Scientometrics (2021) 126:8517–8539
https://doi.org/10.1007/s11192-021-04120-9

Analysis of scientific collaboration network of Italian Institute of Technology

Enrico di Bella[1] · **Luca Gandullia**[1] · **Sara Preti**[1,2]

Received: 14 January 2021 / Accepted: 26 July 2021 / Published online: 18 August 2021

Abstract
It has been proven that collaboration between authors leads to a positive influence on research. This paper aims to analyse the complex structure of the co-authorship network among researchers of the Italian Institute of Technology. In this paper, we examine two different co-authorship networks created starting from the data of the papers published by the Italian Institute of Technology during the period 2006–2019. We apply the main Social Network Analysis techniques to describe the relational structure of the group of researchers and its evolution over time. The structure and characteristics of the networks are analysed both at macro and micro levels, and an attempt is made to identify a possible relationship between the position of researchers in the graphs and their scientific productivity and quality.

Keywords Social network analysis · Scientific collaboration · Research performance · Co-authorship network

Introduction

Nowadays research centers must be constantly monitored to evaluate scientific production, scientific collaboration between researchers, and the degree of openness compared to other national and international research centers. In recent years there has been an increase in scientific collaboration between researchers (Drenth, 1998; Glänzel, 2001; Weeks et al., 2004; Levsky et al., 2007; O'Brien, 2012; Henriksen, 2016; Kuld & O'Hagan, 2018). Cooperation allows us to overcome the growing complexity and specialization of scientific research. Joint work can involve actors working in the same disciplinary field, but also,

✉ Sara Preti
sara.preti@edu.unige.it; sara.preti@iit.it

Enrico di Bella
enrico.dibella@unige.it

Luca Gandullia
luca.gandullia@unige.it

[1] Department of Economics, University of Genoa, Via Vivaldi 5, 16126 Genoa, Italy

[2] Italian Institute of Technology (IIT), Via Morego 30, 16163 Genoa, Italy

and above all, actors from different sectors. This allows each actor involved to influence ideas, share different expertise, and acquire new skills, leading to innovations more easily. An additional benefit of cooperation among scientists is given by the increase of their scientific productivity because synergy could improve the quantity and quality of their research output (De Stefano et al., 2010). Besides, the collaboration between actors may reduce research costs, in particular in experimental research, because it is possible to share the expenses for scientific instrumentation (Huang, 2014). Collaboration could be analysed through co-authorship in scientific publications, as shown by several studies, such as that of Glänzel and Schubert (2004). This paper also uses co-authorship links as a measure of scientific collaboration and, in particular, it analyses the collaborative research network between researchers of one of the main Italian research centers, the Italian Institute of Technology (IIT). IIT is a Foundation established in 2003,[1] publicly financed that conducts scientific research in the public interest by promoting excellence in basic and applied research. The study considers researchers employed in the Central Research Laboratories based in Genoa, where the scientific headquarter of the Institute is located. The total staff of the IIT is comprised of 1762 people and the scientific field is represented approximately 80% of the overall staff. In 2018, State financing received through the Ministry of Economy and Finance amounted to approximately EUR 91 millions (deducted spending review amount), 80% of which were allocated to scientific and technological activities. Moreover, the Institute obtained other monetary resources through participation in competitive and commercial projects, for more than 46.7 million euros. IIT Research Lines are divided into four domains: *Robotics*, *Nanomaterials*, *LifeTech* and *Computational Sciences*. The lines belonging to the first domain are dedicated to the development of new robotic platforms, both hardware, and software. The Research Domain of Nanomaterials, instead, includes activities for the creation of sustainable/biodegradable materials, nanocomposites, 2D materials, nanofabrication technologies and nanodevices, and new colloidal chemistry approaches. The Research Domain of LifeTech is devoted to developing advanced genetic, molecular, electrophysiological, computational, imaging, and perturbation tools for dissecting the microscopic neural processes underlying brain function. Finally, the Computational Sciences domain is focused on massive simulations of physical systems, repeated numerous times to generate robust statistics and data mining of vast datasets to identify unexpected patterns. In addition to the Research Lines, the Institute presents 8 Facilities. Facilities can interact with all Research Lines and provide scientific and technical assistance for basic research and the resulting applications. Each facility offers researchers and students an extensive range of services from essential routine to more advanced technology and consulting services.[2] There are currently 90 independent Research Lines and Facilities at IIT developing the 4 Domains in a fully cross-disciplinary research environment. In this paper, we propose an analysis of the scientific collaboration network between authors belonging to the IIT. The analysis of these networks is very relevant to understand the evolution of the network over time. It is important to understand whether the network of collaborations grows linearly or exponentially. In this work, we describe the data collected via Scopus and those provided directly by the research center. Then, we present the main techniques of the Social Network Analysis that have been used to examine the structure of the graph of co-authorship and its evolution over the years. After we examine the scientific

[1] Italian Institute of Technology was established by Legislative Decree 269/03, converted by Law no. 326/2003 (Article 4 of the Statute).

[2] https://iit.it/research/facilities.

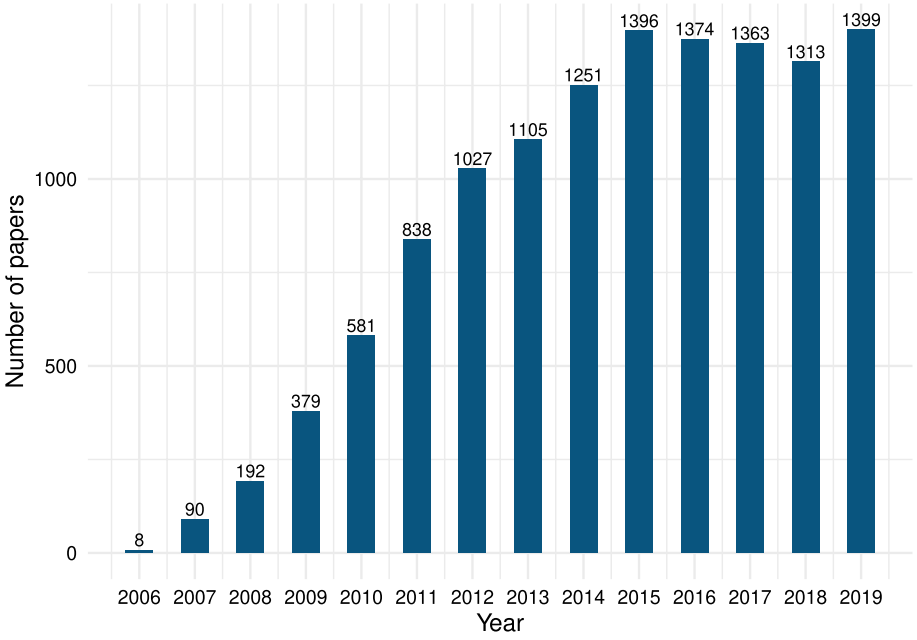


Fig. 1 Evolution of the number of papers published with the affiliation of the Italian Institute of Technology

collaboration network starting from the data of the papers published by the IIT, for this reason, the reference period is between 2006 and 2019. We analyse the structure and the characteristics of the graph (macro perspective), and the most central actors in the network (micro perspective); moreover, we verify the existence of a relationship between centrality measures of each author and their research performance, based on the number of documents published and the number of citations received by each of them. Finally, we discuss the results and draw our conclusions.

Materials and methods

Data sources

The study considers only the papers affiliated with the Italian Institute of Technology to analyse the scientific productivity and the research impact of the organisation. To this aim, we downloaded the data relating to all the documents published by the Italian Institute of Technology for the years 2006–2019 from the bibliographic database of Scopus.[3] There are 12,363 documents in all. The phase of data processing removed 47 papers[4] and the

[3] The Scopus database was produced by the publishing house Elsevier and it was introduced on the market in 2004. It is available at the following link: https://www.scopus.com/search/form.uri?display=basic.

