based clustering techniques [92, 11], for instance, use a prototypical point in the data space, usually, but not exclusively, defined as the barycenter of points weighted by memberships, to represent each cluster; membership of a point in a cluster is decided on a distance-based criterion. This results in clusters which, in crisp cases, are the tiles of a Voronoi tessellation, and are therefore convex; in fuzzy clustering, tile boundaries are fuzzy, but their shape is still convex.

A completely different road is taken by affinity-based techniques, where clustering is performed on the pairwise distance or similarity (affinity) matrix. Here the implied cluster model is that of a component whose points feature a mutual intracluster similarity that is larger than the similarity to points in other clusters. The use of local connectivity may yield very complex overall shapes.

Instances of the latter family range from the single-linkage or minimum spanning trees (MST) clustering [73] to density-based approaches [101]. However in the last decade spectral clustering [186], borrowing from spectral graph partitioning [30], has been undisputably the most prominent example, enjoying a wealth of publications [145, 173, 59].

These two clustering paradigms have their own strength and weaknesses. Unfortunately, the choice is an inherent, ill-defined model selection problem. This is a fundamental limitation of clustering, for which no theoretical approach can provide a thorough answer because of the amount of subjectivity involved.

This thesis explores the combined use of two different clustering paradigms and their combination by means of an ensemble technique. Mixing coefficients are computed on the basis of partition quality, so that the ensemble is automatically tuned so as to give more weight to the best-performing clustering method. To estimate the quality of partitions, we employ well-established

cluster validity indices appropriate to each individual clustering method.

Central clustering works with distances from a reference point acting as a cluster prototype. For metric data, this results in a Voronoi tessellation of the data space, so that each cluster is a Voronoi region (convex, possibly open - although not in the unnormalized "possibilistic" case –, boundary half-way between two centroids). Spectral clustering, on the other hand, can theoretically represent any cluster shape that is an arbitrary manifold in the data space, but is very sensitive to cluster separation.

In practice, when applied to metric data, as opposed to graph data, there are limitations due to non-sparse affinity matrices, especially when working with Euclidean data. The eigensystem is usually not partitionable in an unequivocal way when point density and cluster size differ across the data space by more than some threshold [140]. As a consequence, the method fails to differentiate clusters if they are joined by points with a density that is not sufficiently lower than the inner cluster density. Several ways to cope with this problem have been proposed in the literature, from completely heuristic to deeply grounded in theory [8, 61, 29].

This thesis includes one fuzzy and two crisp centroid-based methods (K-Means, fuzzy C-Means, and Asymmetric Graded Possibilistic C-Means), and 3 variants of spectral clustering.

As representatives of manifold-based cluster representation, we employ the Ng-Jordan-Weiss algorithm [145] (see Sect. 3.1) in three different variations. The variations are related to three different ways to overcome the problem of selecting the kernel parameter, σ_i. The first way, labeled SC(fix), simply uses a fixed value computed by trial and error, i.e., $\sigma_i = \sigma \; \forall i$. Adaptive σ_i is instead computed locally for each point $\mathbf{x}_i$ in two ways.

The computation is based on the ranked list of neighbors. The distances of each point from its top K neighbors (in the experiment we selected $K = 10$) are weighted with a coefficient that depends on the neighbor order, so that the nearest neighbor has the maximum weight, the second nearest a lower weight, and so on until the last neighbor considered. These weights are used to compute an aggregate weighted distance representative of the distribution, that is assumed as the value of σ_i.

The two techniques differ in the decay of the weights as a function of the neighbor's position in the ranked list. The first one, SC(var1), is based on a linear decay: the weight corresponding to any given rank ρ has weight $1 - \frac{\rho-1}{K-1}$. In the second variant, SC(var2), the decay is exponential, so the weight for the same neighbor rank ρ is $e^{\left(\frac{\rho-1}{K-1}-1\right)}$.

All the centroid-based methods considered compute centroids $\mathbf{y}_j$ as the weighted means of points $\mathbf{x}_l$ in each cluster, $\mathbf{y}_j = \sum_l U_{lj}\mathbf{x}_l$, for a suitably defined membership matrix U whose rows correspond to points and columns to clusters.

In the applications that motivate this study the number of data items is possibly very large; on the other hand, the number of base clustering methods is not high, with one partition per method, so

the search for the best match is not computationally demanding. Therefore in [165] we propose an ensemble partition matrix combination approach.

Let U_k^F be the membership matrix obtained from the k-th clustering from family F, where $F = \{C, S\}$ (C for Centroid-based, S for Spectral). The consensus clustering is given by the aggregate membership matrix

$$U = \sum_k \mu_F w_k U_k^F, \tag{3.12}$$

where μ_F is the mixing coefficient for family F, $\mu_C = 1 - \mu_S$, and w_k is the weight computed from the quality index obtained for the k-th clustering, $\sum_k w_k = 1$.

The mixing coefficients μ_F are computed from the quality indices $q(U)$ of all the clusterings in family F (U_k^F for one of the possible values of F), normalized to sum up to 1:

$$\mu'_C = \sum_k q(U_k^C)\ ,\ \ \mu'_S = \sum_k q(U_k^S), \tag{3.13}$$

and

$$\mu_C = \frac{\mu'_C}{\mu'_C + \mu'_S}\ ,\ \ \mu_S = \frac{\mu'_S}{\mu'_C + \mu'_S}, \tag{3.14}$$

where $q(U)$ is the quality index computed on a generic partition matrix U, in the experimental study chosen among those described above. Of course this formulation, there presented for just two families of clustering paradigms, is readily generalizable to more than two.

3.7 The Semantically Enriched Fuzzy c-means Spectral Modularity community detection method

Protein-protein interactions (*PPIs*) refer to physical contacts with molecular docking between proteins that occur in a cell or in a living organism in vivo. The interaction interface is intentional and evolved for a specific purpose distinct from totally generic functions such as protein production, and degradation [41].

In Sect. 3.3 we proposed the *FSM* community detection method for network analysis [125, 126] that estimates the number of communities k using the maximization of modularity procedure depicted by Newman and Girvan in [143] and then performs data clustering in a subspace spanned by the first k eigenvectors of the graph Laplacian matrix [186]. The method follows the spectral clustering approach described in [145], with the main difference consisting in using as a technique of clustering the Fuzzy C-Means [14] that makes possible to identify significant overlapping protein communities.

In [122] we employ a novel pre-processing hybrid technique able to exploit both the available quantitative and the semantic information ("semantic enrichment"). As we shall show, the proposed approach, that we call *Semantically Enriched Fuzzy c-means Spectral Modularity (SE-FSM) community detection* method (see Fig. 3.3), boosts the discriminating capabilities of the *FSM* method.

The *SE-FSM* community detection method infers the overlapping communities using the following steps:

1. The **Protein similarity using an ensemble of quantitative information:** Given a set of l proteins, we can obtain their **quantitative information** on their interactions from *STRING*[1] [187] that is a public on-line repository incorporating different evidence sources for both physical and functional PPIs. *STRING* stores interaction evaluations for each pair of proteins m,n in different spaces (or features), including *homology*, *co-expression*, *experimental results*, *knowledge bases*, and *text mining*. In addition, *STRING* contains a *combined interaction score* between any pair of proteins calculated as:

$$i_{mn} = 1 - \prod_{f=1}^{h} (1 - a_{f,mn}). \tag{3.15}$$

   This score is computed under the assumption of independence for the various sources, in a naive Bayesian fashion, and often has higher confidence than the individual sub-scores $a_{f,mn}$ [187].

   For each feature f we build a *connectivity matrix* (or *similarity matrix*) $A_f = [a_{f,mn}]$ [106, 178, 138]. Then we can combine the connectivity matrices obtaining a *consensus matrix* A with elements:

$$a_{mn} = \sqrt{\sum_{f=1}^{h} \left(a_{f,mn}\right)^2} \qquad f \in \{1,\ldots,h\},\ \ m,n \in \{0,\ldots,l\}, \tag{3.16}$$

   where h is the number of features. From A we can obtain the *quantitative ensemble similarity matrix* Q (see Fig.3.1) making use of a Gaussian kernel:

$$q_{mn} = 1 - e^{-a_{mn}} \qquad m,n \in \{0,\ldots,l\}. \tag{3.17}$$

2. The **Protein semantic similarity**: We gather the **semantic information** from many web repositories containing annotated information about biological processes, molecular func-

[1] http://string-db.org

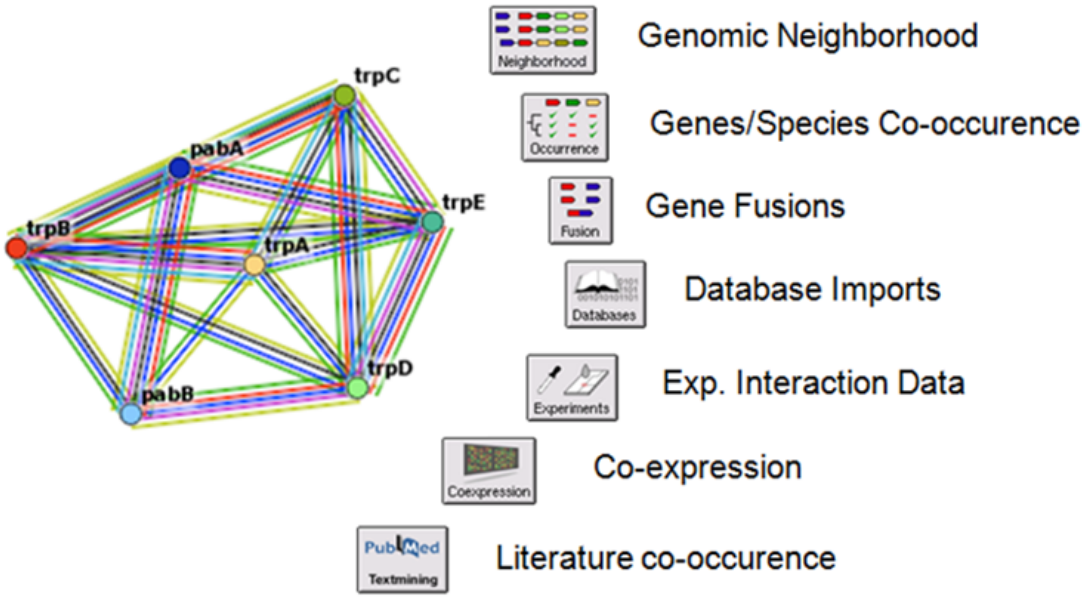


Figure 3.1: *Biological repositories (e.g., STRING) contains different quantitative information about interacting proteins* .

tions and cellular components including, *KEGG* pathway[2], *Reactome* pathway database[3], the pathway interaction database *(PID)* [4], and Gene Ontology *(GO)*[5].

Each Gene Ontology [4] term $t \in [1, p]$ has a set of annotated proteins related to it. The more semantically similar the gene function annotations between the interacting proteins, the more likely the interaction is physiologically relevant. For each protein we can build a binary valued indicator feature vector that refers to whether it contributes in any of the extracted biological terms or not, obtaining in this way a concurrence matrix $C = [c_{tm}]$. Then we measure the *semantic distance* d_{mn} between each pair m,n of analyzed proteins given by:

$$d_{mn} = \sum_{t=1}^{T} |c_{tm} - c_{tn}| \qquad m,n \in \{0,\ldots,l\}, \tag{3.18}$$

where T refers to the number of semantically enriched Gene ontology terms or pathways used. Then we obtain the *semantic similarity matrix* S (see Fig.3.2) whose elements are

[2]http://www.genome.jp/kegg/
[3]http://www.reactome.org/
[4]http://pid.nci.nih.gov/
[5]http://www.geneontology.org

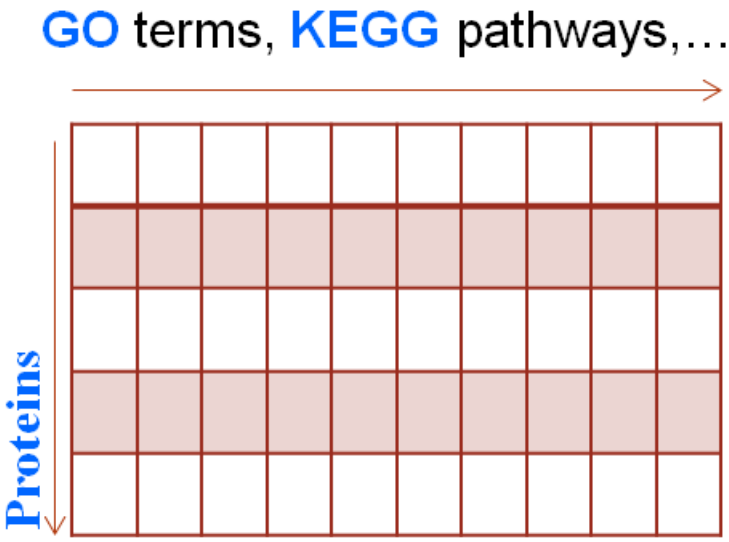


Figure 3.2: *Semantic co-association matrix construction.*

defined as:

$$s_{mn} = e^{\frac{-d_{mn}}{\nu}} \qquad m,n \in \{0,\ldots,l\}, \tag{3.19}$$

where ν is the dispersion parameter, it controls the width of the Gaussian "bell" and depends on the data distribution. There are many approaches to select the spread of the similarity function (ν). We select ν using histogram analysis of d_{mn} another possible choice is to tune the spread as done in [202].

3. Combine the quantitative ensemble similarity matrix Q and the semantic similarity matrix S in a *hybrid similarity matrix* (or *semantically enriched similarity matrix*) H using the *evidence accumulation* approach proposed by Fred and Jain in [66] as shown in Fig. 3.3.

4. Finally, *SE-FSM* applies the ***Fuzzy C-Means Spectral Clustering Modularity*** **(or FSM)** community detection (see, Sec.3.3) to Infers overlapping and semantically significant communities.

It is worth noting that the proposed spreadability cut measure can be used for estimating information spreadability of the network nodes as illustrated in Sec.3.4.

The proposed *SE-FSM* was applied in homo sapiens *PPIs* networks in [119], in particular *HIV-1* (see Sect. 5.3.1) and Leukemia (see Sect. 5.3.2).

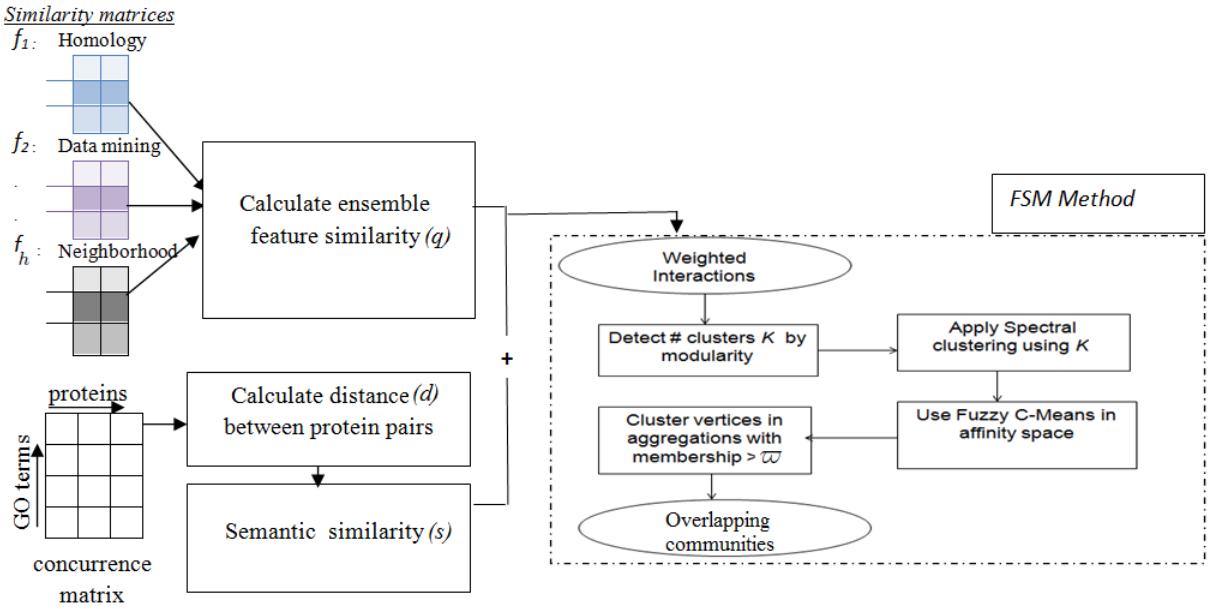


Figure 3.3: *The proposed Semantically Enriched Fuzzy c-means Spectral Modularity (SE-FSM) community detection method.*

3.8 Measuring similarity based on semantic clustering of semantic terms

Lord et al. [115] demonstrated the feasibility of using semantic similarity measures in a biological setting. In this study, the GO (semantic) similarity between two proteins was calculated as the semantic similarity of their annotated GO terms, and they found a strong correlation between GO similarity using annotations found in the UniProt/SwissProt database [6] [3] and their sequence similarity.

Information content based similarities are the most robust measures. Mainly, they are classified into two categories as shown in Fig. 3.4:

1. The annotation based measures consider the annotation of related semantic terms such as [156, 111, 170, 33].
2. The topology based measures consider the intrinsic topology of *GO* such as [136, 205, 190].

However, the performance of the different similarity measures was not uniform over the different aspects of GO, and it was not consistent with previous studies using different corpora either [23].

[6] http://www.uniprot.org/

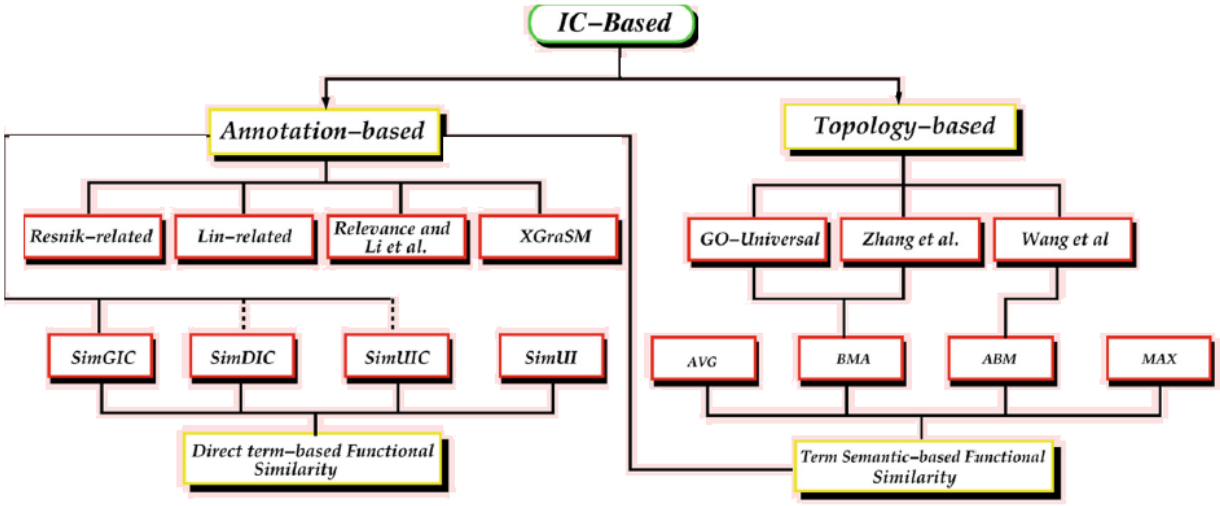


Figure 3.4: *Classification of information content based semantic similarity measures. (From [132])*

For example, Resnik's measure achieved the strongest correlation in the molecular function aspect and the weakest correlation in the biological process aspect.

One explanation for the lack of uniformity and consistency can be the significant number of protein pairs with high GO similarity and low sequence similarity [116]. This was expected, because proteins sharing a biological role do not necessarily have a similar sequence [40].

3.8.1 Annotation based semantic similarity

Graph based Similarity Measure (*GraSM*) [33, 33] and eXtended Graph based Similarity Measure [131] are from the most powerful annotation based semantic similarity measures exist in the literature [132, 130]. They are hybrid approaches in which features of parent and child terms of *GO* are taken into account.

They show a higher correlation between protein families and semantic similarity on all aspects of *GO* unlike other measures relying on the Least informative Common Ancestor (*LCA*) only [156, 111, 96] (see Sect. 2.7) this is due to:

- *LCA* based approaches account only for **one** of the interpretations of the compared terms *a* and *b* and may miss other concepts.
- Considering all terms in *GraSM* and *XGraSM* gives a more realistic representation of *GO* directed acyclic, while (*LCA*) based approaches wrongly consider *GO* as a hierarchical tree [115].

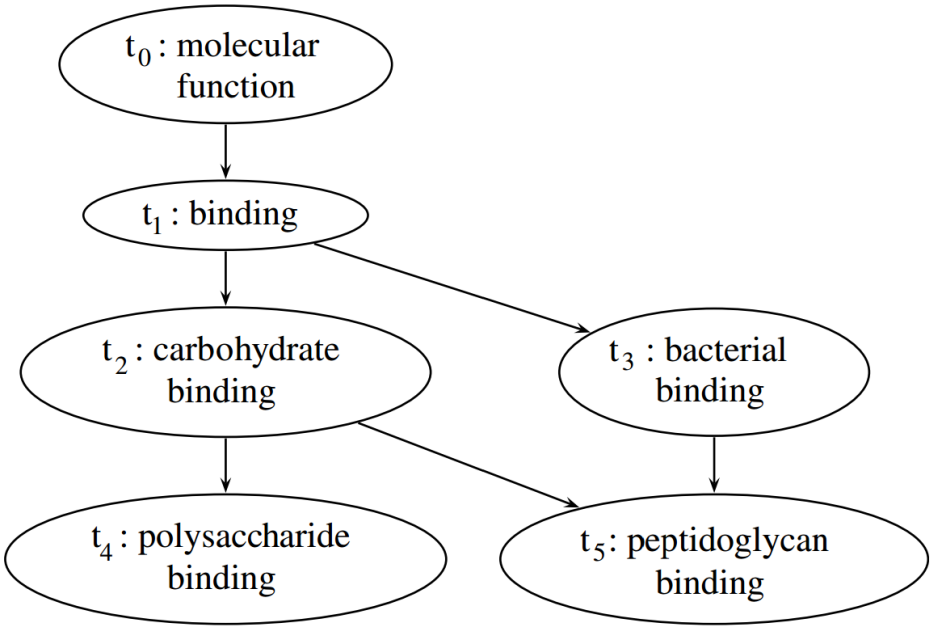


Figure 3.5: *Bacterial binding and carbohydrate binding are two disjunctive ancestors of peptidoglycan binding and polysaccharide binding. (From [33])*

In the rest of this Subsection, we will show the mechanism of *GraSM* and *XGraSM* measures as follows:

- **Graph based Similarity Measure** *GraSM* [33, 34] is an annotation based semantic similarity measure assumes that two common ancestors are **disjunctive** if there are independent paths from both ancestors to the concept (In *GO* the terms $a1$ and $a2$ are considered disjunctive ancestors of a term t if there is a path from $a1$ to t that does not contain $a2$ and there is a path from $a2$ to t that does not contain $a1$.

  For example, in the *GO* subgraph shown in Fig. 3.5, the terms bacterial binding ($t3$) and carbohydrate binding ($t2$) are two disjunctive ancestors of peptidoglycan binding ($t5$) and polysaccharide binding ($t4$) since there are two distinct paths from peptidoglycan binding ($t5$) to carbohydrate binding ($t2$) and binding ($t1$).

  GraSM finds all the common disjunctive ancestors ($CommonDisjAnc(a,b)$) of two concepts in a and b in *GO-DAG*, then *GraSM* defines a new Mutual information ($Mutual_G(a,b)$) for instance by calculating the average Information Content of $CommonDisjAnc(a,b)$ given by:

$$Mutual_G(a,b) = \overline{IC(t)|t \in CommonDisjAnc(a,b)}, \tag{3.20}$$

  where the Information Content $IC(t)$ is obtained using Eq. 2.45. The $Mutual_G(a,b)$ has a

worst-case performance $O(k^2)$, where k is the maximum number of common ancestors of two terms [33].

GraSM derives new semantic similarities such as $sim_{resnikG}$, sim_{linG}, and $sim_{j\&cG}$ instead of the original measures (see Eqs. 2.47, 2.48, 2.49) using $Mutual_G(a,b)$ instead of the original ($Mutual(a,b)$) (see Eq. 2.46).

Tab. 3.3 lists the *GraSM* algorithm [33], it can be summarized as:

1. Start by selecting the common ancestors of both concepts (line 1) and by initializing the list of common disjunctive ancestors as an empty list (line 2).
2. Select each common ancestor in descending order of information content (line 3).
3. For each selected ancestor, check if the ancestor is disjunctive to all the common disjunctive ancestors already selected (lines 4–7).
4. If the ancestor is disjunctive add it to the list of common disjunctive ancestors (line 9).
5. At the end, calculate the average of the information content of all the ancestors in the common disjunctive ancestors list (lines 12–16).

Note that the disjunctive ancestor checking operation ($DisjAnc(c,(a_1,a_2))$) called in (line 6) checks if a pair of ancestors ($a1$,$a2$) are disjunctive ancestors, the algorithm checks if the number of paths from $a1$ to c is larger than the multiplication of the number of paths from $a1$ to $a2$ and from $a2$ to c of a given concept c.

Also note that two disjunctive ancestors of a term represent two distinct interpretations of it.

However, finding the common disjunctive ancestors $CommonDisjAnc(a,b)$ between two GO terms makes the original *GraSM* approach computationally unattractive [131].

- **eXtended Graph based Similarity Measure)** (*XGraSM*) [131] extended the original *GraSM* by considering all informative common ancestors representing all interpretations in the graph structure ($ICA(a,b)$) between two terms a and b in *GO*.

  XGraSM uses the same procedure as *GraSM* (see Tab. 3.3). However, it uses A *XGraSM* based Mutual semantic information ($Mutual_{XG}(a,b)$) based on $ICA(a,b)$ instead of using $CommonDisjAnc(a,b)$ in *GraSM*. Hence *XGraSM* derives new semantic similarities such as $sim_{resnikXG}$, sim_{linXG}, and $sim_{j\&cXG}$.

  Note that *XGraSM* outperforms the original *GraSM* that considers only the disjunctive common ancestors ($CommonDisjAnc(a,b)$).

  In [132, 130] $sim_{resnikXG}$ together with *BMA* showed the best results among the annotation based approaches analyzed (see Sect. 2.7.5). We will use these settings in the experiments shown in Sect. 5.2.4.

Table 3.3: The Graph based Similarity Measure algorithm(*GraSM*) (From [33])

```
1: Anc = CommonAnc(c1, c2)
2: CommonDisjAnc = { }
3: for all a ∈ sortDescByIC(Anc) do
4:     isDisj=true
5:     for all cda ∈ CommonDisjAnc do
6:         isDisj = isDisj ∧ (DisjAnc(c1,(a,cda)) ∨ DisjAnc(c2,(a,cda)))
7:     end for
8:     if isDisj then
9:         addTo(CommonDisjAnc, a)
10:    end if
11: end for
12: shared = 0
13: for all cda ∈ CommonDisjAnc do
14:     shared + =IC(cda)
15: end for
16: return shared/sizeof(CommonDisjAnc)
```

The Disjunctive Ancestor checking operation ($DisjAnc(c,(a_1,a_2))$) in line #6 is:

```
1: Require IC(a1) ⩽ IC(a2)
2: nPaths = |Paths(a1, a2)|
3: nPaths1 = |Paths(a1, c)|
4: nPaths2 = |Paths(a2, c)|
5: return nPaths1 ⩾ nPaths × nPaths2
```

3.8.2 Topology based semantic similarity

Translating the biological content of a given GO term into a numeric value, called the semantic value or topological information, on the basis of its location in the GO-DAG, requires knowledge of the topological position characteristics of its immediate parents.

