Engineering and network analysis

Ing. Pozzi, R. [1]
PhD candidate
LIUC – Cattaneo University
Castellanza, Italy

Prof. Noè, C.
Full professor and Director of the School of Engineering
LIUC – Cattaneo University
Castellanza, Italy

Dr. Strozzi, F.
Researcher
LIUC – Cattaneo University
Castellanza, Italy

Conference Topic: Active Learning

1. INTRODUCTION

The theory of graphs (or networks) has been applied to many different fields [1] and network analysis, the study of the characteristics of the set of nodes and arcs that constitute the graph, has been applied to as many different areas, such as networks of citations, in which research streams are represented as networks [2] and social networks, extensively documented by [3], used to investigate social relationships (involving communities, business organizations and social network).

Due to the recognised importance of network analysis, Industrial Engineering students of LIUC – Cattaneo University attending the Mathematical Methods for Industrial Applications class have dealt with such a topic. In particular, after having attended traditional classes students were asked to group to apply the network analysis techniques to a case of their own interest. For each of the areas investigated by the groups of students (linguistic, collaborations, social, man-made and flavours) previous works concerning them can be find in literature. In [4] it is presented a comprehensive analysis of the literature on the use of graph theory in the context of language. Among the cited works, the study of the structure of the English language done by [5]: with reference to the English Thesaurus dictionary, the authors consider interconnected words that express the same concept. The investigation of the clustering coefficient confirms the expectation of components closely related to each other within the glossary, clearly distinguishing semantic areas. [6] conducted a survey that gives similar results referring to WordNet, a lexical semantic network developed at Princeton by GA Miller, where nodes (words) are connected to each other on the basis of meaning [7]. In literature several studies concern networks of collaborations, meant as working relationships between movies characters. The search of [8], fundamental as it regards the definition of small-world properties (based on the statistical properties of clustering coefficient and path length), takes the graph of collaboration actors-movies as example in order to investigate such properties. Later, [9] take the same example to explain the scale-free networks, and in 2003, [10] defines collaboration networks as "classic examples of social networking”. The same research by [8] uses another example of application of the measure of small-world network man-made: the power distribution network of the USA. Later, [11] study the energy networks of Europe, England and Italy, investigating the degree distribution and the scale-free characteristic and [12] investigates the structural properties (degree distribution, clustering coefficient and preferential attachment) of the energy network of Orissa, a state federated of India. Literature was also in charge of the so-called "flavour networks". First [1] represent the recipes through a network of ingredients and directions for cooking

[1] Corresponding Author
Pozzi, R.

them. Later, [13] study the relationship between ingredients and recipes calculating the degree distribution, while [14] study the impact of aromatic components on ingredients combination.

The present paper first introduces the measures and tools learned and applied by the students. Then, the studied topics are presented and, for the networks that have been extensively analysed, the research questions addressed by the measurements of the theory of complex networks are outlined. Finally, the conclusions that students have drawn with the help of teachers and their usefulness in terms of management are summarized.

1.1. Measures

The measures applied in this work are in the following. They can be distinguished into local measures and global measures, which refer respectively to individual nodes and the network in its entirety.

Among the local measures: *degree centrality*, *betweenness centrality*, *closeness centrality* and *clustering coefficient*.

Degree centrality: it is equal to the number of neighbours of a node (or nodes to which it is connected by a single link), and it is the simplest indicator of centrality.

Betweenness centrality: it is calculated as the number of shortest paths that pass through the node under study in relation to the total number of shortest paths of the network. It indicates if a node can be "intermediary" within the network.

Closeness centrality: it is calculated as the ratio between the number of nodes that make up the network and the sum of the distances of the reference node from the other vertices.

Clustering coefficient: it is the ratio between the number of edges between its neighbour nodes and the maximum possible number of edges between them. It indicates how the neighbours of a node are connected directly together, thus if the node itself is irrelevant to keep the network connected.

Among the global measures: *degree centralization*, *betweenness centralization* and *closeness centralization*.

Degree centralization: it is a measure of the similarity between the structure of the network and a star network (a network in which a single node has a high number of connections with the other nodes of the network).

