The structure, evolution and interaction of multiplex networks of scientific collaboration at a research university

Valerio Leone Sciabolazza [1], Raffaele Vacca [2]

[1]Department of Business and Economics, University of Naples Parthenope
[2]Department of Sociology and Criminology, University of Florida

June 28, 2018

Main motivations

The aim of this paper is to contribute to the understanding of the structural evolution of scientific collaboration networks.

So far, research on science and collaboration conceptualized collaboration as co-authorship. Typical studies look at:

Co-authorship on publications in very **extensive networks**, including multiple institutions (e.g., departments, universities, disciplines, countries, etc.).

Two reasons for the popularity of the idea of collaboration as co-authorship are:

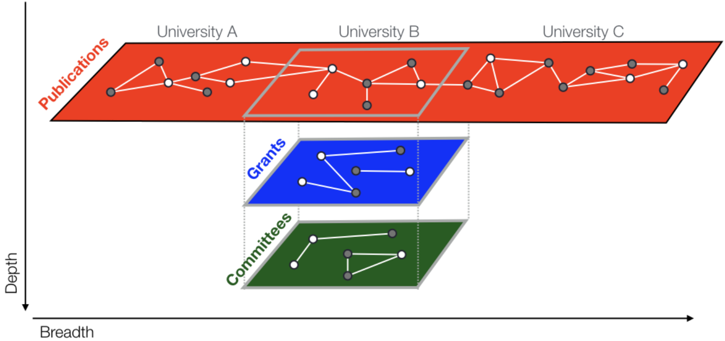
- The preeminence of publications in the measurement and evaluation of research productivity (both for individuals and organizations), and
- the availability of publication data from databases such as the *Web of Science*, *PubMed*, *Google Scholar*, etc. (sometimes with institution-based licenses and access).

However, we know from our experience as faculty and researchers that scientific collaboration can occur in many ways.

Main motivations

We think about this as a problem of dimensions:
Breadth versus **Depth**

- Extensive literature looks at the **red layer**: collaboration networks across different institutions. Not very deep networks (only the publication layer).
- We look at **deeper**, multi-layer networks of collaboration (but limits in data availability and computational power) $\rightarrow$ We reduce the **breadth** of the networks to the University of Florida.



Plan of the day

Data

Consider a multiplex network $G(V, E, D)$, where:

- V: set of nodes representing all UF investigators who have collaborated with at least another UF investigator in one year from 2011 to 2015
- D: set of layers represents a specific instance of scientific collaborations between UF investigators at time t.
- E: set of edges connecting two nodes in one layer, i.e. a collaboration between two investigators.
  Collaborations are proxied by the number of times two researchers were:
  - co-authors on one publication;
  - co-investigators on a grant;
  - members of the same Ph.D. Committee

As a result, for each considered year, we obtain three **undirected weighted networks**.

Data

Publication network

	2011	2012	2013	2014	2015
Number of Nodes	3,658	3,885	3,870	3,470	2,893
Number of Edges	7,582	8,485	7,737	5,980	4,806
Nodes making up to					
50% of collaborations	185	206	211	194	159
90% of collaborations	474	524	526	487	392
Degree					
Average	4.15	4.37	4	3.45	3.32
Standard deviation	5.09	5.19	4.7	3.96	3.98
Clustering coefficient	0.37	0.39	0.40	0.39	0.45
Giant component					
Number of Nodes	1,250	1,287	1,237	1,203	1,188
% of Nodes	0.53	0.54	0.53	0.51	0.51
Isolates					
Number of Nodes	636	652	645	690	691
% of Nodes	27	27	28	29	30

Data

Grant network

	2011	2012	2013	2014	2015
Number of Nodes	2,380	2,381	2,339	2,354	2,318
Number of Edges	2,959	2,966	3,008	3,047	3,403
Nodes making up to					
50% of collaborations	116	114	112	111	99
90% of collaborations	271	269	263	264	234
Degree					
Average	2.49	2.49	2.57	2.59	2.94
Standard deviation	3.28	3.29	3.42	3.44	4.21
Clustering coefficient	0.45	0.42	0.46	0.48	0.53
Giant component					
Number of Nodes	2,264	2,527	2,291	1,697	1,211
% of Nodes	62	65	59	49	42
Isolates					
Number of Nodes	599	619	677	646	617
% of Nodes	16	16	17	19	21

Data

Ph.D. committee network

	2011	2012	2013	2014	2015
Number of Nodes	2,124	2,176	2,220	2,261	2,327
Number of Edges	16,509	17,502	18,329	18,687	19,438
Nodes making up to					
50% of collaborations	179	185	192	198	201
90% of collaborations	408	421	431	443	451
Degree					
Average	15.55	16.09	16.51	16.53	16.71
Standard deviation	12.69	13.05	13.32	13.26	13.53
Clustering coefficient	0.25	0.25	0.25	0.25	0.24
Giant component					
Number of Nodes	2,124	2,176	2,220	2,261	2,327
% of Nodes	100	100	100	100	100
Isolates					
Number of Nodes	0	0	0	0	0
% of Nodes	0	0	0	0	0