This leads to a recursive formula for measuring topological information of a given GO term, in which the child is expected to be more specific than its parents.

The more children a term has, the more specific its children are compared to that term, and the greater the biological difference. In addition, the more parents a term has, the greater the biological difference between this term and each of its parent terms.

The three separate ontologies, namely, molecular function (MF), biological process (BP), and cellular component (CC) with GO Ids GO: 0003674, GO: 0008150, and GO: 0005575 respectively, are roots for the complete ontology, the reference level, and are assumed to be biologically meaningless.

In the rest of this section the level of a term is considered to be the length of the longest path from the root down to that term in order to avoid a given term and its child having the same level.

Define the following:

- N_{GO} is the set of GO terms and links.
- $(a,b) \in L_{GO}$ represents the link or association between a given parent a and its child b, and the level of the link (a, b) is the level of its source node a.
- $[a,b] \in N_{GO}$ indicates that the level of term a is lower than that of b.

The Topological information IC_T of a given term $z \in N_GO$ is given by:

$$IC_T(z) = -ln(\mu(z)), \tag{3.21}$$

where $\mu(z)$ is a s a topological position characteristic of z, recursively obtained using its parents gathered in the set $p_z = \{a : (a,z) \in L_GO\}$ and it is given by:

$$\mu(z) = \begin{cases} 1 & \text{if z is a root,} \\ \prod_{a \in p_z} \frac{\mu(a)}{c_a} & \text{otherwise,} \end{cases} \tag{3.22}$$

where c_a are the number of children of parent term a.

GO-universal [136], is a topology based information content approach given by:

$$sim_{GOu}(a,b) = \frac{IC_T(a,b)}{max\{IC_T(a), IC_T(b)\}}, \tag{3.23}$$

where $IC_T(a,b) = -ln\mu(a,b)$. $sim_{GOu}(a,b)$ induces a distance ($d_{GOu}(a,b)$) or a metric based on information theory that is given by:

$$d_{GOu}(a,b) = 1 - sim_{GOu}(a,b). \tag{3.24}$$

The more topological information two concepts share, the smaller their distance and the more similar they are. Moreover, $sim_{GOu}(a,b)$ emphasizes the importance of the shared *GO* terms by giving more weight to the shared ancestors corrected by the maximum topological information, and thus measuring how similar each *GO* term is to the other.

Thus, for two *GO* terms sharing less informative ancestors the distance is greater and the similarity is smaller, while for two *GO* terms sharing more informative ancestors, they are closer and their similarity is higher.

In [132, 130] sim_{GOu} together with *BMA* mixing strategy (see Sect. 2.7.5) showed the best results among the topology based approaches analyzed. We use $GO - universal$ and *BMA* in *SC-FSM* as well as the experiments shown in Sect. 5.2.4.

3.8.3 The Semantic Clustering Fuzzy Spectral Modularity community detection method (*SC-FSM*)

The Semantic Clustering Fuzzy Spectral Modularity community detection method (*SC-FSM*) measures semantic similarity based on both annotation and topology basis by employing an ensemble of *XGraSM* (see Sect. 3.8.1) and *GO-universal* (see Sect. 3.8.2) similarities to characterize the analyzed proteins.

SC-FSM employs a semantic based clustering in the proposed *SE-FSM* (see Sect. 3.7). We performed a semantic enrichment of data using the well known annotation technique based on *XGraSM* method [34, 33] and Resnik similarity, and we built a consensus similarity [66] combining this metric with the topological similarity based GO-universal approach [136] as shown in Fig. 3.6.

The proposed approach infers significant interaction communities in the spectral space.

The *XGraSM* and *GO-universal* semantic similarity measures were obtained using: Proteins interactions and ontology [7], and *IT-GOM*: Integrated Tool for IC-based *GO* Semantic Similarity Measures [8] publicly available.

Note, many studies showed that analyzing the correlation between semantic similarity and other biological aspects (or dimensions) such as protein pathways, protein complex, protein families

[7] http://www.lasige.di.fc.ul.pt/webtools/proteinon/
[8] http://www.cbio.uct.ac.za/ITGOM/tools/itgom.php

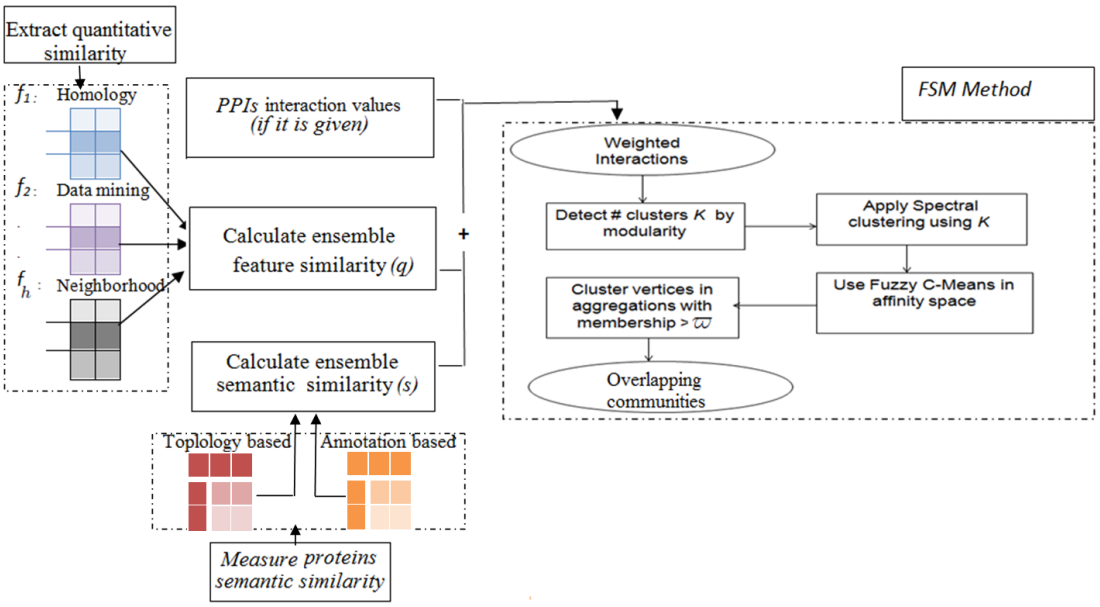


Figure 3.6: The proposed Semantic Clustering Fuzzy Spectral Modularity community detection method (*SC-FSM*).

(*Pfam*), and gene expression are much more interesting to characterize the semantic relations than sequence similarity in many biological scenarios (see Sect. 2.8).

Moreover, the information obtained from these aspects may not be uniform (i.e, in some biological cases some dimensions may contain more interesting information than other), therefore, relying on one dimension only may not be sufficient to infer the intrinsic relations between the biological entities (like, proteins) exist in the network.

Hence, *SC-FSM* as well as *SE-FSM* rely on building an ensemble of this information when measuring the semantic similarity. This helps to infer significant biological results as we will experimentally show in Chapter 5.

Chapter 4

Evaluating the proposed methods

This Chapter presents a comparison of 26 community detection methods (see Sect.2.4 and [195, 141, 65]) including *KSM*, *FSM*, and *PSM* that were proposed in Chapter 3 using synthetic and real world networks used in the literature for testing the quality of community detection methods. The comparison is performed in terms of quality, running time, and similarity to the known ground truth. For comparing the partitions, we solved the general correspondence problem (see Sect. 2.10) using Hungarian algorithm [105].

The methods evaluation is estimated using the following measures:

- Validity indices illustrated in Sect. 2.5 and can be found in detail in [84, 85, 165, 160].
- Partition similarity measures that evaluate the detected partition agreement with the ground truth, such as the normalized mutual information (*NMI*) and the generalized normalized mutual information [107, 65] that supports overlapping partitions, and *Omega* index [75] that is a fuzzy version of *Rand* index introduced in [154].
- Partition quality measures, such as modularity (see Sect. 2.1 and [142, 143]) and some of its variations [146, 195].
- Community goodness measures that depend on the graph topology criteria such as communities cohesion, separation, and cut ratio [197].

The software was developed in Matlab R2009b$^{(C)}$ under Windows 7$^{(C)}$ 32 bits. The experiments were performed on a laptop with 2.00 GHz dual-core processor and 3.25 GB of RAM.

This chapter is organized as follows: Sect. 4.1 shows a classification of the related community detection methods used, Sect. 4.2 evaluates both the methods quality and complexity using the

generated well known Girvan Newman benchmark, while Sect. 4.3 evaluate the proposed methods on real data taken from Zachary karate club benchmark, and Sect. 4.4 shows the communities obtained from the Dolphin benchmark.

4.1 Methods used in the experiments

The analyzed 26 methods can be mainly classified into the following categories:

- Divisive based (D).
- Hierarchical based (H).
- Spectral based (E).
- Fuzzy based (G).
- Quality based (Q).
- Vertex based (V).
- Random walk based (R).
- Topology based (L).

Tab. 4.1list the principle, time complexity and type of the most interesting crisp community detection methods among them, while the overlapping methods are listed in Tab. 4.2.

4.2 Girvan Newman benchmark

Deciding which of the methods listed in the previous section is reliable and shall be used in applications is a tricky question, as the definition of a community embedded in the inference mechanism can differ from a method to another.

Nevertheless, there has been de facto accept a simple network model, the planted ℓ-partition model [31], which is often used in the literature in various versions. In this model one "plants" a partition, consisting of a certain number of groups of nodes. Each node has a probability p_{in} of being connected to nodes of its group and a probability p_{out} of being connected to nodes of different groups. As long as $p_{in} > p_{out}$ the groups are communities, whereas when $p_{in} \leq p_{out}$ the network is essentially a random graph, without community structure.

Table 4.1: Comparison of crisp community detection methods based on maximizing quality (Q), topology (L) and random walk (R).

Algorithm	Complexity	Type	Reference
Betweenness based	$O(m^2n)$	D,H,Q	[Newman et al. 2002, 2004]
• Based on progressive removal of edges. • No quantitative evaluation to resultant communities.			
Blondel	$O(m)$	H,Q	[Blondel et al. 2008]
• Iteratively attempt optimize modularity. • Scalable, but it is more limited by storage demands than computational time. • May not find good optima.			
Multiresolution	$O(n^2(log(n))$	Q	[Duch & Arenas 2005]
• Attempt overcoming modularity resolution limit. • Use tunable scale parameter			
Leading eigen vector	$O(n^2(log(n))$	Q, E	[Newman et al. 2006]
• Maximize modularity using eigenvalues and eigenvectors of adjacency matrix. • Reasonable for bisection graphs, otherwise not good.			
Simulated annealing		Q	[Giumera et al. 2004, 2005]
• Optimize cost function (modularity) until reaching global optima. • Optimization perform local random movement of a node to a different community. • Movement it increases modularity ? • *Slower than previous approaches.*			
FLM	$O(m^3)n$	V	[Fortunato et al. 2004]
• Edge clustering coefficient based. • It is much slower than the algorithm of Newman Girvan. • Edges are removed according to decreasing values of information centrality.			
Link clustering	$O(m^4/n^2)$	D, V	[Radicchi et al. 2004]
• Edge clustering coefficient based approach, . • May fail to identify bridges between communities when the graph has few cycles.			
RN	$O(m^\beta logn)$, $\beta \sim 1.3$	Q	[Rohovde et al. 2009]
• Minimizes the Hamiltonian of a local objective function (the absolute Potts model).			
MCL	$O(nk^2)$, $k<n$	R	[Dongen, 2000]
• Markov Clustering is based on the probability of random walks remaining for a long time in a dense community. • It uses iterative expansion and inflation of network with largest elements k in inflation.			
Infomap	$O(m)$	R	[Rosvall et al. 2008]
• *Maps* of random walks finds communities based on the compression of the description length of the average path of a random walker over the network.			
Donetti	$O(n^3)$	E	[Donetti&Munoz 2004, 2005]
• Based on spectral partitioning using eigen values and graph laplacian.			

Table 4.2: Comparison of overlapping community detection methods based on divisive(D), topology (L) and vertex splitting (V) and fuzzy membership (G).

Algorithm	Complexity	Type	Reference
COGNA	$O(m^3)$	D, H, Q	[Gregory, 2007]
• Expands Newman Girvan approach to support overlap. • Split vertices among clusters if their betweenness exceeds maximum edge betweenness.			
Cfinder (Clique)	$O(exp(n))$	L, V	[Palla et al. 2005]
• Locally search approach depends on network topology (connected components) • based on finding largest cohesive subgraphs explored by rolling k-cliques in the network • For low k values results are reasonable • Robust with network evolution (insertion or removal of links).			
LFM	$O(n^2)$	L	[Lancichinetti et al. 2009]
• Depend s on expanding a random seed to form communities using a fitness function. • The fitness function based on internal and external degree of the community. • Depends on a resolution parameter.			
OSLOM	$O(n^2)$	L, H	[Lancichinetti et al. 2011]
• *Order Statistics Local Optimization Method,* identifies significant communities with respect to a Null model similar to modularity.			
Nepusz	$O(n^2kh)$	G, Q	[Nepusz et al. 2008]
• Based on optimization using simulated annealing. • It uses predefined *weight* and *prior* similarity between nodes, . • It increases the number of communities *(k)* until the structure does not improve. • Using maximization of modified fuzzy modularity based on nodes belonging to communities and bridgeness, where h is the number of iterations for convergence.			
Zhang	$O(mkh+nk^2h+k^3h)$ $+O(nk^2)$	G, Q, E	[Zhang et al. 2007]
• Based on the spectral clustering of [Newman 2006; White and Smyth 2005]. • Requires specifying number of communities k • The top $k-1$ eigenvectors are computed using Fuzzy C-Means (FCM). • Both detection accuracy and computation efficiency rely on the user-specified value k.			

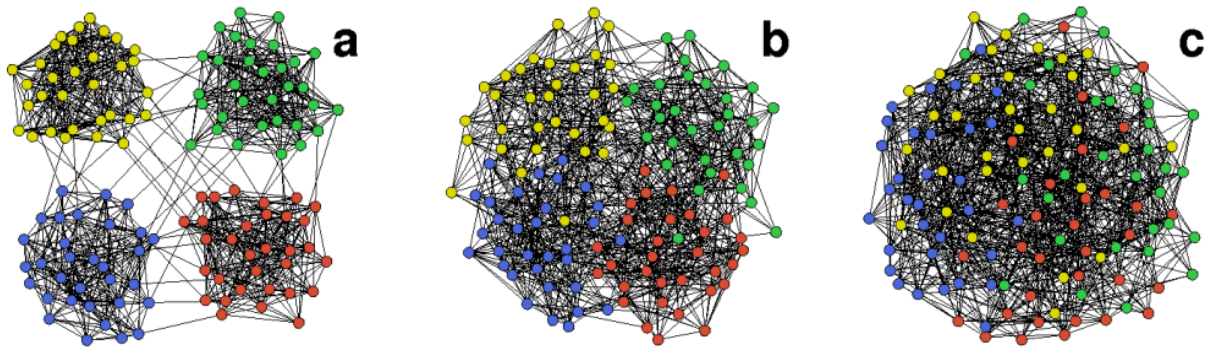


Figure 4.1: Benchmark of Girvan and Newman. (a) k_{in} = 15, k_{in} = 11 (b) and k_{in} = 8 (c). In (c) the four groups are hardly visible. (From [80])

Girvan and Newman (*GN* benchmark) [71, 65] proposed a version of ℓ-partition model (see Fig. 4.1). Here the graph consists of 128 nodes (N), each with expected degree 16, which are divided into four groups (ℓ) of 32 (g) nodes. The *GN* benchmark is regularly used to test algorithms for community detection. Indeed, algorithms can be compared based on their performance on this benchmark. However, the GN benchmark has two drawbacks: 1) all nodes have the same expected degree; 2) all communities have equal size. These features are unrealistic, as complex networks are known to be characterized by heterogeneous distributions of degree and community sizes.

This implies that $p_{in}+3p_{out} \simeq \frac{1}{2}$, so the probabilities p_{in} and p_{out} are not independent parameters. Hence the internal degree is given by $k_{in} = p_{in}(g-1) = 31p_{in}$ and the external degree k_out is given by: $k_{out} = p_{out}g(\ell-1) = 96p_{out}$ for evaluating a community detection method one can increase k_{out} (reduce the strength of community structure), then regenerate the benchmark and check the method accuracy compared to the generated ground truth. The literature reports that most of the methods degrades when k_{out} approaches 6 and may fail at $k_{out} = 8$ [65] due to the weakness of community structure. In the experiments we will present in this section we varied k_{out} from 1 to 10 (see Figs. 4.3, 4.4, 4.5, 4.6).

In Figs. 4.3 evaluates the following partition similarity measures: the fuzzy Omega measure, the overlapping normalized mutual information (overlapping *NMI*, it is a fuzzy version of the normalized mutual information widely usually used in the literature), the relative error, the fuzzy *Rand* index, and the *Jaccard* index [195, 65, 197, 79].

These partition similarity measure requires knowing the benchmark in advance and estimate its agreement to the detected partition by the community detection method. The proposed *PSM* as well as *FSM* showed high partition similarity values (generally approaching 1 when *kout*<6) compared to the other methods.

Note, most of the community detection method give reasonable community and as a consequence

obtain a high partition similarity in GN benchmark when *kout* is less than 6. This is well known characteristic of *GN* benchmark in the literature [65] because the community structure almost becomes vague at this threshold. As long as *kout* increase the community structure becomes weaker and few methods can infer the communities in this settings among them *PSM* and *FSM* that reported high values compared to other methods even if when increasing *kout*.

Fig. 4.4 aims at finding the best methods among the analyzed 26 methods that agree with *GN* benchmarks in node assignment. Note, we already know the correct assignment in the generated *GN* benchmark model in advance (benchmark is known).

For this task, After applying the community detection methods, we plotted the node assignments (i.e., membership label or community index) using all methods. Note we have 128 nodes in GN benchmark and we test *kout* in range [1,10], hence we analyze 128×11 observation shown in X-axis of Figs. 4.4. After that, we iteratively removed the methods that vary from the benchmark.

Many community detection methods obtained noisy results varying from the GN benchmark as shown in Fig.4.4(a). While others inferred node assignment similar to the *GN* benchmark as depicted in Fig.4.4(b).

The Community detection methods that agreed with the benchmark node assignment are:

1. The crisp algorithms: *KSM*, Global density (*GloDens*), Local density (*LocDens*), and *Infomap*.
2. The overlapping algorithms: *Copra*, *PSM*, and *FSM*.

Moreover, It is worth noting that *PSM* result is from the top most results matching benchmark as depicted in Fig.4.4(c).

We measure the modularity and other quality measures such as stability, node membership, global density, local density and distance based shown in Figs. 4.5 that estimate the quality or strength of the detected partition without the need of knowing the benchmark (hence it is usually used to evaluate the community detection algorithms when applied in real world networks (e.g., social, biological networks) where the benchmark is usually unknown). The proposed *PSM* as well as *FSM* showed high quality values compared to the other methods. Moreover *KSM* showed high quality values compared to its crisp peers.

In Figs. 4.6 we show the running time and the number of communities detected by the algorithms when varying *kout*. Note, *PSM*, and *FSM* are unlike other methods that misidentify the number of communities such as *Newman Leading eigen vector* method. Moreover, *PSM*, and *FSM* are more deterministic than these methods (i.e, detect similar results with multiple runs) and the proposed measures are robust as they identify communities accurately even when varying some parameters or intrinsic characteristics such as *kout*. Moreover, they detect the communities in a reasonable time and outperform many overlapping methods such as *COPRA* [76], *COGNA* [78], and *Cilque* (it has exponential complexity) that failed to work in many scenarios.

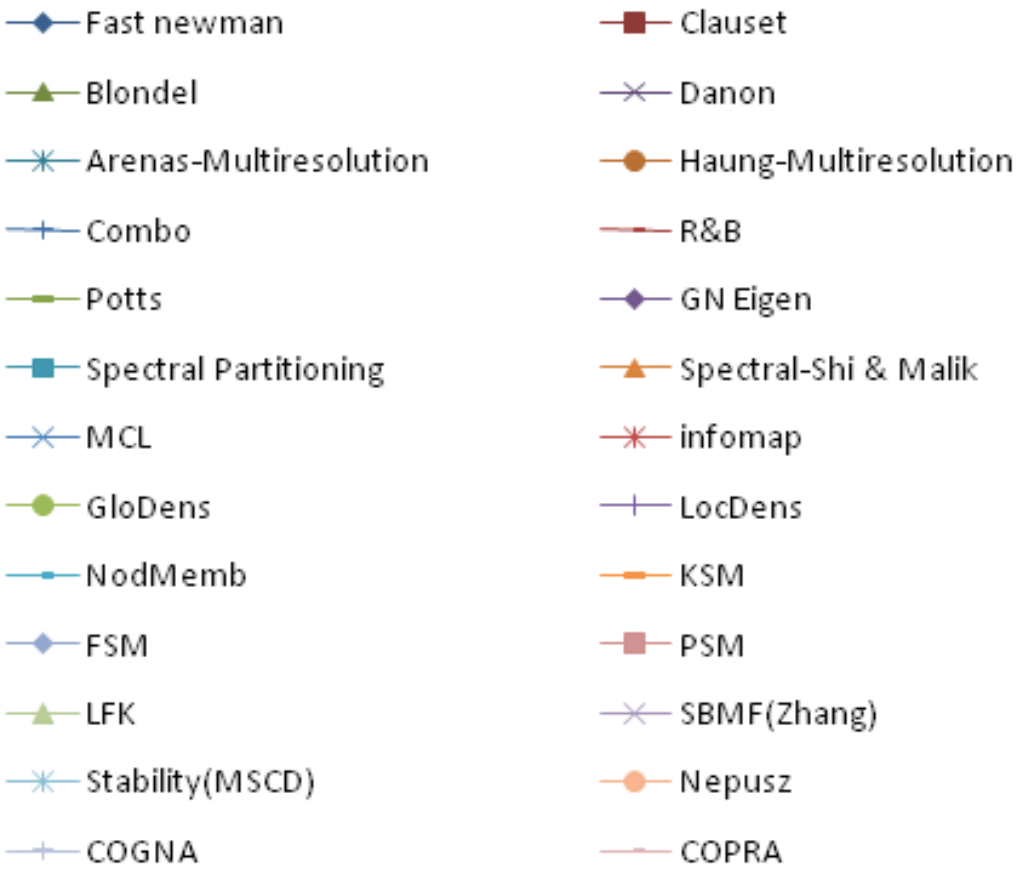


Figure 4.2: The 26 community detection methods compared using Girvan Newman benchmark.

We highlight also that it is more accurate than *Nepusz*[141] that may miss some nodes and fail to assign them to their correct community (it considers them outliers) while they are not, and the proposed methods get a reasonable number of communities by using modularity maximization in spectral space unlike the divisive approaches like *COPRA* or *Newman leading eigenvector*.

We extended our study on more complex networks using the ***LFR*** benchmark [107, 65] known in the literature. We tested the community detection methods on graphs of size 5000 and even 10000 nodes. We found that some methods halted specially the fuzzy methods such as *COGNA* and the methods that use a scaling parameters such as mutliresolution based methods (like, Arenas and Haung) or *POTTS* based methods [195]. Moreover, the crisp methods like *Combo* [175] for instance could not detect the fuzziness in the network, while *FSM*, and *PSM* obtained fuzzy communities in a reasonable time (in average 4 minutes for 10000 nodes network).

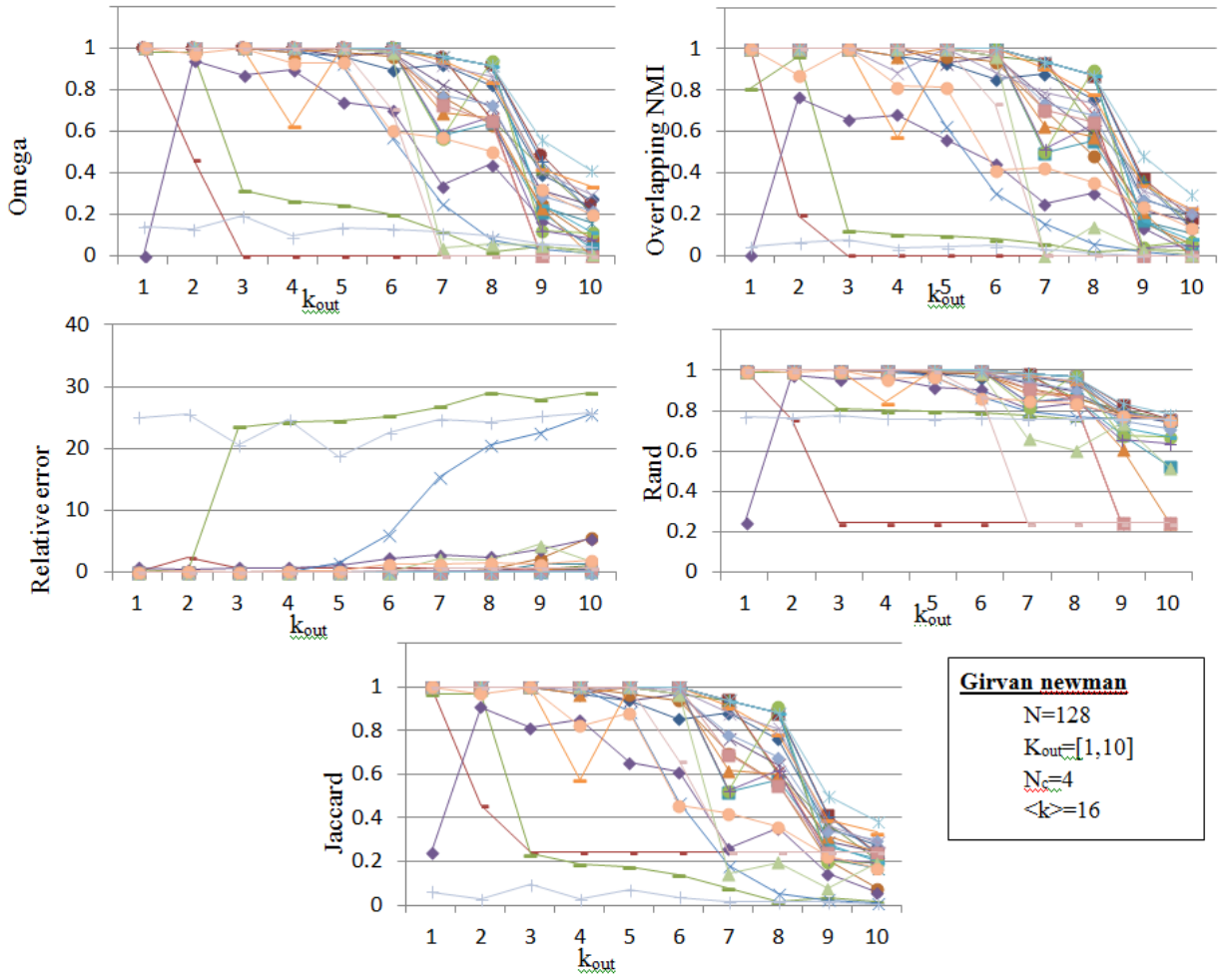


Figure 4.3: Evaluating the similarity between partitions of the 26 methods used and the ground truth of Girvan Newman benchmark using omega, overlapping normalized mutual information (overlapping NMI), relative error, Rand, and Jaccard indices.

