Analysis of Scientific Networks: case studies and paper development.

Issues in the definition and analysis of co-authorship networks. Insights from empirical data

Maria Prosperina Vitale* - *mvitale@unisa.it*

Dept. of Political and Social Studies, University of Salerno (Italy)
* **Joint research project with**:
D. De Stefano and S. Zaccarin (University of Trieste)
V. Fuccella (University of Salerno)

Moscow, 16 July 2019

10th International Summer School "Analysis of Scientific Networks"

ANR-Lab

Talk Outline

Issues in the definition and analysis of co-authorship networks. Insights from empirical data.

- Issues in the analysis of co-authorship networks (De Stefano et al., 2011)
- The use of different data sources in the analysis of co-authorship networks for a target population (De Stefano et al., 2013)
- Improving co-authorship network structures by combining multiple data sources (Fuccella et al., 2016)
- Co-authorship ties in a target population: data quality issues and network analysis (De Stefano et al., work in progress)

Co-authors and publications

Main references

- De Stefano, D., Giordano, G., & Vitale, M.P.: Issues in the analysis of co-authorship networks. *Quality & Quantity*. **45**, 1091-1107 (2011)
- De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*. **35**, 370-381 (2013)
- Fuccella, V., De Stefano, D., Vitale, M. P., Zaccarin, S.: Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians. *Scientometrics*. **107**, 167-184 (2016)
- De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: Co-authorship Ties in a Target Population: Data Quality Issues and Network Analysis (....)

- De Stefano, D. and Zaccarin, S.: Co-authorship networks and scientific performance: an empirical analysis using the generalized extreme value distribution. Journal of Applied Statistics, 43, 262â279 (2016)
- De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: Using web scraping techniques to derive co-authorship data: insights from a case study. In: Book of Short Papers SIS2018. 49th scientific meeting of the Italian Statistical Society (pp. 922-928), Pearson (2018)
- De Stefano, D., Vitale, M. P., Zaccarin, S.: Community structure in co-authorship networks: the case of Italian statisticians. In AA.VV. Statistical Learning of Complex Data Pag.1-8 Heidelberg Springer Nature (2019)
- De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: Web-Based Data Collection and Quality Issues in Co-Authorship Network Analysis. In: Book of Short Papers SIS2019, Smart Statistics for Smart Applications (pp. 811-816), Pearson (2019)

Scientific collaboration

- Scientific collaboration → key element in knowledge advancement for interactions and sharing competences among scholars
- Co-authorship relationship → proxy of scholars' collaborative behaviors

Co-authorship is analysed by means of Social Network Analysis tools, for instance:

- to explore topological properties → small world, scale-free configurations
- to discover clusters → community detection, generalized blockmodeling
- to study the effect of collaboration patterns on the evolution over time of research topics
- to analyse the effect of authors' network position on scientific performance

Data retrieval for co-authorship network extraction

Co-authorship networks are reconstructed from bibliographic data

Co-authorship in a specific scientific field

- International general bibliographic archives: ISI-WoS, Scopus, ...
- Thematic bibliographic archives: Medline, Econlit, Current Index to Statistics, ...

Seminal studies on co-authorship patterns are based on international databases containing mainly high-impact publications (e.g. Moody, 2004; Newman, 2004a; Goyal et al., 2006)

Co-authorship in a specific scientific community or country (target population)

- Individual scientific CVs
- Local/National bibliographic archives $\rightarrow$ good coverage of whole research products of each scientist (Kronegger et al., 2011; De Stefano et al., 2013; Bellotti et al., 2016; Sciabolazza et al., 2017)

Seminal studies on co-authorship networks

Substantial body of works on the analysis of co-authorship patterns at national and international level:

- Physics and Biomedical research (Newman, 2004; Barabasi, 2002) ⇒ MEDLINE and Spires
- Economics (Goyal, 2006; Maggioni and Uberti, 2011) ⇒ Econlit
- Sociology (Moody, 2004; Ferligoj and Kronegger, 2011, 2012, 2016) ⇒ Sociological Abstracts and national archive (COBISS database)

Common aims:

- understanding networks properties through SNA
- implications of collaboration patterns on the evolution over time of topics and methods
- evaluation of scientific productivity and quality
- emergence of network topology structure governing collaboration behaviors

Co-authorship network definition, Newman 2004

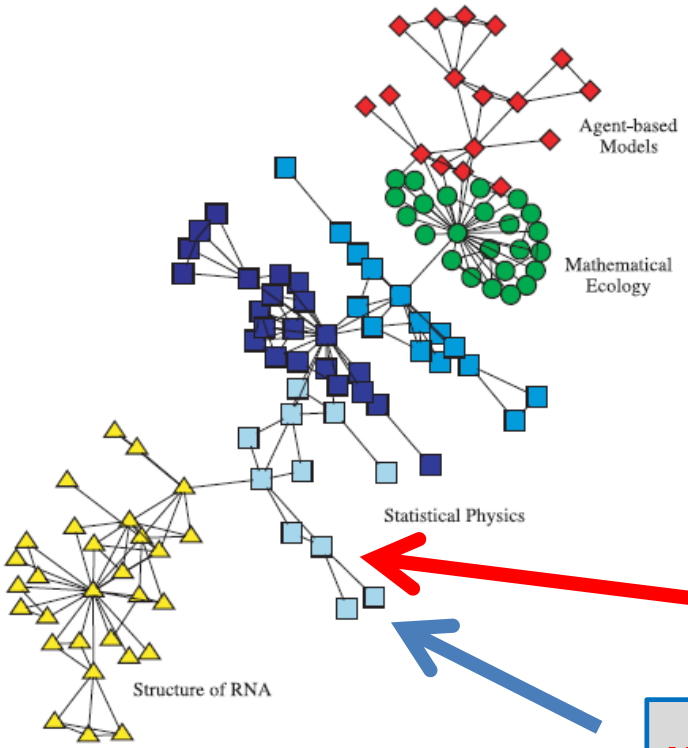


Fig. 1. An example of a small coauthorship network depicting collaborations among scientists at a private research institution. Nodes in the network represent scientists, and a line between two of them indicates they coauthored a paper during the period of study. This particular network appears to divide into a number of subcommunities, as indicated by the shapes of the nodes, and these subcommunities correspond roughly to topics of research, as discussed by Girvan and Newman (37).

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, *101*(suppl 1), 5200-5205.

https://www.pnas.org/content/101/suppl_1/5200

Line between two nodes/scientists indicates they **coauthored** a paper during the period of study

Nodes represent scientists

Issues in co-authorship definition for a target population (De Stefano et al., 2011)

1. Data collection and bibliographic archives
2. Setting network boundaries
3. Definition of the co-authorship data matrix
   - Data cleaning and trasformation of bibliometric data into co-authorship networks
4. Network data analysis and interpreting results

Issues in co-authorship definition for a target population: Data collection and bibliographic archives (1)

Reconstructing co-authorship networks among scientists through bibliographic archives: international *vs* local archives

Issues

- international bibliographic archives are not able to cover all kinds of scientific production (e.g., Hicks, 1999);
- integration of high-impact journals databases with specialized and local bibliographic archives may be the best compromise to obtain a good coverage of whole research products of scientists involved in a specific field.

