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., Fuccella, V., Vitale, M. P., Zaccarin, S.: Co-authorship Ties in a Target Population: Data Quality Issues and Network Analysis (....)

- 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 $\rightarrow$ key element in knowledge advancement for interactions and sharing competences among scholars
- Co-authorship relationship $\rightarrow$ proxy of scholars' collaborative behaviors

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

- to explore topological properties $\rightarrow$ small world, scale-free configurations
- to discover clusters $\rightarrow$ 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)

New bibliographic archive

- Data quality issues
- First results

Iris bibliographic archive

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

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

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

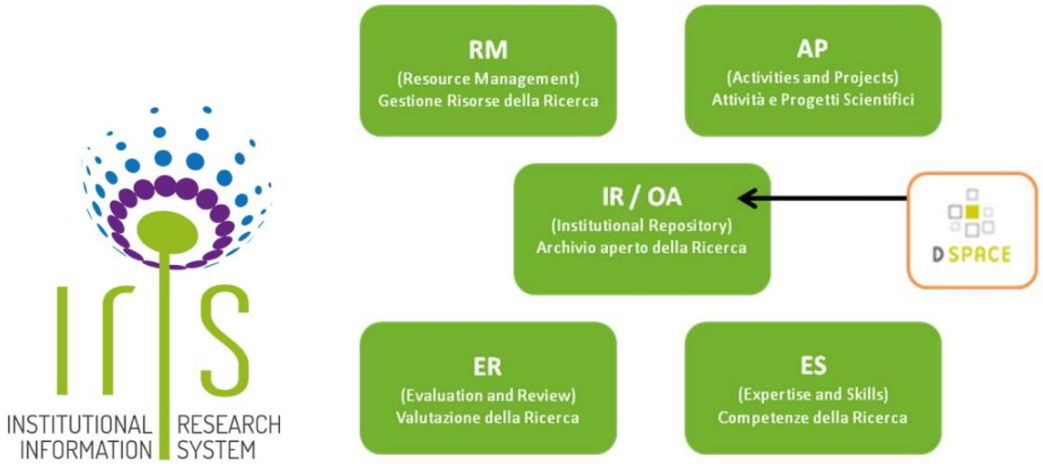
National Registry of Academic scholars - ANPREPS

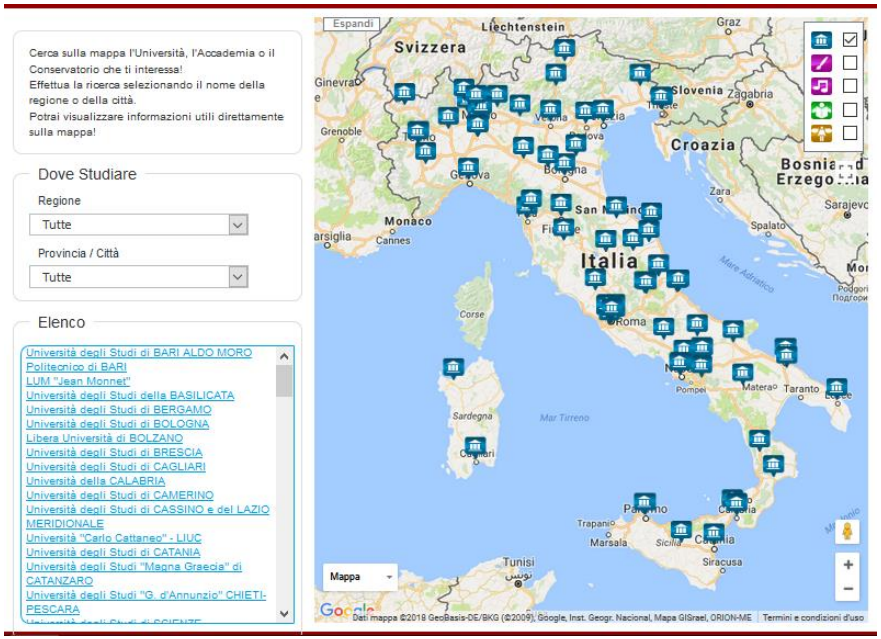
Law n.1, January 2009 by MIUR:

