Statistician Co-Authorship Network

Rabindra Khadka, Alessandro Maregha *SNA,DSSC, University of Trieste*

Abstract

Co-authorship network helps to understand the formation and holding of the co-authorship relationship between professors or researchers associated with different departments in universities.The co-authorship patterns were explored and studied via co-authorship network. The network data was explored by its postional and connectionist features. Then the network cohesiveness ,cut and flow were assessed.Then the communities were detected and inner community density were observed.Finally the ERGM models were created and the received parameters were used to draw underlying probability distributio of our network which was then tested by comparing the simulated results with the observed data.

1.Introduction

Social network is a great tool to explore and discover the dynamics of knowledge flow and collaboration between researchers or professors associated with different departments in universities.This *diffusion of knowledge* between individuals in the scientific community can be extracted by looking at citations in research papers or patents and the *collaborative relations* between researchers or authors can be noted by observing co-authorships,participation in research projects and scientific conferences,co-publishing books,collaboration on proposal writing and some other form of co-producing articles or papers on some topics.

The individual bibiliographic data for this project was retrieved from the national institutional repository and provided to us which highlights the co-authorship between statisticians associated to different departments in italian universities.This affiliation data was used to explore,analyse and discover the colloborative social structure.In the co-authorship network ,nodes represent authors and authors are connected by an edge if they have co-authored one or more papers.These ties are symmetric and this particular network can be vital in detecting communities with their local characteristics which can differ from the global network characteristics.

There are various methods for studying the co-authorship network which includes descriptive statistics,deterministic modelling and others.The choice of the study methods depend upon the objective and the questions posed by the analyst.Some of the basic steps to study co-authorship network could be finding clusters,measuring the size of the components in the network,computing the degree,centrality and closeness measure. Exponential random graph models are also used to understand the dyanmic nature of the co-authorship network and uncover tie formation mechanism. Blockmodelling on co-authorship network can also be applied to unveil the process of formation of network ties over time, dig out some latent pattern and test for homophiliy .

2. Objectives

The objectives of this project include exploring the co-authorship network while drawing **descriptive statistics**, unraveling how the co-authorship network is structured according to **centrality measures**, identifying **influential authors** and discovering how they are tied to each other, ranking by **authors'connectedness and influence** and **detecting communities**.Also the objective is to capture the underlying mechanism of ties formation in the co-authorship network of statisticians

in italian universities using **ERGM models**.Furthermore, understanding the tie formation over time using **block modelling** techniques also comprises the objective of the project work.

Some of the questions posed to motivate our analysis are as follows:

1. Who are the most influential authors in the entire network or so as to say network leaders ?
2. Who frequently collaborates with other individuals between communities?
3. Do actors from cross background work together ?
4. How many authors belong to the largest connected component of the network?
5. How many communities can be detected ?
6. How the local process contributes to the tie formation between authors in our network?

3. The data

The data of authors by papers of scientific collaborations among Italian statisticians was sourced from the national instituional repository.For this project, the data has been divided into two periods , `period 1 (2004-2010) and`period 2 (2011-2017)'. These two periods has been first analyzed seperately then latter the dataset were merged to form a joined network of two periods.

3.1 Descriptive Analysis

The network is `undirected` with network size(vertices=300) for the network `P01`(2004-2010) and network size of vertices equal to 308 for the network `P02`. We took eight attributes for the network namely `SURNAME`,`University`,`Role`,`Localization`,`Faculty`,`Gender`,`Academic_Age`,`Age` and edge weights referring to the number of collaborations two authors had in the network. The output of the summary result can be observed below:

Summary overview of network `P01`

```
## IGRAPH 416f369 UNW- 300 433 --
## + attr: name (v/c), University (v/c), Faculty (v/c), Academic_age
## | (v/n), Role (v/c), weight (e/n)
```

Summary overview of network `P02`

```
## IGRAPH 3e6d915 UNW- 301 461 --
## + attr: name (v/c), University (v/c), Faculty (v/c), Academic_age
## | (v/n), Role (v/c), weight (e/n)
```

The network `P01` is comprised of `88 research assitants`,`113 full professors` and `99 associate professors`.Similarly, network network `P02` is comprised of `87 research assitants`,`115 full professors` and `99 associate professors`. So the distribution of each category of researchers based on their academic roles are fairly similar while comparing two networks.The authors belonged to `28` different faculties from `50` different universities in the network `P01`.Similarly the authors belonged to `29` different faculties from 50 different universities in the network `P02`.