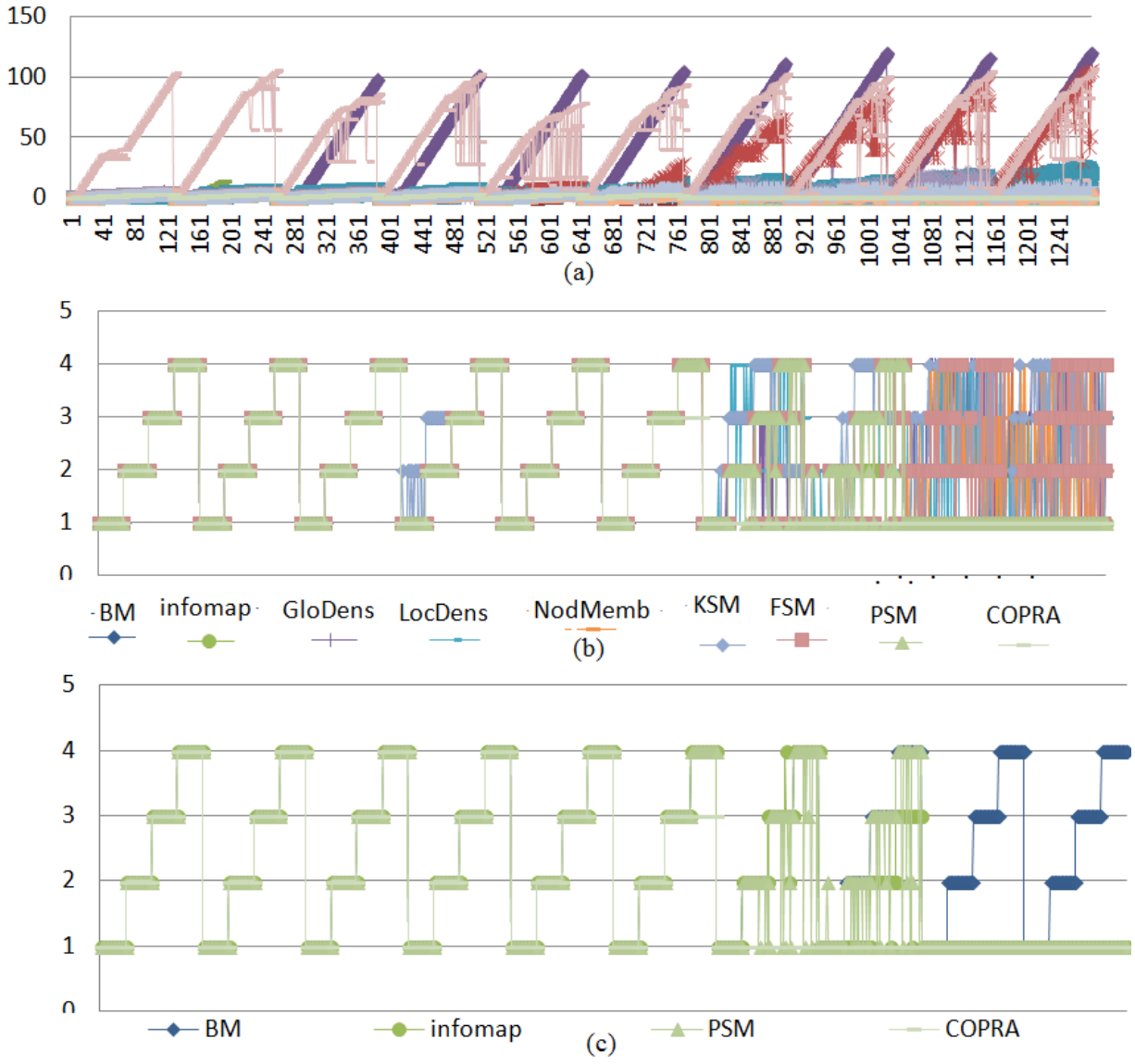


Figure 4.4: The distribution of node memberships for each method compared to the ground truth (BM). (a) shows the 26 methods used, (b) shows methods that obtained number of clusters N_c = 4, and (c) shows *PSM*.

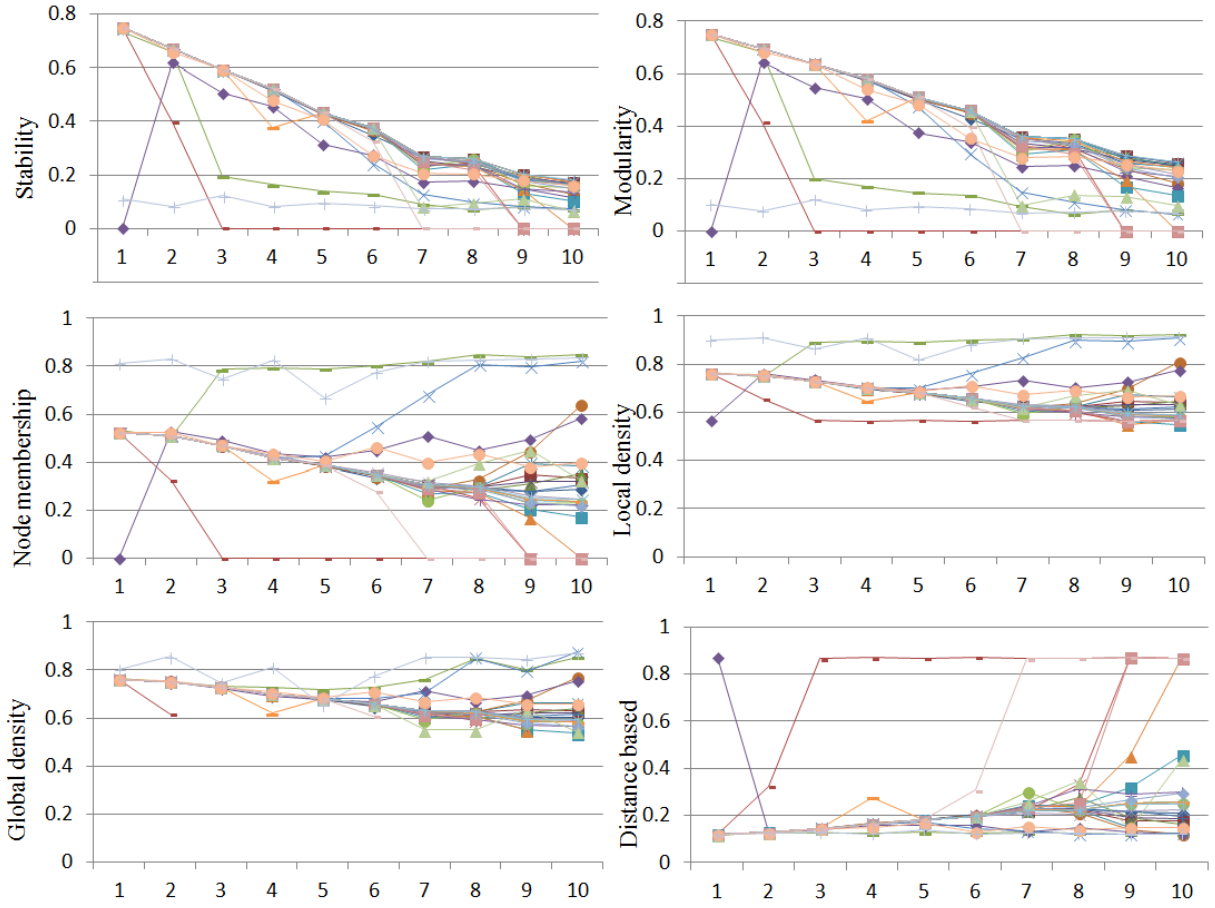


Figure 4.5: Evaluating the methods partition quality on Girvan Newman benchmark using stability, modularity, node membership, local density, global density, and distance based quality measures.

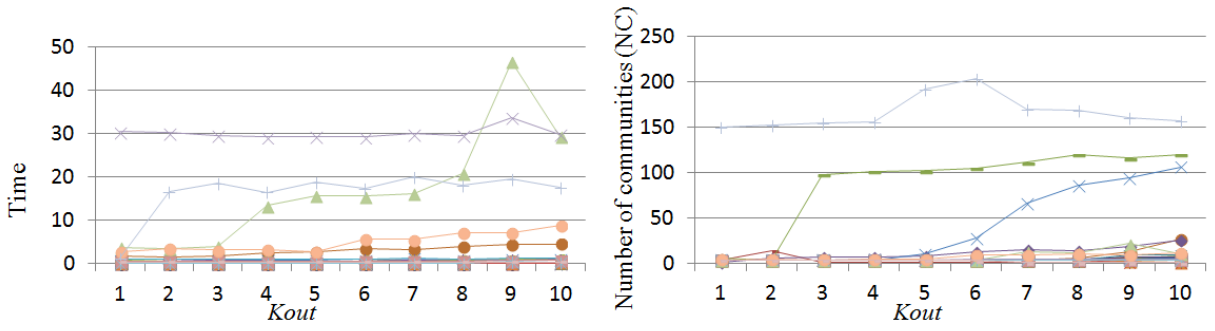


Figure 4.6: Time (left) and number of communities (right) of the analyzed methods by varying k_{out} parameter.

Table 4.3: Characteristics of analyzed datasets. For each dataset we report the number of nodes (n), the number of edges (m), the number of communities estimated as the value maximizing the Newman & Girvan modularity [143] (see Fig. 5.7), and the density (d).

	n	m	C	d
Zachary	34	78	2	0.1390

4.3 Evaluating community discovery in Zachary karate club benchmark

The first data set we present is the Karate Club Social Network (see Tab 4.3) studied by Zachary in [200]. After an argument between the club's administrator and the club's instructor, the network of club members split into two parties. This fighting ended in the instructor established his own club and taking about half of the original club with him. Zachary's paper [200] reports the (binary) social relationships among the 34 members of the karate club, and their club after fission (see "ground truth" in Tab. 4.4). We used this data set as a benchmark for validating the methods we listed in the previous section.

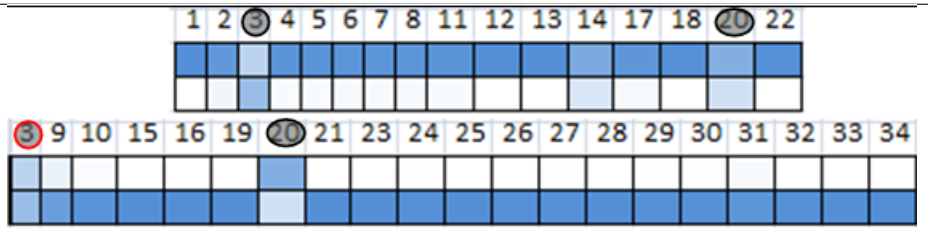
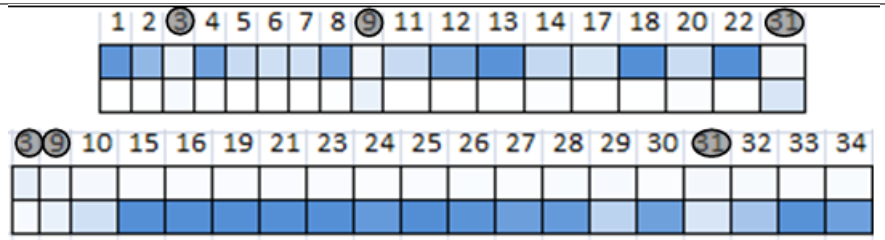
The evaluation of Newman & Girvan modularity [143] versus the number of clusters (or communities) for the Zachary Karate Network is shown in Fig. 5.7. In Tab. 4.3 we report the optimal value of the number of communities, that holds 2. The density of the network given by $d = \frac{2m}{n(n-1)}$ is 0.1390.

In Tab. 4.4 we report the results of our analysis obtained using the methods for community detection proposed by Blondel [15], Newman [144], and the three versions method using Spectral Modularity methods we illustrated in Sect. 3.3.

The first row of data in Tab. 4.4 concerns the "Ground truth" given in the paper by Zachary [200]: after fighting the two clubs contained respectively 16 and 18 members. In the following rows, for each applied method we give the composition of detected communities. Member numbers with font in white color and background in black color are wrongly classified nodes. In case of fuzzy methods nodes with membership vertices bounded with circles are fuzzy nodes exist in more than one community, while red nodes are disagreement with the ground truth (errors). Moreover, we report the error ratio, normalized mutual information, and modularity considering possible overlapping communities between each method and the ground truth in Tab. 4.4.

Among crisp models *KSM* obtained the most accurate results as they varied only in clustering node#3 from the ground truth of Zachary benchmark [200]. Note that the following fuzzy methods found that node#3 has a fuzzy characteristics and lies between the two communities, hence may be clustered in any of them. Blondel method identified instead 12 communities but most

Table 4.4: Comparing results of Blondel, Newman eigenvector, *KSM*, *FSM*, and *PSM* appraches on Zachary karate club benchmark.

	Communities list	C	error
Ground truth	C1={1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22} C2={9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34}	2	0
Blondel	C1={1,2,3,5,10,11,12,13,14,18,20,22}, C2={15,16,19,21,23,27,34}, C3={4,8}, C4={6,17}, C5={24,26}, C6={30,33}, C7={25,32}, C8={7}, C9={9}, C10={28}, C11={29}, C12={31}	12	$\frac{16}{34}$
Newman eigenvector	C1={1,5,6,7,11,12,17} C2={9,10,15,16,19,21,23,27,30,31,33,34} C3={24,25,26,28,29,32} C4={2,18,20,22} C5={3,4,8,13,14}	5	$\frac{15}{34}$
KSM	C1={1,2,4,5,6,7,8,11,12,13,14,17,18,20,22} C2={3,9,10,15,16,19 ,21,23,24,25,26,27,28,29,30,31,32,33,34}	2	$\frac{1}{34}$
FSM		2	$\frac{1}{34}$
PSM		2	0

of them are singletons or are composed of two proteins only. This result highlight the drawback of this hierarchical approach, that iteratively divides and separates the network nodes without giving a clear stopping criterion for the level of decomposition. Newman eigenvector approach detected five communities while as we mentioned the ground truth are only two. We observe that the result of this method on Zachary benchmark could match the ground truth if we combine $C5$ with $C1$, $C3$ with $C2$, and $C4$ with $C1$.

We highlight that the variation of membership values in *PSM* could discriminate the network bridge nodes. We call a nodes as bridge if it is a outlier or a fuzzy node resides between the communities, the removal of these nodes, significantly reduces the cut size, hence maximizes the clustering separation objective. We demonstrate this advantage of our approach using Zachary benchmark by observing the graph cut improvement after progressive removal of the identified bridge nodes based on choice of *PSM* threshold parameter.

We conclude that *PSM* is more robust in identifying the fuzzy and bridge nodes of the network than *FSM*. In addition to inferring nodes discovered by *FSM*, it detected other significant bridge nodes living in considerable topological position in the network. Moreover, the norm value of *PSM* could discover the outlier nodes.

In Tab. 4.4, *KSM* detected fuzziness of nodes#3, 20 with higher membership in $C1$, while *PSM* assigned higher membership to node#3 in $C2$ matching the ground truth results. Moreover, it detected the bridge nodes lies in considerable fraction between communities or outliers such as nodes #3, 9, 14, 31, 32, 20. Our method found nodes # 17, 5, 6, 7, 10, 11 bridge as well, we observed that they have a higher tendency to form a new cluster together.

4.4 Evaluating community discovery in dolphin benchmark

The dolphin benchmark [117, 118] is an ecological study showing that dolphins live and travel together in groups (see Fig. 4.8) by analyzing the interaction among a sample of 61 interacting dolphins. Fig. 4.9 shows the obtained communities by applying *PSM* on the dolphin benchmark. Females are labelled by circles. Dolphin #sn89 has fuzzy membership to both detected communities (aggregations) labeled by black and white. *PSM* results exactly agree with the ground truth reported in [117, 118].

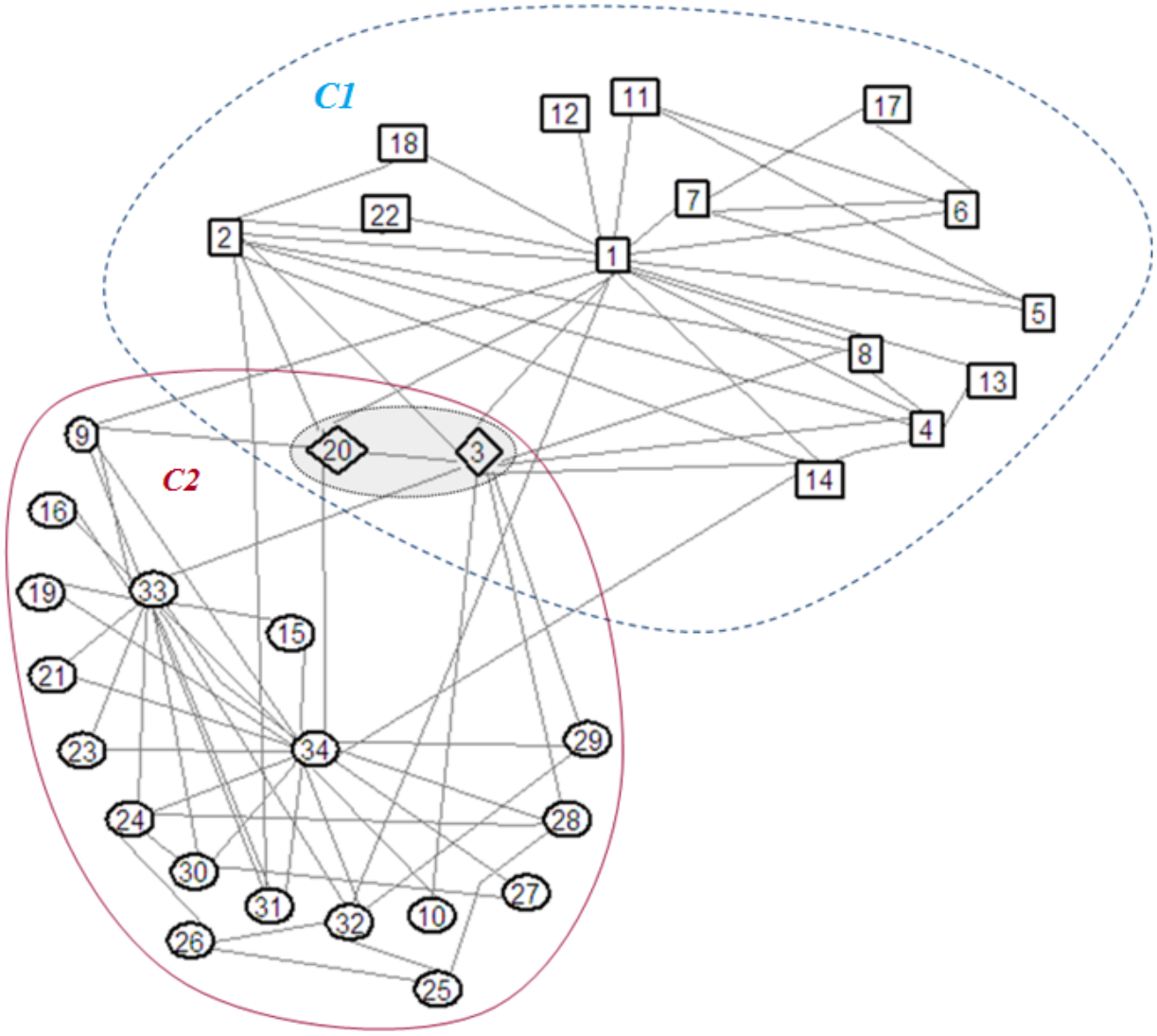


Figure 4.7: Two communities identified by *FSM* on Zachary benchmark. Nodes in gray boundary identified as fuzzy having the following memberships for C1 (the dashed blue border) and C2 (the red border) respectively node #3 (0.4, 0.6), node #20 (0.72, 0.28).

Figure 4.8: Dolphin ecological environment. (From [117])

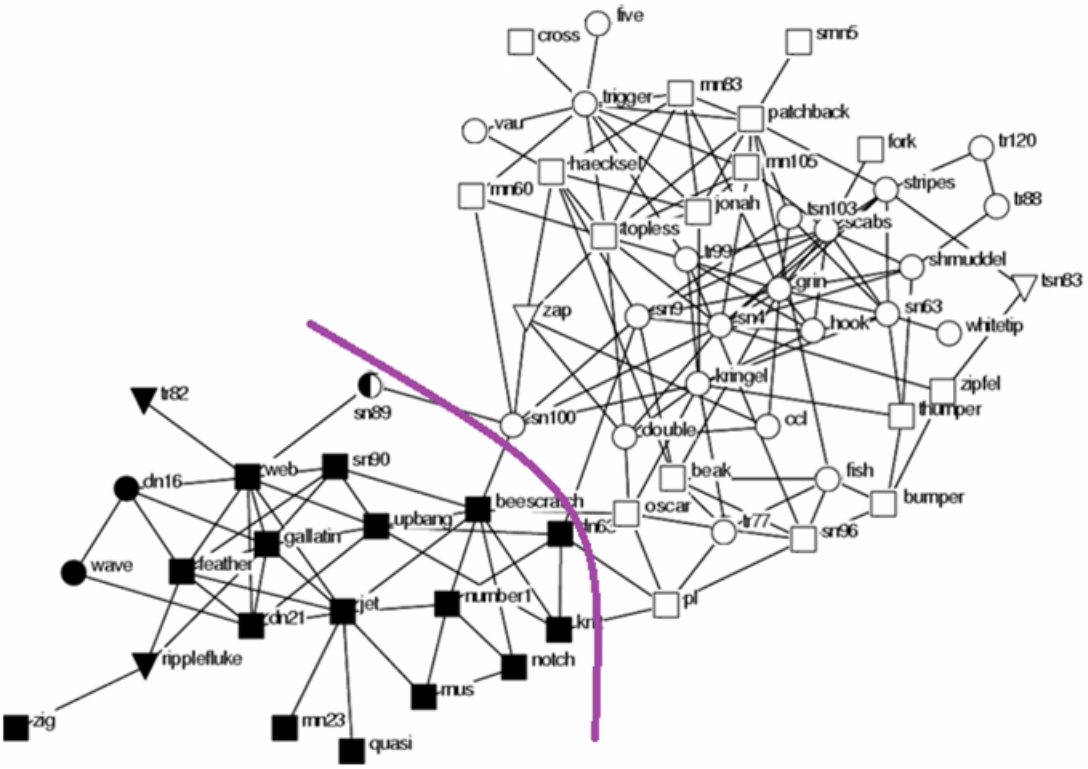


Figure 4.9: Two communities identified by *FSM* on Dolphin benchmark network. (From [117])

Chapter 5

Applications

The software used in the follwoing applications was developed in Matlab R2009b$^{(C)}$ under Windows 7$^{(C)}$ 32 bit. The experiments were performed on a laptop with 2.00 GHz dual-core processor and 3.25 GB of RAM.

5.1 Experimental analysis of fuzzy spectral-possibilistic clustering ensemble paradigm

This section explores the combined use of two different clustering paradigms and their combination by means of an ensemble technique. Mixing coefficients are computed on the basis of partition quality, so that the ensemble is automatically tuned so as to give more weight to the best-performing (in terms of the selected quality indices) clustering method.

A fundamental limitation of the data clustering task is that it has an inherent, ill-defined model selection problem: the choice of a clustering technique also implies some a-priori decision on cluster geometry. This section explores two different clustering paradigms and their combination by means of an ensemble technique. Mixing coefficients are automatically computed on the basis of partition quality, so as to give more weight to the best-performing clustering method.

5.1.1 Experimental setup

For the experimental verification we first applied the method to two synthetic data sets for which different cluster paradigms are clearly required. They are composed of two clusters in the plane, for a total of 300 points. The first data set is composed of three elongated Gaussian blobs of 100 points each. Two blobs are partially overlapping, so that they configure one elongated cluster

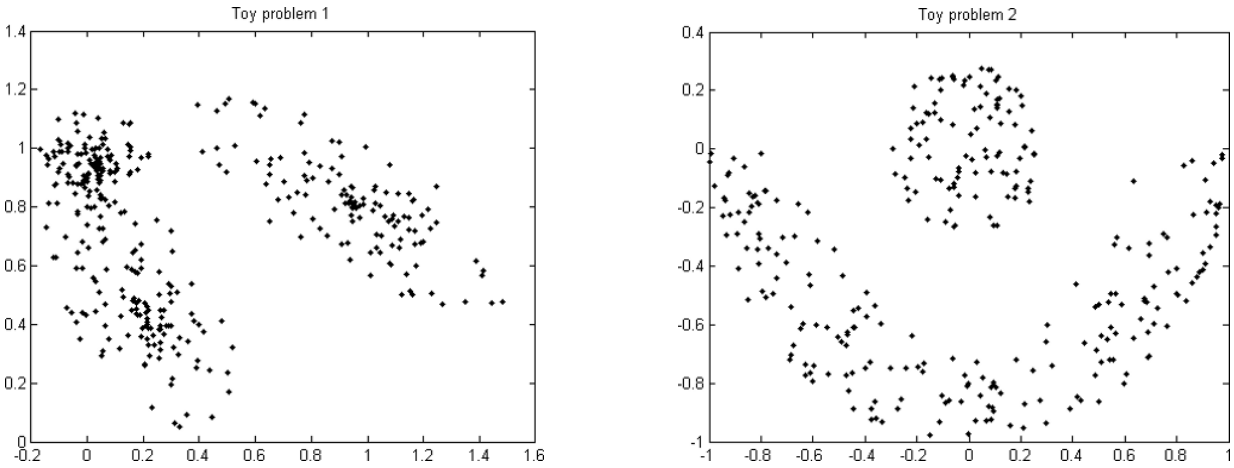


Figure 5.1: Synthetic data sets 1 and 2

with two density peaks. The second data set has no linearly separable clusters, so we expect that it will be more challenging for centroid-based methods. Its two clusters are composed of 100 and 200 points respectively; the smaller one is an isotropic Gaussian cluster, while the larger one is a half-moon shaped distribution. Both data sets are shown in Fig. 5.1.

5.1.2 Experimental results

The quality index values for the individual clustering methods, computed for several choices of the number of clusters c, is shown in Table 5.1 for the two problems. The optimal value is highlighted in bold. It is interesting to note that AGPCM (see Sect. 3.5), being a variation over a *mode seeking* technique rather than a partitional one, refuses to place excess centroids in unsuitable positions, making them overlap instead. This is an inherent cluster validation measure, which not always agrees with other measures, as seen in the table, but tends to reflect the actual cluster structure better than the indices themselves. Stars (*) indicate situations where the values of validity indices are not available because solutions with that particular number of (non-coincident) centroids were not obtained. This automatic validity criterion is useful because the possibilistic clustering model is not a fuzzy generalization of crisp partitions, so the usual hypotheses underlying standard quality indices may not apply.

Table 5.2 indicates the mixing weights obtained for the two problems. In the first case the values are approximately the same; this indicates that the two paradigms reach about the same level of clustering quality.

In the second case, however, the preference for the spectral methods (manifold-based clusters) is clear, with a weight more than triple with respect to the centroidal representation.

Table 5.1: Clustering quality for the two data sets

Problem 1 (blobs)

	KM		**FCM**		**AGPCM**		**SC(fix)**	**SC(var1)**	**SC(var2)**
c	R_{DB}	S_{BX}	R_{DB}	S_{BX}	R_{DB}	S_{BX}	G	G	G
2	**0,290**	**0,119**	**0,363**	**0,096**	**1,044**	**0,871**	**0,0000**	0,0000	0,0000
3	0,319	0,138	0,424	0,099	1,491	1,828	0,0053	**0,0000**	**0,0000**
4	0,502	0,408	0,707	0,142	(*)	(*)	0,0721	0,0002	0,0006
5	0,540	0,270	1,737	0,429	(*)	(*)	0,2710	0,0006	0,0017
6	0,506	0,292	2,286	1,940	(*)	(*)	0,3976	0,0007	0,0020
7	0,603	0,520	3,795	6,003	(*)	(*)	0,4829	0,0011	0,0033
8	0,693	0,669	2,211	0,585	(*)	(*)	0,5208	0,0013	0,0034

Problem 2 (moon)

	KM		**FCM**		**AGPCM**		**SC(fix)**	**SC(var1)**	**SC(var2)**
c	R_{DB}	S_{BX}	R_{DB}	S_{BX}	R_{DB}	S_{BX}	G	G	G
2	0,455	0,273	0,5897	0,208	**0,5205**	**0,148**	**0,0000**	0,0000	0,0000
3	0,463	0,196	0,6437	0,131	1,6058	0,989	0,0036	**0,0000**	**0,0000**
4	**0,416**	**0,092**	**0,5131**	**0,068**	1,5820	0,624	0,0172	0,0001	0,0003
5	0,470	0,150	0,8019	0,121	2,5449	1,491	0,0642	0,0003	0,0009
6	0,537	0,590	1,4408	0,535	(*)	(*)	0,1687	0,0008	0,0020
7	0,539	0,431	3,0569	5,567	(*)	(*)	0,2312	0,0017	0,0041
8	0,505	0,647	3,1629	1,839	(*)	(*)	0,3239	0,0022	0,0048

(*) Values not available for AGPCM. See text.