International databases for co-authorship definition (1)

Table 1. Summary statistics for the three coauthorship networks analyzed here

	Biology	Physics	Mathematics
Number of authors	1,520,251	52,909	253,339
Number of papers	2,163,923	98,502	—
Papers per author	6.4	5.1	6.9
Authors per paper	3.75	2.53	1.45
Average collaborators	18.1	9.7	3.9
Largest component	92%	85%	82%
Average distance	4.6	5.9	7.6
Largest distance	24	20	27
Clustering coefficient	0.066	0.43	0.15
Assortativity	0.13	0.36	0.12

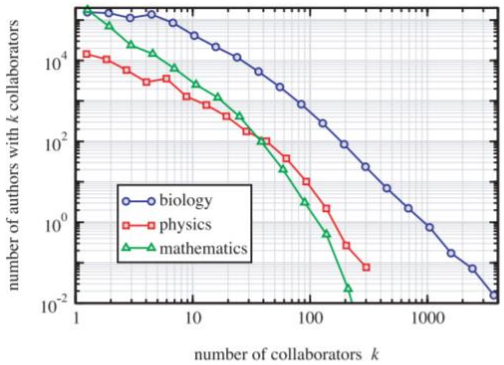


Fig. 2. Histograms of the distribution of numbers of collaborators for scientists in each of three fields studied.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, *101*(suppl 1), 5200-5205.

https://www.pnas.org/content/101/suppl_1/5200

Data source

- A network of coauthorships of papers in the **Medline bibliographical database** from 1995 to 1999, inclusive. Medline is a widely used and compendious database of papers covering biomedical research
- A network of coauthorships of physicists assembled from papers posted on the widely used **Physics E-print Archive** at Cornell University between 1995 and 1999
- A collaboration network of mathematicians compiled from databases maintained by the **journal Mathematical Reviews**

International databases for co-authorship definition (2)

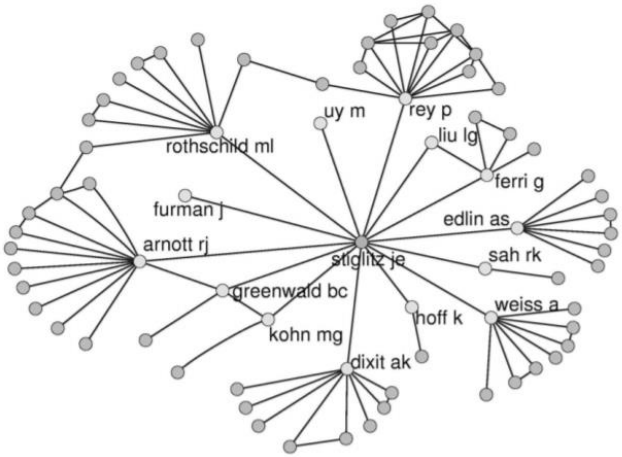


FIG. 2.—Local network of collaboration of Joseph E. Stiglitz in the 1990s. The figure shows all nodes within distance 2 of J. E. Stiglitz as well as the links between them. Some economists might not appear because of misspellings in EconLit. The figure was created by software program Pajek.

Goyal, S., Van Der Leij, M. J., & Moraga-González, J. L. (2006). Economics: An emerging small world. *Journal of political economy*, *114*(2), 403-412.

Data source We study the world of economists who published in journals included in **EconLit**. We cover **all journal articles** that appear in the 10-year windows 1970–79, 1980–89, and 1990–99. The list of journal articles includes all papers in conference proceedings, as well as short papers and notes.
Not cover working papers and work published in books.

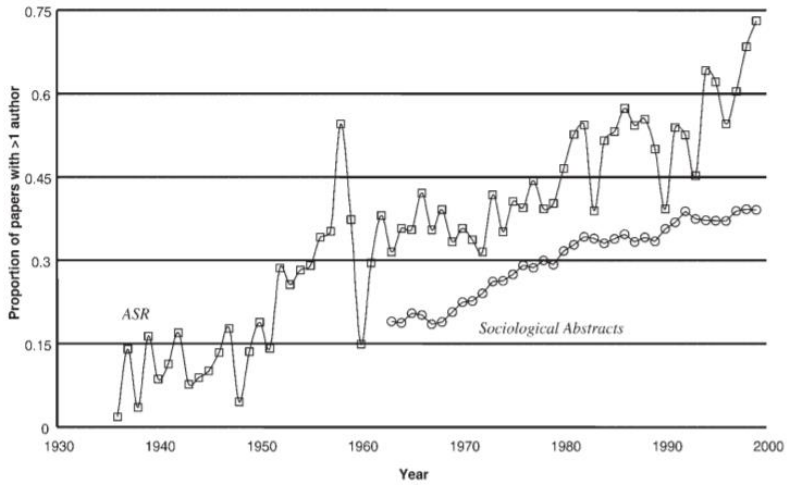


Figure 1. Coauthorship Trends in Sociology

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American sociological review*, *69*(2), 213-238.

Data source **Journal articles** listed in **Sociological Abstracts** published between 1963 and 1999. Sociological Abstracts database covers all journals in sociology proper, and many journals publishing sociologically relevant work in other Fields.
It limits coverage to journal articles, neglecting conference presentations, book reviews, essays, or books.

Co-authorship Ferligoj et al., 2012

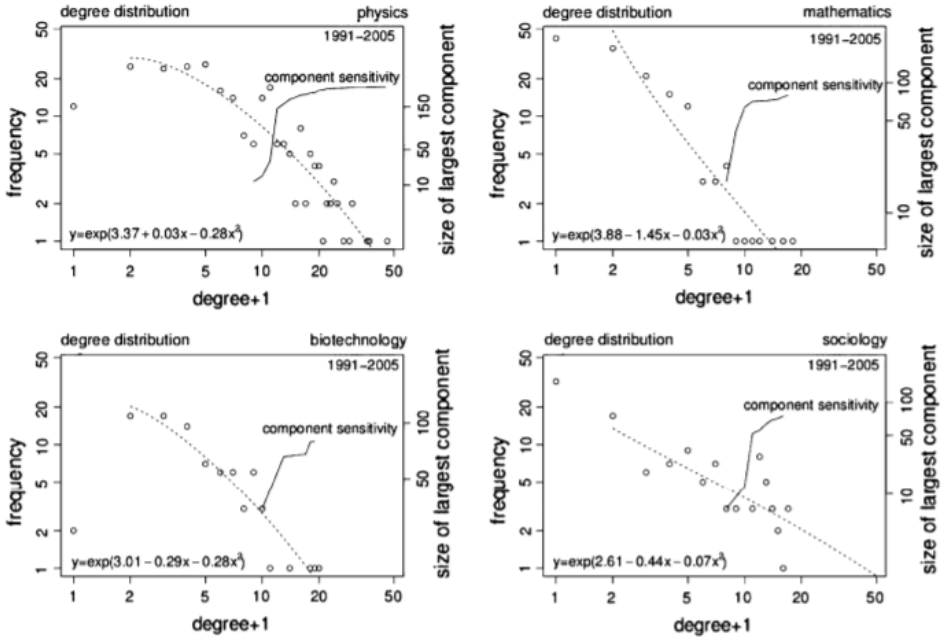


Fig. 1 Degree distribution and component sensitivity within so-authorship networks

Data sources in Slovenia

(i) the Current Research Information System (SICRIS) which includes the information on all active researchers registered at the Slovenian Research Agency and
(ii) the Co-operative On-Line Bibliographic System & Services (COBISS) which contains a database of all publications that can be located through Slovenian libraries

Modeling of network dynamics in four disciplines

- small-world structure of networks
- mechanism of preferential attachment
- stochastic actor-oriented model

Kronegger, L., Mali, F., Ferligoj, A., & Doreian, P. (2012). Collaboration structures in Slovenian scientific communities. *Scientometrics*, *90*(2), 631-647.

Issues in co-authorship definition for a target population: Setting network boundaries (2)

Whole-network approach: boundary specification strategies (Laumann et al. 1989; Marsden 2005)

- Positional approach based on characteristics of actors or formal membership criteria
- Event-based approach founded on participation in some class of relational events
- Relational approach guided by social linkages among actors
  - Relational approach $\Rightarrow$ expanding selection (Doreian and Woodard, 1992), i.e. beginning with a provisional "fixed" list of actors deemed to be in a network, then add actors linked to those on the initial list (as in *snowball sampling design*)

Issues in co-authorship definition for a target population: Definition of the co-authorship data matrix (3)

Network data and weighting systems

- Affiliation matrix in which the generic element is equal to 1 if the i-th author is present in the j-th paper and 0 otherwise
- Adjacency matrix undirected weighted/binary adjacency matrix whose entries are equal to 0 if two authors have never co-authored, elsewhere they hold the number of co-authored papers by pairs of authors
  - *alternative weighting system*, co-authorship relationship is stronger if two scientists are the sole authors of a paper (Newman, 2001)

Co-authorship network definition, Moody 2002 (2)