- creation of a national registry of academic scholars (full and associate professors, lectures) with the list of their scientific publications yearly updated

- eleven years after the registry is not yet available (renewed attention in the recent period)
- general consensus on:
  - registry characteristics $\rightarrow$ openness, accessibility, connection with European/other database, flexibility of research product definition
  - high data quality $\rightarrow$ duplications, missing data and errors $\rightarrow$ certified data validation
  - the potential of the IRIS platform to be used as the registry base

IRIS: bibliographic archive in Italy





Pros and Cons in using IRIS platform

PROS

- Good coverage of Italian universities (66 out of 96 have installed IRIS platform)
- Extraction of all kinds of publications for a target population
- Complete bibliographic data updated by scholars
- Availability of longitudinal bibliographic data
- Publications' coverage near to the individual scientific CVs

CONS

- IRIS heterogeneous systems
  $\rightarrow$ each institution hosts a different system deployment and
- Few mandatory and fixed fields:
  $\rightarrow$ very different data format
- Duplicated records of publications co-authored by scholars belonging to different universities
  $\rightarrow$ No standard procedure to insert some key fields (names of external authors) for the duplication and co-authorship network recognition

Web scraping techniques for extract IRIS bibliographic data (De Stefano et al., 2018, 2019)

Complexities related to scraping data process (Murthy, 2013; Vargiu, 2013; Mitchell, 2015) and data mining tools

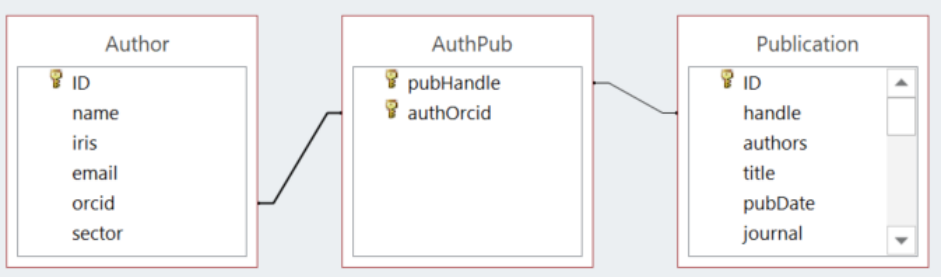
- Data retrieval phase: $\rightarrow$ semi-automated tool to extract the author and publication metadata from IRIS systems
- Data cleaning phase: *Record linkage* of publications and *Author name disambiguation*