Betweenness centralization: it measures the relationship between the *betweenness centrality* variation among vertices divided by the *betweenness centrality* variation among vertices of a star network of the same size. It can indicate the vulnerability of the graph to selective attacks to its nodes (or arcs).

Closeness centralization: it measures the relationship between the *closeness centrality* variation among the nodes and the *closeness centrality* of a star network of the same size.

Clustering coefficient: it is equal to the average of the *clustering* of individual nodes.

1.2. Tools

To perform the analyses the students have used Pajek software (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). The networks have been first represented using the Kamada-Kawai algorithm [15], one of the alternative visualizations proposed by the software to represent complex networks. Referring to the force between the nodes of the network (between two nodes u and v is the number of edges of the shortest path between u and v) and the total energy of the network (due to the force between nodes), the algorithm tries to place each node in a position such as to minimize the total energy. In addition, the positions identified must respect two criteria: (i) minimize the number of intersections between arcs, (ii) to distribute the vertices and edges uniformly.

2. NETWORK ANALYSES

As the purpose of the assignment was increasing the involvement of students in network analysis, the groups of students had the possibility to focus their studies on objects of their own interest. As a consequence, notwithstanding some groups have studied the same topics, a great variety of themes have been involved. Social networks represent the most studied subject, highlighting students' interest in their functioning and in the information that can be obtained from them. Some groups focused on industrial and service topics, involving components failure transmission in a refinery, the characteristics of the Italian distribution network of a famous e-commerce company and Italian airlines networks. The road network represents another shared subject of interest: one group studied the ancient Rome roads, one studied a particular road of the Italian motorway network, while another focused on compressed natural gas (CNG) stations location along the Italian motorway network. Three other groups involved topics extensively studied by literature: glossary, movies and citations networks. Two groups focused on companies organizational aspects: one studied the flow of information internal to an Italian confectionary company, while another studied the organizational evolution of an Italian hospital wards. Two groups involved sports in their analysis:

one investigated the goals made by Italian soccer teams playing the Italian league during the 2012/2013 season, while another is based on the nationalities of all the baseball players in the American league during the 2011/2012 season. Other topics involved by groups are cooking recipes and comics.

2.1. Graph and Digraph glossary

The analysed network has been drawn from the "Graph and Digraph Glossary", produced by Bill Cherowitzo between 1998 and 2001. The glossary contains definitions of terms relating to the area of graphs and directed graphs (digraphs, in fact), most of which are connected to each other. For example, the term "forest" has the following definition: "A graph which contains no cycles. The connected components of a forest are trees", where the underlined words link to other definitions. The correspondent network is not connected (there are components, and even isolated vertices) and links to other definitions make the glossary a direct network (the network is represented in Fig. 1 by using Kamada-Kawai algorithm). Each node represents a glossary term, while any generic link XY (from X term to term Y) exists if Y is the term used to describe term X and/or vice versa, with the following specifications: (i) directed arcs: the definition of X is a reference to Y; (ii) edges: the reference is bidirectional. The research questions to which the analysis seeks for an answer are the following: (i) As part of the analysis of graphs, which terms appear to be "primitive", that can not be described with the use of others, and which are dependent on them? (ii) Which terms most need others to be described? (iii) Is the identification of the areas of concepts closely related possible? (iv) Is deleting "key terms" that isolate areas or conceptual and semantic areas possible? To respond question (i) and (ii) the measure of *degree centrality* is involved. As direct network, the *degree* of the nodes (e.g. node Y) is analysed distinguishing between the input (the term represented by the node Y is used to define another) and the output (the term represented by the node Y makes reference to another to be described). The terms with output degree equal to 0, i.e. those that do not need to be described by other terms, are numerous: sorting them according to their input degree, "graph" (used to define 16 other terms) and "walk" (used to define 7 other terms) result to be the most important. With reference to the terms that depend on others to be defined, the *output degree* calculation shows that "loop" refers to 5 terms, "and condensed bipartite graph" and "forest" terms to 4 others. With regard to the identification of areas of concepts that are closely related, the analysis of *connected components* (*communities*) identifies 9 sub-groups in the network, of which the main part is made of the 83% of the nodes (60/72). Thus, the glossary can be defined as a single predominant conceptual area, contrary to what happens for the Thesaurus and WordNet. *Betweenness* analysis has helped the identification of "key terms" that connect areas. The analysis confirms the result obtained through the communities: there are no "key terms" necessary to keep the network connected, proceeding to the elimination of the 4 most important words ("vertex", "graph", "edge" and "arc"), the network also remains connected.

