Nicotri et al. *EPJ Data Science* (2015) 4:8
DOI 10.1140/epjds/s13688-015-0047-z

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE
Open Access

Complex networks and public funding: the case of the 2007-2013 Italian program

Stefano Nicotri[1*], Eufemia Tinelli[2,3], Nicola Amoroso[1,2], Elena Garuccio[4] and Roberto Bellotti[1,2]

[*]Correspondence:
stefano.nicotri@ba.infn.it
[1]Istituto Nazionale di Fisica Nucleare - Sezione di Bari, via Orabona 4, Bari, I-70125, Italy
Full list of author information is available at the end of the article

Abstract

In this paper we apply techniques of complex network analysis to data sources representing public funding programs and discuss the importance of the considered indicators for program evaluation. Starting from the Open Data repository of the 2007-2013 Italian Program *Programma Operativo Nazionale 'Ricerca e Competitività'* (PON R&C), we build a set of data models and perform network analysis over them. We discuss the obtained experimental results outlining interesting new perspectives that emerge from the application of the proposed methods to the socio-economical evaluation of funded programs.

Keywords: public funding; open data; complex networks; program evaluation

1 Introduction

Since the last years of the past century, the importance of basing policies on evidence, data, and analysis has quickly spread all over the world. The Evidence-Based Policy movement [1–5] has grown enormously, and mainly all public administrations are now focused on maximising utility and show a pragmatic problem-solving approach to socio-economical issues [6]. In this respect, the evaluation of public funding programs is a field of great interest for policymakers and economists. Politicians and technicians need to estimate the impact that funding has on life and society, in order to address future programs and to modify their decisions. Many standard and advanced statistical methods are commonly used for this purpose, such as linear/nonlinear regressions, Bayesian inference, machine learning, data mining, and so on. In this paper we suggest new indicators, coming from network analysis, that can help underlining in a quantitative way important effects that are not usually considered, being them outside the domain of investigation of standard statistical tools. This does certainly not mean that program evaluation cannot be performed without including network analysis, but that valuable insight about public funding programs could hopefully be inferred from such techniques, in order to help increasing objectivity of the extracted results. Recently, a growing interest towards complex network analysis applied to evaluation can be seen both in literature [7–12] and institutional reports [13]. The indicators we suggest can be used by experts in program evaluation for their analyses, giving them the opportunity of considering and quantitatively measuring important features of the funding programs, such as relations between the actors involved in them. Social network analysis is a particularly suitable tool to extract information about relations among the different components of a system. Investigating the relations between the actors participating to a program could be of interest, since can *e.g.* show structural contradictions in the organisation of the different levels involved [10]. Considering the set of projects, research institutions and enterprises that participate to a funding program as a complex dynamical system, it is possible to identify underlying network structures simply defining the edges according to some relations among the components that are of interest for the evaluator. Once the network is constructed, global and local properties can be evaluated and discussed.

From a data collection perspective, the proposed analysis can profit from current emerging technologies and precise guidelines of European governmental institutions to support initiatives such as *Smart Cities & Communities*[a] and their co-related action goals (*Urban and Citizen App*, *e-Government*, *e-Democracy* and so on). All this initiatives have produced a large number of freely available datasets containing information, collected by national governments, which third parties are encouraged to use for their scope, analyse and republish as they wish, without restrictions from any copyright. Recently, Open Government Data (OGD) is emerging as a major movement in knowledge sharing. It promotes transparency and accountability, enables collaboration among stakeholders, encourages novel socio-economic activities and growing of the so-called network economy. Starting from the idea that without sharing information it is not possible to establish a culture of collaboration and participation among the relevant stakeholders, the Linked Open Data (LOD) [14]. Movement, which provides existing data in a machine-readable format, has gained large importance over the last years. From a such perspective, LOD facilitates innovation and knowledge creation from interlinked data, but it also introduces a level of complexity for information management and integration. Considering a good trade-off between data expressiveness and computational cost for data analysis, we have selected only Open Data repository without linked data and RDF[b] triples. Despite the main aim of such movement of reaching the largest possible portion of users, our investigation has outlined that such datasets are usually of heterogeneous quality and size, and that their analysis requires efforts in a pre-processing phase composed of typical ETL (Extract-Transform-Load) [15] and data cleaning procedures. It is worth mentioning that problems are commonly encountered while using network analysis for evaluation, like the concern about anonymity of non-aggregated data (and eventual anonymisation), or the fact that making results public usually interferes with the structure of the network itself [9]. These kind of problems are mitigated by using Open Data, since they are public 'by construction'.

The paper is organised as follows: in Section 2 we introduce the steps composing the schema of the overall analysis process. In Section 3 we describe the structure of the open data repository of the 2007-2013 Italian Program *Programma Operativo Nazionale 'Ricerca e Competitività'* (PON R&C), in order to keep the paper self-contained, and introduce the data model for network analysis; in Section 4 we present features and properties of the analysed network. In order to better discuss the experimental results, we distinguish among local properties, global properties and community structure. Section 5 close the paper.

2 Methodology

The approach followed in this paper consists of two main steps shown in Figure 1:

- processing of data sources (grey blocks),

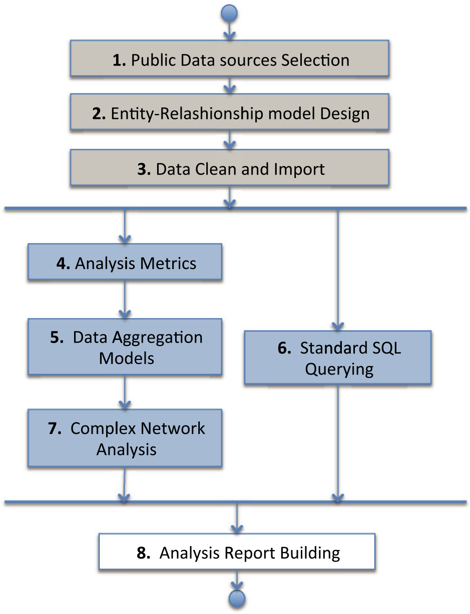

1. Public Data sources Selection
2. Entity-Relashionship model Design
3. Data Clean and Import
4. Analysis Metrics
5. Data Aggregation Models
6. Standard SQL Querying
7. Complex Network Analysis
8. Analysis Report Building

Figure 1 Data-flow schema of the overall analysis process.

- finding analysis models and metrics (blue blocks).

The first blocks of Figure 1 involve the transformation of a general purpose dataset to an analysis-specific one through the implementation of a data model. The newly obtained dataset is then used for network analysis. The graph is constructed identifying the nodes and the properties that define if two nodes are linked or not (in our case: being partner within the same project). Global and local properties are extracted, in order to produce a qualitative and quantitative description of the structure of the network of relationships generated by the program under examination. As represented in Figure 1, the overall process of our studies ends when a report summarising the analysis and its outcomes is produced. Descriptions of the numerical outcomes in terms of social/economical effects are given, in order to provide the evaluator with a useful tool for her/his purposes. We report on the SQL-based [16] modelling approach, which allows to translate a given dataset in Open Data format into the reference set of analysis models (relational tables), and on the selected metrics relevant for executing an effective network analysis. It is worth underlining that we have adopted well-known relational tables to store data in order to ensure integrity of our knowledge base and provide significant results. Our current implementation of data models exploits the open source object-oriented PostgreSQL 9.3 DBMS. For network analysis we have used the Wolfram Mathematica software and the R programming language.

