Scientific collaboration dynamics in a national scientific system

Anuška Ferligoj
Luka Kronegger, Franc Mali
Tom Snijders, Patrick Doreian

SUNBELT 2015
Brighton, June 23-28, 2015

Outline

1. Introduction
2. Hypotheses
3. Model specification
4. Data
5. Results
6. Conclusions

Introduction

Scientific collaboration has been studied systematically since the 1960s.

Slovenian researchers have studied the last seven years scientific collaboration using

- bibliometric analysis (Ferligoj and Kronegger 2009; Kronegger et al. 2011, 2012, 2014)
- survey analysis (Iglič et al. 2014)
- qualitative approach (Groboljšek et al. 2014)

of co-authorship networks using longitudinal data on the Slovenian science system in order to explore and explain their dynamics across four scientific disciplines:

- **Mathematics** - an old discipline where research takes place primarily in offices
- **Physics** - an old discipline where the research occurs mostly organized into research groups within laboratories
- **Sociology** - an old discipline where research also occurs mostly in offices
- **Biotechnology** - a new laboratory discipline

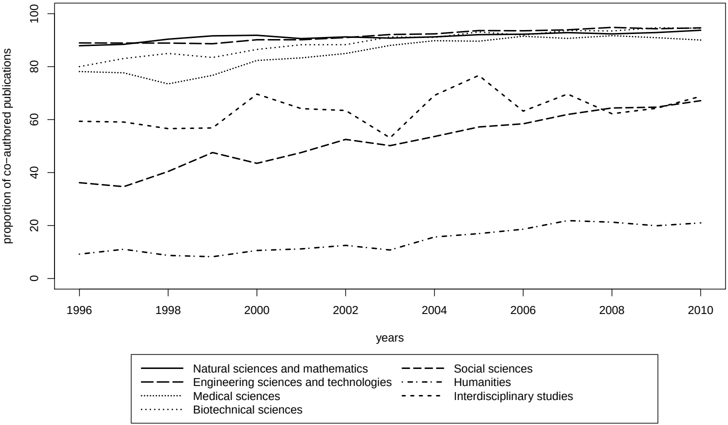
Goal

- The goal of this presentation is to identify the key factors driving collaboration and the main differences in collaboration behavior across **all** scientific fields and disciplines.

Seven scientific fields in Slovenia

ID	Scientific field	No. of disciplines
1	Natural sciences and mathematics	9
2	Engineering sciences and technologies	19
3	Medical sciences	9
4	Biotechnical sciences	6
5	Social sciences	11
6	Humanities	12
7	Interdisciplinary studies	2

Percentages of co-authored publications in seven scientific fields in Slovenia from 1996 to 2010



Theoretical arguments and hypotheses

Since the early work of Price (1963, 1965) and Garfield and Merton (1979), sociologists introduced several theories regarding with scientific collaboration. Here, we focus on

- the theory of cumulative advantage in science, termed the Matthew effect (Merton 1968, 1973; Price 1976) and
- the theory of small-world structure (de Sola Pool and Kochen 1978)

and their applications to the modelling of the dynamics of co-authorship networks.

Small-world model

The small-world model was defined formally by Watts and Strogatz (1998) who introduced an algorithm to construct networks with the following properties:

- having short paths between any two vertices and
- incorporating clustering (small dense parts of the network).

Here, we deal with the clustering level.

H_1 : **The co-authorship networks in the Slovenian scientific community have a high clustering level driven by transitive closure processes, where co-authors of co-authors become, or remain, co-authors.**

Preferential attachement mechanism

The idea of cumulative advantage or preferential attachement implies that excellent scientists are rewarded far more than others in their field. Said *et al.* (2008) noted that young researchers more likely form new co-authorship ties with older, established researchers, usualy their mentors. The formal modelling of preferential attachment as the driving mechanism of co-authorship was examined also by Barabasi (1999).

H_2 : **New co-authorship collaborations of Slovenian researchers are more likely for authors who have more current co-authorships, and for excellent researchers; for co-authorships, this holds both for collaboration within Slovenia and with researchers abroad.**

Individual and organizational contexts

The hypothesis that individual and organizational contexts drive the formation of scientific co-authorship networks was confirmed by Kronegger *et al.* (2012). They showed that the four disciplines were affected in different ways by the organization of local institutions and disciplinary publishing cultures.

H_3 : **Individual and organizational contexts in Slovenia drive the formation of scientific co-authorship networks.**

Model specification

The three hypotheses were tested using an actor-oriented model (Snijders, 2001, 2005; Snijders et al., 2007, 2010) used for longitudinal network data. The model is defined as a continuous-time Markov process.