Data retrieval and Data cleaning phases

Steps of data retrieval

Step 1. URL of IRIS page retrieved for each author

Step 2. IRIS publication metadata retrieved for each author

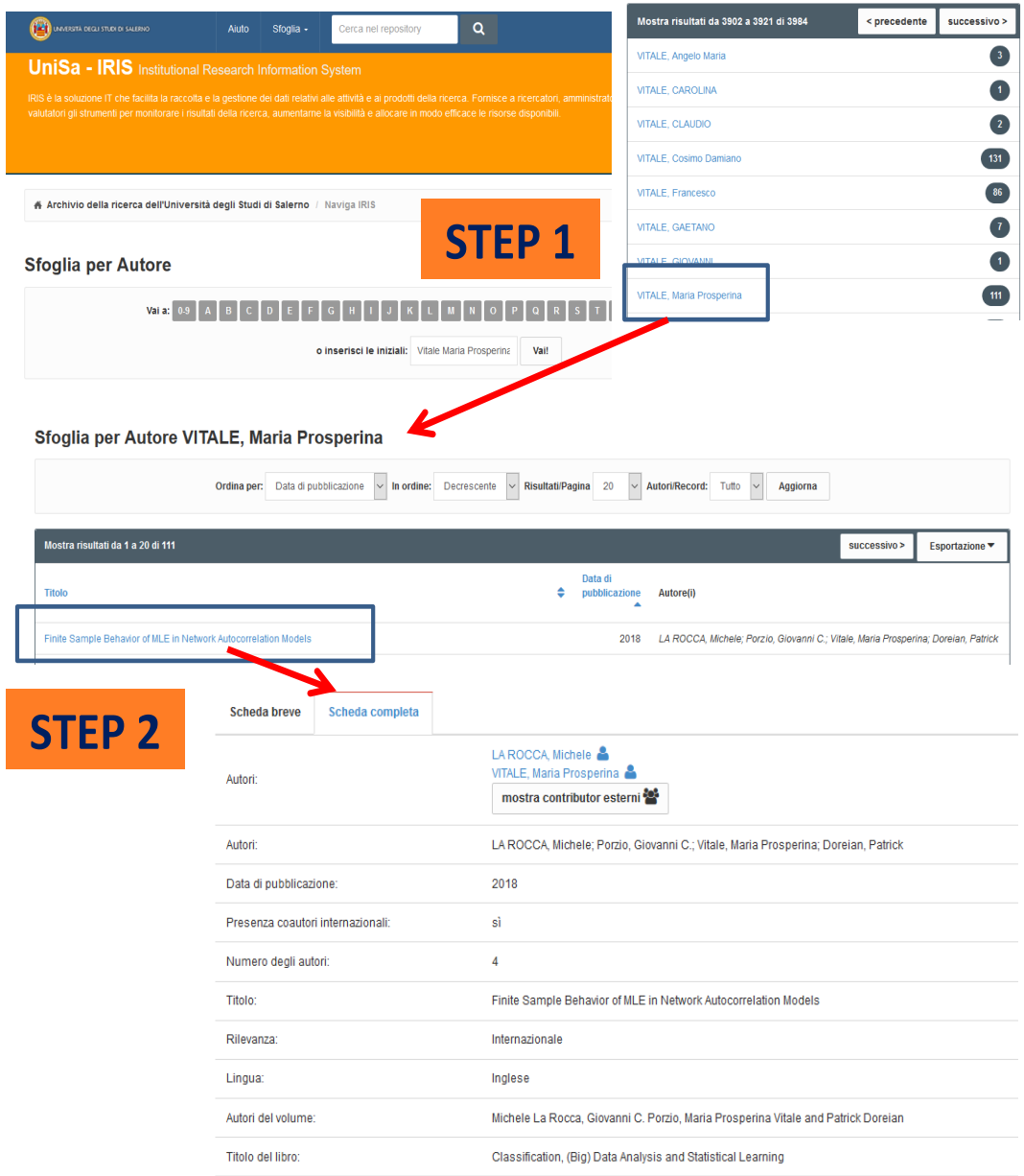


Steps of Data cleaning

Step 3. Publication record linkage ⇒ to combine publications from various IRIS systems by *identifying* and *linking* duplicate records

Step 4. Network-based author name disambiguation ⇒ to deal with author *synonyms* and *homonyms* issues

IRIS platform layout



Web scraping techniques to extract IRIS bibliographic data

Data retrieval: $\rightarrow$ semi-automated tool to extract the author and publication metadata from IRIS systems

- Tool implemented in Java by using standard libraries to download author and publication web pages
- Tagsoup library for parsing even unstructured and malformed HTML

Technical details

- URL of IRIS page is retrieved for each author $\Rightarrow$ author last name is used as a query string. In case of no match or multiple matches, the procedure returned an error.
- Complete database of publications records for each author $\Rightarrow$ each publication is associated to a link to a new page containing publication details (title, authors, venue, year and various identifiers –URL, DOI, ISI codes WoS and Scopus, etc).