3 PON R&C: from datasets to data models

In this section, the data-driven steps shown in blocks 2-6 of Figure 1 are described. As mentioned above, we have selected Open Data about the PON R&C funding program, publicly available at URL http://www.ponrec.it/open-data/. The program, funded with European Structural Funds managed by *Ministero dell'Istruzione, Università e Ricerca* (MIUR, Ministry of Instruction, University and Research) and *Ministero dello Sviluppo*

Economico (MiSE, Ministry of Economic Development), involved four underprivileged regions in Southern Italy: Apulia, Campania, Calabria and Sicily. The main aim of the program consists in promoting socio-economic growth by supporting research and innovation activities, improving quality of life for citizens and competitiveness of small-medium enterprise (SME). The main features of PON R&C can be summarised as follows: 2,962 funded projects, for over 3 billion euro, 11 action programs and 8 action areas: *Health-care*, *Nutrition*, *Energy*, *Environment & Ecology*, *Transportation & Logistics*, *Cultural Heritage & Activities*, *Smart Cities*, *Social Innovation*. The group of all the partners involved in each funded project is called *Temporary Scope Association* (TSA) and plays a fundamental role for our network analysis. The downloaded repository has 3 LOD stars[c] [17], is updated at '2014-06-17', and is composed of 3 datasets (files):

- *Projects* - 10,104 tuples with 52 attributes describing project information about program references, activities, textual description of project scope and objectives, details about partners and so on;
- *Locations* - 11,390 tuples with 8 attributes describing details about geographical localisation of project partners;
- *Budgets* - 5,670 tuples with 13 attributes describing details about amount and state of project funding

and one metadata file describing structure and meaning of each the previous files, according to the Open Data standard. Table 1 shows a sketch of *Projects*, *Locations* and *Budgets* files, representing information useful for the following discussion in form of couples 'attribute/value'. It is important to underline that the approach taken here is somehow different from the ones usually adopted when performing network analysis in other fields. We have adopted intensive database techniques, while, for example, the treatment of authors of scientific papers with the same name in a social network of scientific collaboration is done automatically by computer algorithms, and errors like the correct identification of the same author represented by two different names (*e.g. J. Smith* and *John Smith*) are not

Table 1 Sketch of the structure of original files from PON R&C

PROJECT

UPC	title	smart_cities	social_innovation	…	healthcare	FC	name
PON04a2_A	PRISMA	1	0	…	1	84001850589	INFN
PON04a2_A	PRISMA	1	0	…	1	84001850589	INFN
PON04a2_A	PRISMA	1	0	…	1	80002170720	UNIBA
…	…	…	…	…	…	…	…

LOCATION

UPC	FC	name	kind	region	…
PON04a2_A	84001850589	I.N.F.N.	PRI	Apulia	…
PON04a2_A	80002170720	UNIBA	University	Apulia	…
…	…	…	…	…	…

BUDGET

UPC	FC	name	total_cost	total_funded	…
PON04a2_A	84001850589	INFN - Apulia	2231915.7	1785532.57	…
PON04a2_A	80002170720	University of Bari	2052539	1642031.2	…
…	…	…	…	…	…

Example rows report some features of project titled PRISMA (UPC = PON04a2_A). Note how: 1) *INFN* appears three times with three different values of the name attribute: *INFN*, *I.N.F.N.* and *INFN - Apulia*; 2) in PROJECT table INFN has two duplicate rows.

solved, but just treated statistically [18, 19]. We think that, while this is perfectly reasonable in that case, when studying a productive system like the one we are interested in, all the actors must be correctly identified, and in general errors like these should be reduced to the minimum, in order for the analysis to be reliable and really usable by policy makers.

In order to accomplish this task, we have defined a proper database model able to store all the projects data. First, we have created one table for each of the previous datasets, exploiting the following keys to link tables: `UPC` attribute is the unique identifier of projects (Unified Project Code) and `FC` is the unique identifier of partners (Fiscal Code). In order to improve dataset quality we have solved textual description encoding and numerical value format. Moreover, we have overcome name mismatching by inserting a unique label for each partner. Such label represents a convenient choice among the multiple names associated to the same fiscal code (FC) in the original datases[d] (*e.g.*, between '`I.N.F.N.`' and '`INFN - Apulia`' associated to the same FC '`84001850589`', we have chosen '`INFN`' as label). We underline that this data cleaning step is a key aspect in evaluations based on network analysis, in which results are sensitive to lacking data, and it is not possible to sample the population to extract useful information [20]. We are aware of the fact that such a problem could be much more evident in very large databases (*i.e.* the ones containing millions or more tuples, rather than thousands, like in the case under examination here), and we think the only viable solution is pushing institutions towards producing better open data. After the procedure described above, we have obtained a normalised database designing a many-to-many relation among Project and Partner tables and deleting duplicated and bad-formed tuples. Exploiting SQL standard queries, from our database we have selected 300 projects with at least 2 partners (thus suitable for network analysis). Those involve 769 distinct partners, for a total cost of the projects of 2,500 M euro (around 78% of the total cost of PON R&C projects), divided into Universities (33), Public Research Institutes (21), non-Public Research Institutes (44), Micro Enterprises (203), Small Enterprises (232), Medium Enterprises (58), Large Enterprises (163). It is significant that ∼10% of the total number of funded projects, the ones involving a network of relations, represents ∼78% of the total budget. This is an indication of the importance of relations in the Italian productive system. We note that 24 partners, we name N.C. partners, are not classified (in the original datasets) and that the *Social Innovation* action area does not include any project. In Figure 2, the distribution of fundings for the selected 300 projects is shown. Figure 2 (part (a)) represents cost distribution among different kinds of partners, expressed in percentage w.r.t. the cost of all selected projects, divided for kind of

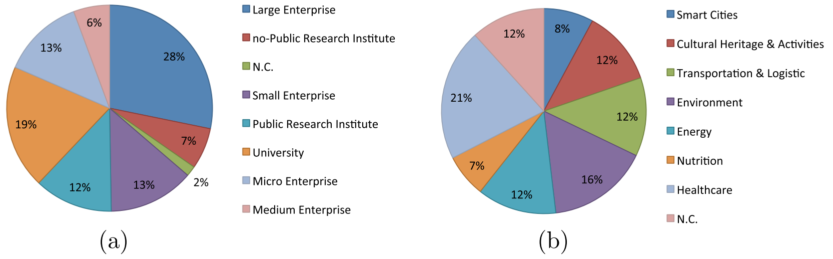
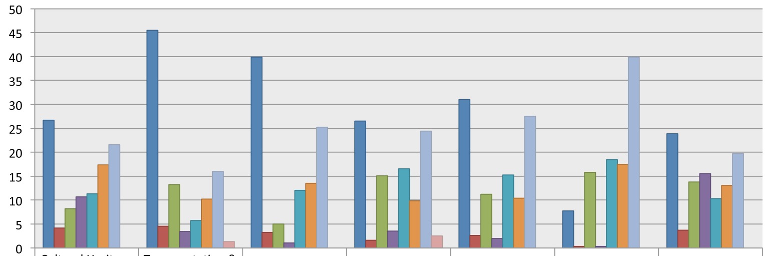


Figure 2 Distribution of fundings for 300 projects within the PON R&C program. Part **(a)** represents cost distribution among different kinds of partners, expressed in percentage w.r.t. the total cost; part **(b)** shows the cost distribution among the action areas mentioned above (exception made for the empty *Social Innovation* one).