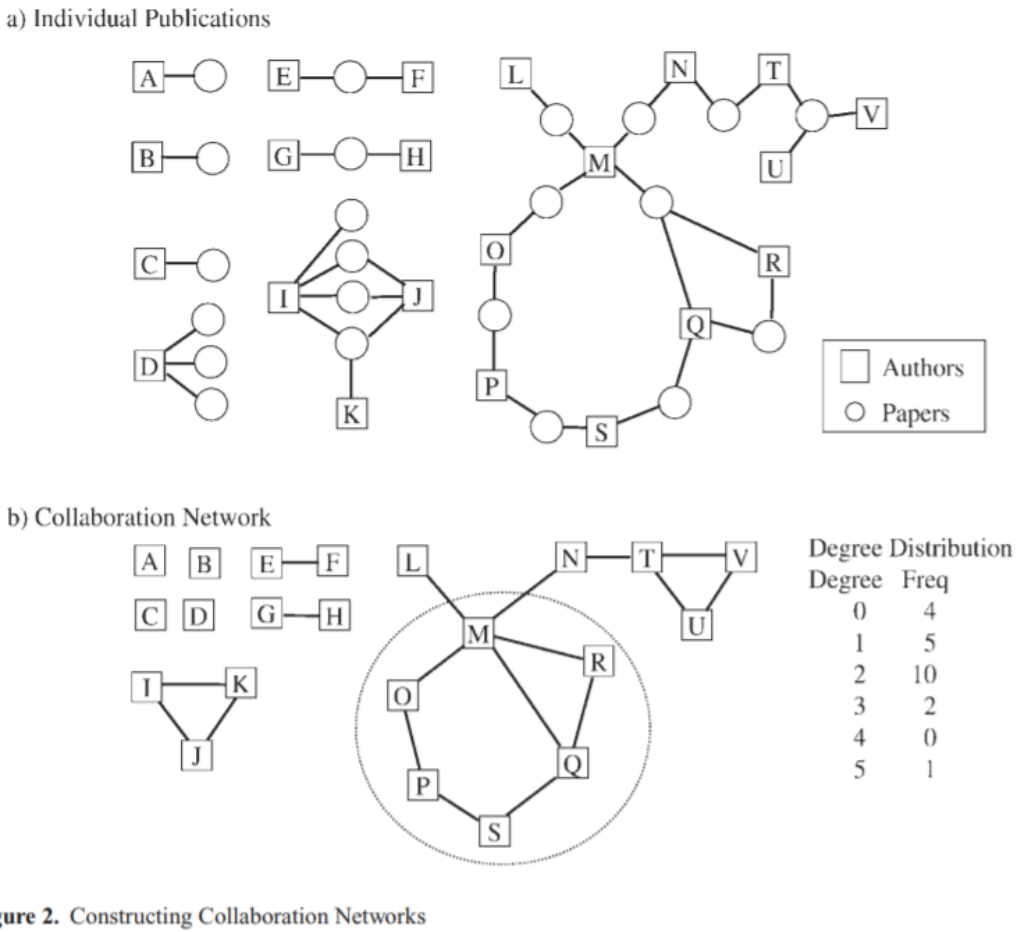


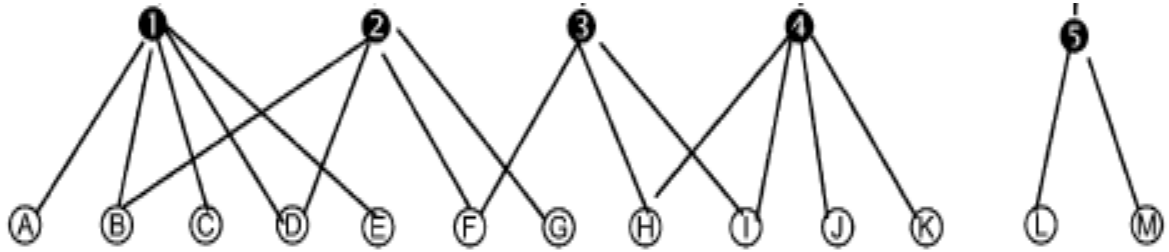
Figure 2. Constructing Collaboration Networks

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American sociological review*, *69*(2), 213-238.

Co-authorship network construction

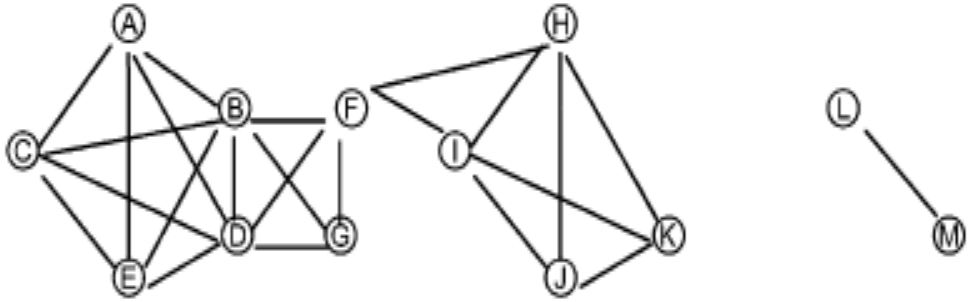
Two-mode network

$\mathbf{AP}_{(n \times m)}$ two sets of nodes:
n **A**uthors $\times$ m **P**apers



One-mode undirected network

$\mathbf{AA}_{(n \times n)} = \mathbf{AP} \times \mathbf{AP}^T$
authors as nodes, papers as links
$a_{ij} \geq 1$ if i-th author have written at least a paper with j-th author



Usually the adjacency matrix **AA** is analysed as a binary network (0/1)

Issues in co-authorship definition for a target population: Network data analysis and interpreting results (4)

Analysis of a co-authorship network for describing:

- structural characteristics of the whole network and topologies
- role played by nodes considering actor-level network measures
- dynamic network configuration over time

Network topologies (1)

Types of Networks: Random, Small-World, Scale-Free

https://noduslabs.com/radar/types-networks-random-small-world-scale-free/

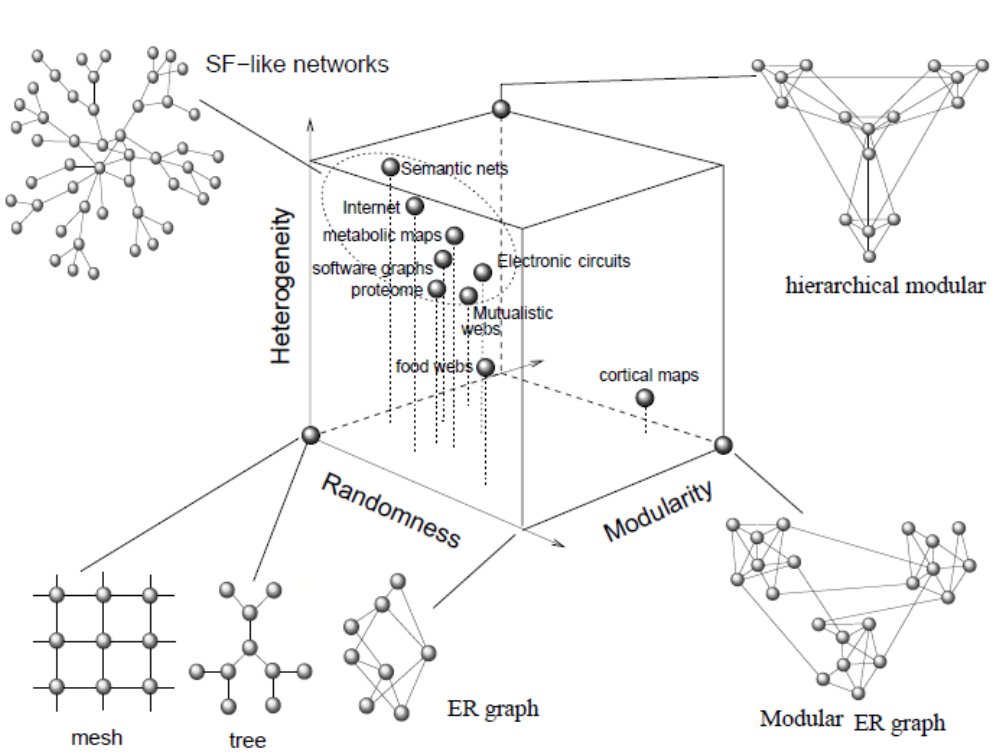
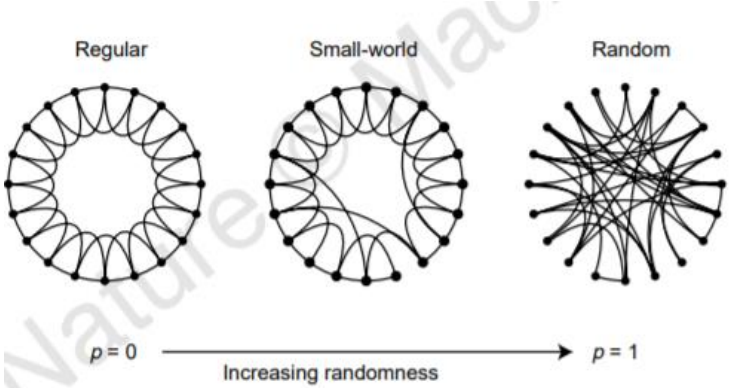


FIG. 3 A zoo of complex networks. In this qualitative space, three relevant characteristics are included: randomness, heterogeneity and modularity. The first introduces the amount of randomness involved in the process of network's building. The second measures how diverse is the link distribution and the third would measure how modular is the architecture. The position of different examples are only a visual guide. The domain of highly heterogeneous, random hierarchical networks appears much more occupied than others. Scale-free like networks belong to this domain.

