A novel methodology to disambiguate organization names: A key step toward a dynamic analysis of EU Framework Programmes

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Andrea Ancona*
Department of Social Sciences and Economics
Sapienza University of Rome

Roy Cerqueti
Department of Social Sciences and Economics
Sapienza University of Rome

Gianluca Vagnani
Department of Management
Sapienza University of Rome

*Presenting author

SAPIENZA
UNIVERSITÀ DI ROMA

Presentation overview

First part

Scientometrics (2023) 128:4447–4474
https://doi.org/10.1007/s11192-023-04746-x

Check for updates

A novel methodology to disambiguate organization names: an application to EU Framework Programmes data

Andrea Ancona[1] · Roy Cerqueti[1,2] · Gianluca Vagnani[3]

Received: 1 October 2022 / Accepted: 10 May 2023 / Published online: 31 May 2023

Abstract
The concept of collaborative R&D has been increasing interest among scholars and policy-makers, making collaboration a pivotal determinant to innovate nowadays. The availability of reliable data is a necessary condition to obtain valuable results. Specifically, in a collaborative environment, we must avoid mistaken identities among organizations. In many datasets, indeed, the same organization can appear in a non-univocal way. Thus its information is shared among multiple entities. In this work, we propose a novel methodology to disambiguate organization names. In particular, we combine supervised and unsupervised techniques to design a "hybrid" methodology that is neither fully automated nor completely manual, and easy to adapt to many different datasets. Thus, the flexibility and potential scalability of the methodology make this paper a worthwhile contribution to different research fields. We provide an empirical application of the methodology to the dataset of participants in projects funded by the first three European Framework Programmes. This choice is because we can test the quality of our procedure by comparing the refined dataset it returns to a well-recognized benchmark (i.e., the EUPRO database) in terms of the connection structure of the collaborative networks. Our results show the advantages of our approach based on the quality of the obtained dataset, and the efficiency of the designed methodology, leaving space for the integration of affiliation hierarchies in the future.

Keywords Organization name disambiguation · Hybrid methodology · Institutions · Labels · Collaborative networks · EU Framework Programmes

- We propose a novel methodology to disambiguate organization names
- "Hybrid" methodology easy to adapt to many different datasets.
- We apply it to the datasets of participants in projects funded by the first three EU Framework Programmes (FPs)
- Our results show the quality of the obtained datasets, and the efficiency of the methodology

Presentation overview

Second part

Uncovering collaborative patterns and transition dynamics in EU Framework Programmes through network modeling

Andrea Ancona[1], Roy Cerqueti[1], Gianluca Vagnani[2]

[1]Department of Social Sciences and Economics
Sapienza University of Rome
[2]Department of Management
Sapienza University of Rome

1 Introduction

A systematic European Research and Technological Development (RTD) policy was established in the 1980s alongside the first European Framework Programme (EU FP). The RTD policy aligns with the broader cohesion policy, which envisions the EU as a common market promoting the free circulation of people, goods, capital, and knowledge exchange. The main goal of the EU FPs is to provide funds mainly to European member states, but also to associate and third countries, in order to promote international research collaboration both at the individual and at the organization level.

Supporting collaborative R&D projects has become increasingly relevant for policy-makers and institutions. Funding is a crucial means to enhance research productivity, even more than collaboration itself. The effects of collaboration indeed, become particularly relevant in the post-funding period, suggesting that it takes time to develop effective collaboration structures that have an impact on research productivity (Defazio et al., 2009). At the European level, collaborations between knowledge-intensive and lagging-behind regions positively affect the innovation capability of the latter ones, demonstrating that a greater openness in the collaboration networks fosters knowledge exchange and spillovers (De Noni et al., 2018). As a consequence, knowledge convergence is gradually emerging among NUTS 2 regions over time, creating an opportunity to integrate knowledge cohesion with the social and economic pillars of the EU cohesion policy (Erdil et al., 2022).

- We analyze the participation dynamics in collaborative projects from FP1 to H2020 through Social Network Analysis (SNA).
- We statistically assess the Markovian nature of the collaboration process.
- We estimate the probability of moving from one level of centrality to another over consective FPs.
- Our results show a quasi-Markovian nature of the process, and shed light on the effectiveness of EU research policies.

A novel methodology to disambiguate organization names: An application to EU Framework Programmes data

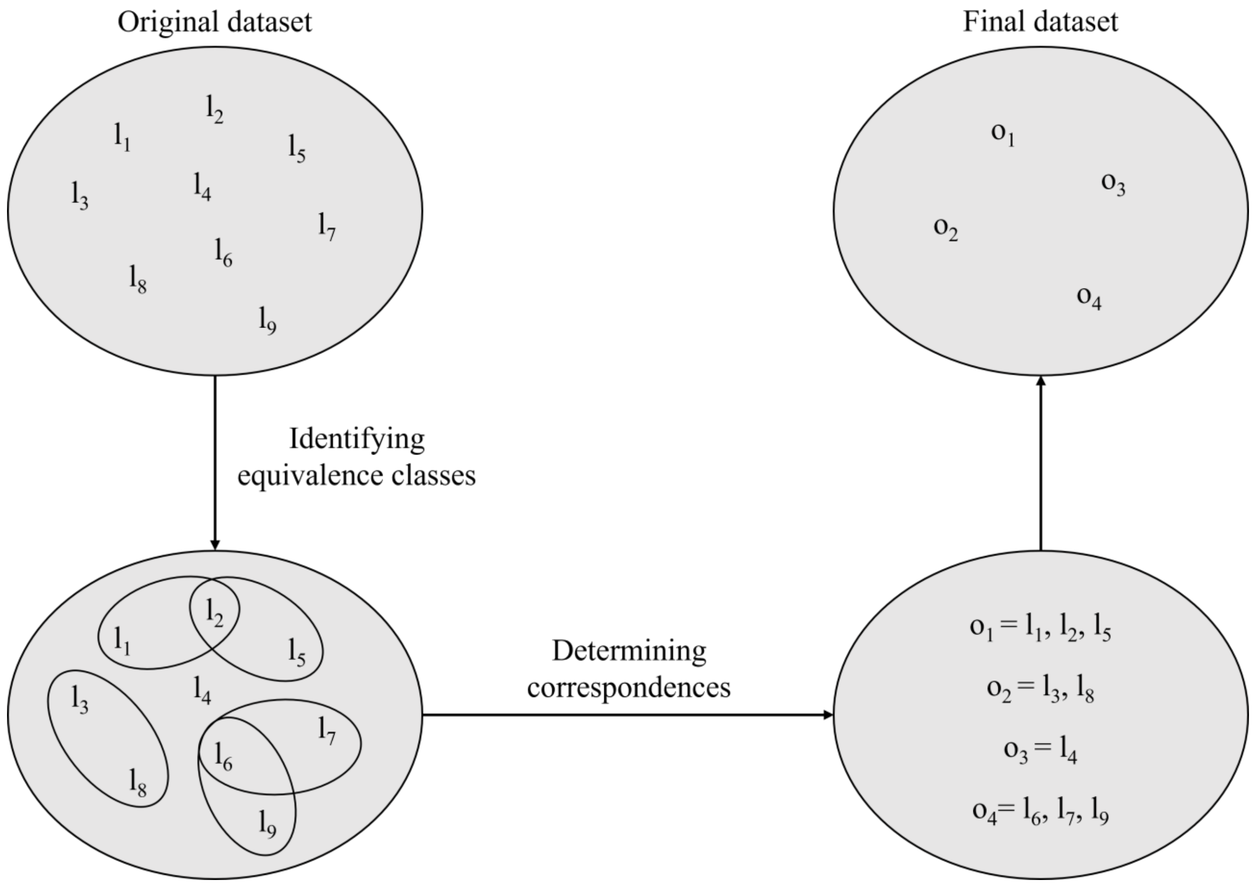
SAPIENZA
UNIVERSITÀ DI ROMA

Research motivation

- The **availability** of **reliable data** is necessary to obtain **valuable** results.
- In **collaborative** environments, we must avoid the **attribution** of **wrong information** to **entities**, and the **aggregation** of **information** related to **distinct entities**.
- Various **reasons** leading to **different labels** to address the **same organization**:
  - **Different languages** (e.g., “Sapienza University of Rome” and “Università degli Studi di Roma La Sapienza”)
  - **Abbreviations** (e.g., “Università degli Studi di Roma La Sapienza” and “Univ di Roma La Sapienza”)
  - **Acronyms** (e.g., “Consiglio Nazionale delle Ricerche (CNR)” and “CNR”)
  - **Punctuation** (e.g., “CEN/SCK” and “CEN – S.C.K.”)
  - **Periphrases** (e.g., “University of Oxford” and “The Chancellor, Masters and Scholars of the Univ of Oxford”)
  - **Linguistic equivalences** (e.g., “University of Aarhus” and “Aarhus University”)
  - **Misspellings** (e.g., “Telefonica Investigacion y Desarfolio” and “Telefonica Investigacion y Desarollo”)

SAPIENZA UNIVERSITÀ DI ROMA

Our approach



1. Data **pre-treatment** through certified lists of organizations (e.g., ROR and OrgReg).
2. Thorough **pre-processing** of **labels** to replace **acronyms**, remove **stopping words**, and include **keywords**.
3. Efficient **automated** part based on **common words, consecutive common characters, cosine similarity**, and **"control" variables**.
4. Final **manual inspection** to disambiguate **borderline** cases.

SAPIENZA UNIVERSITÀ DI ROMA

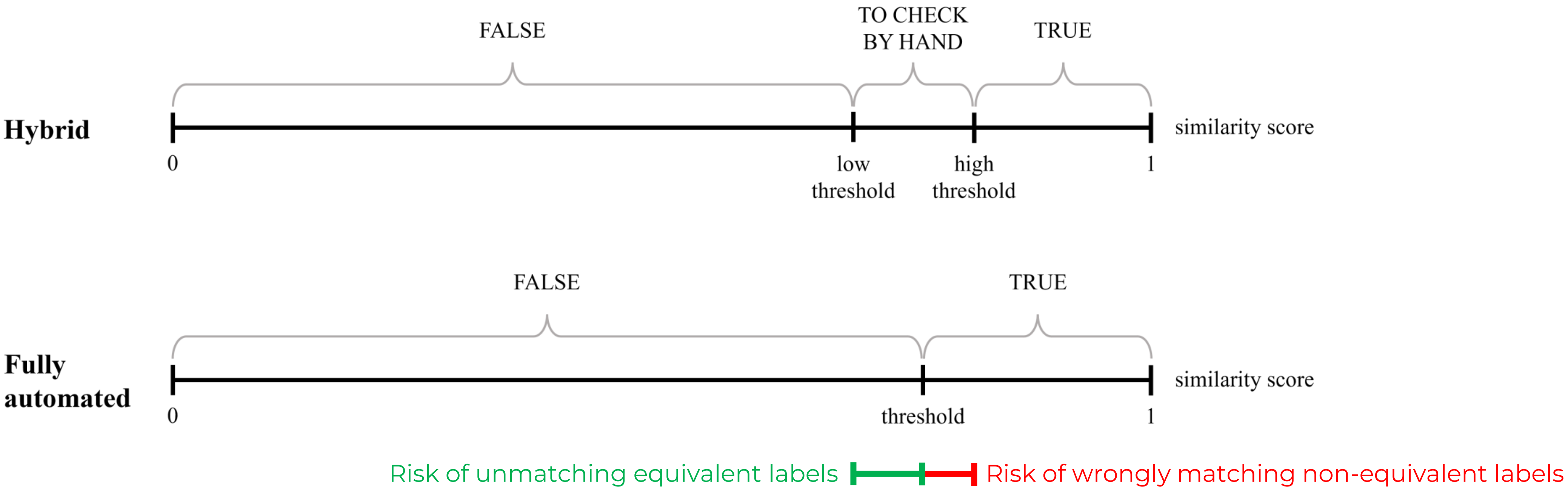
Main similarities and differences with existing approaches

Table 2 Main similarities and differences with key references and services

Reference	Similarities	Differences
Jonnalagadda and Topham (2010).	(1) Pre-processing of institution names by removing blocking words and special characters. (2) Combination of automated methods with manual check.	(1) Domain-specific (i.e., PubMed abstracts). (2) Similarity based on Levenshtein distance.
Huang et al. (2014).	(1) Rule-based approach based on verified conditions.	(1) Similarity based on Jaccard distance and Jaro-Winkler algorithm. (2) Fully automated method.
Zhang et al. (2012), Muñoz et al. (2012), Spina et al. (2013).	No particular similarities.	(1) Fully automated methods. (2) Dependent on Twitter information and external web sources.
Jiang et al. (2011).	(1) Pairwise comparison between all affiliations.	(1) Similarity based on the Normalized Compression Distance. (2) Fully automated method.
Cuxac et al. (2013).	(1) Comparison with manually analyzed reference datasets through a supervised approach.	(1) Bayesian techniques. (2) Discouraged for highly unbalanced data.
Service	**Similarities**	**Differences**
ROR OpenRefine Reconciler	(1) Input table with a column related to organization names. (2) Matchings based on string similarity.	(1) No control variables required. (2) Relying on ROR records. (3) Disambiguating companies with a status of active only. (4) Manual check of all returned cases. (5) Discouraged with a high number of organization names.
AIDA System (Yosef et al., 2011)	(1) Anchoring to keyphrases (similar role as control variables).	(1) Greedy algorithm following a graph-based approach. (2) Dependent on a context (i.e., an input text). (3) Fully automated method.

