Maastricht University

UNITED NATIONS UNIVERSITY

UNU-MERIT

Working Paper Series

#2019-024

Far from random? The role of homophily in student supervision

Giulia Rossello and Robin Cowan

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: info@merit.unu.edu | website: http://www.merit.unu.edu

Boschstraat 24, 6211 AX Maastricht, The Netherlands
Tel: (31) (43) 388 44 00

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology UNU-MERIT

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT to stimulate discussion on the issues raised.

Far from random? The role of homophily in student supervision *

Giulia Rossello[1] and Robin Cowan [1,2,3]

rossello@merit.unu.edu cowan@merit.unu.edu
[1]UNU-MERIT, Maastricht University
[2]BETA, Université de Strasbourg, France
[3]Institut Universitaire de France

June 28, 2019

Abstract

The paper studies racial and gender homophily in student supervision relationships in a context of social transformations, South Africa academia. We develop a technique to separate choice homophily from that induced by the system. Comprising two permutation tests repeated at two levels of aggregation, system and departments. We find clear evidence of homophily in student supervision, along racial lines in particular. Roughly half of the observed homophily is induced by the departments composition and stays constant over time. Overall, choice homophily has similar magnitude along racial and gender dimensions. Further, we ask where choice homophily originates in the demographic groups of students and professors. We find that white (male) students have high tendency to form same-type relations, while among professors it is black (female) who display the higher frequency. Group differences show that choice homophily is likely to originate from students in the former majority.

JEL codes: A14, D71, D85, I23, I24, J15, J16.

Keywords: Academia, South Africa, Student supervision, Induced homophily, Choice homophily, Segregation, Assortativity mixing, Permutation test, Social Transformations, Social Change, System of Organisations, Institutional constraints, Gender ties, Racial ties, University System emerging countries, Racial and Gender Homophily.

*Financial support was provided through the Institut Universitaire de France. We gratefully acknowledge the comments and suggestions of participants of the 2nd Workshop in Economics of Science and Innovation at BETA as well as those of Prof. Pierre Mohnen, Dr. Daniele Giachini, Prof. Alan Kirman, Dr. Mindel van de Laar, and Tatenda Zinyemba.

1 Introduction

Despite decades of targeted intervention, research, and public awareness, labour market outcomes remain unequal along ethnic and gender lines. Compared to white males, blacks and females are paid less, have higher risk of unemployment, and their labour force is concentrated in specific occupations (Neumark, 2018; Cajner et al., 2017; Lang and Lehmann, 2012; Altonji and Blank, 1999). In addition, the lack of representation in top job-positions in many organisations is striking (Phillips et al., 2009; Greenhaus and Parasuraman, 1993; Kossek et al., 2017).

In the business sector, 66% of the Fortune 500 board seats are still held by white men.[1] In politics, Only 37% of the European parliament (elected in 2014) is composed by women.[2] Of the 751 EU parliamentarians, 3 are black, whereas 22% of the European population is black.[3] In universities, women academics held 40.6% of academic positions in the 28 EU-countries in 2013. But few institutions have female heads (20%).[4] In 2016, in USA, 27% of full-time professors are female and 4% are black[5]. In addition to suggesting a social inequity or even injustice, these unequal outcomes represent a large social and economic loss for society. Hsieh et al. (2013) estimate that the decreasing barriers and inequalities for blacks and women observed in US between 1960 and 2010 explains 24 percent of growth in GDP per capita and 6% of economic growth.

Two, often opposing bodies of literature identify the origin of the observed inequalities. The first offers an individual-level explanation; the second an aggregate-level one. The individual-level explanation identifies individual tastes or preferences, or coping with iperfect information as the sources of observed inequalities. For this body of literature the mechanism behind the formation of these preferences are numerous. The main hypotheses are "taste for discrimination", and "statistical discrimination". The first, proposed by Becker (1957), states that, independently from individual productivity or quality, employers have preferences for employees of certain groups. The

[1] The Catalyst 2018 Board Diversity Census of Women and Minorities on the Fortune 500 Boards. Available at `https://www.catalyst.org/research/missing-pieces-report-the-2018-board-diversity-census-of-women-and-minorities-on-fortune-500-boards/`; last access March 2019

[2] Available at `http://www.europarl.europa.eu/elections2014-results/en/gender-balance.html`; last access March 2019

[3] Available at `https://www.theguardian.com/world/2018/aug/29/eu-is-too-white-brexit-likely-to-make-it-worse`; last access March 2019

[4] Source catalyst `https://www.catalyst.org/research`; last access March 2019

[5] Source NCES `https://nces.ed.gov/fastfacts/display.asp?id=61`; last access March 2019

second explanation, pioneered by Arrow et al. (1973) and Phelps (1972) states that employers are not endowed with any specific preference for different groups but such preferences appear to exist because of asymmetric information. When hiring decisions are taken without perfect information on candidate quality, the stereotypical average of candidate quality in the specific group will prevail, causing a self-reinforcing mechanism that penalises under-represented groups in labour market since the average observed quality has larger variance because of the small sample size, and so the confidence in the estimate of the quality of that group is much weaker.

By contrast, the aggregate-level explanation identifies network structures as the main source of observed inequalities. One mechanism that drives the effect is "referral hiring". Because hiring decisions often rely on information coming from agents' ties, if these ties are homogeneous along socio-economic characteristics (as is commonly observed), subsequent hiring decisions will reproduce the same group structure existing in the organisations (Montgomery, 1991).

Whether the source lies at the individual or aggregate level, the outcome presents as homophily in the workplace. That is, we observe a tendency for agents to hire or work alongside, or, generalizing beyond the labour market, to associate with, other agents like themselves in some relevant dimension. Though much of the research in this area has focussed on labour market outcomes and work environments, the phenomenon is observed in a much wider variety of contexts. Work and labour market outcomes are important for individual well-being, which explains much of the interest in that area. However, another locus is also important for individual well-being, namely education. In particular at the university level, homophily is again present. Our concern in this paper is with one specific aspect of homophily in universities, namely the relationships between students and supervisors.

Smith et al. (2014) describe the observed homophily in a system as a summary measure of social distance across time and demographic dimensions. In particular, they define "observed homophily" as the "behavioural expression of institutional segregation, demographic availability, and effective acceptance among categories of people" (Smith et al. (2014) pag. 433).

Implicit here is the idea that an observation that some group is under or over-represented may have several distinct sources. The literature in general expresses this nuance and distinguishes two origins of the homophily observed in social networks. The first, induced homophily, refers to constraints in the structural opportunities for interaction produced by institutional segregation and demographic availability. The second, choice homophily, arises from the effective acceptance among categories of people which are codified by individual preferences or tastes (Kossinets and Watts, 2009). Induced homophily originates from the local homogeneity of the opportunity for interaction, influencing tie formation — it is impossible to hire a white man if there are none present in the region where the job is offered. Indeed, geography, friendship circles, and workplaces, which can be homogeneous in relevant social characteristics, constrain interaction opportunities. Induced homophily will create an unequal distribution of agents over opportunities even if all agents are entirely indifferent with regard to gender, ethnicity...

Choice homophily, by contrast, arises from individual preferences of forming ties with similar others and involves psychological attitudes that can be conscious or unconscious. Here we are speaking "simply" about individuals' tastes to be surrounded by others who are similar to themselves.

This paper studies homophily[6] along racial and gender lines, using as a specific case student-supervisor ties in a context of social transformations, namely South African Academia between 1973 and 2014.[7]

We focus on academia because university study opens doors to many future careers, but at the same time conditions people about what sorts of careers are relevant for them (Gersick et al., 2000; Wenger, 2010). The kinds of relationships people have in university can be formative in the way they view the world and what they expect as a "natural" relationship. For Astin and Astin (2000) university classrooms enable students not only to acquire knowledge, to develop quantitative and writing skills and critical thinking but also to develop personal qualities like self-understanding, listening skills, empathy, honesty, integrity and the ability to work collaboratively. For many degrees, the final requirement is a thesis, and the thesis project is typically supervised or mentored by a faculty member, who guides the student not only about research in this project, but often about much wider subjects such as career prospects, future education and so on, as well.

The literature focusing on the role of mentorship at universities looks at student-professor

[6]In the text "homophily" refers to observed behaviour in a population. At the individual (group) level homophily can refer to the desire to form same-type of ties. We refer to the latter with the term "homophilous preferences".

[7]Formally four "racial" groups are categorised in South Africa: black, white, Indian and coloured. The word "black" is sometimes used to refer to the aggregate of black, Indian and coloured. We use this meaning throughout this text.

ties and identifies 5 components which characterise an effective mentorship relation: (1) a focus on achievements and knowledge acquisition; (2) support, direct assistance, and role modeling; (3) mutual benefit; (4) a personal nature beyond strictly work-related issues; and (5) a professor has within the mentoring setting greater experience, influence, and achievement (Jacobi, 1991; Gersick et al., 2000; Girves et al., 2005; Crisp and Cruz, 2009). These relations have been found to be particularly important (especially in early-stages) for under-represented and minority groups (Girves et al., 2005; Terrell and Hassell, 1994; Pezzoni et al., 2016) and are able to reduce their attrition rate (Terrell and Hassell, 1994). Thus, the student-supervisor relation, because of its role and relevance for under-represented groups, is a potential channel of social transformations.

The case of South African academia could be perceived as an extreme case. But the lack of diversity in academia is persistent also elsewhere (Gasman et al., 2015), and similar mechanisms may influence social transformations in other settings. In South Africa before 1994, the university system was segregated. White universities were well-funded, and specialised in the knowledge production in all fields but in the natural sciences in particular. Black universities were under-funded and specialised in technical education (Herman, 2017). When apartheid ended, the need to reform the education system was urgent. Initially, policy promoted the enrolment of black and female students in former white universities, and encouraged geographic mobility. But going further, in 2002-2004 a systemic reform merged many university departments and created new institutions to foster diversity in the faculty composition (Herman, 2017). However, 25 years after apartheid ended, transformations are still unevenly distributed.

In such a context, our focus on racial and gender homophily in student-supervisor ties not only shows whether and how social transformations change the structure of groups' interactions, but also looks at an interaction that takes place at one of the important moments in an educational career. Moreover, ties between students and professors represent how the old and new generations interact with one other and may give insights about the hiring process. Academia is a highly competitive environment where the student-supervisor relation is the first work-related tie of future professionals, and supervisors (especially for Masters and Ph.D. theses) can be very influential in determining job placement following study. The high level of stratification and low mobility levels commonly found in university systems (Cowan and Rossello, 2018; Burris, 2004; Cruz-Castro and Sanz-Menéndez, 2010) suggest that a first-job in academia is to a very great extent a consequence of Ph.D. thesis work and the social connections developed during it.

This paper contributes to the literature in three ways. First, our method offers a quantitative approach to study social transformations in a system of organisations considering not only changes in the demographic composition of agents but also in the structure of interactions. In particular, we look at how population groups relate to one another, that is, how frequently they form (or not) ties together in terms of homophily. Second, we develop a way to separate induced from choice homophily. Our method, in contrast with previous work (Kossinets and Watts, 2009), analyses homophily at two levels of aggregation, which is key to separating choice from induced homophily. Further, it controls for population sizes, institutional constraints, and confounding factors embodied in network structures (such as preferential attachment, norms and practices, popularity, and perceived quality).

Third, we present a simple statistical model that permits us to estimate the relative strength of choice homophily in different interacting sub-populations. This addresses the issue of where choice homophily originates in the system of supervision, estimating its strength in populations of students and professors. It will thus help to understand the mechanism behind the persistence and diffusion of homophily.

We find strong evidence of homophily in student supervision with race-based homophily stronger than gender-based homophily. However, once induced homophily is controlled for, choice homophily for race and gender are similar. Overall, roughly half of the homophily in the macro system is induced by institutional constraints at the department level.

Differentiating among the four groups of students and professors we find the highest frequency of homophilous preferences in students of the previous majority group (white/male) and in professors of the previously excluded groups (black/female). Homophilous preferences of these two groups hinder the formation of cross-type ties. In many disciplines, it is students who approach faculty seeking supervision (rather than the reverse), so black (female) scholars may face additional induced homophily that we do not fully capture, driven by the fact that white (male) students with homophilous preferences tend not to approach black (female) professors. So some of the homophily we measure as attributed to black (female) professors may originate with white (male) students. As one nuance of our study, we examine a sub-population of students, namely those we can identify as going on to become academics in the future. Within this group, we find that white professors also display homophilous preferences, suggesting additional constraints faced by black students and their entry to academia.

Our findings indicate that a large part of the homophily observed at the aggregate level is induced by past segregation. This suggests that the reforms of 2004 will address some of the imbalance in the system. At the same time though, we seem to observe hints of an increase in choice homophily. In section 10 we address this directly and show that at least part of the measured increase could be mechanically due to changes in the relative sizes of the four sub-populations. We show that even with absolutely no changes in choice homophily levels, under some circumstance a changing population will exhibit apparent increases in choice homophily. Thus empirical results during a transition phase must be interpreted carefully.

2 Homophily, segregation, social transformations and the role of newcomers

Humans are embedded in social structures (Smith et al., 2014; Granovetter, 1973) and, perhaps for this reason, homophily is one of the most compelling and strong empirical regularities in our cultural (Barnett and Benefield, 2017), social (McPherson et al., 2001; Currarini et al., 2009), and economic (Podolny, 1994; Jackson, 2005) life. Homophily is the tendency for agents to connect to those similar to themselves. The nature of similarity has been studied along many dimensions that scholars divide into two main groups: exogenous, such as gender or ethnicity; and endogenous, such as occupation, social status, values and norms (Kossinets and Watts, 2009; Pin and Rogers, 2016).

Our work looks at homophily along two exogenous characteristics — race and gender — in a specific working environment with a history of segregation and an ongoing process of social transformation — South African academia. First it studies the evolution of homophily, separating choice from induced homophily. Second it asks where choice homophily originates in the population groups.

Related to the first part of our work, the interplay between referral hiring, segregation, path dependence, and stratification makes homophily levels, in a system of organisations, likely to change slowly, presenting persistence and high inertia (Huffman and Cohen, 2004; Rydgren, 2004; Granovetter, 1973).

Imperfect information plagues both sides of a hiring decision, and for this reason referrals are common. Referrals work (partially) over social networks. Recruiters, for hiring decisions, will perceive more reliable advice coming from their social networks, where information asymmetries are lower. And job seekers are likely to receive information about job openings from their social contacts. However, social networks exhibit homophily along ethnic, gender, and socio-economic lines. So, hiring decisions are likely to reproduce a similar group structure of the organisation itself, maintaining inequalities and marginalization along the dimensions that exhibit homophily in social networks (Barnard et al., 2016; Montgomery, 1991; Granovetter, 1973; Sherif, 2015).