[4] We removed duplicate papers because Scopus includes both the final versions of the articles and those containing errors and we excluded the papers with a large number of authors for which affiliation information was not available.

final number of documents included in the analysis is 12,316. Figure 1 shows the trend of the Institute's number of publications in the period considered. The number of papers increased significantly until 2015 when it reached a value of 1396; in the following 4 years the growth stopped but the number of publications remained stable with a value of just over 1300. We also used data provided directly by the Institute to implement a qualitative check of the attributes concerning personnel.

The data we processed are relational data, so we can detect patterns and interactions between researchers. In particular, we analysed the overall status of collaborative research among authors, we determine the authors at the core of the cooperative research network, we identify authors that have a strong cooperative relationship, and we verify the existence of a relationship between centrality measures of each author and their research performance, based on the number of documents published and the number of citations received by each of them.

Methods

Our research focuses on co-authorship links among scientists of the Italian Institute of Technology. To study these relations we applied the techniques of the Social Network Analysis (SNA), which is a method of structural analysis applied in many research fields and developed since the 1930s with the work of Jacob Moreno (Freeman, 2004). In recent years, the use of this method has seen exponential growth, also thanks to the increase in data collection and the increase in computational capacity.

Network science allows the measurement and visualization of collaboration networks between individuals through a graph. A graph is represented by a set of V vertices, also called nodes, connected by E edges, also called arcs, links, or ties. In mathematical terms, a network or a graph is defined by the object $G = (V, E)$ where V is the set of nodes and E the set of arcs. In the analysed graph the authors are the vertices and the co-production between authors of a given publication are the edges. The network of co-authorship links is undirected since the relationship between authors who wrote some papers together is symmetric. Furthermore, the graph is weighted and the weight of each link is equal to the number of publications two actors have collaborated on together during the period of interest. The weight reflects the degree of collaboration. It is reasonable to consider that certain relationships are stronger than others because generally, an author prefers to collaborate with another one with whom he has already cooperated, rather than looking for a new peer. We use the weight to compute the average strength of the networks, illustrated and commented below. However, to compute other measures, we do not consider the weight, because it is inappropriate in situations where a smaller weight is preferred. For example, the algorithm employed for the computation of the diameter in a weighted network favors smaller edge weight since the weight of an edge is considered as the cost to travel between two vertices. This implies that the algorithm will select the paths where two authors have fewer collaborations, but in reality, it should be shorter and easier to reach a frequent collaborator rather than a new author (Bian et al., 2014). Another edge attribute is given by the date of publication of each paper. Social Network Analysis includes some measures relevant to the identification of the central actors and the most cohesive relationships within the network. The statistics applied in the current study are described below. Each metric was computed using `igraph`, a network analysis package developed in `R`. For research purposes, centrality measures, density, degree distribution, degree centralisation, clustering coefficient, connectivity, and diameter of the network were evaluated.

Social network metrics

Diameter The first index analysed refers to network connectivity. A graph is said to be *connected* if for each pair of nodes i, j there is a path, i.e. a set of consecutive arcs connecting i and j. In the case of a non-connected network, it is possible to split the graph into components, i.e. in connected sub-graphs. Network components can be made up of a single node, in this case, nodes are called *isolated*, or of a group of connected nodes. *SNA* theory defines a *giant component* a component in which the cardinality of all its nodes is of the order of N, the total number of the nodes in the network. In the studies of complex networks it is generally necessary to work with connected graphs. For this reason, the giant component of the network is typically extracted and analyses are carried out on it. A measure related to network connectivity is the graph diameter. The *diameter* is the longest geodesic distance between any pair of vertices or the distance between the two furthest vertices. Formally, the diameter is shown in Eq. 1.

$$\text{diameter} = \max_{i,j} d(i,j) = \max_i e_i \tag{1}$$

Degree centrality The degree, denoted by k_i corresponds to the number of incident arcs of the ith node. For the *degree centrality* the vertices with a higher degree are more central and, in contrast, the vertices with a lower degree are located in the peripheries of the network. This measure is based on the idea that the central nodes must be the most active since they are mostly connected to the other nodes in the network (Wasserman and Faust, 1994). The general formula of the centrality index based on the degree is shown in Eq. 2:

$$C_D(i) = k_i \tag{2}$$

to compare the degree in different graph, it is possible to standardize the index by dividing by the maximum possible value $N - 1$:

$$C_D(i) = \frac{k_i}{N-1} \tag{3}$$

where the subscript D is for "degree", k_i is the node degree and $N - 1$ is the maximum number of links that the node can have. After converting the multigraph network into a weighted graph, the strength of the vertices was also computed. The strength of a generic node i is equal to the sum of the weights of the arcs connected to it.

Closeness centrality The *closeness centrality* is based on the concept of geodesic distance, which is the shortest path that connects the nodes. According to this index, the central actor is the one who can reach all other nodes most quickly. Hence, the more central a node is, the lower its total distance to all other nodes. Sabidussi (1966) defined the closeness centrality as the reciprocal of the sum of the distances from node i to all other nodes:

$$C_C(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1} \tag{4}$$

in this case, the subscript C stands for "closeness". Again, for comparison purpose, we can standardize the closeness by dividing by the maximum possible value $(N-1)^{-1}$.

Betweenness centrality For the *betweennes centrality* the importance of an actor depends on the probability that he has to be along all the possible shortest paths that

connect the actors of the network. In other terms, an actor is central if he undertakes a mediation role. The betweenness for a node i is computed as follow:

$$C_B(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \tag{5}$$

where g_{jk} is the total number of geodesic paths that link node j to node k, while $g_{jk}(i)$ is the number of geodesics linking the two nodes that involve node i. The betweenness index can be standardized by dividing by the maximum possible value. Unlike closeness centrality, this measure can be applied even if the network is not connected.

Density The *density index*, denoted by Δ, is a graph property that describes the general level of linkage among the actors in a network (Abbasi et al., 2011). The index is defined as the ratio between the number of edges and the maximum possible number of edges in the graph, as shown in Eq. 6.

$$\Delta = \frac{L}{N(N-1)/2} = \frac{2L}{N(N-1)} \tag{6}$$

The Δ varies from zero (there are no links present) to one (all possible links are present). Density values tend to reach zero in sparse networks, and close to one in tightly connected networks (Wu et al., 2019).

Degree distribution Degree distribution, denoted by p_k, represents the probability that a randomly selected node in the network has degree k. This probability is given by the following formula:

$$p_k = \frac{N_k}{N} \tag{7}$$

where N_k is the number of nodes in the graph having degree equal to k and N is the total number of nodes in the graph. This property of the network becomes one of the most important for several reasons. Firstly, to calculate most of the properties of a network it is necessary to know the degree distribution. Secondly, the determination of certain phenomena, such as the robustness of the network, depending on it. Real networks have a highly heterogeneous degree distribution, with low degree nodes and very high degree nodes. Networks with these properties are called *scale-free* since they are not characterized by a specific scale size.

Degree centralisation The *degree centralisation* is an index that detects the dispersion around the average degree value, or in other words, quantifies the range or variability of the degree of each individual actor (Wasserman & Faust, 1994). Therefore, the degree centralisation measures the overall consistency of a network, indicating how much relations are organized around particular star actors. The index is bounded between zero and one; high centralisation values normally indicate the accumulation of relations and ties around a few network actors. Freeman (1979) defined a general mathematical formula for the calculation of centralisation indices, that derive from a transformation of the centrality indices (shown above). For non-direct graphs the degree centralisation index is expressed as follows:

$$C_D = \frac{\sum_{i=1}^{N}(k^* - k_i)}{(N-1)(N-2)} \tag{8}$$

where k_i is the degree of i-th node; N is the total number of nodes, and k^* is the largest observed degree value.

