SAPIENZA - UNIVERSITÀ DI ROMA

ANNALI DEL DIPARTIMENTO DI METODI E MODELLI PER L'ECONOMIA, IL TERRITORIO E LA FINANZA

2016

Perspectives on Behavioural Sciences

ISBN: 978-88-555-3361-4 ISSN: 2385-0825

PÀTRON EDITORE
Bologna 2016

PÀTRON Editore - Via Badini, 12
Quarto Inferiore, 40057 Granarolo dell'Emilia (BO)
Tel. 051.767003
Fax 051.768252
E-mail: info@patroneditore.com
http//www.patroneditore.com
Il catalogo generale è visibile nel sito web. Sono possibili ricerche per autore, titolo, materia e collana. Per ogni volume è presente il sommario, per le novità la copertina dell'opera e una breve descrizione del contenuto.

Stampa: Rabbi s.r.l., Bologna per conto di Pàtron editore, dicembre 2016.

Carlo Drago*

EXPLORING THE COMMUNITY STRUCTURE OF COMPLEX NETWORKS

Abstract: Regarding complex networks, one of the most relevant problems is to understand and to explore community structure. It is important to define in particular the network organization and the functions associated with the different network partitions. In this paper, we consider an approach based on interval data in order to represent the different relevant network components as communities. The method is also useful to represent the network community structure, especially the network hierarchical structure. We apply the methodology on the Italian interlocking directorship network.

Keywords: complex networks, community detection, communities, interval data, interlocking directorates.

1. Introduction

The study of complex networks is very important today because the comprehension of modern systems can be considerably enhanced by considering the different connections between individuals or objects.

The communities are groups of nodes in the networks, maximally connected to each other and weakly connected between the different communities. The communities represent a very important network feature (Fortunato, 2010; Newman, 2006): in fact the community structure allows to understand the concrete

* University of Rome "Niccolò Cusano", Rome, Italy.

functioning of real world systems. In this sense many phenomena can be clarified by taking into account the structural information related to the community. The community structure is typically associated with the activities performed by a single subgroup of nodes (Porter, Onnela, Mucha, 2009). Thus the main question here is: in what way is it possible to measure the different network characteristics and statistically exploit the different relationships between the communities?

The aim of this work is to measure community characteristics and analyse the relationships between different communities. So the analysis is based not only on the nodes of the communities, but also on the communities as different entities. In particular a methodological way to do this is to consider a different approach from that using classical data.

2. The statistical problem

Complex networks are typically characterized by deviations from the random graph (the most simple structure studied among graphs), by unobvious topological characteristics (Newman, 2003), community structure or modularity (Newman, 2006) and finally hierarchical structure (Barabasi et al., 2003). These structures are ubiquitous and this fact represents an important reason that has recently motivated the scientific community to study the mechanisms which can have a relevant impact on the complex network and in particular on their topology (Albert, Barabasi, 2002). In this respect, it is important to take into account the modularity of the complex networks (Newman, 2006). In fact, complex networks are characterized also by multiple communities. It is very relevant in network analysis to identify these communities so allowing one to understand relevant functions inside the networks (Fortunato, 2010). In order to characterize the role of the different communities it is possible to consider the topological features and the characteristics of the attributes for the nodes which are part of the communities.

A fundamental aim in network analysis is to characterize the community structure. Here, following Nishikawa, Motter (2011), it is important, in the context of analysis of the network, to introduce the concept of structural group of nodes. The structural group of nodes is obtained by considering a specific community detection method. In this way we are able to identify group of nodes which are part of a same community. Then an analysis is carried out to analyse the topological features of the groups obtained in the network. This analysis is particularly relevant because the different group of nodes belonging to the different communities which are part of the network are related to some different function of the network. So representing the characteristics of the communities means being able to understand the role or the function of the communities inside a network. Specifically, this means assigning a measure to an entire community considered as an entire group of nodes. In this way each characteristic of the nodes can be read as a characteristic of the single node but considering all the nodes we can consider a way to take into account the entire community.