The network `P01` has 300 loops which represents single authored publications and `866` multiple lines representing co-authored publications.Similarly in network `P02`, it consists of 301 loops which represents single authored publications and `922` multiple lines representing co-authored publications.We also observe that the highest number of authors in the both our networks were associated with the faculty `econmoics` as shown below in `fig 1`.

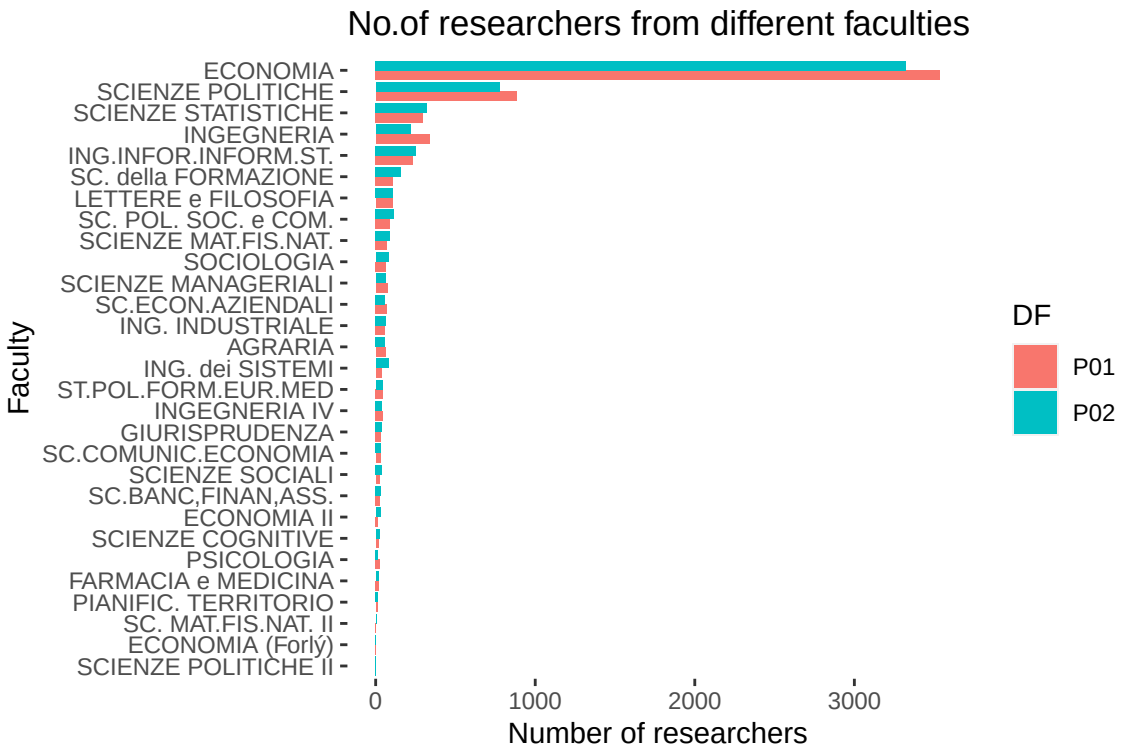

Fig 1: The number of researchers associated with different faculties

We also looked at the trend of the publications over time in both periods.It can be observed in `fig2`below that there was a upward trend of collaborations from the year 2004 to 2007 and downward trend from 2004 -2010. Similarly the collaborations peaked from 2011 to 2012 but declined after 2012 onwards.