Table 5.2: Mixing weights obtained for the two datasets

	μ_C	μ_S
Problem 1	0.44	0.56
Problem 2	0.23	0.77

Table 5.3: Quality, as measured by cluster purity, for the two datasets

Problem 1 (blobs)

	KM	FCM	AGPCM	SC(fix)	SC(var1)	SC(var2)	Ensemble
Cluster 1	93%	93%	93%	100%	100%	100%	100%
Cluster 2	85%	85%	85%	85%	85%	85%	85%
Cluster 3	93%	93%	93%	100%	100%	100%	93%

Problem 2 (moon)

	KM	FCM	AGPCM	SC(fix)	SC(var1)	SC(var2)	Ensemble
Cluster 1	90%	43%	45%	100%	100%	100%	100%
Cluster 2	57%	54%	53%	100%	100%	100%	99%

To measure the ensemble quality, for lack of a suitable internal validation index, we resorted to the external index of cluster purity, the percentage of points in a cluster that belong to the majority class; classes here correspond to the individual distribution components used to generate the data.

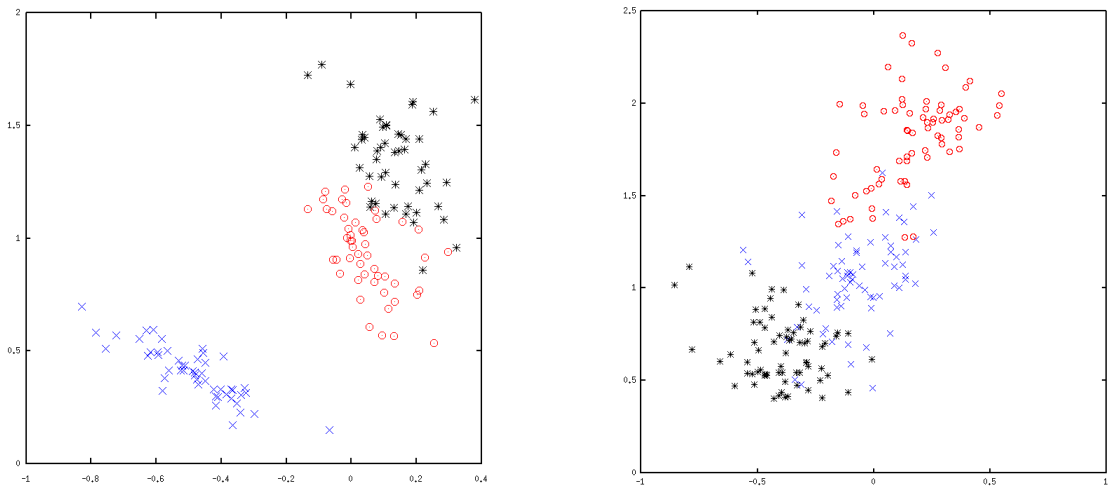
The purity of clusters defined by the ensemble method is compared in Table 5.3 with that of individual clusterings. Although the overall purity of the ensembles is slightly smaller than that of the best performing clusterings, the method clearly points out the most suitable paradigm in each case.

5.1.3 Evaluating Neighbor-based similarities

Due to the large number of variations and the resulting huge number of comparisons that would be needed, in this paper it is impossible to provide a complete experimental validation of the measures described and surveyed. Only selected experiments on some measures will be presented to illustrate a possible experimental approach.

The experiments have been performed on two datasets from the UCI Machine Learning Repository [5], Iris [62] and Seeds [28]. Both are simple and not very extensive, but both present a weak clustering structure, useful to gain understanding on the type of information revealed by similarity measures. we only take fuzzy unsupervised similarities into account, although data are labeled and supervised approaches are also possible.

The well-known Iris data (150 items with 4 attributes) are distributed in three equal-sized classes structured into two clusters. Centroid-based clustering methods can approximate the two touch-



Iris Seeds

Figure 5.2: The datasets used in the experiments, projected onto their first two principal components.

ing classes because the resulting distribution is elongated and such methods make a globular cluster assumption that does not match the distribution very well. Methods that do not rely on this assumption (including heat-kernel based spectral clustering for a range of width parameters) may or may not be able to separate these classes on the basis of structure only (i.e., in an unsupervised setting). On the other hand, the first class is extremely well separated.

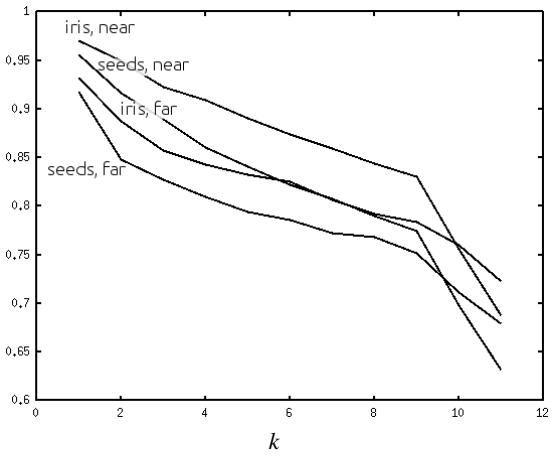
The Seeds data is 210 items with 7 attributes, in 3 classes, and is somewhat similar to two of the three classes of the Iris data in that there are no clear clusters; however the situation here is even worse, since cardinality is similar (70 items per class), dimensionality is higher, and all the three classes are overlapping in one cluster only.

Figure 5.2 shows the two datasets, projected on their first two principal components.

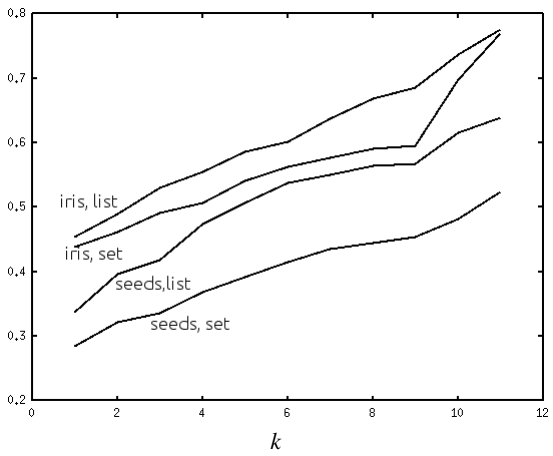
5.1.3.1 Consistency between measures

Figure 5.3 shows the value of the correlation coefficient between the fuzzy measures, as a function of k. This is a consistency test for the measures considered, since correlation is an indication of agreement. The following measures of the "counting" type (`c`) are compared, by computing them on both the Iris and the Seeds data sets: $\hat{s}_{\text{knlc}}$, $\hat{s}_{\text{kflc}}$, $\hat{s}_{\text{knsc}}$, $\hat{s}_{\text{kfsc}}$. These are studies as functions of k. Note that for better readability, "near", "far", "list", and "set" are explicited in the graphs rather than using the corresponding one-character keys.

On the left, the graphs show the correlation of "`s`" measures (using set of neighbors) with "`l`"



Correlation of 'list' vs 'set'



Correlation of 'near' vs 'far'

Figure 5.3: Correlation coefficient between different measures as a function of k.

measures (using list of neighbors) both for "n" (using near neighbors) and "f" (using far neighbors) versions on the two datasets, as indicated in the figure. On the right the graphs show the correlation of "n" measures with "f" measures for "s" and "l" versions, on the two datasets. In both cases we have 2 measures $\times$ 2 datasets = 4 traces. Correlation is decreasing in all traces for the "l" vs "s" case, indicating (as expected) that the "l" criterion is increasingly selective with k. On the other hand, "n" and "f" criteria are increasingly similar, with growing k, in all cases, again confirming a reasonable expectation.

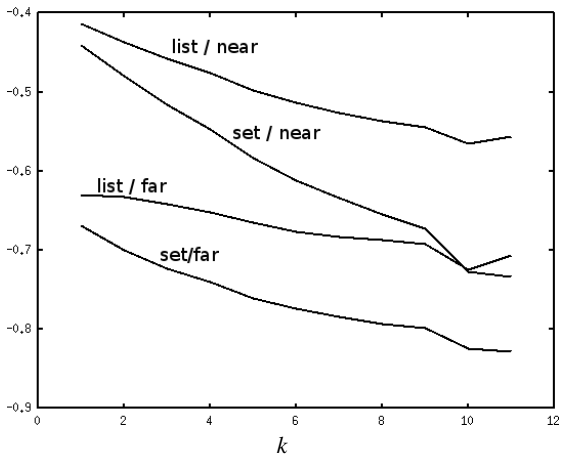
5.1.3.2 Consistency with Euclidean distance

Figure 5.4 shows the value of the correlation coefficient between fuzzy measures and Euclidean distance, as a function of k, another consistency test. The consistency is higher for *more negative* values of the correlation coefficient (since we compare similarities with distances).

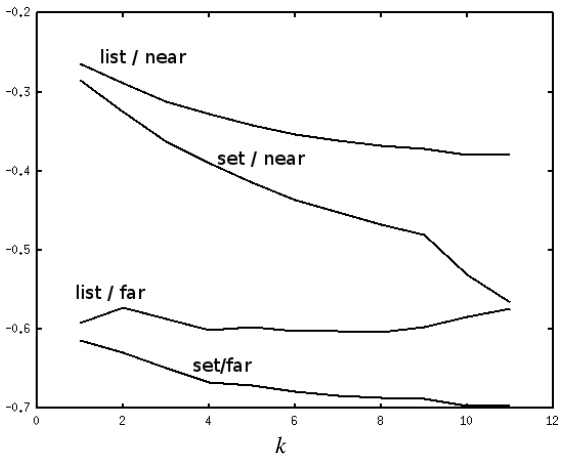
Correlation grows with k, and, somewhat surprisingly, is stronger for measures based on farthest neighbors rather than on nearest neighbors. An expected result, instead, is that correlation for "s" measures is higher than for "l" measures.

5.1.3.3 Experimental analysis with spectral clustering

A sample experiment on the Iris data is presented here to see the actual behavior of a similarity measure when used in spectral clustering, here defined as in [173]. Since (contrary to k-means) we are not enforcing a globular cluster assumption, we expect clusters to be found only for class



Correlation with Euclidean distance, Iris



Correlation with Euclidean distance, Seeds

Figure 5.4: Correlation coefficient of measures with Euclidean distance, as a function of k.

Table 5.4: Example spectral clustering results

Heat kernel, $\sigma = 0.1$

	Setosa	Versicolor	Virginica
cluster 1	50	0	0
cluster 2	0	26	43
cluster 3	0	24	7

$\hat{s}_{\text{knlc}}$, $k = 20$

	Setosa	Versicolor	Virginica
cluster 1	50	0	0
cluster 2	0	27	33
cluster 3	0	23	17

1 (Setosa), while the other classes should correspond much less precisely to clusters.

Confusion matrices between classes and clusters are presented in Table 5.4 for both heat kernel with parameter $\sigma = 1$ (a non-optimal value) and the $\hat{s}_{\text{knlc}}$ similarity (fuzzy unsupervised k-nearest neighbors, rank-sensitive match count). The heat kernel uses all the information available, while $\hat{s}_{\text{knlc}}$ uses only the list of the first 20 nearest neighbors; yet, the results are comparable if one does not take the additional effort to select a suitable heat kernel parameter, a task that in general is known to be tricky [140]. In particular, the well-clustered class is perfectly separated, while the other two classes are mixed in the two other clusters.

5.2 Saccharomyces cerevisiae *PPIs* network study

Inferring significant communities of interacting proteins is a main trend of current biological research, as this task can help in revealing the functionality and the relevance of specific macromolecular assemblies or even in discovering possible proteins affecting a specific biological process. Efficient algorithms able to find suitable communities inside proteins networks may support drug discovery and diseases treatment even in earlier stages.

This section employs spectral and graph clustering methodologies for discovering protein-protein interactions communities in the *Saccharomyces cerevisiae* protein-protein interaction network [125].

The study of the *S. cerevisiae* genetic interactions and their organization by function is the target of many bioinformatic studies [32]. *S. cerevisiae* genome sequence and a set of its deletion mutants represents about 90% of the yeast genome. *S. cerevisiae* PPIs can be used to infer regulation of eukaryotic cells. With some 12 million base pairs and 6,466 genes, at least 31% of *S. cerevisiae* genes have a human homologue [18].

5.2.1 Dataset

We use the *S. cerevisiae* proteins dataset of Krogan et al [104]. In that paper they used a tandem affinity purification to process 4,562 different tagged proteins of the yeast Saccharomyces cerevisiae. Each preparation was analyzed by both matrix-assisted laser desorption/ionization time of flight mass spectrometry and liquid chromatography tandem mass spectrometry. Then they applied an ensemble of decision trees to integrate the mass spectrometry scores and assign probabilities to the protein—protein interactions.

We can represent the dataset as an undirected weighted graph $G = (V, E)$ with V vertices, corresponding to proteins, and E edges indicating protein-protein interaction probabilities (weights) obtained from experiments shown in [104].

We performed our experiments on subgraphs of four different sizes obtained from *S. cerevisiae* dataset (see Tab. 5.5). The subgraphs were chosen on the basis of prior knowledge about protein involved in different biological process.

For instance, protein YAL001C is the largest of six subunits of the RNA polymerase III transcription initiation factor complex (TFIIIC); part of the TauB domain of TFIIIC that binds DNA at the BoxB promoter sites of tRNA and similar genes cooperates with Tfc6p in DNA binding [72].

The evaluation of evaluation of Newman & Girvan modularity [143] versus the number of clusters (or communities) the number of clusters in those sub-graphs is shown in Fig. 5.7 using Newman & Girvan modularity approach [143] and the optimal values are reported in Tab. 4.3. We highlight that the four considered subgraphs are sparse (as noticed by the density evaluations

Table 5.5: Properties of four different subgraphs extracted from *S. cerevisiae* dataset [104] of increasing size and each of them including the smaller subgraphs. For each subgraph we show the number of nodes (proteins), the number of edges and the number of estimated communities using Newman & Girvan modularity [143].

	SG#1	SG#2	SG#3	SG#4
Nodes	31	76	143	257
Edges	30	80	150	300
Communities	2	5	12	19
Density	0.0645	0.0281	0.0148	0.009

shown in Tab 5.5).

5.2.2 Application of graph analysis methods

We analyzed applied various centrality measures and algorithms on SG#1. The evaluation of node centrality degree (Eq. 2.1) allows us to find the centroid YAL001C only (see Fig. 5.5(a)), while closeness (Eq. 2.2) discovers two central nodes: YAL007 and YDR381W (see Fig. 5.5(b)).

The results of the application of the Newman's edge betweenness community detection method [142] are shown in Fig. 5.6. Edges linking proteins YDR381W, YAL007C and YAL001C have highest centralities, but the application of that method is not meaningful on large subgraphs because the random null model underlying modularity becomes unreasonable.

5.2.3 Application of the *FSM*

Before applying the *FSM* community detection method proposed in Sect. 3.3, we evaluted the initial number of clusters step k as shown in Fig. 5.7 using Newman & Girvan modularity approach [143] on the sub-graphs on Tab. 5.5.

Fig. 5.7 shows for each sub-graph the evaluation of modularity versus the number of clusters (or communities). The optimal values are reported on Tab. 5.5.

Then we calculated the affinity or adjacency matrix A between protein pairs s and t with entries defined as:

$$a_{s,t} = \begin{cases} 1 & \text{if } \{s,t\} \in E \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

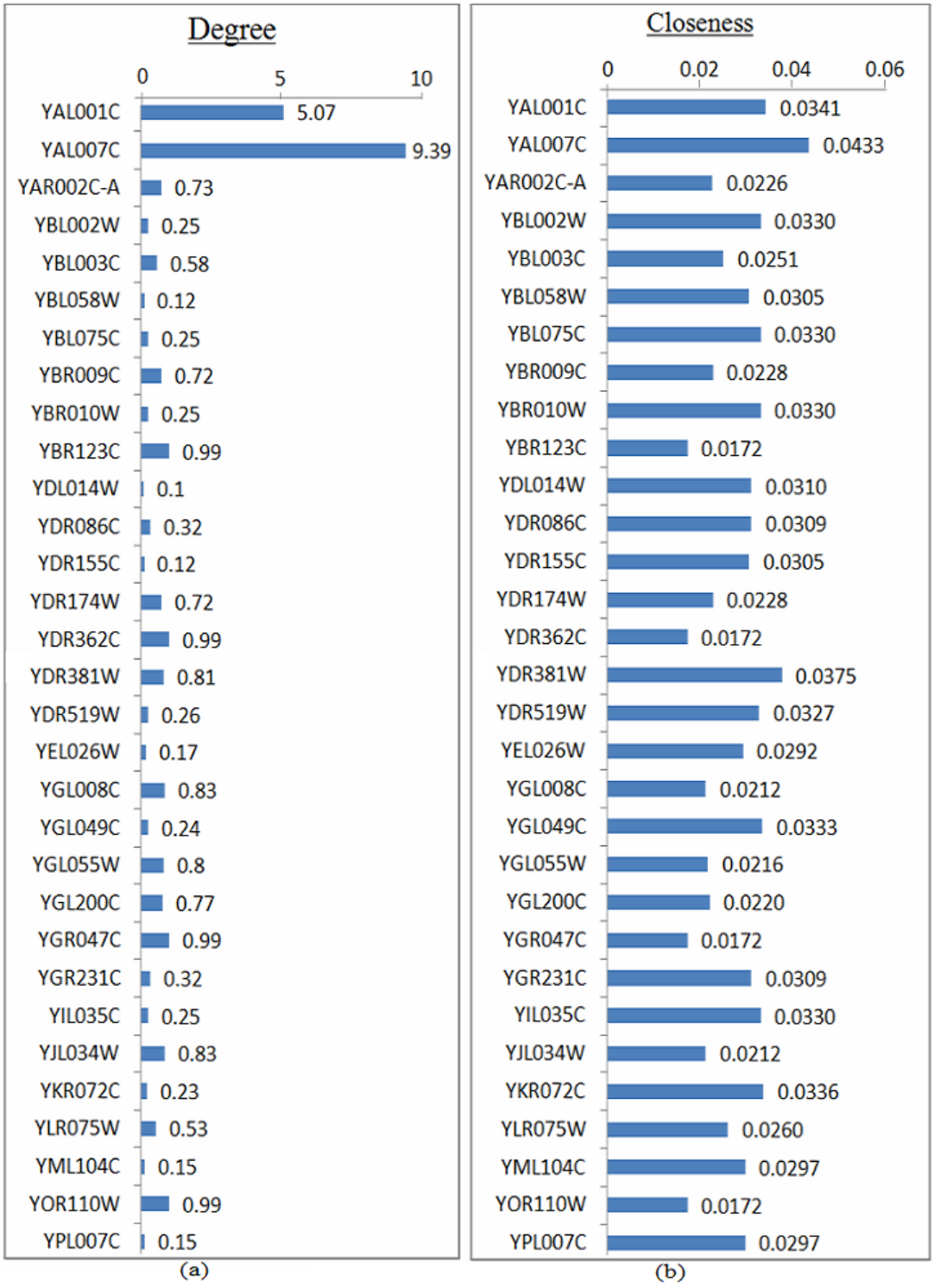


Figure 5.5: Centrality measurements on subgraph SG#1 of the S.cerevisiae *PPI* network. Proteins degree calculation (a); Closeness of proteins (b).

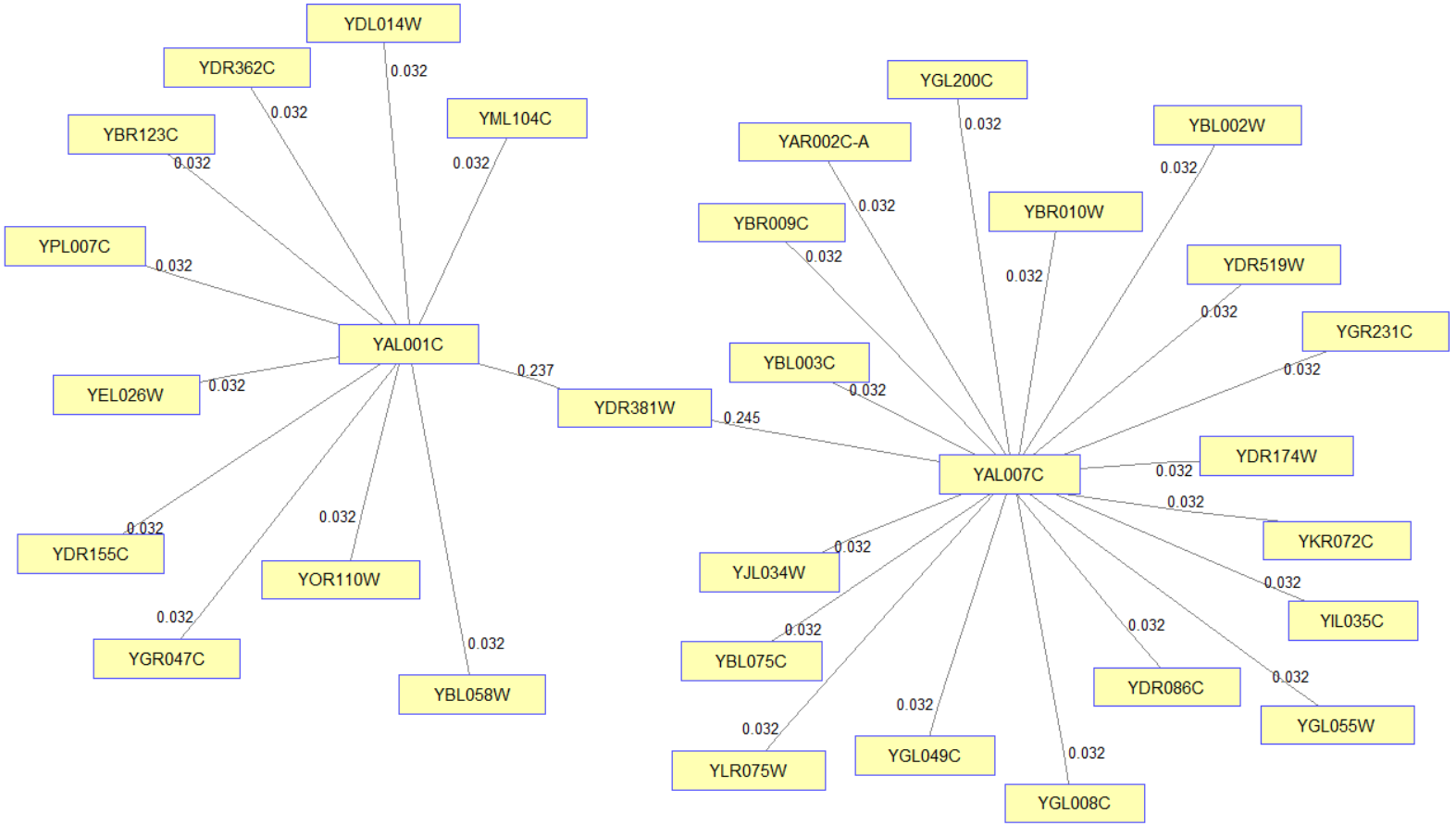


Figure 5.6: Newman edge betweenness evaluations on subgraph SG#1 of the S.cerevisiae *PPI* network.

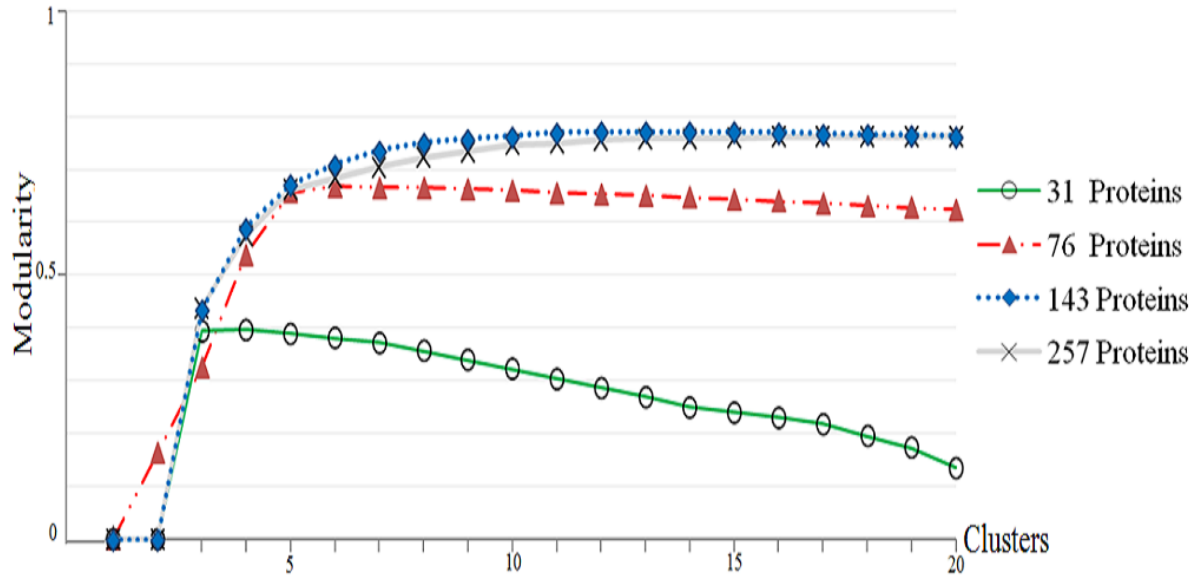


Figure 5.7: Relationship between Newman & Girvan modularity [143] and the choice of the number of clusters (a.k.a. communities) for the four sub-graphs.

The diagonal degree matrix D is obtained by calculating vertices degree. The degree *deg(v)* of each vertex v is the number of edges incident on it.

Fig. 5.8 shows the clustering results on subgraph SG#1 obtained in spectral space using $k = 2$ as the number of clusters to find (corresponding to maximum modularity).

The network is divided into two communities with centroid proteins YAL007C and YAL001C (see Fig. 5.8). We notice that protein YDR381W is assigned with different memberships over a spreadability threshold ϖ (see Sect. 3.4) to both communities: membership $\mu(C_1) = 0.63$ to the community C_1 (with protein names framed with rectangles and centroid protein YAL007C), and $\mu(C_2) = 0.36$ to the community C_2 (with protein names framed with diamonds and centroid protein YAL001C).

It is worth noting that, the results presented in Fig. 5.6 and Fig. 5.8 for Saccharomyces cerevisiae are innovative. In fact, *REACTOME*[1] and *Intact*[2], two known open-sources *PPI* databases, do not replicate the results highlighted in the aforementioned figures.