Clustering coefficient The *clustering coefficient* of a node, defined by Watts and Strogatz (1998), is an index that considers the number of triads in which the corresponding node takes place, compared to the number of potential triads involving that node. In other terms, it measures the transitivity of the node. The index ranges between zero and one. A value close to zero indicates poor transitivity, while a value close to one indicates a high percentage of triangles in the graph.

$$C_i = \frac{2t_i}{k_i(k_i - 1)} \tag{9}$$

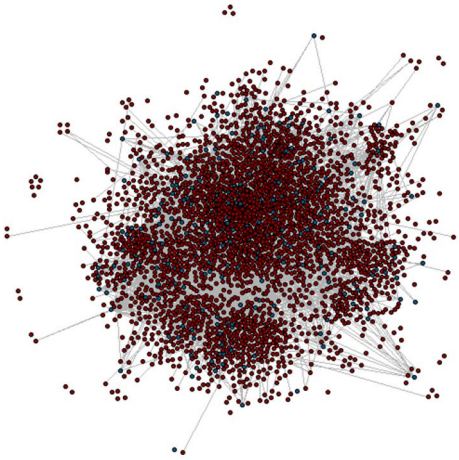
where t_i is the number of links between the set of nodes connected to i. From the nodes clustering coefficient, it is possible to obtain the graph clustering coefficient, whose equation is as follows:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{10}$$

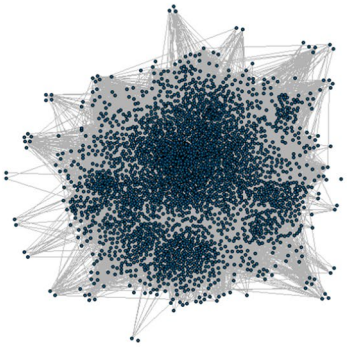
Results

The results presented below, refer to the scientific collaboration network of the Italian Institute of Technology created starting from all the documents published by the research center during the period 2006–2019. The IIT was established in 2003 but the first papers available on Scopus date back to 2006. From the information on the papers, two different Co-Authorship Networks were created. Our data contained a binary variable that takes value 0 if the author was not affiliated with the Italian Institute of Technology, and 1 otherwise. For the first network, we considered only links in which at least one author was affiliated with the Italian Institute of Technology; from this point forward, this graph will be called *heterogeneous network*. The second graph instead was made up only of authors affiliated with the scientific research center and will be called *homogeneous network*. We decided not to consider the ties between authors not affiliated with the IIT because the paper aims to analyse the scientific collaboration network of the IIT. If we had also considered the co-authors of the authors not affiliated with the IIT we would have analysed almost all the papers available on Scopus (according to the Small World theory).[5] We analysed the scientific collaboration networks in 2006 and then by tracking the change of the cumulative network structure in 1-year intervals (e.g. the network from 2006 to 2007, from 2006 to 2008, and so on). Figure 2 shows the cumulative Co-Authorship Networks for the years 2006–2019. The vertices of the graphs are characterized by different colors, the blue ones represent researchers of the IIT, while the red ones researchers with other affiliation. The graphs were created by using the software `R` and the package `igraph`, by applying

[5] The first relevant empirical study of the small-world phenomenon was undertaken by the social psychologist (Easley & Kleinberg, 2010). Small world theory is based on the idea that two individuals will be connected through a series of intermediaries by a small number of hops or steps.



(a) Heterogeneous Network (2006-2019)



(b) Homogeneous Network (2006-2019)

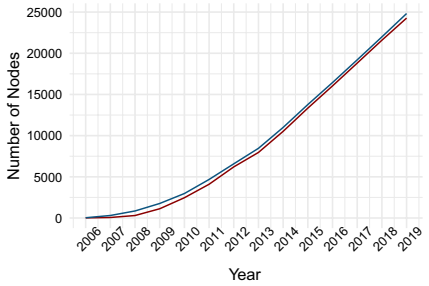
Fig. 2 Cumulative Co-Authorship Networks depicting collaborations among researchers at the Italian Institute of Technology. The blue nodes represent authors of the IIT, while the red ones other co-authors

a force-directed algorithm, Graphopt, created by Michael Schmuhl.[6] The algorithm uses physical analogies to define attracting and repelling forces among the set of nodes and then the physical system is simulated until it reaches an equilibrium.

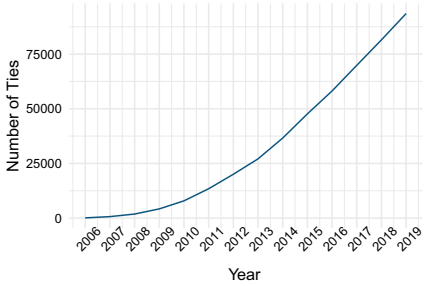
A summary of the statistics of the heterogeneous networks is given in Table 1, and Fig. 3 represents the line plot of each statistic.

The scientific collaboration network has grown over time, in fact from 2006 to 2019, the number of authors has gone from 33 to about 25,000; starting from 2014, the number of nodes increases by more than 2500 units per year. Figure 3a compares the size of the giant component with the whole network. The blue curve denotes the growth of the co-authorship network, while the red one denotes the growth of its largest connected subgraph. These two curves exhibit a similar growth mode and at the end of the reference period, 98% of all the nodes belong to the giant component. The total number of links has also increased over time, moving on from 97 to 93,533, respectively in 2006 and 2019; thus, co-authorship

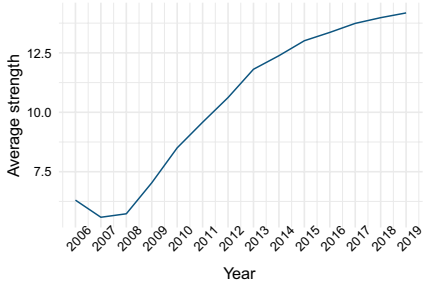
[6] See http://www.schmuhl.org/graphopt/.



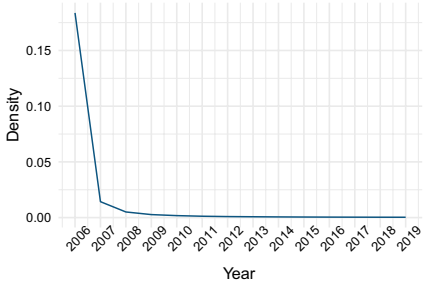
(a) Number of Nodes in the whole network and in the giant component.



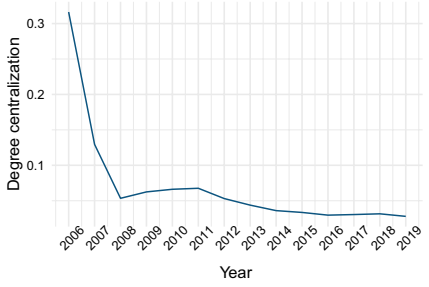
(b) Number of Ties.



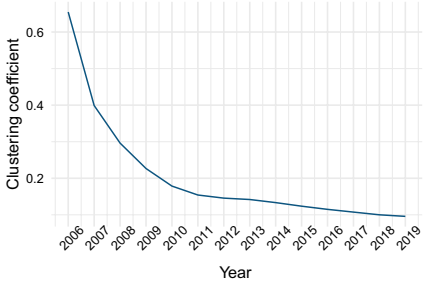
(c) Average Strength values.



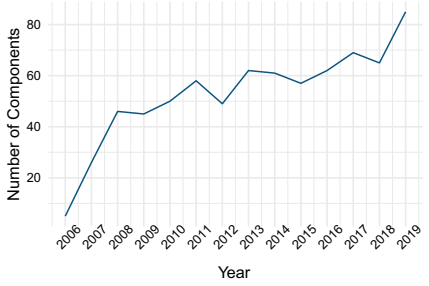
(d) Density values.