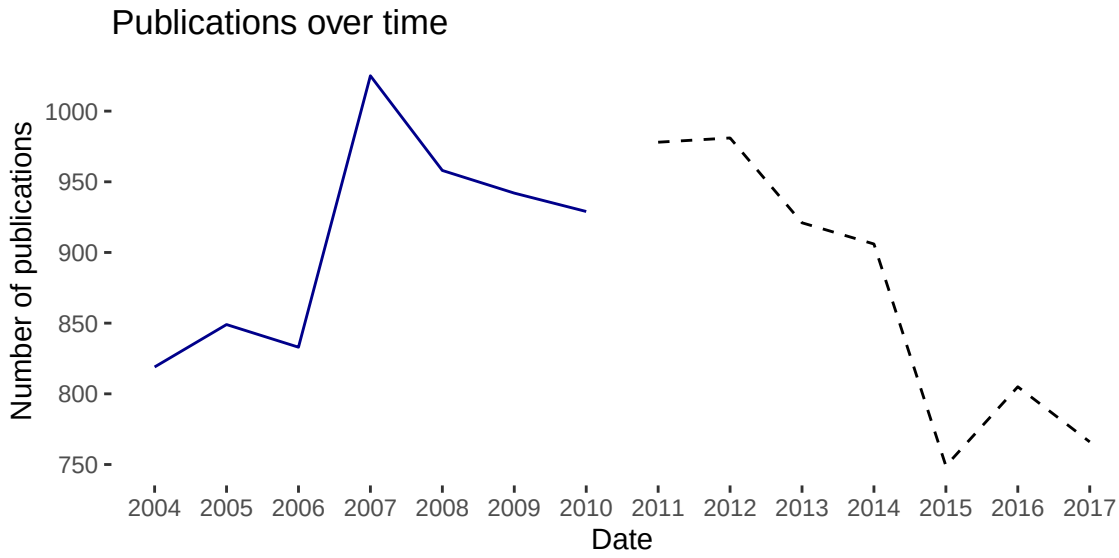

Fig 2: Publications over time from both the network data.

3.2 Network Exploration

Inorder to gain some insight about our networks , we performed some network exploration using tools and libraries namely `igraph` and `statnet`.Network exploration was done with two approaches namely `connectionist feature analysis` and `positional feature analysis`.

3.2.1.Connectionist features

Connectionist features depict how the network has been wired and provide us the potential pattern of information flow in the network. Metrics such as `network size`, `No.of Edges`, `degree`, `density`, `average degree`, `diameter`, `no of clusters`, `transitivity` were observed for both the networks and network formed after joining of `p01` and `P02` .

We observed that both the networks were similar in size with same density but `P02` had lower `average path` and `diameter` which indicates quicker information flow and diffusion of knowledge among authors involved in network period `p02`. However, both the network possess the characteristics of small world network. We saw a marked difference in maximum degree between two networks of different period.Similarly, the joined network reflected similar characterisitics through its connectionist features as obvious reason of being generated form the merging of networks `P01` and `P02`.

Summary Stats. of Perod 1 (2004-2010)

Metrics	Measure
# Vertices	300.00
# Edges	433.00
# Density	0.01
# Diameter	23.00
# Avg_path	6.95
Transitivity	0.36
Max.degree	12.00
# Avg_degree	2.89
# N_clusters	45.00
# Nodes in largest component	220.00
# Isolates	33.00

Metrics	Measure
# Vertices	301.00
# Edges	461.00
# Density	0.01
# Diameter	14.00
# Avg_path	5.45
Transitivity	0.34
Max.degree	20.00
# Avg_degree	3.06
# N_clusters	58.00
# Nodes in largest component	219.00
#Isolates	43.00

Summary Stats. of joined network (2004-2017)

Metrics	Measure
# Vertices	308.00
# Edges	682.00
# Density	0.01
# Diameter	11.00
# Avg_path	4.67
Transitivity	0.32
Max.degree	25.00
# Avg_degree	4.43
# N_clusters	28.00
# Nodes in largest component	278.00
# Isolates	24.00

We also observed the degree distribution of the networks which help us to assess the estimate the probability of having degreek when a node is randomly choosen.The degree distribution for our networks resulted in a right skewed slightly fat tail distribution. This suggests that our networks have a good resilient to random removal of higher degree nodes(hubs).So, removal of a node will have a negligible effect on the cohesion of the network. we also found that the distirbution `asymptotically` followed `power law distribtion` with `alpha` value aroud '2.

We also examined `Avg Neighbourhood degree` vs `Node Degree` and found that the actors with higher degrees tend to link with mostly the actors having degree above the average degree.While actors with low degree are linking up both with lower and higher degree actors. This highlights that experienced, well known authors are collaborating with authors having similar profile mostly . While the authors having few connections are collaborating with prolific authors but mostly with authors similar to their profile.