	Cultural Heritage & Activities	Transportation & Logistic	Smart Cities	Environment	Energy	Nutrition	Healthcare
■ Large Enterprise	27	46	40	26	31	8	24
■ Medium Enterprise	4	5	3	2	3	0	4
■ Micro Enterprise	8	13	5	15	11	16	14
■ no-Public Research Institute	11	3	1	3	2	0	16
■ Public Research Institute	11	6	12	17	15	18	10
■ Small Enterprise	17	10	14	10	10	17	13
■ University	22	16	25	24	28	40	20
■ N.C.	-	1	-	2	-	-	-

Figure 3 Distribution of fundings for each kind of partner, for each action area.

partners and Figure 2 (part (b)) the cost distribution among action areas above mentioned (except the empty *Social Innovation* area). As a general overview, Figure 3 shows the distribution of fundings for each kind of partner, for each action area. In the calculation of such cost distribution, we have considered the attribute `total_cost`, among different ones concerning budgets, since it is the only one having a `NOT NULL` value for each tuple in the original dataset. Starting from concepts underlying the TSA definition, we have built a set of SQL stored procedures performed on our database. Inputs are tables and attributes representing significant elements for the networks construction and outputs are the ad-hoc generated views with aggregated and derived data. The main generated data views are the following:

- *Partner-to-Partner* - the distinct couples of partners involved in the same project;
- *Project-to-Project* - the distinct couples of projects having at least one partner in common, together with the calculated number of shared distinct partners;
- *Project-to-Partner* - the set of distinct involved partners for each project;
- *Partner-to-Funding* - the funding, for each beneficiary, calculated considering all the PON R&C projects it is involved in;
- *Beneficiary-to-Beneficiary* - the distinct couples of beneficiaries having at least one project in common, together with the calculated number of shared distinct projects.

4 Network analysis

This section contains a description of the activities described by points 7-8 of Figure 1. The network analysed here is an *affiliation network* [18, 19], constructed in such a way that every university, research institution or enterprise that has been funded by the program is a vertex of the graph and there is an edge between two vertices if the corresponding participants are part of at least one TSA, for at least one funded project. In this way, the graph is the union of complete, undirected, unweighted graphs, each representing a TSA, in which every node is connected to all the others. The network structure is due to vertices participating to more than one projects, in more than one TSA. In our analysis we have not considered vertices that have been funded without participating in any TSA. The resulting network is shown in Figure 4; it has 769 vertices and 4,868 edges, and is not connected.

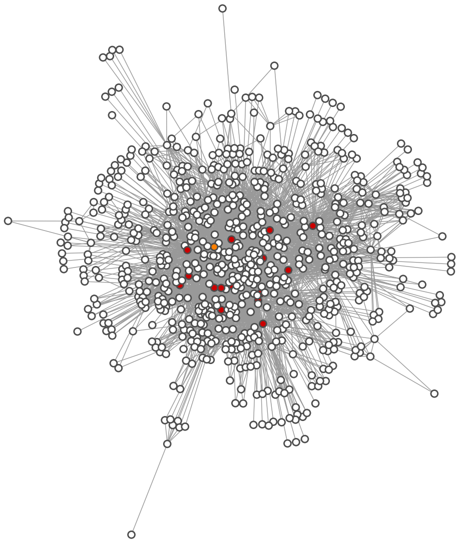


Figure 4 Network structure of the Italian PON R&C funding program. Each vertex is a university, research institution or enterprise funded by the program; two vertices are connected if they are part of a TSA for at least one project. Only the principal giant component is depicted (other connected components have less than six vertices each). The set of nodes constituting the centre of the (giant component of the) PON R&C network is highlighted in red, while the main hub is depicted in orange. The centre include public and private research institutions, all the major Universities involved in the program, and also some large private enterprises.

It is made of one giant component, composed of 744 vertices and 4,845 edges, and 10 small complete graphs of order 5 (one graph), 3 (two graphs) and 2 (seven graphs). The graph has been analysed, and several properties [21, 22] have been extracted to support the evaluation of the PON R&C public funding program. Such properties belong to two main classes: local and global ones. Local properties are features of single vertices or edges, and, in particular, *centrality coefficients* are evaluated, in order to understand the importance of the nodes within the network. Global properties involve the network as a whole instead, and are used to describe the full program, independently of the single nodes.

4.1 Local properties

Properties of vertices evaluated in the present analysis include *degree centrality*, *betweenness centrality*, *closeness centrality*, *eccentricity*, *eigenvector centrality*, *radiality centrality* and *PageRank centrality*, based on the Google PageRank algorithm [23].

The highest values of all centralities is found in correspondence to public research institutions, like universities and specific research centres. In particular, the Italian National Research Centre (*CNR*) shows the best values for all the indicators. It is worth saying that it is a peculiar vertex of the network, being composed of 104 institutes spread over geographically distributed sites (in all the biggest cities in Italy), and covering a large spectrum of activities in many fields, from pure research, to applied disciplines. Probably, it would be better to split such vertex and consider each site, or department, separately, but the dataset does not contain such details. On the contrary, dividing the CNR in many entities would in a certain sense spoil its central nature in the Italian panorama. Resolving this controversy is interesting, but is over the purposes of the present paper, and is left for a future work, when more detailed Open Data will be available. Apart from cases like the one described, the central role of public research institution for the network structure is clear from all the centralities.

Degree centralities are discussed in detail in the next Section 4.2, since global properties of the network can be inferred from the distribution of such quantities, despite being them local in nature.

Betweenness [24] measures the importance of a node for traffic of information across the network. Large betweenness centrality of a vertex indicates that many shortest paths between couples of other vertices pass through that node. The relevance of this quantity for program evaluation stands in the possibility of assessing the role of institutions/enterprises for the eventual aggregation of 'far' nodes. For example, a policymaker interested in promoting a program aimed at aggregating and consolidating the productive system of a region should pay attention not to spoil the edge betweenness of the network of relations between the actors involved in the program.

Closeness centrality indicates whether a node is at a short average distance from every other reachable vertex, with higher closeness meaning shorter distance. A variant is radiality centrality, which gives higher weight to the neighbourhood of the node. From the social/economical point of view these quantities give indication about how easily an institution/enterprise can connect to all the other members of the network (and, so, of the productive system). For example, an enterprise with high closeness centrality could be the right promoter for initiatives like the creation of technological districts, associations or lobbies. Exploiting the information contained in this quantity, a policymaker could more easily head the productive system in the desired direction with focused regulatory interventions.