We also applied the *FSM* community detection method to the other subgraphs of Tab. 5.5, each of them contains the previously introduced subgraphs. The proposed method showed robust results: For example, communities discovered from SG#2 contain those obtained from SG#1

[1] http://www.reactome.org/
[2] http://www.ebi.ac.uk/intact/

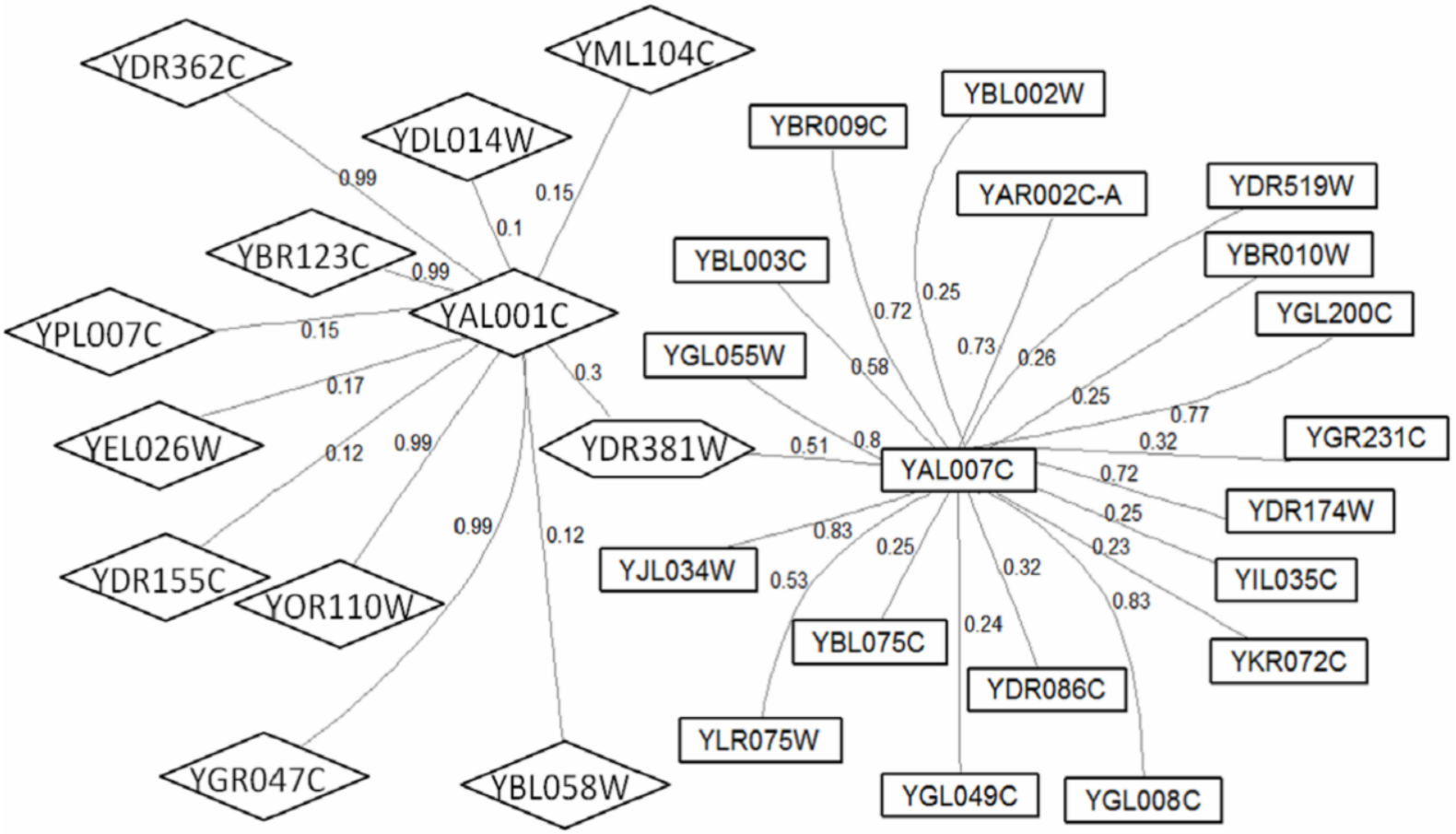


Figure 5.8: Results of the FSM community detection method on SG#1 of 31 *S. cerevisiae* proteins edges are weighted by *PPIs* probabilities measured by Krogan et al [104]. The network is partitioned into two communities, with protein labels framed, respectively, with rectangles and diamonds. Protein YDR381W is framed with an hexagon, as it has significant membership to both communities.

analysis and we still discover the same overlapping proteins and fuzzy structure (see Sect. 5.2.4).

5.2.4 Discovering evolving and overlapping communities

Protein networks can evolve within the changing of the interacting proteins sequence composition, intramolecular contacts, and functions. We applied the *SC-FSM* method (see Sect. 3.8) for characterizing the evolving SG#2 of SG#1 (see previous section).

To this aim, we build the ensemble information content semantic similarity S (see Sect.3.8) by combining the information obtained from both XGraSM [34, 33] annotations, and GO-Universal [136] which considers the topological structure in gene ontology. U

sing the Evidence Accumulation Coding *(EAC)* (see Sect.2.10) [66], we built a consensus similarity matrix using the semantic information and *PPI* interaction measurements given by [104].

Then we calculated the affinity or adjacency matrix A between protein pairs s and t using Eq. 5.1.

Finally, using the *SC-FSM* approach we characterized the fuzzy communities by calculating the fuzzy memberships depicted in Fig.5.9. A protein a is then assigned to a community if it has a significant membership value to it, as discussed in Sect.3.4.

Fig. 5.12 shows the clustering results obtained in spectral space using $k = 5$ that is the number of clusters corresponding to maximum modularity; Fuzzy proteins are labeled by diamonds.

In Figs. 5.10 and 5.11 we depict the semantic similarity between proteins ($sim_{(a,b)}$) for each of the 80 interactions (edges) in the analyzed *S. cerevisiae*'s PPIs network (see Sect.3.8).

For each interaction we measured XGRASm annotation based similarity, and GO-universal topology based semantic similarity through biological process, molecular function, and cellular component directed acyclic graphs of gene ontology.

The results depicted in Figs. 5.10 and 5.11 show higher correlation between *GO-univeral* topological and *XGraSM* annotation based semantic similarity measures as follows: 0.92 for *BP*, 0.79 for *MF*, and 0.74 for *CC*, hence using both of them improve the results specially in biological process specially because it is well defined than others in gene ontology.

We notice that exploiting annotation and topological structure using the proposed semantic enrichment boosts the significance of the detected communities.

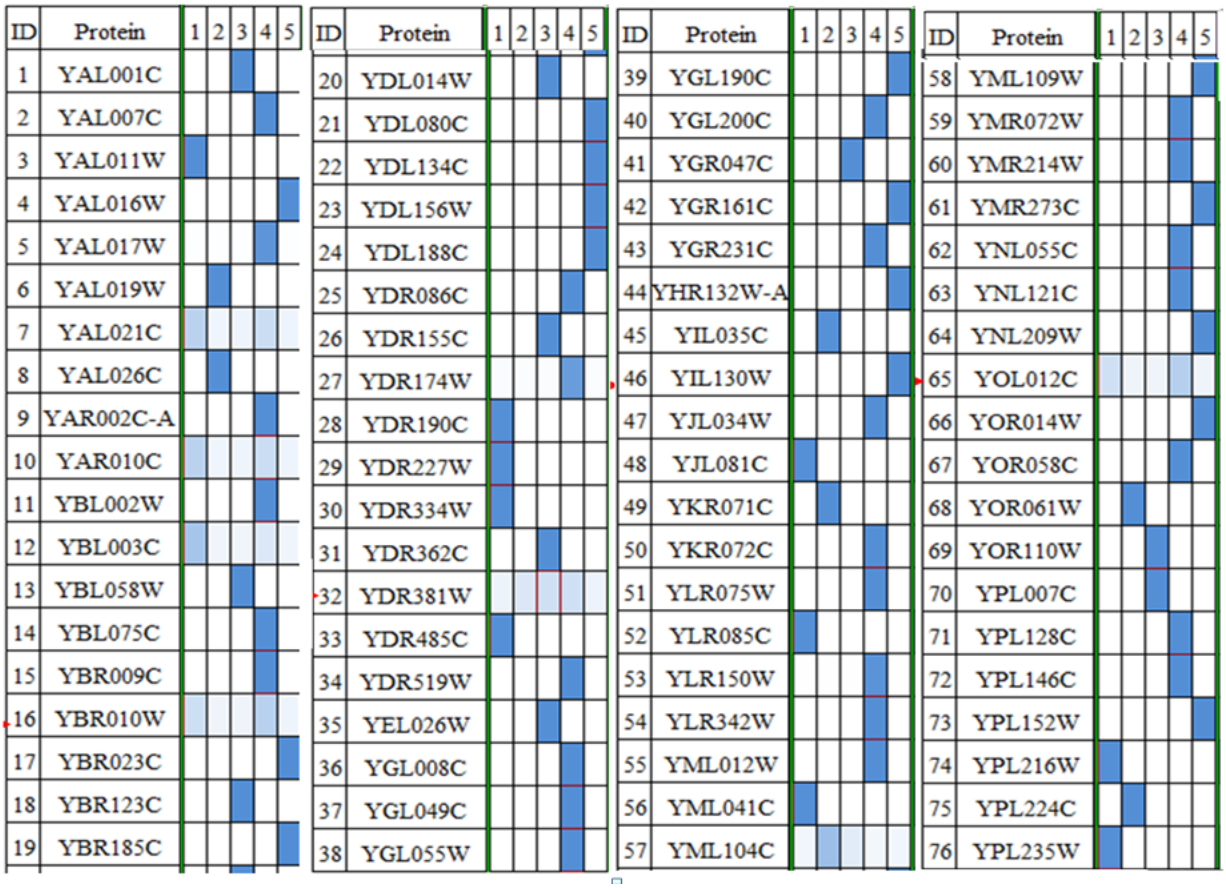


Figure 5.9: Fuzzy membership heatmap of the analyzed *S. cerevisiae* proteins in five communities.

				BP		MF		CC	
Idx	From	To	PPI	XGRASM	GO_Uni	XGRASM	GO_Uni	XGRASM	GO_Uni
1	YAL001C	YBL058W	.12						
2	YAL001C	YBR123C	.99						
3	YAL001C	YDL014W	.10						
4	YAL001C	YDR155C	.12						
5	YAL001C	YDR362C	.99						
6	YAL001C	YDR381W	.30						
7	YAL001C	YEL026W	.17						
8	YAL001C	YGR047C	.99						
9	YAL001C	YML104C	.15						
10	YAL001C	YOR110W	.99						
11	YAL001C	YPL007C	.15						
12	YAL007C	YAR002C-A	.73						
13	YAL007C	YBL002W	.25						
14	YAL007C	YBL003C	.58						
15	YAL007C	YBL075C	.25						
16	YAL007C	YBR009C	.72						
17	YAL007C	YBR010W	.25						
18	YAL007C	YDR086C	.32						
19	YAL007C	YDR174W	.72						
20	YAL007C	YDR381W	.51						
21	YAL007C	YDR519W	.26						
22	YAL007C	YGL008C	.83						
23	YAL007C	YGL049C	.24						
24	YAL007C	YGL055W	.80						
25	YAL007C	YGL200C	.77						
26	YAL007C	YGR231C	.32						
27	YAL007C	YIL035C	.25						
28	YAL007C	YJL034W	.83						
29	YAL007C	YKR072C	.23						
30	YAL007C	YLR075W	.53						
31	YAL007C	YLR150W	.33						
32	YAL007C	YLR342W	.33						
33	YAL007C	YML012W	.93						
34	YAL007C	YMR072W	.25						
35	YAL007C	YMR214W	.74						
36	YAL007C	YNL055C	.33						
37	YAL007C	YNL121C	.25						
38	YAL007C	YOL012C	.24						
39	YAL007C	YOR058C	.23						
40	YAL007C	YPL128C	.25						

Figure 5.10: Part I: Semantic enrichment in yeast S.cerevisiae PPIs of GO biological process (BP), molecular function (MF), and cellular component (CC) aspects. The graphs compare the evaluations obtained using GO-universal and XGraSM.

				BP		MF		CC	
Idx	From	To	PPI	XGRASM	GO_Uni	XGRASM	GO_Uni	XGRASM	GO_Uni
41	YAL007C	YPL146C	.26						
42	YAL011W	YBL003C	.63						
43	YAL011W	YBR010W	.15						
44	YAL011W	YDR174W	.15						
45	YAL011W	YDR190C	.99						
46	YAL011W	YDR227W	.19						
47	YAL011W	YDR334W	.99						
48	YAL011W	YDR485C	.99						
49	YAL011W	YJL081C	.99						
50	YAL011W	YLR085C	.99						
51	YAL011W	YML041C	.99						
52	YAL011W	YOL012C	.15						
53	YAL011W	YPL216W	.19						
54	YAL011W	YPL235W	.99						
55	YAL016W	YBR023C	.15						
56	YAL016W	YBR185C	.17						
57	YAL016W	YDL080C	.16						
58	YAL016W	YDL134C	.99						
59	YAL016W	YDL156W	.16						
60	YAL016W	YDL188C	.99						
61	YAL016W	YGL190C	.99						
62	YAL016W	YGR161C	.55						
63	YAL016W	YHR132W-A	.17						
64	YAL016W	YIL130W	.11						
65	YAL016W	YML109W	.78						
66	YAL016W	YMR273C	.93						
67	YAL016W	YNL209W	.58						
68	YAL016W	YOR014W	.70						
69	YAL016W	YPL152W	.36						
70	YAL017W	YLR150W	.40						
71	YAL019W	YAL026C	.64						
72	YAL019W	YDR381W	.15						
73	YAL019W	YIL035C	.99						
74	YAL019W	YKR071C	.19						
75	YAL019W	YML104C	.15						
76	YAL019W	YOR061W	.12						
77	YAL019W	YPL224C	.15						
78	YAL021C	YAR010C	.16						
79	YAL021C	YBL003C	.12						
80	YAL021C	YBR010W	.12						

Figure 5.11: Part II: Semantic enrichment in yeast S.cerevisiae PPIs of GO biological process (BP), molecular function (MF), and cellular component (CC) aspects. The graphs compare the evaluations obtained using GO-universal and XGraSM.

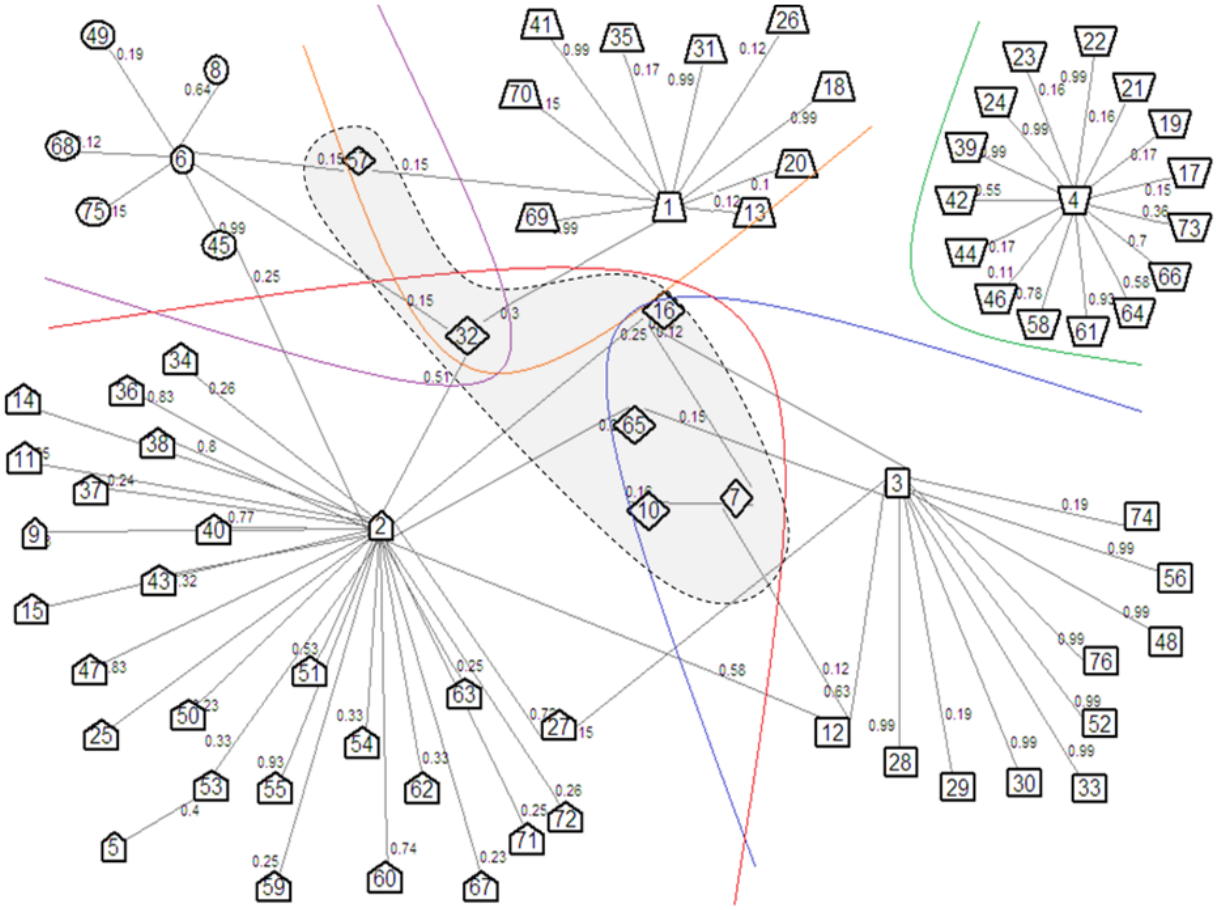


Figure 5.12: Results of the *SC-FSM* community detection method on the analyzed *S. cerevisiae* PPIs network. Edges weights are PPIs probabilities. The network is partitioned into five communities. Proteins in gray region, framed with diamonds, act as bridge nodes with fuzzy memberships.

5.3 *HIV-1* and Leukemia networks study

This section presents an analysis of *PPI* in immunology networks [122, 119].

In this study, we aim at detecting proteins annotated to the biological processes significantly related to Human immunodeficiency virus-1 *(HIV-1)* infection in *Homo sapiens* extracted from NCBI database[3] [180]. *HIV-1* is the etiologic agent of acquired immune deficiency syndrome (AIDS). The number of AIDS-related deaths was 2.1 million in 2007 alone [180].

Leukemia is a group of cancers that usually begins in the bone marrow and results in high numbers of abnormal white blood cells. These white blood cells are not fully developed and are called blasts or leukemia cells. The causes of leukemia is still not clearly identified and they differ from a leukemia type to another. There are four main common types of leukemia: acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL) and chronic myeloid leukemia (CML). In 2012 leukemia developed in 352,000 people globally and caused 265,000 deaths. Leukemia is the most common type of cancer in children [91].

We report the following network characteristics (see Sect.2.1) in Tab. 5.6:

- Number of proteins: $n = |V|$.
- Number of edges: $m = |E|$.
- Average Degree: $C_D = \frac{\sum_{i=1}^{n} C_D(v_i)}{n}$.
- Average Closeness: $C_C = \frac{\sum_{i=1}^{n} C_C(v_i)}{n}$.
- Average Betweenness: $C_B = \frac{\sum_{j=1}^{m} C_B(e_j)}{m}$.
- Average Clustering coefficient: $C_{CC} = \frac{\sum_{i=1}^{n} C_{CC}(v_i)}{n}$.
- Average path length *(APL)*.
- Volume: refers to the sum of degrees of proteins in network.
- Density: given by $\frac{2m}{n(n-1)}$.
- Diameter: refers to the longest shortest path between any two nodes in the network.

For both *HIV-1* and leukemia networks, We applied the proposed *SE-FSM* community detection method (see Sect. 3.7, and validated our results with the cellular proteins induced by HIV-1 infection reported by *QIAGEN*[4].

[3] http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/

[4] http://www.qiagen.com

Table 5.6: *Characteristics of Leukemia and* HIV-1 *networks.*

Dataset	n	m	C_D	C_C	C_B	C_{CC}	APL	$Density$	$Diameter$
$HIV-1$	29	304	6.902	0.034	0.006	0.651	1.092	0.247	2.493
Leukemia	57	720	8.537	0.015	0.003	0.514	1.254	0.152	2.962

5.3.1 *HIV-1* in homo sapiens overlapping community identification

In each experiment we applied the *FSM* community detection method on the *PPI* networks induced by the following similarity matrices (discussed in Sect. 3.7):

- Quantitative information (Q).
- Semantic similarity (S).
- Hybrid similarity matrix (H), that combines Q and S using evidence accumulation [66] (see Sect.2.10).

We started with the analysis of the *PPI* networks induced by Q. The obtained communities are highly consistent, and their Rand index (see Sect. 2.5) [154] is .95.

In order to identify the significant semantic terms annotated to *HIV-1* we considered *HIV-1* significant annotations based on biological processes (BP), molecular functions (MF), and cellular component (CC) of Gene Ontology*(GO)*[5]. Moreover, we selected the relevant protein pathways from *KEGG*[6], *Reactome*[7], and *PID*[8] (see Sect. 2.9).

The analysis of the *PPI HIV-1* infection network using the proposed hybrid similarity metric (H) obtained three functional communities shown in Fig. 5.13 and proteins are assigned to different functional communities :

- proteins member of community $C1$ participate in many processes, including apoptosis, cell death, cell cycle and proliferation activities, cell cycle activities, protein dimerization activity *GO:0046983 (MF)*, amyotrophic lateral sclerosis (ALS), and prostate cancer KEGG pathways;
- proteins member of community $C2$ influence the transcription factors and regulators, STAT transcription factor, DNA-binding, T cell receptor signaling pathway, and Interferon alpha/beta signaling from reactome *RCTM38609*;

[5]http://www.geneontology.org
[6]http://www.genome.jp/kegg/
[7]http://www.reactome.org/
[8]http://pid.nci.nih.gov/

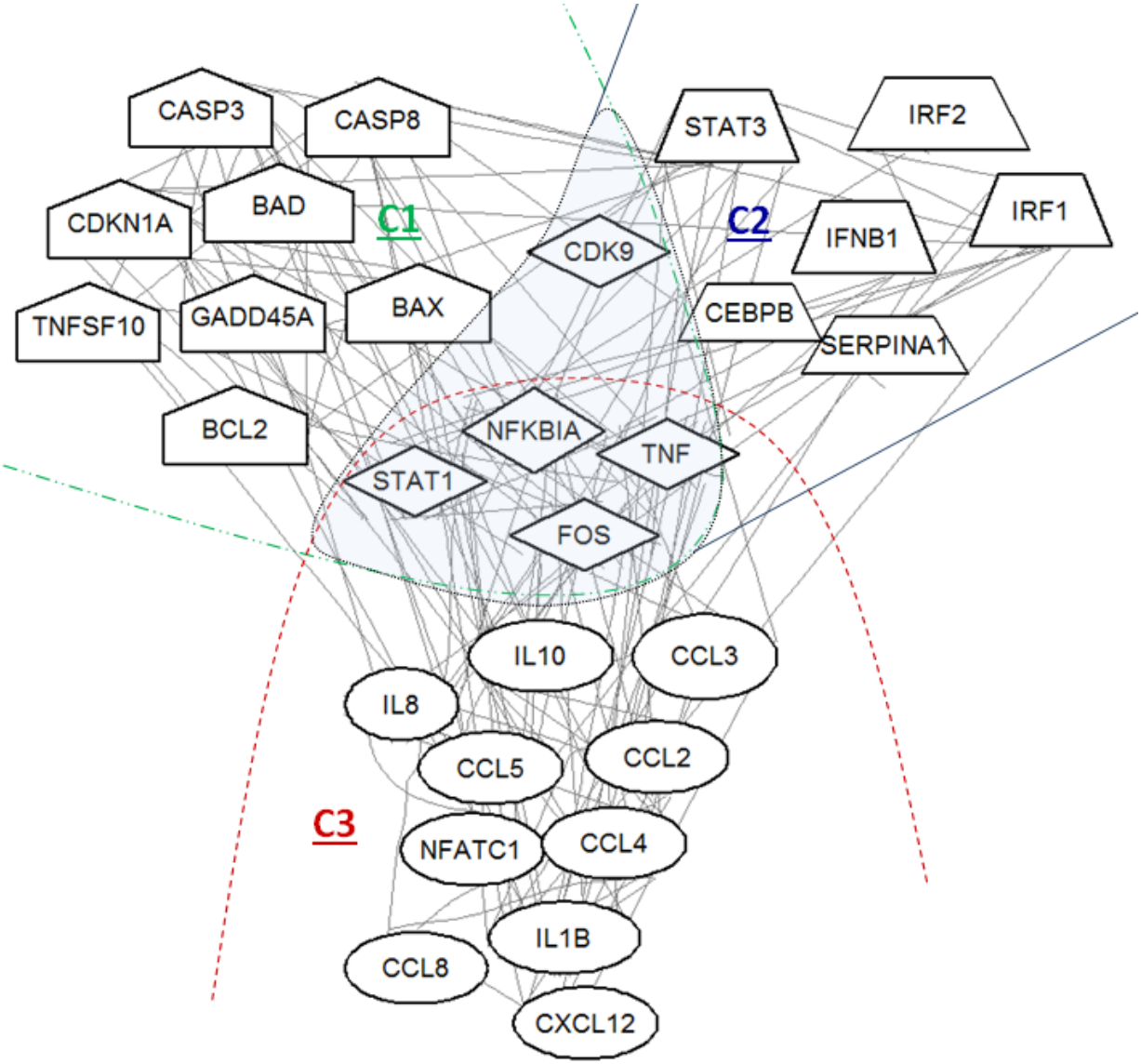


Figure 5.13: *Identified protein-protein interaction communities in HIV-1 biological network induced by (SE-FSM) community detection method. Fuzzy nodes with significant memberships to more than one community are framed by diamonds.*

Table 5.7: *Rand indexes calculated from the results of the analysis performed on the* PPI *networks induced by Q, S, and H.*

Rand	Q	S	H
Q	1	.84	.90
S	.84	1	.78
H	.90	.78	1

- proteins member of community $C3$ participate in viral activities such as response to virus and inflammatory response and homeostasis *GO:0055065 (BP)*, Chemokine signaling pathway, and cytokine activity *GO:0005125 (MF)*.

Proteins framed by diamonds in Fig. 5.13 are fuzzy proteins that our analysis annotated to more than one community. It is interesting to note that the information stored in *QIAGEN* reveal that the *fuzzy nodes* we found participate to the functions of the three communities: For instance, protein *STAT1* participates in protein dimerization activity *GO:0046983 (MF)* in community $C1$; STAT transcription factor and DNA-binding in community $C2$; and Chemokine signaling pathway in community $C3$.

We highlight that, the identified *PPI* communities in the *HIV-1* biological network induced by (*SE-FSM*) community detection method (see, Fig. 5.13) are innovative: only one *PPI* database partially replicates the results described in Fig. 5.13. In fact, when queried for *HIV-1*, the (host interactions of *HIV* factors) pathway identified by the *REACTOME* platform is vaguely similar to the communities described in Fig. 5.13.

In Tab. 5.7 we report the Rand indexes comparing the three communities obtained from the *PPI* networks induced by the similarity matrices Q, S, and H. We highlight that community $C1$ is identified by all approaches and its interactions are robust from the biological viewpoint.

It is worth to note that the semantic enrichment can enhance the *FSM* results, as proteins can be assigned to their significant functional communities as we consider their proteomics pathways as well. For instance, if we take into account the quantitative information only (stored in the similarity matrix Q), protein *CCL3* is assigned to community *C2*, but, when we consider the similarity matrix S including the information related to the semantically enriched terms such as *GO:0005125 (MF)*, the same protein *CCL3* is assigned to community *C3* where it participates in biological activities with higher relevance than those of *C2*.

5.3.2 *Leukemia* in homo sapiens overlapping community identification

We repeat here the results we obtained by applying the proposed *SE-FSM* method in analyzing Leukaemia *PPI*. The *SE-FSM* method detected four overlapping communities ($C1, C2, C3, and\ C4$) depicted in Fig. 5.14. Moreover, we show the fuzzy membership of the interacting proteins in Fig. 5.15.

We note that, the communities we discovered using *SE-FSM* emphasize the proteomics relation reported in many other experimental studies in addition to agree with the results reported in *QIAGEN*. For instance, *Goyama* et al. [74] discovered a dual role of RUNX1 in myeloid leukemogenesis using normal human cord blood cells and those expressing leukemogenic fusion proteins.

