doi.org/10.26398/asaproc.0058

The dynamics of collaboration between researchers of the Italian Institute of Technology: An ERGM approach

Sara Preti [a], Enrico di Bella [b]
[a] Department of Economics, University of Genoa, Genoa, Italy.
[b] Department of Political and International Sciences, University of Genoa, Genoa, Italy.

1. Introduction, data and descriptive analysis

It has been proven that collaboration between authors has positive effects on research, not only for individuals but also for organizations or countries, with benefits that imply greater productivity and innovation. The scientific collaboration network analyzed in this paper is the *Italian Institute of Technology* (IIT) researchers' network. The IIT is one of the primary Italian research centers, publicly financed, it receives approximately 100 million euros per year from the State; additionally, it secures supplementary resources through its involvement in competitive and commercial projects. Currently, the IIT allocates 80% of its total workforce to four main research domains (RDs), Computational sciences, Life technologies, Nanomaterials, and Robotics.

In this study we analyze the collaborative network of researchers at the IIT, focusing on co-authorship as a formal manifestation of intellectual cooperation in scientific research (Acedo et al., 2006). We examine a network considering the authors' research field and socio-demographic characteristics, including gender, nationality, and seniority, that may affect their propensity to form a collaborative relationship. In a graph, the nodes are the 1,469 researchers employed in the Institute's headquarters in 2020, and an edge between two nodes occurs if they have written a paper together. Furthermore, we study the connections between researchers belonging to different RDs, since interdisciplinary research collaboration became necessary to produce advanced knowledge and accelerate innovation.

Finally, an *exponential random graph model* (ERGM) is used to analyze the dynamics of collaboration between authors based on their social and demographic profiles and the structural features of the network. ERGM models the probability of link formation between nodes within a network by studying the configuration of connections between researchers. Currently, link formation represents one of the most promising areas within knowledge and innovation networks (Resce et al., 2022).

The analysis may assess the relevance of sociodemographic characteristics in making collaborative relationships and provide an opportunity for individual researchers to self-evaluate their propensities to cooperate with other scientists. Furthermore, the findings of this study may support the implementation of research policies that may enhance the community organization.

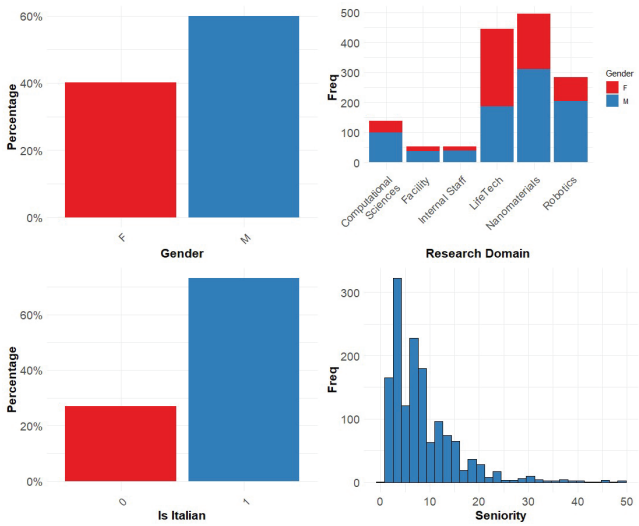
This paper collects information on all the documents published by the IIT researchers, even those published before their hiring at the scientific institute[1]. The starting data source is the full list of researchers with IIT affiliation downloaded using the statistical software *R* through the package called *rscopus* that permitted the extraction of data from the bibliographic database Scopus. This information has been integrated with data provided directly by the Institute, encompassing socio-demographic attributes and variables related to the job position covered by the authors within the organization.

After a cleaning process to address issues related to author name disambiguation[2], we merged

[1] For this reason, the period of interest goes from 1971 to 2020.
[2] The most frequent author name disambiguation problems are the following: surname and given name swapped; full name versus first name use; unicode conversion issue; typos in the author's name.

these datasets and extracted, through the *rscopus* package, the information on all the papers published by each author during their career. Finally, we realized a network starting from these data.

Figure 1: Sample overview of researchers of the Italian Institute of Technology



After the creation of the co-authorship network, we performed a descriptive analysis of the researchers' characteristics. Figure 1 shows a brief overview of our sample. It is worth noting the distribution of the research domains[3] per gender: in some RDs, gender distribution is proportional to the total number of females and males, while others exhibit a gender prevalence.

Since interdisciplinary cooperation fosters innovation, the following section undertakes a community detection analysis to assess the level of heterogeneity within the research communities of the IIT. Finally, to quantify the strength of these intra-group effects and understand how the sociodemographic characteristics of each researcher affect the propensity to form collaborative relationships, an ERGM was implemented.