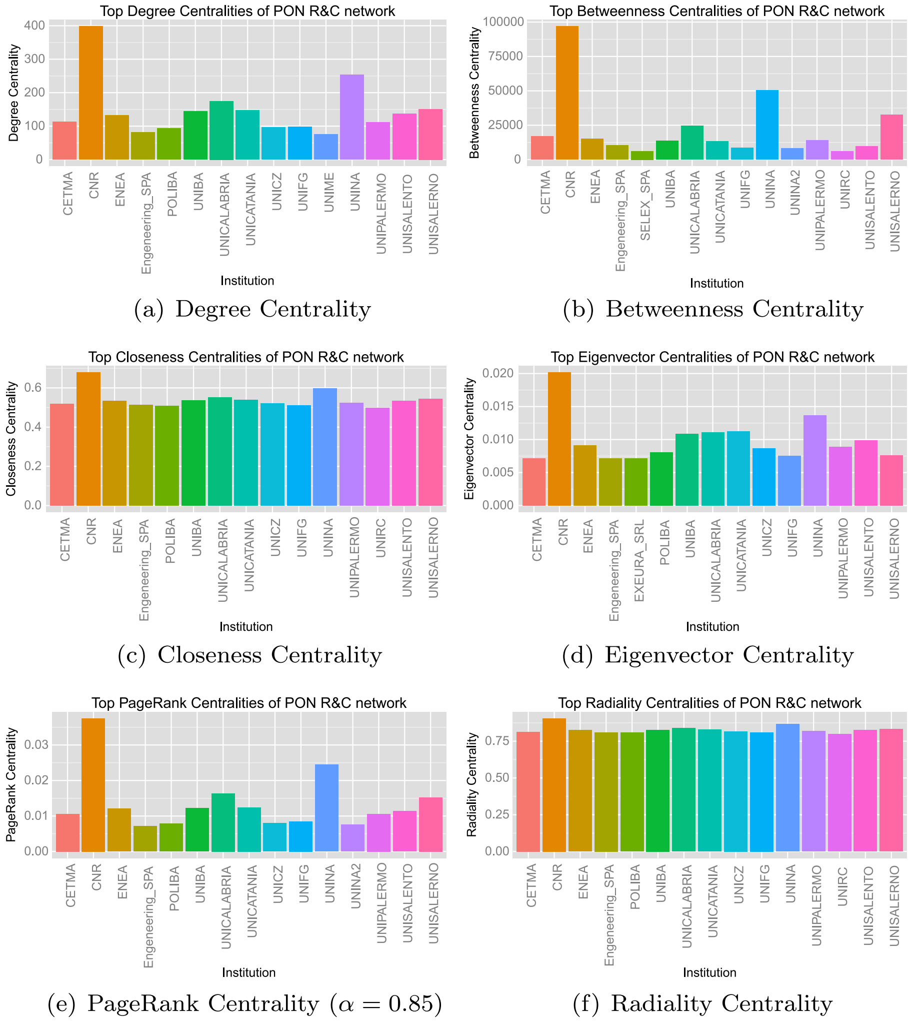
Eccentricity is the maximum value of the distances between a node and any other node in the network. It gives an idea of how central a vertex is within the network, with smallest values corresponding to more central nodes.

High eigenvector centrality is assigned to vertices that are connected to many other well-connected vertices. It can be used to identify the best way to spread a trend within the productive system represented by the network. A variant of eigenvector centrality is PageRank centrality, which is a way of measuring the importance of a node within a graph. The original algorithm was created by Larry Page and Sergey Brin in 1986 at Stanford University [25–27] and is widely used by Google to measure the importance of website pages. The algorithm used here is given by the solutions of:

$$\mathbf{r} = \alpha \mathcal{A}^{\mathsf{T}} \mathcal{H} \mathbf{r}, \tag{1}$$

where $\mathbf{r}$ is the vector of the PageRank centralities for each node, $\mathcal{A}$ is the adjacency matrix of the graph, $\mathcal{H}$ is the diagonal matrix consisting of $1/\max\{1, d_i\}$, d_i being the degree of the ith vertex, and α is the damping factor, an empirical parameter[e] usually set to $\alpha = 0.85$ [23]. This is an example of how fruitful the application of social network analysis can be to the evaluation of the effects of public funding, also by means of well known and successful algorithms, as the PageRank one.

In Figure 5 the largest values of all the centralities of the (giant component of the) PON R&C network are reported. As stated before, the CNR has the highest values for all the centralities, probably due to its being scattered along the whole country. Nevertheless, it is important to observe that public research institutions, mainly universities, occupy the top positions for every centrality, while in the lowest positions we find private enterprises, no matter whether they are large or small (except for the Polytechnic of Bari, which is a public university and has a low value of PageRank centrality). The central role of public research institutions for the network of relations underlying (at least part of) the Italian productive system is clear from Figure 5. Eccentricity is not reported in the figure,



(a) Degree Centrality

(b) Betweenness Centrality

(c) Closeness Centrality

(d) Eigenvector Centrality

(e) PageRank Centrality ($\alpha = 0.85$)

(f) Radiality Centrality

Figure 5 The fifteen largest values of each vertex centrality for the (giant component of the) PON R&C network. The highest positions are occupied by public research institutions.

since a high number of vertices share the same value of this centrality, meaning that the network is somewhat 'equally spaced'. For degree and betweenness centralities only the largest values are reported, since the smallest ones are trivial.

The last centrality considered here is edge betweenness, which is a property of the edges of the network (rather than vertices), and measures how central a link is for the connections between nodes. It is measured by counting the number of shortest paths the edge belongs to, and gives a quantitative idea of how much the relation between two institutions/enterprises is important for the 'communication' between all the actors composing the network. As the number of edges is much larger than the number of vertices, and since it is necessary to evaluate the shortest path between any couple of nodes, the calculation of such centrality is a resource intensive process. The largest values of the edge betweenness for the (giant component of the) PON R&C network are reported in Figure 6. Also in this case, the most important relationships (edges) between the nodes of the network are the ones between public research institutions, while small enterprises give small to no contribution to the geodesics.

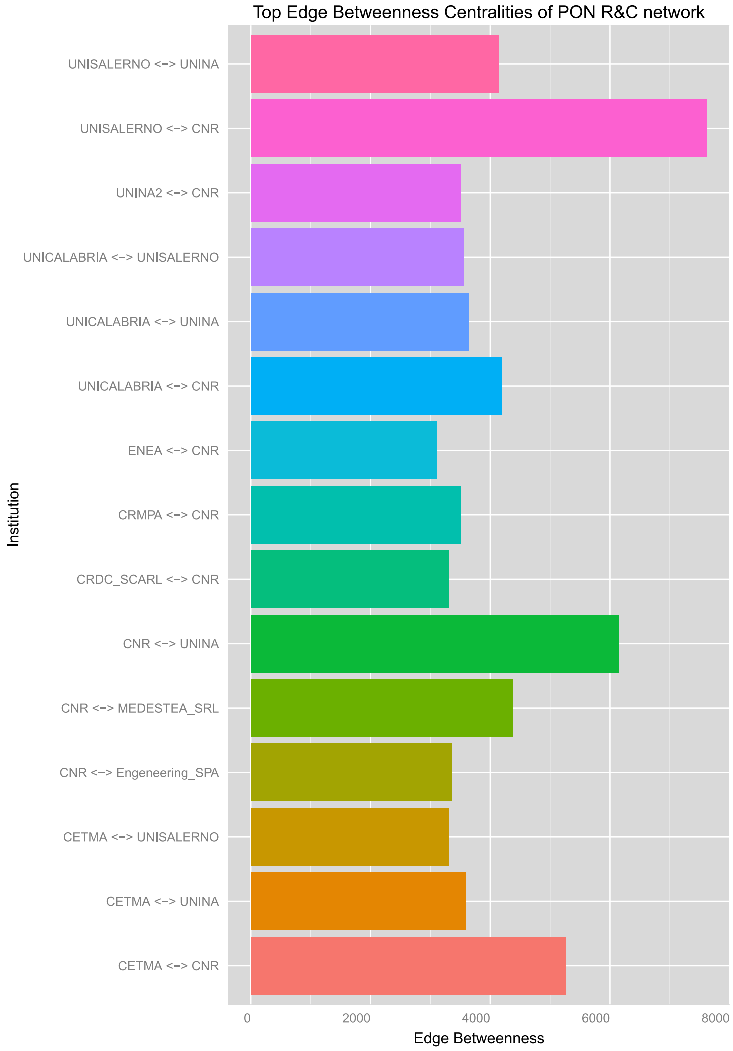


Figure 6 The fifteen largest values of edge betweenness centrality for the (giant component of the) PON R&C network. The highest positions are occupied by links involving public research institutions.