Population coverage

Statisticians (MIUR database, July 2017) found and publications by Statistics subfields

- tool returned 0 publications for some authors ⇒ absence of the IRIS platform (mainly for private and online universities) and limitations to publication record public access (e.g., Bologna). Errors happened also when the author search returned more than one match.

	Authors			Publications		
Subfields	Found	Manual check	% found	Min.	Max.	Median
All (721)	555	82	88.3	1	333	50
Stat (421)	319	54	88.6	1	292	49
Stat for E&T (20)	18	2	100.0	8	126	46
Econ. Stat (145)	110	11	83.4	2	196	42
Demo (70)	55	8	90.0	12	314	55
Social Stat (65)	53	7	92.3	8	333	68

Results

- Good coverage of the target population ⇒ metadata for around 80% of statisticians after data extraction in April 2018
- Manual check to integrate the retrieved metadata ⇒ coverage rates up to 90% for some subfields
- Results in line with the authors' coverage obtained by using three different bibliographic archives (De Stefano et al., 2013)

Recovering duplicated records

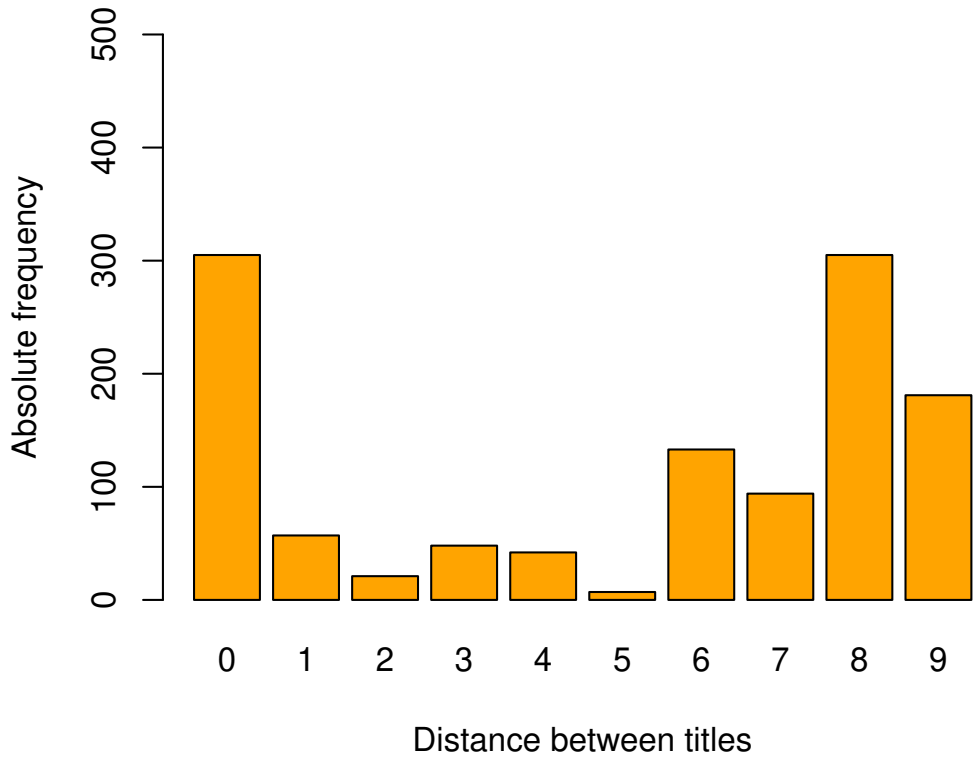
- Duplicated publication records $\Rightarrow$ in principle, # of duplications equal to the number of the coauthors hired by different universities
- to reconstruct a reliable co-authorship network, duplicated records have to be *solved* before author disambiguation step

Some results for Social Statisticians (# 60):