2.2. Marvel heroes

The group of students has drawn and studied "The Social Network of Masked Vigilantes" (Fig. 2), which explores the set of meetings between heroes/characters created by the famous U.S. publisher Marvel Comics. The correspondent network is connected and it is not direct. Each node represents a hero/character of a comic book, and every generic arc exists if there has never been a meeting between two characters in the same book. The questions the analysis aims are the following: (i) Referring to Marvel heroes meetings, which hero is more involved in social relationships? (ii) Which heroes/characters could ask for the help of/help more heroes/characters? (iii) Is the identification of narrative line within the Marvel stories possible? (iv) Which heroes/characters are essential to narrative lines themselves and interweaving among them? To identify which heroes are more involved in social relationships, the number of links for each node is determined, referring to the measure of the degree centrality. The *average degree* of the network is calculated to be 4.8, while the node with the highest *degree* is the one that refers to the node "Captain America" linked to 25 other nodes, followed by "Thor" connected to 22 other nodes. Whit reference to the second question, the possibility of asking for help/help other characters/heroes, the group of students have involved the concept of *closeness centrality*. The above average and highest value is related to the node "Captain America", which, therefore, would be able to get to nodes and to be reached faster than any other. Through the community analysis, in particular using the Louvain Method [16], it is possible to identify six distinct narrative lines led by as many nodes representing "Captain America", "Spiderman", "Human Torch", "Beast", "Thor" and "Hulk". The indispensability of the above mentioned heroes to the success of narrative lines is confirmed by analysis of the *degree centrality*: the six heroes corresponding to the nodes with higher values are the narrative lines leaders. With reference to the last research question, the values of *betweeenness*, *betweenness centrality* and *centralization*, have helped the group of students in analysing the interweaving among the narrative lines of Marvel. In particular, the group has found that eliminating the leader heroes from the stories (i.e. removing nodes from the network) the six narrative lines isolate avoiding the possibility of get in contact with each other. Thus, the group has got benefit from the visualization of the network, which has stimulated some questions concerning market strategy for Marvel: why those nodes have become the most central ones (market policies, historical context, were the most loved by the public, etc.); what encouraged the authors to make those heroes intervene in other stories and how readers have welcomed those interweaves; have the links between different lines maintained over time or have not successful interweaves not recur?

