Formation and persistence of research communities in Middle Income Countries: The case of South Africa

L. Rivera León[a], R. Cowan[b], M. Müller[c]

[a] *UNU-MERIT, Maastricht University, Netherlands*

[b] *BETA, Université de Strasbourg, France; Institut Universitaire de France; UNU-MERIT, Maastricht University, Netherlands*

[c] *BETA, Université de Strasbourg, France*

ABSTRACT

This paper tests empirically the different factors associated with tie formation and tie persistence of network co-authorships using a longitudinal dataset of the National Research Foundation (NRF) of South Africa. We highlight three factors: status homophily and cognitive proximity, organisational proximity, and social and community structure. We find that tie formation is favoured for researchers that are well connected in the co-authorship network, have high cognitive and organisational proximity, and are part of the same social communities. In contrast, our results show that nothing matters apart from cumulative advantage and social structure to make a tie persist.

Keywords: research collaboration, tie formation, tie persistence, scientific communities, middle-income countries, South Africa

1. INTRODUCTION

Many theories have been proposed to explain what drives knowledge creation in the context of innovation and technical change. There have been important changes in the way knowledge is produced, particularly regarding the organisational structure of higher education institutions and public research centres, and the way scientific knowledge is financed and developed. In line with these changes, there is an increasing importance attached to collaboration in the process of knowledge creation: the production of knowledge no longer depends on isolated interests of inventors and scientists, but it depends on and is a result of the exchange and free circulation of knowledge among different actors. This has implications for the spatial diffusion of knowledge, both locally and globally. It is thus central to understand the characteristics of knowledge creation patterns, and the rigidity of the interactions at micro, meso and macro levels that hinder or facilitate knowledge diffusion.

Publications are increasingly important, if not the most important, measure of research productivity, and thus there is more and more interest in understanding the determinants of collaboration and research outputs at individual and collective levels (i.e. departments, universities, countries). Understanding these determinants can help in the design of research policies focusing on how to boost the quality and quantity of publication outputs. This is an issue of particular concern in Middle Income Countries (MICs) as their science systems are typified by scarcity of resources for research, a research system mainly financed with public funds, limited knowledge access, differences in publication polices, limited scientific infrastructure, local research priorities that may differ from global concerns, and so on.

Social network analysis is a tool that can contribute theoretically and empirically to the understanding of these determinants. The dominant concern of existing research is how collaboration happens (i.e. the factors of tie formation) (Snijders, 2001, Levi-Martin and Yeung, 2006, Baldassari and Diani, 2007, Coleman, 1974). However, there is little research focusing on how collaboration persists over time. The determinants of collaborating for the first time may be different of those of sustaining the collaboration. Moreover, given the existence of (relatively) small scientific communities in MICs and tight relationships among researchers — beyond institutional and formal structures, social and community structure, or what Frenken et al. (2009) call *social proximity* — could also be important for understanding research collaborations.

This paper investigates the different factors associated with tie formation and tie persistence affecting collaborative scientific research in South African research networks. While some authors argue that the factors determining the formation of ties are similar to those of the persistence of the relationships (McPherson et al., 2001), other research on interorganisational networks suggest that the two mechanisms are structurally different (Seabright et al., 1992). Our paper gives an important place to investigating the effects of social structure (interdependencies, authority structures, norms, organisations and other features) on scientific collaborations. Following Coleman's *structural individualism*, we are interested in understanding the effects of social organisation by exploring the constituent elements of research collaborations, which lie beneath the system level.

Research gap

In recent years there has been an increase in interest in the economics of science. Most of the work, however, is done in the context of rich countries, with well-developed science systems. Very little work is directly concerned with these issues in middle-income (much less poor) countries. But the constraints under which science systems operate in these countries, and the important role they might play in technological upgrading, and thus in economic growth in general, implies that there I an important lacuna in our understanding. This paper aims at this gap.

Our paper gives an important place to the effects of social structure on scientific collaborations. A researcher's place in the network of collaborations may affect his or her ability or propensity to form or maintain collaborative ties. In particular, we are interested in whether or not membership in a community of researchers has an impact on tie formation and persistence. We adopt some of the relatively new techniques for community detection in networks generally to investigate the meso-level effects of social structure on scientific collaboration.

2. SCIENTIFIC COLLABORATION: FORMATION AND PERSISTENCE

We analyse three different factors of tie formation and tie persistence: (1) status homophily and cognitive proximity, (2) shared organisational foci and institutional constraints, and (3) social and community structure.

Homophily in status and cognitive proximity

While the Platonic form of the researcher as someone selflessly pursuing truth wherever and however he might find it, researchers do, (as well) have personal motivation. Careers and reputation matter. In addition to the goal of finding truths about the world, in their choice of collaborators researchers will pay attention to whether or not the potential collaborator brings valuable resources (most typically in the form of human capital) to the project. Researchers look for the "best" possible partners, where "best" may have a very broad definition. We might expect to see a an assortative pairing, in which better researchers partner with each other, as do less good ones. Thus partnerships that form are likely to exhibit some status homophily. Partnerships that persist, on the other hand, are likely to exhibit *cognitive proximity* given that trust has been built and value has been observed due to the original collaboration.