Solé, R. V., & Valverde, S. (2004). Information theory of complex networks: on evolution and architectural constraints. In *Complex networks* (pp. 189-207). Springer, Berlin, Heidelberg.

Network topologies (2)

Network topologies



Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440.

Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *science*, *325*(5939), 412-413.

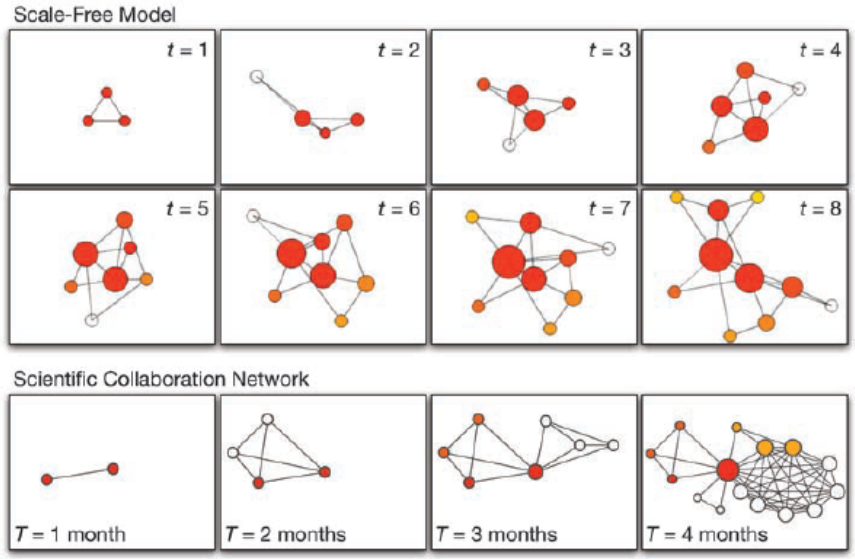


Fig. 1. The birth of a scale-free network. (**Top** and **Middle**) The simplest process that can produce a scale-free topology was introduced a decade ago in (*6*), and it is illustrated in the top two rows. Starting from three connected nodes (top left), in each image a new node (shown as an empty circle) is added to the network. When deciding where to link, new nodes prefer to attach to the more connected nodes, a process known as preferential attachment. Thanks to growth and preferential attachment, a rich-gets-richer process is observed, which means that the highly connected nodes acquire more links than those that are less connected, leading to the natural emergence of a few highly connected hubs. The node size, which was chosen to be proportional to the node's degree, illustrates the natural emergence of hubs as the largest nodes. The degree distribution of the resulting network follows the power law (Eq. 1) with exponent $\gamma = 3$. See also movies S1 to S3. (**Bottom**) Illustration of the growth process in the co-authorship network of physicists. Each node corresponds to an individual author, and two nodes are connected if they co-authored a paper together. The four images show the network's growth at 1-month time intervals, indicating how the network expands in time, leading to the emergence of a clear hub. Once again, the node size was chosen to be proportional to the node's degree. [Credit: D. Wang and G. Palla]

Blockmodeling and co-authorship network

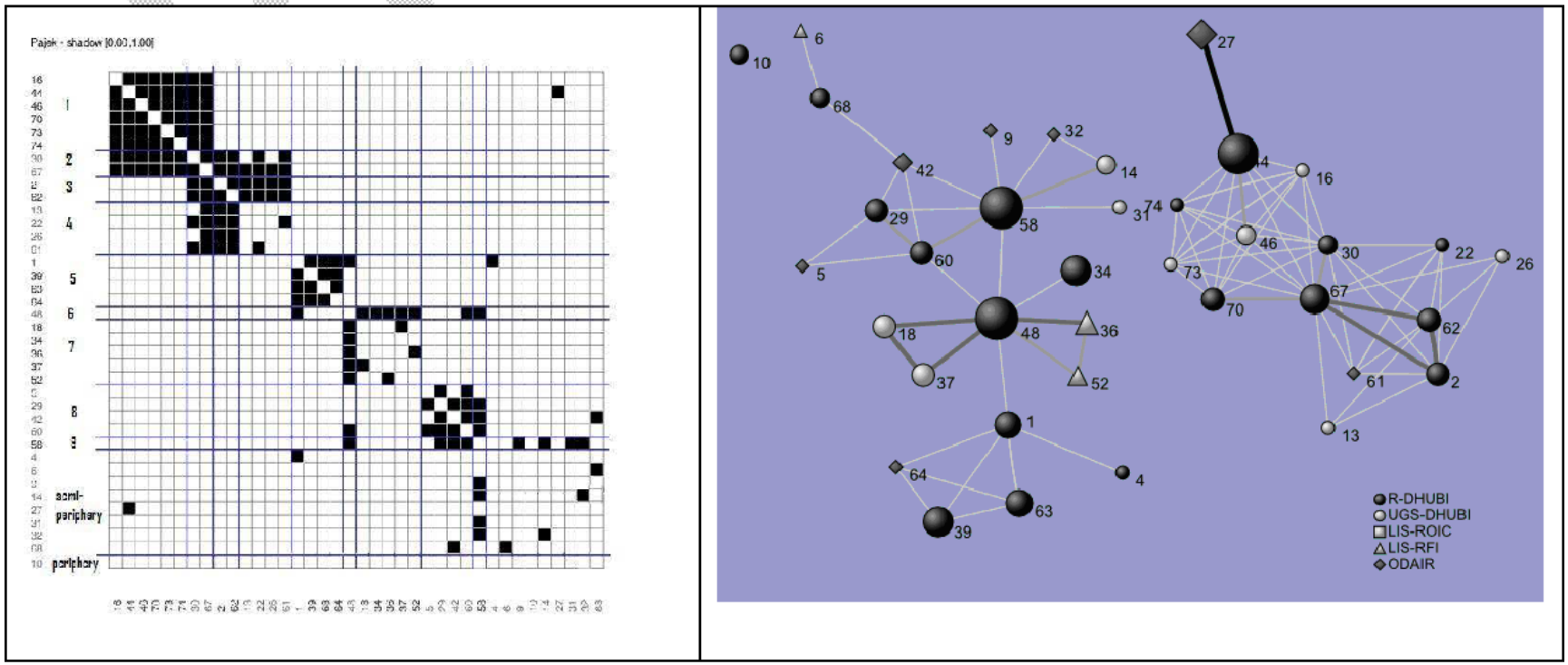


Figure 4. (a) Blockmodel structure and (b) co-authorship network, 2007-2009

blockmodeling in the micro-level study of the internal structure of co-authorship networks over time

data source: corpus of curricula vitae (CV) and the bibliographies of the teaching staff

Chinchilla-Rodríguez, Z., Ferligoj, A., Miguel, S., Kronegger, L., & de Moya-Anegón, F. (2012). Blockmodeling of co-authorship networks in library and information science in Argentina: a case study. *Scientometrics*, *93*(3), 699-717.

Our case study and motivation

Co-authorship patterns in Statistics, focusing on:

Italian academic statisticians

- target population involved in a discipline which is not yet fully explored in terms of its scientific collaboration behaviour (i.e., co-authorship);
- Statistics presents characteristics common to Natural sciences as well as Social sciences, playing a central role in all sciences in view of the importance of statistical methods in everyday applications;
- no unique archive for publications in Statistics and hence interest to trace co-authorship in distinct data sources
- derive a integrated archive to take into account all kind of scientific products (international and national level) by identifying and linking duplicate records (record linkage –RL), and by dealing with issues related to authors name disambiguation –AD;
- adopt web scraping procedure to extract data from a unique national bibliographic archive adopted by most of the Italian universities.