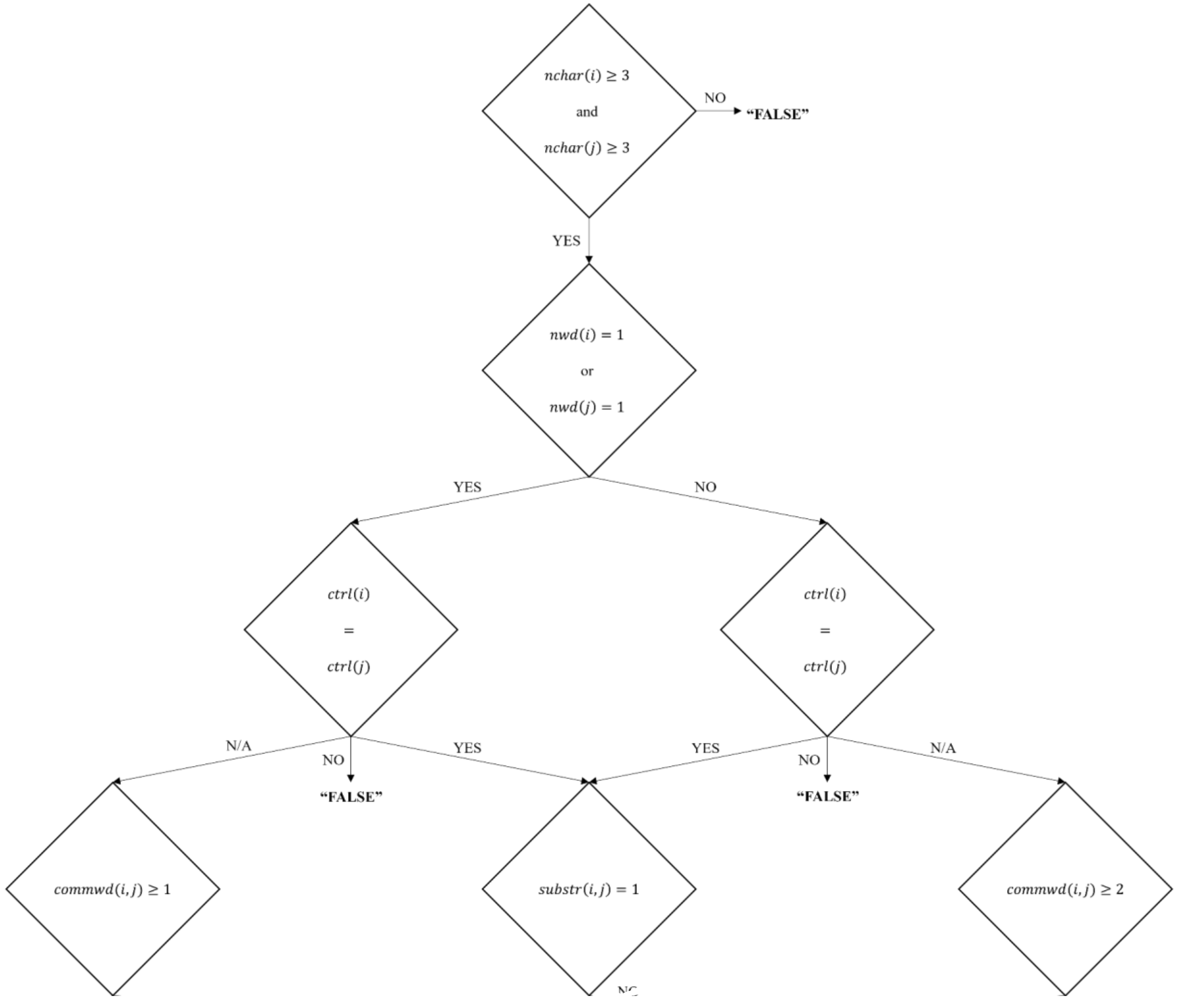
SAPIENZA UNIVERSITÀ DI ROMA

Why a «hybrid» methodology is necessary



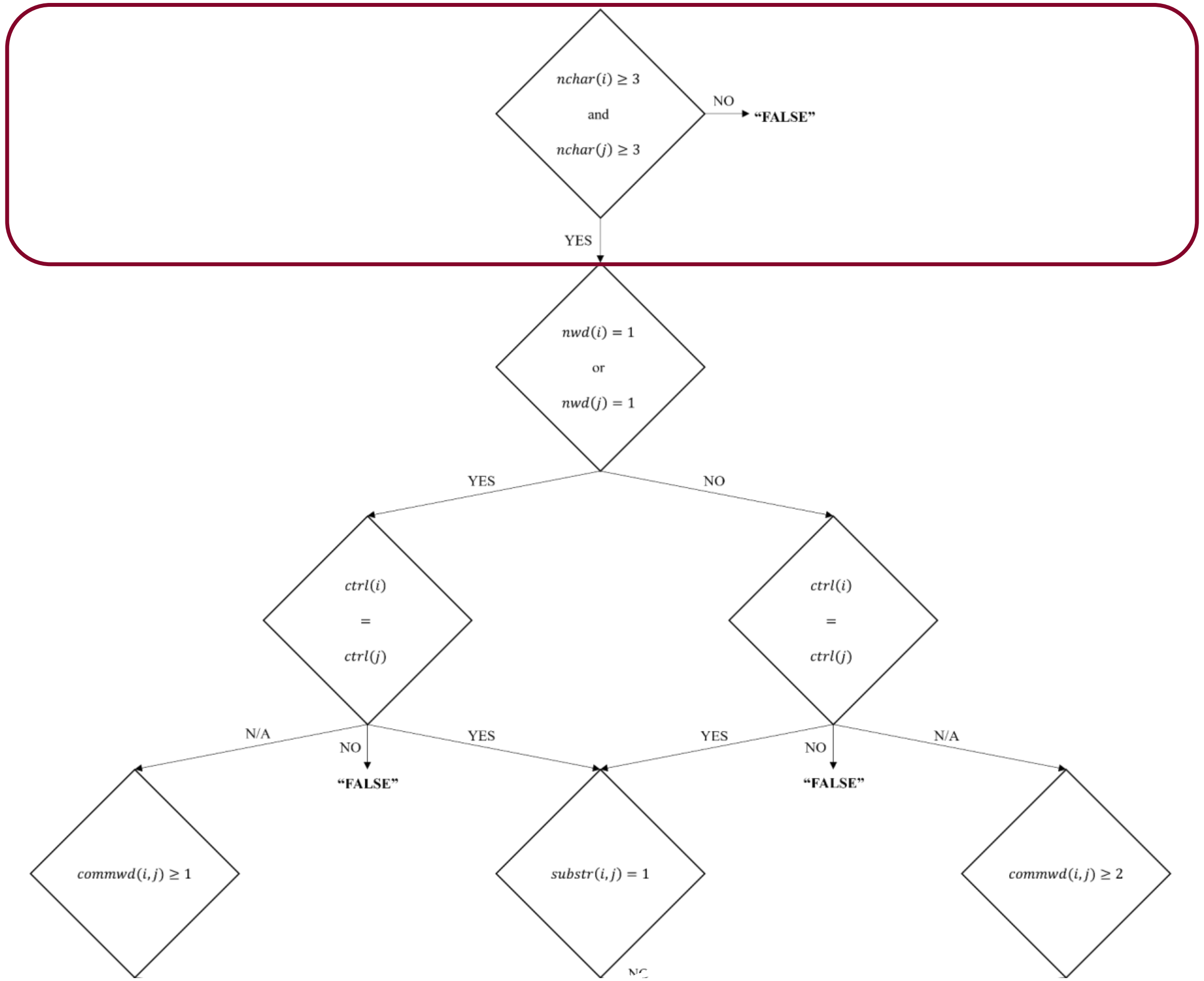
SAPIENZA UNIVERSITÀ DI ROMA

The algorithm



SAPIENZA UNIVERSITÀ DI ROMA

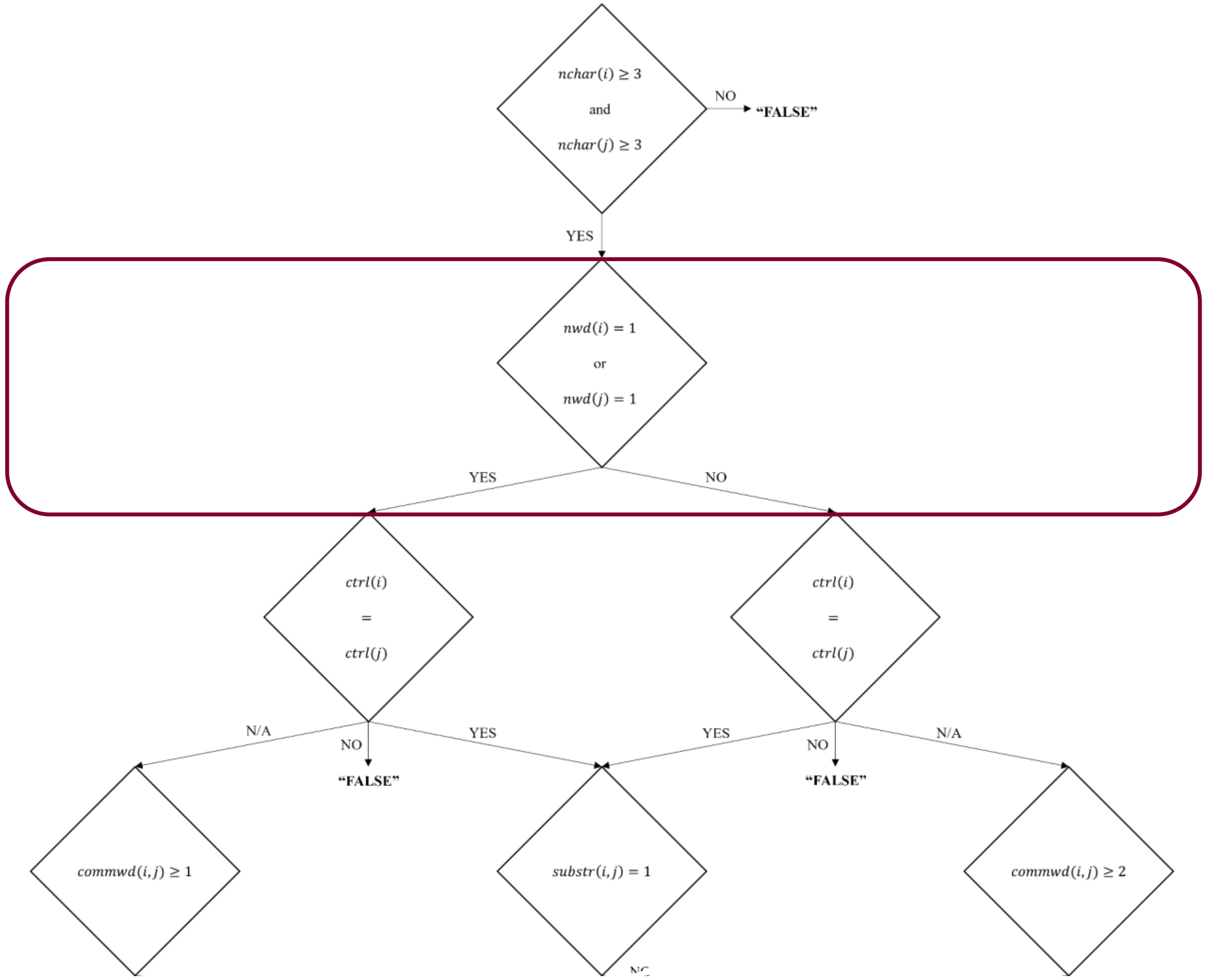
The algorithm



Step 1. Test on the number of characters

SAPIENZA UNIVERSITÀ DI ROMA

The algorithm

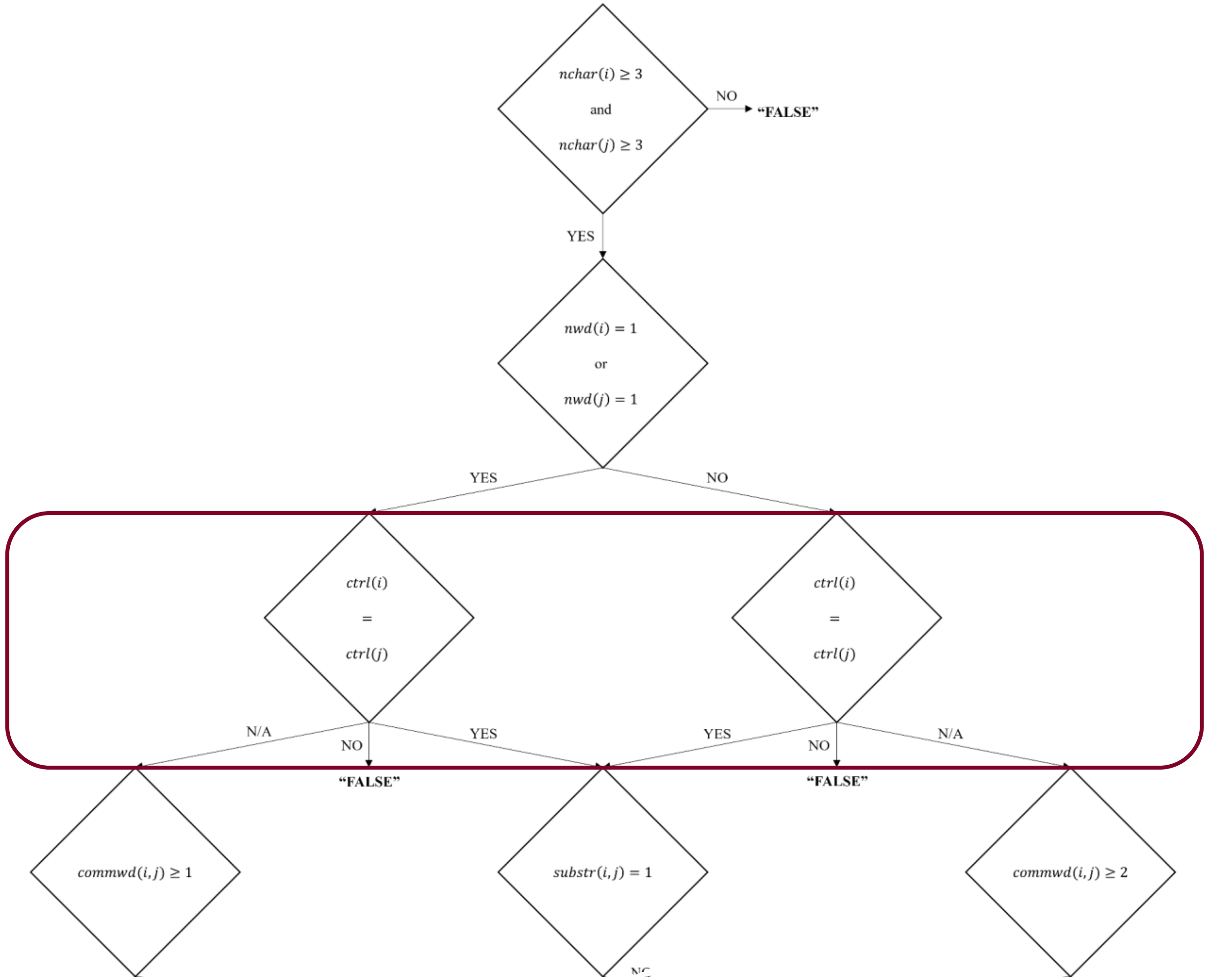