2.3. CNG stations along the Italian motorway network

The group has analysed two networks representing the location of the CNG stations along the Italian motorway network and, in particular, the links that have their origin in Italian administrative centres and their destination in LIUC - Cattaneo University (LIUC), hereinafter network1, depicted by (Fig. 3), and reverse paths, hereinafter network2 (depicted by Fig. 4). Both networks are not *fully connected* and *have direction*. Stations are only accessible for cars moving along the direction of the represented motorway. Each node represents a CNG station, while an arc connects nodes that are less than 193 miles distant from each other (i.e. the distance one can travel by car without refuelling, based on the capacity of the tank mounted on CNG cars most sold in Italy: Fiat Punto and Fiat Panda), and the arc does not exists in case the distance exceeds the threshold value. The questions the analysis wants to answer are the following: (i) is the CNG stations along the Italian motorway network coverage sufficient to travel from/to Italian administrative centres to/from LIUC? (ii) Which stations are fundamental to the journey? (iii) During the strike of CNG suppliers, is the journey along the network1 more risky than network2? The first question is addressed by the study of base statistics that describe the two networks. The coverage level appears low: the *average degree* of the nodes is equal to 2 referring to network1 and equal to 3.5 referring to notwork2, much lower than the value that would assume if the two networks were fully connected, respectively, 29 and 32. With reference to *closeness centrality* measure, the difference in connection between administrative centres and distributors emerges. Focusing on the administrative centres that are part of the network (i.e. the ones that are less than 193 miles distant from a CNG station), the representative nodes have low values of *closeness centrality*, ranging between 0.07 and 0.13. The base statistics analysis clarifies that, despite the number of nodes is very similar between the two networks (30 and 33), network1 is less dense than network2, since the amount of arcs is about half. Moreover, the network coverage is not equivalent in the two directions. In conclusion, you can not make the journey from different administrative centres (e.g. "Bari", "Naples", "Roma") towards LIUC, while you can reach most administrative centres starting the journey from LIUC (including "Napoli" and "Roma”), but not all (such as "Bari"). The analysis of *betweenness centrality* has allowed the identification of the node "Parma" as the main station in keeping connected networks. Eliminating such a node from the networks four administrative centres of network1 and two of network2 are no more connected: if node “Parma” there wasn’t a smaller number of administrative centres could reach LIUC and vice versa. Extending the analysis of communities is necessary to answer the third research question, which network is more risky during the strike of CNG suppliers. In the case of network1, deleting the node with the highest *degree centrality* ("Parma") for four capitals disconnect from the network. In the case of network2, removing it is necessary to simultaneously delete three nodes ("Parma", "Somaglia"and" Assago") to disconnect four capitals. The analysis constitutes a base for a possible strategy to improve the CNG station network, pointing out the characteristics of the current network and its criticalities.

2.4. New Year’s Eve menu

The group has focused their analysis on two networks representing the New Year's Eve menu proposed by a restaurant located close to LIUC. In particular, the analysed menu consists of 7 recipes made by 53 ingredients. Both the networks are connected and *not direct*. In the first analysed network (hereinafter network3), each node represents a recipe, and two nodes are linked if sharing an ingredient. In the second analysed network (hereinafter network4), each node represents an ingredient, and two nodes are linked if belonging to the same recipe. The questions to which the group has responded are the following: (i) which ingredients (if missing) impede the realization of most of the recipes? (ii) One can characterize the variety of the menu? From the analysis of the *clustering coefficient* the lack of salt and pepper (the ingredients with the lowest values) in the kitchen would impede the realization of the menu, while the elimination of any of the recipes from the menu would not impede its realization (in this regard see also the representation of *betweenness* in Fig. 5 and Fig. 6). This result is confirmed by the *degree centralization* analysis: with reference to network3, the value 0.2 indicates that all the nodes are connected in a similar way to each other; with reference to network4, a value equal to 0.64 identifies the greatest similarity to a star network (whose centre node is represented by the nodes "salt" and "pepper"). Thus, if salt and pepper were missing the realization of the menu would be prevented. With reference to the menu variety, the *degree centralization* value of network 3 (0.2) indicates that the network is nearly complete (focusing on the subnet of salty dishes, it definitely is), highlighting strong similarity between the flavours of the recipes, thus lack of variety.

3. CONCLUSION

The present work has discussed the application by engineering students of network analysis to topics of their own interest. In particular, the discussion focuses on four analyses, the ones that have looked into the network in depth, and on the particular and general conclusions that the students, assisted by teachers, have reached. Regarding general conclusions, it is possible to divide them with reference to networks characteristics: “Graph and Digraph Glossary” and “Compressed natural gas stations along the Italian motorway network” are directed networks; “Marvel heroes” and “New Year’s Eve menu” are *undirected* networks. With particular reference to direct networks, students were able to understand that the same measure covers different meanings in relation to the network to which it is applied. While the *input* and *output degrees* of the nodes in the case "Graph and Digraph Glossary" have opposite meanings (the *input/output degree* identifies being necessary to/require the definition of a word), on the contrary, in the case "Compressed natural gas stations along the Italian motorway network" there is no difference in meaning between *input* and *output degrees*. With reference to the *degree* of the nodes of the undirected networks, the *degree* itself has different meanings. Students have observed that in the case of "Marvel Heroes" network it identifies the popularity of the character (the node "Captain America" is the node with the highest *degree* and other measures obtained by the network analysis have highlighted its indispensability), while in the case of "New Year's Eve Menu" network, ingredients with the highest *degree* poorly characterize the recipe (nodes "salt" and "pepper" have the highest *degree*, appearing in almost all of the recipes, so do not give meaningful information about the recipes themselves). In addition, it is worth highlighting a further remark on network analysis measures: some become meaningless when applied to certain networks. This assertion is confirmed in the case of *k-neighbor* measure, which is rich in meaning when applied to social networks (i.e. "Marvel Heroes") or, among the others, to “Compressed natural gas stations along the Italian motorway network” network. However, this measure has no meaning when applied to "New Year's Eve menu" network, as knowing the distance that separates two ingredients gives no value to the analysis. Particular conclusions identify managerial implications that emerged from the individual network analysis. In case of "Marvel Heroes", displaying the network stimulates the above mentioned questions, helping the development of a strategy. With reference to the distribution of CNG, the analysis provides guidance to improve the service. Finally, with reference to "New Year's Eve Menu" network, the recipe variety is clearly understandable, giving indication to the chef for any changes to the recipes.