Plan of the day

Methods

Our methodological toolbox is composed by:

- Connectivity measures:
  - Degree Centrality (DC);
  - Betweenness Centrality (BC);
  - Local Clustering Coefficient (LCC);
  - Average Path Length (APL).
- Methods for cluster analysis:
  - Girvan and Newman community-detection algorithm (Girvan & Newman, 2002).

These are used to uncover the structural features of the network at both global and local level, by combining network science and SNA.

Methods: Degree centrality (1/2)

Asymmetric interactions

A **multiplicative process** (Caldarelli, 2007) fitting a power law distribution.

A network is generated by a power law if its degree distribution fits the probability mass function of a power law:

$$P(X \leq x) = \frac{\xi(\alpha, x)}{\xi(\alpha, x_{min})} \tag{1}$$

Where:

- x is the degree centrality of node i;
- $\xi(\alpha, x) = \sum_{i=0}^{\infty}(n + x_i)^{-\alpha}$, is the generalized zeta function (Abramowitz & Stegun, 1972);
- n is equal to the number of nodes;
- α is a parameter to be estimated. In small world networks, it typically ranges between 2 and 3.

Clauset et al. (2009) test the statistical significance of α in generating a true power law from function (1) $\rightarrow$ *H0*: data is generated from a power law distribution.

Methods: Degree centrality (2/2)

Network attack (Albert *et al.* 2000)

Highly connected nodes might be critical for the architecture of the network and the dynamics of scientific collaborations (Goyal *et al.* 2006): e.g.

Replacing a scientist collaborating with many laboratories may compromise the chances of his/her colleagues to find in the network new collaborators outside their research group.

Compare the consequences of

- Deleting nodes at random vs
- Deleting nodes featuring high degree centrality.

Methods:
Local Clustering Coefficient (C) + Avg. Path Length (L)

Small World behavior (Watts & Strogatz, 1998)

The network (G) highly clustered (C) and the average shortest distance (L) between nodes is low **compared to** an equivalent random network (ERN).

$$\theta = \frac{C_G}{C_{ERN}}\left(\frac{L_G}{L_{ERN}}\right)^{-1} > 1 \tag{2}$$

Note:
The layers of the multiplex are a one-mode projection of a two-mode network $G(X, Y)$:

- X: the set of UF investigators.
- Y: alternatively the set of papers, grants or Ph.D. students.

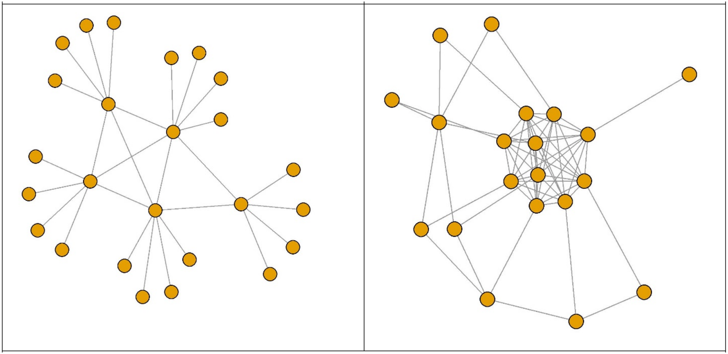
Following Rao *et al.* (1996) and Snijders (2002):

- Reshuffle connections between X and Y, but keep costant nodes' degree centrality (Opsahl *et al.* 2008), obtaining $G(X_1, Y_1)$, then
- Collapse $G(X_1, Y_1)$ into a one mode projection (i.e. the ERN).
  Note: Nodes in G and ERN will have the same degree centrality.

Core-periphery structures and the role of hubs

(Seaquist *et al.* 2014, Leone Sciabolazza, 2018)

Interlinked stars [left panel]:
DC is **positively** correlated with BC and **negatively** correlated with LCC.
Core periphery structure [right panel]:
DC is **positively** correlated with BC and **positively** correlated with LCC.



Methods: DC + BC + LCC (2/2)

Levels of competitiveness, peer pressure, network formation processes
(Burt, 1992; Lindenlaub & Prummer, 2017)

High level of competitiveness: DC is **negatively** correlated with LCC.

- A failed collaboration has small repercussions in the network.
- Agents are risk-lover and compete to to find new collaborators.
- Access to the core of the network is easy.

High level of peer pressure: DC is **positively** correlated with LCC.