Goyama et al. [74] reported that, RUNX1 overexpression inhibits the growth of normal cord blood cells by inducing myeloid differentiation, whereas a certain level of RUNX1 activity is required for the growth of AML1-ETO and MLL-AF9 cells.

Moreover, using a mouse genetic model, they showed that the combined loss of Runx1/Cbfb inhibited leukemia development induced by MLL-AF9. RUNX2 could compensate for the loss of RUNX1. The survival effect of RUNX1 was mediated by BCL2 in MLL fusion leukemia.

In addition, *Goyama* et al. [74] study unveiled an unexpected prosurvival role for RUNX1 in myeloid leukemogenesis. Inhibiting RUNX1 activity rather than enhancing it could be a promising therapeutic strategy for AMLs with leukemogenic fusion proteins.

The results showed that adopting the semantic annotation of the analyzed proteins and exploiting heterogenous biological information sources, such as *homology*, *co-expression*, *experimental results*, *knowledge bases*, and *text mining* enhanced the relevance of the inferred communities and inferred their functional association entailed in Gene Ontology an protein pathways.

It is known that untreated *HIV* infection causes *AIDS* and this major impairment in the immune system is associated with an increased risk of cancer, including leukemia. Recently, it was described a remarkable therapeutic success that connects *HIV* and leukemia. A patient underwent a stem cell transplant for his acute myeloid leukemia (AML) treatment, but ended up with a complete AML remission and his *HIV-1* RNA was undetectable.

Using *SE-FSM* enabled us to infer the functional relationship between *HIV-1* and leukemia networks, we highlight that they share several significant semantic annotations like those shown in Fig. 5.16 [122, 119].

The significant semantic annotations identified in communities of both *HIV-1* and leukemia networks using *SE-FSM* (see, Fig. 5.16) might describe one of the possible system biology scenarios leading to an effective functional *HIV* cure due to leukemia treatment.

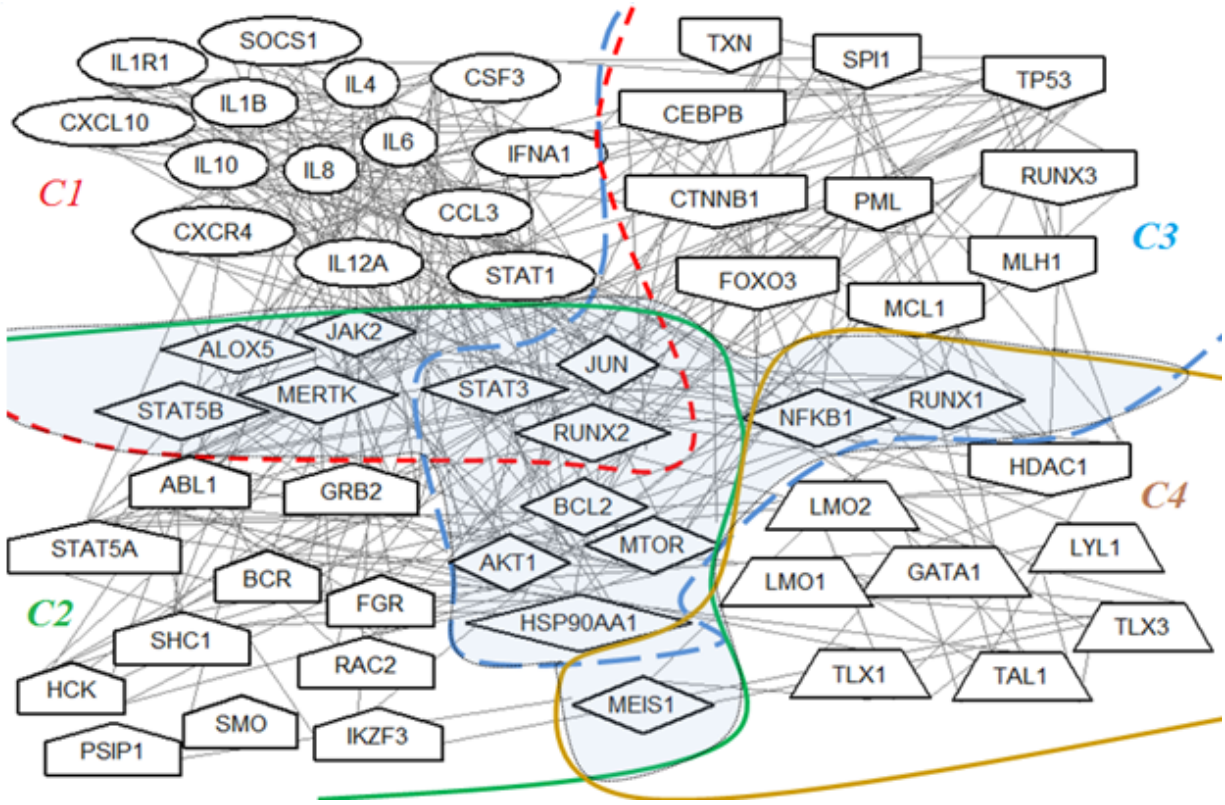


Figure 5.14: *Identified protein-protein interaction communities in leukemia biological network induced by the* SE-FSM *community detection method. Fuzzy nodes significantly annotated to more than one community are framed by diamonds.*

ID	Protein	C1	C2	C3	C4	ϖ	b	χ
1	**ALOX5**	.50	.32	.09	.09	**.16**	.60	3.12
2	**MTOR**	.08	.27	.59	.07	**.18**	.51	2.85
3	STAT1	.92	.04	.02	.02	**.25**	.10	1.43
4	IKZF3	.09	.73	.09	.09	**.22**	.36	2.40
5	FOXO3	.01	.02	.96	.01	**.25**	.06	1.26
6	**JAK2**	.64	.20	.08	.07	**.20**	.47	2.74
7	CXCR4	.99	.00	.00	.00	**.25**	.01	0.00
8	TXN	.01	.02	.94	.02	**.25**	.07	1.32
9	**STAT5B**	.23	.61	.09	.07	**.19**	.50	2.82
10	HCK	.01	.97	.01	.01	**.25**	.04	1.19
11	IFNA1	.94	.02	.02	.02	**.25**	.09	1.36
12	IL6	.96	.02	.01	.01	**.25**	.05	1.21
13	IL10	.95	.02	.01	.02	**.25**	.07	1.29
14	ABL1	.03	.87	.06	.03	**.24**	.17	1.69
15	CTNNB1	.01	.02	.95	.02	**.25**	.07	1.30
16	**JUN**	.22	.19	.47	.12	**.13**	.69	3.51
17	**RUNX2**	.15	.11	.61	.13	**.18**	.52	2.99
18	**STAT3**	.52	.22	.16	.10	**.15**	.63	3.32
19	TAL1	.02	.02	.03	.93	**.25**	.09	1.39
20	SHC1	.03	.92	.03	.02	**.25**	.11	1.45
21	GRB2	.10	.81	.06	.04	**.23**	.26	1.99
22	LYL1	.01	.01	.01	.96	**.25**	.05	1.23
23	STAT5A	.18	.66	.10	.06	**.20**	.44	2.66
24	TLX1	.07	.08	.07	.77	**.23**	.30	2.18
25	TLX3	.10	.12	.09	.70	**.21**	.41	2.58
26	CEBPB	.07	.08	.75	.10	**.22**	.33	2.31
27	BCR	.03	.91	.04	.02	**.25**	.12	1.48
28	FGR	.02	.94	.03	.01	**.25**	.08	1.33
29	RUNX3	.04	.06	.84	.06	**.24**	.21	1.83

ID	Protein	C1	C2	C3	C4	ϖ	b	χ
30	**AKT1**	.14	.24	.54	.09	**.16**	.60	3.21
31	GATA1	.10	.07	.17	.66	**.20**	.45	2.69
32	MCL1	.05	.13	.78	.05	**.23**	.29	2.10
33	HDAC1	.05	.06	.78	.10	**.23**	.29	2.14
34	PSIP1	.13	.61	.11	.15	**.18**	.52	2.99
35	SOCS1	.82	.09	.04	.05	**.24**	.24	1.94
36	**NFKB1**	.27	.08	.34	.31	**.11**	.76	3.60
37	PML	.01	.03	.94	.02	**.25**	.07	1.31
38	IL8	.84	.06	.05	.04	**.24**	.21	1.85
39	CXCL10	.85	.05	.04	.06	**.24**	.20	1.79
40	IL1B	.88	.04	.03	.05	**.24**	.16	1.63
41	RAC2	.04	.86	.07	.03	**.24**	.18	1.72
42	**RUNX1**	.08	.15	.57	.21	**.17**	.56	3.09
43	**HSP90AA1**	.08	.40	.45	.07	**.16**	.60	3.06
44	TP53	.01	.03	.94	.02	**.25**	.07	1.32
45	CSF3	.97	.01	.01	.01	**.25**	.03	1.16
46	SPI1	.04	.06	.83	.07	**.24**	.22	1.89
47	IL12A	.95	.02	.01	.02	**.25**	.07	1.29
48	LMO1	.01	.01	.01	.97	**.25**	.04	1.20
49	LMO2	.01	.01	.01	.97	**.25**	.05	1.20
50	**BCL2**	.09	.41	.43	.07	**.16**	.61	3.12
51	IL1R1	.89	.04	.03	.05	**.24**	.15	1.61
52	IL4	.95	.02	.01	.01	**.25**	.06	1.27
53	CCL3	.80	.06	.05	.09	**.23**	.27	2.05
54	**MEIS1**	.13	.45	.12	.30	**.13**	.69	3.46
55	SMO	.05	.82	.09	.05	**.24**	.24	1.95
56	MLH1	.04	.06	.85	.05	**.24**	.20	1.81
57	**MERTK**	.32	.49	.10	.10	**.15**	.62	3.20

Figure 5.15: *Protein membership in the four communities obtained by applying* SE-FSM *on leukemia* PPI *network. Proteins with fuzzy membership are in bold and their memberships to communities where we assign them are underlined*

	HIV-1 (29 proteins)		Leukemia (58 proteins)	
Semantic enrichment	# proteins	p-value	# proteins	p-value
Biological Process				
GO:0002682 regulation of immune system process	12	1.89E-08	33	5.01E-29
GO:0045595 regulation of cell differentiation	10	3.60E-06	27	8.48E-21
GO:0006955 immune response	15	2.17E-11	26	2.04E-18
GO:2000026 regulation of multicellular organismal development	12	5.11E-08	25	7.22E-18
GO:0042129 regulation of T cell proliferation	4	3.71E-05	7	2.06E-08
GO:0042113 B cell activation	4	4.10E-05	5	1.33E-05
Molecular function				
GO: 0042379 chemokine receptor binding	5	7.69E-09	4	8.96E-07
GO:0005126 cytokine receptor binding	7	1.40E-08	8	1.70E-08
Cellular component				
GO:0000790 nuclear chromatin	3	2.69E-03	10	6.49E-11
GO:0031981 nuclear lumen	8	5.77E-03	23	2.27E-10
GO:0005740 mitochondrial envelope	4	6.09E-03	4	4.10E-02
KEGG pathways				
hsa05200 Pathways in cancer	11	1.71E-08	24	1.48E-29
hsa05220 Chronic myeloid leukemia	3	2.81E-03	11	1.97E-17
hsa04630 Jak-STAT signaling pathway	4	3.40E-03	13	9.47E-17

Figure 5.16: *Significant semantic annotations identified in communities of both HIV-1 and leukemia using SE-FSM.*

5.4 Spreadability analysis

This section shows that identifying fuzzy communities based on the proposed spreadability cut (ϖ) (see Sect. 3.4) is a robust and global measure, unlike other measures such as mean ($\mu = \sum U_{1..k}(s)$) of a node s which is sensitive to noise and membership variation.

Moreover, the proposed ϖ does not have the limitation of other measures e.g, method of Zhang [203], exponential entropy given by $\chi(s) = \Pi_{i=1}^{k} u_i(s)^{-u_i(s)}$, or bridgeness score ($b$) [141] (see Sect. 2.2) that requires an external parameter choice for tuning significant memberships [87].

For example, in the Leukemia study we discussed in Sect. 5.3.2, the proposed spreadability cut (ϖ) can identify fuzziness of Protein #4 shown in Fig. 5.15, while bridgeness (b), and exponential entropy (χ) can not provide a clear membership cut.

In Fig. 5.17, we extend the experiment depicted in Fig. 5.15 and we show that other statistical measures based on the membership vector of each node such as variance (σ^2), standard deviation σ, or measuring $\beta = \mu - \sigma^2$ (note, μ=.25 for all nodes as we have 4 communities) can not always provide a clear cut for node fuzziness.

Hence, the spreadability measure can be generalized (It always detects a significant cut of the node memberships). We proved this experimentally using the following procedure:

- Given a membership matrix (U) (for instance Leukemia membership matrix (see Fig. 5.17).
- We sort U ascending based on nodes fuzziness (i.e., descending based on crispness), such that the top most nodes are crisp or closer to crisp than the following nodes, this is obtained by sorting each membership vector $U_{1..k}(s)$ for each node s and then making lexicographic descending sort of U as shown in Fig. 5.18.
- We plotted the variance (σ^2), the spreadability cut (ϖ), and the standard deviation (σ) as shown in Fig. 5.19.

  We experimentally found that the proposed spreadability cut (ϖ) always decreases as long as fuzziness increases. Moreover, the results confirmed that ϖ always gives a significant fuzziness cut of a node unlike variance or standard deviation.

ID	Protein	C1	C2	C3	C4	max	min	σ	σ^2	β	ϖ	b	χ
1	**ALOX5**	.50	.32	.09	.09	.32	.09	.20	.04	.21	**.16**	.60	3.12
2	**MTOR**	.08	.27	.59	.07	.27	.08	.24	.06	.19	**.18**	.51	2.85
3	STAT1	.92	.04	.02	.02	.92	.04	.45	.20	.05	**.25**	.10	1.43
4	IKZF3	.09	.73	.09	.09	.73	.09	.32	.10	.15	**.22**	.36	2.40
5	FOXO3	.01	.02	.96	.01	.96	.02	.47	.22	.03	**.25**	.06	1.26
6	**JAK2**	.64	.20	.08	.07	.20	.08	.27	.07	.18	**.20**	.47	2.74
7	CXCR4	.99	.00	.00	.00	.99	.00	.49	.24	.01	**.25**	.01	0.00
8	TXN	.01	.02	.94	.02	.94	.02	.46	.21	.04	**.25**	.07	1.32
9	**STAT5B**	.23	.61	.09	.07	.23	.09	.25	.06	.19	**.19**	.50	2.82
10	HCK	.01	.97	.01	.01	.97	.01	.48	.23	.02	**.25**	.04	1.19
11	IFNA1	.94	.02	.02	.02	.94	.02	.46	.21	.04	**.25**	.09	1.36
12	IL6	.96	.02	.01	.01	.96	.02	.48	.23	.02	**.25**	.05	1.21
13	IL10	.95	.02	.01	.02	.95	.02	.47	.22	.03	**.25**	.07	1.29
14	ABL1	.03	.87	.06	.03	.87	.06	.41	.17	.08	**.24**	.17	1.69
15	CTNNB1	.01	.02	.95	.02	.95	.02	.47	.22	.03	**.25**	.07	1.30
16	**JUN**	.22	.19	.47	.12	.19	.12	.15	.02	.23	**.13**	.69	3.51
17	**RUNX2**	.15	.11	.61	.13	.61	.15	.24	.06	.19	**.18**	.52	2.99
18	**STAT3**	.52	.22	.16	.10	.16	.10	.18	.03	.22	**.15**	.63	3.32
19	TAL1	.02	.02	.03	.93	.93	.03	.45	.21	.04	**.25**	.09	1.39
20	SHC1	.03	.92	.03	.02	.92	.03	.44	.20	.05	**.25**	.11	1.45
21	GRB2	.10	.81	.06	.04	.81	.10	.37	.14	.11	**.23**	.26	1.99
22	LYL1	.01	.01	.01	.96	.96	.01	.47	.22	.03	**.25**	.05	1.23
23	STAT5A	.18	.66	.10	.06	.66	.18	.28	.08	.17	**.20**	.44	2.66
24	TLX1	.07	.08	.07	.77	.77	.08	.35	.12	.13	**.23**	.30	2.18
25	TLX3	.10	.12	.09	.70	.70	.12	.30	.09	.16	**.21**	.41	2.58
26	CEBPB	.07	.08	.75	.10	.75	.10	.33	.11	.14	**.22**	.33	2.31
27	BCR	.03	.91	.04	.02	.91	.04	.44	.19	.06	**.25**	.12	1.48
28	FGR	.02	.94	.03	.01	.94	.03	.46	.21	.04	**.25**	.08	1.33
29	RUNX3	.04	.06	.84	.06	.84	.06	.40	.16	.09	**.24**	.21	1.83
30	**AKT1**	.14	.24	.54	.09	.24	.14	.20	.04	.21	**.16**	.60	3.21
31	GATA1	.10	.07	.17	.66	.66	.17	.28	.08	.17	**.20**	.45	2.69
32	MCL1	.05	.13	.78	.05	.78	.13	.35	.13	.12	**.23**	.29	2.10
33	HDAC1	.05	.06	.78	.10	.78	.10	.35	.13	.12	**.23**	.29	2.14
34	PSIP1	.13	.61	.11	.15	.61	.15	.24	.06	.19	**.18**	.52	2.99
35	SOCS1	.82	.09	.04	.05	.82	.09	.38	.14	.11	**.24**	.24	1.94
36	**NFKB1**	.27	.08	.34	.31	.27	.08	.12	.01	.24	**.11**	.76	3.60
37	PML	.01	.03	.94	.02	.94	.03	.46	.21	.04	**.25**	.07	1.31
38	IL8	.84	.06	.05	.04	.84	.06	.39	.15	.10	**.24**	.21	1.85
39	CXCL10	.85	.05	.04	.06	.85	.06	.40	.16	.09	**.24**	.20	1.79
40	IL1B	.88	.04	.03	.05	.88	.05	.42	.18	.07	**.24**	.16	1.63
41	RAC2	.04	.86	.07	.03	.86	.07	.41	.17	.08	**.24**	.18	1.72
42	**RUNX1**	.08	.15	.57	.21	.21	.15	.22	.05	.20	**.17**	.56	3.09
43	**HSP90AA1**	.08	.40	.45	.07	.40	.08	.20	.04	.21	**.16**	.60	3.06
44	TP53	.01	.03	.94	.02	.94	.03	.46	.21	.04	**.25**	.07	1.32
45	CSF3	.97	.01	.01	.01	.97	.01	.48	.23	.02	**.25**	.03	1.16
46	SPI1	.04	.06	.83	.07	.83	.07	.39	.15	.10	**.24**	.22	1.89
47	IL12A	.95	.02	.01	.02	.95	.02	.47	.22	.03	**.25**	.07	1.29
48	LMO1	.01	.01	.01	.97	.97	.01	.48	.23	.02	**.25**	.04	1.20
49	LMO2	.01	.01	.01	.97	.97	.01	.48	.23	.02	**.25**	.05	1.20
50	**BCL2**	.09	.41	.43	.07	.41	.09	.19	.04	.21	**.16**	.61	3.12
51	IL1R1	.89	.04	.03	.05	.89	.05	.42	.18	.07	**.24**	.15	1.61
52	IL4	.95	.02	.01	.01	.95	.02	.47	.22	.03	**.25**	.06	1.27
53	CCL3	.80	.06	.05	.09	.80	.09	.37	.13	.12	**.23**	.27	2.05
54	**MEIS1**	.13	.45	.12	.30	.30	.13	.16	.02	.23	**.13**	.69	3.46
55	SMO	.05	.82	.09	.05	.82	.09	.38	.14	.11	**.24**	.24	1.95
56	MLH1	.04	.06	.85	.05	.85	.06	.40	.16	.09	**.24**	.20	1.81
57	**MERTK**	.32	.49	.10	.10	.32	.10	.19	.04	.21	**.15**	.62	3.20

Figure 5.17: *The proposed spreadability cut could identify overlapping communities better than bridgeness and exponential entropy (shown in the last two columns respectively). Proteins with fuzzy membership are written in bold and their fuzzy communities are underlined. .*

ID	$U_{1..k}(s)$					σ^2	ϖ
7	.99	.00	.00	.00	.493	.243	.250
45	.97	.01	.01	.01	.483	.233	.250
10	.97	.01	.01	.01	.479	.229	.250
48	.97	.01	.01	.01	.478	.229	.250
49	.97	.01	.01	.01	.477	.228	.249
12	.96	.02	.01	.01	.476	.227	.249
22	.96	.01	.01	.01	.474	.224	.249
5	.96	.02	.01	.01	.470	.221	.249
52	.95	.02	.01	.01	.469	.220	.249
13	.95	.02	.02	.01	.467	.218	.249
47	.95	.02	.02	.01	.466	.217	.249
15	.95	.02	.02	.01	.465	.217	.249
37	.94	.03	.02	.01	.463	.215	.249
8	.94	.02	.02	.01	.463	.214	.249
44	.94	.03	.02	.01	.463	.214	.249
28	.94	.03	.02	.01	.461	.213	.249
11	.94	.02	.02	.02	.457	.209	.248
19	.93	.03	.02	.02	.454	.206	.248
3	.92	.04	.02	.02	.448	.201	.247
20	.92	.03	.03	.02	.445	.198	.247
27	.91	.04	.03	.02	.441	.195	.247
51	.89	.04	.03	.05	.424	.180	.244
40	.88	.05	.04	.03	.422	.178	.244
14	.87	.06	.03	.03	.413	.171	.242
41	.86	.07	.04	.03	.408	.167	.242
39	.85	.06	.05	.04	.401	.161	.240
56	.85	.06	.05	.04	.398	.158	.240
29	.84	.06	.06	.04	.396	.157	.239
38	.84	.06	.05	.04	.394	.155	.239

ID	$U_{1..k}(s)$					σ^2	ϖ
46	.83	.07	.06	.04	.388	.151	.237
35	.82	.09	.05	.04	.379	.144	.235
55	.82	.09	.05	.05	.379	.143	.235
21	.81	.10	.06	.04	.372	.138	.234
53	.80	.06	.05	.09	.365	.133	.232
33	.78	.10	.06	.05	.354	.125	.229
32	.78	.13	.05	.05	.355	.126	.229
24	.77	.08	.07	.07	.350	.122	.227
26	.75	.10	.08	.07	.333	.111	.222
4	.73	.09	.09	.09	.322	.104	.218
25	.70	.12	.10	.09	.297	.088	.209
23	.66	.18	.10	.06	.279	.078	.201
31	.66	.17	.10	.07	.277	.077	.200
6	.64	.20	.08	.07	.267	.071	.196
9	.61	.23	.09	.07	.252	.064	.189
34	.61	.15	.13	.11	.240	.058	.183
17	.61	.15	.13	.11	.240	.057	.182
2	.59	.27	.08	.07	.243	.059	.184
42	.57	.21	.15	.08	.218	.047	.170
30	.54	.24	.14	.09	.200	.040	.160
18	.52	.22	.16	.10	.184	.034	.150
1	.50	.32	.09	.09	.201	.040	.160
57	.49	.32	.10	.10	.191	.036	.154
16	.47	.22	.19	.12	.153	.023	.130
43	.45	.40	.08	.07	.202	.041	.161
54	.45	.30	.13	.12	.156	.024	.132
50	.43	.41	.09	.07	.194	.038	.156
36	.34	.31	.27	.08	.120	.014	.105

Figure 5.18: Sorting U ascending based on nodes fuzziness in Leukemia network, the spreadability always decreases as long as fuzziness increases.

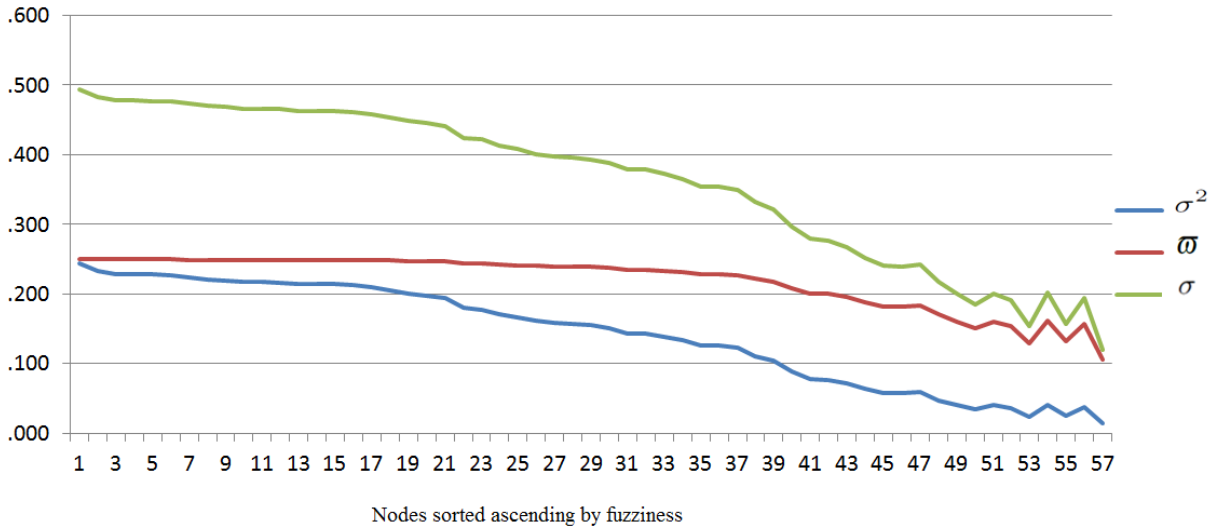


Figure 5.19: Spreadability (ϖ), standard deviation (σ), and variance (σ^2) decrease when fuzziness of nodes increase.

Chapter 6

Stock Market Communities Identification

This chapter will apply the fuzzy and possibilistic approaches to community detection in networks proposed in this thesis to a completely different field: The financial stock market.

In financial stocks, in order to lidentify groups of assets highly affecting the price of each other compared to the rest of the network (i.e., communities) several approaches have been already proposed to identify the structure of the empirical correlation matrix of asset profiles, and to model their reciprocal relationships in terms of hierarchical trees and networks [147, 155, 184]. However, the process of communities identification needs to make use of statistically reliable information, often the correlation matrix is sensitive to several factors, such as the heterogeneity of sampling, the interaction with environment, and the non-stationariety of data sources.

In the following, we will present an analysis of the Italian stock market aimed at identifying companies that act as market drivers, thus influencing the performance in the stock market sector [120].

6.1 Correlation based *FSM*

In order to identify hubs (see Sect. 2.2) and overlapping communities in financial networks, we can apply a correlation based version of *FSM* (see Sect. 3.3) or *PSM* (see Sect. 3.5) that we call *COR-FSM* and *COR-PSM* described in the following steps (see Fig.6.1):

1. Choose the initial time t_0 and the length l of a temporal windows.
2. Within each temporal window, for each asset calculate the corresponding log–returns, defined as:

$$r_h(t) = \ln \frac{p_h(t)}{p_h(t-1)} = \ln p_h(t) - \ln p_h(t-1), \tag{6.1}$$

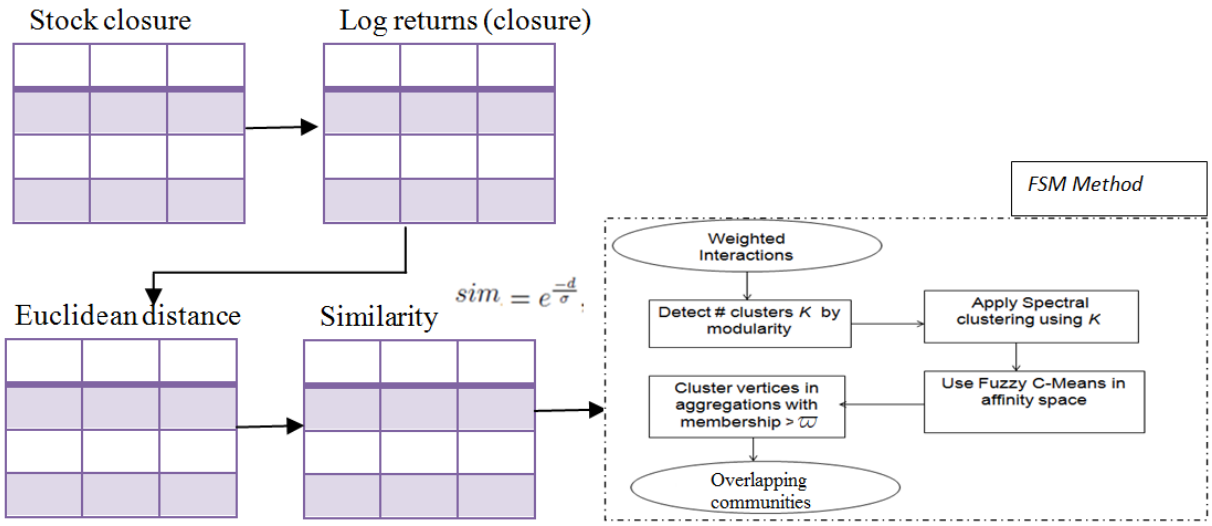


Figure 6.1: The Correlation FSM *(COR-FSM)* Similarity approach.

where $p_h(t)$ is the closure price of asset h at time t ($h = 1,\ldots,N$; $t = 2,\ldots,l$).