Since our data are **non-directed networks**, a modification to the models of Snijders (2001, 2005) is required. To obtain a non-directed network, the assumption is made that at random moments, a randomly chosen actor ('ego') chooses another actor ('alter') to propose a new tie or to drop an existing tie; if a new tie is proposed, alter can decide to accept or reject the proposal (see Snijders, 2008). The choice by ego of alter is a multinomial choice, and the acceptance decision by alter is a binary choice.

Program

The probability models for these choices are based on a linear predictor similar to generalized linear models.

Stochastic-actor-based model (SAOM) is implemented in the SIENA program.

We used RSiena.

Operationalization of clustering

- Intuitively, the clustering can be viewed by the average probability of two co-authors of a researcher will collaborate also. As colloaboration can be viewed as a consequence of transitivity we included the tendency of actors to form **transitive triads** in the model.
- Co-authorship can also be driven by departmental and institutional affiliation, we measured this by working in the **same organizational research group**.

Operationalization of preferential attachement

- Preferential attachment can be observed through the eyes of a single unit and can be modeled as the effect of the **degree parameter** on the production of new ties.
- As the degree parameter of alters captures only collaborations inside the networks (inside scientific fields or scientific disciplines), we also included in the model **alters' collaboration outside the national collaboration network**. This variable was very skewed and so we used its logarithm.

- **Alters' scientific excellence** is measured by a dichotomous variable where 1 means that the researcher has at least one publication published in the top international scientific journals.
- **Scientific age** is defined as the year of an author's first publication
- **Scientific age similarity** is the year of first publication similarity

Data

The data were organized in three 5-year intervals:

- **Period 1, 1996–2000**: a period of harmonization with the European Union (EU) and the OECD standards;
- **Period 2, 2001–2005**: in 2004, Slovenia became a member of the EU. The Slovenian Research Agency was established in the same year followed by many positive effects on R&D evaluation procedures due to its policies;
- **Period 3, 2006–2010**: a more stable period.

Dataset

- **Current Research Information System (SICRIS)** which includes information on all current and former researchers registered with the Slovenian Research Agency and
- **co-operative On-Line Bibliographic System & Services (COBISS)** which is an officially maintained database of all publications available in Slovenian libraries. From this system, we collected complete scientific bibliographies of all Slovenian researchers who had ever been given a research identification number (ARRS ID) by the Slovenian Research Agency.

The network was defined in three consecutive observations and a tie was defined if two researchers appeared together as authors in at least one publication.

- The total number of researchers with an ARRS ID who published in the time period 1996–2010 was 15,424.
- These researchers collaborated with another 48,191 authors not registered with ARRS.
- Together, they published 170,118 publications that are, according to the evaluation criteria of ARRS, treated as scientific outputs.
- The data about discipline memberships were provided by the researchers themselves when they applied for an identification number.

Basic characteristics in networks of research fields

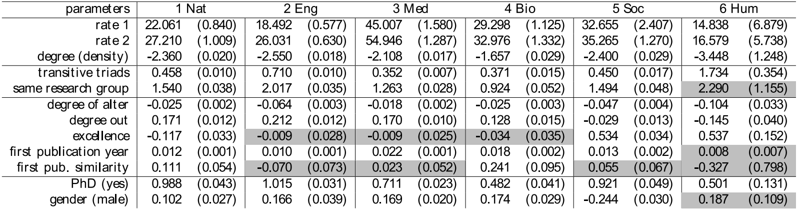
Field		period			Average degree			Researchers		
No.	Name	1	2	3	1	2	3	all	connected	% exclud
1	Natural sciences and mathematics	1538	1795	2089	2.8	3.8	5.13	2585	2294	11
2	Engineering sciences and technologies	2355	2649	2994	2.47	3.25	4.44	4040	3762	7
3	Medical sciences	1470	1636	1720	4.53	5.77	6.62	2144	1978	8
4	Biotechnical sciences	769	797	919	3.03	4.44	5.99	1192	1108	7
5	Social sciences	1309	1648	1830	1.76	2.34	3.23	2193	1718	22
6	Humanities	996	1226	1350	0.39	0.6	1.3	1556	736	53

Results

Stochastic-actor-based model (SAOM) is useful, within a micro-level view, for analyzing observed co-authorship networks.

Scientists collaborating at one point in time can choose their co-authorship tie at a later time. We considered the possibility that ties can be **created or maintained** since this is a feature that characterizes co-authorship networks.