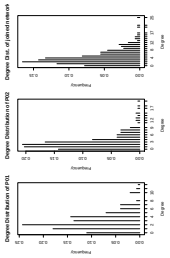

**Fig 4: Degree distributions of network `p01`, `p02` and of theirs merged network .

3.2.2 Positional Features

Positional features are the result of connectionist features i.e how the actors are wired together.These features will help us to get an idea how some specific postitions occupied by nodes are contributing towards the network formation or flow of information.We captured positional features with measures like degree centrality,closeness centrality and betweeness centrality .

`Degree centrality` was used to measure the number of collaboration an author has made and rank the authors based on number of collaborations. Higher degree indicates good communication with others and also the ability to play an influential or central role in the network. The author named `BRENTARI`,`DI CIACCIO` and `VICHI were ranked at first place in networks`P01 , P02 and`joined` nework' with degree centrality of 12,20,25 respectively.

`Betweeness centrality` takes into account the number of times a node appears in the shortest path.The node with high betweeness will act as a bridge or connector between a pair of nodes.Coincidingly, author named `BARTOLUCCI` took the first rank in all our networks `P01`,`P02` and `joined nework` with betweeness centrality of 0.28,0.12,0.24 respectively.

`Closeness centrality` have the notion that actors who are able to reach other actors at shorter path lengths, or who are more reachable by other actors at shorter path lengths have favored positions. It will allow us to measure independence i.e the measure of capacity to reach authors without relying too much on other intermediate authors.The author named `VITTADINI`,`RAMPICHINI` and `DI CIACCIO` were ranked at first place in networks `P01`,`P02` and `joined nework` with closeness centrality of 0.0105,0.108507,0.242266 respectively.

We also observed that most of the top ranking authors in a centrality measures also appeared in the top rank of the rest of the centrality measures.

3.2.3 Edge Betweenness

It captures the notion that a relation is between based on that it is part of the geodesic between pairs of actors. We estimated the importance of collaborations for the flow of information in the

network and observed the top ten collaborations in our network.We computed edge betweenness to assess which co-authorship collaborations are important for the flow of information.

Top 10 most important edges based on edge betweenness for network P01

```
## + 10/433 edges from 416f369 (vertex names):
##  [1] BARTOLUCCI--VITTADINI         BARTOLUCCI--MONTANARI GIORGIO
##  [3] BARTOLUCCI--MIRA              BARTOLUCCI--GRILLI
##  [5] MIRA      --PETRONE           PETRONE   --VERONESE
##  [7] CONSONNI  --VERONESE          VIOLA     --VITTADINI
##  [9] CAMPOBASSO--VIOLA             CONSONNI  --ROVERATO
```

Top 10 most important edges based on edge betweenness for network P02

```
## + 10/433 edges from 416f369 (vertex names):
##  [1] GAETAN           --MASAROTTO     IPPOLITI         --NISSI
##  [3] CAFARELLI        --POLLICE       AMENTA           --BRENTARI
##  [5] BARAGONA         --BATTAGLIA     CIAVOLINO        --D'AMBRA LUIGI
##  [7] MORLINI          --ZANI          CARPITA          --GOLIA
##  [9] GIORDANO GIUSEPPE--MISURACA      PRATESI          --ROCCO
```

Top 10 most important edges based on edge betweenness for joined network

```
## + 10/433 edges from 416f369 (vertex names):
##  [1] GAETAN           --MASAROTTO     IPPOLITI         --NISSI
##  [3] CAFARELLI        --POLLICE       AMENTA           --BRENTARI
##  [5] BARAGONA         --BATTAGLIA     CIAVOLINO        --D'AMBRA LUIGI
##  [7] MORLINI          --ZANI          CARPITA          --GOLIA
##  [9] GIORDANO GIUSEPPE--MISURACA      PRATESI          --ROCCO
```

We observed that the collaboration between `GAETAN` and `MASAROTTO` was annoated with highest edge betweeness indicating a very good flow of knowledge between them.

3.2.4 Connectivity, Cuts, and Flows

Connectivity is significant as it carries the notion of how easy is to go from one node to another and how resilience is the network.Cuts refer to the idea that removing certain set of vertices (edges) will disconnect the graph and a single vertex that disconnects the graph is known as vertex-cut.Analyzing these connectivity and cut points , we can overview how well the information is flowing in the network.