Moreover, segregation could increase the extent of homophily in social networks, by adding additional constraints to agents' interaction. Segregation reduces the number of "relevant" connections between the population groups. Schelling (1971) describes three mechanisms producing segregation: a conscious or unconscious discriminatory individual behaviour; an organised action of a group; and socio-economic inequalities that sort people into different social loci. Even though organised actions of discrimination are banned in many societies, as in the case of South Africa after 1994, the labour market, and indeed any locus where individuals are recruited into an "exclusive group" displays segregation nonetheless because of the interplay between past discriminations, individual behaviour and socio-economic forces (Åslund and Skans, 2010; Altonji and Blank, 1999; Neilson and Ying, 2016).

Homophily and segregation are connected and reinforce each other. On the one hand, when agents are homophilous, even (slightly) preferring to connect with similar types, over time they will form more and more ties to similar agents, and their local networks will become more homogeneous. This local homogeneity will thus increase the social distance between types and segregate the overall network (Schelling, 1971; Schelling et al., 1978; Kossinets and Watts, 2009; Kirman, 2011; Barnett et al., 2016). But further, when networks are segregated agents have fewer opportunities to meet with other types. Thus even when agents form ties randomly with others they meet, their chance of forming ties with other types is low. Over time, due to this path-dependence, their neighbours will become more homogeneous so the level of (observed) homophily will increase (Smith et al., 2014).

In an attempt to clarify what lies behind any observed segregation, the literature identifies two sources of homophily: choice and induced. The first, choice homophily, arises from individual psychological, conscious or unconscious, (perhaps statistically induced) preferences. The latter, induced homophily, is dictated by structural opportunity for interaction, and arises when the potential contact pool of agents is (mostly) homogeneous but not representative of the entire population (McPherson and Smith-Lovin, 1987; Feld, 1982; Blau, 1977).

In theory the two mechanisms are clear and separated. Yet, to distinguish the two empirically it is not easy and is a classic chicken-and-egg problem. This has to do with path dependence and stratification. The social space of each individual is endogenous: the series of repeated choices for association determines its composition; further, each choice may bound by (and sometimes determines) the next, creating path dependence (Sharmeen et al., 2014). Choices of other humans, also from one generation to another, influence individual decisions and position in the social space (Kossinets and Watts, 2009). This process of path dependence creates stratification[8] making any observed behaviour (of individuals or groups) a combination of induced and choice homophily. The formal definition of induced and choice homophily attempts to get around this problem by defining choice homophily as a residual. The overall level of homophily observed is the combination of induced and choice homophily. Induced homophily is the probability of forming ties with agents of the same type, given the group composition. In other words it is the expected random mixing (creating inter-agent ties uniformly at random) given group composition. Choice homophily is the homophily level which exceeds induced homophily (McPherson and Smith-Lovin, 1987; Kossinets and Watts, 2009).

The second part of our work looks at choice homophily in the different sub-populations. Choice homophily may work differently for different demographic groups (Pin and Rogers, 2016) as each group evolves in its own way in response to ongoing social transformations. Indeed, social transformations change the demographic composition of a system and the structure of agents'

[8]In sociology, stratification is the process by which the society and organisations differentiate groups of people in strata of different socio-economic status or prestige (Shavit et al., 2007; Saunders, 2006; Smith et al., 2014).

interactions. However, the demographic change can have a limited scope in social transformations depending on how the different population groups interact with each-other (Smith et al., 2014).

The demographic composition changes (mostly) when a new generation of agents enters. Newcomers in the system are usually young people who, due to lack of time and experience, have few connections. Empirical (Mollica et al., 2003) and theoretical work (Bramoullé et al., 2012) shows how the sparser newcomers' networks have higher homophily levels. In addition, in hierarchical working environments, homophily/heterophily could differ for demographic groups. Incumbents of the (former) majority are more likely to be on top of the hierarchy and each newcomer will have incentives to link with them, to enhance access to resources. When this holds it translates into high homophily levels for newcomers of the (former) majority group and heterophily for newcomers of the under-represented group (Mollica et al., 2003; Main, 2014; McPherson et al., 2001; Ibarra, 1992, 1997).

In more general terms, two opposing bodies of literature offer a framework which links demographic changes and homophilistic behaviour of the groups: contact and conflict theory. In such framework demographic changes shape group identity by affecting the probability of cross-group meeting opportunities which is seen as the main source of homophily in a group (Alderfer and Smith, 1982; Zebrowitz et al., 2008; Hewstone and Swart, 2011; Zhou et al., 2018; Barnard et al., 2016; Jost et al., 2004). For contact theory group identity is a learning process, people learn (unlearn) homophily based on the amount and quality of cross-group meeting opportunities. In other words, a lack of exposure (contact) with members of another group is the primary cause of homophilistic decisions based on group identity (Allport et al., 1954; Pettigrew and Tropp, 2006; Hewstone and Swart, 2011; McKeown and Dixon, 2017; Stainback, 2008; Zhou et al., 2018). When the number of relevant cross-group contacts increases group identity and group homophily decreases since people will base their choices on individual characteristics rather than on group identity (Hewstone and Swart, 2011).

By contrast, for conflict theory group identity is shaped by the tension among groups of different sizes and "status". What drives this tension is the fear of the "majority" group of loosing relevant positions or resources (Sherif et al., 1961; Sherif, 1966; Levine and Campbell, 1972; Brief et al., 2005). In conflict theory a demographic change which increases cross-group contacts rises the tension among groups resulting in homophilistic choices based on group identity.[9]

Both theories are likely to capture important aspects of intra- and cross-group behaviour, yet, experiments demonstrate that group relations can be more complex. Prejudices and stereotypes may have consensus across group boundaries and be highly present also among agents belonging to the most harmed groups (Bian et al., 2017; Cheryan et al., 2017; Beasley and Fischer, 2012; Aronson et al., 1999)

There are many implications and hypotheses related to homophily and social transformations in the literature. However, Smith et al. (2014) underline how little we know about how and whether homophily changes over time. The main bottleneck in this research is the availability of network data with a long time span, where the few databases available are US-based. In our work we study decades of the South African University system looking at racial and gender ties between students and professors which represent thesis supervision. The universities and the society were racially segregated until 1994. But since then social transformations have been taking place.

Besides the historical content of our data, our contribution differs from past literature in two ways. First, we employ a systemic perspective, separating the relative magnitude of induced and choice homophily. Second, we look at choice homophily for the different groups of students and professors.

In particular, the first part of the work studies racial and gender homophily over time, separating choice homophily from that induced by the structure of the system. Key to separating the two is using different levels of aggregation, and creating an appropriate null model of type-blind tie formation. In this respect, we look at tie-types both globally and at the department level. We follow a permutation technique to create a null model with permutations able to include network characteristics, population availability, and institutional constraints. To our knowledge, only Kossinets and Watts (2009) study empirically the relative magnitude of induced and choice homophily. They study the network of e-mail messages of a US university for one academic year. Differently, their methodology focuses on estimating the impact of similarity and social proximity on the probability that two agents form a new tie. They find that similarity governs tie formation only when agents

[9] One can argue that contact and conflict theory are not necessary in contrast but two sides of the same coin. Conflict may prevail when the "quality" of the contact among groups is poor. This may happen when cross-group contact does not translate into more cross-group collaborations and when it does is mostly hierarchical.

are not close in the social space. This result suggests that both induced and choice homophily play a role in tie formation. In contrast we look at choice and induced homophily at two levels of aggregation in an attempt to disentangle the two empirically.

In the second part of the work we further investigate choice homophily, using department level data. We develop a simple model to estimate the strength of homophilous preferences in the different sub-populations. This disaggregation among population groups permits us to ask whether different segments of the population have stronger or weaker preferences for interactions with those similar to themselves.

3 Data

Our data originate with the South African National research Foundation[10] As part of its mission the NRF has a "rating system" in which researchers apply, roughly every 4 years, to have their research output evaluated, and are assigned a "rating" (7 ordered categories), by an expert panel, based on international referee reports. In the application, researchers are asked to submit very complete curriculum vitae information, including details on publications, work history, student supervision, plus characteristics such as race and gender. Strong individual and institutional incentives imply that scholars with a research oriented career usually apply to be rated. NRF data cover the 30% of scholars in the country accounting for roughly 90% of all South African peer-reviewed research outputs (Barnard et al., 2012; León et al., 2016; Cowan and Rossello, 2018). Nonetheless, we should point out that our sample represents only a part of the academic system in South Africa, with a very strong focus on academics who are pursuing a career with a strong research component. This is the part of the system that is foremost in knowledge production, and constitutes most of the "prestigious" part of the system. One reason for focussing on this part of the system is that here transformation may be more challenging as there remains a strong focus on preserving "quality" as defined by international norms of scholarship. The historical legacy of apartheid implies of course that this part of the system until recently has been dominated by white researchers, and so one could argue that it is the part where the transformation issue is both most pressing and most challenging.

[10]NRF (`www.nrf.ac.za`) is a state agency that has as its mission the promotion of research and the development of national research capacity.

The database contains 78081 student-supervisor relationships (with 7432 total supervisors).[11] The data include gender and race of students and professors, student level of education and year, university (39 institutions), and broad scientific field (18 categories). In the South African university system supervision occurs at three levels and all three are present in our data. They include thesis supervision of bachelor ("Honours" in South Africa) (19%), master (56%) and Ph.D. (25%) students where Science Engineering and Technology (SET) represents 73% of the total supervisions and Social Sciences and Humanities (SSH) the remaining 27%.[12]

Our analysis is done considering five time periods: 1973-1995 (3%), 1996-2000 (9%), 2001-2005 (22%), 2006-2010 (39%) and 2011-2014 (27%).[13] We study student-supervisor ties, looking at race and gender. Our interest lies in whether ties display homophily. Supervision relations between agents of same-types are: white-white (WW) and black-black (BB) for race, and male-male (MM) and female-female (FF) for gender. Ties between agents of cross-types are: black-white when a black student has a white supervisor (BW) or white-black for the reverse (WB), and similarly for gender: female-male (FM) or male-female (MF).

4 Preliminaries

In what follows we will make repeated use of two things: a null model of tie-formation; and a particular network statistic. So before we start the analysis we detail them here.

Our population is made of students and supervisors : to avoid confusion, we indicate the first with S and the second with T (teachers). When the S and T are used with subscripts, indicating the dichotomous types b and w (f and m for gender), they represent the relative proportions of types in the population of students or professors. So, for example, S_b indicates the proportion of students who are black, and so on for S_w, T_b, and T_w.

[11] In our sample 68% of the students completed their studies.

[12] Given the source of the data, and how it is collated, if there is an incompleteness in the records, it will be at the lower levels. Academics concerned with presenting a strong research profile are more likely to consider that PhD supervision is more relevant than Bachelor supervision. Comparing students' racial and gender composition of our sample with National statistics we can conclude that our sample is representative of postgraduate students in the country. Looking national statistics for Bachelor students black students are slightly more than in our sample. National statistics are available at `https://www.che.ac.za/sites/default/files/publications/BS\%206\%20National\%20Plan\%20for\%20Higher\%20Education\%20\%282001\%29\%20targets\%20-\%20Final_0.pdf`; `https://www.che.ac.za/focus_areas/higher_education_data/2013/participation`; `https://www.idea-phd.net/images/Doctoral-Education-in-South-Africa-WEB-3.pdf`.

[13] The first two periods should be treated carefully, as both have few observations. Further, the first represents the apartheid era, while the second is after the end of apartheid but before the reform of the university system.

4.1 Null Model

We use a null model of random creation of supervision ties in which a student and a supervisor are drawn randomly and independently from their respective populations. It is straightforward to write the probabilities of observing any of the four tie-types:

$$
\begin{aligned}
Pr(ww) &= S_w T_w \\
Pr(wb) &= S_w T_b \\
Pr(bw) &= S_b T_w \\
Pr(bb) &= S_b T_b.
\end{aligned}
\tag{1}
$$

These represent the probability of observing a certain tie-type given population availability and type-blind tie formation (also called random mixing). $Pr(ww)$ is simply equal to the proportion of students who are white multiplied by the proportion of white supervisors ($S_w T_w$).

4.1.1 Null model with permutations

With a large sample, the law of large numbers implies that the central tendency indicated by the simple model provides a relevant benchmark and so deviations of observed from expected values is a good indication of homophily. With a small sample, however, one need to take more care, as the law of large numbers might not apply, particularly in the case where the sub-population sizes are very different (we have many more students than teachers). Consequently, in much of what follows we estimate the null model numerically to generate a distribution of outcomes of the null model with which we can compare our observed supervisions.

We create a "corrected" null model using permutations. The null model with permutations is the result of tie-types given by repeated permutations of the existing ties. So the probabilities of observing any of the four tie-types is given by permutations results

$$
\begin{aligned}
Pr(ww) =& \frac{1}{J}\sum_{j=1}^{J} W_{perm_j}W \\
Pr(wb) =& \frac{1}{J}\sum_{j=1}^{J} W_{perm_j}B \\
Pr(bw) =& \frac{1}{J}\sum_{j=1}^{J} B_{perm_j}W \\
Pr(bb) =& \frac{1}{J}\sum_{j=1}^{J} B_{perm_j}B.
\end{aligned}
\tag{2}
$$

Where j is a realisation of the permutation, and $perm_j$ indicates a random assignment of students to supervisors. When we do the permutation we retain the number of supervisions that each professor performs at the individual level (in network terms we preserve the degree sequence) and assign students randomly.

4.2 Homophily and Assortativity

To measure observed homophily we use the assortativity coefficient (Newman, 2003). It is a standard measure used to characterise this aspect of social networks. In general terms the assortativity coefficient is

$$
r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}. \tag{3}
$$

Where i indicates the types (w and b); e_{ii} are the observed proportions of ties between agents of the same type (WW and BB in our case) and a_i and b_i are the fraction of each type in the population a and b (S and T in our case). In our case:

$$
Ass. = \frac{(BB+WW)-(S_bT_b+S_wT_w)}{1-(S_bT_b+S_wT_w)}. \tag{4}
$$

Where BB and WW are proportions of same-type ties, S_b is the fraction of black students, T_b is the proportion of black professors, and S_w and T_w are those of white students and professors. The assortativity coefficient measures the distance between observed same-types ties and those predicted by the null model (eq.1). Again, because of our small numbers, when we use assortativity in section 8 we modify Newman's coefficient by replacing the expected value with a permuted value as described above. This gives us a frequency distribution for the assortativity measure rather than a point observation.

With those preliminaries in place, we turn now to the analysis.

