



Culture as configurations of categories: Analyzing peer effects via dual-to-regression modeling

Ronald L. Breiger^{a,*}, Alessandro Lomi^b, Francesca Pallotti^c

^a School of Sociology, University of Arizona, Social Sciences Bldg. 400, Tucson, Arizona 85721, USA

^b Faculty of Economics and Social Network Analysis Research Center, Università della Svizzera italiana, Via Buffi 13, 6900 Lugano, Switzerland

^c Department of Management, University of Exeter Business School, Rennes Dr, Exeter EX4 4PU, UK

ARTICLE INFO

Keywords:

Duality of cases and variables
Regression analysis
Peer effects
Homophily
Social network cohesion
Configurational analysis

ABSTRACT

In this paper we reimagine linear regression modeling as a relational method for cultural analysis. Drawing on the dual-to-regression analytic approach (Schoon, Melamed & Breiger, 2024), we argue that the fundamental building blocks in a regression equation are not single variables, but configurations of variables manifested by clusters of cases. In a study of peer effects and achievement in an academic institution, we show how the regression model itself may be understood as positing a network of pairwise influence relations among social actors that produces the outcome as modeled by the regression. Moreover, this network is appropriate for studying homophily (the tendency for individuals with similar characteristics to have social network connections). We push the new, case-oriented thinking about the regression model of Schoon et al. by incorporating information on networks of social relations connecting the cases. We find that, when profile similarity boosts academic performance, high-density social network clusters are discovered. We demonstrate that it is sometimes useful to consider configurations of cases as the “variables” in a regression model. We argue that this methodological innovation has a distinctive pragmatic value and strong theoretical motivation in the specific empirical context of our study, and beyond.

1. Introduction

Controversies around culture and methods for cultural analysis continue to swirl, in the process generating considerable energy and multiple lines of innovative and rewarding inquiry (Mohr & Rawlings, 2010; Rawlings & Childress, 2021). As John Mohr and co-authors have written, the way forward should involve a capacious dialogue in which students of culture think more together, “across our methodological commitments” (Mohr et al., 2020:274). In this paper we seek to think, in particular, across the methodological lines of relational sociology and linear regression models, to fashion an approach—one approach among many, to be sure—that can make distinctive contributions to advancing the sociology of culture.

Many problems that cultural sociology addresses are amenable to analysis via regression modeling. Andrew Perrin (2004:11) reviews several important examples, while arguing with verve that “half a century of studies using regression and similar techniques offers a rich historical context for evaluating and reevaluating culture as one among many families of variables in causal chains.” The

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author.

E-mail addresses: breiger@arizona.edu (R.L. Breiger), alessandro.lomi@usi.ch (A. Lomi), f.pallotti2@exeter.ac.uk (F. Pallotti).

<https://doi.org/10.1016/j.poetic.2025.102014>

Received 9 November 2023; Received in revised form 10 March 2025; Accepted 15 May 2025

Available online 28 May 2025

0304-422X/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

problem he sees is with a different tradition, that of relational methods for cultural analysis—clustering, metric scaling, Galois lattices, “the geometry of opinion constraint within networks,” and other techniques for analyzing culture as networks of context, such as work developed by Bearman and Stovel (2000), Breiger (2000), Martin (2000), Mische and Pattison (2000), and Mohr (1994), and often published in this journal (see also DiMaggio, 2020). Perrin articulates his concern “that by inventing new methods—or by adapting existing and obscure ones—we marginalize cultural analysis from the rest of sociology” (Perrin, 2004:11).

The usual critique of regression modeling in this discussion (Sonnett & Breiger, 2004; see also Tong, 2019) is that the modeling is disconnected from any goal or vision of cultural sociology. Within its array of variables (such as education, museum visits, liking for opera, race, gender) and cases (whether individuals, organizations, nations, or other types of actors), the cases are viewed as independent, disconnected from one another, and decontextualized. Independence and decontextualization of cases typically sustain an interpretation of culture and cultural identities in terms of the “attribute breakdowns of our everyday categories” (White, Boorman & Breiger, 1976:733), or as the outcome of “sociological cross-tabs” (Padgett, 2018: 406).

Indeed, regression modeling is the study of relations among the variables, and the cases most often are rendered invisible (Shalev, 2007; Rambotti & Breiger, 2020). In a fundamental sense, we could say that making the cases invisible through a set of analytically convenient assumptions is precisely what allows regression to work as a general strategy to learn from data.¹ But even though the variables are given theoretical precedence over the cases in regression modeling (much as, in the seventeenth century, Descartes decreed the precedence of mind over body), it is nigh on impossible for regression models to formulate multi-way interactions of variables (such as a subset of people in a population or sample who are, say, racialized as Latinx, gendered as women, live in south Texas, infrequently visit museums, and vote for Republican Party candidates). The importance of identifying complex interactions of variables such as this one is that context (in the form of intersecting variables) matters greatly in understanding behavior (see also Lee & Martin, 2015; Breiger & Wagner-Pacifi, 2024).

As an illustration of this general argument, consider the case of context-specific interactions among multiple variables that researchers have recurrently parametrized in empirical studies of cultural resources and individual attainment in education—the empirical setting of our study (see, for example, Farkas et al., 1990).

In this setting, influential studies have documented how the tendency of students to partition their environment into discrete local social clusters has consequential implications for relevant individual outcomes such as social integration and academic performance (Coleman, 1961). However defined, discovered, or labeled (Hoffman, Block & Snijders, 2023), clusters of students in which schooling environments are typically partitioned involve intersections between cases (the students) and variables (their attributes). It follows, therefore, that the coalescence of student subcultures into identifiable clusters can be understood only in terms of multiple combinations of variables associated with members in those clusters.

As Fine (2012:167) put it, “Groups are the crucible of cultural creation.” Educational settings are no exception. Membership in specific subcultural clusters, in which educational environments are partitioned and organized, has implications not only for individual academic performance, but also for patterns of social relations within and between members in specific subcultures (Boda, 2018). This observation provides one possible basis for Doehne, McFarland and Moody’s (2024:176) recent claim that the subcultural clusters we have discussed (which they call “social identities” and, occasionally, “peer groups”) “are jointly constructed by attributional features and relational configurations.” Doehne et al. observe, for example (p. 176), that Coleman’s original notion of a “leading crowd” in schools is a combination of “attributional features, for instance rich student athletes,” and “relational attributes such as high internal cohesion and strong admiration from others.” Furthermore, “[t]o the extent that settings admit strong identity divisions, distinct social identities can develop their own relational norms” (p. 176).

With this general discussion as background, in this paper we bring together relational methods and linear regression modeling. In an unusual departure from conventional thinking, we follow Schoon, Melamed and Breiger (2024) in showing how regression analysis can be reimagined as resting on foundations of social networks, field theory, and the duality of cases and variables (Breiger, 2009). We accomplish this for our study of peer effects on academic performance among a class of professional students enrolled in a two-year, full-time Master of Management program at a Swiss university. The class consisted of a highly international group of students with high gender and moderate age diversity and a variety of prior work experiences. In addition to our focus on understanding academic performance, we collected social network data (advice and friendship) and data on students who were connected because they shared aspirations to interview at the same companies. We make use of regression analysis in an unusual way, but one we think is productive of insight about peer effects (Angrist, 2014)—or “student-to-student” effects (Duflo, Dupas & Kremer, 2011)—equivalent terms adopted by researchers in the economics and sociology of education to suggest the detectable presence of non-individual components in outcomes typically understood and represented as individual.

Specifically, we show that the regression model itself may be understood in a new way: as resting on a network of inter-peer influence that directly produces a key model output (fitted values for the outcome variable). Then we identify configurations of intersecting attributes that strongly affect the attainment outcome but that are typically hidden inside regression coefficients. Further, we find that, for those subsets of students whose homophily (similarity) with respect to the intersection of multiple variables is associated with a boost in academic performance, high-density clusters are discovered in their social networks. We argue that studies of peer culture should transcend “general linear reality” (Abbott, 1988, 2001), not by ignoring it, but by exploiting its analytical properties in the new ways that we demonstrate in this paper.

¹ For a general summary discussion see, for example, Meuleman, Loosveldt, & Emonds (2015). For a historical overview of how violations of independence assumptions have stimulated the development of statistical models for networks see, for example, Wasserman & Pattison (1996) and Pattison & Robins (2002).

2. The tension between individual achievement and “fitting in”

The tension between “individual achievement” and “fitting in” is central to sociological research (Akerlof & Kranton, 2002; Goldberg et al., 2016). This tension also makes education a particularly promising setting for exploring the empirical value of our methodological proposal.

In educational organizations, individual achievement involves understanding—and complying with—an institutionalized system of rational evaluation norms, values, and rituals (Meyer & Rowan, 1977; Meyer, Rowan & Meyer, 1978). Fitting in involves searching for—and aligning with—recognizable social identities (Akerlof & Kranton, 2002; Calvó-Armengol, Patacchini & Zenou, 2009).

The tension between these concurrent families of social processes arises from the fact that individual achievement is designed to generate and sustain interpersonal differences made comparable through an institutionalized evaluation metric (Espeland & Stevens, 1998). “Fitting in,” on the other hand, elides the differences created by the processes of achievement and molds them into recognizable social forms. Examples of social forms in this sense include the members of the “leading crowd,” the “activities leader,” the “brilliant student,” and the “athletic star,” described by Coleman (1961). Examples from more recent literature are the “jock,” the “nerd,” and the “burnout” mentioned by Akerlof and Kranton (2002:1168), according to whom: “[a]ssociated with each category is an ideal in terms of physical attributes and behavior.” Coleman (1959; 1960, 1961) considered these social forms as the basic organizational elements of distinct and clearly bounded adolescent subcultures (“identities” in the phrasing of Akerlof and Kranton). Clearly, the meaning of these stylized social forms frequently encountered in educational research depends on the (sub)cultural context in which they are grounded.

2.1. Category as both a social-relational position and a cluster of attributes

In this view, a category is both a social “position” as well as a “cluster of attributes”—constructs (and concepts) that a major theoretical and analytical tradition has long held as antithetical. In the tradition of formal network modeling, perhaps most notably in early work on blockmodel analysis (White, Boorman & Breiger, 1976), “positions” consist of sets of social actors who have similar patterns of social ties (relational connection) to actors in their own and other identified sets. With respect to their social connections, actors within positions are in this sense structurally similar. “It is not the inherent attributes or properties of objects that define them, but rather their relational location within a field of relations,” Mohr and White (2008:489) wrote of this older formalist tradition, in the process of articulating an alternative. More recent developments in the study of categories have reframed networks and culture as mutually constitutive—each understandable only in terms of the other (Mohr, 1994; Mohr & White, 2008; Pachucki & Breiger, 2010). People’s personal networks tend to be homogeneous with regard to many attributes (sociodemographic, behavioral, and intrapersonal characteristics), as much research has shown (McPherson, Smith-Lovin & Cook, 2001). However, less research has been aimed at pinpointing particular intersections or exact mixtures of attributes that, in conjunction with interpersonal network relations, shape or constitute social positions.