Moreover, the possibility of displaying objects of interest as a network has enabled students reasoning on them from a different perspective, understanding the value of network analysis. Thus, the possibility of conducting their study on objects of their own interest has contributed in stimulating students’ interest in reasoning on the meaning of the obtained results, rather than just collecting them, providing curious conclusions.

FIGURES

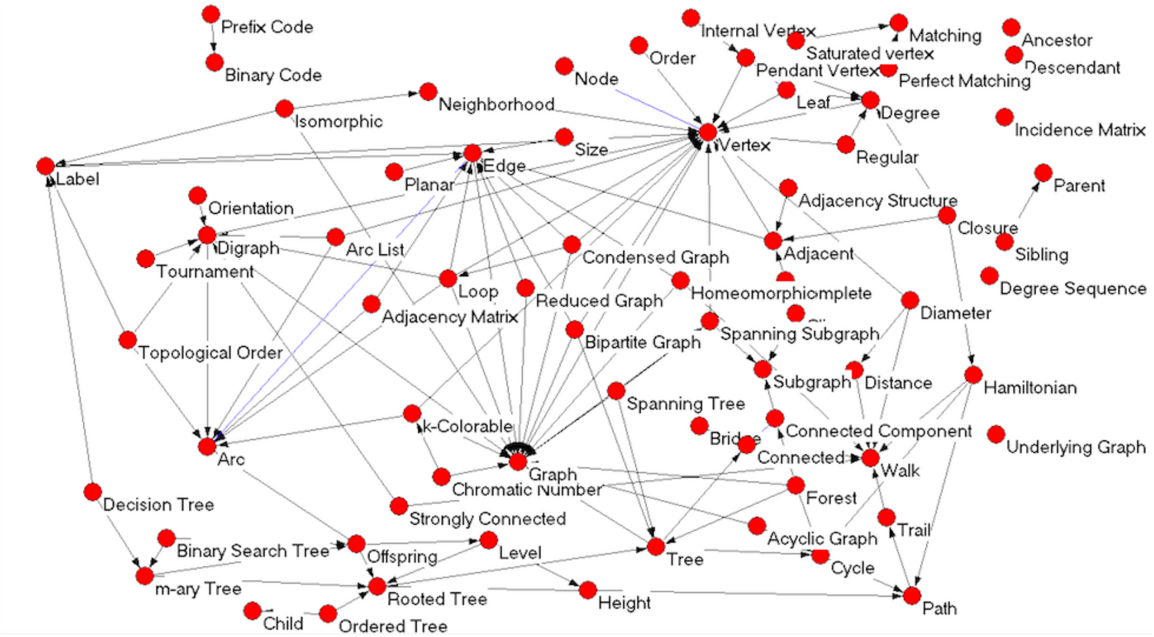

Fig. 1 Graph and Digraph Glossary network

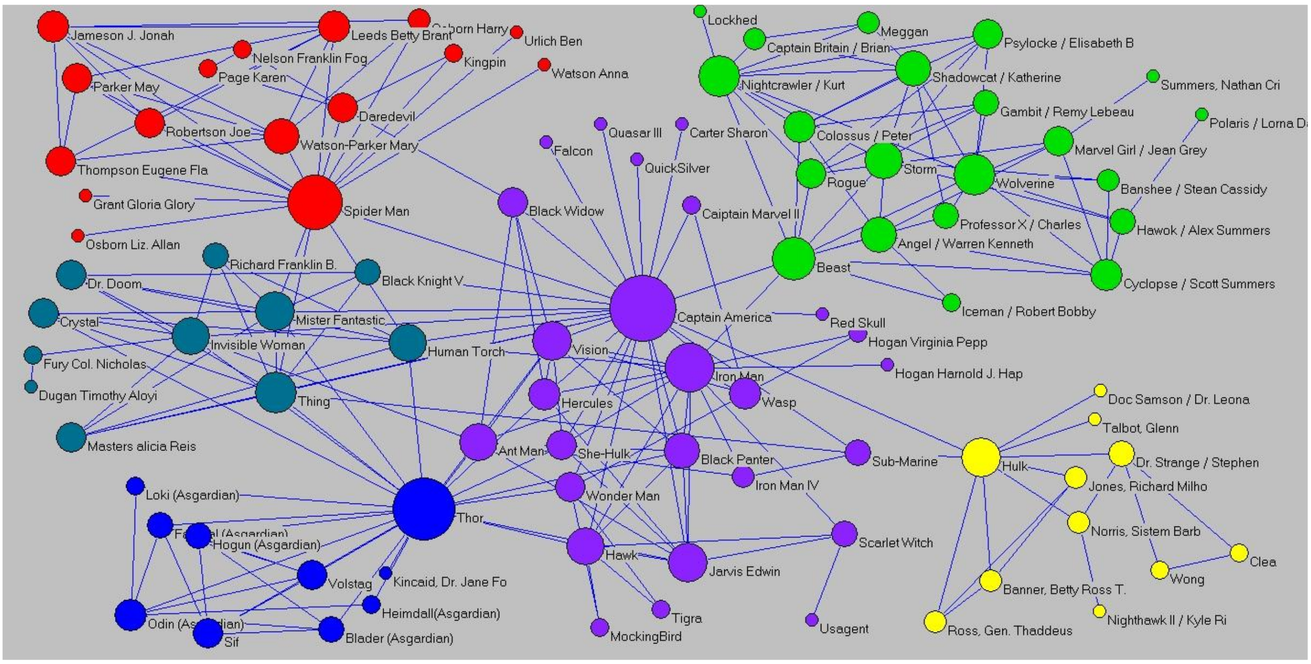

Fig. 2 Marvel heroes network

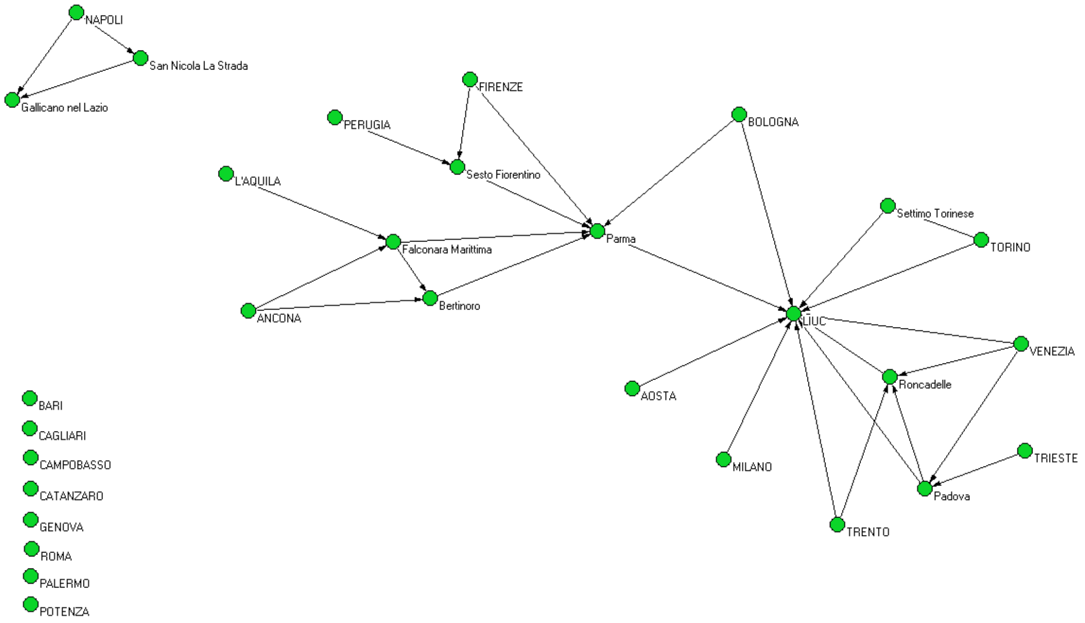

Fig. 3 CNS stations network1

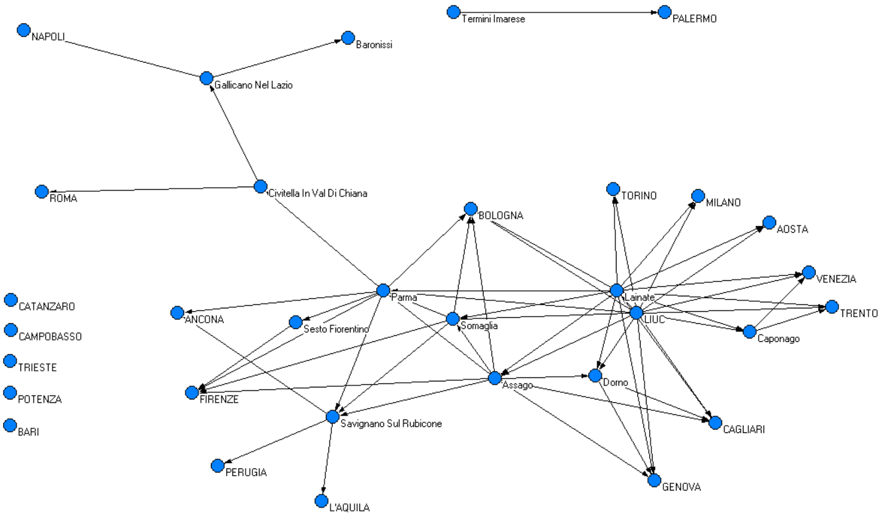

Fig. 4 CNS stations network2

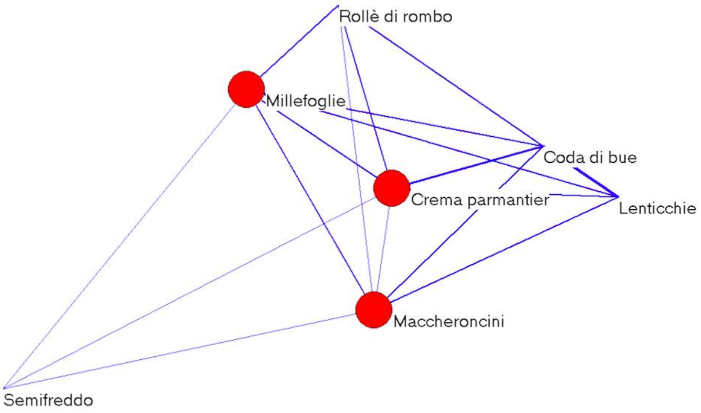

Fig. 5 Network3 *betweenness*

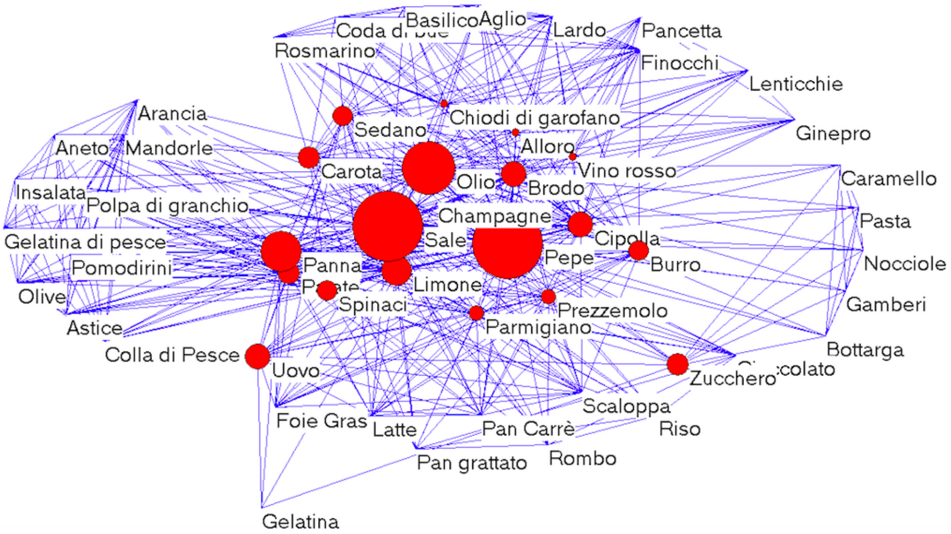