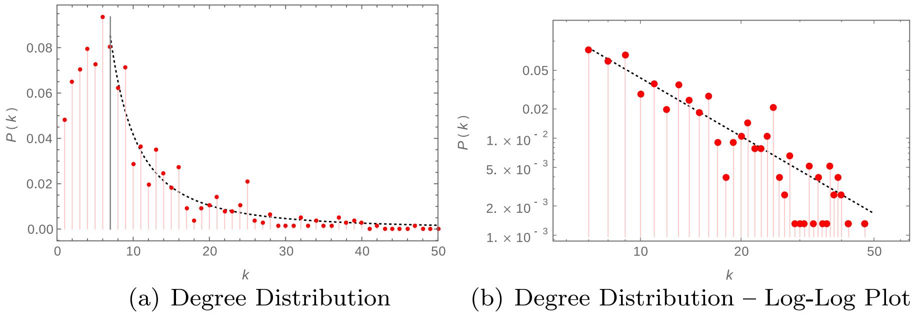
4.2 Global properties

The first property analysed here is the degree distribution of the vertices, *i.e.* the frequencies of the degree centralities described in the previous Section 4.1. The importance of such distribution stands in the possibility of inferring from it information about the topology of the graph, and in particular to understand if the network is *scale-free* [28]. The property of being scale-free is shared by many real networks, showing power law-shaped degree distributions $P(k) = Ak^{-\gamma}$, with exponents usually varying in the range $2 < \gamma < 3$, which have the same form at all scales.

This is of particular interest since power laws are commonly associated with second-order phase transitions in dynamical systems. Phase transitions in complex networks represent an interesting research field [29, 30], but the graph considered here is static, so no considerations can be made in this respect. Anyway, this is an interesting perspective for a future work, in which dynamics can be taken into account.

Scale-free networks have an inhomogeneous degree distribution, with many nodes having more connections than the average (*hubs*). The hubs follow a hierarchy, in which large ones are connected to smaller ones, which are themselves connected to even smaller ones, and so on. This feature makes the network robust against casual failures, since the removal of a random vertex would not systematically affect the main hubs, and connectedness would not be spoiled. Hence, scale-free graphs are a desirable result for policymakers interested in generating a solid network of relationships between productive actors on the territory. Apart from being a strong point for networks, hubs also represent a weakness, since their systematic removal would quickly destroy the network. The property of being scale-free is an important point to be taken into account for an evaluator, as we will show below, in order to monitor and evaluate the results of funding programs. Moreover, it suggests to decision makers that effort should be put in promoting funding program which hubs can profit from.

The degree distribution of the PON R&C network is shown in Figure 7. The tail (starting from the upper bound of the median interval, $m = 7$) is fitted very well to a power-law function of the form $P(k) = Ak^{-\gamma}$ with $A = 4.156 \pm 0.375$ and $\gamma = 1.998 \pm 0.040$. To obtain the fits, a nonlinear regression based on Newton method [31] has been used. A comparison with another fit, to an exponential distribution $P(k) = Ae^{-\gamma k}$ with $A = 0.241 \pm 0.011$ and $\gamma = 0.164 \pm 0.005$, shows that the former fits the distribution slightly better than the latter, with $R^2_{\text{pow}} = 0.935$ and $R^2_{\text{exp}} = 0.929$. Such a small difference between the values of R^2 is not a strong indication of the fact that a power-law fits the distribution better than an exponential law, but together with the fact (shown below) that higher moments grow, it is sufficient to assess the power-law nature of the distribution. In fact, for a power-law distribution with tail of $\mathcal{O}(x^{-\nu})$ the moments of order $n \geq (\nu - 1)$ diverge, and in general higher order moments are larger in size with respect to the lower order ones (this is not true for exponential distributions). Standing the known difficulties in evaluating the nature of the degree distribution, due to noise coming from the finiteness of the sample (especially from boundary values), the present result is satisfactory in assessing the property of the PON R&C network of being scale-free. More refined methods could be used to evaluate the parameter γ with higher precision like *e.g.* the Kolmogorov-Smirnov test [32], but this is outside the purposes of the present work.



(a) Degree Distribution (b) Degree Distribution – Log-Log Plot

Figure 7 The degree distribution of the network based on the PON R&C funding program is shown. The tail is fitted to a power-law function of the form $P(k) = Ak^{-\gamma}$ with $A = 4.156 \pm 0.375$ and $\gamma = 1.998 \pm 0.040$. The vertical grey line in part **(a)** shows the position of the cutoff (corresponding to the upper bound of the median interval) used to fit the tail. Part **(b)** shows the tail and the fitted line in Log-Log scale.

Table 2 Moments of the degree distribution of the PON R&C network

Moment	Value
Expected value	$\langle k\rangle = 12.661$
Mode	$M = 6$
Median	$6 < m < 7$
Standard deviation	$\sigma = 23.781$
Skewness	$s = 8.876$
Kurtosis	$\kappa = 113.882$

The distribution has an expected value $\langle k\rangle = 12.661$, mode $M = 6$, median $6 < m < 7$, standard deviation $\sigma = 23.781$, skewness $s = 8.876$ and kurtosis $\kappa = 113.882$ (all the moments are shown in Table 2).

Once assessed such power-law nature, it is interesting to identify the main hubs. In the PON R&C network considered here, hubs are public research centres, and this represents a strong point for the relationships of the involved productive system. In fact, it is natural, in the lifecycle of a productive system, that some enterprises rise while others fall, resulting, in the language of networks, in the random removal of vertices described above. Anyway, as previously stated, the random removal of vertices from a scale-free network does not spoil connectivity, which happens with the systematical removal of the main hubs instead. In this case, it is unlikely that one of the main hubs, identified here with large public research centres, could disappear, since this would mean *e.g.* the closure of a large public university, a quite rare event. This picture was partly expected, since in many cases it was mandatory to involve public research institutions in the TSAs. Nevertheless it still represents a strong indication for a decision maker, suggesting that it is 'safer' including public research in a future program, since it is the easiest way to keep a solid relationship network within the productive system.

Another way of assessing if a network is scale-free consists in evaluating the distribution of local clustering coefficients, *i.e.* the number of edges connecting the neighbours of each vertex v, divided by the number of edges of a complete graph of the same cardinality of the neighbourhood of v [33]. The PON R&C network represents a special case, in which local clustering coefficients are less important, the majority of them being close to 1 by construction. In fact, since the graph is a union of complete graphs, it is likely that the neighbourhood of a vertex is fully connected, implying the closeness to one of the local clustering coefficient. The global clustering coefficient $\mathcal{C}$, *i.e.* the fraction of paths of length two that are closed (over all paths of length two), is much more significant instead, and it takes a small value $\mathcal{C} = 0.215$ for the giant component, meaning that the network is not strongly clustered. From the political and sociological point of view, this is an interesting point, since the network is made by 'scattered' relationships, despite being composed of 'closed' TSAs.

Other important features that can guide the policymaker in evaluating the effects of the program or planning future ones are *vertex connectivity* V_c and *edge connectivity* E_c, *i.e.* the smallest number of vertices or edges to be removed in order to disconnect the graph, respectively. For the case under examination such quantities take value $V_c = 1$ and $E_c = 1$, meaning that the removal of a single node or edge can be catastrophic for network connectivity. Identifying and monitoring such nodes/edges can be very important in case of low values of such parameters, in order to keep the network of relations tightly connected.