- # of total found publications (with duplicates) $\Rightarrow$ 3366
- mandatory fields in each IRIS: **author(s)**, **title**, **publication venue (journal/publisher)** and **year of publication**
  - inconsistencies due to the different format employed in each IRIS
  - *Edit distance* on title strings in pairs ($= 0$ if two strings are equal, $= 1$ if two strings differ just for one character, and so on)

First results - Social Statisticians/1

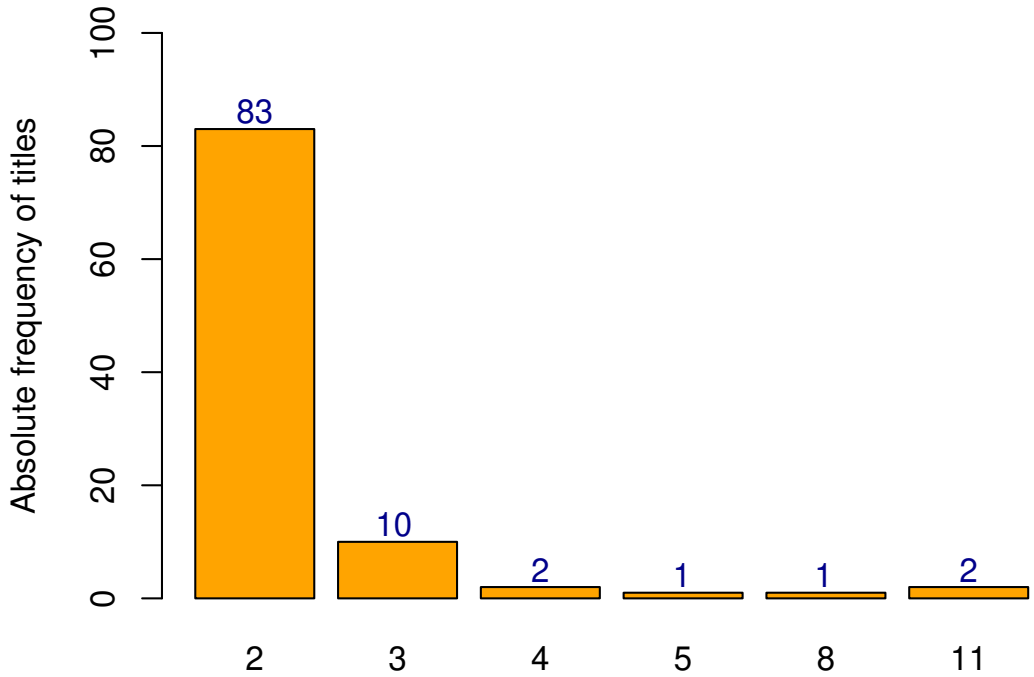
Distribution of distances between titles up to 9



- 9% of pairs have distance on titles = 0 (but can have different venues and years of publications)