Step 1. Test on the number of characters

Step 2. Test on the number of words

SAPIENZA UNIVERSITÀ DI ROMA

The algorithm



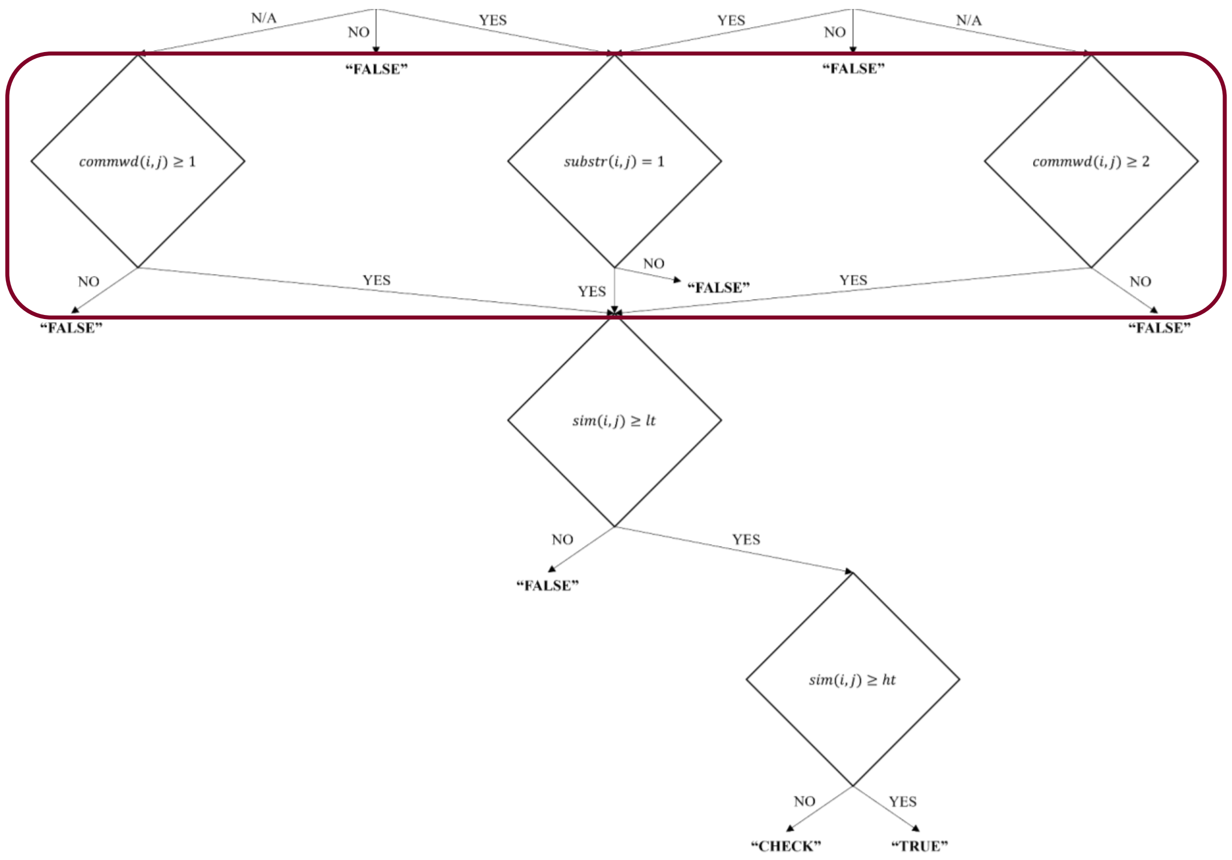
Step 1. Test on the number of characters

Step 2. Test on the number of words

Step 3. Test on the control variables

SAPIENZA UNIVERSITÀ DI ROMA

The algorithm



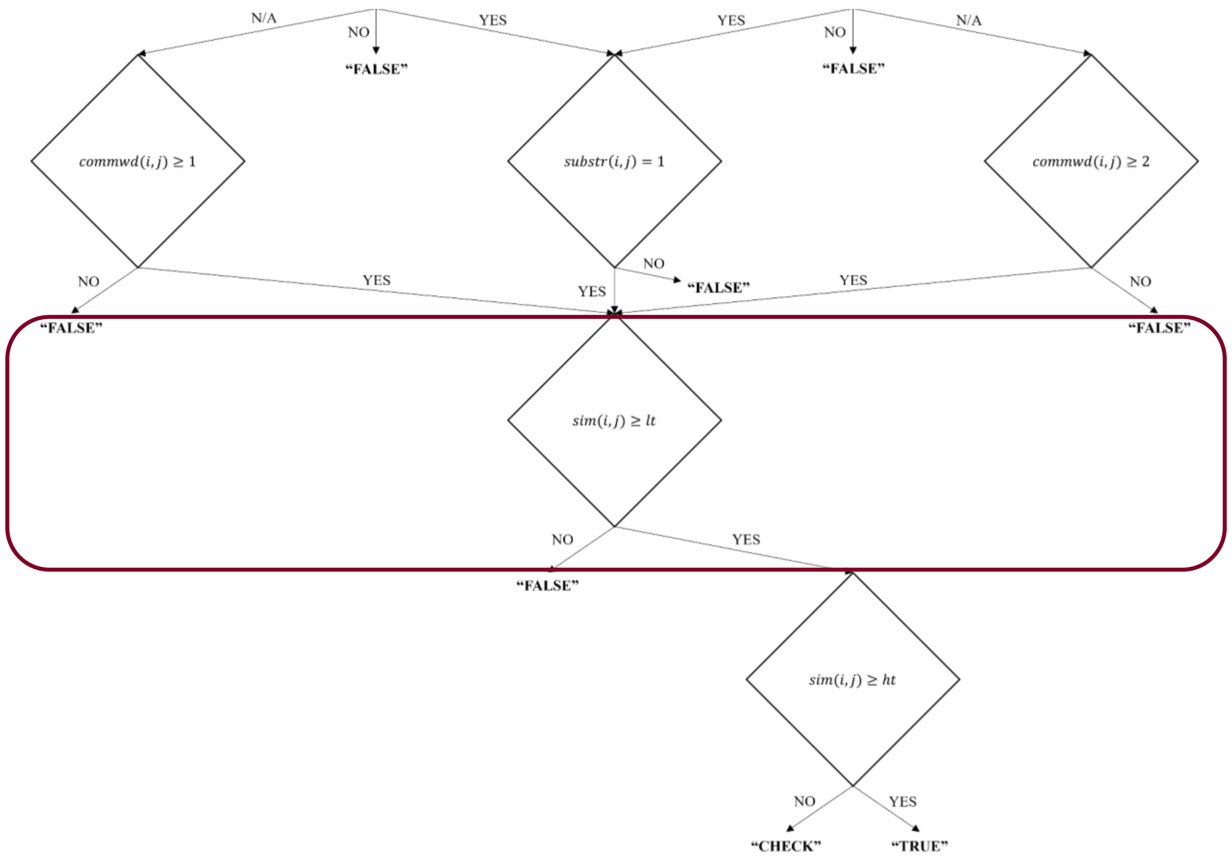
Step 1. Test on the number of characters