2.2. Peer effects

Coleman (1960:339) was among the first to suggest the interdependence between membership in a particular subculture, academic performance, attributes of social actors, and personal and social networks. As he notes in his discussion of gender differences: “... the girls, although better students in every school, do not want to be considered ‘brilliant students.’ They have good reason not to, for the girl in each grade in each of the schools who was most often named as best student has fewer friends and is less often in the leading crowd than is the boy most often named as best student.”² In Coleman’s study academic performance results, in part, from positions or subcultures (membership in, or orientation to, the “leading crowd”) consisting simultaneously of social relations (including number of friends) and of combinations of attributes such as students who are athletes and also rich, in the language of Doehne et al. (2024:176) mentioned earlier.

More recent research has developed and considerably extended Coleman’s original insight that participation in specific subcultures, academic achievement, and social networks are related (Akerlof & Kranton, 2002). Under the general rubric of “peer-effects in education,” or “student-to-student spillovers” (Duflo, Dupas & Kremer, 2011), contemporary empirical research in the economics and sociology of schooling has documented specific aspects of this relation.³

Social relations frequently—although not always explicitly—provide the “raw material” that empirical studies use to construct and test hypotheses about peer effects in education (Angrist & Lang, 2004). In the typical empirical study, social relations are proxied by shared attributes such as family background, ethnicity, geographical location, or prior academic performance (Sacerdote, 2011). Occasionally, studies of peer-effects in education are explicitly relational—i.e., they are designed to examine information contained in social relations reported by students themselves. Lomi, Snijders, Steglich and Torló (2011), for example, show that social selection and social influence on individual performance outcomes co-evolve in a cohort of graduate students enrolled in an MBA program in Italy:

² Although contemporary scholarship on teen peer groups and friendship networks recognizes the nonbinary and fluid nature of gender (Ramirez, Wesche, & Kreager, 2024:219), such was not the case with respect to Coleman’s reliance on rigidly binary gender categories in his pioneering study of “adolescent subculture” in ten secondary schools in 1950s America (Coleman, 1959; 1960; 1961).

³ For a comprehensive recent review of the peer effects literature in education see Paloyo, 2020. For an influential statement see Sacerdote, 2011. For a broad methodological introduction to models and methods for studying peer effects see An, 2011.

students tend to assimilate the average performance of their friends and advisors. At the same time, students attaining similar levels of academic performance are more likely to develop friendship and advice ties.

Using a sample of almost 5000 Swedish adolescent students across 218 classrooms, Raabe et al. (2019) find evidence that students adjust their preferences for subjects to the preferences of their friends, and suggest that network mechanisms amplify gender differences in preferences for STEM subjects (Science, Technology, Engineering, Mathematics).

The empirical identification of peer-effects in education is fraught with methodological difficulties arising from the confounding joint effect of social network mechanisms of peer group formation (Goldsmith-Pinkham & Imbens, 2013; Lomi et al., 2011), and the tendency of individual outcomes to be highly correlated with group average outcomes (Angrist, 2014; Manski, 1993). In consequence, the results reported in the empirical literature are highly dependent on the method adopted to estimate the presence and magnitude of “peer effects” on the outcome of interest. Based on extant research, the impression that a general reader may form is that for every empirical study reporting evidence of peer effect on individual achievement (e.g., Hanushek et al., 2003) there seem to be at least another that reports opposite results (e.g., Arcidiacono & Nicholson, 2005). Evidence of “clean” peer effects is frequently reported in experimental studies (Falk & Ichino, 2006), but such results have proven difficult to replicate in observational settings (Manski, 1993; Shalizi & Thomas, 2011). General agreement is beginning to emerge on the conclusion that the possibility of finding evidence of peer effects is highly context-specific (Paloyo, 2020), due to the multiple ways in which unobservable mechanisms of homophily and social influence combine to produce observed clusters of attributes and outcomes, which are the measured variables used to define and identify “subcultures.” Recent empirical studies (e.g., Lorenz et al., 2020) and theoretical statements (Doehne, McFarland & Moody, 2024) are consistent with this conclusion.

3. Our approach

3.1. Conceptualization

In light of the conceptual and methodological issues just reviewed, our own approach embraces the context-specific nature of peer effects. Our approach rests on two postulates.

- 1) Heterogeneity in behavioral outcomes is organized in terms of “configurations of categories of interest.”
- 2) Peer effects matter because they induce heterogeneity in behavioral outcomes *between* configurations of categories, while leading to similarity *within* configurations.

Concerning the first postulate, heterogeneity in behavioral outcomes is in fact organized along many dimensions, but for the purposes of this paper we focus on the idea that intersecting attributes and categories of individuals—for example, boys in a religiously-oriented school who “went out for football this Fall;” girls in public schools in work-class suburbs who attended many “home basketball games” (Coleman, 1960:338; 1961: Question 11 of the appended “boys’ questionnaire” and Question 12 of the one for “girls”)—indeed organize behavioral outcomes in peer groups ranging from secondary schools to juvenile gangs to corporate offices. An empirical research problem on which we focus is how to identify the multi-category intersections that actually do organize behavior (while avoiding the myriad of complex intersections of attributes and categories that have no bearing on behavior).

Our second postulate states why (within the context of the present study) certain configurations of categories matter. The complex, multi-category intersections of attributes that we seek to identify by means of a search procedure are those configurations that lead to maximally similar behavior internally while exhibiting maximal distinctiveness externally. For example, we would like to identify configurations of students’ attributes that lead to high probabilities of social network friendship or advice ties among students who share a given configuration, and relatively lower probabilities of network choices between students who manifest different configurations.

3.2. Method: dual-to-regression modeling

These conceptual developments have been accompanied and reinforced by parallel progress on the methodological front. The “regression inside out” approach of Schoon, Melamed and Breiger (2024), more specifically, may be thought of as one possible attempt to reconcile in practice what theory frequently divides. In this modeling framework, combinations or interactions of variables reveal, and at the same time reflect, clusters of cases. In this way, regression analysis is re-functionalized to facilitate the discovery of (and hypothesis testing about) how statistical effects may differ across clusters of cases. Regression coefficients can now be interpreted as sums across clusters of cases—or network “positions,” as defined earlier—facilitating the recovery of information about the cases from regression estimates.

In this paper we push the new, case-oriented thinking about the regression model of Schoon et al. (2024) beyond its current ability to reveal the duality of clusters of cases and statistical interactions, by incorporating information on networks of social relations connecting the cases. We argue and show that this methodological innovation has distinctive pragmatic value and strong theoretical motivation in the specific empirical context of our study. Our focus in this paper is on recasting linear regression modeling as resting on a network of profile similarity, as well as investigating the association between this profile similarity network and social networks.

3.3. Setting

We are fortunate in being able to study peer effects on academic achievement in a context that affords us both attribute data and data on multiple social networks (see also [Parker, Pallotti & Lomi, 2022](#)). We collected data on 139 graduate students enrolled in a two-year, full-time Master in Management program at a public Swiss university in Ticino, the only canton (state) where Italian is the sole official language. Students typically enter the program immediately following completion of their undergraduate degree. The class we analyzed consisted of a highly international group of students with a diversity of academic backgrounds.

The data were collected during a first-year core course that requires full-time attendance and offers a practical and interactive approach to learning based on a variety of pedagogical tools such as role-based games and case-based studies. Students attending the course were assessed through a midterm exam and a final exam.

Approximately one week before the midterm exam, the students were invited to complete an on-line questionnaire designed to elicit information on their academic background, work experience, networking behavior, and career preferences. Students were fully informed about the nature and objectives of the study and reassured that their privacy would be protected. Participation was voluntary and accounted for as part of the class engagement score (5 % of the overall course grade). The quality of the information provided by participants was carefully monitored during the data collection process. Individual reminders were sent to students when inconsistencies, duplications or missing data were discovered in the questionnaires. The response rate was 100 % with no missing observations. One student dropped out of the program prior to the midterm exam and was excluded from analysis. The final sample included 138 students. The age range was 21–55 years ($M = 23.98$) with 67.4 % identifying as women. The students came from 32 countries and had varied academic backgrounds with 56.5 % having undergraduate degrees in the social sciences and management, 39.1 % in arts and humanities, and 2.9 % in sciences and engineering. One student had a background in computer science, and one in legal studies.

The first section of the questionnaire gathered demographic data, as well as data on the students' academic background, performance (i.e., bachelor exit grade), and work experience prior to joining the program. The second and third sections were designed to collect social network data through the well-established roster method ([Kilduff & Krackhardt, 2008:140](#)). [Table 1](#) briefly defines and describes the variables on which we will focus most of our analysis.

Operationalizing a measure of students' academic ability required special attention. We used a measure of students' performance in their Bachelor degree programs to assess the quality of past work that each student brought with them when commencing the Master in Management program that we study. However, since the students come from more than thirty different countries, finding a standardized measure of prior quality of academic work was problematic. In the end we opted to create a dummy indicator variable that takes the value of 1 if the student completed college with the maximum possible marks. This variable is labeled "BAQual" in [Table 1](#).

Building on prior research on social networks and academic achievement (e.g., [Baldwin et al., 1997](#); [Lomi et al., 2011](#)) we also considered two distinct relational contents: face-to-face study-related advice, and friendship via social media. For advice, we asked participants to "please indicate the names of your classmates to whom you would go for help and advice on course-related issues." The questionnaire specified examples of possible issues for which students might seek help and advice, such as asking for class notes, discussing course material that the students found unclear or hard to understand, or borrowing textbooks. Respondents were presented with a complete list of names in alphabetical order (by surname) and asked to check the box next to the classmates they went to for advice. For friendship relations, we asked respondents to "indicate the names of the classmates whom you have included as friends in Facebook (or other social networking platforms you use, e.g., Twitter, LinkedIn, Google +)." Respondents were once again presented with a complete list of names and asked to check the box next to the classmates included in their lists of friends on social media.

In addition to the two social networks, we replicated the design of [Snijders, Lomi and Torló \(2013\)](#) to obtain data on a two-mode network connecting these management students to companies. Our questionnaire included an item inviting students to select from a list the names of the companies at which they would consider applying for a job after graduation. Students were invited to think about companies they would like to work with, from which they would accept a job offer or to which they would send a curriculum vitae for evaluation. The list was provided by the Placement Office of the university and included companies that routinely hire students from Swiss universities. The companies were grouped into 7 categories corresponding to public and private sectors. Students could select as many companies as they deemed appropriate.

In our study, the individual behavioral outcome of interest is academic performance, as measured by the student's grade in the midterm exam, which was given very close in time to (just a week or so after) our collection of attribute, social network, and job interview preference data.

4. Analysis

4.1. OLS regression

We begin by modeling students' score on the midterm exam as a function of the variables described in [Table 1](#). The list of independent variables is somewhat ad hoc, but more-or-less reasonable with respect to our goal of exploring academic achievement in this Master in Management program. We use linear modeling (ordinary least squares, or OLS, multiple regression) with a quadratic term for students' age. The model estimates are given in [Table 2](#).

Although we report significance tests ("p-values") in [Table 2](#), we interpret them basically as descriptive measures of the effects of the respective variables on the outcome, rather than as inferences or hypothesis tests. This is because we use the data to help guide model choice, and therefore uncertainty cannot be precisely quantified ([Tong, 2019](#)), as elaborated in Appendix B.1 of the

Table 1

Variable definitions and descriptives.