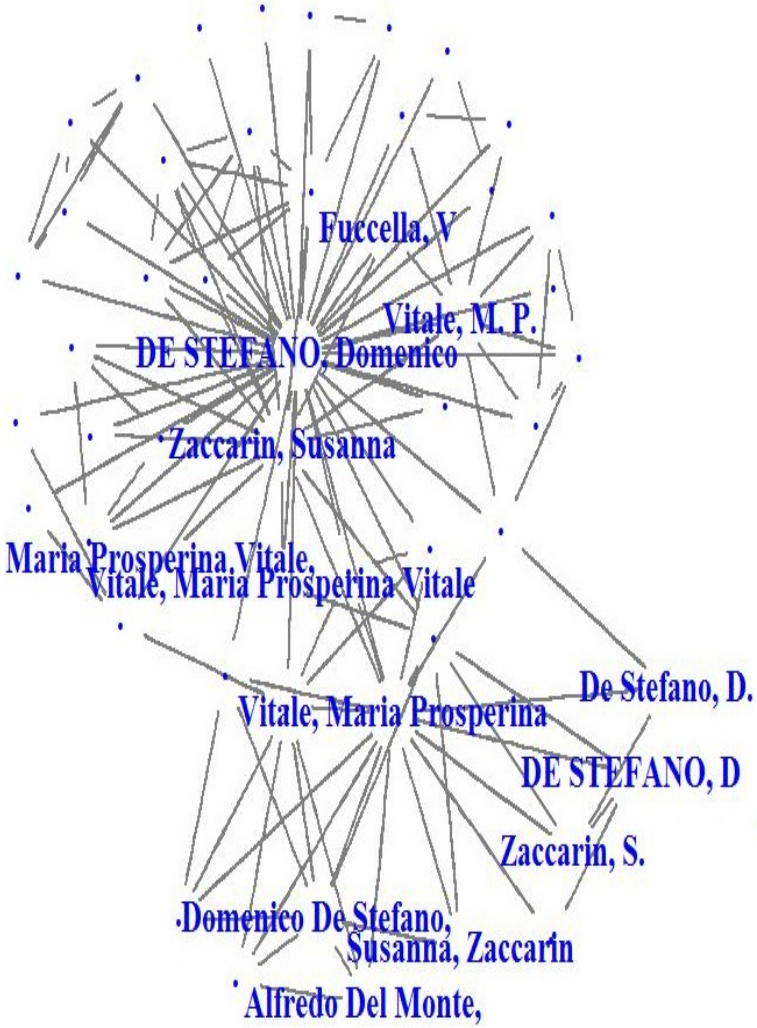
First results - Social Statisticians/2

Distribution of title duplicates

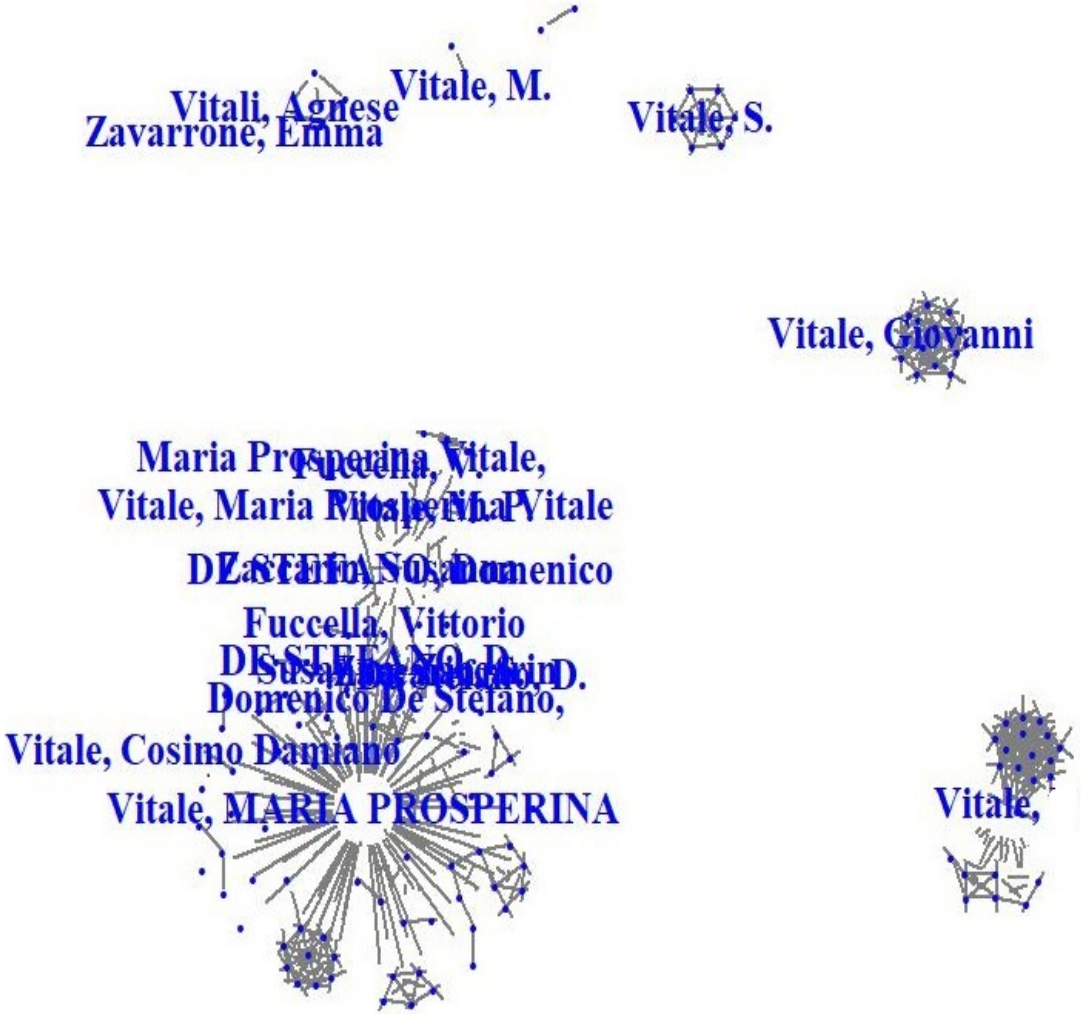


- 83 titles are double, 10 are triplets, few have more replicates
- high numbers of replicates: evidence of non-informative titles (e.g. *Conclusions*, *Introduction*)
- evidence of a low propensity of coauthoring with Social Statisticians hired in other universities (in line with previous results from different data sources)

Ego-centred co-authorship network: example 1



De Stasio, (S)	{De Stasio, (S)=1}	https://art.torvergata.it/handle/2108/158207
De Stefano, Domenico (D)	{De Stefano, Domenico (D)=76, Domenico De Stefano, ()=7}	https://arts.units.it/handle/11368/2877781
De Togni, Aldo (A)	{De Togni, Aldo (A)=1}	https://iris.unipv.it/handle/11571/27216



Vitale, Maria Prosperina (MP)	{Vitale, Maria Prosperina (MP)=36, Vitale, Maria Prosperina Vitale (MPV)=1, Maria Prosperina Vitale, ()=1}
Vitale, (M)	{Vitale, (M) = 1}
Vitale, ()	{Vitale, () = 1}

Work in progress

Open issues affecting co-authorship data quality in IRIS

- publication identifiers (e.g., title, publication year, venue, etc.) could be not reliable, especially for publications in non-indexed journal (in WoS, Scopus)
- no automatic procedure that allows to match the same publication coauthored by authors enrolled in different institutions. Therefore for each co-authored publication, a number of duplication of the same product equal to the number of the coauthors hired by different universities can be found.
- no standard procedure to insert some key fields for the co-authorship network recognition (e.g., institutional external co-authors)
- both product and author name duplications should be addressed before the co-authorship network construction by adapting the procedure proposed in Fuccella et al. (2016)

Work in progress

- Extending duplicated record analysis to other scientific fields
- Improving current author name disambiguation to include useful information on same author identities detected by the duplicated records
- Analysing the provisional network and its relational information to recover the same identities $\Rightarrow$ community detection algorithms

Next project

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca — anvur — National Agency for the Evaluation of Universities and Research Institutes

III Concorso Pubblico Idee di ricerca (bando dell'11 maggio 2017)

Elenco dei beneficiari (ai sensi della delibera del Consiglio Direttivo n. 70 del 18 aprile 2018)

RESEARCH PROJECT 2018-2019

SNEval - Social Network tools for the evaluation of individual and group scientific performance

Research Group: Domenico De Stefano (Coordinator); Luka Kronegger, Valerio Leone Sciabolazza, Maria Prosperina Vitale, Susanna Zaccarin

Publications submitted for the Italian research assessment exercise

VQR 2011–2014

Thanks

Thank you for your attention!