In addition to status, shared attributes are a factor determining collaboration, and research has found that people tend to associate with others of similar age, gender and ethnicity (McPherson et al., 2001). Shared attributes are also more relevant for tie formation, and less important for persistence (Dahlander and McFarland, 2013), once *cognitive proximity* is built.

Organisational Foci and Institutional constraints

Some kind of proximity (geographical, organisational, cognitive, in terms of goals) is always necessary for collaborations to take place. Institutional structures bring together people who would otherwise not be in proximity, and the proximity arising from institutional structures

helps in creating social bonds. This in turn can increase the dimensions in which (a pair of) researchers are proximate.

There are several factors that determine whether researchers collaborate with peers in their own institutions. Frenken et al. (2009) provide a detailed review of the factors behind the spatial distribution and biases of research outputs. *Institutional proximity* allows for face-to-face interactions, common values, and shared scientific standards (Evans et al., 2011). These institutional constraints make it easier for tacit knowledge to be transferred, and informal contacts allow for increasing commitment to collaboration.

Inter-institutional ties are partly driven by scientists seeking and using expertise from several different sources. However, we would expect that ties spanning different institutions/organisations are hard(er) to form and sustain. Organisational foci and institutional constraints are positively associated with both tie formation and persistence, but we might expect the degree of importance to be different. Being part of the same institution is likely to make formation simpler, but, given that a collaboration exists (and is successful) institutional proximity may not matter. If a cross-institution collaboration somehow forms, and is successful, then there is incentive to continue, regardless of the costs of crossing institutional borders.

Communities and social structure

Researchers who belong to the same social community (Wellman et al., 1997) are more likely to collaborate. Ties that are built on shared social structure and high *social proximity* are more likely to persist, because of the nature of the relationship, the exchange of trust and the frequent contact. We aim to understand whether social interaction changes the future behaviour and/or preferences of researchers (i.e. even beyond dyadic/triadic relationships).

A stream of research from social sciences and economics uses network theory and network communities as a proxy for unobserved similarity of actors' intrinsic characteristics (Evans et. al., 2011). In this line of research, fit is mostly given when actors are similar, yielding homophily interaction, and less often when they are complementary. Whether homophily or complementarity are the 'true' drivers of interaction depends on the level of detail. For example, a plumber and an engineer complement each other - one plans, the other implements. Yet they are similar as they both work in the construction industry. Hence, they are more likely to work together than say an engineer and a painting artist. Most of this research however, considers only individuals, e.g. the prior partners, a common friend, or the number of agents within the shortest path of two agents. For example, Fafchamps et.al. (2009) found that new collaborations emerge faster between two researchers if they are "closer" in their co-authorship network. They estimate that being in a network distance of 2 instead of 3 can raise the probability of tie formation by 27 percent.

When internal characteristics are exogenous, then unequal distribution as well as correlation of characteristics within individuals is going to create clustered interaction. In a process where internal characteristics are further endogenously becoming more similar through interaction, clustering of interaction is going to be reinforced. Whether intrinsic characterstics are endogenous or exogenous to interaction does however not change the role communities take in this research where communities simply reflect aggregate interaction which is completely organised at an individual level (there is a one-way direction here - individuals create ties which

results in patterns (here communities). To the extent that these patterns are not explained with observed internal characteristics (or individual level social environment such as referrals), there existence is assumed to result from unobserved internal characteristics.

Our approach to social distance and structure is somehow different, as we aim not to consider the co-authorship network only as a system of pipes channeling information (e.g. over the shortest path). Instead, we aim to use the network of collaboration as a proxy of the formation of social groups by individuals, which affects interests, preferences, tastes and information on a social level beyond the individual and thereby affects subsequent groupings (again through tie formation among individuals). Under this approach, the community is a self-standing social entity which gives identity and a social context to the people within and outside of the community and thereby directs (inter-)action. Thus, we aim to investigate the micro-macro bridge, or how actions at individual level create social entities and how that feeds back on individual action.

3. DATA AND METHODS

3.1 The data

We use data on rated researchers from the National Research Foundation (NRF) of South Africa. The NRF is a state agency that has as its mission the production of research and the development of national research capacity. One of its key roles is to facilitate the "ranking" of researchers and universities and other public research institutions (Barnard et al., 2011). The system works *de-facto* as a proxy to assess the national and international standing of researchers, and the NRF aims to use the rating process as a way to increase the country's research capacity. Many South African universities use the outcomes of the rating process as a way to categorise themselves as being research-intensive (NRF, 2011). Moreover, many of the national programmes funding research require the applicants to be NRF-rated. The rating process is focused only on research, centrally administered and valid only for a set period (five years in most cases). NRF rankings are useful to understand the quality of research in the academic community in South Africa because the NRF filings require extensive detail and suffer little missing information. Moreover, nearly all top-researchers in the South African academic research community are NRF-ranked (estimates of coverage suggest that about 90% of all South African peer-reviewed research outputs were written by researchers who are NRF-rated) and the rigour of the review process further suggests that the rankings are reliable.