Step 2. Test on the number of words

Step 3. Test on the control variables

Step 4. Test on the number of common words or consecutive common characters

The algorithm



Step 1. Test on the number of characters

Step 2. Test on the number of words

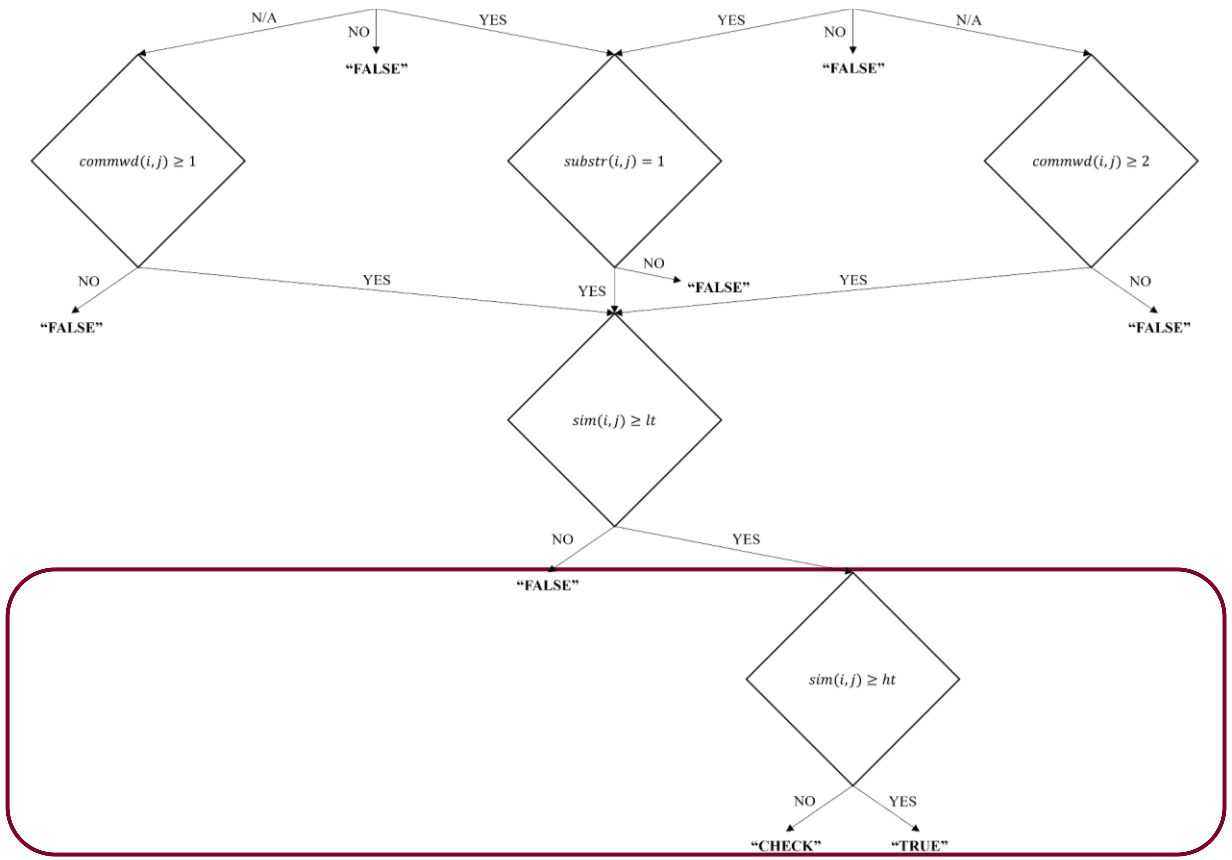
Step 3. Test on the control variables

Step 4. Test on the number of common words or consecutive common characters

Step 5. First test on the similarity score

SAPIENZA UNIVERSITÀ DI ROMA

The algorithm



Step 1. Test on the number of characters

Step 2. Test on the number of words

Step 3. Test on the control variables

Step 4. Test on the number of common words or consecutive common characters

Step 5. First test on the similarity score

Step 6. Second test on the similarity score

The first three EU FPs

- One of the most **relevant case studies** in the field of **collaborative R&D**.
- **Data** are **publicly available** on the **CORDIS** website.
- Datasets on the **first three EU FPs** are the **most unbalanced** and the **less standardized** among all FPs.
- We can rely on a **high quality dataset** as a **benchmark** to assess the **efficiency** of our **methodology**, i.e., the **EUPRO database** (Roediger-Schluga & Barber, 2008).
- We dowloaded **CORDIS data** on **October 1st, 2021**; at the same time, we requested and obtained access to the **EUPRO database**.
- After **removing** rows with **no organization names**, we obtained the **final sample**:
  - **Raw data**: 7,900 participations in FP1; 19,054 participations in FP2; 31,348 participations in FP3.
  - **EUPRO**: 7,818 participations in FP1; 19,126 participations in FP2; 30,732 participations in FP3.

SAPIENZA UNIVERSITÀ DI ROMA

Results in terms of network metrics

- By comparing **raw** (i.e., from CORDIS data), **refined** (i.e., obtained through the application of the methodology), and **EUPRO networks** (i.e., from the EUPRO database) we are able to assess if the **connection structure** of the **obtained networks** is moving **closer** to the **EUPRO ones**.
- In this way, we can determine if the application of the **methodology** has contributed to **improve** the **quality** of the **original dataset**, making it **reliable** to map the **collaboration process**.

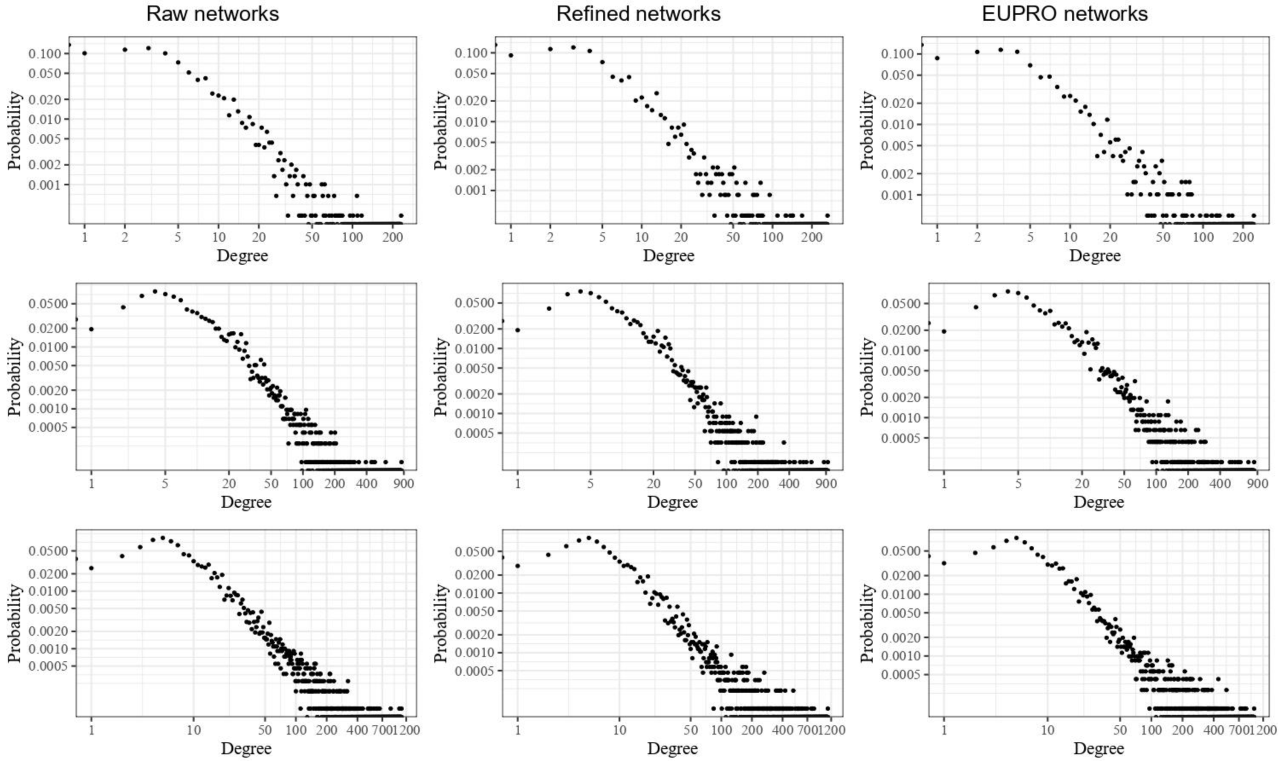
	Raw network			Refined network			Eupro network		
	FP1	FP2	FP3	FP1	FP2	FP3	FP1	FP2	FP3
Distinct organizations	2, 977	7, 280	10, 969	2, 319	5, 596	8, 644	1, 972	4, 587	7, 095
Giant component	2, 301	6, 840	9, 942	1, 860	5, 320	7, 881	1, 593	4, 379	6, 520
Edges	10, 236	66, 085	113, 627	9, 159	60, 453	97, 890	8, 429	55, 825	88, 966
Density	0.23%	0.25%	0.19%	0.34%	0.39%	0.26%	0.43%	0.53%	0.35%
Diameter	9	8	9	8	7	8	7	6	7
Average shortest path	3.9	3.3	3.3	3.4	3	3.1	3.3	2.9	3
Clustering coefficient	0.25	0.23	0.19	0.20	0.19	0.20	0.20	0.20	0.22
Mean degree[a]	8.6	19.2	22.6	9.7	22.7	24.8	10.4	25.4	27.2

[a] Referred to giant component only

SAPIENZA UNIVERSITÀ DI ROMA

Results in terms of network metrics

- To provide a whole picture of the **connection structure**, we estimate the **exponent α** of the respective **degree distribution functions** following a **power-law** ($P(x) \sim x^{-\alpha}$).



	Raw network	Refined network	Eupro network
FP1	2.9	2.5	2.2
FP2	2.5	2.3	2.2
FP3	2.1	2.1	2

SAPIENZA UNIVERSITÀ DI ROMA

Error analysis

- Some **differences** between **refined** and **EUPRO networks** still hold as not all **organizations** can be determined **distinctively** by identifying **equivalences** between **labels**.
  - **Research institutes** that are **aggregated** in **EUPRO** under the name of the **research center** they belong.
  - **Organizations** involved in **mergers and acquisitions**.
  - Possible **rebranding**.
- We compute the **pairwise-Precision (pP)** and the **pairwise-Recall (pR)** for the aggregated dataset (i.e., including all three FPs) to conduct a **systematic error analysis**.

$$pP = \frac{|Pairs_{disambiguated} \cap Pairs_{labeled}|}{|Pairs_{disambiguated}|}, \qquad pR = \frac{|Pairs_{disambiguated} \cap Pairs_{labeled}|}{|Pairs_{labeled}|}$$

SAPIENZA UNIVERSITÀ DI ROMA

Error analysis

- We find **97%** of **disambiguated pairs** to be **correct** in comparison with EUPRO ($pP = 0.97$), and **82%** of all **labeled pairs identified** ($pR = 0.82$).
- We also analyze the values of $\boldsymbol{pR}$ by **country** and **activity type** to unveil the nature of **unidentified matchings**.

	FR	DE	UK	IT	NL
N	1, 458, 922	414, 353	364, 613	313, 597	294, 108
pR	0.93	0.75	0.81	0.72	0.69

	Research	Education	Industry	Governmental	Other
N	2, 108, 273	1, 111, 044	290, 024	34, 276	32, 151
pR	0.87	0.82	0.50	0.91	0.51

- **Dutch names** are the most **difficult** to identify through **linguistic criteria**
- **Half** of **pairwise equivalences** between **private organization** names are **not** identified

SAPIENZA UNIVERSITÀ DI ROMA

Efficiency of the methodology

- The **algorithm** returns only **0.005%** of all **possible pairwise equivalences** to **check by hand**.
- The use of **“cosine” distance** to assess **string similarity** is supported by our results.

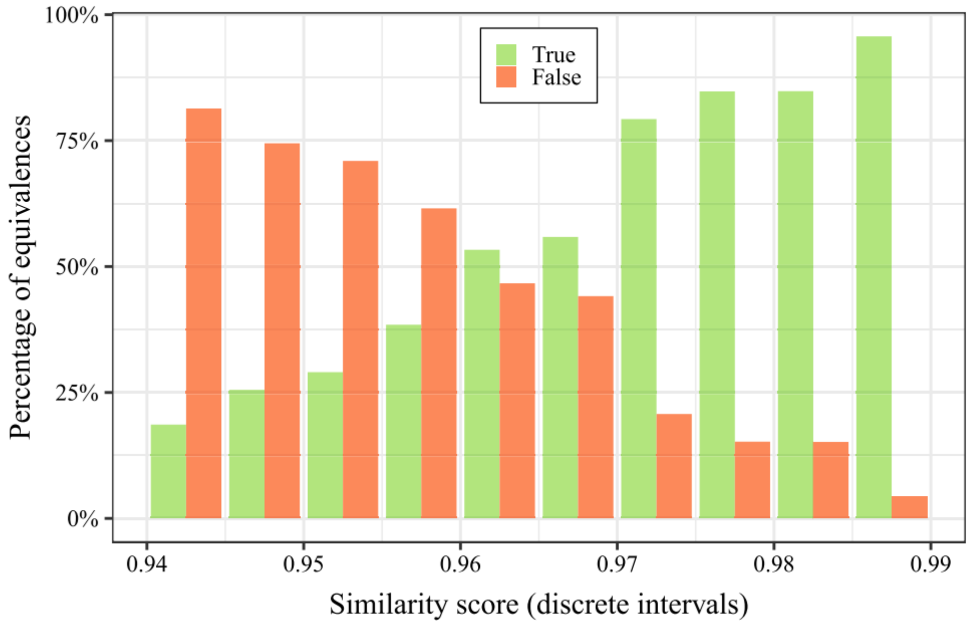


Fig. 5 Percentage of equivalences between labels classified as true or false by hand, over discrete intervals of similarity score

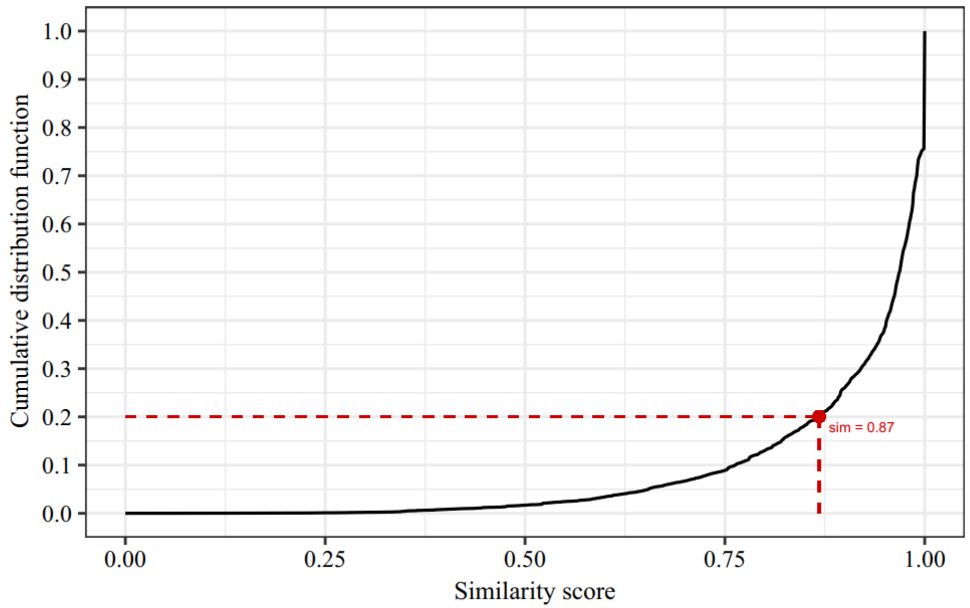


Fig. 7 Cumulative distribution function of the similarity score between corresponding original and EUPRO labels

SAPIENZA UNIVERSITÀ DI ROMA

Conclusions

- We propose a **hybrid methodology** which classifies **almost** all **equivalences automatically** and returns just a **small portion** to **check by hand**.
- It is **not domain-specific**, and it does **not** require a **context** to **disambiguate**. Moreover, it can be applied to **unbalanced data** with a **considerable** number of organizations.
- The dataset we obtain moves **closer** to the **EUPRO** database through a **reduced time consuming** procedure.
- The methodology is shown to be **efficient** and **reliable** thanks to the implementation of **«cosine» distance**.
- **Remaining** unmatched cases are mainly due to **sources of errors** that we are **not** addressing. However, the methodology can be **integrated** with **external registers** (such as FirmReg) to consider **private companies dynamics**.
- The **relevance** of **unidentified matchings** depends however on the specific **research objectives**.