5 Student-Supervisor ties

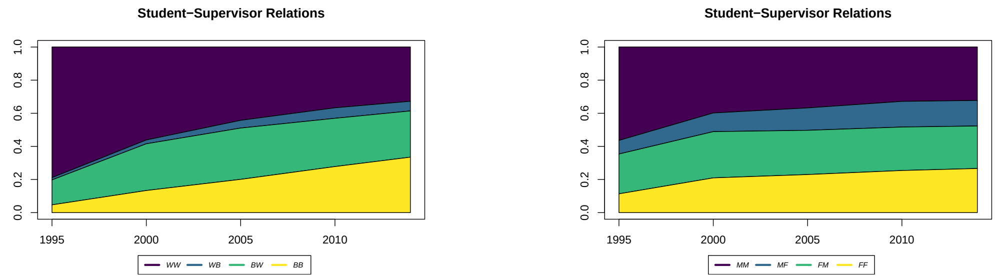
Table 1: Student supervision composition. Where w indicates white, b black, m male, f female, S represents students, and T supervisors. $Ass.$ is an abbreviation for assortativity. Expected values of random mixing (eq. 1) are in parenthesis. Columns 1 and 2 should be treated carefully as both have few observations. Column 1 represents the apartheid era and column 2 is before the reform of the university system.

	1973-1995	1996-2000	2001-2005	2006-2010	2011-2014
			Populations		
S_w	0.80	0.58	0.49	0.43	0.39
S_b	0.20	0.42	0.51	0.57	0.61
T_w	0.90	0.80	0.71	0.61	0.56
T_b	0.10	0.20	0.29	0.39	0.44
			Supervisions		
WW	0.79 (0.72)	0.56 (0.46)	0.44 (0.35)	0.37 (0.26)	0.33 (0.22)
WB	0.01 (0.08)	0.02 (0.12)	0.05 (0.14)	0.06 (0.17)	0.06 (0.17)
BW	0.15 (0.18)	0.28 (0.34)	0.31 (0.36)	0.29 (0.35)	0.28 (0.34)
BB	0.05 (0.02)	0.13 (0.08)	0.20 (0.15)	0.28 (0.22)	0.33 (0.27)
$Ass.b/w$	0.38	0.31	0.29	0.32	0.34
			Populations		
S_m	0.65	0.51	0.50	0.48	0.48
S_f	0.35	0.49	0.50	0.52	0.52
T_m	0.74	0.67	0.62	0.57	0.56
T_f	0.26	0.33	0.38	0.43	0.44
			Supervisions		
MM	0.56 (0.48)	0.40 (0.34)	0.37(0.31)	0.33 (0.27)	0.32 (0.27)
MF	0.08 (0.17)	0.11 (0.17)	0.14 (0.19)	0.15 (0.21)	0.15 (0.21)
FM	0.24 (0.26)	0.28 (0.33)	0.27 (0.31)	0.26 (0.30)	0.26 (0.29)
FF	0.11 (0.09)	0.21 (0.16)	0.23 (0.19)	0.25 (0.22)	0.27 (0.23)
$Ass.m/f$	0.23	0.21	0.2	0.16	0.18

Figure 1 presents a stack plot of tie-type composition over time. The social transformations of the system at the aggregate level, in particular the growth of black participation (and to a lesser extent that of women), in academia and the growth of cross-type ties, is evident. Table 1 presents some descriptive statistics on populations, supervisions, and the network assortativity coefficient.

Figure 1: Stack plot of Student-Supervisor relations 1973-2014. Ties are labels ordered as student, supervisor. Racial ties WW, WB, BW, BB (1a) and Gender ties MM, MF, FM, FF (1b). For 5 time periods: 1973-1995; 1996-2000; 2001-2005; 2006-2010; and 2011-2014

(a) Racial ties **(b)** Gender ties



The table shows the social transformations of the university system, from a system dominated by white males to one showing more diversity. We see clearly an increase in the presence of black (female) students in this part of the university system: from $S_b = 20\%$ ($S_f = 35\%$) in 1973-94 to $S_b = 61\%$ ($S_f = 52\%$) in 2011-14. A similar, though less striking change also occurs in the professoriate: the proportion of black (female) professors (T_b and T_f) increases from 10% (26%) to 44% (44%) over the same time span.

As would be expected with such changes in the composition of students and supervisors, in the second parts of each panel we observe a strong increase in the number of cross-type types. For the period 1973-1995 white professors have 94% of the total supervisions, and supervisors are 80% male. This is not surprising given that this period is from the apartheid era. In 2011-2014, by contrast, white and male professors supervise respectively 61% and 58% of the theses. Overall, 84% of student-supervisors ties are WW or BB in 1973-1995 falling to 66% in 2011-2014. Similarly, along gender lines, MM and FF supervisions fall from 67% in 1973-1995 to 59% in 2011-2014. In parentheses in each cell we have indicated the proportion of supervisions of that type expected under the null model (eq. 1). This permits us to observe that although cross-type ties are growing over time, ties of white (male) students with black (female) supervisors are relatively rare even today.[14] For both race and gender, we see prima facie evidence of homophily because same(cross)-ties are more (less) common than random mixing.

The final row in each of the panels shows Newman's assortativity coefficient. Both race and gender assortativity are positive, indicating a positive tendency (at the aggregate level at least) for

[14] A simple explanation may relate to the availability of black and female professors for thesis supervision. We take this up below.

agents to connect with those similar to them. This provides prima facie evidence of homophily.

Race assortativity does decrease before the period of the university reform (2001-2005) but increases again after it, while, gender assortativity appears more stable. We should notice that even though the representation of formerly excluded groups increases, assortativity rises over time. At first glance this seems to indicate a (distressing) growing preference for similarity.[15] It is possible, though, that this observation is mechanically due to changes in the population composition. We examine this possibility in section 10 below.

5.1 Future Academics

In our data it is possible to identify students who go on to have a (research-active) career in a South African university. They represent a possibly interesting sub-group, as they may be identified by potential supervisors as promising students. We refer to them as future academics.[16] They constitute 13% of the sample of students. We consider the sample of future academics as key to representing the entry process of research-oriented scholars into the system. Table 3 in Appendix A shows summary statistics for this sub-population of students. The population composition and supervisions of the sub-sample are in line with the overall population, yet social transformation appears slightly lower along racial lines.

6 Methodology

Key to isolating choice homophily in a system of organisations is using various levels of aggregation. The university system is comprised of a large group of institutions, the universities, each one further divided into departments. In principle universities and departments are porous and interact with each-other. However, there are various institutional constraints governing these interactions. For students' supervision, in particular, each university department has the duty and the incentive to provide internal supervision for their students. In this case ties form (mostly) at the level of university departments. The department represents the potential meeting pool of the agents in close proximity and likely to form ties together. Indeed, departments represent the institutional con-

[15]Though one that would be consistent with the conflict hypothesis (Levine and Campbell, 1972; Brief et al., 2005; Sherif, 2015; Hewstone and Swart, 2011; Zhou et al., 2018).

[16]We identify future academics exploiting professor information in our database: we match professor data with student data by name, surname, university, field and degree year.

straints which create induced homophily within the system, because they are the organisational "limits" to interaction expressed by norms, the division of labour, and specialisations. Thus if we restrict attention to populations of students and supervisors within a single department, we remove most (though possibly not all) of the induced homophily — any student within a department can take any faculty within that department as a supervisor.[17]

In the first step of our analysis we use a permutation technique repeated at two levels of aggregation: system and department. Our data do not specify university departments reliably, so we identify university departments by the pair: university name and broad scientific field. Scientific fields, of which we have 18 in our data, are broad, so our proxy for departments is not perfect, and thus our measure of choice homophily will in fact include some induced homophily. Using the permutation results, which gives a null model of randomly generated ties, we compute homophily levels in the two cases. Our measure of homophily is a modified version of the assortativity coefficient (presented in section 4) and looks at the gap between the null model with permutations and observed ties (eq. 5). The measure of homophily done at the aggregated, system level comprises both choice and induced. The second estimate, done at the department level removes (much of) the homophily induced by the system and can be taken to represents choice homophily.[18]

In section 9 below, we refine the analysis to study where choice homophily originates, we look at the strength of homophilous preference in the population of students and professors. We develop a simple model of tie-formation where-in each sub-population has a tendency to form ties with similar others.

We use our model to estimate the strength of homophilous preference in the four demographic groups: black and white students and black and white professors (male and female for gender). To do this we use department level data. Further, in section 10, we discuss, using predictions of our model, how the measure of assortativity can be prone to misleading interpretations in the presence of social transformations.

[17] We are over-stating the case here to some extent. Physics departments, for example, might include both theoretical and experimental physicists. A student doing a theoretical thesis would not be supervised by an experimenter. So there will remain some induced homophily for which we cannot correct. This is particularly true given the way we are forced to operationalize "department", as we discuss below.

[18] Because of imperfection in our definition of departments, this may be an over-estimate of choice homophily.

7 Permutation Test

As described in section 4 we use the permutation test to create our null model with permutations (eq. 2) to test whether observed tie-types are more or less likely than a model of random tie-formation would predict. We perform the permutation test at the aggregate level, and then at the department level.

At the aggregate level for each period, we permute the students 100 times in the following way. We retain the actual population of students and professors and the number of supervisions each supervisor performs (maintaining supervisor's degree sequence). The permutation consists in randomly assigning students to "supervision slots".[19] From this set we compute averages and confidence intervals for the four tie-types and we compare them with the observed tie-types in the system.

If there is no homophily in actual supervision relationships, the observed tie-types should lie within the confidence intervals of our permutation test.

At the aggregate level, some of the divergence between the permutation test and observed proportions of tie-types is that implicitly the permutation test assumes that each professor can supervise anyone: a physicist in Cape Town can supervise a student of law in Limpopo. But many actual constraints (geographic, disciplinary...) prevent the formation of such ties, making part of the observed divergence induced.

We can remove some of the induced homophily by repeating the permutation test at the department level. In particular, using the same basic procedure, we restrict the permutation to run within the university department. Our proxy for university departments is the combination of university name and scientific field of study. Our definition of departments is probably too broad to get precisely the right level of disaggregation. An example of university department is the couple "Economic sciences" and "University of Cape Town", a definition which considers business and economics together. Unfortunately our data do not permit a more precise definition of department. In particular, we consider 39 South African institutions and 18 scientific fields (broadly defined).[20]

[19] Our procedure preserves the degree sequence of student supervision, randomising the demographic composition of the ties. In this way it accounts for preferential attachment and other confounding factors like: willingness of supervisors to accept students, budget constraints, and faculty practices and norms.

[20] For South African Institutions we use post 2004 merger names listed below: University Stellenbosch, University Cape Town, University Pretoria, University Fort Hare, University Western Cape, University Free State, University KwaZulu Natal, University Johannesburg, University Limpopo, Durban Institute Technology, Tshwane University

21

In figure 2 we show permutation results with and without institutional constraints and for different sub-samples of the data. The left column of panels treats the entire country as one integrated system, imposing no constraints on who can supervise whom. The right column of panels constrains supervisions to take place within one "department". The sub-samples of the population are: identified students who enter academia (Future Academics), Science Engineering and Technology (SET), Social Sciences and Humanities (SSH), top institutions[22], and PhD students (the latter two are in Appendix B).

7.1 Permutation Test results

The permutation results for the different samples are consistent.[23] Figure 2 (figure 8 in the Appendix for gender) shows that student supervisor relations do display homophily along racial (gender) lines. The observed frequency of tie-types (solid lines) differs from those predicted by the permutation test (dashed lines). In particular, same-type supervision (WW, BB and MM, FF for gender) ties are more frequent than type-blind tie formation would imply and are well outside any reasonable confidence intervals; while cross-type ties (WB, BW and MF, FM for gender) are less frequent than those predicted by the null model with permutations.

Comparing permutation test results with and without institutional constraints (figures 2 and 8 from the right column of panels to the left column) the divergence between observed tie-type and the null model is much lower when institutional constraints are taken into account, though it remains significant. This shows the presence of a large amount of induced homophily, as we might expect in the circumstances.

Technology, Nelson Mandela Metropolitan University, Rhodes University, University Venda, Cape Peninsula University Technology, University South Africa, University Zululand, North West University, Walter Sisulu University, Vaal University Technology, University Witwatersrand, Vista University, Central University Technology, CGS, CSIR, ARC, SAAO, Natal Museum, Monash SA University, HMO, HSRC, NECSA, NHLS, MRC, NRF, Mintek, National Museum Bloemfontein, Sasol, DBSA.

[21] Scientific fields considered are: Agricultural sciences, Health Sciences, Biological sciences, Pharmaceutical Sciences, Chemical sciences, Technologies and applied sciences, Law, Engineering sciences, Physical sciences, Social Sciences, Medical sciences: Basic, Arts, Humanities, Earth and marine sciences, Mathematical sciences, Information and Computer science, Economic sciences, Medical sciences: Clinical.

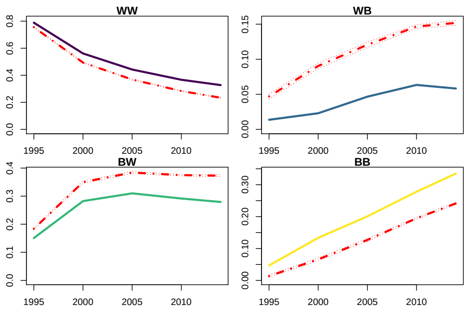
[22] As top institutions we include: Cape Town; Pretoria KwaZuluNatal; Stellenbosch; Rhodes; Witwatersrand; and Western Cape University.

[23] As a robustness check, we performed the analysis permuting ties in each year and then aggregating into our time periods. Results are the consistent.

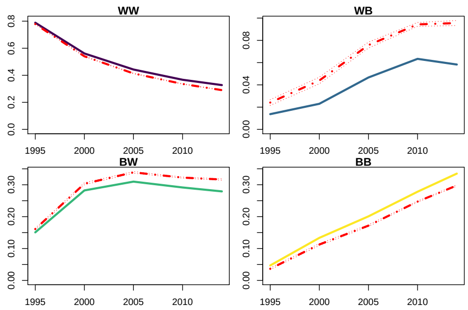
In the next section we use the modified assortativity coefficient to address the difference between induced and choice homophily more precisely.

Figure 2: Permutation tests for student supervisor data 1973-2014. The permutation is done keeping fixed the number of ties per supervisor and permuting the students. Permutation without constraint (left) and with institutional constraints (right). The Permutation is repeated 100 times for each of time period. The results of the permutation are plotted using dashed lines with two standard deviations on either side of the mean shown in the same colour. Solid lines in each plot show the proportions observed in our data. Tie-type labels (BB, BW, WW, WB) are read as student-supervisor.