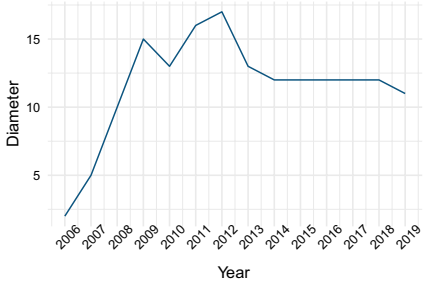
(e) Degree Centralisation values.



(f) Clustering Coefficient values.

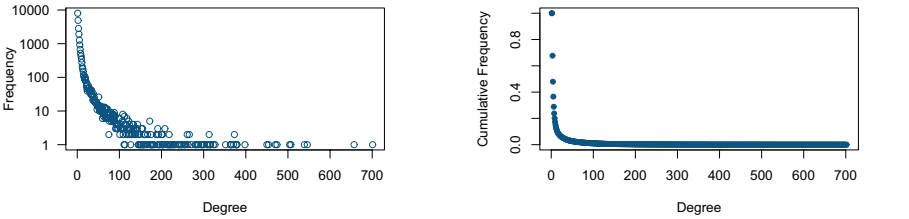


(g) Number of Components.



(h) Diameter values.

Fig. 3 Line plots of the Heterogeneous Cumulative Co-Authorship network measures



(a) Degree distribution on a logarithmic scale

(b) Cumulative degree distribution

Fig. 4 Degree distribution of the Heterogeneous cumulative Co-authorship Network

has become a common practice in the Institute. From 2006 to 2007 the number of new ties was 574, the value has increased each year and from 2008 to 2019 the number of new ties reached the value of 12,030. The increase in ties just described does not take into account the consolidation of existing connections, but only the presence of new links. However, even considering the weight, given by the number of papers that two actors have coauthored, the growth is confirmed by the trend of the average strength: the average number of scientific collaborations per author has increased from 6.30 in 2006 to 14.18 in 2019. We also computed the degree distribution for the final cumulative co-authorship network (2006–2019) because it contains important information about the nature of the network. Real networks have a highly heterogeneous degree distribution, with many nodes with only a few links (low degree) and few hubs with a large number of links (high degree), as in our case study (Fig. 4). Networks with these properties are called *scale-free* since they are not characterized by a specific scale size.

As mentioned in Sect. 2.2, the increase in the number of nodes determines an increase in the number of the maximum possible distinct links. If we look at the density values, they decrease over time from 18% in 2006 to 0.03% in 2019. However, it is worth noting that the links between the authors who are not affiliated with the Italian Institute of Technology have not been taken into consideration, therefore with the increase in the number of external collaborators, the density can only decrease; the same is valid for the clustering coefficient. For this reason, it is worth analyzing these two metrics for the network in which all the nodes are affiliated with the Italian Institute of Technology. Degree centralisation is not constant over time but remains relatively low, except for the first 2 years in which its values are equal respectively to 32% and 13%. In the following years, the values fluctuate and then begin to decrease starting from 2011, going from 7 to 3%. Degree centralisation indicates that the whole network is not organized around a single central node but it has many influential authors. This means that the co-authorship network of the Italian Institute of Technology is not based on a single author, but its research impact is determined by the sum of the individual authors; therefore even eliminating the most important nodes from the graph, the structure would remain the same.[7] The clustering coefficient values decrease with time, the probability that two of a researcher's collaborators have themselves

[7] We have implemented a robustness check, eliminating a random sample equal to 10% of the total network nodes. We calculated the degree centralisation metric on the new network, obtaining values identical to those of the complete network. We also tried to eliminate the ten authors with the highest degree one at a time and we saw that the network structure remains stable.

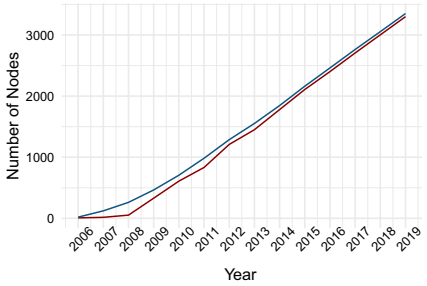
collaborated is equal to 0.65 in 2006, and only to 0.10 in 2019. The last values reported in Table 1 refer to graph connectivity. The values of diameter increase with time from 2 (2006) to 17 (2006–2012) but in subsequent networks the values of the diameter decrease, and this implies that the efficiency of information flow on the network is increased. The Co-Authorship Network is not a single connected graph; the number of components in which the network is divided is not regular in the period considered, but in general, it increases, going from 5 in 2006 to 85 in 2019. According to the theory, as the number of total links increases, there comes a point in which a giant component forms (Sabidussi, 1966). Even the trend of the percentage of nodes that make up the largest component is not constant with time, but it has grown substantially and in recent years the 98% of the authors are included in the giant component. So 98% of the nodes are connected by a direct or non-direct collaboration network.

To analyse how much the researchers of the Italian Institute of Technology tend to collaborate with each other, we also analysed the homogeneous network constituted exclusively by the authors that have been affiliated with the scientific research center. Table 2 shows the main statistics for the homogeneous network, and Fig. 5 represents the line plot of each statistic.

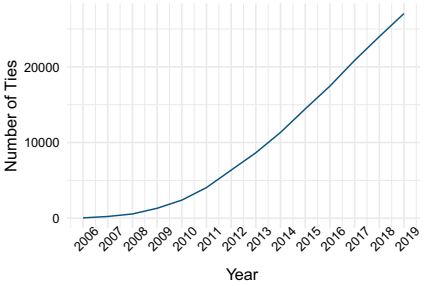
As for the heterogeneous network, the size of the graph has increased gradually, there are 3351 authors in 2019 and only 19 in 2006. The number of links has grown even more rapidly, going from 39 to 27,051, respectively in 2006 and 2019. The speed with which collaborations between researchers affiliated to the Institute increases is the same as that observed in the previous network, as shown in Fig. 6, where the red line corresponds to the ties of the heterogeneous network, while the blue one to those of the homogeneous graph. Comparing the number of links between the two networks, it is clear that the percentage of collaborators affiliated with IIT represents about 30% of the total collaborations. So, it is evident that the growth of internal collaborations has gone together with that of external collaborations.

The values of other statistics are relatively higher than those reported in Table 1. Compared to the previous network, cooperation between researchers of the Italian Institute of Technology is stronger, the average number of scientific collaboration per author reached values above 40 since 2017. Comparing strength and degree values in the two networks, it is possible to note that the frequency with which two authors collaborate is higher for internal collaborations (co-authorship ties between members of the institute), this means that although the number of internal and external collaborators grows equally in the two networks, the average number of times two authors cooperate is higher among members of the Italian Institute of Technology. Figure 7 shows that even the final homogeneous cumulative co-authorship network (2006–2019) is *scale-free*.

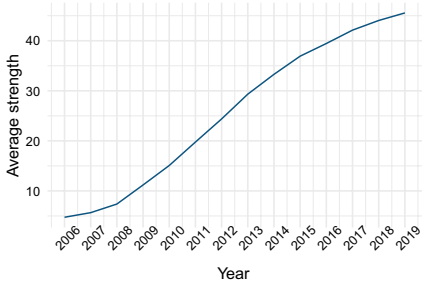
Density decreases with increasing network size, values going from 23% in 2006 to 0.48% in 2019, but slowly compared to the previous network. It means that the increase of the existing links is less than that of the number of possible connections between the nodes. This result confirms the theory according to which the levels of density tend to decrease in larger graphs (Kim & Diesner, 2016). Furthermore, real complex networks typically present a low level of cohesion. Also, the observed values for the clustering coefficient decrease with time, from 0.91 in 2006 to 0.25 in 2019; however, a propensity of authors to form cliques equal to 25% shows a fairly high local cohesion. Degree centralisation fluctuates in the first years but starting from 2011 values decrease until they reach 7%. This metric indicates that the structure of the graph remains sparse and therefore it is not based on particular focal nodes, but on multiple authors, as for the previous network. Finally, the number of components in which the new graph is divided is significantly lower, and after



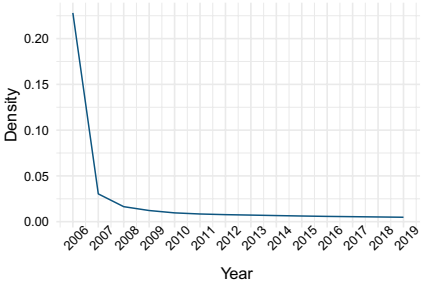
(a) Number of Nodes in the whole network and in the giant component.