The number of connected components were observed in all of our participant networks.

```
##
##   1   2   3   5   6   8 220
##  33   3   2   3   2   1   1

##
##   1   2   3   4   7 219
##  43   9   2   2   1   1
```

```
##
##   1   2 278
##  24   3   1
```

We found that the largest component in our`P01` network has 220 nodes which is $220/300 \approx 73.33\%$; similarly for `P02` largest component took $\approx 73\%$ of the total share of nodes. The joined network have 3 connected components with $\approx 93\%$ of the total nodes in the network.

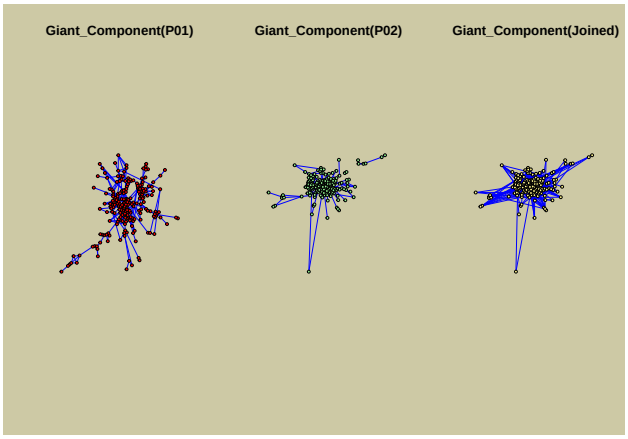

Fig 6: Networks of giant component of our three networks.

We also observed the density and average path of the giant compoents .We found the clustering is high and average path distance is low which resembles the small world network characteristics.We also observed that the giant component can represent well the entire network as the way it had connected to its nodes resembled how the entire network had been connected.

In terms of the connectivity , the giant components had the vertex connectivity and edge connectivity is equal to 1 for the networks **thus requiring the removal of only a single well-chosen node (author) or 1 collaboration ties in order to break the subgraph into additional components**. We also observed the set of vertex cut points for all our candidate networks.These are known as `articulation points` which can be regarded as the set of most important authors in our network.

Articulation point for network P01

```
## + 57/220 vertices, named, from c75bf01:
##  [1] PETRELLA          TONELLATO         TANCREDI          CONIGLIANI
##  [5] LISEO             COZZUCOLI         DE BATTISTI       SALINI
##  [9] PERRI             ALFO'             BARAGONA          BOCCI LAURA
## [13] PICCOLO           ROCCI             GIRONE            SPADA
## [17] PALUMBO           PROVASI           GRIGOLETTO        GAETAN
## [21] MASAROTTO         GUSEO             DALLA VALLE       FURLAN
## [25] VENTURA           GIUMMOLE          TORELLI           PAGANI
## [29] DELDOSSI          ZAPPA             INGRASSIA         MUCCIARDI
## [33] MAZZA             LOPERFIDO         MINOZZO           BARTOLUCCI
## [37] FREDERIC          PATERLINI         POLI              LOVISON
## + ... omitted several vertices
```

Articulation point for netowork P02

```
## + 60/219 vertices, named, from 714b4c0:
##  [1] BRENTARI         PICCOLO          BATTAGLIA        MUCCIARDI
##  [5] OTRANTO          MAZZA            INGRASSIA        MAROZZI
##  [9] FONTANA          TORELLI          LA ROCCA LUCA    CONSONNI
## [13] DELDOSSI         ZAPPA            ZANAROTTI        PAGANI
## [17] DOMMA            GIORDANO SABRINA BARABESI         DE IACO
## [21] MARIELLA         CIAVOLINO        GALLO            MOLA
## [25] LA ROCCA MICHELE PORZIO           RAGOZINI         D'URSO
## [29] VICHI            LAGAZIO          BIGGERI ANNIBALE MUSIO
## [33] VENTURA          AGOSTINELLI      GRECO LUCA       PETRONE
## [37] GRIGOLETTO       VICARI           PALUMBO          CAFARELLI
## + ... omitted several vertices
```