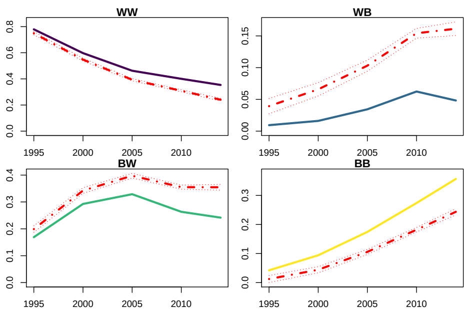
(a) All



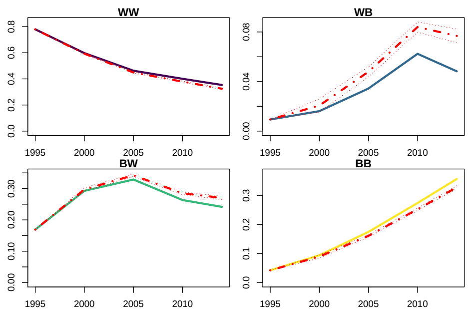
(b) All, University & Field



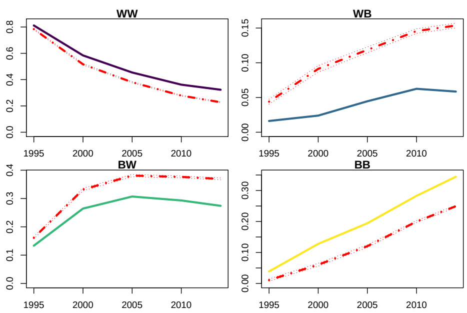
(c) Future Academics



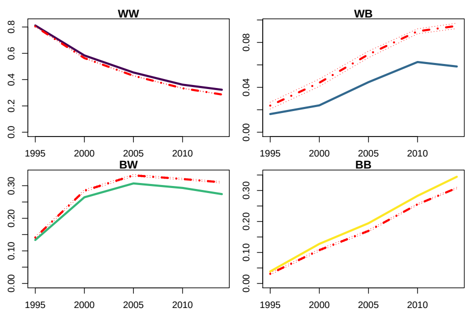
(d) Future Academics, University & Field



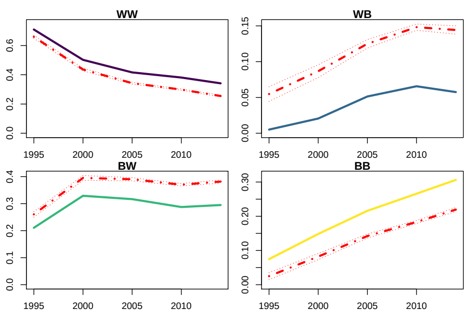
(e) SET only



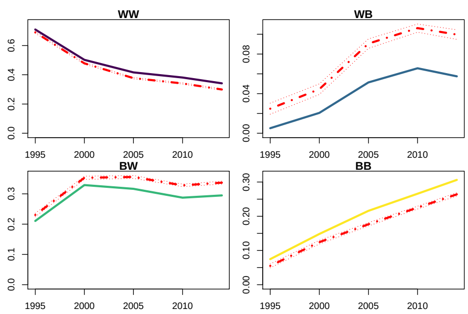
(f) SET only; University & Field



(g) SSH only



(h) SSH only; University & Field



8 Choice and Induced homophily

In this section we use the permutations at the two levels of aggregation to separate choice homophily from that induced by the system.

To interpret the results presented in figure 3 we recall the discussion in section 4. Expected values work well when populations are large (so you would be likely to observe the expected values assuming that we are close to a symmetric, uni-modal distribution) but for small populations, like ours, it is safer to draw several realisations of the null model with permutations to compare observation with the distribution of outcomes the null model would produce.

To measure homophily, we modify Newman's assortativity coefficient (in eq. 4) as follows. Instead of computing it as the distance between observed proportions of ties and expected values of random mixing in the population, we compare observed ties distribution with permutation results. The measure becomes

$$Ass_j = \frac{(BB_{obs} + WW_{obs}) - (B_{perm_j}B + W_{perm_j}W)}{1 - (B_{perm_j}B + W_{perm_j}W)}. \tag{5}$$

Where j is a realization of the permutation.[24] BB and WW are proportions of tie-types between agents of same type, and *obs* and *perm* stands for observed and permutation results. We present these realisations using boxplots.

In figure 3, in any single column in a time period we can compare upper and lower boxes. The upper shows assortativity at the system level, the lower assortativity when calculated with department level constraints. In the lower box we have supervisions constrained to take place within a department, which approximates a situation of zero induced homophily—in principle any student can be supervised by any professor. The lower box thus approximates choice homophily, and the difference between the two boxes is an estimate of induced homophily.

Figures 3(a) and 3(b) show choice and induced homophily along racial lines comparing the whole sample (grey background) with future academics (white background), and SET (grey background) with SSH (white background). The figures show that homophily computed at system level (upper boxes) is higher and roughly double that with department level constraints (lower boxes).

[24] The notation $B_{perm_j}B$ is adopted to indicate that students are randomly assigned to supervisors, retaining the degree sequence for supervisors.

This illustrates that a large part of the observed homophily is induced by department composition. Overall, choice homophily (lower boxes) shows a slightly increasing trend. This may be partially an artefact of social transformations, and we explore this further in section 10. In Figure 3(a), after 2001-2005, future academics (white background - upper boxes) show higher aggregate homophily than we see in the whole sample (grey background — upper boxes); by contrast, choice homophily (lower boxes) has similar levels. This implies that future academics experience larger institutional constraints and so higher induced homophily. This could be explained if particular departments specialize in certain sub-fields, and that there is a correlation between sub-field and racial composition of the academics in it. Such a correlation could arise from demonstration effects, wherein when choosing an academic discipline students ask whether "people like me" prosper in that discipline. This is often suggested as an explanation for the under-representation of women in SET subjects (Steele and Ambady, 2006; Bian et al., 2017; Blau and Kahn, 2017; Cheryan et al., 2017).

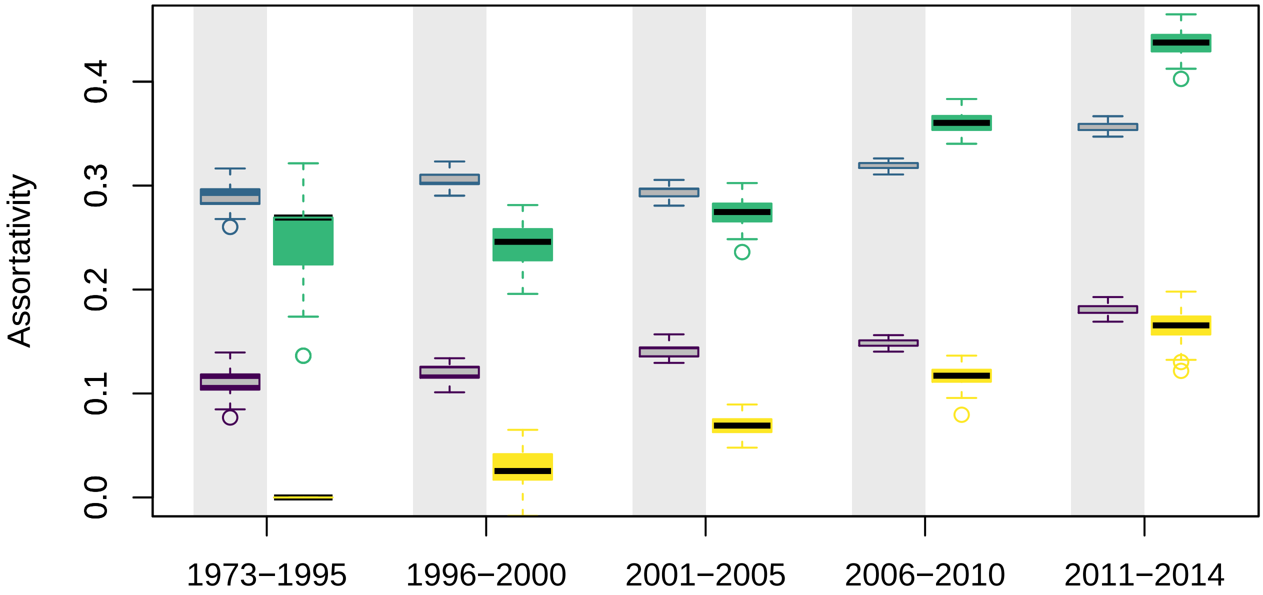
In figure 3(b) globally we find similar levels of homophily in SET (grey) and SSH (white), but a different composition. In SET (grey) the gap between total homophily and choice homophily appears to be larger than it is in SSH, implying stronger induced homophily. If a demonstration effect is at work, this suggests it is stronger in SET than in SSH.

Figure 10 in the Appendix gives the analogous results for gender. Gender-based homophily tends to be lower than race-based homophily in general. But trends over time and estimates of choice homophily are very similar for race and gender.

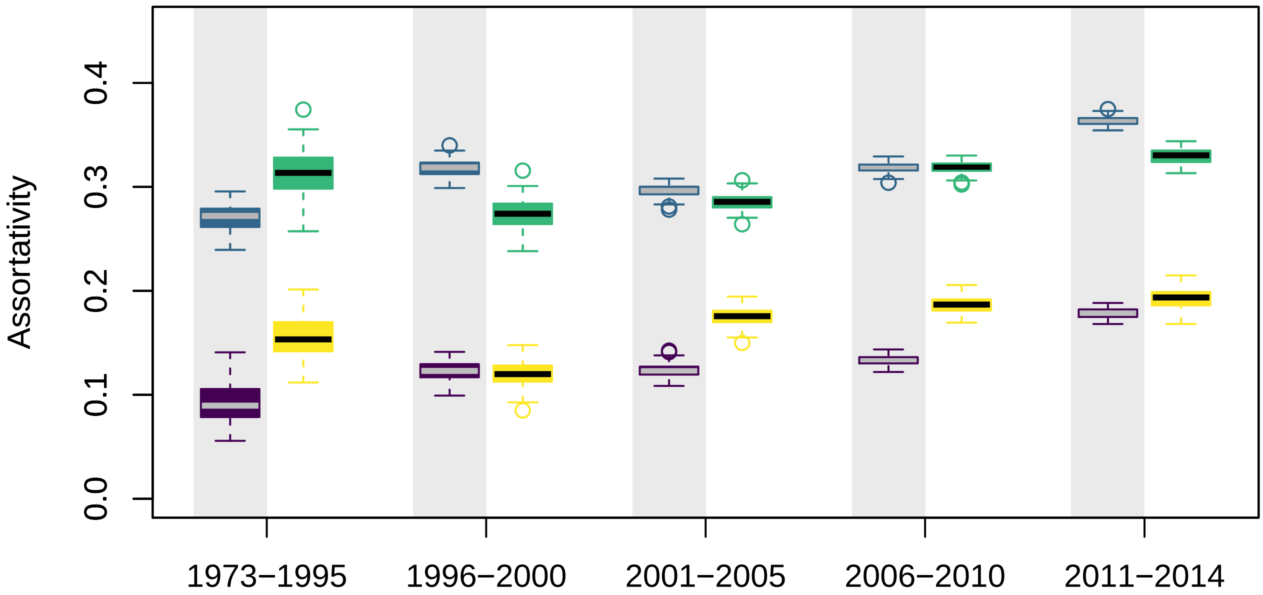
Our results underline that large part of racial and gender homophily is induced by the departments composition (university-field constraints). Additionally, the diverse composition of homophily between SET and SSH (showing higher levels of induced homophily in SET) suggest an unequal distribution of social transformation in the South African university system. In line with this Herman (2017) suggests that the knowledge divide created by the apartheid regime is still present. Partly with this in mind, South Africa academia was reformed in 2002-2004. The number of universities was reduced from 36 to 23 with mergers to redistribute resources and to remove the racial constraints and inequalities inherited from apartheid (Herman, 2017). However, newly merged and previously disadvantaged universities increased the number of PhDs in "soft/ or soft-applied fields such as education, philosophy or business, while the previously white universities continue to produce PhDs in hard/ or hard/applied subjects” (Herman, 2017) page. 1452.

Figure 3: Induced and Choice Homophily comparison looking racial ties in student supervisor relations 1973-2014. We compare results of assortativity of different sub-samples: All vs. Future Academics (a), SET vs. SSH (b). Each panel contains 4 series; In Fig.(a): aggregated (upper series) versus department level (lower series) crossed with total sample (white background) versus future academics (grey background). And similarly for Fig. (b). Upper series are system-level permutations, including both induced and choice homophily; lower series are department-level permutations, excluding much of the induced homophily. Each box plot represents 100 permutations and associated assortativity calculation, as described in section 4.2, and equation 4.

(a) All vs. Future Academics



(b) SET vs. SSH



9 Where does choice homophily originate?

In the previous sections, we measured choice homophily over time, having assumed that all types of participants (black, white, student, supervisor) display the same degree of homophily in their preferences. Given the history of South Africa, this might not be the case. In this section we present a method for distinguishing homophily levels among the different groups. To do this we expand our model of random tie formation to include distinct homophilous preferences for each sub-population of students and supervisors. Second, at the department level, we estimate the proportion of agents with homophilous preferences in the different populations of students and professors. Further we include a simple econometric analysis predicting the likelihood of same-type ties as robustness check. Lastly, we discuss the limitations of the assortativity coefficient in presence of social transformations.[25]

To simplify exposition of the model, we consider that each sub-population contains some members with strictly homophilous preferences (will only form same-type ties) and some who are completely "colour-blind" (form ties at random).[26] The four population groups are: white (male) students/professors, black (female) students/professors. We apply the model at the department level, thus eliminating induced homophily.

9.1 A Model of homophilous preferences of the different groups

The focus of this section is our simple model of random supervision tie formation, that we use to estimate the aggregate homophilous preferences of the population groups. In particular, we have a population of students and professors of two types w or b. We describe the event of student supervision ties as a two-stage process that we model as a probability tree. Students and professors are paired by drawing one member randomly from each population with replacement. First, we draw a student and then the professor.

This mimics the fact that in many disciplines (particularly at the lower levels) the student approaches professor for supervision.[27] The probability of observing tie-types are the possible

[25] Ideally one would like to allow yet another dimension: homophily among a particular group, white students for example, might be different in different contexts, in this case, at different universities, or even in different departments. In principle that is present in the model we present. However our data are not strong enough to include that aspect in our estimates, so we assume that university-level effects are not present.

[26] The model could as well be explicated in terms of "tendencies" for (non-)homophilous tie formation at the individual level. The analysis would be identical.

[27] In some disciplines at higher levels (Master and especially Ph.D.) professors do approach students. 27 % of our

outcomes. In the type-blind model, without homophilous preferences, they depend only on relative sizes of the different groups, as in equations 1. In this model, though, we add homophilous preferences.