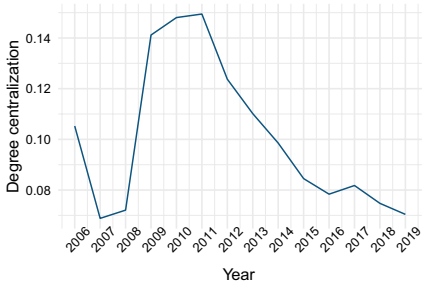
(b) Number of Ties.



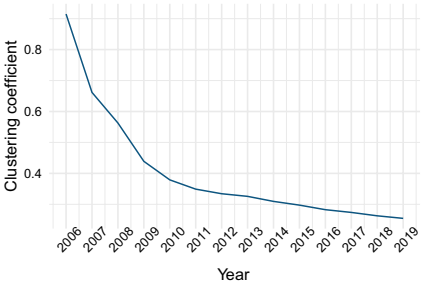
(c) Average Strength values.



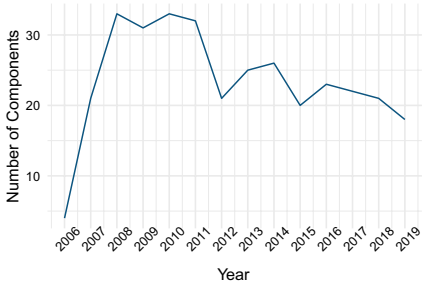
(d) Density values.



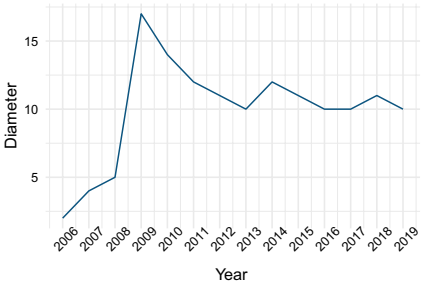
(e) Degree Centralisation values.



(f) Clustering Coefficient values.



(g) Number of Components.



(h) Diameter values.

Fig. 5 Line plots of the Heterogeneous Cumulative Co-Authorship network measures

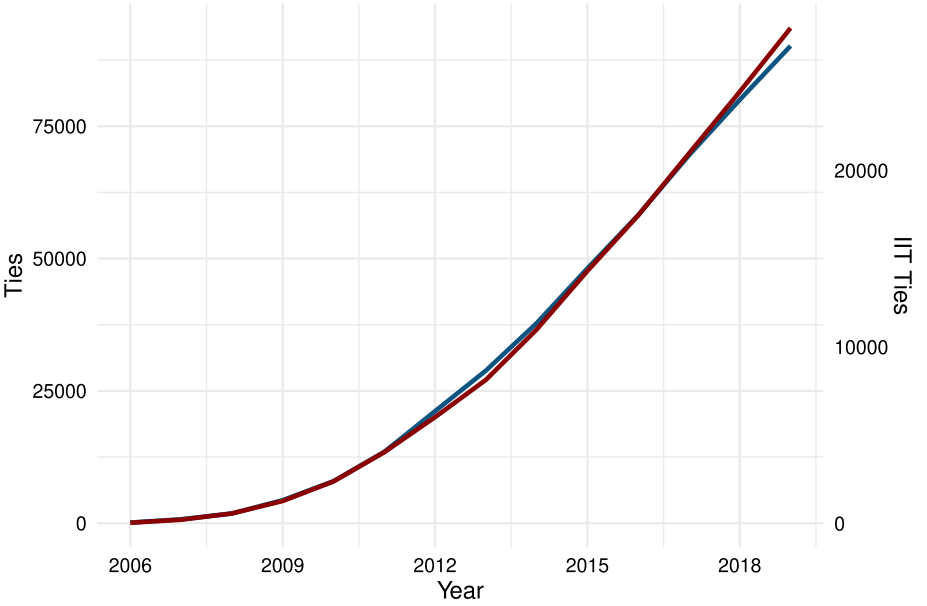
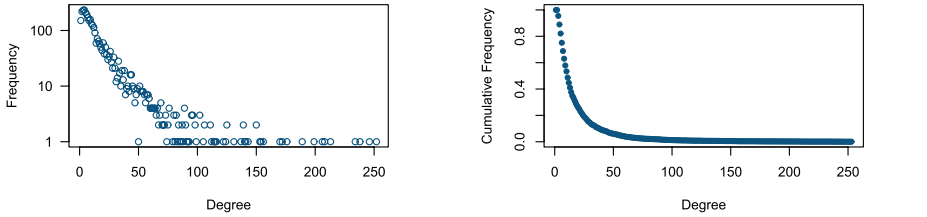


Fig. 6 Growth curve of internal and external collaborations of the IIT



(a) Degree distribution on a logarithmic scale

(b) Cumulative degree distribution

Fig. 7 Degree distribution of the Homogeneous cumulative Co-authorship Network

an initial period of fluctuations, starting from 2014 the number of components decreases, reaching a value of 18 in 2019, hence in this network researchers are well connected. As for the heterogeneous network, the giant component contains almost all the nodes (98.5%).

Up to now, the scientific collaboration networks have been studied at a macro level, since we have examined the global structure and characteristics of graphs. Now, we will study the networks from a micro-perspective, through different centrality measures that explain the behavior of researchers based on their relationships with other researchers (Bordons et al., 2015), also identifying the most influential nodes in the scientific collaboration network. These metrics are computed for individual authors on the giant component of the last two networks. The frequency distributions of the centrality measures for the years 2007, 2011, 2015, and 2019 are shown in Figs. 10 and 11. The first one relates to the heterogeneous network, in which a link is made up of at least one author belonging to the Italian Institute of Technology, while the second figure relates to the homogeneous network that includes only authors affiliated to the research organisation. For both networks,

Table 1 Statistical properties of the heterogeneous cumulative Co-Authorship Networks

Period	No. of Nodes	No. of Ties	Density	Average strength	Degree centralisation (%)	Clustering coefficient	No. of components	Diameter
2006	33	97	0.18371	6.30	31.63	0.65	5	2
2006–2007	307	671	0.01429	5.58	12.95	0.40	26	5
2006–2008	856	1850	0.00506	5.73	5.34	0.30	46	10
2006–2009	1766	4203	0.00270	7.03	6.25	0.23	45	15
2006–2010	2974	7891	0.00178	8.50	6.62	0.18	50	13
2006–2011	4681	13,416	0.00122	9.58	6.76	0.15	58	16
2006–2012	6578	20,044	0.00093	10.62	5.31	0.15	49	17
2006–2013	8461	27,081	0.00076	11.81	4.38	0.14	62	13
2006–2014	10,999	36,594	0.00061	12.38	3.60	0.13	61	12
2006–2015	13,775	47,656	0.00050	13.01	3.35	0.12	57	12
2006–2016	16,445	58,140	0.00043	13.36	2.97	0.11	62	12
2006–2017	19,201	69,853	0.00038	13.74	3.05	0.11	69	12
2006–2018	21,988	81,503	0.00034	13.98	3.15	0.10	65	12
2006–2019	24,822	93,533	0.00030	14.18	2.79	0.10	85	11

the frequency distributions of degree and betweenness centrality follow a gamma distribution, most authors have low centrality values, while a few authors are characterized by high values. This type of distribution for degree centrality indicates that these co-authorship networks have a scale-free character (Barabási & Albert, 1999). Conversely, the values of closeness centrality follow the Gaussian function. Tables 3 and 4 report the position of the top-ranked authors of the cumulative networks 2006–2019, according to the centrality measures.