Another important property of scale-free networks is that they are *small world* networks [34]. This means that relatively short paths exist between any two nodes (with respect to the large size of the graph), with an average shortest path length[f] $L \sim \mathcal{O}(\log N)$, N being the total number of vertices. This is due to the existence of links between vertices belonging to farther parts of the graph, having the role of connecting them and reducing distances to few hops. Usually, in scale-free networks such vertices are the hubs and the small-world property is enhanced when $2 < \gamma < 3$ where $L \sim \mathcal{O}(\log\log N)$ (while it is $L \sim \mathcal{O}(\log N)$ when $\gamma > 3$) [35]. For the PON R&C network, $\gamma = 1.998 \pm 0.040$, $L = 2.532$, $\log N = 6.645$ and $\log\log N = 1.889$, so the small world property is enhanced, as expected when the vertex distribution follows a power-law with $2 < \gamma < 3$. Again, this cannot be considered a smoking gun proving that the network is scale-free, but just another indication in addition to the ones mentioned above.

Other global features of the network are the radius $\mathcal{R}$ and diameter $\mathcal{D}$ of the graph, defined as the minimum and the maximum eccentricity of all vertices, respectively, the eccentricity being the longest shortest path from a source node to every other vertex in the graph. For the PON R&C network $\mathcal{D} = 5$ and $\mathcal{R} = 3$, meaning that no vertex is more than 5 hops far from any other node, and that the farthest destination is never closer than 3 hops from any source. From the point of view of program evaluation, this means that PON R&C has been successful in creating (or intersecting) a network of close relationships between the funded actors. Being interested in promoting such a relationship network while defining the program, these could be good *ex post* indicators of the goodness of the obtained results.

The centre of the graph[g] is shown in Figure 4. It includes public research centres like CNR (which is also the main hub) and ENEA, all the major Universities involved in the program (Bari, Calabria, Catanzaro, Foggia, Naples, Palermo, Salento, Salerno), private research centres like CETMA, and also some large private enterprises like Avio S.p.A., Engeneering S.p.A., IBM, SELEX S.p.A., and EXEURA S.r.l. This is a strong indication that the network of funded projects gravitates around large poles involving research centres (public and private), which turn out to have a key role in aggregating entities. This can also be an explanation for the scale-free property of the graph, since *preferential attachment* is known to be a generating mechanism for this kind of networks [36–38], in which nodes prefer to link to vertices with high degree. It is reasonable to imagine that many small actors prefer forming TSAs including large research organisations, which are usually able to get more funds, rather than form TSAs between themselves. From the point of view of social networks and relationships, it is particularly interesting to study such feature side-by-side with assortativity [39], which indicates whether nodes of the graph tend to connect with their connectivity peers (vertices with similar degree) or not. In the first case the network is said to be *assortative*, while in the second case it is *anti-assortative*. This feature is quantitatively measured through the *assortativity coefficient* r, whose range is $-1 \leq r \leq 1$, $r = 1$ (-1) meaning a perfectly (anti-)assortative graph and $r = 0$ indicating no particular preference for the majority of the nodes. In the present network, $r = -0.173$, meaning that the graph is slightly anti-assortative. This means that the productive system funded by such program has a little tendency not to form *lobbies* among important actors, but to associate strongly connected hubs to smaller and less connected enterprises/institutions. From the socio-economical point of view, it seems reasonable that small enterprises turn to larger ones or to big research centres to benefit from sharing and collaborations. This

Table 3 Global properties of the PON R&C network

Property	Value
Radius	$\mathcal{R} = 3$
Diameter	$\mathcal{D} = 5$
Density	$\rho = 0.017$
Global clustering coefficient	$\mathcal{C} = 0.215$
Vertex connectivity	$V_c = 1$
Edge connectivity	$E_c = 1$
Average shortest path length	$L = 2.532$ ($\mathcal{O}(\log \log N)$)
Link efficiency	$\xi = 0.999$
Assortativity coefficient	$r = -0.173$

is an interesting result, since most social networks show assortative mixing by node degree [40], and it also has some implications on the topology of the network. First, anti-assortative networks are more susceptible to the removal of high-degree nodes (here represented by universities and research centres), which is an indication for the policymaker of the importance that public research has in the productive system, and of the possible disruptive effect of its underestimation. Second, in anti-assortative networks epidemics span to larger portions of the nodes than in similar assortative ones. This means that being anti-assortative is preferable for the spreading of knowledge and know-how in the productive system, making it more efficient. It is worth noting that in a recent work [12] a social network similar to the one studied here, concerning the funding of FP7 (Seventh Framework Programme) European research projects, has been found to be anti-assortative as well, and conclusions close to the ones put foward here are drawn. This could be an indication of some structural feature shared by graphs constructed starting from public funding programs, and we plan to further investigate this point in a future work. Lastly, *link efficiency* is a measure of traffic capacity within the network, representing how efficiently information can be transmitted along the graph. This parameter takes the very high value $\xi = 0.999$ in the PON R&C network, which is a strong indicator of robustness for the relations between vertices, especially in a graph with small density $\rho = 0.017$ as the one under examination.

The study of global properties of the PON R&C graph is given as an example showing how network analysis provides a concrete way of examining the role of funded actors within a program, supporting its *ex post* evaluation with the introduction of rather innovative indicators. In particular, it is able to describe the structure of the productive system, highlighting the key nodes for network connectivity, or vertices that have a central role, through quantitative (thus evaluable) indicators, which for the present case are summarised in Table 3.

4.3 Community structure

The community structure of a graph is a global property, but a separate section is dedicated to it, since it has a special role, of particular importance for program evaluation. In the PON R&C network 15 communities are found with the Newman-Girvan algorithm [41], composed of 207, 136, 129, 113, 75, 29, 8, 7, 7, 7, 6, 6, 5, 5, and 4 vertices, respectively. The algorithm consists in recursively removing from the graph the edges with the highest edge betweenness and recalculating the edge betweenness for the new graph obtained at each step. This procedure generates a dendrogram of sets of communities, from which the set with largest modularity is chosen. The community structure of the PON R&C network

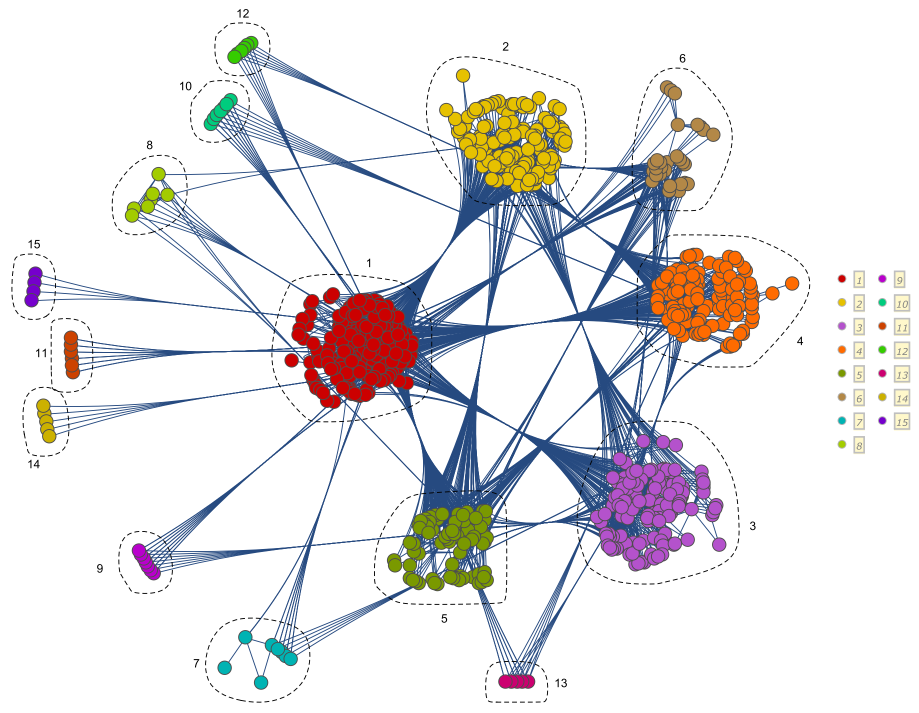


Figure 8 The community structure of the (giant component of the) PON R&C network. 14 communities are highlighted, found with the Newman-Girvan algorithm.