In our model we assume that each sub-population has its own level of homophilous preferences. The level of homophilous preferences of a sub-population can be formalised as the frequency of agents in the sub-population who have strictly homophilous preferences, so they form ties only with those similar to them.[28]

So agents with homophilous preferences create links only with those of the same type and refuse links with dissimilar types while the agents with non-homophilous preferences link at random in their potential pool. For example the probability of observing a WW type is equal to the probability of an homophilous w student ($S_w h_{sw}$) plus the probability that a non-homophilous w student links with a w professor ($(S_w(1-h_{sw})\frac{Tw}{1-h_{tb}Tb})$). It follows that the probabilities of the different tie-types are simply

$$\begin{aligned} Pr(ww) =& S_w\left(\frac{T_w + h_{sw}(1-h_{tb})T_b}{1-h_{tb}T_b}\right) \\ Pr(wb) =& S_w T_b\left(\frac{(1-h_{sw})(1-h_{tb})}{1-h_{tb}T_b}\right) \\ Pr(bw) =& S_b T_w\left(\frac{(1-h_{sb})(1-h_{tw})}{1-h_{tw}T_w}\right) \\ Pr(bb) =& S_b\left(\frac{T_b + h_{sb}(1-h_{tw})T_w}{1-h_{tw}T_w}\right). \end{aligned} \tag{6}$$

Here h_{sw} and h_{sb} are respectively proportions of agents with homophilous preferences in the population of white and black students. h_{tw} and h_{tb} are those of white and black supervisors.

We use equation 6 to estimate the different hs of 4 sub-populations at the department level each period.

9.2 Choice Homophily of different groups

In this section we estimate the hs in the model presented above. Given the population composition of a department, for any assumed vector of $h = (h_{sw}, h_{tw}, h_{sb}, h_{tb})$ equation 6 predicts the

data are Ph.D. supervisions.

[28] Skvoretz (2013) underlines the importance of including in a model of intergroup relations two mechanism: attraction to similar and repulsion from dissimilar. In a simple way our model captures both mechanism as agents with strictly homophilous preferences form only same-type ties.

composition of supervision ties.

Thus, we compute in our data for each time period and department the proportions of tie-types and those of the different sub-population. We remove departments with fewer than 10 supervisions per period and without population variability. Specifically, we include departments with more than 10 supervisions per period which satisfy $0.1 < T_w, T_b, S_w, S_b < 0.9$. We estimate the four hs using the predictions of equation 6.

We follow a bootstrap re-sampling technique (Efron and Tibshirani, 1986). In each period, we create 10000 bootstrap samples $U' = \{u'_1, ..., u'_{10000}\}$ from the set U of size M of university departments.[29] For each bootstrap sample u' we estimate $h = (h_{sw}, h_{tw}, h_{sb}, h_{tb})$ minimising the average relative entropy according to the formula (Kullback and Leibler, 1951)

$$\overline{D_{KL}(p||q)} = \sum_{j \in J} \sum_{i \in I} p(i|j) log \frac{p(i|j)}{q(i|j)}. \tag{7}$$

Where $I = \{ww, wb, bw, bb\}$, $J = \{\text{departments in the bootstrap sample } u'\}$, p are empirical probabilities and q are predictions of our model.

In other words we minimize the average loss of information between observed proportions of tie-types and those predicted by our model in equation 6 at department level; obtaining each time an estimate of the relative proportions of agents with strictly homophilous preferences in each sub-population. We present the bootstrap distributions for our estimates. [30]

Figures 4 and 12 show the results for race and gender. In both cases, we find that students of the former majority (white and male) have the highest estimated homophilous preferences. Among professors, those previously excluded (black and female) have larger levels. This is not the case looking the sample of future academic students along racial lines. In this case, black professors have lower homophilous preferences than white professors in the last period, 2011-2014, in particular.

Our findings suggest that homophily originates from white (male) students and black (female) professors. Even though black (female) professors display homophilous preferences, their student counterparts do not. Since the creation of student supervisor ties (mostly) starts with a student

[29] M is the number of department observed in the data in the period.

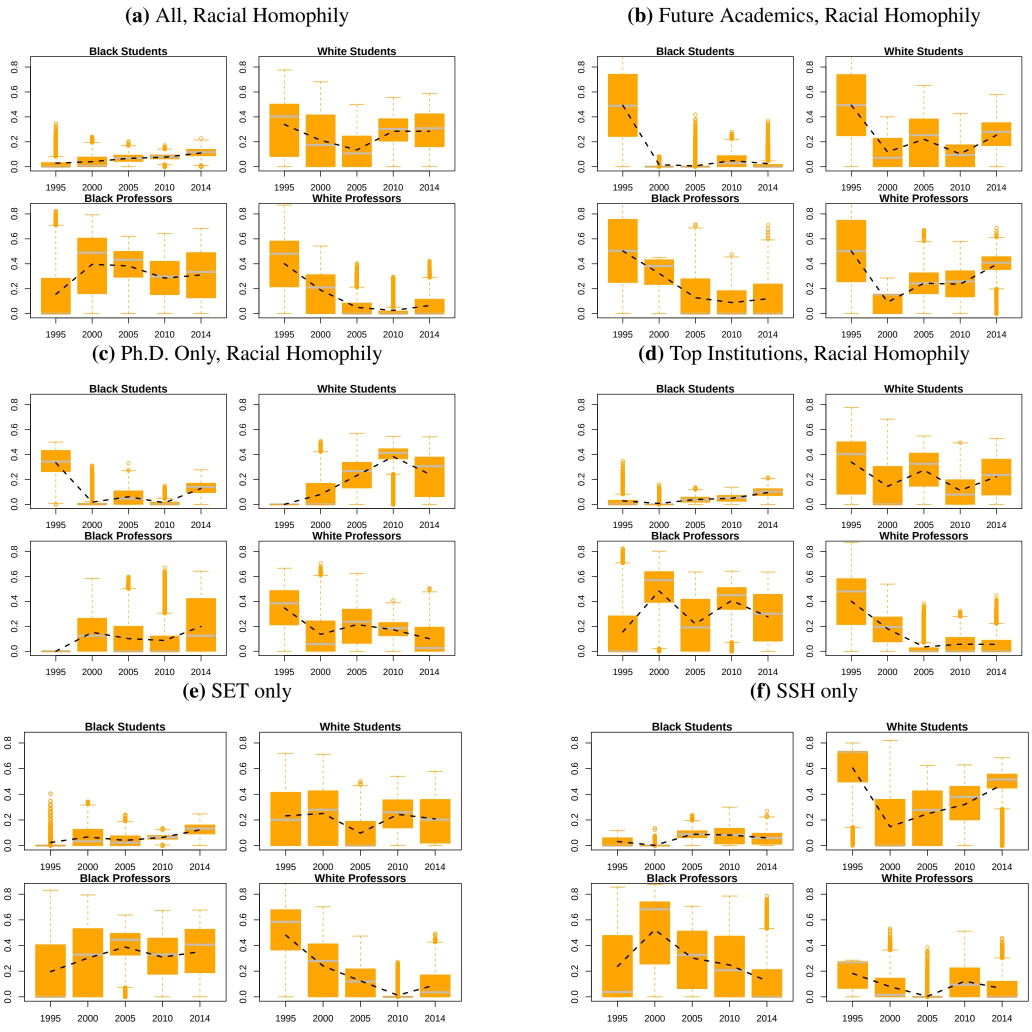
[30] For the minimization we used the limited memory algorithm for bound constrained optimization as in Byrd et al. (1995), under the optim R function.

who asks a professor to supervise his/her work, black (female) professors may face additional induced homophily that we do not capture. To the extent this is true, white (male) students drive tie formation and by this mechanism they may transmit homophily. For future academics this mechanism may be reinforced also by the high presence of white professors with homophilous preferences.

With respect to same-type of supervisions among under-represented groups the psychological literature underlines the importance of these ties in helping students of these groups and in particular in reducing their attrition rate (usually high for under-represented and minority groups). Main (2014), for example, finds that same-type supervision can be important for mentoring reasons and psychological considerations like empathy and identification. Therefore black (female) professors, likely to be aware (because of their personal experience) about the difficulties faced by under-represented groups of students, can be guided by other motives in tie formation. For example their formation for same-type ties may well be an expression of emphatic or mentoring motives rather than racial (gender) preferences.

The results for white (male) students are harder to rationalize. They are consistent, though, with the findings of Moolman (2010). In a survey of Afrikaner South Africans he found that younger people (aged 21-30) have high level of in-group identification and they perceive "inclusion policies" as threatening. This suggests that in that group at least the conflict hypothesis dominates. Further, the extent to which they experience threat is highly correlated with attitudes toward policies such as affirmative action and to its beneficiaries. Slightly older respondents to the survey were much more open to policies that favour the formerly disadvantaged. This general structure is present in our results.

Figure 4: Race Homophily of the different groups 1973-2014. The estimation is done minimising the average relative entropy of tie-types between observed and predicted from equations 6. We use data at the department level with 10000 samples with replacement of department observations. Black dashed lines are averages. Faculties with fewer than 10 observations and without population variability are excluded.



9.2.1 Robustness Check

As robustness check, to further access the presence of group differences in terms of homophilous preferences, we look at supervision at tie level yearly. For consistency we remove departments with fewer than 10 supervisions per period and without population variability, as above. We present logistic regression models that predict the likelihood of same-type tie versus cross-type tie with the following structure:

$$logit(p_{ij}) = \alpha + \beta X_i + \gamma X_j + \delta X_{ij} \tag{8}$$

Where ij is the tie between student i and professor j. X_i, X_j are agents' covariates and X_{ij} are tie covariates. Our main independent variables are the dummies student and professor race (gender), both equal to 1 for white (male) and zero otherwise. Additional controls are student and professor gender (race), registration year, proportions of black (female) students and professors in the system (computed in the 5 time periods considered), professor age, professor rating[31] (a control for professor "quality") [32], and department dummies (institutional constraints of university and field). In terms of group differences in homophilous preferences we interpret, for example, a positive coefficient of student race as an indication that white students are more homophilous than black - more likely to form same-type ties compared to black students.

Table 2 models 1 and 2 show results with same-race ties as the binary dependent variable, and model 3 and 4 for same-gender. Models 2 and 4 are for the sub-sample of future academics. The results we found in the previous section are confirmed: white (male) students are more likely to form same-race (same-gender) ties than black (female) students; while white (male) professors are less likely than black (female) to form same-race (same-gender) ties. Model 2, looking at same-race ties, confirms the larger homophilous preferences of white professors compared to black professors for the sub-sample of future academics. In the Appendix, table 4 for same-race and table 5 for same-gender ties we include other model specifications, controlling for population sizes and results are unchanged.

[31] Professor ratings are individual ratings for years 1983 – 2012, which measures individuals' academic performance. The NRF grades researchers following a rigorous examination of a candidate's research output. The process involves international referees who evaluate the CV and published papers of the professor. This process ends with a rating: one of 7 ordered categories. See Section 3.

[32] We have to remark that this variable reduces our sample. To account for this, we estimated the presented models also excluding the variable and we found the same results. The rating variable does not show a systematic effect.

Table 2: Results of logistic regressions of yearly student-supervisor tie data. Dependent variable same-race (same-gender) is 1 for same-race (same-gender) ties and zero for cross-race (cross-gender) ties. Models 2 and 4 are for the sub-sample of future scholars. Faculties with fewer than 10 observations and without population variability are excluded. Where variable studrace (studgender) is a students' dummy 1 for white (male); race (gender) is a professors' dummy 1 for white (male). Age is professor's age; YearFirstRegistration is the students' registration year. Universities + Field are department dummies, and Rating is professors' NRF rating (7 rating categories).

	Dependent variable:			
	same-race ties		same-gender ties	
	(1)	(2)	(3)	(4)
studrace	3.026^{***}	2.655^{***}	0.058^{**}	0.114^{*}
	(0.050)	(0.105)	(0.026)	(0.065)
race	-0.834^{***}	1.052^{***}	0.023	-0.142^{*}
	(0.050)	(0.117)	(0.031)	(0.082)
studgender	-0.092^{**}	-0.024	0.893^{***}	1.141^{***}
	(0.043)	(0.103)	(0.024)	(0.060)
gender	0.197^{***}	0.193^{*}	-0.605^{***}	-0.960^{***}
	(0.045)	(0.104)	(0.026)	(0.063)
age	0.001	-0.003	-0.002	-0.014^{***}
	(0.003)	(0.006)	(0.001)	(0.004)
YearFirstRegistration	0.002	0.026	-0.004	0.014^{*}
	(0.006)	(0.018)	(0.003)	(0.008)
Constant	-3.696	-54.197	7.161	-27.032
	(11.367)	(37.039)	(6.619)	(16.499)
Universities + Field	*yes*	*yes*	*yes*	*yes*
Rating	*yes*	*yes*	*yes*	*yes*
Observations	14,924	2,934	33,310	6,026
Log Likelihood	-7,352.243	-1,411.708	-21,570.380	-3,752.652
Akaike Inf. Crit.	14,830.490	2,935.415	43,322.770	7,687.304
Note:			*p<0.1; **p<0.05; ***p<0.01.	

10 An simple model with changing populations and unchanging homophily

Earlier results in the paper suggest that the preference for similarity is increasing over time. This would be somewhat distressing if true. However this could be a composition effect of changes in the sup-population.[33] To illustrate this effect consider a simple model with $S_w = T_w$ and the same unchanging level of homophilous preferences in all groups $h_{sw} = h_{sb} = h_{tw} = h_{tb} = 0.3$. We start with a homogeneous population of 500 students and 500 supervisors with initially $S_w = T_w = 1$. We progressively add the previously excluded groups, so that S_b and T_b, which begin equal to zero, slowly rise.[34] We present the results as a function of N_b, simply the number of blacks in the system.[35] We make supervision assignments using the model in section 9.1 and then calculate proportions of tie-types and assortativity mixing.

Figure 5 shows model predictions and assortativity as a function of the number of blacks on the x-axis. Figure 5(a) relates to figure 2 and figure 5(b) to figure 3. Even though by construction tastes-for-similarity do not change over time, we observe that in the initial phase of the introduction of the excluded groups there is a strong change in all panels: Over time, as the number of blacks in the system (N_b) increases, the gaps between the null model (eq. 1) predictions and "observed" tie-types initially rises; assortativity makes a steep rise, before it falls slowly;

This underlines that at the beginning of the process of social transformations assortativity mechanically increases even when the level of homophilous preferences stays constant. Thus, some of the increasing trend in homophily found in section 8 could be an artefact of the increasing number of the formerly under-represented having entered the system (particularly into the formerly white part of it). It suggests that social transformations in South Africa academia may still be in the transition phase. This section suggests an additional complication in the study of homophily in a context of social transformations, which is that standard measures could be open to misleading interpretations and apparent changes in preferences might in fact, paradoxically, be due simply to changes in population composition and the interaction of thereof.

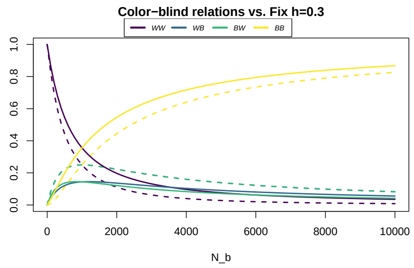
[33] We thank Prof. Alan Kirman for his suggestions related to this section.