2. Methods and results

The Institute presents four main research domains, along with the facilities and the internal staff, leading to the classification of researchers into six distinct clusters. Information on the RDs can be used to evaluate *homophily*[4] through the *modularity* measure[5]. In addition to this domain-based node classification, the network's structure can be used to identify an alternative clustering. For graph partitioning, we applied the Girvan-Newman betweenness algorithm that computes the edge betweenness for each link[6], progressively eliminating those with the highest values to detect clusters, as illustrated in Figure 2. Modularity measures are notably high in both partitions: 0.53 for RD-based clusters and 0.65 for the divisive hierarchical clustering algorithm. Consequently, we can assert that the collaboration network of the IIT shows domain-based homophily.

[3] In addition to the four main research areas, we analyzed the group of facilities that are lines of support for the main RDs, and the internal staff.

[4] Homophily refers to the tendency of a node to form links with other nodes sharing similar properties. It is a first demonstration of how the context of a social network can change the formation of the link within it.

[5] Mathematically, the modularity is the fraction of the edges that fall within a cluster *c* minus the expected fraction if edges were randomly distributed. This measure, like a correlation coefficient, ranges between -1 and 1.

[6] *Edge betweenness centrality* is a measure describing the number of shortest paths that go through an edge between pairs of nodes in a graph.

Figure 2: Functioning of the Girvan-Newman betweenness algorithm

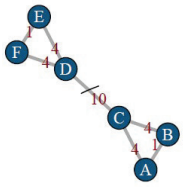


Figure 3: Giant component of the IIT co-authorship network 2006-2020. Note: vertices of the same color belong to the same cluster

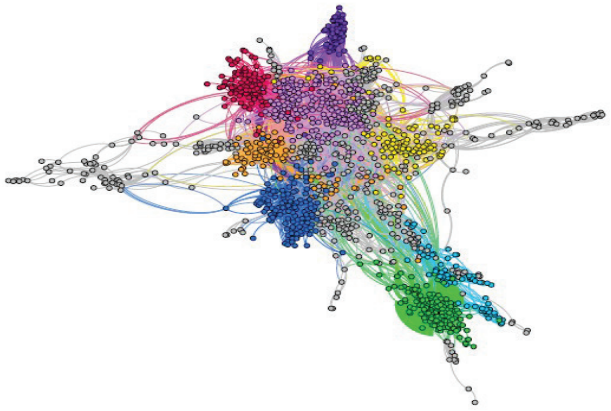


Table 1: Contingency table of two partitions

Cluster	*Computational Sciences*	*Facility*	*Internal Staff*	*LifeTech*	*Nanomaterials*	*Robotics*	*Total*
1	5	15	2	12	183	4	221
3	28	4	8	139	4	8	191
19	3	1	1	1	0	129	135
2	2	0	0	10	85	1	98
6	29	6	0	42	6	2	85
10	1	5	13	27	2	32	80
8	1	14	8	0	0	46	69
9	1	77	3	9	40	0	60
5	1	0	0	58	0	0	59
7	0	0	0	53	2	0	55
4	0	0	3	0	40	0	43
13	0	0	1	1	39	2	43
18	0	0	1	35	0	0	36
15	1	0	1	0	32	1	35
12	0	0	0	0	4	29	33
26	0	0	0	1	20	0	21
22	13	0	0	0	0	0	13
24	0	0	3	2	0	7	12
14	0	0	0	0	0	11	11
28	10	0	0	0	0	0	10

Table 1 reports the results of a two-cluster analysis, confirming researchers' tendency to collaborate within the same domain. The largest cluster mainly comprises Nanomaterials researchers, the second LifeTech, and the third Robotics. Computational sciences authors show wider distribution but a strong tendency to cooperate with the Life tech domain, as exemplified in clusters 3 and 6. Figure 3 illustrates the largest network communities identified by the hierarchical algorithm.

We hypothesized that the attributes of the authors who make up our graph vertices may affect their propensity to form ties. The probability of a tie's existence could be determined by the structural features of the graph. To test these hypotheses and identify the extent to which each variable influences the probability of a link between two authors, we performed an ERGM. This model probabilistically explains network links, clarifying the specific structure over alternative configurations. Equation 1 outlines the general model form:

$$\Pr(Y = y|X) = \frac{\exp(\theta' g(y, X)}{k(\theta)} \quad (1)$$

where y is a network realization in the set of all possible networks Y; $g(y, X)$ is a vector of statistics and additional covariate information about the network; θ is a vector of coefficients for those statistics; $k(\theta)$ represents a normalizing constant that ensures that Equation 1 is a legitimate probability distribution.