Variable	Definition	Range	Mean/Proportion	s.d.
Midterm	Outcome variable: score on midterm	(5, 20)	14.64	3.20
Age	Age in years	(21, 55)	23.98	3.68
AgeSq	Quadratic term	(441, 3025)	588.40	251.24
WorkEx	Student has work experience	(0 or 1)	0.746	0.44
BAQual	Student completed college with the maximum possible grade	(0 or 1)	0.058	0.23
Female	Student self-identifies gender as female	(0 or 1)	0.674	0.47
EconMgmt	Student's prior academic background was in Economics or Management	(0 or 1)	0.478	0.50
Swiss	Student's citizenship is Swiss	(0 or 1)	0.225	0.42
EngLLang	Student's first language was English	(0 or 1)	0.036	0.19
ItalLang	Student's first language was Italian	(0 or 1)	0.514	0.50

Table 2

Regression coefficients, probabilities, fit statistics.

Variable	Metric Coefficients Estimate	Pr(> t)	
(Intercept)	31.7659	0.000	***
Age	−1.1553	0.011	*
AgeSq	0.0169	0.011	*
WorkEx	0.4764	0.462	
BAQual	1.7830	0.124	
Female	−0.6013	0.299	
EconMgmt	−0.7720	0.156	
Swiss	1.8285	0.007	**
EngLLang	0.0782	0.961	.
ItalLang	1.0421	0.066	■
R ² = 0.1613 (Adj R ² = 0.1023)			
F(9, 128) = 2.735 (p = .006)			

Note: Outcome (Y) = score on midterm exam.

N = 138, *** $p < .001$, ** $p < .01$, * $p < .05$, ■ $p < .10$.

Supplementary materials.⁴ While ours is therefore largely a descriptive study of peer effects on academic attainment in one context, “to mistakenly treat description as ‘routine’ is almost surely to botch the job” (Moses, 1992:117).

The observed average (mean) midterm score by (years of) age for these 138 students is reported in Fig. 1, along with the quadratic predicted effect of age on score (net of the other variables in Table 2). The sole outlier on age was 55 years old and earned a midterm score (19 out of 20 possible points) that was just inside the highest 10 % of scores.⁵ All other students had ages ranging from 21 to 38 and, in general, on average younger students tended to perform better on the test (Fig. 1). Net of other predictors, Swiss nationals on average had scores 1.8 points higher than students with citizenship in other countries (Table 2). Students whose first language was Italian scored about 1.0 point higher than those from other language backgrounds. Students who had completed their previous Bachelor studies with the maximum possible grade (“BAQual”) scored 1.8 points higher on the midterm (a net effect that is “not significant” according to the conventional test). From Table 2, other net effects that are “not significant” were for students who had academic backgrounds in economics or management (0.8 points lower), those who self-reported as women (0.6 points lower), and those who had work experience (0.5 points higher). The result that master in management students with a bachelor degree in business and economics tend to exhibit lower levels of performance (on average) may seem puzzling, but is not unusual; see, for example, Christensen, Nance and White (2012) for a similar result in the context of MBA programs, and Shaffer et al. (2016) for a comparable result in the context of biology education.

One potential explanation could be that the course from which the data were collected (i.e., Organizational Behavior) involves notions and ideas considered of peripheral interest by students with undergraduate business-oriented degrees, whose disciplinary backbone typically involves courses in corporate finance, accounting, business strategy, and perhaps marketing management. Students with undergraduate business degrees may be less prepared than students with a background in the social sciences (broadly construed) to benefit from graduate courses inspired by basic behavioral science principles. Of course, this is only a plausible conjecture—one, however, that the analytical framework we articulate in this paper can easily turn into a testable hypothesis.

This is a puzzle we will further explore with the new methods that we will soon present. But first we introduce our re-envisioning of the regression model itself, based on our discussion of peer effects.

⁴ In Appendix B of the Supplementary Materials we offer elaboration of several aspects of our methods and findings.

⁵ Of the 138 students, 14, or 10.1%, earned scores of 19 or 20 on the midterm (20 is the highest possible score). Fig. 1 reports the average (mean) midterm score for each year of age.

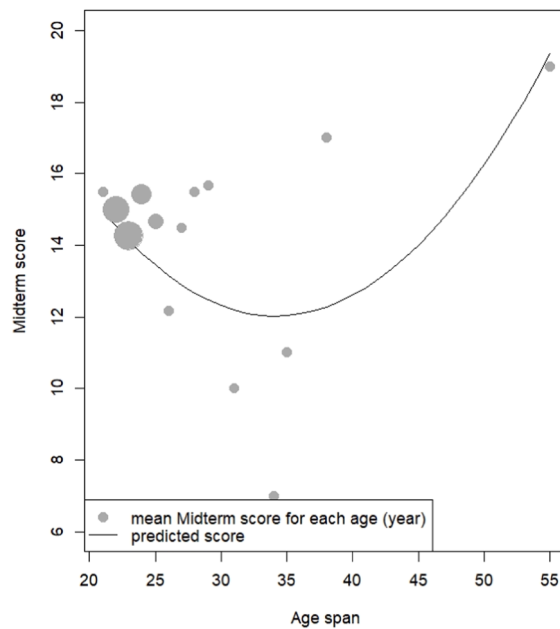


Fig. 1. Midterm score by student age: observed mean score by year (dots) and predicted score (line) from the regression model of Table 2. Larger dots represent more students of a given age (in years).

4.2. Fitted values in regression are produced by a network of interpersonal influence based on attribute similarity

Sharp distinctions are often drawn between social network and variable-centered approaches to social science inquiry, and (for the most part) rightly so. According to the classic regression model, each individual (or “case”) is a randomly selected representative of the population, and in this particular sense is entirely interchangeable with any other case (Schoon et al., 2024:4). As usually practiced, linear modeling thus tears individuals from their social contexts and becomes a “sociological meatgrinder” (Barton, 1968) in the view of many relational analysts. This and related features of regression modeling led prominent theorists to call for rejection of the construal of the social world as a “general liner reality” (Abbott, 1988:181). Variable-centered analysis is often seen as an example of “substantialist thinking” placed in opposition to the “relational” perspective (Emirbayer, 1997:288).

We now illustrate how the boundary separating linear regression and relational perspectives may be more porous than is usually imagined. We consider the vector of fitted values ($\hat{\mathbf{y}}$) produced by a linear, multiple regression model. For the model reported in Table 2,⁶ $\hat{\mathbf{y}}$ consists of the midterm score predicted for each of the 138 students; we will refer to their observed (actual) midterm scores as the vector $\mathbf{y}$. There is a square, symmetric matrix that projects $\mathbf{y}$ into $\hat{\mathbf{y}}$. Hence it is occasionally called the projection matrix, and it is also widely known as the influence matrix and as the Hat matrix (because it puts the “hat” on $\mathbf{y}$), denoted $\mathbf{H}$,

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \quad (1)$$

where the product is matrix multiplication. The above equation is well-known (Belsley et al., 2004:16). In our study, matrix $\mathbf{H}$ has 138 rows and the same number of columns, one for each student. A small submatrix of our Hat matrix is given in Fig. 2, which zooms in on four of our 138 students (with names disguised). Entries on the main diagonal of $\mathbf{H}$ (the number at the intersection of any person’s row and the same person’s column), denoted $\mathbf{H}_{ii}$, are widely used in regression diagnostics as measures of the influence of each respective individual on the fit of the regression model. The off-diagonal entries (numbers at the intersection of any person and any other person in Fig. 2) are almost always ignored in regression modeling (one exception being Gray & Ling, 1984). Each off-diagonal entry shows, for a pair of people, how much influence one person has on producing the predicted value ($\hat{Y}_i$) of another person (more generally, we replace “person” with “case”). In this sense, a key output of the regression model ($\hat{\mathbf{Y}}$) is a function of “peer effects” among the cases.

In our view an adequate model of peer effects (Section 2.2 above) would synthesize similarity in social network position and shared profiles across attributes and variables, showing the effect of such similarity in producing an outcome. Remarkably, at the core of every regression model is a network of peer effects. However, these peer effects are based entirely upon profile similarity (across values on the predictor variables, including attributes) for all pairs of cases, without reference to social relations. Because social network position is ignored, the Hat matrix is far from our ideal model of peer effects. However, it is the model of peer effects that is “built in” to every

⁶ We confine this discussion to multiple regression OLS modeling, but the results we discuss generalize to other forms of the General Linear Model (Schoon et al., 2024:ch 3).

$$\mathbf{H}\mathbf{y} = \hat{\mathbf{y}}$$

	Anton	Bernd	Chiara	Dario	[...]		
Anton	0.0740	0.0619	-0.0182	-0.0245	...	y_1	$\hat{y}_1$
Bernd	0.0619	0.0593	-0.0182	-0.0195	...	y_2	$\hat{y}_2$
Chiara	-0.0182	-0.0182	0.0338	0.0339	...	y_3	$\hat{y}_3$
Dario	-0.0245	-0.0195	0.0339	0.0365	...	y_4	$\hat{y}_4$
[...]	...	...	...	...	...	...	...

Fig. 2. Illustration of a Hat matrix, showing that $\mathbf{H}\mathbf{y} = \hat{\mathbf{y}}$.

regression model, and—as we will illustrate—analysts can benefit from recognizing this feature of the regression model.

The regression model constructs the vector of fitted values ($\hat{\mathbf{y}}$) so as to maximize their correlation (more precisely, the square of their correlation) with the observed outcome ($\mathbf{y}$). From eq. (1), we see that the Hat matrix ($\mathbf{H}$) is the adjacency matrix of a network among the cases that, when post-multiplied by $\mathbf{y}$, produces $\hat{\mathbf{y}}$ so as to maximize the squared correlation of $\mathbf{y}$ and $\hat{\mathbf{y}}$. Moreover, if $\mathbf{y}$ consisted of an entirely different outcome (for example, if we used the variables in Table 1 to predict some different outcome, perhaps the wages of these students' first jobs after graduation), matrix $\mathbf{H}$ would remain the same, because $\mathbf{H}$ depends only on relations among the predictor variables, and not on the outcome. In this sense—because it maximizes the squared correlation between the observed and the fitted values, no matter what values $\mathbf{y}$ consists of—we understand $\mathbf{H}$ as an “optimal” network among the cases (Schoon et al., 2024:212). Just as standard regression modeling finds the “best” linear combination of the predictor variables, the one that maximizes the squared correlation of $\mathbf{y}$ with $\hat{\mathbf{y}}$, from a dual perspective $\mathbf{H}$ is the network of dyadic influence among the cases that (see eq. (1)) produces the $\hat{\mathbf{y}}$ fitted values of the regression model. In brief: every time an analyst punches the “compute regression” button for OLS regression and many of its generalizations, the resulting fitted values ($\hat{\mathbf{y}}$) come from the peer effects network ($\mathbf{H}$) we have described.