However a specific problem exists: representing the different communities with a specific value related to the different community characteristics can lead to relevant information loss. In this case a representation which preserves this information is required. In this context the different communities are represented by intervals or symbolic data in which we consider both the lower and the upper bound for the values considered in the same community. In this framework the topological features and the attributes of nodes in the same community are represented as intervals, that is, interval data can be considered a way to measure quantitatively the network structures. In this way, a possible solution is to consider intervals of values to represent the different communities and thus be able to represent the network. The characterization of the communities on the network can be very useful in allowing to predict the future behaviour of the network as a whole.

3. Network representations

There has been an increasing interest in analytical techniques which consider complex data, and in particular on data characterized by specific internal variations. Particularly relevant contributions in the field have been those by Billard, Diday (2003) and Diday, Noirhomme-Fraiture (2008). Interval data (Billard, 2008), in particular, can be very useful in order to measure the different structural characteristics of the communities as a whole. In fact, communities are typically characterized by heterogeneous node characteristics and this heterogeneity can be usefully represented by interval data. These data types have their mathematical properties and appropriate statistical methods (Gioia, Lauro, 2005; Brito, Duarte, 2012; Lauro; Palumbo, 2000 among others). In network analysis this approach was also considered by Giordano, Brito (2012) who measured and compared entire network using histogram data.

In our work, we will consider the network communities with the aim to characterize the community structure as a whole and analyse the different roles of the single nodes in the community. Interval data has been chosen as the approach because it allows the measurement of the characteristics of the communities considered. At the same time the use of the interval data has been preferred because the number of the nodes of the communities are not so large and so an interval seems a reasonable way to represent the characteristics of the group considered. Now we will consider both the network and the different ways to represent the network and the community structure. We start with an undirected graph: each node can be differently characterized by considering for example their centrality features.

An important characteristic of the networks is the modularity which measures the degree of the possibility of dividing the network into different modules. The modularity can be used to identify the different communities of the networks (Newman, 2006; Reichardt, Bornholdt, 2006). In this sense the communities are relevant characteristics of the networks and allow us to understand the organization of the network as a whole (Porter, Onnela, Mucha, 2009).

But there is not a unique specific definition of community. We can consider the definition in Fortunato (2010). The different communities are part of the community structure (Newman, Girvan, 2004). Various ways to detect communities are proposed in literature (Fortunato, Lancichinetti, 2009; Lancichenetti, Fortunato, 2012; Leskovec et al., 2010; Drago, Balzanella, 2014). The aim is to represent the community structure in an adequate manner in order to discover the latent information which is possible to observe. In particular, it is very important to understand from the data the role of the different communities on the whole network. In order to perform this task we can consider interval data based on the single communities. In this way the communities are represented by also considering interval data on the member features of the communities.

4. The community structure representation

Each network can be studied by observing their communities and a different community can be characterized by the topological features related to the different node members of the community. In particular we can consider the different members of the communities and the different characteristics or features of the communities. Interval data can be used to represent the different communities. In fact when using the classical data we are not able to represent adequately the variations of the features. In this respect by using a mean we are losing information. In order to consider the communities as a whole we can consider a related interval data for each individual community. In that way the data matrix related to each single community can be represented by the lower and upper bounds of the communities. Then we can consider the obtained intervals in the visualization process. So we have:

$$I[Y_k^c] = [\underline{Y}_k^c, \overline{Y}_k^c] \tag{1}$$

where *k* is the community considered and *c* is the feature considered (for example the Freeman degree, or the betweenness). We can consider the single interval as the way to measure a single characteristic for the entire community. Intervals can provide important information about the network structure (for example the summary statistics of the intervals which can be obtained) and we are able to obtain a data table to visualize the interval data. The interval data for each community represents the upper bound and the lower bound characterizing each community inside the network. The different intervals can show not easily detectable structures. In order to represent the network communities, we first of all detect the different communities by using the fast greedy method (Clauset et al., 2004). Then we represent each community by using an interval data. At this point we are able to visualize the data by using an interval scatterplot (Bock, Diday, 2000) in order to discover the community network structure.