[34] For simplicity students and professors are always of equal size, yet the main result here is consistent with a case with a different rates of sub-population growth.

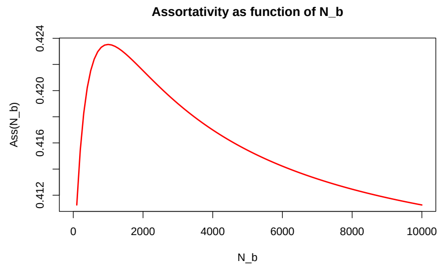
[35] N_b is equal to the number of black students plus black supervisors where both have equal size.

Figure 5: Model predictions as a function of number of black in the population with a starting population of 1000 white. The model is with one population (Sw=Tw) and h=0.3 for all groups. Solid lines for model with h=0.3 and dashed lines for model h=0.

(a) Tie types



(b) Assortativity mixing



11 Discussion

In this paper we have examined the composition of student-supervisor relationships in South African universities. Following the end of apartheid and the formal de-segregation of the universities, we observe an ongoing transformation in its transition phase. In this context student-supervision relationships can be central in determining the path of a student's future career, both if a student leaves academia and (and perhaps especially) if the student follows a career within the university system. Consequently understanding what, if any, role of homophily plays in those relationships may be important.

Homophily, that is a tendency to form links with individuals similar to oneself, is ubiquitous in social life. It can be of value, if information is very asymmetric or incomplete, in that it provides a schema on which to make generalizations. But at the same time it can be harmful, as it can engender and perpetuate ascripitive inequality. Homophily can make a transition from a segregated society (such as apartheid South Africa) to an integrated society, difficult, and very lengthy.

In this paper we developed a method to measure choice homophily that controls for opportunities for cross-type relations and network structures, (that is, removing induced homophily). In line with past research in other contexts, our results suggest that race and gender homophily is present in student-supervisor relationships. By decomposing homophily we find that induced and choice homophily have a similar magnitude. This underlines how both can play a role.

In our results there also appear to be trends. There is some evidence that the strength of homophily is student-supervisor relationships is increasing over time. This observation must be treated very carefully though. First, because of the nature of our data. There are few observations in the first period (1970-1994), and to a lesser extent second (1994-2000). Second there was a major reform of the universities in 2004, so there was a break in the very structure of the system which was intended to change the ethnic composition of universities. Finally, particularly in figure 4 where we make group-specific estimates of homophily levels, there is considerable variation within our observations, so trends based on mean values must be treated with considerable caution.

We observe that, though these caveats notwithstanding, there appears to be some evidence of increasing homophily. We can observe this in the simple permutation test (section 7.1) where the gap between what we would expect in a type-blind assignment of students to supervisors and what we actually observe appears to be growing. We make similar observations in the assortativity measures (section 8) and in some of the panels in Figure 4. This hints at a hardening of attitudes, as suggested by Moolman (2010). These results should be interpreted carefully, however, because some of this increasing divergence is probably due simply to a change in the composition of the population of students and supervisors and their interaction. Even with no change in homophily levels, we observe in our simple model in section 10 that in the short run there will be an increase in divergence from "type-blindness" as previously excluded groups enter the system.

When we estimate homophily levels for each sub-population (students and supervisors, black and white) we observe that white professors have the lowest measured homophily levels. The highest levels are found in white students, who exhibit a preference for white supervisors.

Several cautionary notes should be added here. First is that in some cases we have small samples. This is particularly so in the first and second of our periods. (And we should note that the first period is very anomalous as it is during the apartheid era.) Even in our third and fourth periods, though, when we disaggregated to the department level to remove the induced homophily, some of our departments are quite small. Thus caution should be exercised when drawing very general conclusions, particularly as regards trends.

Second, we should observe that our data are drawn from a particular part of the system. The basic observations are drawn from faculty members who have a strong research career. This places a bias both at the individual level and at the university level. So universities or faculty members who have decided to focus on, for example, undergraduate teaching will to a large extent be missing from our analysis. Thus we cannot draw conclusions about the entire system, only a particular part of it. On the other hand, one could argue that the part of the system under-represented in our data has, due to history, become more rapidly representative of the population as a whole, and so the issue of transformation is less severe there. And further, it is the "prestigious" part of the system that is most important symbolically, academically, politically and socially, and so the focus of efforts on transformation should be on that part of the system.

Third our definition of "department" is somewhat imprecise. We do not have data on formal department affiliation of our faculty, we only have a statement of their broad research interests. Thus in our disaggregation from system to "department" level we will not have removed all of the induced homophily. So our estimates of the level of choice homophily are likely to be upper bounds. It is very difficult to tell whether this over-estimation has any temporal pattern though.

Whether or not high choice homophily is good or bad for student supervision, it is not ideal for producing new ideas, which usually need the recombination of different knowledge and perspectives (Bertrand and Duflo, 2017). This is particularly relevant for social sciences and humanities. Moreover, the observation that homophily is strongest in white, male students could make the transition phase of social transformations longer, preventing the formation of cross-type ties. In a dynamic perspective, since disciplines are chosen by students, if homophily among students is strong, we will see a segregation at the discipline level. This is already present to some extent, with black students tending to choose subjects in SET (Barnard et al., 2016). But again, future development of South African society will demand a change in social perspective which will be encouraged by a stronger presence of black scholars in the humanities and social sciences. This observation is more troubling when we look at future academics, where patterns of homophilous relations seem stronger.

South African society is changing, slowly de-segregating, from a society legally divided on racial lines. The transformation is not complete, and in many places there remains a considerable gap between the current situation and aspirations. The formerly white universities are transforming but the process in still ongoing, and how relations among students and supervisors evolve across race and gender lines is central to how quickly and effectively transformation proceeds.

12 Conclusion

Our results suggest that targeting together induced and choice homophily could accelerate the transition phase of social transformations. Further, looking at group interactions, we show that homophily in the system might originates from the group that feels more threatened and perhaps less able to conceptualize the potential gains that an inclusion policy brings for all. For this, future programmes and research aim at decreasing homophily, targeted to this group could be very effective. In this respect, the review of Cooper et al. (1999) looked at 8 cooperative learning programs focusing at team outcomes for students at school and found that they can increase the prevalence and quality of interracial friendships. However little research has been conducted at the university level. In the context of academia in particular, encouraging cross-type supervision and mentoring (also across department boundaries) could be a potential channel for transformations and foster collaborations between universities. The slow path of social transformations harms many societies', to address it with a more holistic view beyond looking only at demographic changes could have a large potential and create a better and more productive environment for all.

References

Alderfer, C. P. and K. K. Smith (1982). Studying intergroup relations embedded in organizations. *Administrative Science Quarterly*, 35–65.

Allport, G. W., K. Clark, and T. Pettigrew (1954). The nature of prejudice.

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. *Handbook of labor economics 3*, 3143–3259.

Aronson, J., M. J. Lustina, C. Good, K. Keough, C. M. Steele, and J. Brown (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of experimental social psychology 35*(1), 29–46.

Arrow, K. et al. (1973). The theory of discrimination. *Discrimination in labor markets 3*(10), 3–33.

Åslund, O. and O. N. Skans (2010). Will i see you at work? ethnic workplace segregation in sweden, 1985–2002. *ILR Review 63*(3), 471–493.

Astin, A. W. and H. S. Astin (2000). Leadership reconsidered: Engaging higher education in social change.

Barnard, H., R. Cowan, and M. Müller (2012). Global excellence at the expense of local diffusion, or a bridge between two worlds? research in science and technology in the developing world. *Research Policy 41*(4), 756–769.

Barnard, H., R. A. Cowan, A. P. Kirman, and M. Müller (2016). Including excluded groups: The slow racial transformation of the south african university system. Technical report, Working Paper Series in Economics, Karlsruher Institut für Technologie (KIT).

Barnett, G. A. and G. A. Benefield (2017). Predicting international facebook ties through cultural homophily and other factors. *new media & society 19*(2), 217–239.

Barnett, I., T. Khanna, and J.-P. Onnela (2016). Social and spatial clustering of people at humanity's largest gathering. *PloS one 11*(6), e0156794.

Beasley, M. A. and M. J. Fischer (2012). Why they leave: The impact of stereotype threat on the attrition of women and minorities from science, math and engineering majors. *Social Psychology of Education 15*(4), 427–448.

Becker, G. S. (1957). 1971. the economics of discrimination.

Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. In *Handbook of Economic Field Experiments*, Volume 1, pp. 309–393. Elsevier.

Bian, L., S.-J. Leslie, and A. Cimpian (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science 355*(6323), 389–391.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Blau, P. M. (1977). *Inequality and heterogeneity: A primitive theory of social structure*, Volume 7. Free Press New York.

Bramoullé, Y., S. Currarini, M. O. Jackson, P. Pin, and B. W. Rogers (2012). Homophily and long-run integration in social networks. *Journal of Economic Theory 147*(5), 1754–1786.

Brief, A. P., E. E. Umphress, J. Dietz, J. W. Burrows, R. M. Butz, and L. Scholten (2005). Community matters: Realistic group conflict theory and the impact of diversity. *Academy of Management Journal 48*(5), 830–844.

Burris, V. (2004). The academic caste system: Prestige hierarchies in phd exchange networks. *American sociological review 69*(2), 239–264.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing 16*(5), 1190–1208.

Cajner, T., T. Radler, D. Ratner, and I. Vidangos (2017). Racial gaps in labor market outcomes in the last four decades and over the business cycle. *Finance and Economics Discussion Series 2017-071*.

Cheryan, S., S. A. Ziegler, A. K. Montoya, and L. Jiang (2017). Why are some stem fields more gender balanced than others? *Psychological Bulletin 143*(1), 1.

Cooper, R. et al. (1999). Improving intergroup relations: Lessons learned from cooperative learning programs. *Journal of Social issues 55*(4), 647–663.

Cowan, R. and G. Rossello (2018). Emergent structures in faculty hiring networks, and the effects of mobility on academic performance. *Scientometrics 117*(1), 527–562.

Crisp, G. and I. Cruz (2009). Mentoring college students: A critical review of the literature between 1990 and 2007. *Research in higher education 50*(6), 525–545.

Cruz-Castro, L. and L. Sanz-Menéndez (2010). Mobility versus job stability: Assessing tenure and productivity outcomes. *Research Policy 39*(1), 27–38.

Currarini, S., M. O. Jackson, and P. Pin (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica 77*(4), 1003–1045.

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, 54–75.

Feld, S. L. (1982). Social structural determinants of similarity among associates. *American sociological review*, 797–801.

Gasman, M., U. Abiola, and C. Travers (2015). Diversity and senior leadership at elite institutions of higher education. *Journal of diversity in higher education 8*(1), 1.

Gersick, C. J., J. E. Dutton, and J. M. Bartunek (2000). Learning from academia: The importance of relationships in professional life. *Academy of Management Journal 43*(6), 1026–1044.

Girves, J. E., Y. Zepeda, and J. K. Gwathmey (2005). Mentoring in a post-affirmative action world. *Journal of Social Issues 61*(3), 449–479.

Granovetter, M. S. (1973). The strength of weak ties1. *American Journal of Sociology 78*(6), 1360–1380.

Greenhaus, J. H. and S. Parasuraman (1993). Job performance attributions and career advancement prospects: An examination of gender and race effects. *Organizational Behavior and Human Decision Processes 55*(2), 273–297.

Herman, C. (2017). Looking back at doctoral education in south africa. *Studies in Higher Education 42*(8), 1437–1454.

Hewstone, M. and H. Swart (2011). Fifty-odd years of inter-group contact: From hypothesis to integrated theory. *British Journal of Social Psychology 50*(3), 374–386.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2013). The allocation of talent and us economic growth. Technical report, National Bureau of Economic Research.

Huffman, M. L. and P. N. Cohen (2004). Racial wage inequality: Job segregation and devaluation across us labor markets. *American Journal of Sociology 109*(4), 902–936.

Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative science quarterly*, 422–447.

Ibarra, H. (1997). Paving an alternative route: Gender differences in managerial networks. *Social psychology quarterly*, 91–102.

Jackson, M. O. (2005). A survey of network formation models: stability and efficiency. *Group Formation in Economics: Networks, Clubs, and Coalitions*, 11–49.

Jacobi, M. (1991). Mentoring and undergraduate academic success: A literature review. *Review of educational research 61*(4), 505–532.

Jost, J. T., M. R. Banaji, and B. A. Nosek (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political psychology 25*(6), 881–919.

Kirman, A. (2011). Complex economics. *NY: Routledge*.

Kossek, E. E., R. Su, and L. Wu (2017). "opting out" or "pushed out"? integrating perspectives on women's career equality for gender inclusion and interventions. *Journal of Management 43*(1), 228–254.

Kossinets, G. and D. J. Watts (2009). Origins of homophily in an evolving social network. *American journal of sociology 115*(2), 405–450.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The annals of mathematical statistics 22*(1), 79–86.

Lang, K. and J.-Y. K. Lehmann (2012). Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature 50*(4), 959–1006.

León, L., R. Cowan, and M. Müller (2016). Formation and persistence of research communities in middle income countries: The case of south africa. *UNUMERIT Working paper*.

Levine, R. A. and D. T. Campbell (1972). Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior.

Main, J. B. (2014). Gender homophily, ph. d. completion, and time to degree in the humanities and humanistic social sciences. *The Review of Higher Education 37*(3), 349–375.

McKeown, S. and J. Dixon (2017). The "contact hypothesis": Critical reflections and future directions. *Social and Personality Psychology Compass 11*(1), e12295.

McPherson, J. M. and L. Smith-Lovin (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, 370–379.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology 27*(1), 415–444.

Mollica, K. A., B. Gray, and L. K. Treviño (2003). Racial homophily and its persistence in newcomers' social networks. *Organization Science 14*(2), 123–136.

Montgomery, J. D. (1991). Social networks and labor-market outcomes: Toward an economic analysis. *The American economic review 81*(5), 1408–1418.

Moolman, J. F. (2010). *The role of threat on Afrikaner attitude towards affirmative action and its beneficiaries*. Ph. D. thesis, University of Pretoria.

Neilson, W. and S. Ying (2016). From taste-based to statistical discrimination. *Journal of Economic Behavior & Organization 129*, 116–128.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature 56*(3), 799–866.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E 67*(2), 026126.

Pettigrew, T. F. and L. R. Tropp (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology 90*(5), 751.

Pezzoni, M., J. Mairesse, P. Stephan, and J. Lane (2016). Gender and the publication output of graduate students: A case study. *PLoS One 11*(1), e0145146.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, 659–661.