We analyse individual NRF-rated scholars on the basis of their peer-reviewed publications and the networks of collaborators evident from those publications, as well as their NRF ratings and scientific domains at moment of application. We construct our working sample by restricting the data set to researchers who have received a NRF rating in the period 2000-2011 in all scientific domains, which comprises a total of 3,532 researchers.

When necessary, we fill in missing information of co-authors by crossing the NRF data with a list of authors affiliated with South African institutions sourced from Web of Science. We use the employment history of each researcher to link their institutional affiliation at moment of publication and test for organisational foci/institutional constraints.

Variables

Tie formation. We developed a yearly dataset of each tie created between two NRF rated researchers as reported to the NRF. The variable was coded as a dummy taking the value of 1 if a tie between individuals *i* and *j* was formed in year *t*.

Tie persistence. The second analysis focuses on the persistence of the first tie initiated between two co-authors. We coded a dummy variable to take the value of 1 if the tie between two researchers was repeated in year *t* through at least one joint publication. A tie that is never repeated is treated as right-censored.

Homophily of status and cognitive proximity. We tested cognitive proximity by using several variables that proxy for similar social characteristics of co-authors as well as by similar academic interests and values.

Gender, age and ethnicity. Evidence has shown that researchers with the same gender are more likely to collaborate, and to make their collaborations persist (McPherson et al., 2001). In our models we use a dummy variable called *different gender* that captures whether a pair of co-authors have a different gender. Age difference is another homophily variable that causes ties to form and persist. Our variable *different age* measures whether co-authors with a similar age, measured by a variable taking into account 5-year age groups, are more like to form ties and to make them persist. Finally, we measured *different ethnicity* with a dummy that indicates whether a pair of co-authors has the same or different ethnicities, taking a value of 1 if co-authors have a different ethnicity and 0 otherwise.

Scientific discipline. Research has shown that ties and co-authorships across scientific disciplines are rare and that interdisciplinary collaborations are limited. We use scientific discipline as rough measure of cognitive proximity. We developed a dummy variable that measures whether two co-authors are from the same broad scientific discipline in year *t*.[1]

A first look to the data (Table 2) suggests that the most homophilous co-authorships are within the *hard sciences* including the Physical, Chemical and Biological sciences. In contrast, heterophilous co-authorships happen within disciplines that have rather complementary specialties, such as pairs between the medical and health sciences, the agricultural and biological sciences, the arts and humanities and the social sciences, the economic and social sciences, the pharmaceutical and the chemical sciences, and the technologies and applied sciences and the chemical sciences. Importantly, the shares of the most common types of co-authorships increase in most cases for tie persistence compared to tie formation, suggesting that collaboration within these disciplines becomes stronger when the collaboration is repeated.

Status. Collaboration is more likely to occur between researchers who are at similar stages of their research careers. But it is also the case that collaboration across researchers with different tenure status might happen when more junior researchers join forces with seniors through symbiotic relations in which a junior brings new methodological skills and a senior offers access to resources (Dahlander and McFarland, 2013). We measure achieved status by the NRF rating received by the researcher. We use a measure of differences in tenure status by

[1] In the data there are 18 broad academic disciplines, covering all academic research fields. See Table 2.

coding a dummy variable as 1 when co-authors hold a different NRF rating at moment of publication.

Organisational foci and institutional constraints. It is often suggested that ties across different institutions are rare mainly because of organisational differences, ways of working (e.g. work philosophy, work values, etc.) and due to lack of physical proximity. As in the case of the broad scientific discipline, we developed a dummy variable that measured whether two co-authors are from the same research institute or university. The university system in South Africa underwent a major reform around 2004, and in those reforms several universities merged or changed names. We use the most recent name of the university to account for comparability before and after 2004.