The *R* package *ergm*, from the *statnet* suite was used for ERGMs fitting (Hunter et al., 2008). For analysis purposes, four models were developed: two simpler models having as regressors only the sociodemographic attributes of each author and their interaction terms (e.g., Sex*RD or Sex*Seniority)[7], and two models including network structure statistics, such as subgraphs or triadic closure, in addition to the variables of the simplest models[8].

In what follows, we describe the results of the better-fit model according to the log-likelihood and Akaike's information criterion (AIC) values and the implementation of some goodness-of-fit diagnostics for ERGMs[9]. This model controls not only for exogenous attributes but also for endogenous ones.

Table 2 reports the main results, including estimated coefficients, their standard errors, and significance levels. ERGMs operate on a logic analogous to that of a logistic regression model, predicting a binary outcome, i.e., the existence of a link, based on selected predictor regressors that are supposed to explain the observed outcome. The model includes a term for the overall number of edges in y several terms related to nodal properties, and a network statistic variable (*GWESP*). The nodal covariates represent homophily terms, reflecting the growing body of empirical research highlighting the influence of node similarity on collaborative relationship formation (Lü et al., 2009). The *nodematch* and *absdiff* commands for homophily terms permit comparison of log-odds to a reference point or a baseline, as in a conventional logistics model.

Table 2: Estimated coefficients and standard errors for ERGM parameters

Exponential Random Graph Model	
Edges	-35.90(0.83)***
GWESP	24.81(0.64)***
Nodematch Sex (F)	-0.08(0.02)***
Nodematch Sex (M)	0.07(0.01)***
Nodematch RD Internal Staff	0.38(0.13)**
Nodematch RD Computational	0.95(0.04)***
Nodematch RD Facility	1.23(0.07)***
Nodematch RD LifeTech	0.77(0.02)***
Nodematch RD Nanomaterials	0.69(0.02)***
Nodematch RD Robotics	0.56(0.03)***
Nodematch Nationality (Other)	0.06(0.01)***
Nodematch Nationality (IT)	0.23(0.02)***
Absdiff Seniority	0.00(0.00)***
AIC	76538.81
Log likelihood	-38256.40
*** p <0.001; ** p<0.01; * p<0.05	

The baseline (*Edges* in Table 2) represents a collaborative relationship between two authors

[7] These models, which encompass solely nodal covariates as regressors are defined as *dyadic-independence models* since links and their values are stochastically independent given the model parameters.

[8] Conversely, these models are referred to as *dyadic-dependence models* since the presence or absence of a link lead to changes in the network statistics. These models are more complex and require the *Markov chain Montecarlo* (MCMC) algorithm instead of the *Maximum likelihood* for the parameter estimates.

[9] The goodness-of-fit diagnostics consist of performing graphical goodness-of-fit tests. The approach entails simulating a large number of potential networks using estimated model parameters via an MCMC algorithm. Subsequently, various structural measures are computed from these graphs. Typically, these measures encompass degree counts, edgewise shared partner counts, and minimum geodesic distances since these measure phenomena such as sociality, clustering, transitivity, and centrality. Finally, the goodness-of-fit involves comparing the distributions of network statistics from the simulated graphs to the observed graph.

with differing sex, nationality, RD membership, and identical career lengths (Seniority).

Firstly, it is evident that all the regressors hold statistical significance, with RD playing a more relevant role than other variables, aligning with community detection results: Collaboration is more likely to occur among researchers within the same domain. Furthermore, the model highlights a higher probability of collaboration among facilities or computational sciences researchers than in the other RDs. Regarding sex, holding other variables constant, the likelihood of a cooperative relationship between two males surpasses that between a male and a female or between two females; this suggests that females are less likely to cooperate. For nationality attribute, Italian authors are around 1.26 times (in fact, *exp(0.23)* ≈ *1.26*) more likely to connect in the network (ceteris paribus). Finally, a larger difference in seniority enhances the likelihood of collaboration, making co-authorship more probable between a junior and a senior researcher than between two juniors or two seniors.

Up to this point, we have seen that two authors can match based on exogenous attributes (socio-demographic characteristics); however, it is necessary to consider endogenous attributes (network structural features) as well. For instance, local clustering may arise from actors forming collaborative relationships with shared partners. To represent the network structure more accurately, and since it is well known that triangles abound in social networks, the model also includes a term for transitivity, the *geometrically weighted edgewise shared partnership* (GWESP) statistic. GWESP is a function that uses a curved exponential family form to represent the shared partner distributions[10]. This variable represents a term for triadic closure, or network transitivity, according to which the presence of a link between actors i and j (Y_{ij}=1), and between actors j and k (Y_{jk}=1) increases the probability of a social relationship between actors i and k (Y_{ik}=1$) (Levy et al., 2018) since previous relationships can influence the formation of links in the cluster (Capone et al., 2020). The GWESP statistic is given by the following equation:

$$GWESP(y;\alpha) = e^{\alpha} \sum_{k-1}^{n-2} \{1 - (1 - e^{-\alpha})^k\} \, ESP_k(y) \tag{2}$$

where α is a decay parameter, the higher its value, the slower the decay will be[11]; $ESP_k(y)$ represents the number of edges between two vertices in y with exactly k shared (edgewise) partners; n is the number of nodes in the network; $n-2$ is the maximum number of edgewise-shared partners for any pair of vertices in the network.