5. An application on real data

We will now consider a dataset related to interlocking directorships in Italy (year 2012). The network community structure allows us to determine the number of relevant communities in a network and the different nodes which are part of the different communities. Thereby we can obtain the different communities and then we can represent them as communities. At this point we decide to represent them as interval scatter plot. The different community structure is represented on the interval scatter plot diagram, especially where we expect to find a particular data structure for the network.

It is important to consider the shape of the different observations on the scatter plot. If there is higher centralization we can expect a particular network structure. In terms of the Freeman degree and the betweenness we can have a higher value for some statistical units. If there is a higher difference on the different observations it means that the network as a community structure becomes more centralized. So by considering the different structure of the intervals we are able to identify the centralization level of the community structure. On observing the network we can observe the general network structure which tends to be characterized by a central structure. This structure seems to be coherent with previous results in literature (Piccardi et al., 2010; Bellenzier, Grassi, 2014; Drago et al., 2014, 2015). In particular by considering the levels of betweenness and Freeman degree the results seem to be coherent with previous results for each node.

In order to consider community detection we also use the fast greedy algorithm which performs well with large networks. In this case we observe that there are at least 17 communities (Table 1). These communities share different characteristics on betweenness and Freeman degree. In particular we can observe that there are three communities which are the most relevant. The first one (community 4) is characterized by the highest values of betweenness and Freeman degree. Community 5 shows a higher level than community 6 of betwenness but lower levels on the Freeman degree. This means that community 5 tends to be more globally than locally centered. At the same time community 6 tends to be characterized by a more local than global centrality. Both the communities have lower values of Freeman degree and betweenness than community 4. The community with a higher level of betweenness and Freeman degree represents the most centralized companies in Italian capitalism, this has also been observed in other studies.

The most relevant result is that the structure of the Italian interlocking directorship network seems to be clearer by observing the interval scatterplot diagram. In particular we are able to detect three relevant communities (4, 5 and 6) whilst at the same time we are able to detect the highest community (Table 1). The communities 5 and 6 show higher levels than other communities, which in general reveals an equilibrium on the levels of the centrality measures.

Tab. 1 - First twelve communities by betweenness and degree interval values.

Community	Min. Betw	Max. Betw	Min. Deg	Max. Deg
1	0	1175.873	1	13
2	0	889.7801	1	6
3	0	1099.6	1	10
4	0	2076.065	1	20
5	0	1684.352	1	15
6	0	1590.881	1	19
7	0	243.9954	2	9
8	0	618.0377	1	5
9	0	345.6612	1	6
10	0	994.6088	1	12
11	0	622.6268	1	7
12	0	485	1	3

6. Conclusions

In this work we have proposed a new approach for analysing complex networks. This approach consists in analysing the network by decomposing it into different communities. The different communities are characterized by considering interval data in order to allow the variation existing between the different nodes. In this way we are able to detect the structure of the network.

The results obtained are encouraging. In fact the network seems to be well represented by the intervals while using a mean seems to reduce the information extracted from the networks and from the different communities. The final conclusion of the application is that this method allows us to identify the correct structure of the network. For example, by considering the network of Italian interlocking directorships, we are able to identify the structure of the communities and in particular the different roles of the different communities.

References

ALBERT R., BARABASI A.L. (2002), Statistical mechanics of complex networks, *Reviews of modern physics* **74**, 1, pp. 47-94.

BARABASI A.L., DEZSO Z., RAVASZ E., YOOK S.H., OLTVAI Z. (2003), Scale-Free and Hierarchical Structures in Complex Networks, in *Modelling of Complex Systems: Seventh Granada Lectures*, Vol. 661, No. 1, AIP Publishing, pp. 1-16.