Table 2. Tie Formation and Persistence by scientific field of co-authorships

		Author j			
		Tie Formation		Tie Persistence	
		Main field of co-authorship	Share of total	Main field of co-authorship	Share of total
Author i	Agricultural sciences	Biological sciences	31,2%	Biological sciences	35,8%
	Arts	Social Sciences	22,2%	Social Sciences	29,0%
	Biological sciences	Biological sciences	42,7%	Biological sciences	42,1%
	Chemical sciences	Chemical sciences	47,9%	Chemical sciences	53,8%
	Earth and marine sciences	Earth and marine sciences	38,6%	Earth and marine sciences	39,9%
	Economic sciences	Social Sciences	20,5%	Social Sciences	20,8%
	Engineering sciences	Engineering sciences	26,0%	Engineering sciences	30,4%
	Health Sciences	Health Sciences	36,0%	Health Sciences	42,9%
	Humanities	Social Sciences	18,8%	Social Sciences	20,8%
	Information and Computer science	Information and Computer science	22,5%	Information and Computer science	25,5%
	Law	Law	21,1%	Law	23,7%
	Mathematical sciences	Mathematical sciences	38,1%	Mathematical sciences	42,3%
	Medical sciences: Basic	Health Sciences	26,5%	Health Sciences	29,6%
	Medical sciences: Clinical	Health Sciences	38,5%	Health Sciences	54,4%
	Pharmaceutical Sciences	Chemical sciences	20,0%	Chemical sciences	14,0%
	Physical sciences	Physical sciences	52,7%	Physical sciences	62,8%
	Social Sciences	Social Sciences	23,2%	Social Sciences	27,4%
	Technologies and applied sciences	Chemical sciences	16,3%	Chemical sciences	16,7%

Community affiliation. In network theory, communities are defined as sub-networks that are locally dense even though the network as whole is sparse. Members are assigned to a community on the basis of having more links within it than outside it. Communities typically correspond to functional sub-units, namely sets of vertexes that have a property or function in common. Because most community detection algorithms are not deterministic, and different algorithms use different criteria for community assignment, we test different algorithms (fast greedy modularity optimisation, walktrap community, multi-level optimisation of modularity,

and spinglass community) and incorporate the differences in membership affiliations between a pair of co-authors of each community detection algorithm as independent variables in our econometric models. By using these methods, we aim to identify different types of groups of researchers: those interacting in close networks, arguably fostering trust and the transfer of tacit knowledge; and those interacting in open networks, with structural holes and facilitating knowledge creation (Lambiotte and Panzarasa, 2009).

Table 3 presents some statistics of the communities identified through the different detection algorithms. Two algorithms identified a similar number of communities because of the relative closeness of their methods and functions: the *Greedy modularity optimisation method* (FastGreedy) and the *multi-level optimisation of modularity* (Multilevel).

Table 3. Descriptive statistics of scientific communities detected through different community detection algorithms

	FastGreedy	Walktrap	Multilevel	Spinglass
Total number of communities	58	571	60	10
Number of communities that account for a cumulative percentage of at least 50%	18	88	37	6
	Number of researchers in top communities			
Top 1	257	398	365	442
Top 2	206	262	215	425
Top 3	165	52	160	405
Top 4	159	49	138	384
Top 5	152	47	137	367

Community detection through the Walktrap algorithm identified a very large number of communities, with many of them including only one researcher, and with the top five communities including about 23% of all researchers. Finally, we chose a total of ten steps to identify communities using the spinglass algorithm.

Cumulative advantage. Finally, as a complement to the identification of communities of researchers we include variables of cumulative advantage into our models. We expect individuals to have advantage as a result of collaboration centrality in the network of co-authors. To measure *collaboration centrality*, we calculate the respective eigenvector centrality in the co-authorship network. In addition, we include a variable of the intensity of collaboration, or *propensity to collaborate*, which accounts for the number of times a given tie was repeated in a given year.

3.2 The methods

We use a longitudinal dataset to analyse the formation and persistence of ties of all NRF rated researchers in the period 2000-2011. To control for left censoring, we also included all those researchers who were at risk of forming ties in 2000. As NRF ratings have an average validity of minimum three years, we included all NRF researchers that received a rating back to 1997. We made sure that all ties in this sample were in reality first formed in the period of analysis (since

we only observe a fraction of the 'publication life' of each researcher). All other ties were excluded from the sample.

We analyse two sets of models, one estimating tie formation; and a second one for tie persistence. We use a probit model to estimate the probability of a first collaboration, and for those ties that were formed; we use a Cox proportional hazard regression model in order to estimate time to consequent collaborations in the period of analysis. Under our approach, a possible pair enters our models when both co-authors received a NRF rating in the period of study, and they are considered at "risk" to collaborate in any of the years of the period of study.

For the tie formation model we include time dummies for each year following the initial exposure, with t1 = 2000, and where the event is considered to happen the event dummy takes a value of 1 for the year of first publication. For tie persistence, the event takes a value of 1 in the year where the collaboration is repeated.

The Cox hazard tie persistence models use clustered standard errors in each of the co-authors.

4. RESULTS

Table 4 shows the results for the probit tie formation models. The samples of the models are different depending on the members of the top communities identified with each algorithm. All variables tested are the same with the exception of the community detection method used for identifying social communities of researchers (social proximity). Four different algorithms are tested: Fast-greedy modularity optimisation (Model 1), Walktrap community (Model 2), Multilevel optimisation of modularity (Model 3), and Spinglass community (Model 4). Note that the coefficients remain broadly similar in signs and magnitudes across all models. However, the significance changes for some of them depending on the community detection method used. Specifically, this is the case of the homophily variables *different rating* (non-significant in Model 1), *different race* (non-significant in Models 1 and 4) and *different gender* (non-significant in Model 2).