4.3. The regression model as the basis of a model for homophily

We should also ask: can we use $\mathbf{H}$ to predict the ties in a social network? Theorists and analysts of homophily (McPherson, Smith-Lovin & Cook, 2001) find that social network ties often reflect similarity across multiple attributes. Recent work of McPherson and Smith (2019) is particularly generative, and particularly close to our approach in spirit, in that these researchers use many attitude and attribute variables from survey data to create profile similarity ties that they see as suggestive of social network ties. But how should those various attributes and variables be combined? As the adjacency matrix of a network that optimizes the (squared) correlation of $\mathbf{y}$ and $\hat{\mathbf{y}}$, the Hat matrix ($\mathbf{H}$) offers the optimal answer to this question, from the specific point of view of the regression model.

Fortunately, our dataset includes social network data on the same population of individuals (with names disguised) analyzed in the regression model of Table 2. Again following Schoon et al. (2024:47–48), the Hat matrix has an interesting decomposition that we illustrate in Fig. 3 for the example subset of students depicted in the previous figure. An entry on the diagonal of the Hat matrix (such as the number 0.0740 at the intersection of Anton and himself in the Hat matrix of Fig. 2) may be thought of as the squared distance from the origin of a point representing Anton in a multidimensional space of cases. Thus, the distance of Anton from the origin is 0.2721 (the square root of 0.0740, with a small bit of rounding error). Similarly, the distance of Chiara from the origin (the square root of 0.0338) is 0.1839 (with a bit of rounding error). When we remove these distances from matrix $\mathbf{H}$, what remains are similarities among individuals in the form of cosines that, just like correlations, vary between -1 and 1 . For example, the intersection of Anton and Chiara in matrix $\mathbf{H}$ (Fig. 2) is -0.0182 , and -0.0182 divided by $(0.2721 * 0.1839)$ is (with a bit of rounding error) -0.3645 , reported in Fig. 3 as the cosine (degree of similarity) between Anton and Chiara across all the variables in our regression study. We often use the matrix of cosines (illustrated in Fig. 3) as our network of profile similarity among the students.

In our illustrative example, we find that the first-listed two students are quite similar across all the attributes in our regression model (cosine above 0.90), as are the last-listed two students. However, these two sets are quite different from each other (exhibiting negative cosines, indicating dissimilarity in attributes).

We wish to emphasize that this network of relations is “built in” to the regression model. A main point of our paper is to interpret regression models in a new way: as resting upon a network among the cases. As we see it, regression analysis is therefore well-aligned with conceptions of culture (and cultural preferences more specifically) as embedded in networks of social relations (Lewis & Kaufman, 2018; Lizardo, 2006).

The networks referenced just above are networks of profile similarity among all pairs of cases, where profiles consist of values on multiple variables. The reader will recall from Section 3.3 that we have, in addition, two social networks among the students (advice and friendship), with data collected about a week prior to the midterm exam. How similar are the social networks to the (cosine) network of profile similarity that was introduced just above? For the Advice network among these 138 students, those pairs of students connected by a tie of advice (that is, either student reports seeking advice from the other, whether or not the tie is reciprocated) have an average (cosine) profile similarity (with respect to the variables in Table 1) of $+0.322$. For pairs of students who seek advice from each

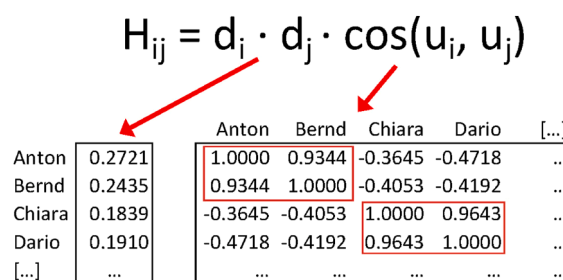


Fig. 3. Illustration of the decomposition of the Hat matrix of Fig. 2 into distances and cosines.

other (reciprocated ties), the profile similarity rises to +0.430, which is 0.92 standard deviations above the mean.⁷ For pairs of students not connected by advice (neither student reports seeking advice from the other), their average (cosine) profile similarity is lower, at +0.119. To summarize: across all pairs of these 138 students, greater homophily (across all the variables in Table 1) is associated with advice ties (+0.322) and, even more strongly, with reciprocated advice (+0.430), in comparison with no advice seeking (+0.119). Qualitatively similar results, not quite so strong, are found for the Friend network (Appendix B.2).

We also find a significant association between the profile similarity network and each of the social networks, as elaborated in Appendix B.3.

4.4. Cases as variables

We provide a visualization of the (cosine) similarity matrix in Fig. 4. Coordinates for nodes are given by eigenvectors of the cosine matrix, which may be thought of as a scaffolding for the network.⁸ Numerical values for cosines within the matrix are visualized as lines of varying thickness in Fig. 4 that connect pairs of students as nodes. Negative cosines are represented with red lines, and positive cosines with green lines. Notice from the Figure that red lines (reflecting negative similarity) tend to connect nodes at great distance from each other, whereas green lines (especially those that are thick, representing the highest similarity) tend to connect nodes closer to one another in the Figure. This suggests a clustering of the students based on the similarity of their attributes.

Identifying clusters from Fig. 4 seems a daunting task, however. For example, nodes in the upper-left quadrant of the Figure tend to have Italian as their first language (dark color) and to have academic backgrounds in Economics and Management (large size), but they mix Swiss and other citizens (red vs. blue color) and they mix genders (circles vs. triangles). In the upper-right quadrant we see another clump of students who have academic backgrounds in Economics and Management, but this clump tends to be not Swiss and not first-language Italian speakers (i.e., pink color), and to exhibit sub-clustering with respect to gender. For both quadrants, the two attributes language and academic background seem to sustain "fitting in" processes as their combination identifies distinct students' subcultures, or social identities (i.e., "the Italians", "the management students") coalescing into observable clusters of students. Similarly intricate complexity characterizes the other two quadrants as well.

Based on Fig. 4, we decided to divide each dimension (eigenvector) according to whether its sign was positive or negative. The cross-cutting of these binarized dimensions gave us four sets of students. We refer to each set as a "quadrant." In order to characterize each quadrant by a combination of variables, we made use of a core insight of Schoon et al. (2024; see also Rambotti & Breiger, 2020). Regression coefficients are functions of the variables in the regression equation. However, the insight of the above-cited researchers is that the same regression coefficients are sums across cases, and (therefore) also sums across sets of cases (if the sets partition the cases). Using procedures detailed in Schoon et al. (2024:25–30), we may write each regression coefficient (Table 2) as a sum across the four sets (quadrants) of students described above in this paragraph. This gives us the array in Table 3. Each numerical entry in this table is what Schoon et al. call an "intensity" or (synonymously) a "contribution" of a particular one of our four quadrants to the regression coefficient for a particular variable: note that the sum across any row of Table 3 gives the regression coefficient for that variable (Table 2). Thus, this (or any other) partition of the cases yields the exact same regression coefficients as those produced by standard analyses. The value-added of our non-standard approach is that we can relate variables to the cases (and sets of cases) that comprise them, thus revealing what regression models typically hide. (See also Appendix B.4.)

We wanted to characterize each cluster by the variables that seem salient to it. In our inductive analysis, we explored several different ways of associating variables with clusters, based on the intensities in Table 3, where our final choices are shown in bold. Ignoring the intercept, we associated the highest "intensities" in any column of Table 3 with the respective variable being present (or, in the case of age, a student who was older than average), and conversely for the lowest values, which we associated with the absence of that variable. We also associated having work experience with cluster C because it took on a higher magnitude in that cluster than in any other. We made one other adjustment. Our previous discussion of Fig. 1 revealed one student who was an outlier with respect to

⁷ For all $(138 \times 137)/2 = 9,453$ pairs of students, the mean profile similarity is +.133, with a standard deviation of .324.

⁸ The eigenvector corresponding to the highest eigenvalue of our matrix reflects the centrality of nodes, whereas the second emphasizes clustering (see also Susca et al., 2021). Therefore, we use the eigenvectors corresponding to the second- and third-highest eigenvalues for the coordinates of nodes (students) in Fig. 4.

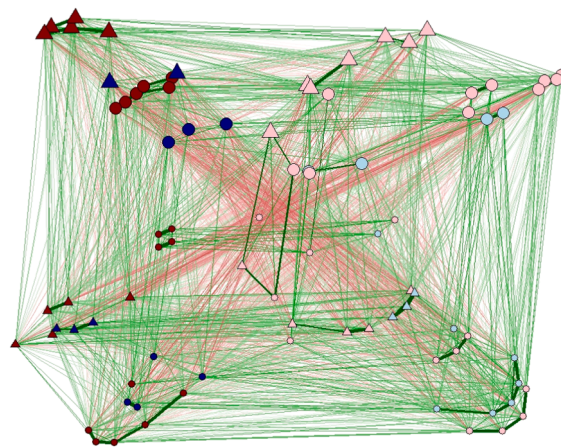


Fig. 4. A visualization of the network of (cosine) profile similarity. Cosines greater than 0 drawn with green lines; negative cosines, with red lines. *Note:* With reference to several of the variables in Table 1: dark blue nodes are students who have Swiss citizenship and whose first language was Italian; light blue nodes are those who have Swiss citizenship and whose first language was other than Italian; dark red nodes are students who are not Swiss citizens but whose first language was Italian; pink nodes are those who are not Swiss and have a first language other than Italian; large nodes are students whose academic background is in economics or management; small nodes are those with academic background in other fields; circles are women; triangles are men.

Table 3

Contribution to regression coefficients, by quadrant.

Variable	Quadrat				Sum == regression coefficient
	A	B	C	D	
(intercept)	−7.4345	20.8024	14.5552	3.8428	31.7659
Age	0.6914	−0.5474	−1.2002	−0.0990	−1.1553
AgeSq	−0.0101	0.0043	0.0170	0.0057	0.0169
WorkEx	−0.0713	0.2875	0.7780	−0.5178	0.4764
BAQual	0.4737	0.2197	−0.1767	1.2663	1.7830
Female	0.2613	−0.2140	−0.6907	0.0420	−0.6013
EconMgmt	6.2022	− 6.5094	7.9900	− 8.4548	−0.7720
Swiss	− 0.8767	2.7721	1.7113	− 1.7782	1.8285
EngLang	0.5199	− 4.0353	−0.2527	3.8463	0.0782
ItalLang	− 6.0510	− 7.9508	6.9918	8.0522	1.0421

Note: The sums in the final column are the regression coefficients in Table 2.

age, and we placed that student in a fifth set by herself.

Each column of Table 3 identifies a set of students, one for each quadrant of Fig. 4. We identify “configurations of cases” that are centrally representative of each (respective) quadrant, consisting of those students at the *intersection* of the categories most distinctively associated with that quadrant, based on the bolded entries in Table 3. For example, for column A of Table 3 we define a subset of our 138 students who are old (an age of at least 23, which is the median age), report their gender as female, have an undergraduate academic background in Economics or Management, are *not* Swiss citizens (their “intensity” in column A is a negative number), *and* do

Table 4

Regression coefficients with configurations of cases as the variables.