Except for a few authors, the names in both tables are the same and all the authors are affiliated with the Italian Institute of Technology. Note that the author who appears in the highest position in all rankings is almost always the same; hence he is the central actor of the analysed co-authorship networks. In the heterogeneous network, the degree centrality for this author indicates that he collaborated with 504 co-authors, but in the homogeneous network he occupies the first position of the ranking with 252 collaborations; this means that the author preferred internal collaborators than external ones. For a scientific community, the greater is the number of collaborators, the higher is the level of degrees of intra-community influence (Yan & Ding, 2009). The authors that are in these tables play a pivotal role in the network because they are characterized by high communication activity and

Table 2 Statistical properties of the homogeneous cumulative Co-Authorship Networks

Period	No. of nodes	No. of ties	Density	Average strength	Degree centralisation (%)	Clustering coefficient	No. of components	Diameter
2006	19	39	0.22807	4.74	10.53	0.91	4	2
2006–2007	122	224	0.03035	5.66	6.88	0.66	21	4
2006–2008	261	556	0.01639	7.38	7.21	0.56	33	5
2006–2009	464	1308	0.01218	11.19	14.12	0.44	31	17
2006–2010	705	2385	0.00961	15.08	14.81	0.38	33	14
2006–2011	983	4047	0.00838	19.76	14.95	0.35	32	12
2006–2012	1287	6351	0.00767	24.38	12.37	0.33	21	11
2006–2013	1552	8649	0.00719	29.33	11.02	0.33	25	10
2006–2014	1845	11,339	0.00667	33.29	9.85	0.31	26	12
2006–2015	2162	14,443	0.00618	36.92	8.45	0.30	20	11
2006–2016	2461	17,463	0.00577	39.45	7.84	0.28	23	10
2006–2017	2763	20,854	0.00547	42.12	8.18	0.27	22	10
2006–2018	3055	23,998	0.00514	44.05	7.48	0.26	21	11
2006–2019	3351	27,051	0.00482	45.56	7.04	0.25	18	10

Table 3 List of the top ten authors based on centrality measures

ID	Degree	ID	Closeness	ID	Betweenness
57214432011	701	36038586400	0.355	36038586400	0.091
57208455894	657	7004555976	0.348	57214432011	0.066
7005221987	546	57208455894	0.340	7004555976	0.061
7004555976	539	57214432011	0.339	6603680863	0.045
7004723880	507	57203054720	0.338	7101898098	0.045
36038586400	504	7005225151	0.336	57208455894	0.045
57211227572	474	56999237200	0.336	57201554399	0.039
57201554399	470	7006790590	0.335	57211227572	0.038
7101898098	454	7005221987	0.334	6507663930	0.036
23100242200	450	24502683600	0.332	7006790590	0.036

The authors belong to the heterogeneous network

Table 4 List of the top ten authors based on centrality measures

ID	Degree	ID	Closeness	ID	Betweenness
36038586400	252	36038586400	0.416	36038586400	0.127
57211227572	246	7004555976	0.397	57211227572	0.086
57208455894	238	7005225151	0.393	57201554399	0.059
7005221987	234	57208455894	0.385	7004555976	0.054
7005225151	213	57203054720	0.385	57214432011	0.049
57214432011	208	57214432011	0.383	6603680863	0.048
7004266208	206	55894681500	0.380	7006595496	0.048
7004555976	199	8545961300	0.379	7101898098	0.045
24502683600	189	7005221987	0.379	7005225151	0.041
57201554399	176	7005551012	0.378	57203054720	0.038

The authors belong to the homogeneous network

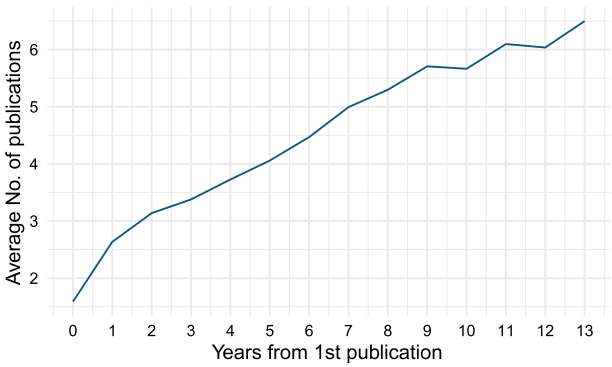


Fig. 8 Average number of publications from the first year of publication

popularity. Researchers with the highest closeness centrality values had the highest opportunity to exchange and propagate information to establish a cooperative relationship with other authors within the network. The closeness rate explains how long information flow takes from a given node to others (Yan & Ding, 2009). Finally, the authors with the highest betweenness centrality values play a crucial role to connect different groups as a broker, hence they possessed a large number of research sources and they can control collaborative relationships (Wu et al., 2019). It is worth noticing that all the top-ranked authors in Table 4 have published for several years for the Italian Institute of Technology. Each of these authors has been publishing for the IIT for more than 8 years, hence they consolidated their role within the organisation over the years. This result leads us to analyse the relationship between the number of publications and the seniority of the authors, measured as the number of years since the first publication for the research organisation. Figure 8 and Table 5 show the average number of publications per seniority.

On average authors published 1.59 papers in their first year, the average increases by 65% in their second year, and continues to grow in the subsequent years. This confirms that productivity grows with seniority. Authors with the fewest years from first publication are also included in subsequent years, but not vice versa. All authors contribute to the first-year average, but only eight authors have published for 14 years for the Institute. For this reason, we have decided to repeat the analysis considering only the 501 authors who have

Table 5 Average number of publications from the first year of publication

Years from 1st publication	Average No. of publications
0	1.59
1	2.64
2	3.14
3	3.38
4	3.73
5	4.06
6	4.47
7	4.99
8	5.30
9	5.71
10	5.66
11	6.10
12	6.04
13	6.50

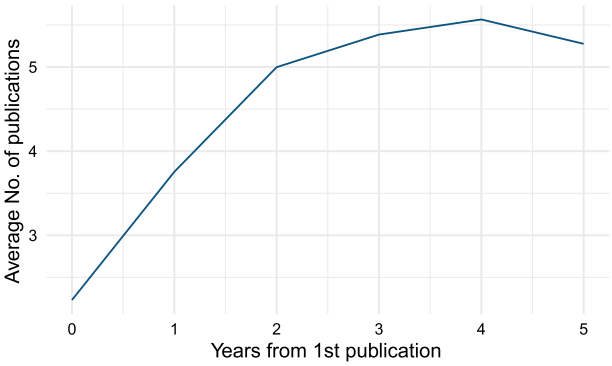


Fig. 9 Average number of publications from the first year of publication

published for at least 6 years so that the average is always calculated on the same authors. The results are shown in Fig. 9. Even though the average values are higher than the previous ones, the growth rate is the same. In fact, in the first year, the average is 2.23 and in the second year, it grows by 68%.