Uncovering collaborative patterns and transition dynamics
in EU Framework Programmes through network modeling

SAPIENZA UNIVERSITÀ DI ROMA

Context and background

- The **EU Research and Technological Development (RTD) policy** was established in the 1980s to promote **international research collaboration**.
- Supporting **collaborative R&D projects** has become increasingly relevant for **policy-makers** and **institutions**. At the same time, the **innovation performance** of **organizations** benefits from **collaborative R&D**.
- **Various approaches** have been applied to the study of **collaborative R&D**. The **dynamic nature** of **socio-economic** processes necessitates examining them through the new lens of **economic complexity** (Balland et al., 2022).
- **SNA** especially, has been increasingly adopted to investigate the **behaviors** of **collaborative relationships** (Cerqueti et al., 2023), analyze the **structure** of **innovation systems** (Ancona et al., 2023), and identify the **key actors** in **collaboration networks** (Cinelli et al., 2022).

SAPIENZA UNIVERSITÀ DI ROMA

Research motivations

- When considering the **evolution** of **collaborative networks**, most of the works employ **Stochastic Actor-Oriented Models (SAOMs)** (Giuliani, 2013; Cao et al., 2017) since they are efficient in processing **longitudinal network data** (Broekel et al., 2014).

- This method assumes that **network structures evolve** as a **Markov chain outcome** (Snijders, 2017), whereas in many **real networks**, **evolution** often displays a **non-Markovian** behavior (Williams et al., 2022).

- Previous **studies on EU FPs** focus mainly on **macro-level analyses** (i.e., at the country or regional level) and **average dynamics** of **specific FPs**.

- **Low attention** has been paid to the **micro-dynamics** at the **participant level**, especially over an **extended period**, spanning **multiple FPs**.

SAPIENZA UNIVERSITÀ DI ROMA

Our approach

- We explore the **participation dynamics** in **collaborative research projects** funded by the **first eight EU FPs**, i.e., from **FP1** to **Horizon 2020** (H2020).
- We map the **local behaviors** of **single actors** in terms of their **position** in the **collaborative networks** through **centrality measures**.
- We **statistically assess** whether the **dynamics** of **collaborations** among the **organizations** receiving funds in **all** the first eight EU FPs have a **Markovian nature**.
- We employ an **innovative method** to **partition** the **rankings** of **organizations** based on the values of **strength centrality** and we estimate the **probability of moving** from one **level** of **centrality** to another over **consecutive FPs**.

SAPIENZA
UNIVERSITÀ DI ROMA

Dynamics of collaborative research networks

- **Co-evolution** of **network position** and **research performance** (Zhang & Chen, 2022).
- **"Core-periphery"** structure and **preferential attachment** mechanisms (Wagner & Leydersdorff, 2005; Cao et al., 2017; Xie and Su, 2021; Zirulia, 2023).
- **Liberal democracy** and **governance similarities** positively affect **international research collaboration** (Whetsell, 2023).
- **Science and technology policies** influence the **dynamics** of **collaborative networks** (Park & Leyedesdoff, 2010) and their **small-world** structure (Zhang et al., 2016).
- The **collaboration dynamics** depend on the **dimension** of the **research group** (Palla et al., 2007) and the **technological dynamism** of the **industry** organizations belong to (Tatarynowicz et al., 2016).

SAPIENZA UNIVERSITÀ DI ROMA

EU-funded projects

- **Participation** in **EU research projects** enhances the **scientific performance** of **organizations** (Calignano, 2021).
- The **scientific reputation** of organizations in turn impacts the **likelihood** of **receiving funds** (Enger & Castellacci, 2016) and the **collaboration structure** of **HEIs** (Lepori et al., 2015).
- The **centrality** of **organizations** in **collaborative networks** affects the **probability** of **applying** and **being funded** (Enger, 2018). Particularly **central organizations** have generally access to **more funds** (Cinelli et al., 2022).
- **HEIs** and **participants** from **EU-15 countries** exhibit **higher centrality** values (Balland et al., 2019).
- **Consortium characteristics** (Wanzeböck et al., 2020), and **geographical heterogeneity** among partners (de Arroyabe et al., 2021) are relevant for the **project success**.
- **Geographical**, **social**, and **technological proximities** affect **collaboration patterns** (Scherngell & Barber, 2011; Paier & Scherngell, 2011; Heringa et al., 2016).

SAPIENZA UNIVERSITÀ DI ROMA

Sample and data

Main source: **EUPRO database** (Heller-Schuh et al., 2020).

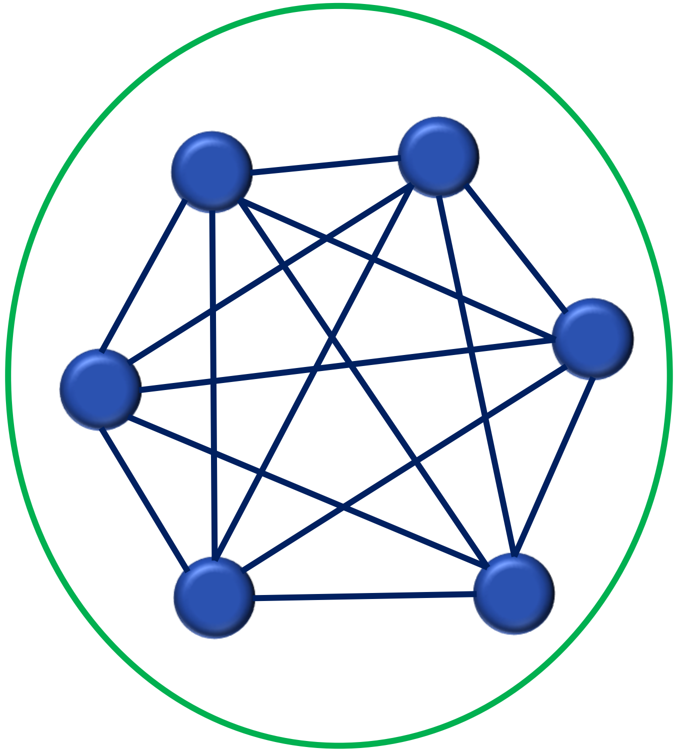
	Distinct projects	Distinct organizations
FP1 (1984-1987)	3,266	1,972
FP2 (1987-1991)	3,972	4,587
FP3 (1990-1994)	5,461	7,095
FP4 (1994-1998)	14,493	19,255
FP5 (1998-2002)	15,091	22,862
FP6 (2002-2006)	10,100	20,582
FP7 (2007-2013)	25,778	29,334
Horizon 2020 (2014-2020)	25,604*	31,319*

*Currently updated

For our purpose, we consider the **organizations** participating in **all the eight FPs**, in order to analyze the whole **dynamics** from **FP1** to **H2020**. The final amount of selected participants is equal to **509** organizations.

SAPIENZA UNIVERSITÀ DI ROMA

Network modeling



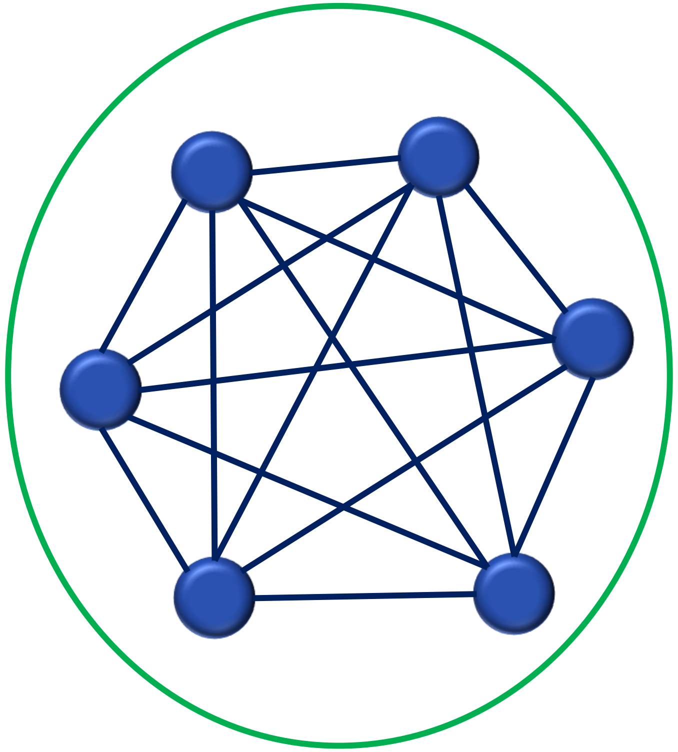
PROJECT X

RISIS RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

SAPIENZA UNIVERSITÀ DI ROMA

Network modeling

$$\boxed{d_i = 5 \ \forall \ i \in V_X, \quad w_{ij} = 1 \forall \ (i,j) \in E_X}$$