In terms of network structure, the magnitude of the triangle statistic coefficient indicates a non-trivial transitivity effect. This implies that a link that closes a triangle is now more likely to occur than one link that does not close a triangle. Therefore, the IIT researchers' collaborative network reflects a triadic structure, indicating a social proximity process where links form between acquaintances and actors connected through an intermediary (Capone et al., 2020).

3. Conclusions

Nowadays, it is essential to maintain ongoing assessments of research centers, encompassing the evaluation of scientific output, inter-researcher collaboration, and the level of openness relative to national and international peers. For this reason, the present study focuses on the Italian Institute of Technology, examining collaboration within the center and across its research domains.

We verified the assumption according to which researchers from the same domain tend to write papers together. However, we also identified a pattern of collaboration between different domains; for instance, the Computational science domain demonstrates a strong inclination to cooperate with the Life Tech research domain. The transversal impact of the Computational

[10] For details on the curved exponential-family models for graphs see Hunter and Handcock (2006) and Hunter (2007).
[11] In the present work, we set a fixed decay parameter α = 0.25 following common ERGM modeling practice.

sciences' RD on the other IIT's scientific activities represents a major asset for the research of the Institute since interdisciplinary cooperation could lead to innovative research outputs and outcomes and potentially drive progress in specific scientific areas. The findings provide individual researchers with a means to assess their cooperative inclinations and may inform the development of research policies with potential implications for organizations' communities.

Furthermore, the paper contributes to the literature on network formation. Our results show gender differences in the propensity to collaborate, also searching for potential differences across research fields. Precisely, females are less likely to cooperate within the Institute. The empirical findings of the study offer valuable support to policymakers in designing and evaluating the effectiveness of interventions intended to favor gender equity in collaboration. In addition, the results concerning the network structure variable (GWESP statistic) show that there is also evidence for a non-trivial transitivity effect, and this result highlights the relevance of the role played by trust and previous relationships in the formation of a collaborative relationship. This culture of trust should be extended to the research community to facilitate more effective and productive collaboration.

It is fitting to acknowledge certain limitations in this study. Firstly, co-authorship data is just one indicator of scientific collaboration; not all collaborations result in publications (Sampaio et al., 2016). Secondly, our data collection relied solely on a bibliometric source, suggesting our dataset doesn't cover the entire IIT collaborative network. In fact, other databases may contain publications in different journals or languages, contributing to a more comprehensive representation (Abbasi et al., 2011).

To extend our research, we intend to include additional variables in the ERGM, such as the position covered by researchers within the IIT, to confirm the fact that junior and senior researchers are more likely to cooperate than two juniors or seniors and demonstrate that triangles are not formed exclusively by male researchers in top positions.

References

Abbasi, A., Hossain, L., Uddin, S., Rasmussen, K.J. (2011). Evolutionary dynamics of scientific collaboration networks: Multi-levels and cross-time analysis. *Scientometrics,* **89**(2), pp. 687-710.

Acedo, F.J., Barroso, C., Casanueva, C., Galán, J.L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies,* **43**(5), pp. 957-983.

Capone, F., Zampi, V. (2020). Proximity and centrality in inter-organisational collaborations for innovation: A study on an aerospace cluster in Italy. *Management Decision,* **58**(2), pp. 239-254.

Hunter, D.R., Handcock, M.S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics,* **15**(3), pp. 565-583.

Hunter, D.R. (2007). Curved exponential family models for social networks. *Social Networks,* **29**(2), pp. 216-230.

Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software,* **24**(3), nihpa54860.

Levy, M.A., Lubell, M.N. (2018). Innovation, cooperation, and the structure of three regional sustainable agriculture networks in California. *Regional Environmental Change,* **18**, pp. 1235-1246.

Lü, L., Jin, C.H., Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Physical Review E,* **80**(4), 046122.

Resce, G., Zinilli, A., Cerulli, G. (2022). Machine learning prediction of academic collaboration networks. *Scientific Reports,* **12**, 21993.

Sampaio, R.B., Fonseca, M.V.D.A., Zicker, F. (2016). Co-authorship network analysis in health research: Method and potential use. *Health Research Policy and Systems,* **89**(2), pp. 1-10.