Config. Name	(N)	Configuration	Metric coefficients		
			Estimate	Pr(> t)	
C0	(59)	(Intercept)	13.9831	0.000	***
C1	(14)	EconMgmt; ~ItalLang; ~Swiss; Female; Older	−2.2688	0.010	*
C2	(21)	EconMgmt; ItalLang; WorkEx	1.3979	0.062	■
C3	(29)	~EconMgmt; ItalLang; ~Swiss	1.7066	0.011	*
C4	(14)	~EconMgmt; ~ItalLang; Swiss; ~EngLang	2.8027	0.002	**
C5	(1)	Age=55	5.0169	0.091	■
	(138)				
$R^2 = 0.1945$ (Adj $R^2 = 0.1640$)					
$F_{(5, 132)} = 5.819$ ($p < .001$)					

Note: $N = 138$, *** $p < .001$, ** $p < .01$, * $p < .05$, ■ $p < .10$.

not have Italian as their first language. We refer to this complex intersection as a “configuration” of cases. We proceed similarly with the other columns of Table 3, but we pull out the one outlier on age as a separate one-person configuration. Furthermore, the clusters so identified do not include all 138 students. We group together the 59 students who belong to none of our emphasized configurations and assign them to a set of their own.

As a counterpart to the regression model we discussed previously (Table 2), we then computed a second regression model, where “variables” were replaced by the configurations of cases whose construction we described above. The results are reported in Table 4.

A remarkable feature of the regression model of Table 4 is that, using these configurations of cases as predictors, our model accounts for more variance than when we used the variables. The adjusted R^2 for the model of Table 2 (i.e., the proportion of variance accounted for by the regression, taking into account the degrees of freedom expended) was 0.1023, whereas for the model of Table 4 it is 0.1640, which is an improvement of 60 %.

Because we identified the clusters of Table 4 from our inductive examination of the patterning (in Table 3) of contributions to the regression coefficients, the question naturally arises as to whether the improvement of the Table 4 regression model reflects overfitting. How robust are our discovered clusters to sampling variability? We explore this question in the Supplementary Materials (Appendix A). Our answer, with appropriate qualification, is that, for the most appropriate statistic for non-nested model selection, overfitting can likely be ruled out.

Nonetheless, our intention in this paper is not to present “a method” for finding clusters, or to add to the vast and sprawling existing methods literatures on clustering, factor analysis, community detection, and related inductive pattern-finding procedures. We readily admit that many such methods exist, some of which may be applicable (Fortunato, 2010). Instead, our focus in this paper is on how the regression model itself is constituted so as to advance cultural analysis. Table 4 reports a demonstration of something that is possible, namely: using the decomposition of regression coefficients (Table 3) to identify complex configurations of categories that can provide parsimonious predictors of an outcome variables such as (in our study) performance on a midterm exam.

The regression model of Table 4 helps us to understand a “puzzle” that we presented at the end of Section 4.1, namely: since our outcome variable is a midterm exam in a Master of Management program, why did the students whose undergraduate academic background was in Economics or Management perform somewhat below average, as seen in our earlier regression model coefficient (Table 2)? We see from Table 4 that the 21 students with an Economics or Management background *and* who had work experience *and* whose first language was Italian (i.e., Cluster C2 in Table 4) scored an average of 1.4 points higher than students in our reference cluster (C0). However, the 14 students with an Economics or Management background who were older *and* reported their gender as female *and* were citizens of countries other than Switzerland *and* did not have Italian as their first language (Cluster C1) scored on average 2.3 points lower than students in C0. Unlike the usual assumption that is made—that each variable is sufficient unto itself as a building block (predictor) in a regression equation—our results are consistent with an alternate assumption, namely, that each single variable needs to be understood in the context of the other variables and categories with which it combines. Building blocks are not single variables, in this alternative view, but configurations of variables manifested by clusters of cases.

4.5. Peer effects on midterm scores

As motivated in Section 4.2 above, the Hat matrix allows us to identify the influence of each person on every other person’s outcome ($\hat{Y}$). Implementing eq. (1) leads to Table 5. Specifically, we multiply each student’s row in the Hat matrix by the vector of midterm scores. Then (by eq. (1)) each resulting row sums to a given student’s predicted midterm score from the regression model of Table 2. For the four students illustrated in Fig. 2, Table 5 aggregates each student’s row (as described above) by the clusters identified in Table 4. For example, more than half (9.8 points) of Anton’s predicted Midterm score of 16.4 points (see the first row of Table 5) is due to the influence of other Cluster 4 members on Anton. This result tends to hold across the 138 students: the largest share of their predicted midterm scores come from fellow members of their own cluster. (We note that Anton and Bernd are in cluster C4, while Chiara and Dario are in C2.) We explore this point further in Appendix B.5.

Any of the students’ “self” scores in Table 5 increases to the extent that the student has a distinctive set of attributes (see the d_i

Table 5

Peer effects on achievement, for four students.

name	self or sum within Cluster							Sum = Fitted Midterm Score
	self	C0	C1	C2	C3	C4	C5	
Anton	0.8885	8.4241	−1.0532	−0.6065	−1.2109	9.8260	0.1733	16.4412
Bernd	0.8303	9.5421	−0.8452	−1.4054	−1.1426	8.8754	−0.1013	15.7534
Chiara	0.6428	1.7779	2.2788	8.3722	3.6988	−2.3896	−0.0993	14.2816
Dario	0.5109	2.3618	2.3918	8.0884	3.7366	−2.9087	−0.2601	13.9207

Note: Each row of the Hat matrix (size 138 rows by 138 columns) was multiplied by the outcome variable (Y), with respect to the regression model of Table 2. From all 138 rows, the four rows for the students shown in Figs. 2 and 3 were selected for illustration. (Anton and Bernd are in Cluster C4, and Chiara and Dario are in Cluster C2, of Table 4.) Shown here are the contributions to the fitted midterm scores ($\hat{Y}$) for each student himself or herself (column “self”) and sums of contributions to each student’s fitted midterm score from Clusters C0 through C5 (corresponding to the six clusters identified in Table 4). The sum of each row is that student’s predicted score (Y-hat) from the regression model of Table 2. This table illustrates a novel application of Eq. (1).

distances in Fig. 3). If each of the 138 students contributed equally to Anton's predicted midterm score, each student would add just a fraction of a point, $0.1191 (= 16.4412 / 138)$, to Anton's predicted score. His "self" value (0.8885) is >7 times higher than the expected value of this crude null model. Broadly similar results are found for the other students. Anton's "self" value (0.8885) is also higher than the mean contribution of the 13 other Cluster 4 students to his score, which is $0.7558 (= 9.8260 / 13)$. Mean contributions by cluster, as well as the network of mean influence among the clusters, are reported and discussed in Appendix B.5 (Supplementary Materials).

4.6. When profile similarity boosts performance, high-density social network clusters are discovered

No information about students' networks of social relations was used in constructing the configurations that were put to work in Table 4. Do these configurations of cases identify sets of social peers? And if so, do these sets of peers affect academic performance?

Table 6 adds to Table 4 the social network density—on both Advice and Friend networks—for each configuration identified in Table 4. We see that, for the Advice network overall, the average density is 4.6 % (meaning that the number of advice-seeking ties in this network is 4.6 % of the number of ordered pairs among these 138 students, which is $138 \times 137 = 18,906$ pairs in all). The 59 students who do not belong to any of our configurations of interest have a density of advice-seeking ties (4.1 %) that is just about the overall average. In this sense, these 59 students do not seem to be cohesive.

The students in the configuration that performs less well than average on the midterm test (with scores on average 2.27 points lower than the first-mentioned group of 59) are only slightly more cohesive than average, with an internal density of 8.8 %. However, the sets of students who perform *better* are much more cohesive than average. The configurations of students who score (respectively) 1.40, 1.71, and 2.80 points above those in the first-mentioned group have internal densities in the advice-seeking network of 13.6 %, 15.7 %, and 24.7 % (respectively), which are substantially above the overall density of 4.4 %. When we turn to the network of friendship ties (in the final column of Table 6), the results are similar in that the 59 students who are not members of any of our configurations of interest are not a cohesive group (their density is 9.7 %, about as high as the overall network density of 9.2 %), those who perform more poorly than average on the midterm are not very cohesive (with a friendship density of 11 %), but those configurations of students who score highest on the exam are considerably more cohesive in their friendship choices (with internal densities ranging from 19.2 % to 26.9 %). In brief: when profile similarity boosts academic achievement, high-density social network clusters are discovered.

Additional findings in the same direction—when profile similarity boosts performance, high density social network clusters are discovered—are detailed in Appendix B.6 with respect to strong (reciprocated) vs. weak ties. The same Appendix presents further findings along these lines with respect to data on shared aspirations among these management students to interview at the same companies. Here too, higher attainment on the midterm is associated with a higher level of shared career aspirations.

The findings here and in the Appendix relate to a main point of our paper, namely: the regression model (and in particular, the Hat and cosine matrices illustrated in Figs. 2 and 3, the decomposition of regression coefficients in Table 3, and the decomposition of fitted midterm scores in Table 5) can serve as a basis for the sociological study of homophily in ways that have not been heretofore explored. The new insight follows from our emphasis on the regression model itself as resting on a duality of cases and variables.

We found that homophily (increased density of social network ties within clusters of students who have relatively high-profile similarity on shared attributes) is stronger in the advice-seeking network than for ties in the friendship network. We should point out that both the model specification that we use to define homophily (the regression model of Table 2), and the question wording used to construct the social networks strongly shape the degree to which we find homophily to be associated with observed network ties.

5. Conclusion: let the cases be seen!

Heterogeneity in behavioral outcomes is organized in terms of "configurations of categories of interest." Peer effects matter because they amplify heterogeneity in behavioral outcomes *between* configurations of categories, while leading to similarity *within* configurations. Beginning with these postulates, we asked how to reduce the myriad number of potential (and actual) intersections of multiple categories and attributes to a parsimonious set that organizes a behavioral outcome: academic performance on an examination. In the process, we hope to have weakened the boundary between linear regression analysis and relational approaches such as those based on

Table 6
When profile similarity boosts performance, high-density social network clusters are discovered.

Config. Name	(N)	Configuration	coef.	Social Network Density	
				Advice	Friend
C0	(59)	(Intercept)	13.9831	4.1 %	9.7 %
C1	(14)	EconMgmt; ~ItallLang; ~Swiss; Female; Older	-2.2688	8.8 %	11.0 %
C2	(21)	EconMgmt; ItallLang; WorkEx; ~BAQual	1.3979	13.6 %	26.9 %
C3	(29)	~EconMgmt; ItallLang; ~Swiss	1.7066	15.7 %	22.9 %
C4	(14)	~EconMgmt; ~ItallLang; Swiss; ~EngLang	2.8027	24.7 %	19.2 %
C5	(1)	Age=55	5.0169	—	—
	(138)	Overall social network density		4.6 %	9.2 %

$R^2 = 0.1945$ (Adj $R^2 = 0.1640$)
 $F_{(5, 132)} = 5.819$ ($p < .001$)

networks and fields (Martin, 2003), while demonstrating the relevance of the dual-to-regression approach of Schoon et al. (2024) for identifying configurations of interest in a specific context (subcultures in an academic institution) that we believe has wider applicability for social and culture analysis.

We think that the analytical approach that we developed and illustrated in this paper holds considerable promise to advance our sociological understanding of culture across different and, occasionally, rival theoretical visions. One such vision locates the origin of culture as a specific aspect of social order in social interaction within small groups of the kind we have examined (Fine, 1979). Crucial to the articulation of this “local sociology” perspective on culture (Fine, 2012) is the identification of the actual boundaries around small groups—and the implications of such boundaries. This is particularly important because “sociologists may study groups without defining their research as being about groups” (Fine, 2012:159).