BELLENZIER L., GRASSI R. (2014), Interlocking directorates in Italy: persistent links in network dynamics, *Journal of Economic Interaction and Coordination* **9**, 2, pp. 183-202.

BILLARD L. (2008), Some analysis of interval data, *CIT - Journal of Computing and Information Technology* **16**, 4, pp. 225-233.

BILLARD L., DIDAY E. (2003), From the statistics of data to the statistics of knowledge: symbolic data analysis, *Journal of the American Statistical Association* **98**, 462, pp. 470-487.

BOCK H.H., DIDAY E. (Eds.) (2000), *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*, Springer Science & Business Media.

BRITO P., DUARTE SILVA A.P. (2012), Modelling interval data with Normal and Skew-Normal distributions, *Journal of Applied Statistics* **39**, 1, pp. 3-20.

CLAUSET A., NEWMAN M.E., MOORE C. (2004), Finding community structure in very large networks, *Physical review E* **70**, 6, 066111.

CSARDI G., NEPUSZ T. (2006), The igraph software package for complex network research, *Inter. Journal of Complex Systems* **1695** (http://igraph.org).

DIDAY E., NOIRHOMME-FRAITURE M. (Eds.) (2008), *Symbolic data analysis and the SODAS software*, J. Wiley & Sons, pp. 1-457.

DRAGO C., BALZANELLA A. (2015), Non-metric MDS Consensus Community Detection, in MORLINI I., MINERVA T., VICHI M. (Eds.), *Advances in Statistical Models for Data Analysis*, 08/2015, Springer, pp. 97-105.

DRAGO C., MILLO F., RICCIUTI R., SANTELLA P. (2014), Corporate Governance Reforms, Interlocking Directorship and Company Performance in Italy, *International Review of Law and Economics* **41**, March 2015, pp. 38-49.

DRAGO C., RICCIUTI R., SANTELLA P. (2015), An Attempt to Disperse the Italian Interlocking Directorship Network: Analyzing the Effects of the 2011 Reform, *Working Paper*, Available at SSRN: http://ssrn.com/abstract=2580700.

ERDOS P., RENYI A. (1959), On random graphs, *Publicationes Mathematicae* **6**, pp. 290-297.

FORTUNATO S. (2010), Community detection in graphs, *Physics Reports* **486**, 3, pp. 75-174.

GIOIA F., LAURO C.N. (2005), Basic statistical methods for interval data, *Statistica applicata* **17**, 1-29.

GIORDANO G., BRITO P. (2014), Social Networks as Symbolic Data, in VICARI D., OKADA A., RAGOZINI G., WEIHS C. (Eds.), *Analysis and Modeling of Complex Data in Behavioral and Social Science*, Springer: Heidelberg, pp. 133-142.

LANCICHINETTI A., FORTUNATO S. (2009), Community detection algorithms: a comparative analysis, *Physical review E* **80**, 5, 056117.

LANCICHINETTI A., FORTUNATO S. (2012), Consensus clustering in complex networks, *Scientific reports* **2**.

LAURO C.N., PALUMBO F. (2000), Principal component analysis of interval data: a symbolic data analysis approach, *Computational Statistics* **15**, 1, pp. 73-87.

LESKOVEC J., KLEINBERG J., FALOUTSOS C. (2005), Graphs over time: densification laws, shrinking diameters and possible explanations, *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177-187.

LESKOVEC J., LANG K.J., MAHONEY M. (2010), Empirical comparison of algorithms for network community detection, *Proceedings of the 19th international conference on World wide web*, ACM, pp. 631-640.

NEWMAN M.E. (2003), The structure and function of complex networks, *SIAM review* **45**, 2, pp. 167-256.

NEWMAN M.E. (2006), Modularity and community structure in networks, *Proceedings of the National Academy of Sciences* **103**, 23, pp. 8577-8582.