Co-authorship network definition and analysis

- from different data sources (De Stefano et al., 2013) ...
- to unique database (Fuccella et al., 2016)

De Stefano et al., 2013

1. To explore the pattern of the scientific co-authorship network among Italian academic statisticians (792 grouped in **5 subfields** at March 2010)
   - S-/01 - Statistics
   - S-/02 - Stat. for E&T Research
   - S-/03 - Economic Statistics
   - S-/04 - Demography
   - S-/05 - Social Statistics
2. To reveal the influence of distinct data sources (different kinds of publications) on co-authorship patterns
   - International general: ISI Web of Science (WoS)
   - International thematic: Current Index to Statistics (CIS)
   - National/thematic: MIUR database of granted national projects (PRIN)
3. To model the impact of the individual network position on scientific performance (measured by *h*-index)

Data sources for co-authorship of Italian statisticians

- WoS high-impact journals (> 10000) and conference proceedings (> 111000) in all disciplines published since 1900

  (oldest publication for an Italian statistician: 1984)

- CIS 160 core statistical journals, around 1200 additional journals with statistical oriented articles and 10000 books in statistics published since 1975

  (oldest publication for an Italian statistician: 1973)

- PRIN publications (of statisticians) included in granted research projects PRIN coordinated by a statistician (2000-2008)
  2000–2006: max 30 publications
  since 2007: publications of all project members

  (oldest publication for an Italian statistician: 1966)

Heavy data cleaning phase to handle misreporting problems: different names for the same author, same name for different authors, duplicated entries ...

Co-authorship patterns in statistics: research hypotheses

- H1: co-authored publications by Italian academic statisticians are growing faster than single-authored publications, as observed in other disciplines (in all data sources)
- H2: The collaboration style of the overall Italian statistician community resembles the typical style observed in the literature for social sciences (e.g., Economics)
- H3: The subfields of Statistics exhibit different collaboration styles (different topological structures)
- H4: The scientific performance of Italian statisticians is related to authors' position in co-authorship networks

H2, H3, H4: expected different results for the three data sources

Authors coverage rates (%)

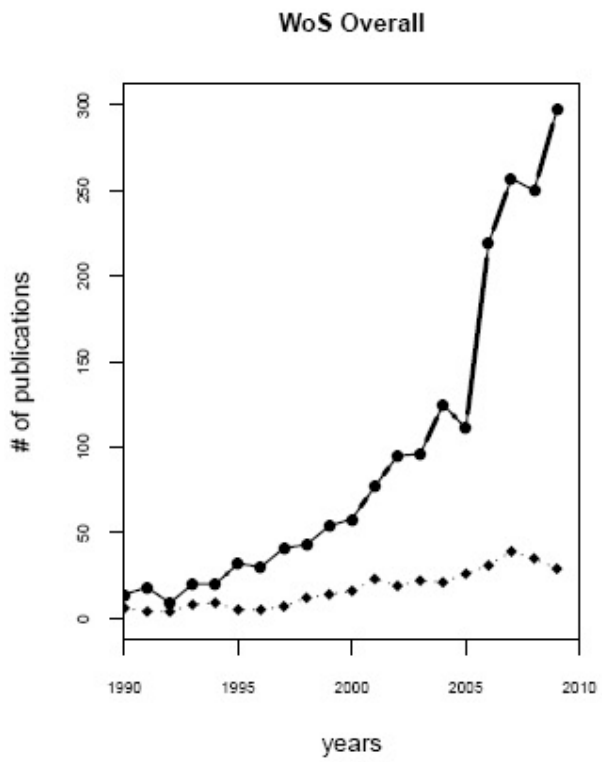
Subfields	WoS	CIS	Prin	*Never found*
S-/01 - Statistics (443)	71.3	85.1	72.7	7.9
S-/02 - Stat. for E&T Research (30)	86.7	60.0	73.3	13.3
S-/03 - Economic Statistics (160)	42.5	65.0	59.4	20.0
S-/04 - Demography (85)	40.0	48.2	67.1	27.1
S-/05 - Social Statistics (74)	50.0	55.4	81.1	12.2
All statisticians (792)	60.7	73.4	70.2	13.0

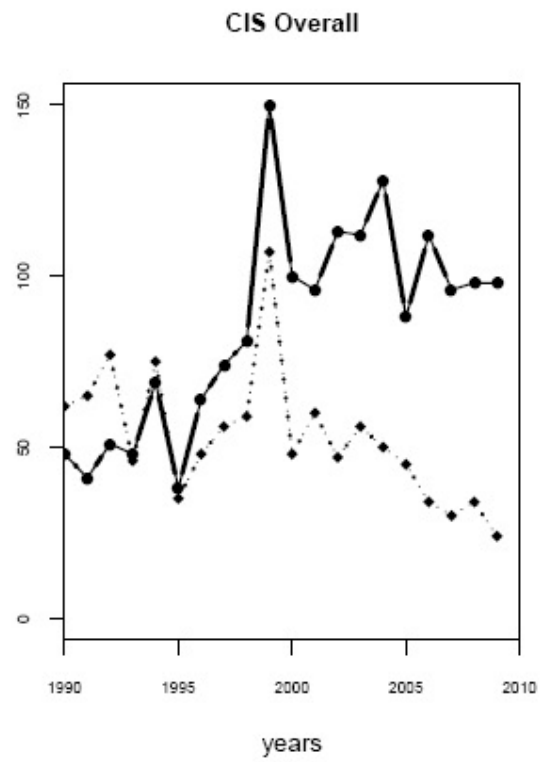
Subfields	WoS				CIS				Prin			
	# Papers	%CP	A×P	P×A	# Papers	%CP	A×P	P×A	# Papers	%CP	A×P	P×A
S-/01	1491	81.6	4.3	6.0	2927	56.0	2.4	7.8	3301	76.2	2.7	10.2
S-/02	361	99.2	49.2	15.7	123	79.7	2.9	6.8	394	83.5	4.1	17.9
S-/03	224	80.4	3.2	3.9	427	57.1	2.3	4.1	799	74.3	2.6	8.4
S-/04	122	83.6	3.6	5.2	82	70.7	2.7	2.0	612	65.0	2.6	10.5
S-/05	190	95.8	7.2	5.3	134	62.7	2.6	3.3	834	70.1	2.9	13.9
Total	2289	84.6	12.6	6.1	3518	55.3	2.4	6.4	5608	71.2	2.8	10.7

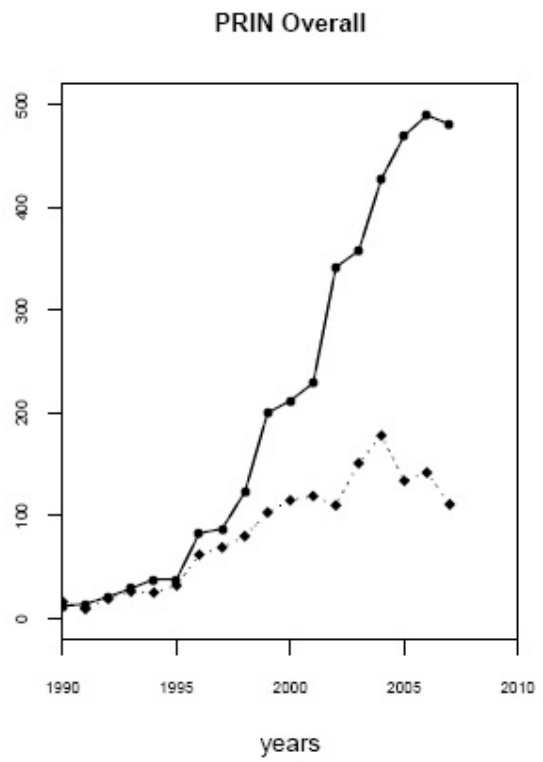
- WoS: highest % of co-authored papers (CP)
- PRIN: highest number of papers × authors (P×A)
- WoS: Stat. for E&T Res., highest number of authors × papers (A×P)