- A failed collaboration would lead to frictions not only between project partners, but also between them and their common collaborators.
- Agents are risk adverse and put higher effort in projects characterized by certainty.
- Access to the core of the network is difficult.

Methods: Cluster Analysis

Network partitions (Girvan & Newman, 2002) are used to investigate:

- Temporal effects
  → Investigators maintaining existing collaborations and remaining consistently within the same cluster over the years .
- Overlapping partitions
  → Investigators being in the same cluster in two different layers (e.g. publication and grant network).
- Disciplinary divide:
  → Investigators with common affiliation (department) are within the same cluster.

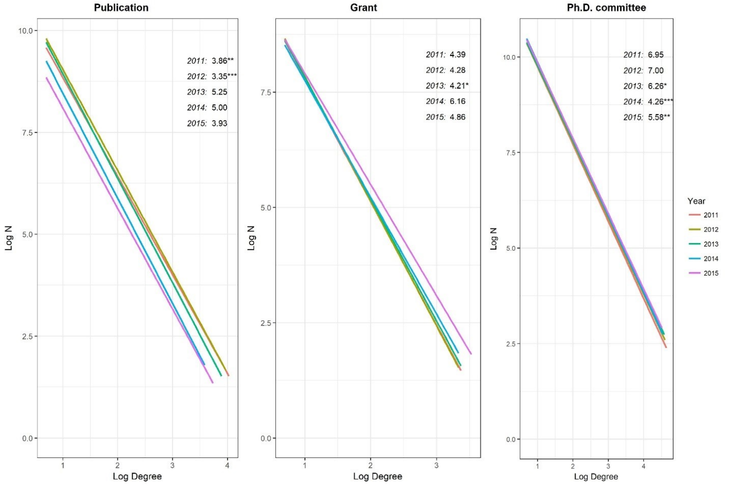
Normalized Mutual Information (NMI) index (Danon *et al.* 2005): it measures the extent to which one partition is explained by another one. It goes from:

- 0: investigators are in different clusters for different partitions (e.g., there is no overlap between cluster of co-authors and co-PIs).
- 1: overlap between different partitions.

Plan of the day

Results: Asymmetric interactions

- In all layers there is a few, but significant number of nodes with many connections, and a trailing tail of nodes with very few connections.
- DC in publication layer is evolving towards a power law.



Results: Network attack test

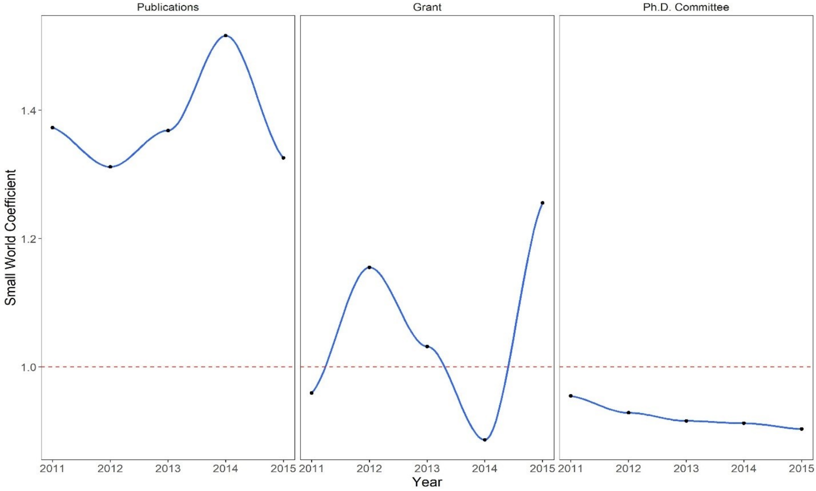
- Random attack: remove 5% of the edges at random.
- Target attack: remove 5% of edges targeting nodes with highest degree centrality.

\# of components generated by the attack

Network	Attack	2011	2012	2013	2014	2015
Publication	Target	64	80	81	79	54
	Random	18	25	31	14	16
Grant	Target	85	69	69	59	45
	Random	20	17	14	20	17
Ph.D. Committee	Target	4	6	6	3	6
	Random	1	1	1	1	1

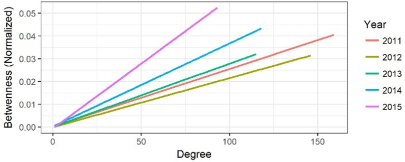
Results: Small World Hypothesis

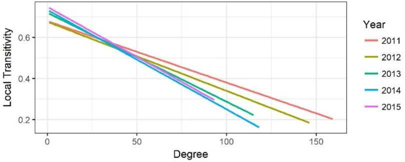
- The publication layer is a small world (Newman, 2001).
- The Ph.D. committee layer is a random network.