In this paper we have shown that the dual-to-regression modeling approach of Schoon et al. (2024) turns this theoretical problem into a wealth of concrete analytical possibilities that support the identification of meaningful clusters of dependence relations inherent in any sociological definition of “group” (Freeman, 1992). But group identification per se was not the only, or even the main purpose of our empirical examination. Our focus in this paper was on how membership in subgroups produces differences in individual outcomes of interest—and how the organization of these differences is affected by ways in which group members partition their social environment.

A second theoretical vision traces the roots of culture back to jointly occupied network positions (Anheier et al., 1995; Faulkner, 1983; Podolny & Baron, 1997:690). This more structural interpretation typically takes the notion of field (rather than group) as a primitive concept to explain culture within—and cultural variation between—network positions (DiMaggio, 2011). In this perspective, positions identified in the network structure of fields “induce shared subjectivities or ‘culture’ as we call it” (Martin, 2003:42).

We note that despite their appreciable differences, both theoretical visions posit that “people use culture to make connections with one another” (DiMaggio, 1987:442). Building on this general understanding of culture conversion (Lizardo, 2006), in the present study we have framed variations in individual academic performance as an outcome of the interaction between cultural and social forms of capital (Lizardo, 2024). By revealing a clear link between our definitions of culture (as configuration of variables) and social relations (as configurations of network nodes and edges), the models we have discussed may be usefully adopted to forward theoretical visions making very different ontological commitments to explain cultural phenomena. All we did, in a sense, was to let the cases previously masked by combinations of variables be seen.

In summary and conclusion: Regression modeling rests on a network of influence (the Hat matrix), which in turn rests on a network of profile similarity among the cases (the cosine matrix), that is, the extent to which each pair of cases is similar with respect to all the predictor variables in the regression equation. We have expanded the approach of Schoon et al. (2024) by relating this network to homophily with respect to social network cohesion. We have found that, when pairwise profile similarity boosts academic performance, high-density social network clusters are discovered. Moreover, we found this result while examining two social networks (advice-seeking and friendship), the strength of ties in each network, and a third network of shared aspirations (namely, for each pair of students, the shared number of companies at which both members of the pair would consider applying for a job after graduation; Appendix B.6).

To the extent that peer groups sustain resonant social identities, i.e., identities that “capture or activate powerful distinctions” (Baron, 2004:11), group membership will tend to increase homogeneity within and—in consequence—heterogeneity between peer groups. Over time, these coupled processes of heterogeneity reduction and amplification produce the polarization that is so frequently observed in empirical studies of on-line communities, political blogs, and social media more generally (Adamic & Glance, 2005; Dvir-Gvirman, 2017; Santos et al., 2021).

When measures of average academic performance vary significantly across peer groups (as in the academic context we examined), a conjecture suggested by our study is that individual differences in academic performance are likely to be accentuated by peer-group membership—a prediction consistent with results reported in early studies of the relation between student subcultures and academic achievement (Coleman, 1960).

Unlike the usual assumption that is made in regression modeling—that each variable is sufficient unto itself as a building block (predictor) in a regression equation—our results are consistent with an alternate assumption, namely, that each single variable can only be understood in the context of the other variables and categories with which it combines. Building blocks are not single variables, in this alternate view that we and colleagues are developing, but configurations of variables manifested by clusters of cases. We suggest that it is sometimes useful to consider configurations of cases as the “variables” in a regression model.

In the first section of this paper, we discussed Anrew Perrin’s (2004) essay in the newsletter of a national association’s Sociology of Culture section. Titled “Who’s Afraid of General Linear Regression?”, the essay argued that sociologists of culture should privilege standard techniques of quantification (“linear regression and its cousins”) over formal modeling of culture as relational. “Our colleagues know how to interpret regression coefficients, but they are (perhaps correctly) highly suspicious of scaling methods that ask them to interpret a ‘cluster’ of points on a graph” (Perrin, 2004:11). Now at the end of our paper we endorse the use of regression models in cultural analysis—but we hope to have demonstrated the added insight that comes from reimagining regression models themselves as relational at their core.

As most often practiced, regression modeling makes the cases invisible. We have argued and shown that this is unnecessary, because regression analysis at its core is about a network of profile similarity among the cases, and about dual relations between cases and variables. Rather than (as in common practice) using the cases to learn about the variables, we want as well to use the variables to learn about the cases, and thus to let the cases be seen.

CRediT authorship contribution statement

Ronald L. Breiger: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Alessandro Lomi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Francesca Pallotti:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.poetic.2025.102014](https://doi.org/10.1016/j.poetic.2025.102014).

References

- Abbott, A. (1988). Transcending general linear reality. *Sociological Theory*, 6(2), 169–186. <http://www.jstor.org/stable/202114>.
- Abbott, A. (2001). *Time matters : On theory and method*. University of Chicago Press.
- Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 36–43). <https://doi.org/10.1145/1134271.1134277>
- Akerlof, G. A., & Kranton, R. E. (2002). Identity and schooling: Some lessons for the economics of education. *Journal of Economic Literature*, 40(4), 1167–1201. <https://doi.org/10.1257/002205102762203585>
- An, W. (2011). Models and methods to identify peer effects. In J. Scott, & P. J. Carrington (Eds.), *SAGE handbook of social network analysis* (1st ed., pp. 514–532). London: SAGE Publications Ltd. <https://doi.org/10.4135/9781446294413.n34>.
- Angrist, J. D. (2014). The perils of peer effects. *Labour Economics*, 30, 98–108. <https://doi.org/10.1016/j.labeco.2014.05.008>
- Angrist, J. D., & Lang, K. (2004). Does school integration generate peer effects? Evidence from Boston's Metco Program. *American Economic Review*, 94(5), 1613–1634. <https://doi.org/10.1257/0002828043052169>
- Anheier, H. K., Gerhards, J., & Romo, F. P. (1995). Forms of capital and social structure in cultural fields: Examining Bourdieu's social topography. *The American Journal of Sociology*, 100(4), 859–903. <http://www.jstor.org/stable/2782154>.
- Arcidiacono, P., & Nicholson, S. (2005). Peer effects in medical school. *Journal of Public Economics*, 89(2–3), 327–350. <https://doi.org/10.1016/j.jpubeco.2003.10.006>
- Baldwin, T. T., Bedell, M. D., & Johnson, J. L. (1997). The social fabric of a team-based M.B.A. program: Network effects on student satisfaction and performance. *Academy of Management Journal*, 40(6), 1369–1397. <https://doi.org/10.5465/257037>
- Baron, J. N. (2004). Employing identities in organizational ecology. *Industrial and Corporate Change*, 13(1), 3–32. <https://doi.org/10.1093/icc/13.1.3>
- Barton, A. E. (1968). Bringing society back in: Survey research and macro-methodology. *American Behavioral Scientist* (Philadelphia, Pa.) 1, 12(2), 1–9.
- Bearman, P. S., & Stovel, K. (2000). Becoming a Nazi: A model for narrative networks. *Poetics (Hague, Netherlands)*, 27(2–3), 69. [https://doi.org/10.1016/S0304-422X\(99\)00022-4](https://doi.org/10.1016/S0304-422X(99)00022-4)
- Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.
- Boda, Z. (2011). Social influence on observed race. *Sociological Science*, 5(3), 29–57. <https://doi.org/10.15195/v5.a3>
- Breiger, R. L. (2000). A tool kit for practice theory. *Poetics (Hague, Netherlands)*, 27(2–3), 91–115. [https://doi.org/10.1016/S0304-422X\(99\)00026-1](https://doi.org/10.1016/S0304-422X(99)00026-1)
- Breiger, R. L. (2009). On the duality of cases and variables. In D. Byrne, & C. C. Ragin (Eds.), *The sage handbook of case-based methods* (pp. 243–259). Sage. <https://doi.org/10.4135/9781446249413>.
- Breiger, R. L., & Wagner-Pacifi, R. (2024). Social networks and social categories. In J. McLevey, J. Scott, & P. J. Carrington (Eds.), *SAGE handbook of social network analysis* (2nd ed., pp. 32–43). SAGE Publications Ltd.
- Calvó-Armengol, A., Patacchini, E., & Zenou, Y. (2009). Peer effects and social networks in education. *Review of Economic Studies*, 76(4), 1239–1267. <https://doi.org/10.1111/j.1467-937X.2009.00550.x>
- Christensen, D. G., Nance, W. R., & White, D. W. (2012). Academic performance in MBA programs: Do prerequisites really matter? *Journal of Education for Business*, 87(1), 42–47. <https://doi.org/10.1080/08833233.2011.555790>
- Coleman, J. S. (1959). Academic achievement and the structure of competition. *Harvard Educational Review*, 29, 330–351.
- Coleman, J. S. (1960). The adolescent subculture and academic achievement. *American Journal of Sociology*, 65(4), 337–347. <https://doi.org/10.1086/222722>
- Coleman, J. S. (1961). *The adolescent society: The social life of the teenager and its impact on education*. Free Press of Glencoe.
- DiMaggio, P. (1987). Classification in art. *American Sociological Review*, 52(4), 440–455. <http://www.jstor.org/stable/2095290>.
- DiMaggio, P. (2011). Cultural networks. In J. Scott, & P. J. Carrington (Eds.), *The sage handbook of social network analysis* (pp. 286–300). Sage.
- DiMaggio, P. (2020). Duality and relationality: The cultural matrix of John W. Mohr. In *Poetics*, 78, Article 101438. <https://doi.org/10.1016/j.poetic.2020.101438>.
- Doehne, M., McFarland, D. A., & Moody, J. (2024). Network ecology: Tie fitness in social context(s). *Social Networks*, 76, 174–190. <https://doi.org/10.1016/j.socnet.2023.09.005>
- Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5), 1739–1774. <https://doi.org/10.1257/aer.101.5.1739>
- Dvir-Gvirman, S. (2017). Media audience homophily: Partisan websites, audience identity and polarization processes. *New Media & Society*, 19(7), 1072–1091. <https://doi.org/10.1177/146144815625945>
- Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology*, 103(2), 281–317. <http://www.jstor.org/stable/2782992>.
- Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, 24, 313–343. <http://www.jstor.org/stable/223484>.
- Falk, A., & Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1), 39–57. <https://doi.org/10.1086/497818>
- Farkas, G., Grobe, R. P., Sheehan, D., & Shuan, Y. (1990). Cultural resources and school success: Gender, ethnicity, and poverty groups within an urban school district. *American Sociological Review*, 55(1), 127–142. <https://doi.org/10.2307/2095708>
- Faulkner, R. R. (1983). *Music on demand : Composers and careers in the hollywood film industry*. Transaction Publishers.
- Fine, G. A. (1979). Small groups and culture creation: The idioculture of Little League baseball teams. *American Sociological Review*, 44(5), 733–745. <http://www.jstor.org/stable/2094525>.
- Fine, G. A. (2012). Group culture and the interaction order: Local sociology on the meso-level. *Annual Review of Sociology*, 38(1), 159–179. <https://doi.org/10.1146/annurev-soc-071811-145518>
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5), 75–174.