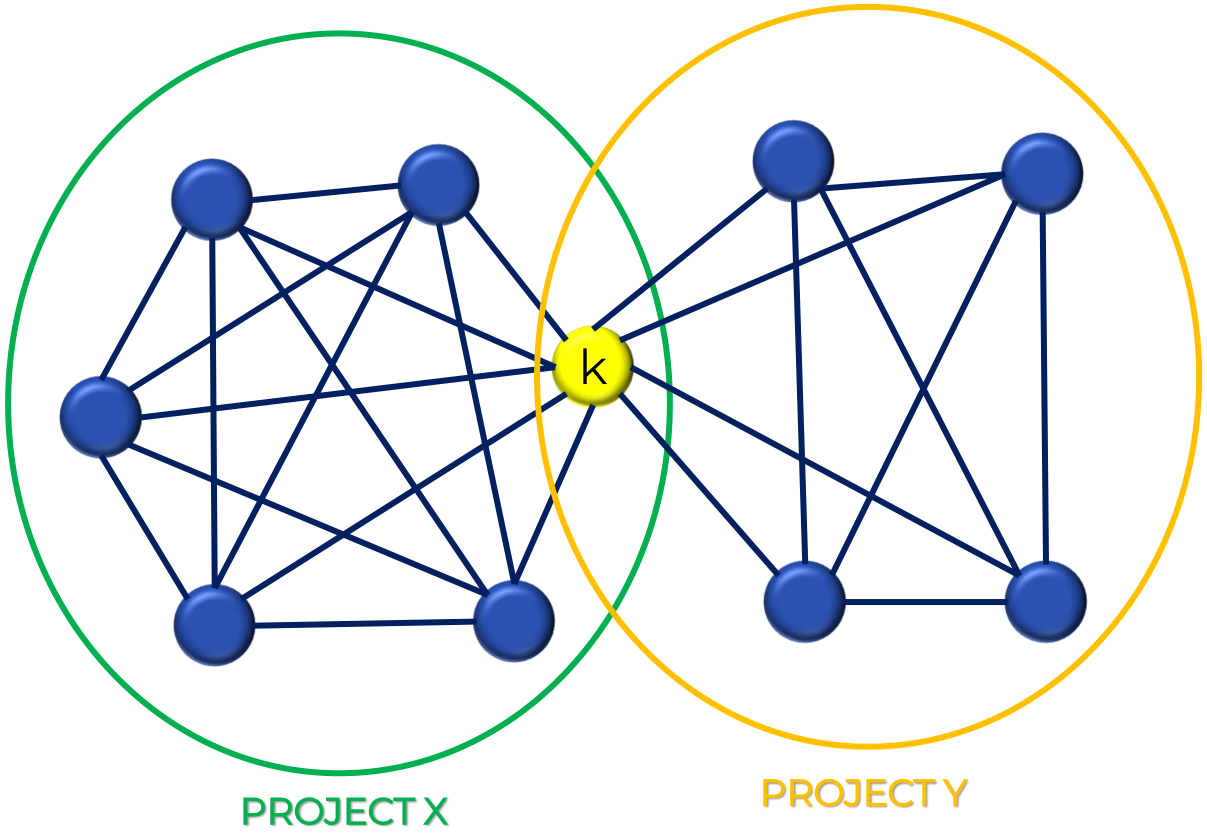


PROJECT X

RISIS RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

SAPIENZA
UNIVERSITÀ DI ROMA

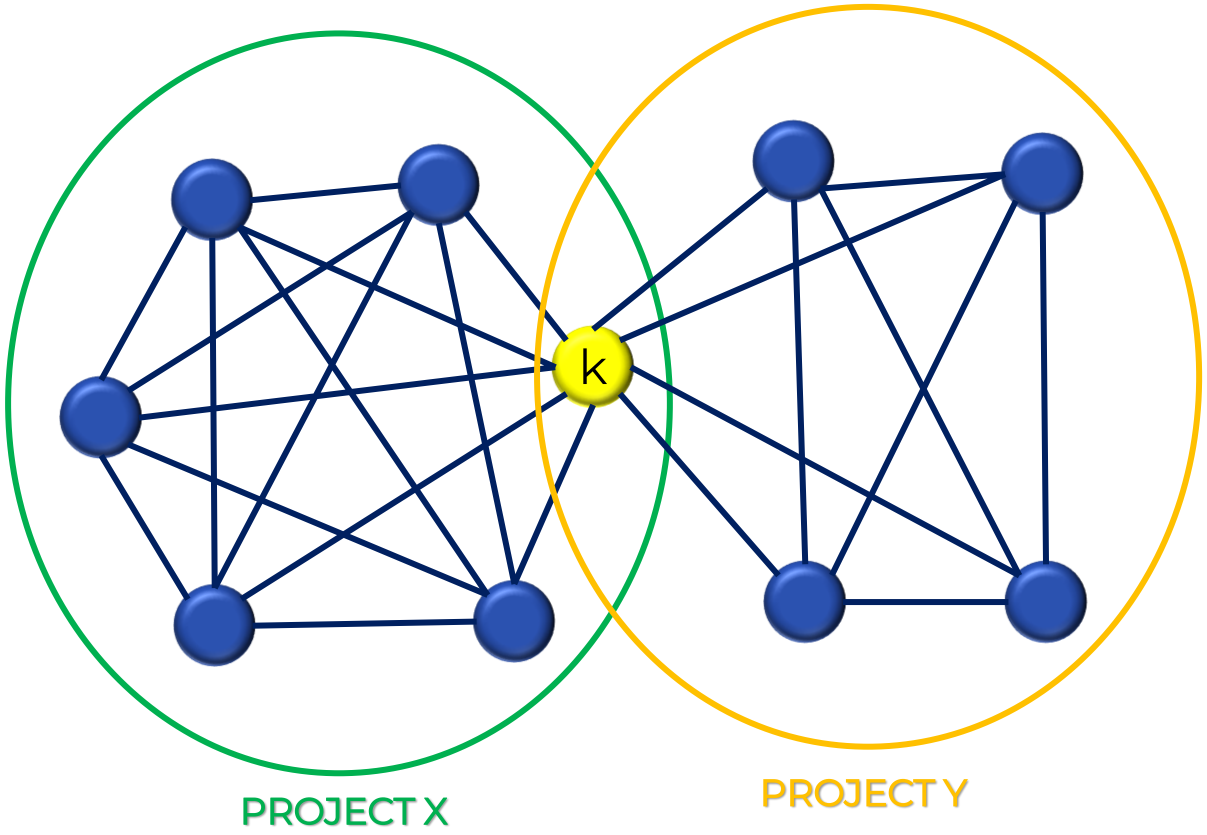
Network modeling



SAPIENZA UNIVERSITÀ DI ROMA

Network modeling

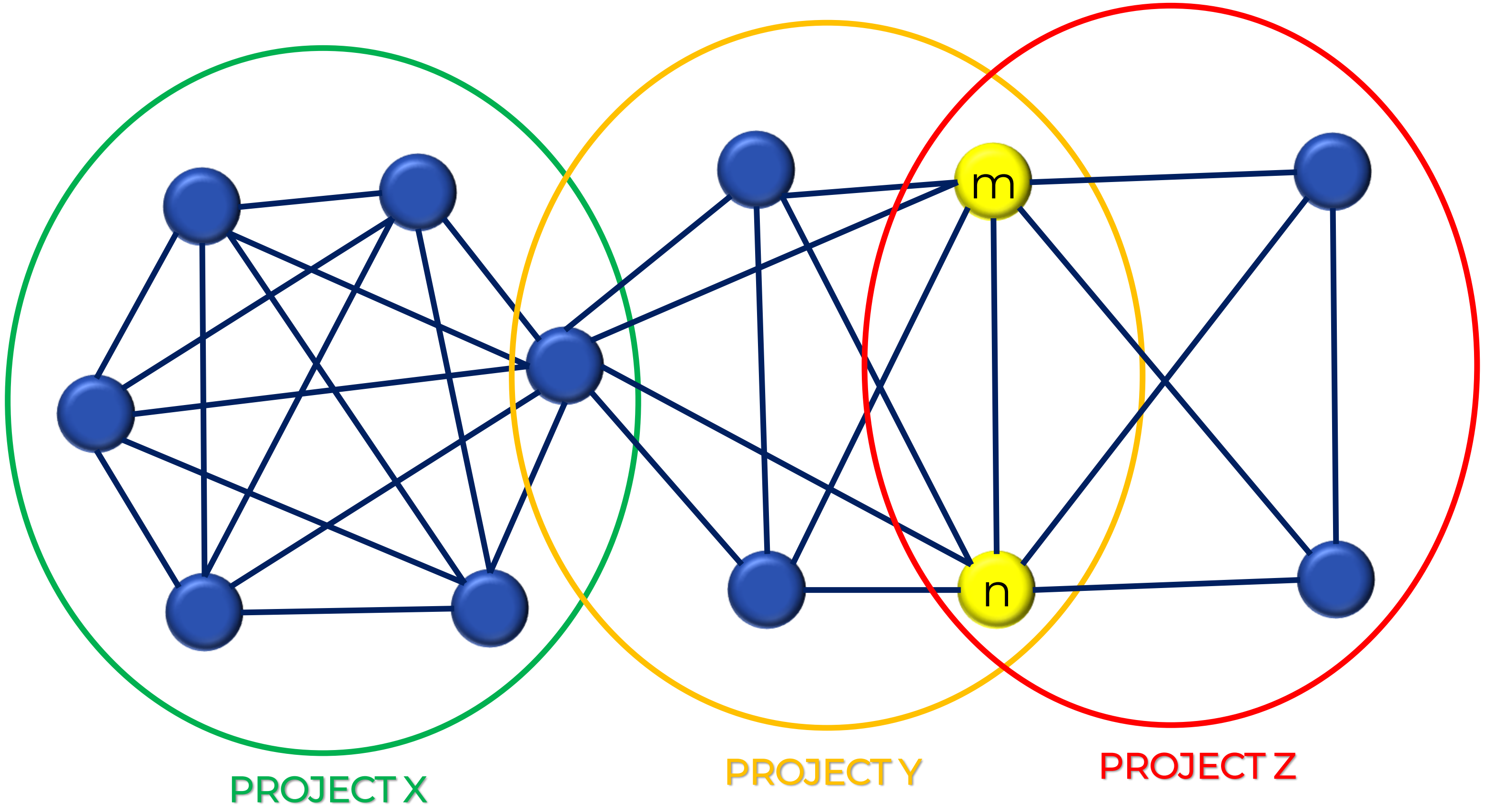
$d_k \uparrow, \quad w_{ij} = 1 \forall (i,j) \in E_X \cup E_Y$



RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

SAPIENZA UNIVERSITÀ DI ROMA

Network modeling

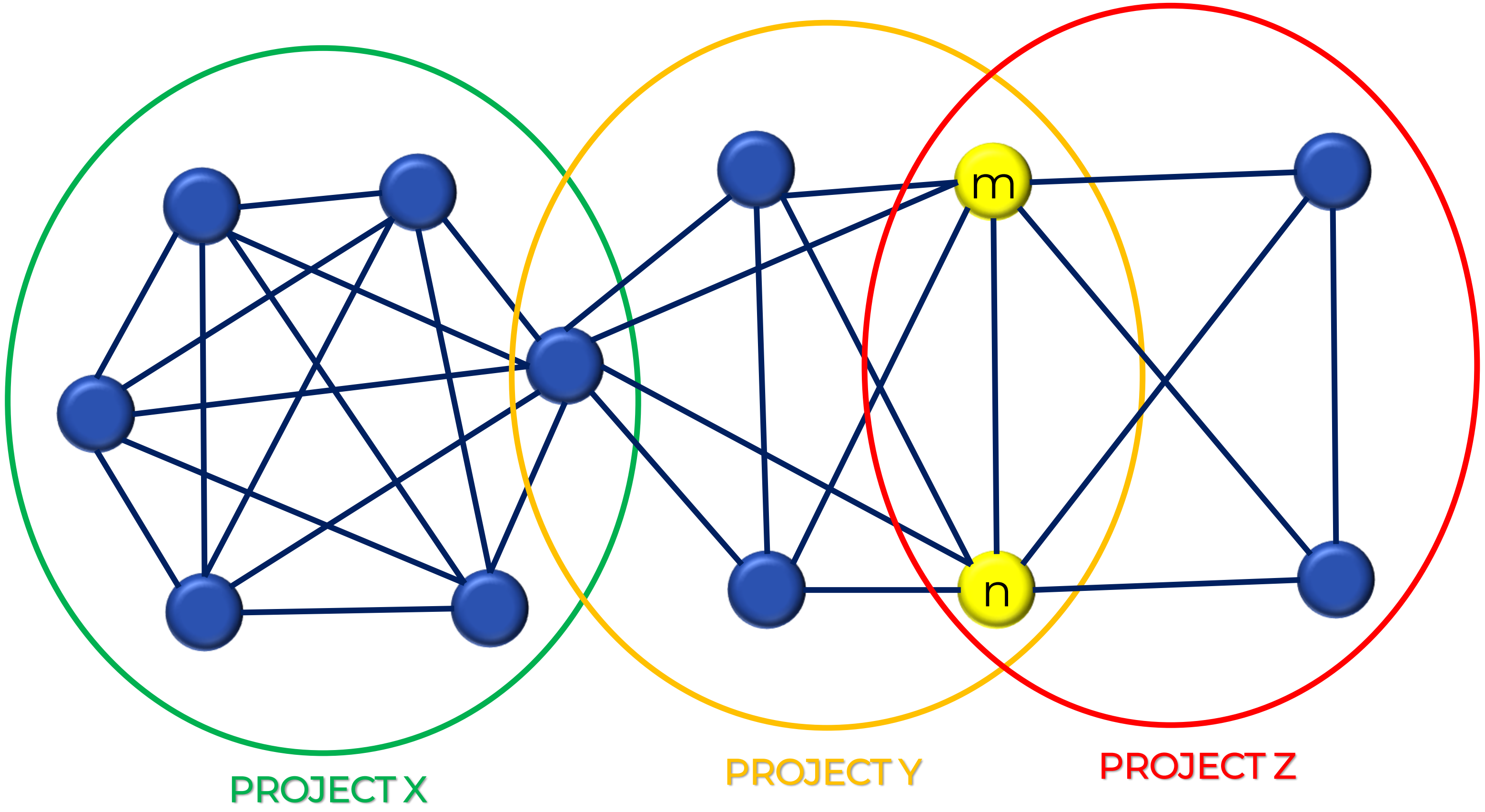


RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

SAPIENZA UNIVERSITÀ DI ROMA

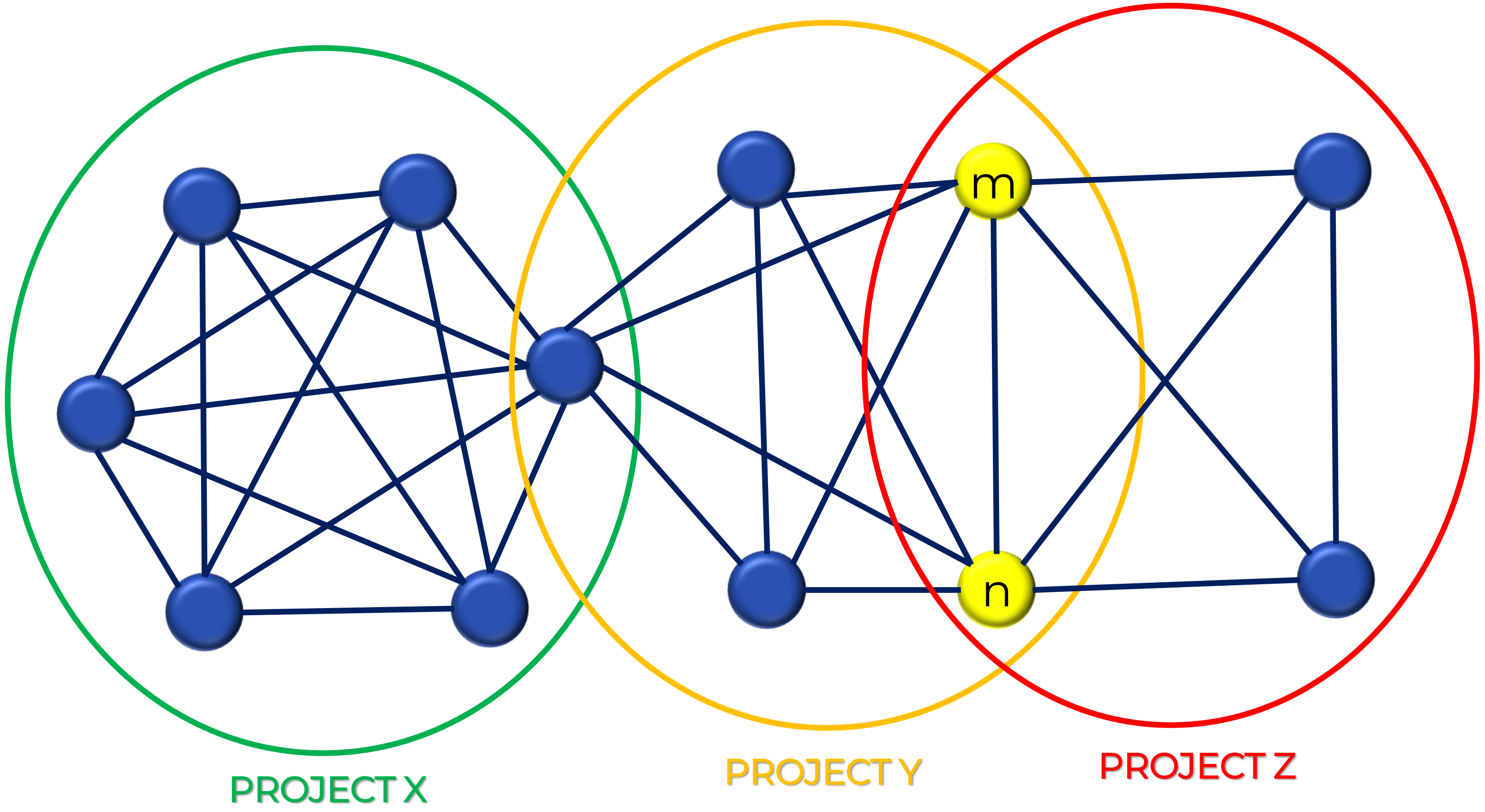
Network modeling

$d_m, d_n \uparrow, \quad w_{mn} \uparrow$



SAPIENZA UNIVERSITÀ DI ROMA

Network modeling



Key centrality measure:

$$s_i = \sum_{j=1}^{N} w_{ij} \, ,$$

where N is the number of nodes in the network

SAPIENZA UNIVERSITÀ DI ROMA

Statistical analysis of process Markovianity

- A **first-order Markov** process is a **memory-less** process, i.e., the **probability** of becoming one of the states of the chain in the **next step** depends only on the **present state** (Gudivada et al., 2015).
- Such a property represents the **stochasticity** of some **phenomenon evolution**, with relevant **implications** in the context of **forecasting**.
- How do we assess it?
  - Given a **discrete-time stochastic process** $X = (X(t): t \in N)$ taking values in a set of **ranks**, and P the related **probability law**, we can say that X is a **Markov chain** of **order one** if we have (Friedrich et al., 2011):

$$\underbrace{P(X(t+1) = i_{t+1} \mid X(t) = i_t)}_{First-order\ transition\ matrix} = \underbrace{P(X(t+1) = i_{t+1} \mid X(t) = i_t, X(t-1) = i_{t-1})}_{Second-order\ transition\ matrix}$$

Statistical analysis of process Markovianity

- A **first-order Markov** process is a **memory-less** process, i.e., the **probability** of becoming one of the states of the chain in the **next step** depends only on the **present state** (Gudivada et al., 2015).

- Such a property represents the **stochasticity** of some **phenomenon evolution**, with relevant **implications** in the context of **forecasting**.

- How do we assess it?

  - Given a **discrete-time stochastic process** $X = (X(t): t \in N)$ taking values in a set of **ranks**, and P the related **probability law**, we can say that X is a **Markov chain** of **order one** if we have (Friedrich et al., 2011):

$$P(X(t+1) = i_{t+1} \mid X(t) = i_t) = P(X(t+1) = i_{t+1} \mid X(t) = i_t, X(t-1) = i_{t-1})$$

  - We run **1000 simulations** for both **order one** and **two transition matrices** and pairwise compare them via **Kolmogorov–Smirnov (KS) test**.

SAPIENZA UNIVERSITÀ DI ROMA

Partitioning the transition matrices

- For each pair of consecutive FPs, we rely on the **empirical probability matrix** $\boldsymbol{P}$, whose dimension is equal to $n \times m$, where n is the number of **distinct values of strength** in a **specific FP**, and m is the number of **distinct values of strength** in the **subsequent FP**.

- Let us consider the **empirical probability matrix** $\boldsymbol{P}$ from FP1 to FP2:

$$P^{1,2} = \begin{pmatrix} p_{1,1} & \cdots & p_{1,m} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,m} \end{pmatrix}$$

- The generic element $\boldsymbol{p_{i,j}}$ is equal to the **probability** that an **organization** whose **strength** in FP1 is $\boldsymbol{s_i}$, has a value of **strength** in FP2 equal to $\boldsymbol{s'_j}$, which is computed as:

$$p_{i,j} = \frac{k_{i,j}}{\sum_{j=1}^{m} k_{i,j}} \ ,$$

where $k_{i,j}$ corresponds to the number of times s_i is associated to s'_j.

SAPIENZA UNIVERSITÀ DI ROMA

Partitioning the transition matrices

- We adapt a **methodology** proposed in (Cerqueti et al., 2017) to **endogenously partition** the **transition matrices** by identifying **three different classes** of **strength** for each FP, i.e., **low**, **medium**, and **high**.

$$Low = \{s_i \mid s_i \leq t_1\}$$

$$Medium = \{s_i \mid t_1 < s_i \leq t_2\}$$

$$High = \{s_i \mid t_2 < s_i\}$$

Where t_1 and t_2 are the **optimal thresholds** to be determined accordingly.

.

- For each pair of **consecutive FPs**, we obtain the following matrix.

$$\pi = \begin{pmatrix} \pi_{L,L} & \pi_{L,M} & \pi_{L,H} \\ \pi_{M,L} & \pi_{M,M} & \pi_{M,H} \\ \pi_{H,L} & \pi_{H,M} & \pi_{H,H} \end{pmatrix}$$

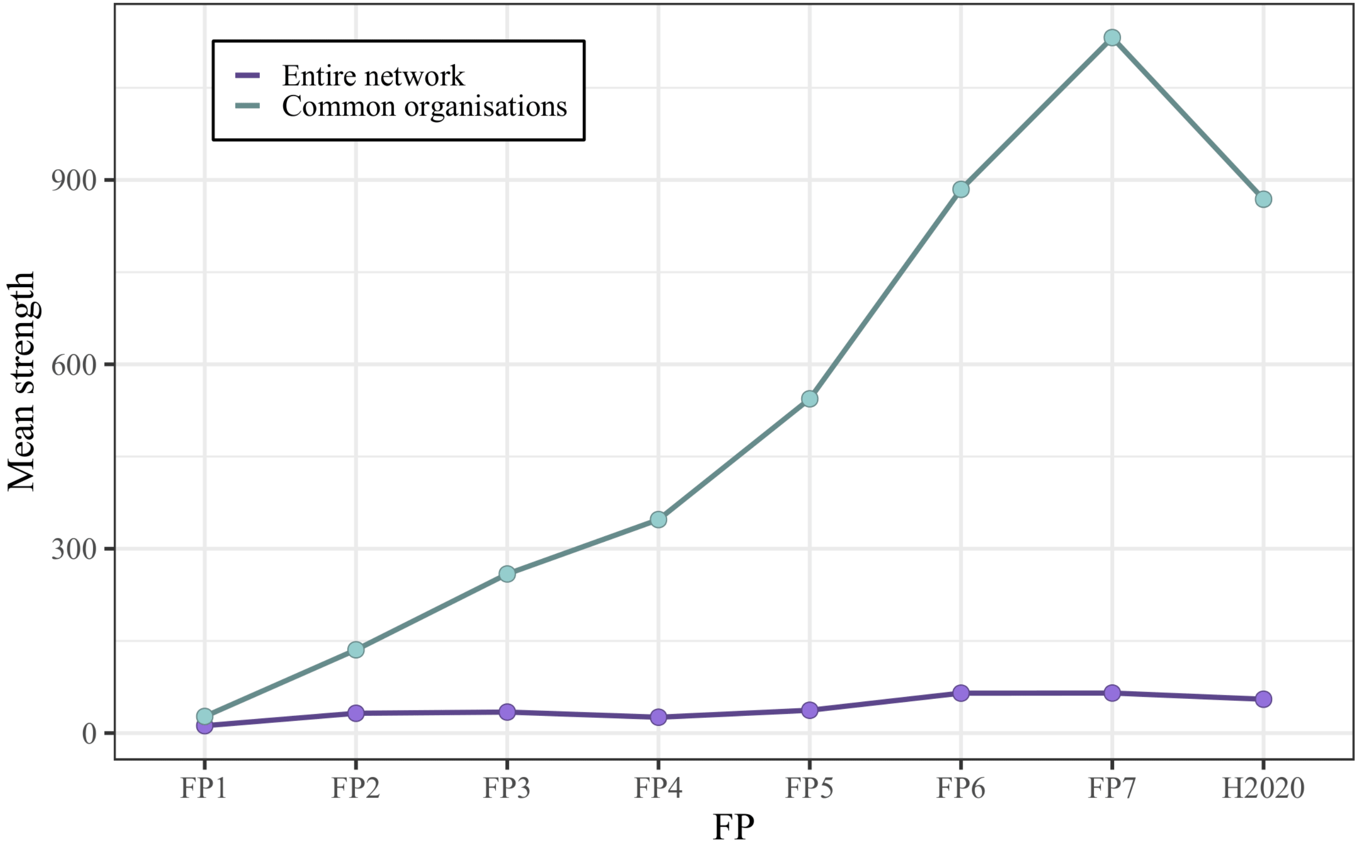
Probability of decreasing the level of centrality