According to the literature, collaborations have a positive correlation with the number of publications (Lee & Bozeman, 2005; McFadyen & Cannella, 2004) and the publications' impact (He et al., 2009; Wuchty et al., 2007). We try to verify the existence of a relationship between the position of authors in the cumulative scientific collaboration network of 2006–2019 and their performance based on the number of documents published and the number of citations received from each author. In particular, we compare Table 3 with Table 6 that reports the list of the top ten researchers based on the number of citations and publications. From the comparison of the two lists, it is clear that 8/10 of the top-ranked authors for centrality measures are in the top ten rankings for the number of publications, whereas only 4/10 are in the top ranking for the number of citations. The difference between the two tables is because greater scientific productivity is not necessarily

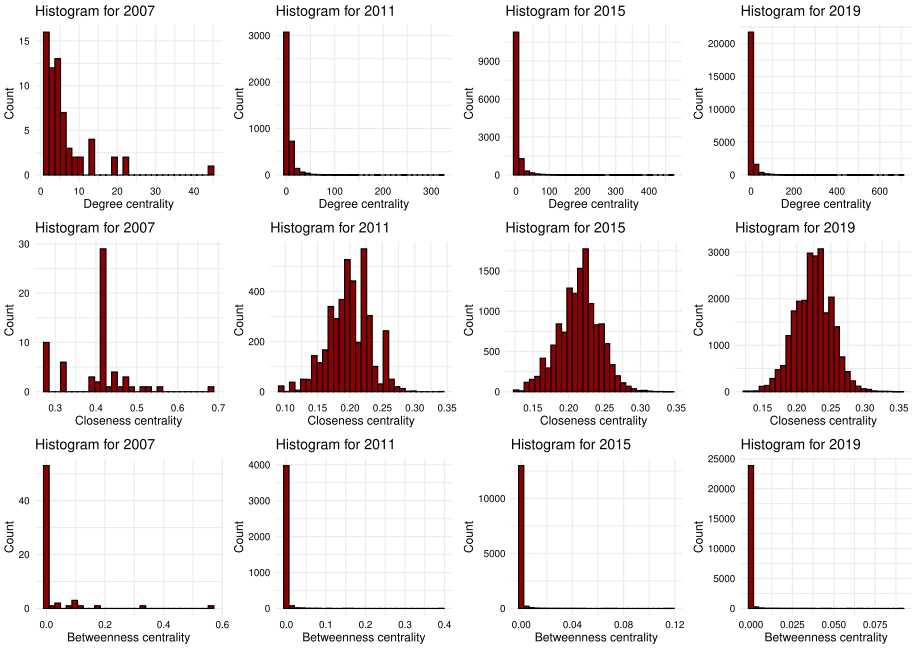


Fig. 10 Frequency distributions of degree, closeness and betweenness centrality of the heterogeneous network

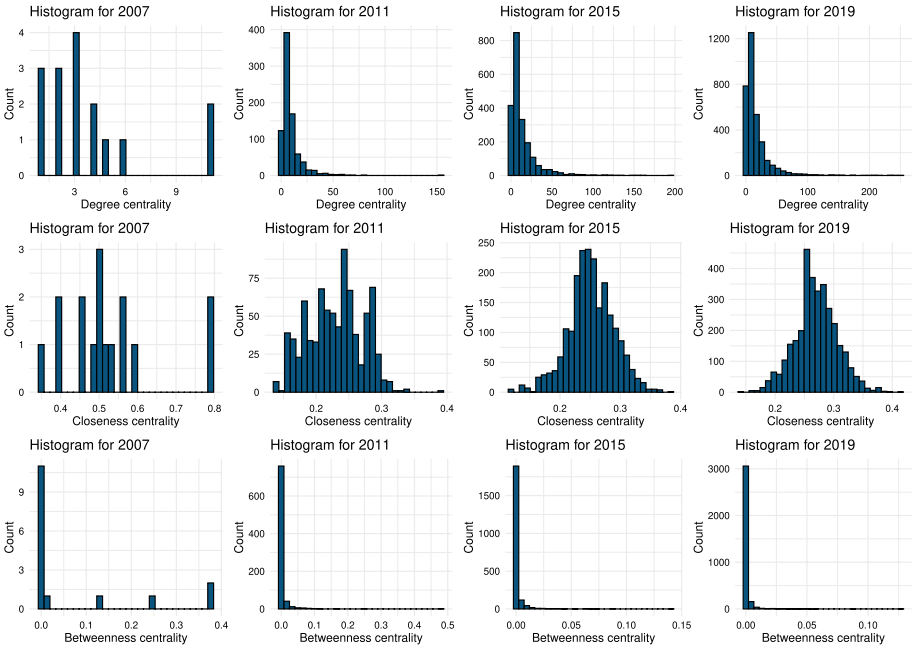


Fig. 11 Frequency distributions of degree, closeness and betweenness centrality of the homogeneous network

Table 6 List of the top ten authors based on citation and publication counts

Ranking	ID	Publication	ID	Citation
1	57211227572	555	24344803200	15,273
2	56275172000	347	7005221987	15,244
3	57201554399	305	8833978100	10,105
4	7005221987	264	57211227572	9986
5	7006595496	256	6508101563	9793
6	6602884280	256	36038586400	9459
7	7004555976	250	23003424200	7353
8	7005225151	233	7006790590	7169
9	36038586400	224	55053316100	7166
10	57203054720	215	35224069200	7144

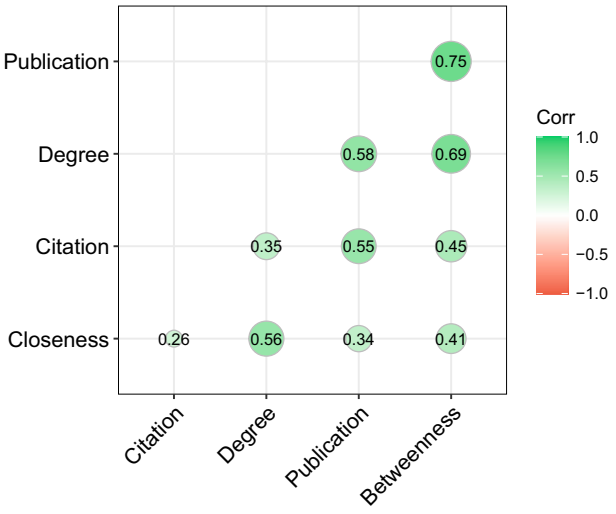


Fig. 12 Spearman's correlations between centrality measures and publication and citation counts

linked to a greater number of collaborators; some authors may have fewer collaborators or tend to cooperate with the same colleagues, they are not cut-points, and accordingly, they are in the periphery of the co-authorship network. As done by previous studies [for example those of Hou et al., (2008), Yan and Ding (2009)], we compute the Spearman's correlation between the number of publications and citations and centrality measures for all the authors in the giant component of the cumulative co-authorship network 2006–2019. Figure 12 shows that there are only positive correlations between metrics. The most correlated are betweenness and degree centrality measures with publication counts, with a correlation value equal respectively to 0.75 and 0.58, whereas the correlation value between closeness centrality and the number of publications is lower and equal to 0.34. Centrality measures present less correlation with citations than with publications, values correspond to 0.45, 0.35, and 0.26 respectively for betweenness, degree, and closeness centrality. This is normal because the co-authorship network is indirectly created from data concerning the documents published by the research organisation. However, although citations, publications, and centralities measure different contents, the positive correlations of centrality

measures with the number of publications and citations demonstrate that centrality measures are able not only to assess the importance of the actors within a network but in some way also their scientific productivity and in a secondary way their quality. In particular, these results suggest that researchers who play a bridging role between different groups of authors (high betweenness) or those with a higher number of collaborators (high degree) are likely to obtain better performance results. It is worth noting that these metrics are used to measure the structure and the characteristics of the organisation as a whole and cannot be used to make a single author assessment. These measures do not take into account the academic career, the degree of seniority, and the position held by each author within the research organisation.