Estimated parameters for the six scientific fields



parameters	1 Nat		2 Eng		3 Med		4 Bio		5 Soc		6 Hum	
rate 1	22.061	(0.840)	18.492	(0.577)	45.007	(1.580)	29.298	(1.125)	32.655	(2.407)	14.838	(6.879)
rate 2	27.210	(1.009)	26.031	(0.630)	54.946	(1.287)	32.976	(1.332)	35.265	(1.270)	16.579	(5.738)
degree (density)	-2.360	(0.020)	-2.550	(0.018)	-2.108	(0.017)	-1.657	(0.029)	-2.400	(0.029)	-3.448	(1.248)
transitive triads	0.458	(0.010)	0.710	(0.010)	0.352	(0.007)	0.371	(0.015)	0.450	(0.017)	1.734	(0.354)
same research group	1.540	(0.038)	2.017	(0.035)	1.263	(0.028)	0.924	(0.052)	1.494	(0.048)	2.290	(1.155)
degree of alter	-0.025	(0.002)	-0.064	(0.003)	-0.018	(0.002)	-0.025	(0.003)	-0.047	(0.004)	-0.104	(0.033)
degree out	0.171	(0.012)	0.212	(0.012)	0.170	(0.010)	0.128	(0.015)	-0.029	(0.013)	-0.145	(0.040)
excellence	-0.117	(0.033)	-0.009	(0.028)	-0.009	(0.025)	-0.034	(0.035)	0.534	(0.034)	0.537	(0.152)
first publication year	0.012	(0.001)	0.010	(0.001)	0.022	(0.001)	0.018	(0.002)	0.013	(0.002)	0.008	(0.007)
first pub. similarity	0.111	(0.054)	-0.070	(0.073)	0.023	(0.052)	0.241	(0.095)	0.055	(0.067)	-0.327	(0.798)
PhD (yes)	0.988	(0.043)	1.015	(0.031)	0.711	(0.023)	0.482	(0.041)	0.921	(0.049)	0.501	(0.131)
gender (male)	0.102	(0.027)	0.166	(0.039)	0.169	(0.020)	0.174	(0.029)	-0.244	(0.030)	0.187	(0.109)

Shaded estimates are not statistically significant;

there are standard errors in parentheses.

The first three parameters

The first three parameters are technical requirements of the stochastic-actor-based model:

- the rate parameter for the first transition;
- the rate parameter for the second transition, and
- the density parameter.

Estimated parameters for small-world

The **transitive triads** effect is positive and significant for *all* fields showing that scientists tend to form new co-authorship ties with the co-authors of their co-authors inside the scientific field.
The estimated parameter on '**belonging to the same research group**' is also positive and significant in five scientific fields (it is positive but not significant in the Humanities).

These estimates provide irrefutable confirmation of a high level of clustering within co-authorship networks in the Slovenian scientific community.

Estimated parameters for preferential attachement

- The parameter for '**degree of alter**' is significant and negative for all fields. Researchers do *not* tend to form new ties with those researchers who collaborate more within the field.
- When considering the parameter '**publishing out of the field**', the Social sciences and the Humanities contradict H_2 while it is supported in the remaining scientific fields .
- '**Publication excellence**' has no effect on co-authorship in the Technical, Medical and Biological sciences. It has a negative effect in the Natural sciences. However, the effect of excellence on tie formation is positive in the Social sciences and the Humanities.

Estimated parameters for student - mentor relationship

The estimated parameter for the '**year of authors' first publication**', is positive and significant for all field except the Humanities. The effect of '**first publication similarity**' is also positive and significant for the Natural science and mathematics and Biotechnology. In these two fields young researchers more likely establish co-authorship ties with young not yet excellent colleagues. In the other scientific fields a similar tendency of connecting young researchers with younger collegues is present but with lesser effects.

This result does *not* follow the standard hypothesis claiming young researchers form new co-authorship ties with scientifically excellent older scientists.

Scientific disciplines

The next question is whether these results hold for the scientific disciplines. For this purpose, we estimated the SAOM models for most of the disciplines.