Trend single and coauthored publications

The number of co-authored publications is growing faster than number of single authored publications in all subfields since the end of '90 (with a ten-year delay with respect to other disciplines)



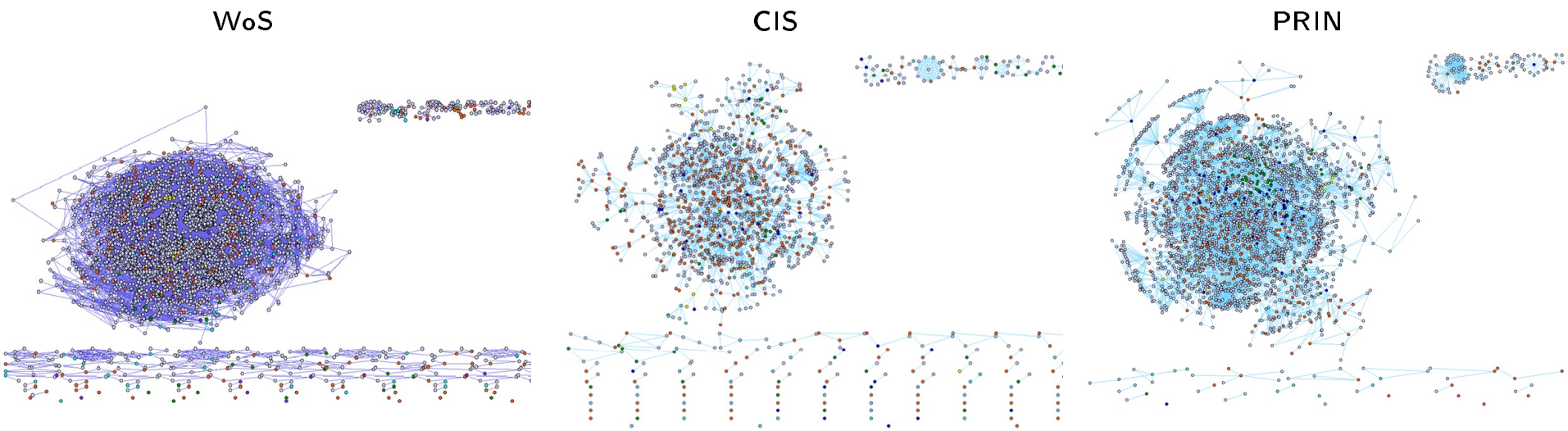




The pattern of the scientific co-authorship network

	WoS	CIS	Prin
N. of authors (secs)	5291 (481)	1525 (581)	2839 (556)
N. of isolated nodes	26	60	7
N. of edges (inside secs)	427238 (403)	2534 (631)	9379 (999)
Largest component (%)	91.7	87.7	94.9
N. of components ≥ 1	77	54	20
Density	0.031	0.002	0.002
Average degree	161.5	3.3	6.6
Largest distance	16	19	17
Average Path Length	5.47	7.15	6.52
Clustering coefficient	0.91	0.30	0.54

The pattern of the scientific co-authorship network



Main assumption: Collaborative behaviour within a scientific community closely depends on the topological features of the co-authorship network

- Network topology appears to be strongly related to data source characteristics ⇒ data source is not neutral
- The collaboration style of Italian academic statisticians resembles features partly observed in both social and natural sciences ⇒ Small-worldliness with relevant star actors roles
  - Likewise in Economics also for Statistics in Italy, highly linked authors act like interconnected stars (with different roles in the data sources) and their removal greatly increases the distance in the network.

Assessment of Small World and Scale Free properties

Overall Networks	WoS	CIS	PRIN
Small world			
$\ell(G)/\ell(ER)$	2.769	1.166	1.473
$\Gamma(G)/\Gamma(ER)$	29.663	138.195	231.842
$\ell(G)/\ell(CM)$	2.135	1.382	1.632
$\Gamma(G)/\Gamma(CM)$	2.510	45.903	50.188
Scale free			
- *power law*			
C	0.240	0.494	0.391
$\hat{\alpha}$	1.281	1716	1.515
- *truncated power law*			
$_{min}$	3	3	11
$\hat{\alpha}$	1.500	2.610	3.100

[a]Significant parameter at: $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$

WoS: only Economics Statistics and Demography subfields are Small Worlds
CIS: overall and all subfields are Small Worlds and 4 out of 5 subfields (except Statistics) are consistent with a truncated power law
PRIN: Statistics for E&T Research and Demography (border values in CM model) are Small Worlds; Demography consistent with truncated power law

H2: partially confirmed. Fully confirmed for CIS *overall* and other subfields only
H3: confirmed. The effect of data sources on subfield is also confirmed

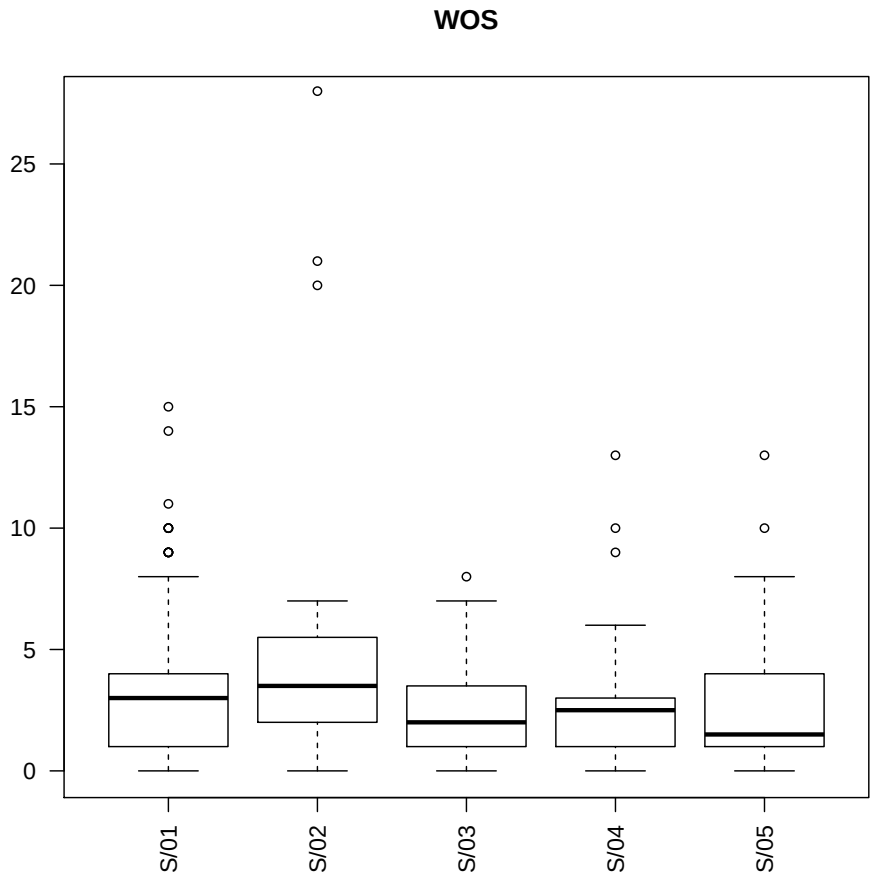
Co-authorship network positions and scientific performance

H4: The scientific performance of Italian statisticians is related to authors' position in co-authorship networks

- to model the effect of individual network position on scientific performance
- to use *h*-index as scientific performance measure
  - it is a bibliometric indicator which combines both productivity (# of publications) and impact (citations)
  - it is available for all members in our target population
- it also has some drawbacks:
  - the different productivity and citation practices of different disciplines
  - its value is dependent on the duration of each scientist's career (i.e., academic seniority)
  - its distribution is generally highly skewed and heavy tailed

h-index distribution per statistics subfields

Highly skewed and with zero inflation but very similar across subfields



... for these reasons we use a Generalized Extreme Value Distribution (GEV) and fit a single model for the overall statisticians network in the three sources

The model

- GEV is a family of distributions combining the Gumbel, Fréchet and Weibull:

$$F(z;\mu,\sigma,\xi) = exp\{-[1+\xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\}$$

  - μ = location parameter, σ = scale parameter, ξ = shape parameter (the higher ξ, the heavier the tail)