is shown in Figure 8. The reason why it is so important is that it represents an unbiased way of discovering the existence of groups within a certain network of relationships, and highlighting such groups can be very important for the analysis of a productive system like the one described by the graph under examination. The PON R&C network shows strongly heterogeneous communities, with hugely populated groups and very small ones. An important point, that can be interesting for an *ex post* program evaluator, is that when communities grow in size, they tend to include important nodes. For example, the biggest community, made of 186 vertices, include the CNR in it, which shows the record values for all the centralities, as stated in Section 4.1.

Moreover, comparing the community structure coming from network analysis with the one expected on the basis of external (economical, political, and social) considerations can enrich the evaluation by introducing a different point of view on the system under examination, not driven by 'human' considerations, but purely mathematical in nature. The distribution in percentage of action areas within each community is shown in Figure 9. It can be seen that the communities found with network analysis are not directly linked to action areas, at least the largest ones. Other algorithms can be used to extract the community structure of a graph, like the leading eigenvector [42], the multi-level modularity [43] or the spin-glass [44], and refined methods can be used to which one gives the most significant results, like *e.g.* a consensus analysis [45]. This kind of study is outside the purposes of the present paper, since it must be policy-driven, rather than research-driven, in order to be of interest for program evaluation. We plan to develop similar analyses in future work.

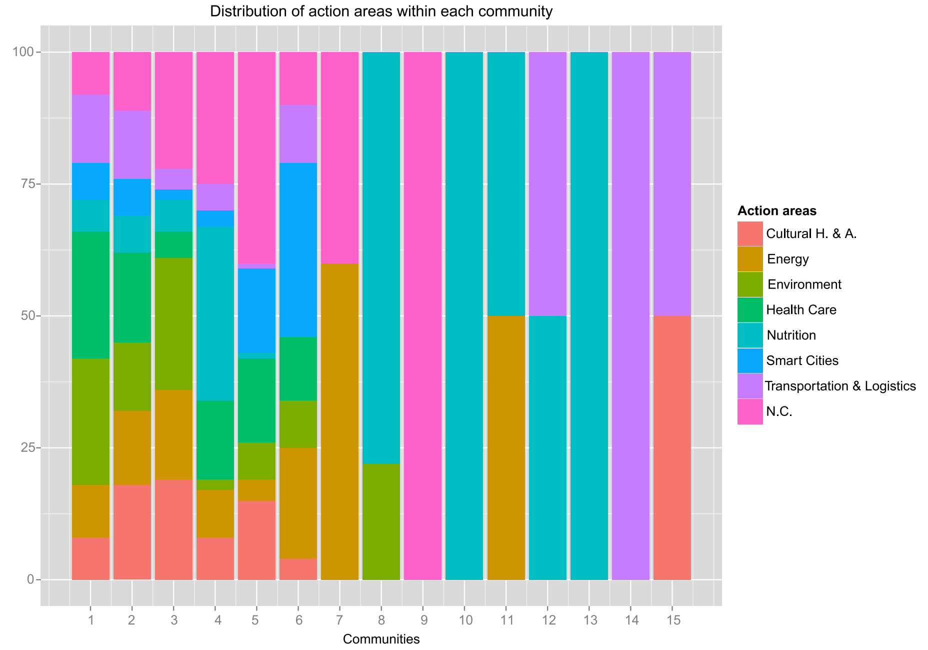


Figure 9 Distribution of action areas within each community.

5 Conclusions and perspectives

In this paper we have used techniques borrowed from complex network analysis to evaluate the effects of a public funding program on the relations between the funded 'actors'. The PON R&C program involves a large number of actors and is extended over a period of seven years (2007-2013). The dataset is completely made of Open Data, and we have shown a way of concretely using information made available by Governments, in the spirit promoted by current global guidelines. We have described the full process of knowledge management, from data acquisition, to cleaning, model building and querying. The whole chain is data oriented and is focused in retaining every piece of available information, in order for the output of the analysis to show the highest possible accuracy.

The processed PON R&C data have been used for complex network analysis, and the resulting network has 769 vertices and 4,868 edges. We have evaluated the most important centralities for each node, plus some relevant global properties of the graph. The outcome of our analysis shows a dominant role of public (and, but less importantly, also private) research institutions within the Italian productive panorama, at least for the part portrayed by the program under examination. Universities and research centres play the role of the 'glue' for this particular program, *i.e.* they are responsible of the connectedness of the network, and a failure involving some of them would be disruptive for the whole productive system. This picture was partly expected, due to the way the program has been realised, since in many cases it was mandatory to involve public research institutions in the projects. Nevertheless it can be useful to use such result as an *ex-post* indicator. Moreover, we have found that the PON R&C network is anti-assortative, an unusual feature of social networks, shared only with other cases involving FP7 public program of European research funding, preferable than the most common assortative mixing for the spreading of knowledge and know-how in the productive system and for its efficiency.

We have shown that social network analysis can produce useful results for program evaluators, since it allows to consider, in a quantitative fashion, very important aspects that are usually ignored, due to common difficulties in quantifying them. A mathematical description of the structure of relations generated by a national funding program is the example shown in this work. Indicators such as vertex and edge centralities have been used to generate a ranking between the main actors involved in the program, as shown in Figures 5 and 6. We hope that the procedure and the results described in the present paper can help opening interesting new perspectives from new indicators for decision and policy makers and program evaluators, providing them with an useful tool.

Many possibilities are left open by the present work. First of all, as mentioned in Section 3, around ∼78% of the total budget is concentrated into ∼10% of the funded projects. This suggests that introducing information about the financial aspect into the network analysis could be interesting and meaningful for the evaluator. This could be done in many different ways, from simple visualisation techniques in sociograms, like relating the size of the nodes to the total funding received by the actor, to more refined analysis, like defining weighted networks with weights related to budgets. We plan to investigate these directions in future works. Other planned future activities include the introduction of dynamical networks, involving the study of temporal series, refining of network analysis techniques, *e.g.* by introducing different kinds of weighted networks and related features, and generalising the analysis extending it to different levels [7, 12]. Moreover, the expected improvement in quality of Open Data (for example increasing the level of detail within public research institution, *e.g.* discriminating among single departments rather than universities) could lead to many interesting improvements of the present analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors designed research. SN designed the network model. ET designed the data model. SN and ET prepared the data for network analysis. All authors analysed data. All authors read and approved the final manuscript.

Author details

[1]Istituto Nazionale di Fisica Nucleare - Sezione di Bari, via Orabona 4, Bari, I-70125, Italy. [2]Dipartimento Interateneo di Fisica '*M. Merlin*', Università degli Studi di Bari '*A. Moro*', via Orabona 4, Bari, I-70125, Italy. [3]Comune di Bari - Ripartizione Innovazione Tecnologica, Sistemi Informativi e TLC, Corso Vittorio Emanuele II, 143, Bari, I-70122, Italy. [4]Dipartimento di Scienze Fisiche dell'Ambiente e della Terra, Università degli Studi di Siena 1240, via Roma 56, Siena, I-53100, Italy.