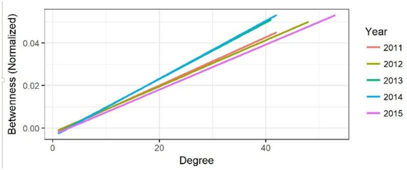
Results: Interlinked stars vs Core-periphery structure

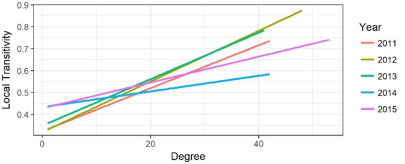
The publication layer is composed by interlinked stars.



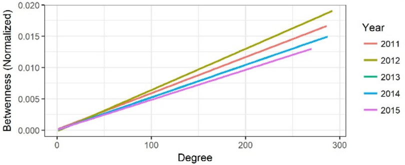


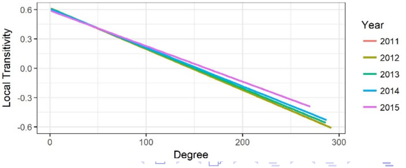
The grant layer is a core-periphery network.



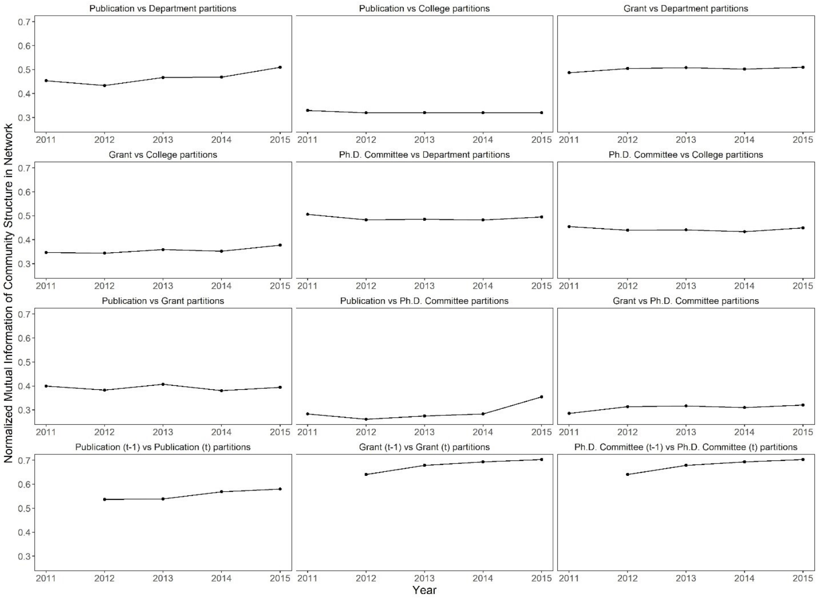


The Ph.D. committee layer is composed by interlinked stars.





Results: Determinants of Partitions (NMI)



Plan of the day

Discussions

The analysis of DC and LCC at the global and local level hints to different features in each layer.

	Publication	Grant	Ph.D. committee
Size	Large number of investigators	Low share of population (access is granted only to those at higher stages of carrer)	
Structure	Interlinked Stars	Core-Periphery	Interlinked Stars
DC vs LCC	Competition (Power Law)	Peer Pressure	Competition
Access	Easy (but time constrainsts)	Hard (risk adversion)	Easy (but time constrainsts)
Trend (NMI)	Small groups becoming denser (small world)	Stable	Stable (random network)

Discussions

The global structure of the network is also hinting to different styles of collaborations in each layer.

- Publication & Grant layer:
  Separated academic sylos (many components).
- Ph.D. Committee layer:
  High rate of interactions across all disciplines (one component).

By looking at partitions, we find that:

- Publication & Grant layer:
  Partitions are strictly related → Investigators tend to stay within their comfort zone (Fortunato *et al.* 2018).
  Note: Co-advising the same students is a form of interaction that rarely creates opportunity for other kinds of collaborations.
- Grant & Ph.D. Committee layer:
  Intra-department relations are as important as inter-department relations → Same level of inter-disciplinarity, different structural organization.

Discussions

The fragilities in publication and grant layers' topology has relevant implications in terms of policy.

Azoulay et al. (2010) found that the sudden loss of highly connected scientists leads to a lasting **5% to 8% decrease** in quality of publications, and it is likely to negatively affect other forms of collaborations.

Academic research networks would benefit from a system of **incentives for highly-connected scholars** to:

- **Remain** in the university maintaining an efficient network of collaborations.
- **Increase the involvement of their collaborators** in research projects, in order to reduce the dependency of the overall network from their own work.

Thank you

valerio.leonesciabolazza@uniparthenope.it
www.valerioleonesciabolazza.com