- Freeman, L. C. (1992). The sociological concept of "group": An empirical test of two models. *American Journal of Sociology*, 98(1), 152–166. <https://doi.org/10.1086/229972>
- Goldberg, A., Srivastava, S. B., Manian, V. G., Monroe, W., & Potts, C. (2016). Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *American Sociological Review*, 81(6), 1190–1222. <https://doi.org/10.1177/0003122416671873>
- Goldsmith-Pinkham, P., & Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3), 253–264. <https://doi.org/10.1080/07350015.2013.801251>
- Gray, J. B., & Ling, R. F. (1984). K-clustering as a detection tool for influential subsets in regression. *Technometrics: A journal of statistics for the physical, chemical, and engineering sciences*, 26(4), 305–318. <http://www.jstor.org/stable/1269490>
- Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5), 527–544. <https://doi.org/10.1002/jae.741>
- Hoffman, M., Block, P., & Snijders, T. A. B. (2023). Modeling partitions of individuals. *Sociological Methodology*, 53(1), 1–41. <https://doi.org/10.1177/00811750221145166>
- Kilduff, M., & Krackhardt, D. (2008). *Interpersonal networks in organizations: Cognition, personality, dynamics, and culture*. Cambridge.
- Lee, M., & Martin, J. L. (2015). Surfeit and surface. *Big Data and Society*, 2(2), 1–3. <https://doi.org/10.1177/2053951715604334>
- Lewis, K., & Kaufman, J. (2018). The conversion of cultural tastes into social network ties. *American Journal of Sociology*, 123(6), 1684–1742.
- Lizardo, O. (2006). How cultural tastes shape personal networks. *American Sociological Review*, 71, 778–807. <https://doi.org/10.1177/000312240607100504>
- Lizardo, O. (2024). Culture and social capital. In S. McDonald, R. Côté, & J. Shen (Eds.), *Handbook on inequality and social capital* (pp. 140–156). Elgar.
- Lomi, A., Snijders, T. A. B., Steglich, C. E. G., & Torló, V. J. (2011). Why are some more peer than others? Evidence from a longitudinal study of social networks and individual academic performance. *Social Science Research*, 40(6), 1506–1520. <https://doi.org/10.1016/j.ssresearch.2011.06.010>
- Lorenz, G., Boda, Z., Salikutluk, Z., & Jansen, M. (2020). Social influence or selection? Peer effects on the development of adolescents' educational expectations in Germany. *British Journal of Sociology of Education*, 41(5), 643–669. <https://doi.org/10.1080/01425692.2020.1763163>
- Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3), 531–542. <https://doi.org/10.2307/2298123>
- Martin, J. L. (2000). What do animals do all day?: The division of labor, class bodies, and totemic thinking in the popular imagination. *Poetics (Hague, Netherlands)*, 27(2–3), 195–231. [https://doi.org/10.1016/S0304-422X\(99\)00025-X](https://doi.org/10.1016/S0304-422X(99)00025-X)
- Martin, J. L. (2003). What is field theory? *American Journal of Sociology*, 109(1), 1–49. <https://doi.org/10.1086/375201>
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415–444. <http://www.jstor.org/stable/2678628>
- McPherson, M., & Smith, J. A. (2019). Network effects in Blau space: Imputing social context from survey data. *Socius*, 5, 1–21. <https://doi.org/10.1177/2378023119868591>
- Meuleman, B., Loosveldt, G., & Emonds, V. (2015). Regression analysis: Assumptions and diagnostics. In H. Best, & C. Wolf (Eds.), *The sage handbook of regression analysis and causal inference* (pp. 83–110). Sage. <https://doi.org/10.4135/9781446288146>
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2), 340–363. <https://doi.org/10.1086/226550>
- Meyer, J. W., Rowan, B., & Meyer, M. W. (1978). The structure of educational organizations. In J. Ballantine, & J. Z. Spade (Eds.), *Schools and society: A sociological approach to education* (pp. 217–225). Pine Forge Press.
- Mische, A., & Pattison, P. (2000). Composing a civic arena: Publics, projects, and social settings. *Poetics (Hague, Netherlands)*, 27(2–3), 163–194. [https://doi.org/10.1016/S0304-422X\(99\)00024-8](https://doi.org/10.1016/S0304-422X(99)00024-8)
- Mohr, J. W. (1994). Soldiers, mothers, tramps and others: Discourse roles in the 1907 New York City charity directory. *Poetics (Hague, Netherlands)*, 22(4), 327–357. [https://doi.org/10.1016/0304-422X\(94\)90013-2](https://doi.org/10.1016/0304-422X(94)90013-2)
- Mohr, J. W., Bail, C. A., Frye, M., Lena, J. C., Lizardo, O., McDonnell, T. E., Mische, A., Tavory, I., & Wherry, F. F. (2020). *Measuring culture*. Columbia University Press.
- Mohr, J. W., & Rawlings, C. (2010). In J. C. Alexander, R. Jacobs, & P. Smith (Eds.), *The use of formal models in cultural sociology*. The Oxford Handbook of Cultural Sociology.
- Mohr, J. W., & White, H. C. (2008). How to model an institution. *Theory and Society*, 37(5), 485–512. <https://doi.org/10.1007/s11186-008-9066-0>
- Moses, L. E. (1992). The reasoning of statistical inference. In D. C. Hoaglin, & D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107–122). American Mathematical Society.
- Pachucki, M. A., & Breiger, R. L. (2010). Cultural holes: Beyond relationality in social networks and culture. *Annual Review of Sociology*, 36, 205–224. <https://doi.org/10.1146/annurev.soc.012809.102615>
- Padgett, J. F. (2018). Faulkner's assembly of memories into history: Narrative networks in multiple times. *American Journal of Sociology*, 124(2), 406–478. <https://doi.org/10.1086/700831>
- Paloyo, A. R. (2020). Peer effects in education: Recent empirical evidence. In S. Bradley, & C. Green (Eds.), *The economics of education* (2nd ed., pp. 291–305). Elsevier. <https://doi.org/10.1016/B978-0-12-815391-8.00021-5>
- Parker, A., Pallotti, F., & Lomi, A. (2022). New network models for the analysis of social contagion in organizations: An introduction to autologistic actor attribute models. *Organizational Research Methods*, 25(3), 513–540. <https://doi.org/10.1177/10944281211005167>
- Pattison, P., & Robins, G. (2002). Neighborhood-based models for social networks. *Sociological Methodology*, 32(1), 301–337. <https://doi.org/10.1111/1467-9531.00119>
- Perrin, A. J. (2004). Who's afraid of general linear regression? *Newsletter of the Sociology of Culture Section of the American Sociological Association*, 18(3), 11–13, 1. https://asaculturesection.files.wordpress.com/2018/03/2004b_spring.pdf
- Podolny, J. M., & Baron, J. N. (1997). Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review*, 62(5), 673–693. <http://www.jstor.org/stable/2657354>
- Raabe, I. J., Boda, Z., & Stadtfeld, C. (2019). The social pipeline: How friend influence and peer exposure widen the STEM gender gap. *Sociology of Education*, 92(2), 105–123. <https://doi.org/10.1177/0038040718824095>
- Rambotti, S., & Breiger, R. L. (2020). Extreme and inconsistent: A case-oriented regression analysis of health, inequality, and poverty. *Socius*, 6, 1–13. <https://doi.org/10.1177/2378023120906064>
- Ramirez, N. G., Wesche, R., & Kreager, D. A. (2024). Dating, sexuality, and adolescent friendship networks. In M. E. Feinberg, & D. W. Osgood (Eds.), *Teen friendship networks, development, and risky behavior* (pp. 218–246). Oxford. <https://doi.org/10.1093/oso/9780197602317.003.0009>
- Rawlings, C. M., & Childress, C. (2021). Measure Mohr culture. *Poetics (Hague, Netherlands)*, 88(1), Article 101611. <https://doi.org/10.1016/J.POETIC.2021.101611>
- Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In E. A. Hanushek, L. Woessmann, & S. J. Machin (Eds.), *Handbook of the economics of education* (pp. 249–277). Elsevier. <https://doi.org/10.1016/B978-0-444-53429-3.00004-1>
- Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50), Article e2102141118. <https://doi.org/10.1073/pnas.2102141118>
- Schoon, E. W., Melamed, D., & Breiger, R. L. (2024). *Regression inside Out*. Cambridge. <https://doi.org/10.1017/9781108887205>
- Shaffer, J. F., Dang, J. V., Lee, A. K., Dacanay, S. J., Alam, U., Wong, H. Y., Richards, G. J., Kadandale, P., & Sato, B. K. (2016). A familiar(ity) problem: Assessing the impact of prerequisites and content familiarity on student learning. *PLOS ONE*, 11(1), Article e0148051. <https://doi.org/10.1371/journal.pone.0148051>
- Shalev, M. (2007). Limits and alternatives to multiple regression in comparative research. In L. Mjøsset, & T. H. Clausen (Eds.), *Limits and alternatives to multiple regression in comparative research*. *Comparative Social Research*, 24, 261–308.
- Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2), 211–239. <https://doi.org/10.1177/0049124111404820>

- Snijders, T. A. B., Lomi, A., & Torló, V. J. (2013). A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. *Social Networks*, 35(2), 265–276. <https://doi.org/10.1016/j.socnet.2012.05.005>
- Sonnett, J., & Breiger, R. (2004). How relational methods matter: A reply to Perrin. *Newsletter of the Sociology of Culture Section of the American Sociological Association*, 19(1), 8–10, 1 https://asaculturesection.files.wordpress.com/2018/03/2004c_fall.pdf.
- Susca, V. A. R., Vivo, P., & Kühn, R. (2021). Second largest eigenpair statistics for sparse graphs. *Journal of Physics A: Mathematical and Theoretical*, 54(1), Article 015004. <https://doi.org/10.1088/1751-8121/abcba4>
- Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician*, 73, 246–261. <https://doi.org/10.1080/00031305.2018.1518264>
- Wasserman, S., & Pattison, P. E. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika*, 61(3), 401–425. <https://doi.org/10.1007/BF02294547>
- White, H., Boorman, S., & Breiger, R. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology*, 81(4), 730–780. <https://www.jstor.org/stable/2777596>.

Ronald Breiger is Regents Professor of Sociology Emeritus at the University of Arizona. His interests are in social network analysis, culture, and measurement issues in social and cultural research. He is a recipient of the Simmel Award (INSNA) and the J.S. Coleman and P.F. Lazarsfeld distinguished career awards of (respectively) the American Sociological Association Mathematical Sociology and Methodology Sections.

Alessandro Lomi is a Professor at the Università della Svizzera italiana (USI) in Lugano (Switzerland). His current interests include statistical models for social and other networks. He serves as an associate editor at the Journal of the Royal Statistical Society (Series A, Statistics and Society). He is a Fellow of the European Academy of Sociology, and a life member of Clare Hall College, University of Cambridge (UK).

Francesca Pallotti is an Associate Professor at the Department of Management, University of Exeter Business School. Her research investigates how social relations develop and affect individual behavior within formal organizations. Her current research focuses on social identity and organizational change, coordination within teams, and the emergence of informal social relations in hierarchically structured settings.