Table 5 presents the results for the tie persistence analyses. We used a proportional hazard model in which time hazard dummies of years after first collaboration estimate tie persistence (the dummy takes the value of 1 in each year where the same pair of NRF authors repeated their collaboration through a NRF rated publication). Because ties in this analysis are conditional upon a tie having been formed, only those ties that actually formed in the first place are included in the hazard model. We mimicked the approach as for the tie formation model, testing different community detection algorithms from Models 5 to 8. All models use clustered standard errors in each of the co-authors.

Table 4. Tie Formation Results

Variable	Model 1 Fastgreedy	Model 2 Walktrap	Model 3 Multilevel	Model 4 Spinglass
Hypothesis 1: Status homophily and cognitive proximity				
Different NRF rating	0.02240	0.061754***	0.05377***	0.08669***
	(0.01957)	(0.015717)	(0.01586)	(0.01930)
Different scientific discipline	-0.22017***	-0.229069***	-0.27461***	-0.30454***
	(0.01991)	(0.015858)	(0.01644)	(0.01981)
Different gender	-0.05288**	0.007099	-0.08502***	-0.06301**
	(0.01921)	(0.015452)	(0.01554)	(0.01919)
Different ethnicity	-0.01258	0.063218***	-0.06271***	0.01280
	(0.02148)	(0.018221)	(0.01776)	(0.02147)
Different age	0.43113***	0.419963***	0.45135***	0.41224***
	(0.02383)	(0.020205)	(0.01997)	(0.02372)
Hypothesis 2: Organisational Foci and Institutional constraints				
Different institutional affiliation	-0.07735***	-0.231765***	-0.08713***	-0.13925***
	(0.02217)	(0.018445)	(0.01807)	(0.02195)
Hypothesis 3: Social Structure and communities				
Cumulative Advantage				
i's collaboration centrality	1.85297***	1.279568***	0.82946***	1.54033***
	(0.07871)	(0.053131)	(0.04951)	(0.07221)
j's collaboration centrality	0.44425***	0.329046***	0.71882***	0.78531***
	(0.06659)	(0.049346)	(0.05255)	(0.06638)
Propensity to Collaborate	10.19326	9.875522	10.31642	10.01227
	(38.03265)	(31.800847)	(31.14045)	(38.10209)
Different social communities				
Fast-greedy modularity optimisation	-1.46533***			
	(0.02006)			
Walktrap community		-1.433011***		
		(0.017620)		
Multilevel optimisation of modularity			-1.65450***	
			(0.01727)	
Spinglass community				-1.35784***
				(0.01925)
Control variables				
One co-author is Male	0.48358**	0.521256***	0.65373***	0.49124***
	(0.02720)	(0.023845)	(0.02317)	(0.02676)
Year dummies	Yes	Yes	Yes	Yes
Number of observations	28159	44950	45742	27942

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Standard errors are in parenthesis.

Below, we use the results presented in Tables 4 and 5 as well as some descriptive statistics on the dataset to interpret our findings.

Status homophily and cognitive proximity

We hypothesised that status homophily might become irrelevant once the tie is formed; while complementarity through value homophily grows stronger once the tie is established. Our measure of status homophily is the ranking given by the NRF. Senior researchers may obtain ratings of A, B, or C, while junior researchers may obtain P, Y, or L where the order is always from highest to lowest rating. We incorporate three variables of cognitive proximity — gender, ethnicity and scientific field.

In contrast with our reasoning, we found that status homophily is not a significant predictor of tie formation. In fact, our results suggest the importance of heterophily, hinting that researchers look for complementary skills and knowledge when collaborating, which is in line with the increasing dominance of teams in publications patterns (Wuchty et al., 2007), giving rise to *specialisation* and gains from the collaboration. Status homophily is not significant for tie persistence in any model.

The above argument is backed by the descriptive statistics from our dataset. When looking at the characteristics of co-authors in our sample of formed ties we note that the majority of dyads are formed between researchers that have different NRF ratings. All authors tend to co-author with others of different rating with the exception of C-rated researchers and P-rated researchers that form dyads mostly with other C-rated (41%) and with B-rated researchers (30%) respectively. The results of the Cox-regression analysis may however simply mimic the structural characteristics of the community of NRF rated researchers. In fact, of all NRF researchers that received a rating in the period 2000-11, 47% of them received a C, followed by B-rated researchers (18% of total), and those that were not successful in obtaining a rating (13%). Interestingly is that P-rated researchers and L-rated researchers represent only 1% of total researchers in the period respectively.

With respect to *cognitive proximity*, we argued that researchers who share interests and similar social characteristics have more in common and are more likely to form ties. But if this hypothesis is true, then links that exist are likely to be homophilous, and thus that variable should have little influence on tie persistence (i.e. as these variables cannot change with time and thus cannot grow stronger). We incorporate as variables of cognitive proximity whether a dyad of researchers has a different ethnicity, a different gender and if they do research in a different scientific field. Our regression results support (partially) our hypothesis of tie formation. Having a different scientific field affects negatively tie formation. However, these results suggesting disciplinary homophily need to be taken with caution, as our database defines scientific disciplines very broadly, with only 18 academic fields in the entire panoply of research.[2] However, even if this is the case, as our descriptive statistics suggest, only an average of 30% of ties formed are homophilous, in comparison to only 34% of ties persisted.