Acknowledgements

We thank N Coniglio for valuable suggestions and discussions. SN thanks R Anglani and V Mariani for discussions. SN, and NA acknowledge funding by the Italian MIUR grant *PON PRISMA* Cod. PON04a2_A. ET acknowledges partial funding by the Italian MIUR grant *PON PRISMA* Cod. PON04a2_A.

Endnotes

[a] http://ec.europa.eu/eip/smartcities/.

[b] Resource Description Framework - http://www.w3.org/standards/techs/rdf#w3c_all.

[c] The path from Open Data to Linked Open Data was firstly introduced by Sir Tim Berners-Lee in the 5 Stars Model at the Gov 2.0 Expo in Washington DC in 2010, where costs and benefits for both publishers and consumers of LOD are explained.

[d] See *e.g.* the *name* attribute in Table 1.

[e] Representing the probability that a traveler randomly navigating the network continues doing it at a given point.

[f] *I.e.* the average length of all shortest paths between couples of vertices of the graph.

[g] *I.e.* the set of vertices with minimum eccentricity.

Received: 27 October 2014 Accepted: 23 June 2015 Published online: 09 July 2015

References

1. Smith AFM (1996) Mad cows and ecstasy: chance and choice in an evidence-based society. J R Stat Soc, Ser A, Stat Soc 159(3):367-383. doi:10.2307/2983324
2. Nutley SM, Davies HTO, Smith PC (2000) What works?: evidence-based policy and practice in public services. Policy Press
3. Solesbury W, Policy EUCfEB, Practice (2001) Evidence based policy: whence it came and where it's going. ESRC UK Centre for Evidence Based Policy and Practice
4. Young K, Ashby D, Boaz A, Grayson L (2002) Social science and the evidence-based policy movement. Soc Policy Soc 1:215-224. doi:10.1017/S1474746402003068
5. Kay A (2011) Evidence-based policy-making: the elusive search for rational public administration. Aust J Public Adm 70(3):236-245. doi:10.1111/j.1467-8500.2011.00728.x
6. Legrand T (2012) Overseas and over here: policy transfer and evidence-based policy-making. Policy Stud J 33(4):329-348. doi:10.1080/01442872.2012.695945
7. Provan KG, Milward HB (2001) Do networks really work? A framework for evaluating public-sector organizational networks. Public Adm Rev 61(4):414-423. doi:10.1111/0033-3352.00045
8. Ferlie E, Lynn LE, Pollitt C, Klijn EH (2007) Networks and inter-organizational management: challenging, steering, evaluation, and the role of public actors in public management. Oxford University Press, London http://www.oxfordhandbooks.com/10.1093/oxfordhb/9780199226443.001.0001/oxfordhb-9780199226443-e-12
9. Penuel WR, Sussex W, Korbak C, Hoadley C (2006) Investigating the potential of using social network analysis in educational evaluation. Amer J Eval 27(4):437-451. doi:10.1177/1098214006294307
10. Horelli L (2009) Network evaluation from the everyday life perspective: a tool for capacity-building and voice. Evaluation 15(2):205-223. doi:10.1177/1356389008101969
11. Ploszaj A (2011) Networks in evaluation. In: Olejniczak K, Kozak M, Bienias S (eds) Evaluating the effects of regional interventions. A look beyond current structural funds practice. MRR, Warszawa
12. Tsouchnika M, Argyrakis P (2014) Network of participants in European research: accepted versus rejected proposals. Eur Phys J B 87(12):292. doi:10.1140/epjb/e2014-50450-4
13. Regione Puglia, Area Politiche per il Lavoro Sviluppo e Innovazione Servizio ricerca Industriale e Innovazione, InnovaPuglia S.p.A. (2014) Agenda Digitale Puglia 2020. Bollettino Ufficiale della Regione Puglia (BURP) 128:33423-33502
14. Bizer C, Heath T, Berners-Lee T (2009) Linked data-the story so far. Int J Semantic Web Inf Syst 5(3):1-22
15. Lenzerini M, Vassiliou Y, Vassiliadis P, Jarke M (2003) Fundamentals of data warehouses. Springer, Berlin
16. Abiteboul S, Hull R, Vianu V (1995) Foundations of databases, vol 8. Addison-Wesley, Reading
17. Janowicz K, Hitzler P, Adams B, Kolas D, Vardeman C (2014) Five stars of linked data vocabulary use. Semant Web 5(3):173-176
18. Newman MEJ (2001) Scientific collaboration networks. I. Network construction and fundamental results. Phys Rev E 64:016131. doi:10.1103/PhysRevE.64.016131
19. Newman MEJ (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E 64:016132. doi:10.1103/PhysRevE.64.016132
20. Olejniczak K, Bienias S, Kozak M (2012) Evaluating the effects of regional interventions. A look beyond current structural funds practice. Ministry of Regional Development, Republic of Poland
21. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: structure and dynamics. Phys Rep 424(4-5):175-308. doi:10.1016/j.physrep.2005.10.009
22. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47-97. doi:10.1103/RevModPhys.74.47
23. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst 30(1-7):107-117. Proceedings of the Seventh International World Wide Web Conference. doi:10.1016/S0169-7552(98)00110-X
24. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35-41
25. Page L (2001) Method for node ranking in a linked database. Google patents. US Patent 6,285,999. http://www.google.com/patents/US6285999
26. Page L (2011) Annotating links in a document based on the ranks of documents pointed to by the links. Google patents. US Patent 7,908,277. http://www.google.com/patents/US7908277
27. Page L (2014) Scoring documents in a linked database. Google patents. US Patent 8,725,726. http://www.google.com/patents/US8725726
28. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509-512. doi:10.1126/science.286.5439.509
29. Albert R, Barabási A-L (2000) Topology of evolving networks: local events and universality. Phys Rev Lett 85:5234-5237. doi:10.1103/PhysRevLett.85.5234
30. Iglói F, Turban L (2002) First- and second-order phase transitions in scale-free networks. Phys Rev E 66:036140. doi:10.1103/PhysRevE.66.036140
31. Avriel M (2003) Nonlinear programming: analysis and methods. Dover books on computer science series. Dover Publications, New York
32. Eadie WT, Drijard D, James FE, Roos M, Sadoulet B (1971) Statistical methods in experimental physics. North-Holland, Amsterdam
33. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440-442
34. Milgram S (1967) The small world problem. Psychol Today 2(1):60-67
35. Chung F, Lu L (2002) The average distances in random graphs with given expected degrees. Proc Natl Acad Sci USA 99(25):15879-15882. doi:10.1073/pnas.252631999
36. Barabási A-L, Albert R, Jeong H (1999) Mean-field theory for scale-free random networks. Physica A 272(1-2):173-187. doi:10.1016/S0378-4371(99)00291-5
37. Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51(4):1079-1187. doi:10.1080/00018730110112519

38. Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. Phys Rev Lett 85:4629-4632. doi:10.1103/PhysRevLett.85.4629
39. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701. doi:10.1103/PhysRevLett.89.208701
40. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67:026126. doi:10.1103/PhysRevE.67.026126
41. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113. doi:10.1103/PhysRevE.69.026113
42. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104. doi:10.1103/PhysRevE.74.036104
43. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
44. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74:016110. doi:10.1103/PhysRevE.74.016110
45. Lancichinetti A, Fortunato S (2012) Consensus clustering in complex networks. Sci Rep 2. doi:10.1038/srep00336