Fig. 6 Network4 *betweenness*

REFERENCES

[1] Wang, L., Li, Q., Li, N., Dong, G. and Yang Y. (2008), Substructure similarity measurement in Chinese recipes. Proceeding of the 17th International Conference on World Wide Web ACM, 2008. pp. 979-988.

[2] Strozzi, F. and Colicchia C. (2012), Literature review on complex network methods applied to measure robustness in supply chain design, Liuc Papers.

[3] Scott J. (2010), Social network analysis: developments, advances, and prospects., Social Network Analysis and Mining, Vol. 1, No. 1, pp. 21–26.

[4] Altieri, N., Gruenenfelder, T. and Pisoni D. B. (2010), Clustering coefficients of lexical neighborhoods: Does neighborhood structure matter in spoken word recognition?, The mental lexicon, Vol. 5, No. 1, pp. 1–18.

[5] Motter, A. E., Moura, A. P. S. De, Lai, Y. and Dasgupta P. (2008), Topology of the conceptual network of language, Physical Review E, 65, 065102.

[6] Steyvers, M. and Tenenbaum, J. B. (2005), The large-scale structure of semantic networks: statistical analyses and a model of semantic growth., Cognitive Science, Vol. 29, No. 1, pp. 41–78.

[7] Fellbaum, C., (1999), WordNet, an electronic lexical database. Cambridge, MA.

[8] Watts, D. J. and Strogatz S. H. (1998), Collective dynamics of “small-world” networks. Nature, Vol. 393, No. 6684, pp. 440–2.

[9] Barabasi A. and Albert R. (1999), Emergence of Scaling in Random Networks, Science, Vol. 286, No. 5439, pp. 509-512.

[10] Newman M. E. J. (2003), The structure and function of complex networks, SIAM review, Vol. 45, No. 2, pp. 167-256.

[11] Rosas-Casals, M., Valverde, S. and Solé R. V. (2007), Topological Vulnerability of the European Power Grid Under Errors and Attacks, International Journal of Bifurcation and Chaos, Vol. 17, No. 07, pp. 2465–2475.

[12] Roy D. S. (2013), The Topological Structure of the Odisha Power Grid: A Complex Network Analysis, IJMCA, Vol. 1, No. 1, pp. 12–16.

[13] Kinouchi, O., Diez-Garcia, R., Holanda, A., Zambianchi, P. and Roque A. (2008),The non-equilibrium nature of culinary evolution. New Journal of Physics, Vol. 10, No. 7.

[14] Ahn, Y.Y., Ahnert, S. E., Bagrow, J. P. and Barabási A. L. (2011), Flavor network and the principles of food pairing, Scientific Reports, Vol. 1.

[15] Kamada, T. and Kawai S. (1989), An algorithm for drawing general undirected graphs, Information processing letters, Vol. 31, No. 1, pp. 7-15.

[16] Blondel, V. D., Guillaume, J. L., Lambiotte, R. and Lefebvre E. (2008), Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, Vol. 10, P10008.