[2] In further work we intend to incorporate finer details on academic field.

Table 5. Tie Persistence Results

Variable	Model 5 Fastgreedy	Model 6 Walktrap	Model 7 Multilevel	Model 8 Spinglass
Hypothesis 1: Status homophily and cognitive proximity				
Different NRF rating	-6.953e-02	-4.655e-02	-3.461e-02	-1.115e-02
	(4.576e-02)	(4.138e-02)	(3.879e-02)	(4.572e-02)
Different scientific discipline	2.608e-02	5.339e-02	3.035e-02	-6.210e-03
	(4.151e-02)	(3.797e-02)	(3.570e-02)	(4.152e-02)
Different gender	-2.931e-02	-3.382e-02	-4.894e-02	-1.049e-04
	(4.258e-02)	(3.893e-02)	(3.679e-02)	(4.263e-02)
Different ethnicity	-1.018e-01	-1.011e-01	-1.032e-01	-1.328e-01
	(4.997e-02)	(4.676e-02)	(4.296e-02)	(4.987e-02)
Different age	6.585e-02	6.952e-02	1.053e-01	6.866e-02
	(5.917e-02)	(5.309e-02)	(5.045e-02)	(5.913e-02)
Hypothesis 2: Organisational Foci and Institutional constraints				
Different institutional affiliation	-4.276e-02	-1.301e-01*	-8.040e-02	-6.820e-02
	(4.188e-02)	(3.795e-02)	(3.559e-02)	(4.184e-02)
Hypothesis 3: Social Structure and communities				
Cumulative Advantage				
i's collaboration centrality	-1.019e+00**	-7.366e-01*	-8.417e-01**	-1.103e+00**
	(2.151e-01)	(1.765e-01)	(1.741e-01)	(2.156e-01)
j's collaboration centrality	-2.996e+00***	-2.338e+00***	-2.251e+00***	-2.815e+00***
	(2.711e-01)	(2.153e-01)	(2.085e-01)	(2.694e-01)
Propensity to Collaborate	1.206e-01***	1.259e-01***	1.258e-01***	1.244e-01***
	(3.902e-03)	(3.324e-03)	(3.233e-03)	(3.840e-03)
Different social communities				
Fast-greedy modularity optimisation	-1.177e+00***			
	(5.942e-02)			
Walktrap community		-1.243e+00***		
		(4.535e-02)		
Multilevel optimisation of modularity			-1.574e+00***	
			(5.093e-02)	
Spinglass community				-1.221e+00***
				(6.109e-02)
Control variables				
Year dummies	Yes	Yes	Yes	Yes
Number of observations	5645	7960	8601	5888

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Standard errors are in parenthesis. Time-hazard dummies are suppressed

Having a different ethnicity, a different age, and a different gender are significant for tie formation but are not for tie persistence. In fact, our hypothesis for tie persistence is confirmed, as none of the cognitive proximity variables are significant. These results suggest that once a collaboration dyad is formed, and authors find a cognitive 'fit', nothing else is needed to explain why co-authors make their tie persist.

Organisational foci and institutional constraints

We include in our model a variable of whether co-authors are researchers in the same institution or not. About 50% of all NRF rated researchers in the period 2000-11 are concentrated in only five academic institutions: the University of Cape Town (12%), the University of Pretoria (11%), the University of Stellenbosch (10%), the University of Witwatersrand (9%), and the University of Kwa-Zulu Natal (8%). Table 6 presents descriptive statistics on the institutional affiliations of all dyads of co-authors included in the tie formation and tie persistence models. The table shows that about 30% of all dyads initially formed happen between authors within institutions of each of the main institutional affiliations of NRF researchers. This share increases to between 28% and 41% in the case of persisting collaborations.

Table 6. Tie Formation and Persistence among co-authors of main institutional affiliation of NRF rated researchers

		Author j				
		Tie Formation				
		University Of Cape Town	**University Of Kwa Zulu Natal**	**University Of Pretoria**	**University Of Stellenbosch**	**University Of Witwatersrand**
Author i	**University Of Cape Town**	31%	2%	2%	7%	5%
	University Of Kwa Zulu Natal	3%	30%	4%	6%	7%
	University Of Pretoria	5%	1%	31%	5%	2%
	University Of Stellenbosch	9%	4%	5%	27%	2%
	University Of Witwatersrand	8%	3%	2%	5%	30%
		Tie Persistence				
	University Of Cape Town	37%	2%	2%	5%	5%
	University Of Kwa Zulu Natal	3%	30%	3%	5%	5%
	University Of Pretoria	3%	1%	36%	5%	2%
	University Of Stellenbosch	6%	3%	5%	28%	2%
	University Of Witwatersrand	7%	3%	2%	3%	41%