Excluded disciplines (14 out of 72)

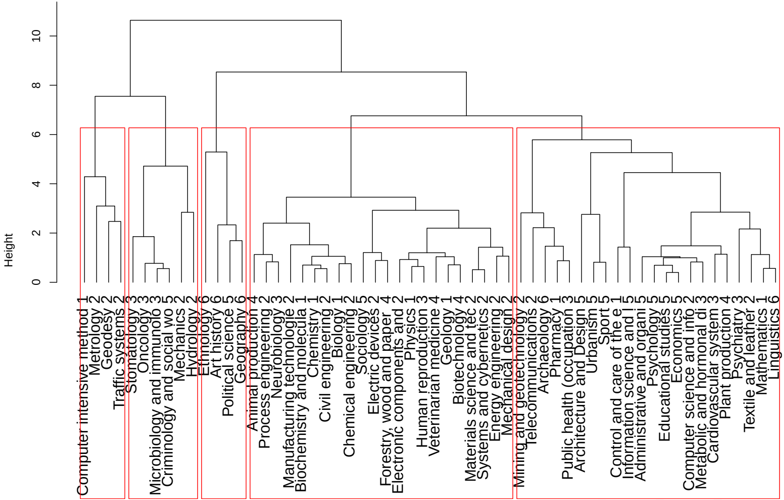
- Technology driven physics (30), Communications technology (31), Landscape design (45) and Ethnic studies (57) - too small numbers of reseachers.
- Anthropology (62), Culturology (65), Literary (66), Musicology (67), Philosophy (69), and Theology (70) - having few researchers (all less than 30) and few co-publications (average degrees less than 1).
- Law (51) - deviating data structure: each wave had papers having a very high number of authors.
- Historiography (60) - high proportion of missing values in variables for actor properties.
- The NCKS Research programme (72) and Interdisciplinary research (73) - lacks an established field structure (Interdisciplinary studies).

Clustering scientific disciplines

The estimated parameters are not directly comparable across disciplines due to variations in the size of the disciplines.

- While the starting point is the set of the estimated parameters, we transformed them to measure the importance of the estimated parameters using the proposed method of Indlekofer and Brandes (2013).
- These values ignore the sign of the estimated parameters for disciplines. For disciplines having negative estimated parameters, the sign of the importance measure was multiplied by -1.
- These measures were standardized before obtaining the Euclidean distances for each pair of disciplines.
- The clustering used Ward's hierarchical clustering procedure (Ward, 1963).

Hierarchical clustering of the scientific disciplines



Obtaned clusters

1. ENG,NAT - four disciplines from the technical and natural sciences.
2. ENG,MED - all but one are disciplines from the technical and medical fields.
3. HUM,SOC - disciplines from the social science and humanities.
4. NAT,ENG,MED,BIO - disciplines that can be viewed as natural and technological sciences with one exception (Sociology)
5. RESID - residual cluster with disciplines from all of the first six fields.

Averages of the importance coefficients for each obtained cluster

	triads	same rG	deg alter	degOut	Excel
1 ENG,NAT	0.12	0.48	-0.23	0.30	-0.25
2 ENG,MED	0.19	0.29	-0.53	0.13	-0.01
3 HUM,SOC	0.15	0.38	-0.17	-0.12	0.30
4 NAT,ENG,MED,BIO	0.12	0.27	-0.18	0.09	-0.05
5 RESID	0.07	0.18	-0.08	0.05	-0.00

Conclusions

- The first hypothesis about the presence of a clustering level as a dimension of a small-world structure was confirmed.
- The evidence regarding the second hypothesis concerning preferential attachment as the driving mechanism of co-authorship is mixed. Our results show the **distance between researchers who collaborate matters**:
  - Alters' high degree of co-authorship **inside** the field or discipline has a *negative* effect on new tie formation in all scientific fields and nearly all scientific disciplines.
  - Alters' higher degree of **outside** collaboration has a negative effect on new tie formation in the Social sciences and Humanities but a positive effect in the other fields.

- Alters' publication excellence on formaing collaborative ties has a positive effect in the Social sciences and Humanities but the negative effect in the other fields.
- The standard hypothesis that young researchers form new co-authorship ties with scientifically excelent older scientists was not confirmed.
- The third hypothesis was confirmed and the evidence demonstrates that the scientific fields and disciplines are affected by the organization of local institutions and publishing cultures.

These findings hold for broad fields and also for the disciplines.

The differences between the two basic pools of scientific knowledge (i.e., between the natural and technical sciences, and the social sciences and humanities) according to the mechanism of preferential attachment can also be the result of contextual (research policy) factors operating in Slovenia:

- In the former socialist era, due to ideological pressure on the social sciences and the humanities, these disciplines were less internationalized and much less oriented to publish in high-ranking international journals.
- After 1991, Slovenian R&D policy gradually began to introduce the criteria of (international) excellence into R&D evaluation procedures. It seems that in adapting to the more demanding R&D evaluation criteria there were different impacts on the social sciences and humanities compared to the other fields.