In detail, we model the h-index distribution (h) as: $h \sim GEV(\mu_i, \sigma_i, \xi_i)$

$$\mu_i = const_1 + \alpha_1 d_i + \alpha_2 c_i + \alpha_3 b_i + \alpha_4 \Gamma_i + \alpha_5 EI_i + \alpha_6 FP_i$$
$$\sigma_i = \sigma$$
$$\xi_i = const_2 + \beta_1 d_i + \beta_2 c_i + \beta_3 b_i + \beta_4 \Gamma_i + \beta_5 EI_i + \beta_6 FP_i$$

Influence of actor relational covariates on scientific performance (overall statisticians networks):

Degree (d_i)	$\Rightarrow$	number of different co-authors
Closeness (c_i)	$\Rightarrow$	Proximity in the network
Betweenness (b_i)	$\Rightarrow$	bridging otherwise disconnected groups of authors
Clustering Coef. (Γ_i)	$\Rightarrow$	propensity to work in "closed" subgroups
E-I index (EI_i)	$\Rightarrow$	proxy for propensity for intedisciplinarity
Dummy variable "Full Professorship" (FP_i)	$\Rightarrow$	proxy for academic seniority

GEV (final) model parameter estimates

Model on the overall statisticians networks

Parameters	**WoS**	**CIS**	**PRIN**
$const_1(\mu)$	2.14(0.07)***	1.99(0.09)***	1.95(0.08)***
α_1- Degree (d_i)	0.31(0.09)***	0.43(0.07)***	0.61(0.07)***
α_2- Clos. (c_i)	0.25(0.07)***	0.15(0.08)**	-
α_3- Bet. (b_i)	1.25(0.07)***	-	-
α_4- Γ (Γ_i)	-0.14(0.06)***	-	-0.14(0.06)***
α_5- E-I index (EI_i)	0.18(0.07)***	0.16(0.07)***	-
α_6- Full Professor (FP_i)	-	-0.31(0.14)***	-
σ	1.38(0.05)***	1.39(0.06)***	1.54(0.07)***
$const_2(\xi)$	0.06(0.03)**	0.08(0.05)*	0.16(0.04)***
β_1- Degree (d_i)	-	-	0.05(0.03)**
β_2- Clos. (c_i)	-	0.07(0.04)**	-
β_3- Bet. (b_i)	-	-	-
β_4- Γ (Γ_i)	-	-	0.07(0.04)**
β_5- E-I index (EI_i)	-	-	-0.06(0.04)*
β_6- Full Professor (FP_i)	-	0.15(0.09)**	-

[a]Significant parameter at: $*p < .1$, $**p < .05$, $***p < .01$.

Towards a combined co-authorship network for Italian statisticians...unique database

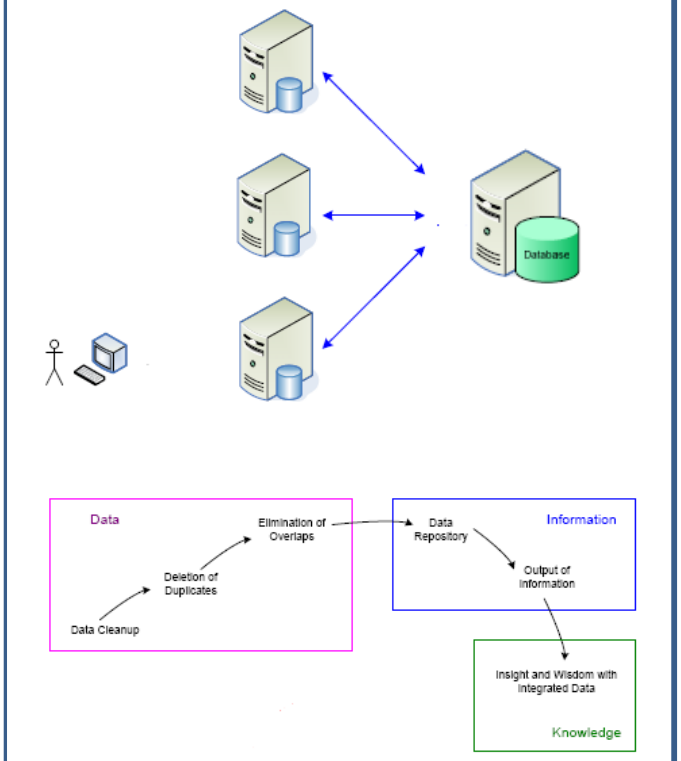
A possible strategy to obtain a combined archive from distinct data sources with the aim to achieve a better quality of co-authorship data in network analysis

Two main challenges have to be addressed:

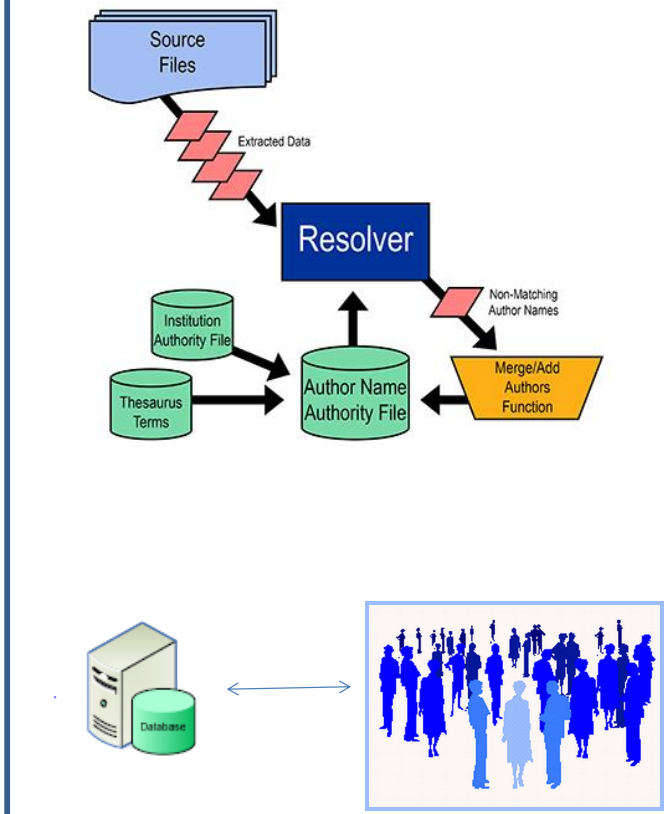
a) how to combine information from heterogeneous sources by identifying and linking duplicate records
b) how to deal with issues related to authors *synonyms* and *homonimies*

Towards a combined co-autorship network: unique publications archive

Record linkage



Author Name Disambiguation



Two-step procedure to merge data

To take advantage of data sources heterogeneity, we aimed at merging bibliographic data

A two-step procedure is used:

1. a semi-automatic method to pairwise match sources, by evaluating similarity of two records, through distance functions on key fields $\Rightarrow$ record linkage (RL)
2. author disambiguation through an unsupervised method the procedure following network-based approach $\Rightarrow$ author name disambiguation (AD)

Record linkage step

Record linkage: distance functions on key fields by pairs of sources

- Co-authors: *Jaccard* distance of authors' surnames of two publications (d_A)
- Title: error rate measure derived from the edit distance between two titles, $d_T = Ld(t_1, t_2)/max(|t_1|, |t_2|)$
- Year: absolute difference between years of publication (d_Y)

if $d_A < 10\%$, $d_T = 0$, and $d_Y = 0$: couples were automatically linked

Results

8735 publications by 677 statisticians and their co-authors

- Total author coverage rate: 85.5%
- Small overlapping of publications retrieved in the three data sources: 5.0%
- $>$ 40% of publications in PRIN, 24.6% in CIS and 13.0% in WoS: high heterogeneity of scientific production of Italian statisticians

Results after record linkage

8735 publications by 677 statisticians and their co-authors

Table: Number of linked records in the pairs of sources before reconciliation.