From the regressions, our hypothesis is confirmed for tie formation, as being affiliated to different institutions is significant and affects negatively the formation of ties, suggesting that physical proximity is key for making researchers get to know each other and that organisational foci exposes individuals to one another. In contrast, different institutions is not significant for making ties persist, suggesting that once two individuals know each other, physical proximity is not a requirement for making their collaboration persist over time. Thus, NRF researchers seek at a first instance mostly collaborators within their organisational boundaries. However, for making these ties persist they do not follow the rule "*out of sight, out of mind*", contradicting the results of recent research (Dahlander and McFarland, 2013, Evans et al., 2011, Reagans, 2011).

Social structure and scientific communities

For testing the idea that researchers that belong to the same social community are more likely to collaborate together, we apply recent community-detection methods to partition the network into communities and uncover the effects of these on scientific collaborations. In addition, we also introduce the eigenvector centrality explicitly into the model, as well as the variable *Propensity to Collaborate*, or the intensity of the collaboration, both accounting as variables of cumulative advantage.

All variables including the differences of community membership between a pair of co-authors are estimated to be highly significant both for tie formation and tie persistence, proving our hypothesis that social structure, beyond formal institutional and organisational structures, matter for tie formation and tie persistence.

Finally, we look at each individual's collaboration centrality in the publication network, as well as the intensity of collaboration of a pair of co-authors in order to understand the effects of cumulative advantage on scientific collaboration. The coefficients of centrality are significant and positive for tie formation across all models, suggesting that the importance of the authors in the network structure determines their formation of ties and that a researcher that is well connected is most likely to collaborate with other well-connected researchers for the first time. In contrast, and surprisingly, centrality becomes significant and negative for all tie persistence models, suggesting that as a given author becomes more central in the network, it is more likely to have more peers or options of collaboration, which hinders to some extent the persistence of their formed ties. The intensity of collaboration variable is not significant in all tie formation models, and is positive and significant in all tie persistence models. This suggests that cumulative advantage, or the intensity of the collaboration is of high importance for making ties persist.

Robustness Checks

We plan to conduct several robustness checks in the following weeks in order to strengthen our inferences.

Our main concern is to increase the robustness of the social structure and community effects results. One of the main challenges of the existing community detection methods in networks is their stochastic nature. Moreover, for all the algorithms currently tested each vertex is assigned to a single cluster, while in (real) social networks vertices, or researchers, are likely to be shared between two or more communities.

As a way to account for this challenge, we are currently testing the OSLOM (Order Statistics Local Optimisation Method) community detection algorithm, which is a method based on the local optimisation of a fitness function expressing the statistical significance of clusters with respect to random fluctuations, estimated with tools of Extreme and Order statistics. Thus, OSLOM results are somehow controlled for stochasticity, as it is able to compute the statistical significance of social communities using a tolerance p-value that is pre-defined as a fitness measure. Starting from one agent it adds and subtracts agents until the partition is rejected at some level p. The higher p, the smaller the communities are, because then communities are more easily rejected (likeliness to reject the null model - a random network).

Furthermore, Lancichinetti et.al. (2011), the research team behind the OSLOM algorithm, remind us that 'in principle' one may never be able to sort out the influence of unobserved (potentially non-stable) characteristics of the agents from the effect of social interaction (per se) within the community. First, we would need to control for all internal characteristics (non time variant characteristics could be controlled using fixed effects, and in fact they are already controlled for to some extent in our tie persistence models); but also time-varying characteristics; and most importantly lower level network measures such as prior ties, common partners, network distance, etc.

Thus, our aim is to sort out these both factors, the unobserved characteristics of agents effect vs. the social interaction effect, as good as possible. For doing so, we are proposing testing two additional models: an *uncertainty model* and a *community characteristics model*.

The *uncertainty model* would introduce a measure of how uncertain we (i.e. the statistician) are about whether two agents are in a same community. There are two possible approaches for doing so. A first test we are currently conducting is running the OSLOM algorithm at increasing p-values (from 0.01, 0.05, 0.1, 0.2 etc.) for every year in order to identify at what *p*-level two agents are considered to belong to the same community. A second approach would be identifying the probability of membership to the same community. This would follow McNerney et.al. (2013) approach by producing a co-authorship matrix (co-classification matrix in McNerney's paper) equal to the frequency with which an author *i* is grouped with author *j*. In practice, this would require running the OSLOM community detection algorithm *x*-number of times (e.g. 100 times) and extracting the same number of community partitions. If certain authors, or group of authors are frequently grouped together, they would appear with high frequencies in the co-authorship matrix. The contrary would happen if the community memberships were highly variable.