NEWMAN M.E., GIRVAN M. (2004), Finding and evaluating community structure in networks, *Physical review E* **69**, 2, 026113.

NISHIKAWA T., MOTTER A.E. (2011), Discovering network structure beyond communities, *Scientific Reports* **1**, 151.

OLDEMAR RODRIGUEZ R. (with contributions from Calderon O. and Zuniga R.) (2014), *RSDA: R to Symbolic Data Analysis*, R package version 1.2, http://CRAN.R-project.org/package=RSDA.

PICCARDI C., CALATRONI L., BERTONI F. (2010), Communities in Italian corporate networks, *Physica A: Statistical Mechanics and its Applications* **389**, 22, pp. 5247-5258.
PORTER M.A., ONNELA J.P., MUCHA P.J. (2009), Communities in networks, *Notices of the AMS* **56**, 9, pp. 1082-1097.
REICHARDT J., BORNHOLDT S. (2006), Statistical mechanics of community detection, *Physical Review E* **74**, 1, 016110.

Appendix: Companies belonging the first three communities by betweenness and Freeman degree interval values

Community 4: A2A SPA, ADF SPA, ARNOLDO MONDADORI EDITORE SPA, ATLANTIA SPA, AUTOGRILL SPA, BANCA POPOLARE DI MILANO SCRL, CAMFIN SPA, CLASS EDITORI SPA, COBRA AUTOMOTIVE TECHNOLOGIES SPA, COMPAGNIA IMMOBILIARE AZIONARIA - CIA SPA, DADA SPA, EL.EN. SPA, ENGINEERING - INGEGNERIA INFORMATICA - SPA, ERGYCAPITAL SPA, FIERA MILANO SPA, GEMINA SPA - GENERALE MOBILIARE INTERESSENZE AZIONARIE, INTEK GROUP SPA, INTESA SANPAOLO SPA, MAIRE TECNIMONT SPA, MEDIASET SPA, MEDIOBANCA SPA, MEDIOLANUM SPA, MOLECULAR MEDICINE SPA, PIRELLI & C. SPA, PRELIOS SPA, SALVATORE FERRAGAMO SPA, SNAM SPA, TELECOM ITALIA SPA, VITTORIA ASSICURAZIONI SPA.

Community 5: ASTALDI SPA, BANCA PROFILO SPA, CEMBRE SPA, CIR SPA - COMPAGNIE INDUSTRIALI RIUNITE, COFIDE - GRUPPO DE BENEDETTI SPA, GEOX SPA, GRUPPO EDITORIALE L'ESPRESSO SPA, IMMSI SPA, M&C SPA, MEDIACONTECH SPA, PIAGGIO & C. SPA, PREMUDA SPA, SOGEFI SPA, TREVI - FINANZIARIA INDUSTRIALE SPA.

Community 6: BE THINK, SOLVE, EXECUTE SPA, BOLZONI SPA, CREDITO EMILIANO SPA, DATALOGIC SPA DAVIDE CAMPARI - MILANO SPA, DELCLIMA SPA, DE LONGHI SPA, ENEL SPA, GAS PLUS SPA, INTERPUMP GROUP SPA, IREN SPA, ITALCEMENTI SPA FABBRICHE RIUNITE CEMENTO, ITALMOBILIARE SPA, MITTEL SPA, NOEMALIFE SPA, PRIMA INDUSTRIE SPA, PRYSMIAN SPA, RCS MEDIAGROUP SPA, SOCIETA' CATTOLICA DI ASSICURAZIONE SOCIETA' COOPERATIVA, SORIN SPA, TAMBURI INVESTMENT PARTNERS SPAUNIONE DI BANCHE ITALIANE SCPA, ZIGNAGO VETRO SPA.

Acknowledgment: We would like to thank Alessandra Reale and an anonymous referee for their helpful comments and suggestions.