3. Normalize the log return data.

4. Calculate the pairwise Euclidean distance d_{hi} between the returns profiles of assets h and i (within each temporal windows), for each pair of assets:

$$d_{hi} = \sqrt{\sum_{p=0}^{l-1} (r_h(t_p) - r_i(t_p))^2}. \tag{6.2}$$

In case of missing observations, the assets are discarded; undefined values are replaced by zero in each asset.

5. Estimate the similarity between asset profiles [161] as:

$$sim_{hi} = \exp\left(\frac{-d_{hi}}{s}\right), \tag{6.3}$$

where s refers to the dispersion of data distribution and can be estimated from histogram analysis.

6. Apply the FSM-community detection method to the asset profile similarities matrix $\Sigma = [sim_{hi}]$.

Table 6.1: Network features of Italian Stock market between 15/3/2004 and 15/3/2005

	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.
Assets	171	171	174	174	174	176	178	179	179	179	180	182
Edges	29241	29241	30276	30276	30276	30976	31684	32041	32041	32041	32400	33124
APL	0.043	0.012	0.013	0.003	0.088	0.067	0.023	0.027	0.128	0.142	0.178	0.054
Q	0.384	0.384	0.469	0.479	0.434	0.425	0.404	0.389	0.384	0.411	0.407	0.416

6.2 Financial network dataset

The approach described in section 6.1 was applied to the closure prices of the Italian Stock Exchange observed in the period from 15 March 2004 to 15 March 2014. Our dataset contains 171 assets actively traded on Milan Stock Exchange (MSE), in addition to 13 evolving assets depicted in Fig. 6.2, classified into 37 categories.

6.3 Experimental study

We used a time window of 30 days, a fuzziness parameter $m = 2$, and a dispersion parameter $s = .5$ experimentally evaluated from histogram analysis.

We used modularity maximization to obtain the number of clusters. In Tab. 6.1 we report the modularity values (Q) of the communities identified together with the number of assets, the number of edges, and the average path length *(APL)* (see Eq. 2.6) in each time window of 30 days.

Tab. 6.2 shows the degree distribution during one year (from 15 March 2004 to 15 March 2005) on monthly basis. It is worth noting that, the degree and the closeness measures can not identify time stable hubs and is not sufficient to characterize its internal structure of this network. On the other hand, the proposed *COR-FSM* could characterize the network dynamical structure.

Figs. 6.3 and 6.4 list the composition of the five communities obtained by running the *COR-FSM* approach, and shows the obtained hubs (in bold) and bridge (denoted as "fuzzy") assets belonging to more than one community. Note that assets belonging to the same category show higher tendency to be grouped together.

Moreover, we remark that some sectors affecting the others, such as electricity (EL), industrial engineering (IE), and Industrial transportation (IT) sectors. Fig. 6.5 shows a detailed three-dimensional representation of the number of assets for each category to the five communities.

To study the stability of obtained communities, we constructed the contingency matrix containing the fuzzy Rand index RI_f (Eq. 2.36) between each pair of temporal windows (30 days) (see

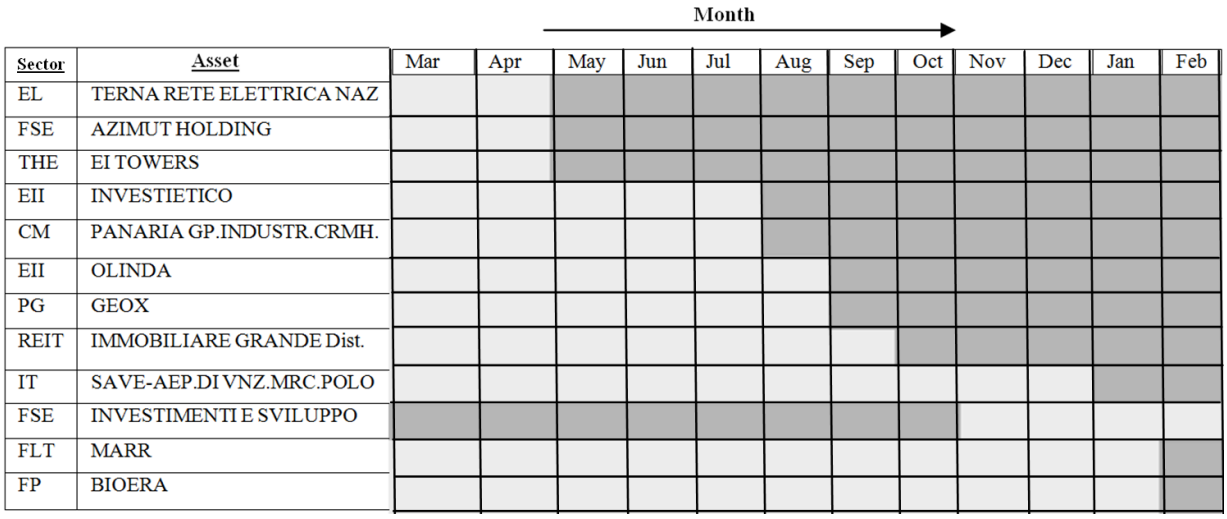


Figure 6.2: Monthly stock market evolution from 15/3/2004 to 15/3/2005. Light cells indicate that an asset observations are missed during this month. Dark cells refer to new assets involved in market.

Table 6.2: Degree distribution of Italian stock market assets in 12 months between 15/3/2004 and 15/5/2005

Month

	1	2	3	4	5	6	7	8	9	10	11	12
10	3	2	4	5	3	1	1	1	1	1	0	2
20	3	2	0	2	1	2	1	1	4	4	3	3
30	4	9	4	3	4	3	1	6	5	6	6	5
40	9	4	2	9	4	2	5	7	7	4	5	4
50	22	26	9	11	9	7	7	14	18	19	17	17
60	37	47	21	24	25	16	11	28	33	27	32	20
70	68	61	41	44	44	27	29	63	76	71	63	41
80	25	20	59	55	65	69	45	53	31	42	46	64
90	0	0	33	21	19	44	56	6	4	5	8	26
100	0	0	1	0	0	5	22	0	0	0	0	0

C1			C2		
Asset Name	sector	Fuzzy	Asset Name	sector	Fuzzy
'A2A'	'EL'		ACEA'	'EL'	2,1
'ENEL'	'EL'		'FRENI BREMBO'	'AUP'	
'FALCK RENEWABLES'	'EL'		'FINMECCANICA'	'AED'	
'SOGEFI'	'AUP'		'KINEXIA'	'AEN'	
'DAVIDE CAMPARI MILANO'	'BEV'		**'BANCA MONTE DEI PASCHI'**	'BKS'	
'AMUNDI RE EUROPA'	'EII'		'BANCA POPOLARE DI MILANO'	'BKS'	
'AMUNDI RE ITALIA'	'EII'		'BANCA PROFILO'	'BKS'	
'BANCA CARIGE'	'BKS'		'BANCO DI SARDEGNA RSP'	'BKS'	
'BANCA PPO.ETRURIA LAZIO'	'BKS'		**'BNC.DI DESIO E DELB.'**	'BKS'	
'BANCA PPO.DI SONDRIO'	'BKS'		'CREDITO EMILIANO'	'BKS'	
'BANCA PPO.EMILIA ROMAGNA'	'BKS'		'MEDIOBANCA BC.FIN'	'BKS'	
'BCA.PICCOLO CDT.VALTELL'	'BKS'		'UNIONE DI BANCHE ITALIAN'	'BKS'	
'ASTALDI'	'CM'		'UNICREDIT'	'BKS'	
'CALTAGIRONE'	'CM'		**'BUZZI UNICEM'**	'CM'	
'ITALCEMENTI FABBRICHE RIUNITE'	'CM'		**'CEMENTIR HOLDING'**	'CM'	
'CEMBRE'	'EEE'		'DATALOGIC'	'EEE'	
'EL EN'	'EEE'		'PRIMA INDUSTRIE'	'EEE'	
'IRCE'	'EEE'		'MITTEL'	'FSE'	
'DEA CAPITAL'	'FSE'		'CENTRALE DEL LATTE DI TRO.'	'FP'	
'BONIFICHE FERRARESI'	'FP'		'IREN'	'GWM'	2,3
'SNAM'	'GWM'		'CENTRO HL DISTRIBUZIONE'	'GR'	
'EMAK'	'HGHC'		'AMPLIFON'	'HCES'	
'CARRARO'	'IE'		'DE LONGHI'	'HGHC'	
'DANIELI'	'IE'		'VINCENZO ZUCCHI'	'HGHC'	
'IMA INDUA.MACCHINE'	'IE'		**'AEROPORTO DI FIRENZE'**	'IT'	
'SABAF'	'IE'		'CATTOLICA ASSICURAZIONI'	'LINS'	
'ATLANTIA'	'IT'		'CALTAGIRONE EDITORE'	'MED'	
'ASTM'	'IT'		'MONDO TV'	'MED'	
'SIAS'	'IE'		'POLIGRAFICI EDITORIALE'	'MED'	
'CAIRO COMMUNICATION'	'MTEL'		'TELECOM ITALIA MEDIA'	'MED'	
MEDIACONTECH'	'MED'	1,5,2	'UNIPOLSAI'	'NLI'	
'UNIPOL GRUPPO FINANZIARI'	'NLI'		'ASSICURAZIONI GENERALI'	'NLI'	
'VITTORIA ASSICURAZIONI'	'NLI'		'GIOVANNI CRESPI'	'PG'	
'ENI'	'OG'		'CSP INTERNATIONAL'	'PG'	
'ERG'	'OG'		BRIOSCHI SVILUPPO IMMBL'	'REIS'	2,1
'CIA'	'REIS'		**'CAD IT'**	'SCS'	
'NOVA RE'	'REIS'		'DADA'	'SCS'	
'RISANAMENTO'	'REIS'		'ENGR.INGEGNERIA INFORMA'	'SCS'	
'BENI STABILI'	'REIS'		'TXT E-SOLUTION'	'SCS'	
'AUTOGRILL'	'TLE'	1,4,5	'ESPRINET'	'THE'	
'SS LAZIO'	'TLE'		'FNM'	'TLE'	
			'GRANDI VIAGGI'	'TLE'	
			'JUVENTUS FOOTBALL CLUB'	'TLE'	
			'SNAI'	'TLE'	

Figure 6.3: Communities $C1$ and $C2$ identified using *FSM* over a sample 30 days long time window. Hubs are labeled by bold, while for fuzzy assets the indices of their overlapping communities are listed.

C3			C4			C5		
Asset Name	sector	Fuzzy	Asset Name	sector	Fuzzy	Asset Name	sector	Fuzzy
ISAGRO'	'CHE'		'ALERION CLEAN POWER'	'EL'	4,1	'FIAT'	'AUP'	
'SOL'	'CHE'		'K R ENERGY'	'EL'	4,1,3	'PIRELLI'	'AUP'	
'CREDITO BERGAMASCO'	'BKS'		'IMMSI'	'AUP'		**'BANCA FINNAT EURAMERICA'**	'BKS'	
'BEGHELLI'	'EEE'		'PININFARINA'	'AUP'		'BANCA PPO.DI SPOLETO'	'BKS'	
'GEFRAN'	'EEE'		'BOERO BARTOLOMEO'	'CM'		'BANCO POPOLARE'	'BKS'	
'BANCA INTERMOBILIARE'	'FSE'		'GRUPPO CERAMICHE RICCHET'	'CM'		**'INTESA SANPAOLO'**	'BKS'	
ACQUE POTABILI'	'GWM'		'TREVI FIN INDUSTRIALE'	'CM'		**'IMPREGILO'**	'CM'	
ACSM-AGAM'	'GWM'		'VIANINI INDR.'	'CM'		'ITALMOBILIARE'	'CM'	
'HERA'	'GWM'		'VIANINI LAVORI'	'CM'		'SAES GETTERS'	'EEE'	
CICCOLELLA'	'GR'	3,5	'BANCA IFIS'	'FSE'		**'RETI TELEMATICHE ITALIAN'**	'FSE'	
BORGOSESIA RSP'	'IE'	3,1	'LVENTURE GROUP'	'FSE'		**'TELECOM ITALIA'**	'FLT'	
'MONRIF'	'MED'		'INVESTIMENTI E SVILUPPO'	'FSE'		**'ARENA'**	'FDR'	
'SEAT PAGINE GIALLE'	'MED'	3,2,5	**'LA DORIA'**	'FP'		'EDISON RSP'	'GWM'	5,2
'SAIPEM'	'MM'		'RENO DE MEDICI'	'GI'		'CIR CIE.INDI.RIUN.'	'GI'	
'BASICNET'	'PG'		**'SORIN'**	'HCES'		**'COFIDE GRUPPO DE BENEDET'**	'GI'	
'RATTI'	'PG'		'BIESSE'	'IE'		'DMAIL GROUP'	'GR'	
'RECORDATI INDUA.CHIMICA'	'PHB'	3,1	'FIDIA'	'IE'		**'INDESIT COMPANY'**	'HGHC'	
'GABETTI PROPERTY SLTN.'	'REIS'		'AUTOSTRADE MERIDIONALI'	'IT'		**'INTERPUMP GROUP'**	'IE'	
'BASTOGI'	'SCS'		FEDON (PAR)'	'PG'		'INTEK GROUP'	'IE'	
FIERA MILANO'	'SSER'		'AEDES LIGURE LOMBARDA'	'PHB'		PREMUDA'	'IT'	5,1
'POLIGRAFICA S F'	'SSER'		**'BEE TEAM'**	'REIT'		**'DIGITAL BROS'**	'LG'	
'AMBIENTHESIS'	'SSER'		'EXPRIVIA'	'SCS'		'MEDIOLANUM'	'LINS'	
'OLIDATA'	'THE'		'FULLSIX'	'SCS'		'ACOTEL GROUP'	'MTEL'	
AS ROMA'	'TLE'		'SINTESI SOCIETA DI INVMI'	'SSER'		**'CLASS EDITORI'**	'MED'	
'GTECH'	'TLE'	3,2				'GRUPPO EDIT.L''ESPRESSO'	'MED'	
						'MEDIASET'	'MED'	
						'ARNOLDO MONDADORI EDI.'	'MED'	
						'RCS MEDIA GROUP'	'MED'	
						'INTEK GROUP'	'MED'	
						'LUXOTTICA'	'PG'	
						'STEFANEL'	'PG'	
						'TOD''S'	'PG'	
						PRELIOS'	'REIS'	5,1
						'REPLY'	'SCS'	
						'TAS TGA.AVANZATA SISTEMI'	'SCS'	
						'TISCALI'	'SCS'	
						'IT WAY'	'THE'	

Figure 6.4: Communities $C3$, $C4$, and $C5$ identified using *FSM* over a sample 30 days long time window. Hubs are labeled by bold, while for fuzzy assets the indices of their overlapping communities are listed.

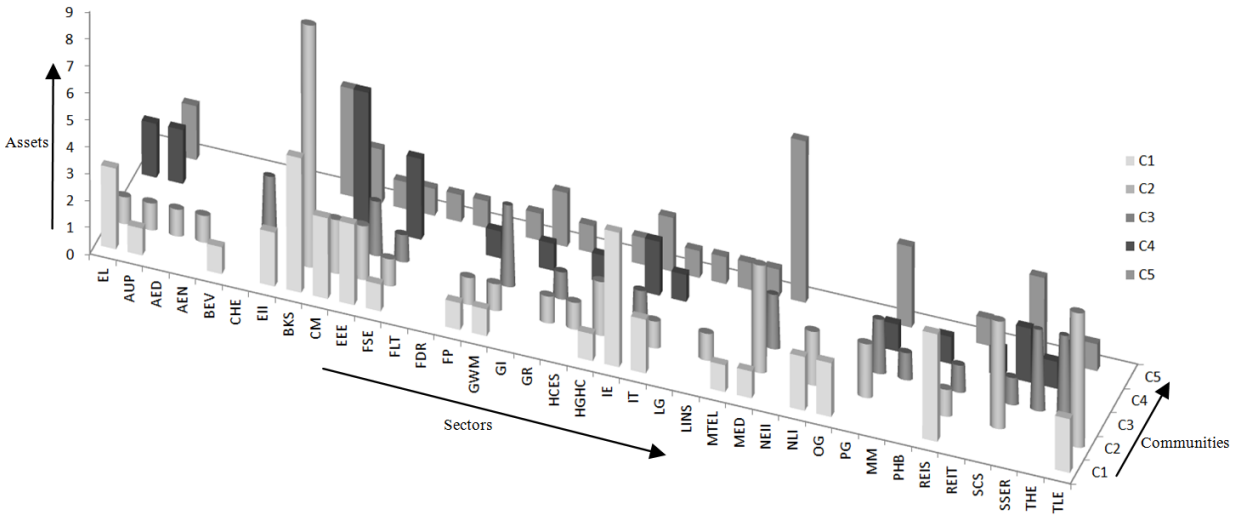


Figure 6.5: Distribution of Italian stock 37 sectors in 5 clusters, as resulting from the *COR-FSM* method over a sample 30 days long time window.

Fig. 6.6). The RI_f shows small variance all over the network time windows. Moreover, the variance is proportional to the network temporal evolution: In facts, generally the closer the time windows are, the higher the fuzzy Rand index between the identified communities.

The stability analysis is obtained in the following way:

1. Perform repeated *FSM* for detecting communities in each temporal window (t) (e.g., 30 days).
2. Evaluate the similarity measure (or its average, in the case of multiple starts) of each pair of detected communities.
3. Arrange the similarity values in a $N_{steps} \times N_{steps}$ similarity matrix.
4. Convert the similarity matrix to a heat map image (see Fig. 6.6).
5. Rank the columns or rows index corresponding to maximum stability.
6. Retrieve the values, corresponding to most correlated market behaviors at these time windows to that t.

We highlight that this approach is general and other indexes may be employed (e.g., the fuzzy Jaccard index, as we used in [160]) instead of the one in use. It is worth noting that our experimental results confirm those observed in [184] and [147], but are more robust to network evolution and missing observations because they are not sensitive to noise artifacts, and moreover, they support overlapping communities.

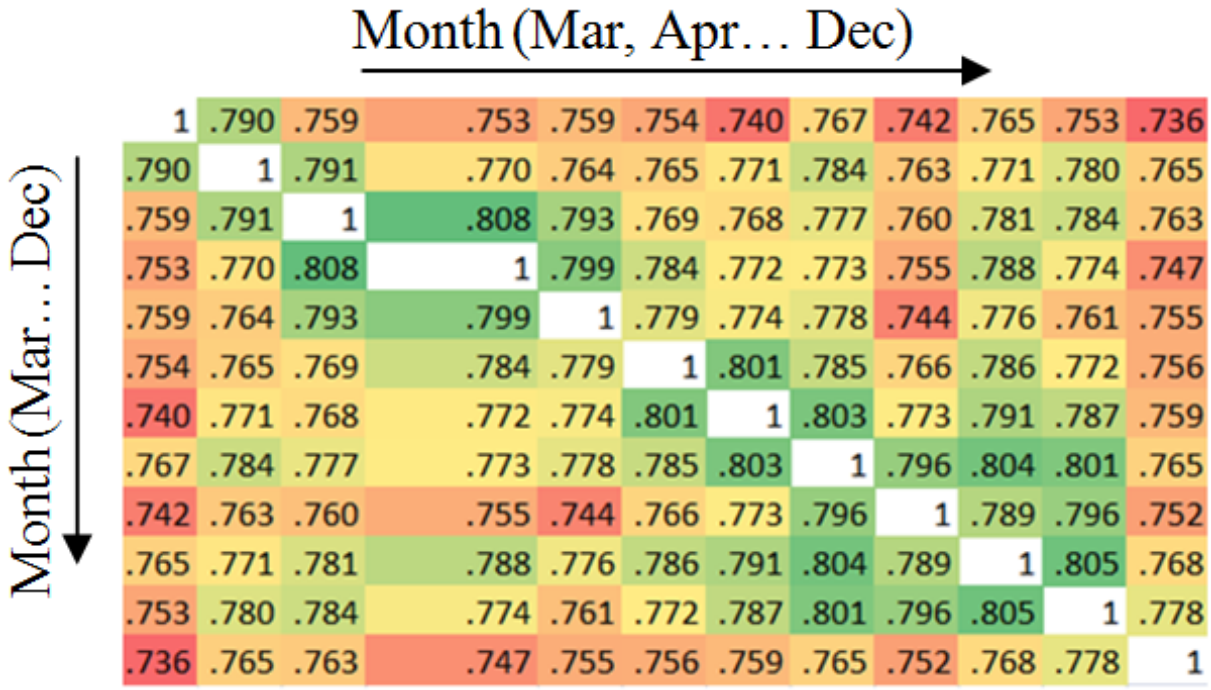


Figure 6.6: Heat map of fuzzy Rand contingency matrix between *FSM* monthly memberships during one year.

Chapter 7

Conclusion

Community detection approaches allow us to extract dense aggregation of nodes from networks. They can be used in knowledge inference of different application domains such as social network analysis, and biological networks. Understanding such communities may help to understand the communication flow and the role played by entities inside networks. These approaches when applied in biology for instance can improve the research and development of drug discovery, motif finding, and cancer understanding.

This thesis proposes novel community detection approaches that can characterize overlapping and temporal communities. The proposed approaches can be categorized as spectral, modularity, fuzzy (see Sect. 3.3), possibilistic (see Sect. 3.5), ensemble (see Sect. 3.6), semantic (see Sects. 3.7 and 3.8), and dynamic based approaches (see Sect. 6.1).

Moreover, this thesis reviews and analyzes the graph based measures for analyzing networks (see Chapter 2), in addition to introducing a new measure called spreadability (see Sect. 3.4) that is employed in finding a significant membership cut of the fuzzy and possibilistic approaches proposed.

In addition to comparing the proposed methods with state of the art disjoint (crisp) and overlapping community detection approaches in terms of their accuracy and complexity (see Chapter 4 and Sect. 2.4). To this aim, this thesis evaluates the proposed methods in social [200], and ecological [117, 118] real world networks. In addition to testing them on syntactic benchmarks such as the planted ℓ-partition model [31].

The ensemble procedure presented in this thesis (see Sect. 3.6) leaves room for several developments. Mixing coefficients are computed as (normalized) average qualities; other aggregation schemes may be devised, where different computations of the μ_F may be more appropriate. The ensemble and semantic similarity measures proposed here can be applied even in the absence of an explicitly computable primary measure. They can be used to turn an asymmetric similarity

structure into a symmetric, positive definite similarity matrix even for non-Euclidean, and fuzzy data.

Acknowledging that no clustering paradigm is adequate for all data sets, the methods presented in this thesis help in deciding the most suitable representation scheme on an empirical basis, by evaluating the quality of clusterings and by providing an ensemble method that balances its components according to their clustering quality, as measured by quality indices (see Sect. 2.5).

The proposed methods were applied in characterizing the overlapping and evolving biological communities, such as *S. cerevisiae* (see Section 5.2) protein-protein interaction network and in homo sapiens *HIV-1* (see Sect. 5.3.1), leukemia (see Sect. 5.3.2) protein-protein interaction networks using protein pathways and Gene ontology terms.

This experimental study showed that the proposed methods allowed us to infer the communities not only relying on the topological structure of interactome or the biological information, but also based on the semantic enrichment entailed in Gene Ontology and protein pathways. The analysis done in this study found that *PPI* in *HIV-1* and Leukemia significantly related to each other [122, 119]. Those results demonstrated that our proposed approach can boost the functional significance of the identified communities in overlapping protein interaction environments. Moreover, this enabled us to find a significant semantic overlap between the detected communities of *HIV-1* and leukemia networks.

The proposed methods were applied in other different domains such as analyzing temporal communities of financial networks (see Chapter. 6). To this aim the proposed *PSM* and *FSM* overlapping community detection methods were enhanced to maintain the similarity between interacting nodes profile, with an application in Italian stock market, and to infer significant communities on them based on graph theory and on fuzzy spectral clustering.

Bibliography

[1] Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data. Data Mining and Knowledge Discovery, 11(1), 5-33, (2005).

[2] Anderson, D. T., Bezdek, J. C., Popescu, M., Keller, J. M.: Comparing fuzzy, probabilistic, and possibilistic partitions. Fuzzy Systems, IEEE Transactions on, 18(5), 906–918, (2010)

[3] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Yeh, L. S. L.: UniProt: the universal protein knowledgebase. Nucleic acids research, 32(1), 115–119, (2004)

[4] Ashburner, M., Ball, C. A., Blake, J. A. et al.: Gene ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29, (2000)

[5] Asuncion, A., Newman, D.J.: UCI machine learning repository, (2007)

[6] Ayad, H. G., Kamel, M. S.: On voting-based consensus of cluster ensembles. Pattern Recognition, 43(5), 1943–1953, (2010).

[7] Azuaje, F., Wang, H., Bodenreider, O.: Ontology-driven similarity approaches to supporting gene functional assessment. In Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies, 9-10, June, (2005)

[8] Bach, F.R., Jordan, M.I.: Learning spectral clustering. Tech. Rep. UCB/CSD-03-1249, EECS Department, University of California, Berkeley, (2003)

[9] Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Machine Learning **56**(1-3), 89–113, (2004)

[10] Barabàsi, A. L., Albert, R.: Emergence of scaling in random networks. science, 286(5439), 509–512, (1999)

[11] Baraldi, A., Blonda, P.: A survey of fuzzy clustering algorithms for pattern recognition. I. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), 29, 778–785, (1999)

[12] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., GriffithsJones, S., et al.: The Pfam protein families database. Nucleic acids research, 32(1), 138–141, (2004)

[13] Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., Devignes, M. D.: IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC bioinformatics, 11(1), 588, (2010)

[14] Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA, USA, (1981)

[15] Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10, 10008, (2008)

[16] Bohm, C., Railing, K., Kriegel, H. P., Kroger, P.: Density connected clustering with local subspace preferences. IEEE 13th International Conference on Data Mining (ICDM04), 27–34, doi:10.1109/ICDM.2004.10087, (2004)

[17] Bonacich, P.: Power and centrality: A family of measures. American journal of sociology 92, 1170–1182, (1987)

[18] Botstein, D., Chervitz, SA, Cherry, JM: Yeast as a model organism. Science 277(5330), 1259–1260, (1997)

[19] Brandes, U.: On variants of shortest-path betweenness centrality and their generic computation. Social networks 30, 136–145, (2008)

[20] Brandes, U.: A faster algorithm for betweenness centrality. Journal of mathematical sociology 25(2), 163–177, (2001)

[21] Breiman, L.: Bagging predictors. Machine Learning. 24 (2), 123–140, (1996)

[22] Brouwer, R. K.: Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. Journal of Intelligent Information Systems, 32(3), 213–235, (2009)

[23] Budanitsky, A., Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resource,2, (2001).