Finally, the *community characteristics model* would introduce measures that characterise the community of which nodes *ij* are members with certainty or without measurement error, including a measurement of the effect of community size for pairs in the same community, and a measure of the effect of the significance for pairs in the same community (e.g. using OSLOM's computed significance values at a given *p*). This would give an approximation of the encapsulation or non-openness of the community, because it is related to how unlikely a given community would be in a random network with degree sequence fixed. Our hypothesis is that in practice, communities with many within-links and few external-links will be highly significant, indicating some sort of inside/outside orientation of the members.

5. CONCLUSIONS

A large body of research argues that there are positive outcomes to research collaborations. To contribute to this body of research from the perspective of Middle Income Countries (MICs), we tested for the determinants of tie formation and persistence using a unique dataset on research collaborations from the National Research Foundation (NRF) of South Africa.

We found that tie formation is favoured by researchers being well-connected in the co-authorship network; for most disciplines, the plurality of collaborations are within discipline, though in the raw, descriptive tables, surprisingly many are made across disciplines. Cognitive proximity is an important factor for tie formation, but not for tie persistence, whereas difference is status is positive for tie formation, but is not significant for tie persistence. Being part of the same social community is however significant for both tie formation and tie persistence.

Interestingly, the crude interpretation of tie persistence is that nothing matters apart from cumulative advantage and social structure, not even organisational proximity. What this suggests is that once a good fit between partners is found (in terms of the other explanatory variables) a tie will naturally persist, since most of the explanatory variables do not change with time, so once a "good fit" is found, it is always a "good fit".

6. REFERENCES

BALDASSARI, D. & DIANI, M. 2007. The integrative power of civic networks. *American Journal of Sociology,* 113**,** 735-780 p.p.

BRAUNERHJELM, P. & FELDMAN, M. 2006. *Cluster Genesis: Technology-based Industrial Development,* Oxford, Oxford University Press.

CHUNG, S., SINGH, H. & LEE, K. 2000. Complementarity, status similarity and social capital as drivers of alliance formation. *Strategic Management Journal,* 21**,** 1-22.

COLEMAN, J. C. 1974. *Power and the Structure of Society,* New York, Norton.

DAHLANDER, L. & MCFARLAND, D. A. 2013. Ties That Last: Tie Formation and Persistence in Research Collaborations over Time. *Administrative Science Quarterly,* 58**,** 69-110.

EVANS, T. S., LAMBIOTTE, R. & PANZARASA, P. 2011. Community Structure and Patterns of Scientific Collaboration in Business Management. *Scientometrics*.

FAFCHAMPS, M., GOYAL, S. & LEIJ, M. J. V. D. 2009. Marching and network effects. *Instituto Valenciano de Investigaciones Económicas, S.A.* Universidad de Alicante: WP-AD 2009-15.

FRENKEN, K., HARDEMAN, S. & HOEKMAN, J. 2009. Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics,* 3**,** 222-232.

LANCICHINETTI, A., RADICCHI, F., RAMASCO, J. J. & FORTUNATO, S. 2011. Finding statistically significant communities in networks. *PLoS ONE,* 6.

LAMBIOTTE, R. & PANZARASA, P. 2009. Communities, knowledge creation and information diffusion. *Journal of Informetrics,* 3.

LEVI-MARTIN, J. & YEUNG, K. T. 2006. Persistence of close personal ties over a 12-year period. *Social Networks,* 28**,** 331-362 p.p.

MCNERNEY, J., FATH, B. D. & SILVERBERG, G. 2013. Network structure of inter-industry flows. *Physica A,* 392**,** 6427-6441.

MCPHERSON, M., SMITH-LOVIN, L. & COOK, J. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology,* 27**,** 415-444 p.p. .

NEWMAN, M. E. J. 2004. Fast algorithm for detecting community structure in networks. *Physical Review,* 69.

PODOLNY, J. M. 1994. Market uncertainty and the social character of economic exchange. *Administrative Science Quarterly,* 39**,** 458-483.

REAGANS, R. 2011. Close encounters: Analyzing how social similarity and propinquity contribute to strong network connections. *Organization Science,* 22**,** 835-849.

RUEF, M., ALDRICH, H. E. & CARTER, N. M. 2003. The structure of founding teams: Homophily, strong ties, and isolation among U.S. entrepreneurs. *American Sociological Review,* 68**,** 195-222 p.p. .

SEABRIGHT, M., LEVINTHAL, D. & FICHMAN, M. 1992. Role of individual attachments in the dissolution of inter-organisational relationships. *Academy of Management Journal,* 35**,** 122-160 p.p. .

SNIJDERS, T. A. B. 2001. The statistical evaluation of social network dynamics. *Sociological Methodology,* 31**,** 361-395 p.p. .

WELLMAN, B., WONG, R. Y., TINDALL, D. & NAZER, N. 1997. A decade of network change: turnover, persistence and stability in personal communities. *Social Networks,* 19**,** 27-50 p.p. .

WUCHTY, S., JONES, B. F. & UZZI, B. 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science,* 316**,** 1036-1039.