Phillips, K. W., N. P. Rothbard, and T. L. Dumas (2009). To disclose or not to disclose? status distance and self-disclosure in diverse environments. *Academy of Management Review 34*(4), 710–732.

Pin, P. and B. W. Rogers (2016). Stochastic network formation and homophily.

Podolny, J. M. (1994). Market uncertainty and the social character of economic exchange. *Administrative science quarterly*, 458–483.

Rydgren, J. (2004). Mechanisms of exclusion: ethnic discrimination in the swedish labour market. *Journal of Ethnic and Migration Studies 30*(4), 697–716.

Saunders, P. (2006). *Social class and stratification*. Routledge.

Schelling, T. et al. (1978). Micromotives and macrobehavior. *WW Norton & Company, New York,*.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology 1*(2), 143–186.

Sharmeen, F., T. Arentze, and H. Timmermans (2014). Dynamics of face-to-face social interaction frequency: role of accessibility, urbanization, changes in geographical distance and path dependence. *Journal of Transport Geography 34*, 211–220.

Shavit, Y. et al. (2007). *Stratification in higher education: A comparative study*. Stanford University Press.

Sherif, M. (1966). *Group conflict and co-operation: Their social psychology*. London: Routledge and Kegan Paul.

Sherif, M. (2015). *Group conflict and co-operation: Their social psychology*. Psychology Press.

Sherif, M., O. Harvey, W. Hood, and C. Sherif (1961). Intergroup conflict and group relations: The robbers cave experiment. *The Institute of Group Relations, Norman, OK*.

Skvoretz, J. (2013). Diversity, integration, and social ties: Attraction versus repulsion as drivers of intra-and intergroup relations. *American Journal of Sociology 119*(2), 486–517.

Smith, J. A., M. McPherson, and L. Smith-Lovin (2014). Social distance in the united states: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004. *American Sociological Review 79*(3), 432–456.

Stainback, K. (2008). Social contacts and race/ethnic job matching. *Social Forces 87*(2), 857–886.

Steele, J. R. and N. Ambady (2006). "math is hard!" the effect of gender priming on women's attitudes. *Journal of Experimental Social Psychology 42*(4), 428–436.

Terrell, M. C. and R. K. Hassell (1994). Mentoring undergraduate minority students: An overview, survey, and model program. *New directions for teaching and learning 1994*(57), 35–45.

Wenger, E. (2010). Communities of practice and social learning systems: the career of a concept. In *Social learning systems and communities of practice*, pp. 179–198. Springer.

Zebrowitz, L. A., B. White, and K. Wieneke (2008). Mere exposure and racial prejudice: Exposure to other-race faces increases liking for strangers of that race. *Social cognition 26*(3), 259–275.

Zhou, S., E. Page-Gould, A. Aron, A. Moyer, and M. Hewstone (2018). The extended contact hypothesis: A meta-analysis on 20 years of research. *Personality and Social Psychology Review*, 1088868318762647.

A Future Academics

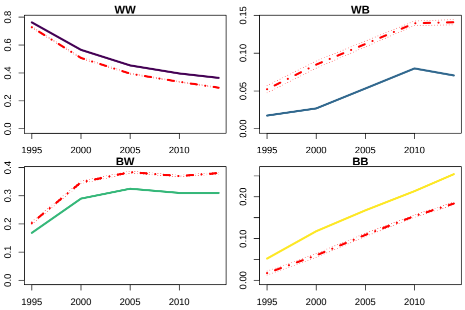
Table 3: Student supervision composition subsample of Future Academics. Where w for white, b for black, m for male, f for female, S for students, and T for professors. Ass. is abbreviation for assortativity. Expected values of random mixing are in parenthesis. To note: column 1 and 2 should be treated carefully, both have few observations. Column 1 represents the apartheid era and column 2 is before the reform of the university system.

	1973-1995	1996-2000	2001-2005	2006-2010	2011-2014
			Populations		
S_w	0.79	0.61	0.50	0.46	0.40
S_b	0.21	0.39	0.50	0.54	0.60
T_w	0.93	0.87	0.75	0.64	0.58
T_b	0.07	0.13	0.25	0.36	0.42
			Supervisions		
WW	0.78 (0.73)	0.60 (0.53)	0.46 (0.38)	0.40 (0.29)	0.35 (0.23)
WB	0.01 (0.06)	0.02 (0.08)	0.03 (0.13)	0.06 (0.17)	0.05 (0.17)
BW	0.17 (0.20)	0.29 (0.34)	0.33 (0.38)	0.26 (0.35)	0.24 (0.35)
BB	0.04 (0.01)	0.09 (0.05)	0.17 (0.13)	0.27 (0.19)	0.36 (0.25)
$Ass.w/b$	0.28	0.26	0.26	0.35	0.44
			Populations		
S_m	0.67	0.51	0.51	0.46	0.40
S_f	0.33	0.49	0.49	0.54	0.60
T_m	0.79	0.73	0.67	0.61	0.58
T_f	0.21	0.27	0.33	0.39	0.42
			Supervisions		
MM	0.60 (0.53)	0.40 (0.37)	0.37 (0.34)	0.33 (0.28)	0.29 (0.23)
MF	0.07 (0.14)	0.11 (0.14)	0.14 (0.17)	0.12 (0.18)	0.11 (0.17)
FM	0.23 (0.26)	0.32 (0.36)	0.30 (0.33)	0.30 (0.33)	0.29 (0.35)
FF	0.10 (0.07)	0.17 (0.13)	0.19 (0.16)	0.24 (0.21)	0.30 (0.25)
$Ass.m/f$	0.25	0.13	0.11	0.15	0.21

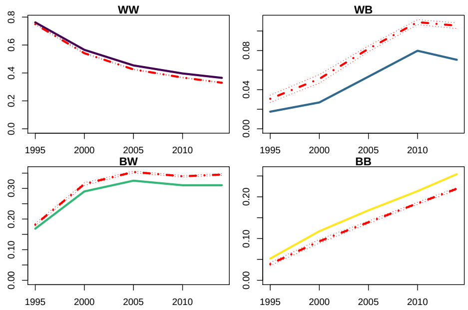
B Permutation Test results, Choice and Induced Homophily

Figure 6: Permutation test for student supervisor data 1973-2014 for top institutions and Phd students. The permutation is done keeping fixed the number of ties and the supervisor and permuting the students. Permutation without constraint (left) and with institutional constraints (right). Permutation is repeated 100 times for each of time period. The Observed proportions of racial student-supervisor relations (solid lines) are plotted versus the results of the permutation test (dashed lines) with respective two standard deviation from the mean. As top institutions we include: Cape Town; Pretoria KwaZuluNatal; Stellenbosch; Rhodes; Witwatersrand; and Western Cape University;"

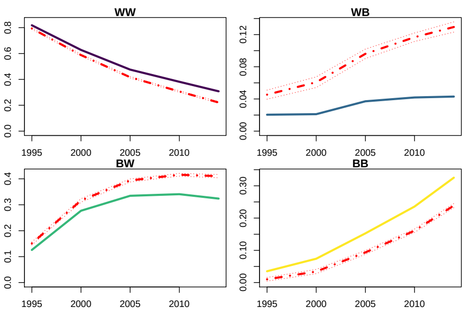
(a) Top Institutions



(b) Top Institutions; University & Field



(c) PhD Students only



(d) PhD Students only; University & Field

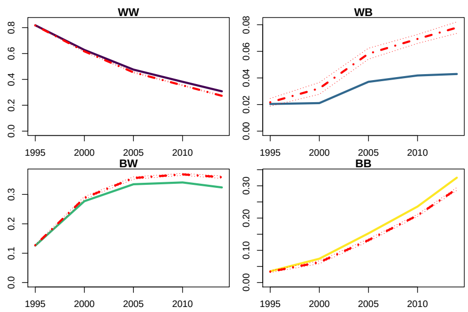
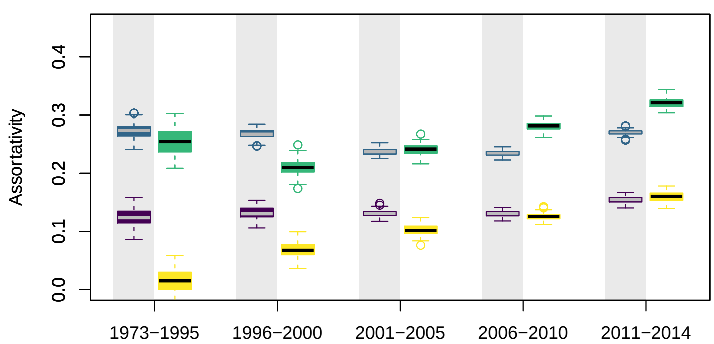


Figure 7: Induced and Choice Homophily comparison looking racial ties in student supervisor relations 1973-2014. We compare results of assortativity of two sub-samples: Top Universities vs. Ph.D. Sub-sample . The panel contains 4 series: aggregated (upper series) versus department level (lower series) crossed with Top Universities (white background) versus Ph.D. (grey background). Upper series are system-level permutations, including both induced and choice homophily; lower series are department-level permutations, excluding much of the induced homophily. Each box plot represents 100 permutations and associated assortativity calculation, as described in section 4.2, and equation 4.

(a) Top Universities vs. Ph.D.



C Robustness check

Table 4: Results of logistic regressions of yearly student-supervisor racial tie data. Dependent variable same-race is 1 for same-race ties and zero for cross-race ties. Models 3 and 4 are for the sub-sample of future scholars. Faculties with fewer than 10 observations and without population variability are excluded. Where variable studrace (studgender) is a students' dummy 1 for white (male); race (gender) is a professors' dummy 1 for white (male). Age is professor's age; YearFirstRegistration is the students' registration year. Sb and Tb are proportions of black students and professors aggregated for each time period. Universities + Field are department dummies, and Rating is professors' NRF rating (7 rating categories).

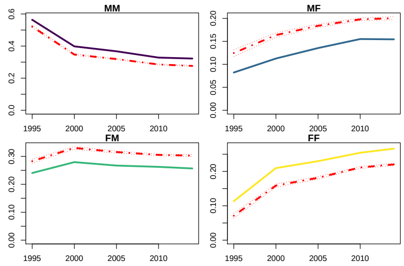
	Dependent variable:			
	same-race ties			
	(1)	(2)	(3)	(4)
studrace	3.025***	3.027***	2.655***	2.659***
	(0.050)	(0.050)	(0.105)	(0.105)
race	−0.834***	−0.837***	1.053***	1.052***
	(0.050)	(0.050)	(0.117)	(0.117)
studgender	−0.092**	−0.092**	−0.024	−0.021
	(0.043)	(0.043)	(0.103)	(0.103)
gender	0.197***	0.195***	0.193*	0.193*
	(0.045)	(0.045)	(0.104)	(0.104)
age	0.001	0.002	−0.003	−0.003
	(0.003)	(0.003)	(0.006)	(0.006)
YearFirstRegistration	−0.008	−0.016	0.022	0.008
	(0.015)	(0.016)	(0.037)	(0.038)
Sb	−1.257	−0.213	−0.902	−5.404
	(1.838)	(1.927)	(11.861)	(12.286)
Tb	1.526	−8.331	0.921	−16.390
	(1.584)	(5.564)	(8.285)	(14.659)
Sb:Tb		12.410*		27.080
		(6.715)		(18.899)
Constant	14.864	32.915	−46.270	−14.099
	(30.474)	(32.008)	(72.927)	(76.235)
Universities + Field	*yes*	*yes*	*yes*	*yes*
Rating	*yes*	*yes*	*yes*	*yes*
Observations	14,924	14,924	2,934	2,934
Log Likelihood	-7,351.759	-7,350.050	-1,411.695	-1,410.666
Akaike Inf. Crit.	14,833.520	14,832.100	2,939.390	2,939.332
Note:				*p<0.1; **p<0.05; ***p<0.01

D Gender Analysis

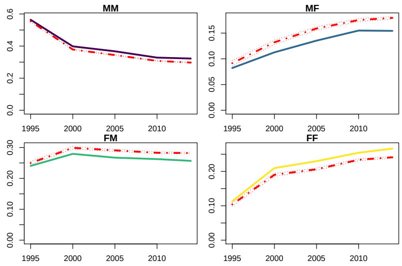
D.1 Permutation Test

Figure 8: Permutation test for student supervisor data 1973-2014. The permutation is done keeping fixed the number of ties and the supervisor and permuting the students. Permutation without constraint (left) and with institutional constraints (right). Permutation is repeated 100 times for each of time period. The Observed proportions of gender student-supervisor relations (solid lines) are plotted versus the results of the permutation test (dashed lines) with respective two standard deviation from the mean.