Sources	Matches	Possible matches	Finally linked records
(WOS, CIS)	782	71	827
(CIS, PRIN)	729	209	917
(PRIN, WOS)	612	166	756

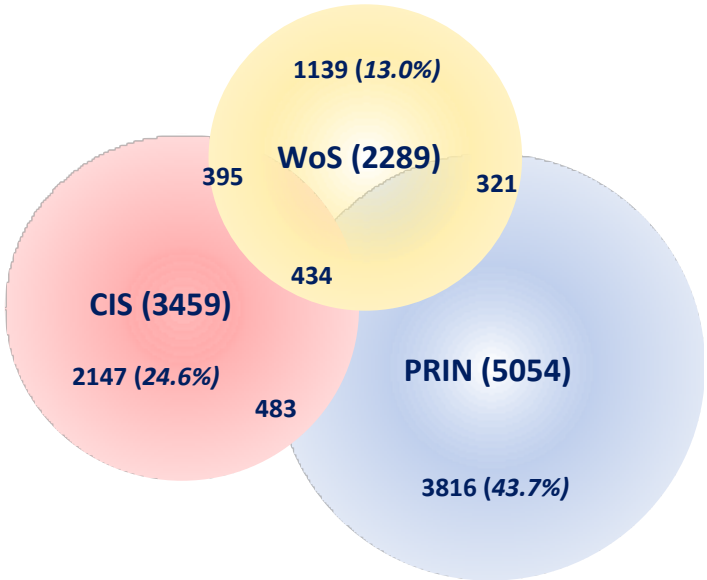


Figure: Number of publications after record linkage by archives

Author name disambiguation step

Unsupervised method: Strotmann *et al.* algorithm (2009)

- **graph-based** representation of author occurrences (vertices). Edges added when occurrences show some *evidence* of belonging to the same identity.

  Output: same identities are obtained by observing graph connected components

  - **Advantages**: it requires a restricted set of record attributes (*identifier*, *co-author names*)
  - **Limitations**: lack of misprints handling for compatibility checking and pessimistic behavior in merging identities (more identities than the real ones)

Our improvement of Strotmann *et al.* algorithm

- handling misspellings and double names/surnames
- enhanced use of record data to merge identities considering the title of the publication and the identifier of the query to retrieve records

Evidence measure

We calculated an evidence measure as:

$$0 \leq E = w_a \times e_a + w_v \times e_v + w_q \times e_q + w_t \times e_t \leq 1$$

Fuction	Values	Formula
Shared co-authors (e_a)	[0,1]	*Jaccard* coefficient on co-authors set
Publication venue (e_v)	{0,1}	Same venue = 1, 0 otherwise
Query ID (e_q)	{0,1}	Retrieved in the same query = 1, 0 otherwise
title keywords (e_t)	[0,1]	TF-IDF similarity between titles

weights w_a, w_v, w_q and w_t set up to .25

For each checked occurrence, the associated vertex is only connected to the vertex with the highest evidence E.

AD procedure: results and evaluation

Results

AD procedure returned a total of 7230 identities: 808 associated to statisticians

True positive (TP) – right identities returned by algorithm (489)
False positive (FP) – incorrect identities by merging separate authors (102)
False negative (FN) – incorrect identities by splitting unique author (112)

Evaluation of the AD procedure: evaluation metrics

Performance measures

precision (P), recall (R) and the harmonic mean of P and R metrics F_1

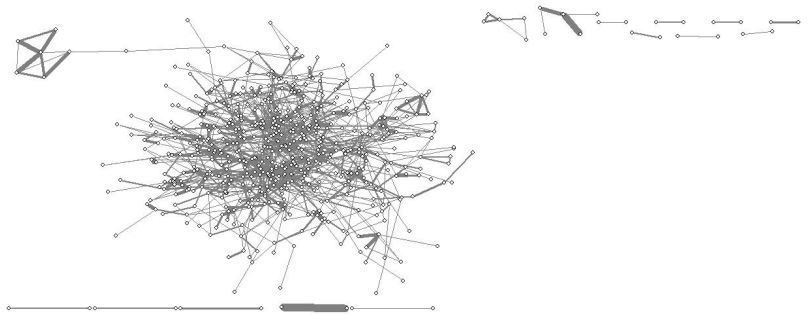
1. matching surnames and initials of all authors included in the target population with the identities returned by the algorithm
2. extracting a sample of statisticians (the 5% of 677 statisticians retrieved after RL) and furnishing the exact number of FP and FN

Metrics	Formula	Statisticians	Sample of Stats.	External authors
P	$\frac{TP}{TP+FP}$	.83	.93	.95
R	$\frac{TP}{TP+FN}$	.81	.85	.96
F_1	$\frac{2 \times P \times R}{P+R}$	.82	.89	.96

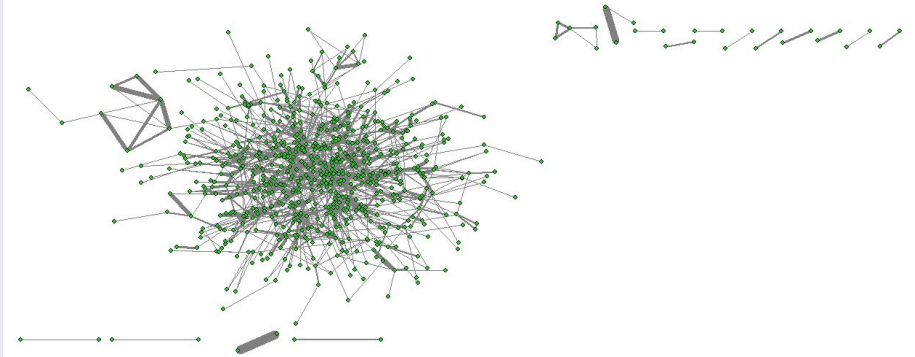
Values in line to the success rates of the original algorithm and quite comparable to the best results others have reported

Co-authorship network of statisticians (only internal links)

RECORD LINKAGE



AUTHOR NAME DISAMBIGUATION



Network statistics

	Stats –RL	*Stats –AD*		*Stats–RL*	*Stats –AD*
# authors	677	808	Largest distance	13	14
# isolated	92	121	Average Path Length	5.46	5.51
# edges	1197	1328	Clustering Coeff.	*0.39*	*0.39*
Density	0.005	0.004	# of components (size >1)	16	16
Average degree	3.54	3.29	Giant component (%)	81.24	80.82

Network-level analysis

RL_{NET} and AD_{NET} net statistics (all authors and statisticians only)

	RL	AD		RL	AD
All authors					
# authors	7332	7230	Largest distance	14	16
# isolated	42	31	Average Path Length	5.29	5.17
# edges	474478	424545	Clustering Coeff.	0.88	0.91
Density	0.018	0.008	# of components (>1 node)	35	58
Average degree	129.43	117.44	Giant component (%)	97.64	95.59
Statisticians					
# authors	677	808	Largest distance	13	14
# isolated	92	116	Average Path Length	5.46	5.53
# edges	1197	1346	Clustering Coeff.	0.26	0.24
Density	0.005	0.003	# of components (>1 node)	16	15
Average degree	3.54	3.33	Giant component (%)	81.24	81.68

Two main interacting effects are at work in shaping the network structures.
For all authors:

- merging of identities (# of authors and # of links both lower in the case of AD)
- splitting of identities (reduction of isolates and increasing of components)

For statisticians: main effect of

- splitting of identities (higher # of nodes and edges with an increase of isolates - exclusion of external authors who cannot connect couples of statisticians

...IRIS bibliographic archive

To reconstruct the co-authorship network among the Italian academic statisticians belonging to five subfields → target population

- using a reliable and updated database with all kind of publications in Italy → IRIS
- analysing the changes of collaborative behaviors before and after the national evaluation of research quality (VQR 2011-2014, ANVUR) → policy implications

Previous analysis

- Multiple bibliographic archives (De Stefano et al., 2013) and their integration to obtain a unique co-authorship network (Fuccella et al., 2016)
  - Three bibliographic archives → Web of Science and Current Index to Statistics– and a national archive based on publications attached to the nationally funded grants (PRIN projects)

Current analysis → IRIS platform https://www.cineca.it/it/content/

- Institutional Research Information System (IRIS) in Italy developed by Cineca consortium

  → a database to store and manage research products

  → used by most of the Italian universities

IRIS: bibliographic archive in Italy

INSTITUTIONAL RESEARCH INFORMATION SYSTEM

RM
(Resource Management)
Gestione Risorse della Ricerca

AP
(Activities and Projects)
Attività e Progetti Scientifici

IR / OA
(Institutional Repository)
Archivio aperto della Ricerca

DSPACE

ER
(Evaluation and Review)
Valutazione della Ricerca

ES
(Expertise and Skills)
Competenze della Ricerca