[24] Bulotta, S., Mahmoud, H., Masulli, F., Palummeri, E., Rovetta, S.: Fall Detection Using an Ensemble of Learning Machines. 22nd Italian Workshop on Neural Networks, WIRN2012, Vietri, Italy, Neural Nets and Surroundings - Smart Innovation, Systems and Technologies, Springer, 19, 81–90, (2013)

[25] Buckingham, Steven D.: Data mining for protein–protein interactions in invertebrate model organisms. Invertebrate Neuroscience, 5(3-4), 183–187, (2005)

[26] Campello, R. J.: A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters, 28(7), 833–841, (2007)

[27] Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Apweiler, R.: The Gene Ontology annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. Nucleic acids research, 32(1), 262–266, (2004)

[28] Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Żak, S.: Complete gradient clustering algorithm for features analysis of x-ray images. In: Information Technologies in Biomedicine. Springer 15–24, (2010)

[29] Chaudhuri, K., Chung, F., Tsiatas, A.: Spectral clustering of graphs with general degrees in the extended planted partition model. Journal of Machine Learning Research, 1–23, (2012)

[30] Chung, F.R.K.: Spectral Graph Theory Washington Conference Board of the Mathematical Sciences, 92, 849–856, (1997)

[31] Condon, A., Karp, R. M.: Algorithms for graph partitioning on the planted partition model. Random Structures and Algorithms, 18(2), 116–140, (2001)

[32] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S.: The genetic landscape of a cell. Science 327(5964), 425–431, (2010)

[33] Couto, F. M., Silva, M. J., Coutinho, P. M.:Measuring semantic similarity between Gene Ontology terms. Data and Knowledge Engineering 61(1), 137–152, (2007)

[34] Couto, F. M., Silva, M. J., Coutinho, P. M.: Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05), 343–344, (2005)

[35] Couto, F. M., Silva, M. J., Coutinho, P. M.: Measuring semantic similarity between Gene Ontology terms. Data & knowledge engineering, 61(1), 137–152, (2007)

[36] Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Trans. Info. Theory **IT-13**, 21–27, January, (1967)

[37] Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines (and Other Kernel-based Learning Methods). Cambridge University Press, (2000)

[38] Davies, D.L., Bouldin, D.W.: A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1(2), 224–227, (1979)

[39] de Silva, V., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In Becker, S., Thrun, S., Obermayer, K., eds.: Proceedings NIPS, 15, 721–728, (2003)

[40] Devos, D., Valencia, A.: Intrinsic errors in genome annotation, TRENDS in Genetics, 17(8), 429-431, (2001)

[41] De Las Rivas, J., Fontanillo, C.: Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. PLOS comput biol, 6(6), 1–7, e1000807, doi:10.1371/journal.pcbi.1000807, (2010)

[42] Dimitriadou, E., Weingessel, A., Hornik, K.: Voting-merging: An ensemble method for clustering. In Artificial Neural Networks—ICANN 2001, Vienna, Springer Berlin Heidelberg. , 217–224, (2001)

[43] Ding, C. H., He, X., Zha, H., Gu, M., Simon, H. D.: A min-max cut algorithm for graph partitioning and data clustering. Proceedings IEEE International Conference on Data Mining (ICDM01), 107–114, (2001)

[44] Donath, W. E., Hoffman, A. J.: Lower bounds for the partitioning of graphs. IBM journal of research and development 17(5964), 420-425, (1973)

[45] Donetti, L., Munoz, M. A.: Detecting network communities: a new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, 2004(10), P10012, (2004)

[46] Dongen, S.: Performance criteria for graph clustering and Markov cluster experiments, Technical Report, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, (2000)

[47] du Plessis, L., Škunca, N., Dessimoz.: The what, where, how and why of gene ontology—a primer for bioinformaticians. Briefings in bioinformatics, bbr002, C, (2011)

[48] Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley and Sons, New York (USA),3, (1973)

[49] Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics, 19(9), 1090-1099, (2003)

[50] Duin, R.P., Pękalska, E.: The dissimilarity space: Bridging structural and statistical pattern recognition. Pattern Recognition Letters, **33**(7),826 – 832, (2012)

[51] Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 3, 32–57, (1974)

[52] Dunn, J. C.: Some recent investigations of a new fuzzy partitioning algorithm and its application to pattern classification problems. Journal of cybernetics 4(2), 1–15, (1974)

[53] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863–14868, (1998)

[54] Erdős, P., Rényi, A.: On random graphs. Publicationes Mathematicae Debrecen, 6, 290–297, (1959)

[55] Fern, X. Z., Brodley, C. E.: Solving cluster ensemble problems by bipartite graph partitioning. In Proceedings of the twenty-first international conference on Machine learning , ACM, p. 36, (2004)

[56] Fiedler, M.: Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, 23(2), 298–305, (1973)

[57] Filippone, M., Masulli, F., Rovetta, S.: Clustering in the membership embedding space. International Journal of Knowledge Engineering and Soft Data Paradigms **1**(4), 363–375, (2009)

[58] Filippone, M., Masulli, F., Rovetta, S.: Comparing Fuzzy Approaches Biclustering. Lecture notes on Computer Science, 5488, 91–101, ISSN: 0302-9743, (2009)

[59] Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern recognition 40(1), 176–190, ISSN: 0031–3203, (2008)

[60] Fischer, B., Buhmann, J.M.: Bagging for path-based clustering. IEEE Transactions on Pattern Analysis Machine Intelligence 25 (11), 1411–1415, (2003)

[61] Fischer, I., Poland, J.: New methods for spectral clustering. Tech. rep., IDSIA/USI-SUPSI, (2004)

[62] Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annual Eugenics **7, part II**, 179–188, (1936)

[63] Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. Project Number 21-49-004 4, USAF School of Aviation Medicine, Randolph Field, Texas, USA, Feb., (1951)

[64] Fowlkes, E. B., Mallows, C. L.: A method for comparing two hierarchical clusterings. Journal of the American statistical association,78(383), 553-569, (1983).

[65] Fortunato, S.: Community detection in graphs. Physics Reports, 486(3), 75–174, (2010)

[66] Fred, A. L., Jain, A. K.: Combining multiple clusterings using evidence accumulation. Pattern Analysis and Machine Intelligence, IEEE Transactions, 27(6), 835–850, (2005)

[67] Fred, A.L.N., Jain, A.K.: Data clustering using evidence accumulation. Pattern Recognition, 4, (2002)

[68] Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: a measure of betweenness based on network flow. Social networks 13(2), 141–154, (1991)

[69] Freeman, L. C.: A set of measures of centrality based on betweenness. Sociometry, 40,35–41, (1977)

[70] Frigui, H., Hwang, C., Chung-Hoon Rhee, F.: Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition, 40(11), 3053–3068, (2007)

[71] Girvan, M., Newman, M. E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826,(2002)

[72] Geiduschek, E. P., Kassavetis, G. A.: The RNA polymerase III transcription apparatus. J mol biol 310(1), 1–26, (2001)

[73] Gower, J.C., Ross, G.J.S.: Minimum spanning trees and single linkage cluster analysis. Journal of the Royal Statistical Society. Series C (Applied Statistics) 18(1), 54–64 (1969)

[74] Goyama, S., Schibler, J., Cunningham, L., Zhang, Y., Rao, Y., Nishimoto, N., Mulloy, J. C.: Transcription factor RUNX1 promotes survival of acute myeloid leukemia cells. The Journal of clinical investigation, 123(9), 3876, (2013)

[75] Gregory, S. Fuzzy overlapping communities in networks. Journal of Statistical Mechanics: Theory and Experiment, 2, (2011)

[76] Gregory, S. . Finding overlapping communities in networks by label propagation. New J. Phys. 12, 10, (2010)

[77] Gregory, S.: Finding overlapping communities using disjoint community detection algorithms, CompleNet 207, 47–61, (2009)

[78] Gregory, S.: A fast algorithm to find overlapping communities in networks. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, 5211, Springer, 408–423, (2008)

[79] Gregory, S.:An Algorithm to Find Overlapping Community Structure in Networks. In Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD 2007, Springer-Verlag, 91–102, (2007)

[80] Guimerra, R., Amaral, L. A. N.:Functional cartography of complex metabolic networks. Nature, 433, 895–900, (2005)

[81] Guo, X., Liu, R., Shriver, C. D., Hu, H., Liebman, M. N.: Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics, 22(8), 967–973, (2006)

[82] Guzzi, P. H., Mina, M., Guerra, C., Cannataro, M.: Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in bioinformatics, 13(5), 569-585, (2012)

[83] Hage, P. and Harary, F.: Eccentricity and centrality in networks. Social Networks 17(1),57–63, (1995)

[84] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: part I. ACM Sigmod Record, 31(2), 40-45.(2002)

[85] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: part II. ACM Sigmod Record, 31(3), 19–27, (2002)

[86] Hartigan, J. A.; Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society, Series C Applied Statistics 28(1), 100–108, JSTOR 2346830, (1979)

[87] Havens, T. C., Bezdek, J. C., Leckie, C., Ramamohanarao, K., Palaniswami, M.: A soft modularity function for detecting fuzzy communities in social networks. IEEE Transactions on Fuzzy Systems, 21(6), 1170–1175, (2013)

[88] Hein, M., Audibert, J. Y., Von Luxburg, U.: Graph Laplacians and their convergence on random neighborhood graphs. arXiv preprint math/0608522, (2006)

[89] Hüllermeier, E., Rifqi, M.: A Fuzzy Variant of the Rand Index for Comparing Clustering Structures. IFSA/EUSFLAT Conf., 1294-1298, (2009)

[90] Minaei-Bidgoli, B., Topchy, A., Punch, W.F.: A comparison of resampling methods for clustering ensembles. In International Conference on Machine Learning, Models, Technologies and Applications (MLMTA 2004), 939–945, (2004)

[91] Hutter, J. J.. Childhood leukemia. Pediatrics in Review, 31, 6, 234–241, (2010)

[92] Jain, A. K., Dubes, R. C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey, USA, (1988)

[93] Jain, S., Bader, G. D.: An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC bioinformatics, 11(1), 562, (2010)

[94] Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE Transactions on Computers **C22**, 1025–1034, (1973)

[95] Jia, J., Xiao, X., Liu, B., Jiao, L.: Bagging-based spectral clustering ensemble selection. Pattern Recognition Letters, 32(10), 1456–1467, (2011)

[96] Jiang, J. J.,Conrath, D. W.:Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In International Conference on Research in Computational Linguistics, ROCLING X, 19–33, (1997)

[97] Keller, J.M., Gray, M.R., Givens, J.A.: A fuzzy k-nearest neighbor algorithm. IEEE Transactions on Systems, Man and Cybernetics, (4), 580–585, (1985)

[98] Kernighan, B. W. , Lin, S.: An efficient heuristic procedure for partitioning graphs, Bell System Technical Journal, 49(2),291–307, (1970) doi:10.1002/j.1538-7305.1970.tb01770.x.

[99] Koschützki D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., Zlotowski, O.: Centrality Indices, In Brandes, U. and Erlebach, T. (Eds.) Network Analysis: Methodological Foundations. (LNCS), Springer-Verlag., 3418,16–61, (2005)

[100] Krause, A. et al: Large scale hierarchical clustering of protein sequences. BMC Bioinformatics, 6,6–15, (2005)

[101] Kriegel, H.P., Kröger, P., Sander, J., Zimek, A.: Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(3), 231–240, (2011)

[102] Krishnapuram, R., Keller, J. M.: The possibilistic c-means algorithm: insights and recommendations, IEEE Transactions on Fuzzy Systems, 4(3), 385–393, (1996)

[103] Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems, 1(2), 98–110, (1993)

[104] Krogan, N. et al: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637–643, (2006)

[105] Kuhn, H. W.: The Hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2), 83–97, (1955)

[106] Kuncheva, L.: Combining pattern classifiers. Methods and Algorithms. Wiley, Chichester, (2004)

[107] Lancichinetti, A., Fortunato, S., Kertesz, J.: Detecting the overlapping and hierarchical community structure of complex networks. New Journal of Physics, 11(033015), (2009)

[108] Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In WordNet: An electronic lexical database MIT Press, 265–283, (1998)

[109] Li, B., Wang, J. Z., Feltus, F. A., Zhou, J., Luo, F.: Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins. arXiv preprint arXiv:1001.0958, (2010)

[110] Li, Y., Bandar, Z., McLean, D: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, 15, 871–882, (2003)

[111] Lin, D.: An information-theoretic definition of similarity In J. Shavlik (Ed.), Fifteenth International Conference on Machine Learning, ICML Madison, 296–304, (1998)

[112] Lin, M. K., Lee, Y. J., Lough, T. J., Phinney, B. S., Lucas, W. J.: Analysis of the pumpkin phloem proteome provides insights into angiosperm sieve tube function. Molecular and Cellular Proteomics, 8(2), 343–356, (2009)

[113] Liu, J.: Fuzzy modularity and fuzzy community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems, 77(4), 547–557, (2010)

[114] Lloyd, S. P.: Least square quantization in PCM, Bell Telephone Laboratories, Murray Hill (1957) Reprinted in: IEEE Transactions on Information Theory, 28(2), 129–137, (1982)

[115] Lord, P. W., Stevens, R. D., Brass, A., Goble, C. A.: Semantic similarity measures as tools for exploring the gene ontology. In Pacific Symposium on Biocomputing, 8, 601–612, (2003)

[116] Lord, P. W., Stevens, R. D., Brass, A., Goble, C. A.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 19(10), 1275–1283, (2003)

[117] Lusseau, D. The emergent properties of a dolphin social network. Proceedings of the Royal Society of London B: Biological Sciences, 270(2), 186–188, (2003)

[118] Lusseau, D., Newman, M. E. Identifying the role that animals play in their social networks. Proceedings of the Royal Society of London. Series B: Biological Sciences, 271(6), 477–481, (2004)

[119] Mahmoud, H., Masulli, F., Rovetta, S., Russo, G.: Detecting overlapping protein communities in disease networks. Lecture Notes in Computer Science (LNCS), Computational Intelligence Methods for Bioinformatics and Biostatistics, (in press)

[120] Mahmoud, H., Masulli, F., Marina, R., Rovetta, S., Abdulatif, A.: Hubs and Communities Identification in Dynamical Financial Networks. Neural Nets and Surroundings Smart Innovation, Systems and Technologies, 24th Italian workshop on Neural Networks, WIRN2014, Vietri sul Mare, Salerno, Italy (in press)

[121] Mahmoud, H., Masulli, F., Rovetta, S., Russo, G.: Identifying overlapping Interactome communities using spectral and semantic approaches. In Quarta Giornata Ligure di Bioinformatica (GLIB2014). Congress Centre CBA IRCCS AOU San Martino IST, Genoa, Italy, Dec 19, (2014)

[122] Mahmoud, H., Masulli, F., Rovetta, S., Russo, G.: Exploiting Quantitative and Semantic Information in Protein-Protein Interactions Networks Analysis. Computational Intelligence Methods for Bioinformatics and Biostatistics-11th International Meeting (CIBB 2014), Cambridge, UK, June 26-28, (2014)

[123] Mahmoud, H., Masulli, F., Rovetta, S., Russo, G.: Characterizing evolving protein communities. Poster presentation, Bioinformatics Italian society eleventh annual meeting (BITS 2014), Rome, Italy, PP 80, Feb 26-28, (2014) http://bits2014.uniroma2.it/

[124] Mahmoud H., Masulli F., Rovetta S.: Finding fuzzy biological and ecological aggregations in spectral space, Poster presentation, Giornata Ligure di Bioinformatica 2013 (GLIB2013), Genova, Italy, Nov, (2013)

[125] Mahmoud, H., Masulli, F., Rovetta, S., Russo, G.: Community Detection in Protein-Protein Interaction Networks Using Spectral and Graph Approaches. Computational Intelligence Methods for Bioinformatics and Biostatistics-10th International Meeting (CIBB 2013), France, Lecture Notes in Bioinformatics (LNBI), Springer, (2013)

[126] Mahmoud, H., Masulli, F., Rovetta, S., Russo, G.: Community detection in Protein-Protein Interaction networks. Computational Intelligence Methods for Bioinformatics and Biostatistics-10th International Meeting (CIBB 2013), Nice, France, Jun 19-22, (2013)

[127] Mahmoud, H., Masulli, F., Rovetta, S.: A Fuzzy Clustering Segmentation Approach for Feature-Based Medical Image Registration. Ninth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, CIBB2012, Houston,Texas,USA, Lecture Notes in Computer Science (LNCS) - Springer, 7845, 37–47, (2013)

[128] Mao, X., Cai, T., Olyarchuk, J. G., Wei, L.: Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics, 21(19), 3787–3793, (2005)

[129] Masulli, F., Rovetta, S.: Soft transition from probabilistic to possibilistic fuzzy clustering. IEEE Transactions on Fuzzy Systems, 14(4), 516–527, (2006)

[130] Mazandu, G. K., Mulder, N. J.: Information content-based gene ontology functional similarity measures: which one to use for a given biological data type?. PloS one, 9(12), e113859, (2014)

[131] Mazandu, G. K., Mulder, N. J.: Information content-based gene ontology semantic similarity approaches: toward a unified framework theory. BioMed research international, (2013)

[132] Mazandu, G. K., Mulder, N. J..: DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content measures. BMC bioinformatics, 14(1), 284, (2013)

[133] Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Weil, B., et al.: MIPS: a database for genomes and protein sequences. Nucleic acids research, 30(1), 31–34, (2002)

[134] Mistry, M., Pavlidis, P.: Gene Ontology term overlap as a measure of gene functional similarity. BMC bioinformatics, 9(1), 327, (2008)

[135] Maier, M., Hein, M., Von Luxburg, U.: Cluster identification in nearest-neighbor graphs. In Algorithmic Learning Theory. Springer Berlin Heidelberg, 196–210, (2007)

[136] Mazandu, G.K., Mulder, N.J.: A Topology-Based Metric for Measuring Term Similarity in the Gene Ontology. Advances in Bioinformatics, (2012)

[137] Meila, M., Shi, J.: A random walks view of spectral segmentation. Artificial Intelligence and Statistics (AISTATS), (2001)

[138] Monti, S., Tamayo, P., Mesirov, J. , Golub, T.: Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data. Machine Learning, 52, 91–118, (2003)

[139] Morozova, O., Marra, M. A.: Applications of next-generation sequencing technologies in functional genomics. Genomics, 92(5), 255–264, (2008)

[140] Nadler, B., Galun, M.: Fundamental limitations of spectral clustering. Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 19, 1017–1024, (2007)

[141] Nepusz, T., Petróczi, A., Négyessy, L., Bazsó, F.: Fuzzy communities and the concept of bridgeness in complex networks. Physical Review E, 77(1), 016107, (2008)

[142] Newman, M.E.J.: Detecting community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems, 38(2), 321–330, (2004)

[143] Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113, (2004)

[144] Newman, M. E. J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences (PNAS), 103(23) 8577–8582, (2006)

[145] Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14. MIT Press, Cambridge, MA, 849–856, (2002)

[146] Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri, M.: Extending the definition of modularity to directed graphs with overlapping communities. Journal of Statistical Mechanics: Theory and Experiment, 03, P03024, (2009)

[147] Onnela, J. P, Kaski, K., Kertész, J.: Clustering and information in correlation based financial networks. The European Physical Journal B-Condensed Matter and Complex Systems 38(2), 353–362, (2004)

[148] Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435(7043), 814–818, (2005)

[149] Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcào, A. O., Couto, F. M.: Metrics for GO based protein semantic similarity: a systematic evaluation. BMC bioinformatics, 9(5), S4, (2008)

[150] Pesquita, C., Faria, D., Bastos, H., Falcào, A., Couto, F.: Evaluating go-based semantic similarity measures. In Proc. 10th Annual Bio-Ontologies Meeting, 37(40), 38, July, (2007)

[151] Pesquita, C., Pessoa, D., Faria, D., Couto, F.: CESSM: Collaborative evaluation of semantic similarity measures. JB2009: Challenges in Bioinformatics, 157, (2009)

[152] Rada, R., Mili, H., Bichnell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 9, 17–30, (1989)

[153] Radicchi, F.,Castellano, C., Cecconi, F., Loreto, V., Parisi,D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA, 101, 2658–2663, (2004)

[154] Rand, W. M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), 846–850, (1971)

[155] Resta, M.: On a data mining framework for the identification of frequent pattern trends. C. Perna and M. Sibillo (Edrs): Mathematical and Statistical Methods for Actuarial Sciences and Financial Markets, Springer International Publishing, 173–176, (2014)

[156] Resnik, P. : Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In C. S. Mellish (Ed.), 14th International Joint Conference on Artificial Intelligence, IJCAI, 1, 448–453, (1995)

[157] Ridella, S., Rovetta, S., Zunino, R.: K-winner machines for pattern classification. IEEE Transactions on Neural Networks, **12**(2), 371–385, March, (2001)

[158] Ridella, S., Rovetta, S., Zunino, R.: IAVQ–interval-arithmetic vector quantization for image compression. IEEE Transactions on Circuits and Systems, Part II **47**(12), 1378–1390, December, (2000)

[159] Rivals, I., Personnaz, L., Taing, L., Potier, M. C.: Enrichment or depletion of a GO category within a class of genes: which test?. Bioinformatics, 23(4), 401–407, (2007)

[160] Rovetta, S., Masulli, F.: Visual stability analysis for model selection in graded possibilistic clustering. Information Sciences, 279, 37–51, (2014)

[161] Rovetta, S., Masulli, F., Mahmoud, H.: Neighbor-based similarities. Tenth International Workshop on Fuzzy Logic and Applications (WILF 2013), Lecture Notes in Computer Science (LNCS) - Springer, 8256, 161–170, (2013)

[162] Rovetta, S., Masulli, F.: Shared farthest neighbor approach to clustering of high dimensionality, low cardinality data. Pattern Recognition, **39**(12), 2415–2425, December, (2006)

[163] Rovetta, S., Masulli, F.: Vector quantization and fuzzy ranks for image reconstruction. Image and Vision Computing **25**(2), 204–213, (2006)

[164] Rodriguez, M. A., Egenhofer, M. J.: Determining semantic similarity among entity lasses from different ontologies. IEEE Transactions on Knowledge and Data Engineering, 15, 442–456, (2003)

[165] Rosasco, R., Mahmoud, H., Rovetta, S., Masulli, F.:A quality-driven ensemble approach to automatic model selection in clustering. 23nd Italian Workshop on Neural Networks, WIRN2013, Vietri, Italy, (2013)

[166] Sabidussi, G.: The centrality index of a graph. Psychometrika, 31(4), 581–603, (1966)

[167] Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications, 39(9), 7718-7728, (2012)

[168] Sánchez, D., Batet, M., Isern, D.: Ontology-based information content computation. Knowledge-Based Systems, 24(2), 297-303, (2011)

[169] Schapire, R.E.: The strength of weak learn ability. Machine Learning. 5 (2), 197–227, (1990)

[170] Schlicker, A., Domingues, F. S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics, 7(1), 302, (2006)

[171] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Wanker, E. E.,et.al: A human protein-protein interaction network. a resource for annotating the proteome. Cell, 122(6), 957-968, (2005)

[172] Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A.,et al.: Correlation between gene expression and GO semantic similarity. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2(4), 330-338, (2005)

[173] Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888–905, (2000)

[174] Shimbel, A.: Structural parameters of communication networks. The Bulletin of Mathematical Biophysics, 15(4), 501–507, (1953)

[175] Sobolevsky, S., Campari, R., Belyi, A., Ratti, C.: General optimization technique for high-quality community detection in complex networks. Physical Review E, 90(1), 012811, (2014)

[176] Steinhaus, H.: Sur la division des corp materiels en parties. Bull. acad. Polon. sci, 1, 801–804, (1956)

[177] Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res., 3, 583–617, (2003)

[178] Strehl, A.,Ghosh, J.: Cluster ensembles – A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3, 583–618, (2002)

[179] Tao, Y., Sam, L., Li, J., Friedman, C., Lussier, Y. A.: Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics, 23(13), 529–538, (2007)

[180] Tastan, O., Qi, Y., Carbonell, JG., Klein-Seetharaman, J.: Prediction of interactions between HIV-1 and human proteins by information integration. Pacific Symposium on Biocomputing, 14, PubMed PMID: 19209727; PubMed Central PMCID: PMC3263379, 516–527, (2009)

[181] Timm, H., Borgelt, C., Döring, C., Kruse, R.: Fuzzy cluster analysis with cluster repulsion, in: Proc. European Symposium on Intelligent Technologies (EUNITE, Tenerife, Spain), on CDROM. Verlag Mainz, Aachen, Germany, Citeseer, (2001)

[182] Topchy, A., Jain, A. K., Punch, W.: Clustering ensembles: Models of consensus and weak partitions. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(12), 1866–1881, (2005)

[183] Tumer, K., Agogino, A. K.: Ensemble clustering with voting active clusters. Pattern Recognition Letters, 29(14), 1947–1953, (2008)

[184] Tumminello, M., Coronnello, C., Lillo, F., Micciche, S., Mantegna, R. N.: Spanning trees and bootstrap reliability estimation in correlation-based networks. International Journal of Bifurcation and Chaos, 17(07), 2319–2329, (2007)

[185] Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 25(03), 337–372, (2011)

[186] Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing, 17(4), 395–416, (2007)

[187] Von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., Bork, P.: STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res, 33(1), 433–437, (2005)

[188] Wagner, D., Wagner, F.: Between min cut and graph bisection, Springer Berlin Heidelberg, 744-750, (1993)

[189] Wang, H., Azuaje, F., Bodenreider, O., Dopazo, J.: Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on IEEE, 25—31, Oct., (2004)

[190] Wang, J. Z., Du, Z., Payattakool, R., Philip, S. Y., Chen, C. F.: A new method to measure the semantic similarity of GO terms. Bioinformatics, 23(10), 1274-1281, (2007)

[191] Watts, D. J., Strogatz, S.: Collective dynamics of small world networks. Nature, 393(6684), 440–442, (1998)

[192] Wu, X., Zhu, L., Guo, J., Zhang, D. Y., Lin, K.: Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. Nucleic acids research, 34(7), 2137-2150, (2006)

[193] Wu, Z., Palmer, M.: Verb semantics and lexical selection. In 32nd annual Meeting of the Association for Computational Linguistics, 133–138, Las Cruces, New Mexico: Association for Computational Linguistics, (1994)

[194] Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on pattern analysis and machine intelligence, 13(8), 841–847, (1991)

[195] Xie, J., Kelley, S., Szymanski, B. K.: Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Computing Surveys (CSUR), 45(4), 43, (2013)

[196] Xu, T., Du, L., Zhou, Y.: Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC bioinformatics, 9(1), 472, (2008)

[197] Yang, J., Leskovec, J..: Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems, 42(1), 181–213, (2015)

[198] Yager, R.R., Detyniecki, M., Bouchon-Meunier, B.: A context-dependent method for ordering fuzzy numbers using probabilities. Information Sciences, 138(1), 237–255, (2001)

[199] Yu, H., Jansen, R., Gerstein, M.: Developing a similarity measure in biological function space. Bioinformatics, (2007)

[200] Zachary, W.W.: An information flow model for conflict and fission in small groups, Journal of anthropological research, 452–473, (1977)

[201] Zadeh, L.A.: Similarity relations and fuzzy orderings. Information sciences,3(2), 177–200, (1971)

[202] Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In Advances in neural information processing systems, 1601–1608, (2004)

[203] Zhang, S., Wang, R. S.,Zhang, X. S: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications, 374(1), 483–490, (2007)

[204] Zhang, S., Wang, R. S.,Zhang, X. S: Uncovering fuzzy community structure in complex networks. Phys. Rev, 76(4), (2007)

[205] Zhang, P., Zhang, J., Sheng, H., Russo, J. J., Osborne, B., Buetow, K.: Gene functional similarity search tool (GFSST). BMC bioinformatics, 7(1), 135, (2006)