Articulation point for joined network

```
## + 40/278 vertices, named, from 4756456:
##  [1] COZZUCOLI        PERONE PACIFICO  DE SANTIS FULVIO VENTURA
##  [5] FURLAN           MASAROTTO        PETRELLA         TONELLATO
##  [9] PICCOLO          FREDERIC         PATERLINI        DE IACO
## [13] MARIELLA         BERTOLI BARSOTTI RIBECCO          DI CIACCIO
## [17] PROVASI          TORELLI          VICHI            DE BATTISTI
## [21] SALINI           VICARD           TARANTOLA        VERONESE
## [25] PETRONE          MEZZETTI         MULIERE          D'URSO
## [29] MUCCIARDI        CAROTA           TANCREDI         ROGANTIN
## [33] FONTANA          SCLOCCO          SPADA            COCCHI
## [37] BARABESI         LAGAZIO          PACINI           BARTOLUCCI
```

This shows that the network vulnerabality in `P01`,`P02`and `P03` is dependent on 19%,20% and 12.9% of the total nodes in networks `P011` ,P02`and`joined network'.

4. Community Detection

We applied various community detection algorithm such as `Louvain`, `Grivan-Newman` and `label propagation`. Louvain algorithm was choosen based on the best modularity given by the algorithm. We applied community detection algorithm to the giant component of our network. Then the denisty of communities along with the author having maximum degree in the community was extracted as shown below.

Modularity of different algorithm applied on giant component

Modularity Comparison (Giant Comp. joined)

Method	Communities	Modularity
Louvain	**22**	**0.8531343**
Grivan-Newman	15	0.6640077
Label Propagation	66	0.4317382

Density and author with maximum degree of different communities of GC

Summary (Giant Comp. joined)

Communities	Community.sizes	Freq	Density	Author	Max_degree
C1	1	16	0.0123077	COSTANZO	2
C2	2	5	0.0210526	AGRO'	1
C3	3	12	0.0000000	ALFO'	0
C4	4	7	0.0065359	GIUDICI PAOLO	1
C5	5	3	0.0250000	AMENDOLA	2

5.Exponential Random Graph Model :

Exponential Random Graph Model are models based on exponential family which helps us to draw the probability distribution of our network.We defined first the null model and then gradually included covariates representing homophily, triads and other choosen structural features. Then the coefficent parameter were obtained and the probability was deduced for each parameter.Then we simulated the new network from the underlying distribution given by the fitted model and finally the goodness of fit were observed.

The result of our ergm models is seen below:

Parameter coefficients of various models fit on giant component of the joined network.

	Model 1	Model 2	Model 3	Model 4
edges	-3.68 ***	-5.84 ***	-3.67 ***	-5.75 ***
	(0.04)	(0.14)	(0.07)	(0.15)
gwesp.fixed.0.25		2.21 ***		2.16 ***
		(0.11)		(0.11)
absdiff.Academic_age			-0.00	-0.00
			(0.01)	(0.00)
nodematch.Faculty				0.09
				(0.05)
AIC	5765.35	4936.75	5767.30	4942.18
BIC	5773.48	4953.00	5783.55	4974.68
Log Likelihood	-2881.68	-2466.37	-2881.65	-2467.09

*** p < 0.001; ** p < 0.01; * p < 0.05

We can interpret the coefficients of this model in terms of the probabilities of different types of ties by taking expit or inverse logit: the probability of a tie that is completely heterogeneous i.e the two members differ from each other in academic age is 0.023, the probability of a tie that is homogeneous by minimum academic age only is 0.0245 and maximum academic age only is 0.022 similarly one that is homogeneous in all three attributes is interms of log odds can be calculated as -2.30 .

To assess how our models captured the structure of our original network, we performed simulation by using the model fit generated above.We then tried to fit our model on another new feature which was degree distribution and compare to see how our model performed.

6. Summary

We observed that the connected component represented the entire network very well with large proportion of nodes w.r.t entire network .So the largest connected component clearly possesd community structure of the co-authorship network. We also found that the communities inside the connected component had fairly good density and authors had a great connnections inside the detected community.So, we can state with good confidence that the authors from a commmunity more often collaborated with authors from the same group.There was also biconnected component which provided a fairly stable component which was used for ergm application. The coefficient of

ergm models could be improved by increasing the iterations,more data nodes and manipulating MCMC diagnostics. Additionally , egocentric sampling of data could also help us to improve and extend our model to the entire network or larger population with limited compute capacity.

Finally, it would also be intersting to do further work on to explore how the network evolved over time by looking at the publication of articles at different time periods.