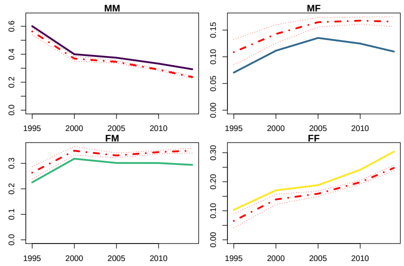
(a) All, no constraints



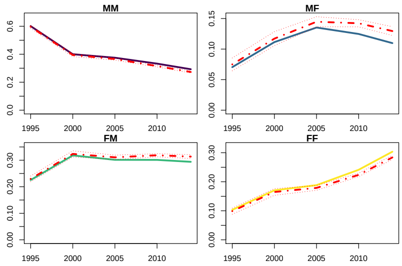
(b) All, University & Field



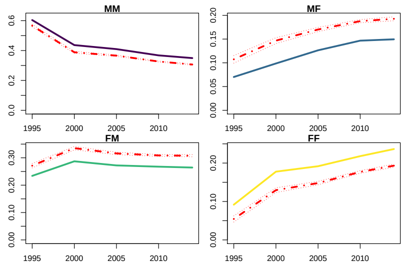
(c) Future Academics, no constraints



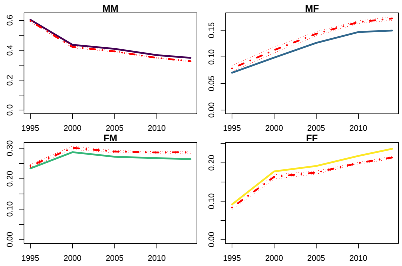
(d) Future Academics, University & Field



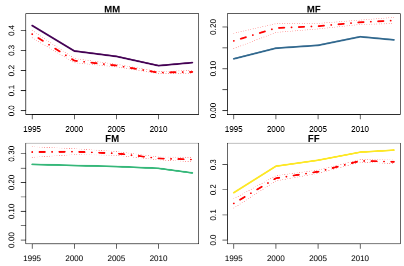
(e) SET only



(f) SET only; University & Field



(g) SSH only



(h) SSH only; University & Field

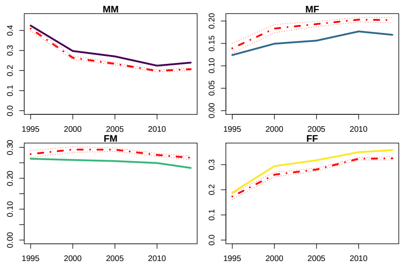
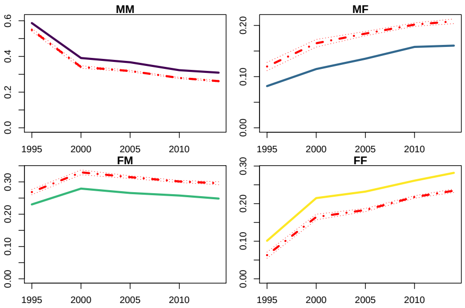
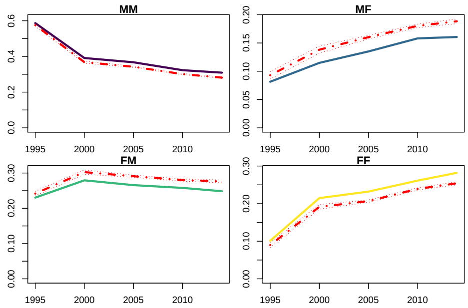


Figure 9: Permutation test for student supervisor data 1973-2014 for top institutions and Phd students. The permutation is done keeping fixed the number of ties and the supervisor and permuting the students. Permutation without constraint (left) and with institutional constraints (right). Permutation is repeated 100 times for each of time period. The Observed proportions of gender student-supervisor relations (solid lines) are plotted versus the results of the permutation test (dashed lines) with respective two standard deviation from the mean. As top institutions we include: Cape Town; Pretoria KwaZuluNatal; Stellenbosch; Rhodes; Witwatersrand; and Western Cape University;"

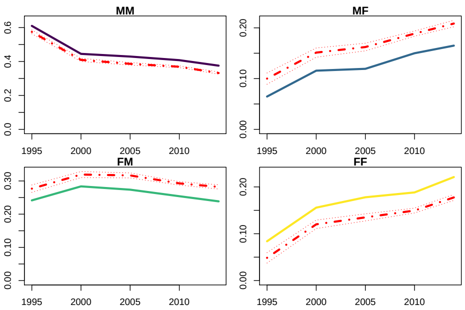
(a) Top Institutions



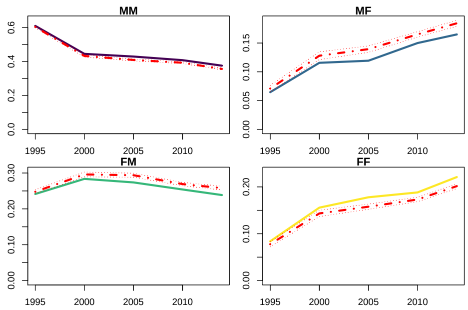
(b) Top Institutions; University & Field



(c) PhD Students only



(d) PhD Students only; University & Field



D.2 Choice and Induced Homophily

Figure 10: Induced and Choice Homophily comparison looking gender ties in student supervisor relations 1973-2014. We compare results of assortativity of different sub-samples: All vs. Future Academics (a), SET vs. SSH (b). Each panel contains 4 series; In Fig.(a): aggregated (upper series) versus department level (lower series) crossed with total sample (white background) versus future academics (grey background). And similarly for Fig. (b). Upper series are system-level permutations, including both induced and choice homophily; lower series are department-level permutations, excluding much of the induced homophily. Each box plot represents 100 permutations and associated assortativity calculation, as described in section 4.2, and equation 4.

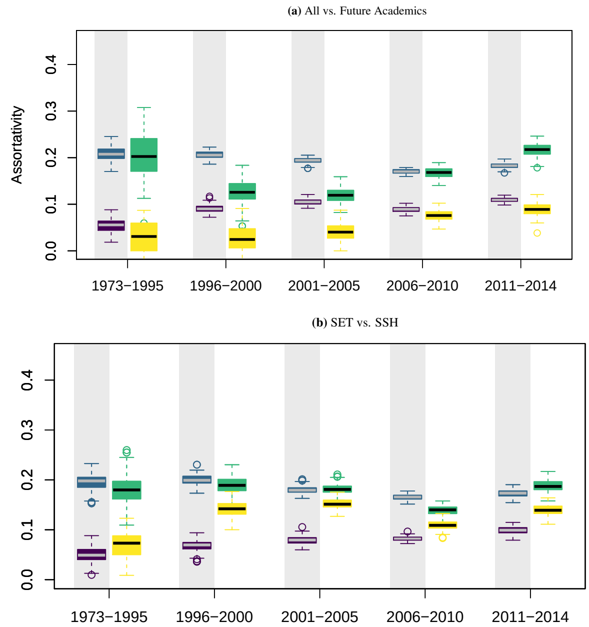
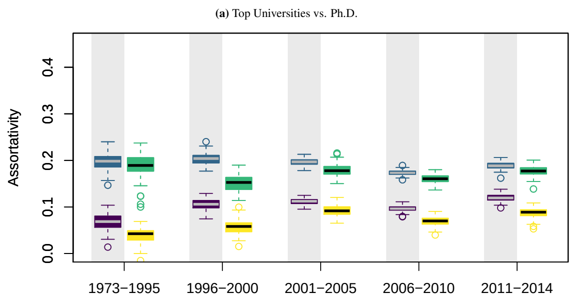


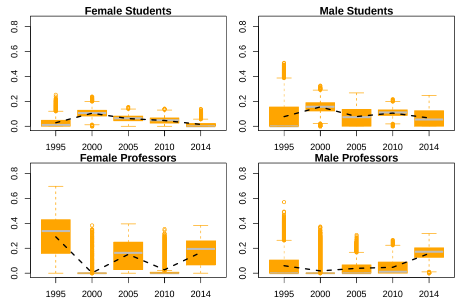
Figure 11: Induced and Choice Homophily comparison looking gender ties in student supervisor relations 1973-2014. We compare results of assortativity of two sub-samples: Top Universities vs. Ph.D. Sub-sample . The panel contains 4 series: aggregated (upper series) versus department level (lower series) crossed with Top Universities (white background) versus Ph.D. (grey background). Upper series are system-level permutations, including both induced and choice homophily; lower series are department-level permutations, excluding much of the induced homophily. Each box plot represents 100 permutations and associated assortativity calculation, as described in section 4.2, and equation 4.



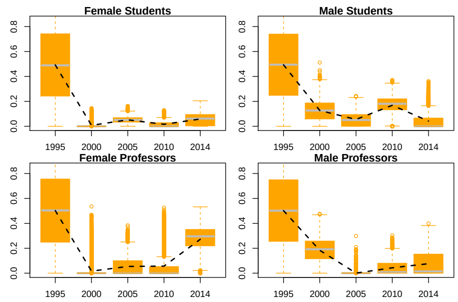
D.3 Choice Homophily of different groups

Figure 12: Gender Homophily of the different groups 1973-2014. The estimation is done minimising the average relative entropy of tie-types between observed and predicted from equations 6. We use data at department level under 10000 samples with replacement of department observations. Black dashed are averages. Faculties with less than 10 observations and without population variability are excluded.

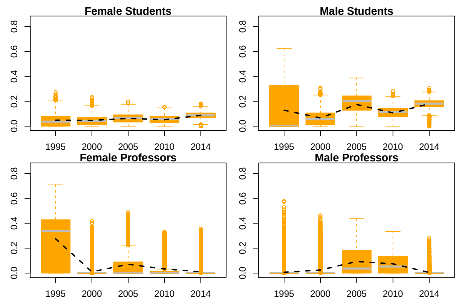
(a) All, Gender Homophily



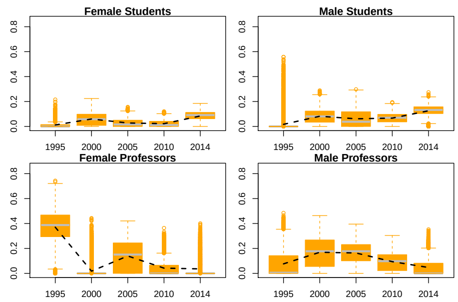
(b) Future Academics, Gender Homophily



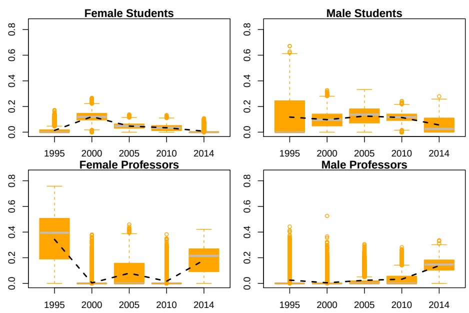
(c) Ph.D. Only, Gender Homophily



(d) Top Institutions, Gender Homophily



(e) SET only



(f) SSH only

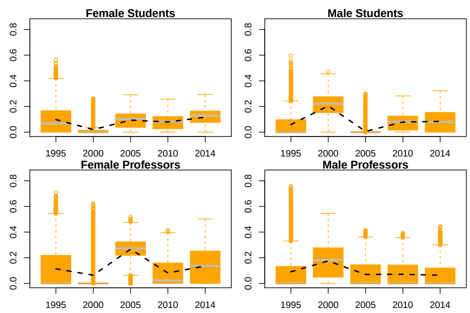


Table 5: Results of logistic regressions of yearly student-supervisor gender tie data. Dependent variable same-gender is 1 for same-gender ties and zero for cross-race ties. Models 3 and 4 are for the sub-sample of future scholars. Faculties with fewer than 10 observations and without population variability are excluded. Where variables: studrace (studgender) is a students' dummy 1 for white (male); race (gender) is a professors' dummy 1 for white (male). Age is professor's age; YearFirstRegistration is the students' registration year. Sf and Tf are proportions of female students and professors aggregated for each time period. Universities + Field are department dummies, and Rating is professors' NRF rating (7 rating categories).

	Dependent variable:			
	same-gender ties			
	(1)	(2)	(3)	(4)
studrace	0.058^{**}	0.058^{**}	0.114^{*}	0.112^{*}
	(0.026)	(0.026)	(0.065)	(0.065)
race	0.024	0.024	-0.141^{*}	-0.140^{*}
	(0.031)	(0.031)	(0.082)	(0.082)
studgender	0.893^{***}	0.893^{***}	1.140^{***}	1.141^{***}
	(0.024)	(0.024)	(0.060)	(0.060)
gender	-0.605^{***}	-0.604^{***}	-0.960^{***}	-0.959^{***}
	(0.026)	(0.026)	(0.063)	(0.063)
age	-0.002	-0.002	-0.014^{***}	-0.013^{***}
	(0.002)	(0.002)	(0.004)	(0.004)
YearFirstRegistration	-0.011	-0.016^{*}	0.006	-0.011
	(0.007)	(0.009)	(0.019)	(0.022)
Sf	1.440	-40.396	-0.712	-138.266
	(6.722)	(39.172)	(15.845)	(93.336)
Tf	0.636	-40.419	1.321	-131.649
	(2.603)	(37.963)	(6.234)	(89.135)
Sf:Tf		86.907		282.468
		(80.169)		(188.883)
Constant	20.846	51.648	-11.425	88.826
	(15.588)	(32.408)	(39.345)	(77.716)
Universities + Field	*yes*	*yes*	*yes*	*yes*
Rating	*yes*	*yes*	*yes*	*yes*
Observations	33,310	33,310	6,026	6,026
Log Likelihood	-21,569.580	-21,568.990	-3,752.535	-3,751.416
Akaike Inf. Crit.	43,325.160	43,325.980	7,691.070	7,690.832
Note:				*p<0.1; **p<0.05; ***p<0.01

The UNU-MERIT WORKING Paper Series

2019-01 *From "destructive creation" to "creative destruction": Rethinking Science, Technology and innovation in a global context* by Luc Soete
2019-02 *Do young innovative companies create more jobs? Evidence from Pakistani textile firms* by Waqar Wadho, Micheline Goedhuys and Azam Chaudhry
2019-03 *What gains and distributional implications result from trade liberalization?* by Maria Bas and Caroline Paunov
2019-04 *FDI, multinationals and structural change in developing countries* by André Pineli, Rajneesh Narula and Rene Belderbos
2019-05 *The race against the robots and the fallacy of the giant cheesecake: Immediate and imagined impacts of artificial intelligence* Wim Naudé
2019-06 *The middle-technology trap: The case of the automotive industry in Turkey* by Ibrahim Semih Akçomak and Serkan Bürken
2019-07 *The impact of a mathematics computer-assisted learning platform on students' mathematics test scores* by Marcelo Perera and Diego Aboal
2019-08 *Health insurance and self-employment transitions in Vietnam* by Nga Le, Wim Groot, Sonila M. Tomini and Florian Tomini
2019-09 *Knowledge economy and economic development in the Arab region* by Samia Mohamed Nour
2019-10 *Migration of higher education students from the North Africa region* by Samia Mohamed Nour
2019-11 *Job automation risk, economic structure and trade: a European perspective* by Neil Foster-McGregor, Önder Nomaler an Bart Verspagen
2019-12 *The breadth of preferential trade agreements and the margins of* exports by Rod Falvey and Neil Foster-McGregor
2019-13 *What a firm produces matters: diversification, coherence and performance of Indian manufacturing firms* by Giovanni Dosi, Nanditha Mathew and Emanuele Pugliese
2019-14 *Brazilian exporters and the rise of Global Value Chains: an empirical assessment* by Caio Torres Mazzi
2019-15 *How has globalisation affected the economic growth, structural change and poverty reduction linkages? Insights from international comparisons* by Aradhna Aggarwal
2019-16 *R&D, innovation and productivity* by Pierre Mohnen
2019-17 *Domestic intellectual property rights protection and exports: Accessing the credit channel* by Gideon Ndubuisi
2019-18 *The role of early-career university prestige stratification on the future academic performance of scholars* by Mario Gonzalez-Sauri and Giulia Rossello
2019-19 *The employment impact of product innovations in sub-Saharan Africa: Firm-level evidence* by Elvis Korku Avenyo, Maty Konte and Pierre Mohnen
2019-20 *Embodied and disembodied technological change: the sectoral patterns of job-creation and job-destruction* by G. Dosi, M. Piva, M. E. Virgillito and M. Vivarelli
2019-21 *Can we have growth when population is stagnant? Testing linear growth rate formulas and their cross-unit cointegration of non-scale endogenous growth models* by Thomas H.W. Ziesemer

2019-22 *Technical progress and structural change: a long-term view* by Alessandro Nuvolari and Emanuele Russo

2019-23 *No evidence of an oil curse: Natural resource abundance, capital formation and productivity* by Mueid al Raee, Denis Crombrughe and Jo Ritzen

2019-24 Far from random? The role of homophily in student supervision by Giulia Rossello and Robin Cowan