Discussions and conclusions

Compared to the existing literature, this paper does not focus on scientific collaboration across different disciplines but collaboration within an Italian research center, the Italian Institute of Technology. The main objective of this paper is to analyse the structure and evolution of the co-authorship network of the IIT, the actors' position and attributes to identify some interesting features of the research organisation, and the areas where it is most productive. In this paper, we have analysed different co-authorship networks in which the nodes are authors, while the edges represent the connection between two actors that have co-authored a paper. In particular, we have examined two graphs that are created starting from the data of the papers published by the Italian Institute of Technology during the period 2006–2019: the first network is made up of ties in which at least one author is affiliated with the research organisation (external collaboration), while the second network is made up of ties only between the Italian Institute of Technology members (internal collaboration). The structure of the graph is rapidly expanding, and since 2014 growth has started to stabilize. Generally, external collaborations represent 70% of the total collaborations, while internal ones only 30%. The growth of collaboration is constant over the years and external and internal collaborations grow together. However, when considering the relationship between degree and average strength, an author is more likely to collaborate more frequently with another author affiliated with the institute than with an external researcher. The networks are decentralised, so if the Institute loses one of its focal authors, the structure of the networks would not change, and this result was confirmed by the robustness checks. The networks have been analysed not only at a macro-level but also at a micro-level. In particular, we identify the researchers with a more strategic position within the network, through the computation of centrality measures. We found out that the most central actors are those that have belonged to the Institute for the longest time. So we analysed the link between the seniority and the average value of publications and we discovered that the values increase over the years. For this reason, it would be convenient to keep their researchers for a longer period rather than favor the turnover. For a more in-depth analysis of the link between seniority and productivity, the papers published by the researchers should also be analysed before entering the institute. We also try to figure out which practices determine higher performance, studying the relationship of the centrality measures with the number of publications and citations. Our research shows that there is a relationship between the centrality measures of each author and their research performance, but there is not a perfect correlation. This is because centrality measures, publication, and citation counts measure different contents. In particular, the number of publications measures

the author's productivity, the number of citations the quality, and the impact of papers, instead, centralities measure not only the author's productivity and impact, but also his role within the community and his influence in spreading information. However, since there is a positive correlation between the metrics, it is possible to affirm that centrality measures can mainly evaluate the scientific productivity of a researcher and only partially the quality of his work. These metrics can be used jointly to evaluate the characteristics of the network, but it is necessary to note that these cannot be used to assess individual actors. These metrics do not consider determining factors for an objective assessment, such as the degree of seniority, the academic career, and the position held by each author. Finally, it is worth noting that we have found a positive relationship between the centrality measures and the number of citations and publications, but correlation does not imply a cause-effect relationship but the tendency of one variable to change as a function of another.

Our study presents some limitations. First of all, we used only a bibliometric source, the Scopus database, to collect publication data, hence the dataset does not represent the complete production of the IIT (other databases could contain publications in other journals and/or languages) (Abbasi et al., 2011). Secondly, we have used only the citation and publication counts to evaluate research performance, other bibliometric indicators could have been used. In general, it is possible to state that the results obtained can allow the individual researchers to self-evaluate their propensities to cooperate with other scientists (Abramo et al., 2013). Besides, the results can support the implementation of research policies and can have a potential impact on the community organisation. Further studies should consider all the documents published by researchers employed at the IIT, also the papers published before their hiring at the scientific institute; the prestige of a research center is given not only by the scientific results obtained, but also by its human capital, that is, by the people who make up the organisation, so it is necessary to consider the entire career of the researchers. Finally, we would like to make a national or international comparison among different research institutions; it would be interesting to compare the style of collaboration in research organisations of countries with different cultures.

Acknowledgements We thank particularly Alessandro Roscini, Paolo Pelori, and Massimo Pittaluga of the Italian Institute of Technology to help us in obtaining a part of the data used for this study and for their valuable comments.

Funding Open access funding provided by Università degli Studi di Genova within the CRUI-CARE Agreement.

Declarations

Conflict of interest The Italian Institute of Technology funded 50% of the postdoc grant of XX. The authors do not have any competing interests to declare. Data and materials are available from the authors on request. Code is available from the authors on request.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2011.05.007.

Abbasi, A., Hossain, L., Uddin, S., & Rasmussen, K. J. (2011). Evolutionary dynamics of scientific collaboration networks: Multi-levels and cross-time analysis. *Scientometrics, 89,* 687–710.

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). The collaboration behaviors of scientists in Italy: A field level analysis. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2013.01.009.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*. https://doi.org/10.1126/science.286.5439.509.

Bian, J., Xie, M., Topaloglu, U., Hudson, T., Eswaran, H., & Hogan, W. (2014). Social network analysis of biomedical research collaboration networks in a CTSA institution. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2014.01.015.

Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2014.12.001.

De Stefano, D., Vitale, M. P., & Zaccarin, S. (2010). The scientific collaboration network of Italian academic statisticians. In *Paper presented at the 45th scientific meeting of the Italian Statistical Society*, Italy: Padua.

Drenth, J. P. (1998). Multiple authorship: The contribution of senior authors. *Journal of the American Medical Association*. https://doi.org/10.1001/jama.280.3.219.

Easley, D., & Kleinberg, J. (2010). The small-world phenomenon. In *Networks, crowds, and markets: Reasoning about a Highly Connected World* (pp.611–642). Cambridge University Press.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*. https://doi.org/10.1016/0378-8733(78)90021-7.

Freeman, L. C. (2004). *The development of social network analysis—A study in the sociology of science*. Empirical Press.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*. https://doi.org/10.1023/A:1010512628145.

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research* (pp. 257–279). Kluwer Academic Publisher.

He, Z. L., Geng, X. S., & Campbell-Hunt, C. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand University. *Research Policy*. https://doi.org/10.1016/j.respol.2008.11.011.

Henriksen, D. (2016). The rise in co-authorship in the social sciences (1980–2013). *Scientometrics*. https://doi.org/10.1007/s11192-016-1849-x.

Hou, H., Kretschmer, H., & Liu, Z. (2008). The structure of scientific collaboration networks in Scientometrics. *Scientometrics*. https://doi.org/10.1007/s11192-007-1771-3.

Huang, J. S. (2014). Building Research Collaboration Networks—An interpersonal perspective for research capacity building. *Journal of Research Administration, XLV*(2), 89–112.

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.23489.

Kuld, L., & O'Hagan, J. (2018). Rise of multi-authored papers in economics: Demise of the 'lone star' and why? *Scientometrics*. https://doi.org/10.1007/s11192-017-2588-3.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*. https://doi.org/10.1177/0306312705052359.

Levsky, M. E., Rosin, A., Coon, T. P., Enslow, W. L., & Miller, M. A. (2007). A descriptive analysis of authorship within medical journals, 1995–2005. *Southern Medical Journal, 100,* 371–375.

McFadyen, M. A., & Cannella, A. A. J. (2004). Social capital and knowledge creation: Diminishing returns of the number and strength of exchange. *The Academy of Management Journal, 47*(5), 735–746.

O'Brien, T. L. (2012). Change in academic coauthorship, 1953–2003 Science. *Technology & Human Values*. https://doi.org/10.1177/0162243911406744.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*. https://doi.org/10.1007/BF02289527.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small world" networks. *Nature,393,* 440–442.

Weeks, W. B., Wallace, A. E., & Kimberly, B. C. (2004). Changes in authorship patterns in prestigious US medical journals. *Social Science & Medicine*. https://doi.org/10.1016/j.socscimed.2004.02.029.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*. https://doi.org/10.1126/science.1136099.

Wu, W., Xie, Y., Liu, X., Gu, Y., Yuting Zhang, X. T., & Tan, X. (2019). Analysis of scientific collaboration networks among authors, institutions, and countries studying adolescent myopia prevention and control: A review article. *Iranian Journal of Public Health, 48*(4), 621–631.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*. https://doi.org/10.1002/asi.21128.