Probability of maintaining the same level of centrality

Probability of increasing the level of centrality

SAPIENZA UNIVERSITÀ DI ROMA

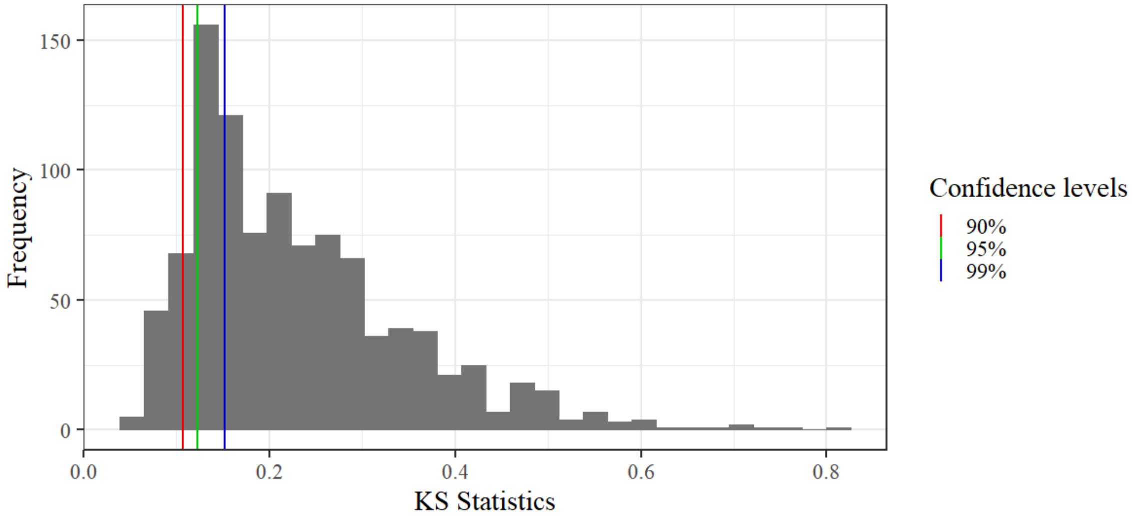
Unveiling dynamic patterns of EU FPs

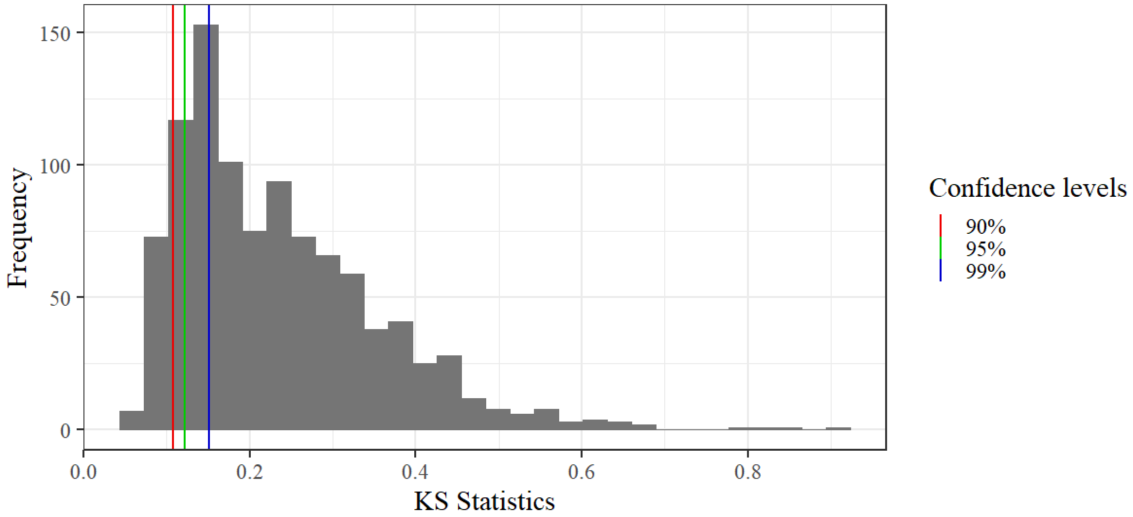


- The value of **mean strength** has been dramatically **increasing** over time for the **organizations** taking part in **all** the eight FPs, whereas it is almost **stable** when considering the **entire network**.
- **New incumbents** tend to collaborate with **experienced organizations**, augmenting the **gap** between **old** and **new** participants.

SAPIENZA UNIVERSITÀ DI ROMA

Assessing the Markovianity of participation dynamics





- The **outcome** of the **KS test** is not obvious: just a **portion** of **KS values** confirm the **Markovianity** of the **process**.
- However, considering that the **mode** of the **distribution** is in the 99% confidence interval, we can say that the process is likely to be generated from a **Markov chain**.
- We can then observe a **quasi-Markovian** nature of the **participation dynamics** in EU FPs.

SAPIENZA UNIVERSITÀ DI ROMA

The effects of EU initiative on transition matrices

- The estimated **transition probability matrices** are reported as follows:

$$\pi^{1,2} = \begin{pmatrix} 0.36 & 0.14 & 0.5 \\ 0.15 & 0.09 & 0.76 \\ 0.04 & 0.02 & 0.94 \end{pmatrix} \quad \pi^{2,3} = \begin{pmatrix} 0.96 & 0.00 & 0.04 \\ 0.84 & 0.01 & 0.15 \\ 0.24 & 0.01 & 0.75 \end{pmatrix} \quad \pi^{3,4} = \begin{pmatrix} 0.48 & 0.01 & 0.51 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix}$$

$$\pi^{4,5} = \begin{pmatrix} 0.33 & 0.02 & 0.65 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix} \quad \pi^{5,6} = \begin{pmatrix} 0.92 & 0.00 & 0.08 \\ 1.00 & 0.00 & 0.00 \\ 0.28 & 0.01 & 0.71 \end{pmatrix} \quad \pi^{6,7} = \begin{pmatrix} 0.58 & 0.01 & 0.41 \\ 0.17 & 0.00 & 0.83 \\ 0.02 & 0.01 & 0.97 \end{pmatrix}$$

SAPIENZA UNIVERSITÀ DI ROMA

The effects of EU initiative on transition matrices

- The estimated **transition probability matrices** are reported as follows:

$$\pi^{1,2} = \begin{pmatrix} 0.36 & 0.14 & 0.5 \\ 0.15 & 0.09 & 0.76 \\ 0.04 & 0.02 & 0.94 \end{pmatrix} \quad \pi^{2,3} = \begin{pmatrix} 0.96 & 0.00 & 0.04 \\ 0.84 & 0.01 & 0.15 \\ 0.24 & 0.01 & 0.75 \end{pmatrix} \quad \pi^{3,4} = \begin{pmatrix} 0.48 & 0.01 & 0.51 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix}$$

$$\pi^{4,5} = \begin{pmatrix} 0.33 & 0.02 & 0.65 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix} \quad \pi^{5,6} = \begin{pmatrix} 0.92 & 0.00 & 0.08 \\ 1.00 & 0.00 & 0.00 \\ 0.28 & 0.01 & 0.71 \end{pmatrix} \quad \pi^{6,7} = \begin{pmatrix} 0.58 & 0.01 & 0.41 \\ 0.17 & 0.00 & 0.83 \\ 0.02 & 0.01 & 0.97 \end{pmatrix}$$

- It is **rather hard** for a participant with a **high level** of **strength** to shift toward a **less central position** in the following FP.

SAPIENZA UNIVERSITÀ DI ROMA

The effects of EU initiative on transition matrices

- The estimated **transition probability matrices** are reported as follows:

$$\pi^{1,2} = \begin{pmatrix} 0.36 & 0.14 & 0.5 \\ 0.15 & 0.09 & 0.76 \\ 0.04 & 0.02 & 0.94 \end{pmatrix} \quad \pi^{2,3} = \begin{pmatrix} 0.96 & 0.00 & 0.04 \\ 0.84 & 0.01 & 0.15 \\ 0.24 & 0.01 & 0.75 \end{pmatrix} \quad \pi^{3,4} = \begin{pmatrix} 0.48 & 0.01 & 0.51 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix}$$

$$\pi^{4,5} = \begin{pmatrix} 0.33 & 0.02 & 0.65 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix} \quad \pi^{5,6} = \begin{pmatrix} 0.92 & 0.00 & 0.08 \\ 1.00 & 0.00 & 0.00 \\ 0.28 & 0.01 & 0.71 \end{pmatrix} \quad \pi^{6,7} = \begin{pmatrix} 0.58 & 0.01 & 0.41 \\ 0.17 & 0.00 & 0.83 \\ 0.02 & 0.01 & 0.97 \end{pmatrix}$$

- It is **rather hard** for a participant with a **high level** of **strength** to shift toward a **less central position** in the following FP.

- A **participant** with a **low level** of **strength** is more likely to **increase** its **centrality** over consecutive FPs, **except** for the **transition** from **FP2 to FP3** and from **FP5 to FP6**.

SAPIENZA UNIVERSITÀ DI ROMA

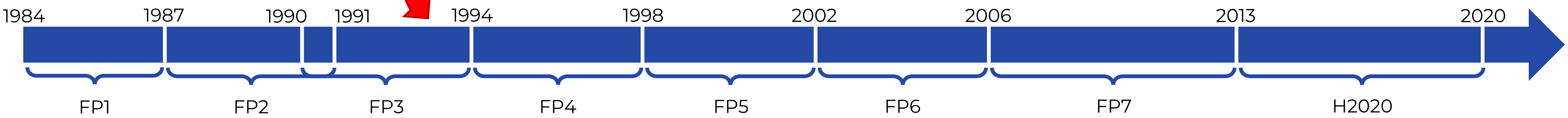
The effects of EU initiative on transition matrices

- The estimated **transition probability matrices** are reported as follows:

$$\pi^{1,2} = \begin{pmatrix} 0.36 & 0.14 & 0.5 \\ 0.15 & 0.09 & 0.76 \\ 0.04 & 0.02 & 0.94 \end{pmatrix} \quad \pi^{2,3} = \begin{pmatrix} 0.96 & 0.00 & 0.04 \\ 0.84 & 0.01 & 0.15 \\ 0.24 & 0.01 & 0.75 \end{pmatrix} \quad \pi^{3,4} = \begin{pmatrix} 0.48 & 0.01 & 0.51 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix}$$

$$\pi^{4,5} = \begin{pmatrix} 0.33 & 0.02 & 0.65 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix} \quad \pi^{5,6} = \begin{pmatrix} 0.92 & 0.00 & 0.08 \\ 1.00 & 0.00 & 0.00 \\ 0.28 & 0.01 & 0.71 \end{pmatrix} \quad \pi^{6,7} = \begin{pmatrix} 0.58 & 0.01 & 0.41 \\ 0.17 & 0.00 & 0.83 \\ 0.02 & 0.01 & 0.97 \end{pmatrix}$$

Maastricht Treaty



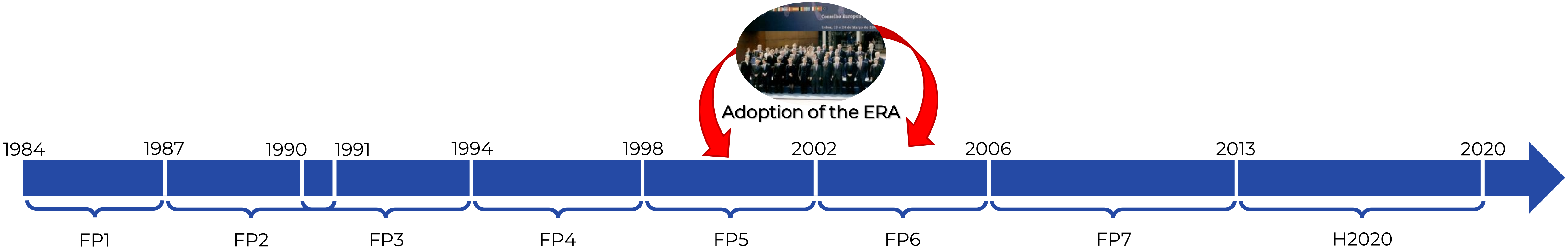
SAPIENZA UNIVERSITÀ DI ROMA

The effects of EU initiative on transition matrices

- The estimated **transition probability matrices** are reported as follows:

$$\pi^{1,2} = \begin{pmatrix} 0.36 & 0.14 & 0.5 \\ 0.15 & 0.09 & 0.76 \\ 0.04 & 0.02 & 0.94 \end{pmatrix} \qquad \pi^{2,3} = \begin{pmatrix} 0.96 & 0.00 & 0.04 \\ 0.84 & 0.01 & 0.15 \\ 0.24 & 0.01 & 0.75 \end{pmatrix} \qquad \pi^{3,4} = \begin{pmatrix} 0.48 & 0.01 & 0.51 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix}$$

$$\pi^{4,5} = \begin{pmatrix} 0.33 & 0.02 & 0.65 \\ 0.00 & 0.00 & 1.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix} \qquad \pi^{5,6} = \begin{pmatrix} 0.92 & 0.00 & 0.08 \\ 1.00 & 0.00 & 0.00 \\ 0.28 & 0.01 & 0.71 \end{pmatrix} \qquad \pi^{6,7} = \begin{pmatrix} 0.58 & 0.01 & 0.41 \\ 0.17 & 0.00 & 0.83 \\ 0.02 & 0.01 & 0.97 \end{pmatrix}$$



SAPIENZA
UNIVERSITÀ DI ROMA

Discussion and conclusion

- **Innovative approach** to analyze **collaborative patterns** and **participation dynamics** in **EU-funded projects**, comprising elements from **SNA** and **statistics**.
- The **participation process** exhibits a **quasi-Markovian nature**, opening the space for accurate fo**recasting procedures**.
- The **Treaty of Maastricht** first, and the **promotion of the ERA** then, emerge as the most **crucial events** determining the **openness** and the "**democratization"** of **European research funds**.
- **Policy actions** are needed to avoid **“predatory” behaviors** and **exclusive access** to **European funds**.
- The analysis relies on **statistical assumptions** that can be challenged in **future research**.

THANKS FOR YOUR ATTENTION

andrea.ancona@uniroma1.it

Q&A Session