Culture as Configurations of Categories: Analyzing Peer Effects
via Dual-to-Regression Modeling

Appendix

Contents

Appendix A: Model Selection.....	2
Appendix B: Elaboration of Methods and Findings	6

Appendix A: Model Selection

Because we identified the clusters of Table 4 from our inductive examination of the patterning of contributions to the regression coefficients (in Table 3), the question naturally arises as to whether the improvement of the Table 4 regression model reflects overfitting. How robust are our discovered clusters to sampling variability?

Appropriate statistics for model selection in the case of non-nested models include the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). The adjusted R^2 is not designed for comparison of non-nested models. Further, as James et al. (2021: 235) point out, despite its popularity, and even though it is quite intuitive, the adjusted R^2 is not as well motivated in statistical theory as are AIC and BIC. In general, BIC tends to penalize models with a large number of parameters more severely than AIC, so it is often used when the goal is to find a parsimonious model (James et al., 2021: 234). Below we will look at applications to our dataset of both the BIC and the AIC measures. For the latter we use the bias-corrected (or “second-order”) Akaike Information Criterion AIC, denoted AICc, defined in eq. 3.4 of Hurvich & Tsai (1989: 300), computed via the multi-model inference (MuMIn) package in R. The AICc statistic is recommended by Hurvich & Tsai (1989) when sample size is small in comparison to the number of estimated parameters.¹

We first examine the change in these measures (AICc and BIC), referred to below as delta(AICc) and delta(BIC) respectively, for the regression model of Table 4, in comparison to that of Table 2, in application to our dataset. Since overfitting might be a problem, however, we then examine the change in these measures for a large number of bootstrapped samples (10,000 samples) from our dataset (which comprises 138 students), where each bootstrapped sample is a random sample (with replacement) of size 138 from our observed dataset (see Efron & Hastie, 2016, on the logic of the bootstrap and other resampling schemes). Thus, each of our bootstrapped samples is likely just by chance to omit some of the students in our dataset, and to include some others multiple times. For each of the 138 “students” (observation) in each bootstrapped sample, we identify the cluster of that observation on the basis of its characteristics, according to the clusters identified in Table 4. (For example, all observations that include {a major in Economics or Management AND a reported gender of female AND a first language that is not Italian AND a nationality that is not Swiss AND an age of 23 or older} are assigned to the second-listed cluster of Table 4, cluster C1).

For our observed dataset, the regression model of Table 4 results in a BIC statistic that is 25.28 less than the BIC for the regression model of Table 2 (delta(BIC) = 25.28), and also results in an AICc statistic that is 14.81 lower (delta(AICc) = 14.81). Lower values of each statistic are better (indicate better model fit, while taking into account the number of parameters expended in fitting the model), so both statistics favor the regression model of Table 4. However, these results do not take account of possible overfitting. For that purpose, we turn to our 10,000 bootstrapped samples.

¹ Burnham & Anderson (2002) recommend the use of AICc when the ratio of sample size to number of estimated parameters is less than 40.

For each bootstrapped sample, we run the regression model of Table 2, and we run the regression model of Table 4 on that particular sample. We compute both AICc and BIC scores for each model, on each sample. For each sample, we compute the difference in AICc scores – delta(AICc) is notation for the AICc score for the model of Table 2 minus the AICc score for the model of Table 4 – and similarly for the BIC scores. We thus have 10,000 bootstrapped differences in AICc scores, and 10,000 bootstrapped differences in BIC scores. Differences less than zero (i.e., negative differences) support the null hypothesis that the Table 2 model fits better than the Table 4 model. Differences greater than zero (i.e., positive differences) support the alternative hypothesis that the regression model of Table 4 has the better fit.

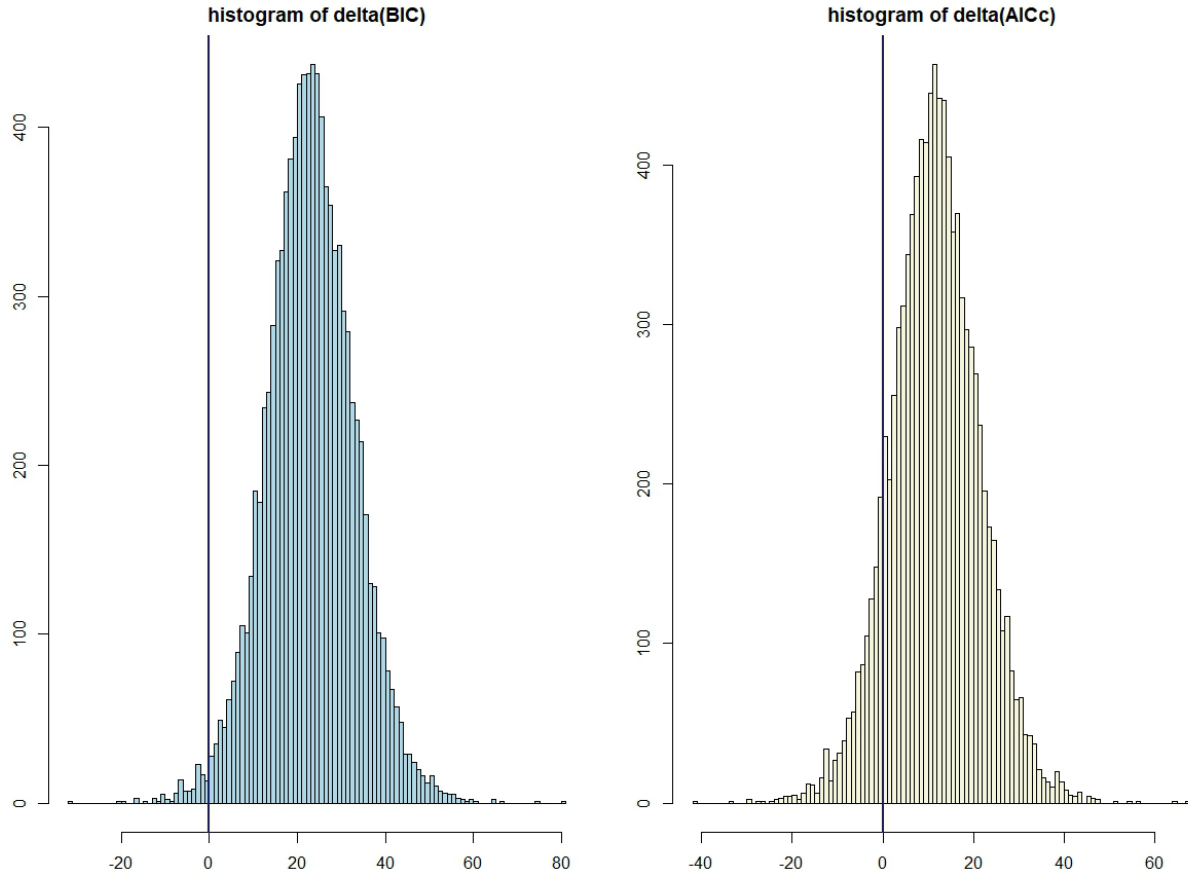


Fig. A.1. (a) Histogram for 10,000 differences: BIC scores for the regression model of Table 2 “minus” BIC for the regression model of Table 4. (b) Histogram for 10,000 differences using AICc scores. In both (a) and (b), differences of 0 are indicated by a vertical line.

We report in Fig. A.1 histograms for the differences in BIC and AICc scores (respectively), where a difference of 0 is marked by a dark vertical line in each panel. From panel (a), the p-value for testing the null hypothesis using BIC is .0114 (because 114 of our 10,000 random bootstrapped samples have a difference in BIC scores that is less than 0). From panel (b), the p-value for testing the same null hypothesis using AICc is just under .11 (that is: 1,075 of our 10,000 bootstrapped samples have a difference in AICc scores that is less than 0).

In Table A.1, we report various descriptive features of the two distributions shown in Fig. A.1.

Table A.1

Distribution of delta(BIC) and delta(AICc) for 10,000 bootstrapped samples.

Criterion	Min.	0.025	1st Q	Median	Mean	3rd Q	0.975	Max.
delta(BIC)	-31.347	3.500	16.704	22.926	23.073	29.434	42.860	80.578
delta(AICc)	-41.822	-7.535	5.344	11.526	11.632	17.842	31.134	67.396

From Table A.1, we see that a 95% confidence interval for the change in BIC scores – the BIC for the regression model of Table 2, minus the BIC for the regression model of Table 4, across our 10,000 bootstrapped samples – is $3.50 < \text{delta}(\text{BIC}) < 42.86$. The comparable 95% confidence interval based on AICc scores is $-7.535 < \text{delta}(\text{AICc}) < 31.134$. Using BIC scores, we reject the null hypothesis (which is a preference for the regression model of Table 2) with 95% confidence (because 0 is not in the confidence interval for delta(BIC), and, as we have seen in the preceding paragraph, $p = .0114$). However, using AICc scores, we cannot reject the null hypothesis with 95% confidence (and, as we saw above, $p = .1075$).

Interestingly, there is a test due to Vuong (1989) that computes confidence intervals for differences in BIC scores, and differences in AIC scores, from the two original models (in our case, the models of Table 2 and Table 4, respectively), without going to the trouble of bootstrapping. Using the Vuong test (as implemented by the function `icci()` in R library “nonnest2”), the 95% confidence intervals are, respectively,

$$8.526 < \text{delta}(\text{BIC}) < 42.042 \quad (\text{Vuong test})$$

and

$$-3.183 < \text{delta}(\text{AIC}) < 30.333 \quad (\text{Vuong test}).$$

(Note that the Vuong test as implemented in the R function we used works with AIC scores but not with corrected AICc scores.) Although Vuong’s test rests on assumptions about asymptotics that are not imposed by the bootstrapping procedure, the confidence intervals constructed using the Vuong test lead us to the same conclusion as those constructed from the bootstrapped samples. Namely: Using BIC, we find that the regression model of Table 4 significantly improves fit in comparison to the regression model of Table 2 (significant at the .05 level, $p = .0114$) and is robust to sampling variability. However, using AICc (in our bootstrapped samples) or AIC (in the Vuong test), the Table 4 model is not a significant improvement in the face of sampling variability (the 95% confidence interval includes 0, and $p = .1075$).

Interpretation: We will discuss two possible interpretations. One possible and reasonable interpretation of these results is that they are inconsistent, in that, using the BIC criterion, we can say with confidence that the improvement in model fit of the regression model using clusters of cases (the model of Table 4) is robust across the sampling variability that is tested by the bootstrapped samples, which lends support to an argument that the improved fit of the Table 4 model was not due to overfitting. However, we are less than 95% confident in preferring the model of Table 4 if we use the AICc (or, in the Vuong test, the AIC) statistic. Hence the inconsistency.

But there is also a case to be made for privileging the BIC statistic. As reviewed in the second paragraph of this Appendix, BIC and the AIC statistics give different values for a reason, namely that BIC tends to penalize models with a large number of parameters more severely than does AIC, so it is often used when the goal is to find a parsimonious model. The model of Table 4 uses fewer parameters than the model of Table 2 (6 clusters are the independent variables in the Table 4 model, while 10 parameters, including the intercept, are the independent variable in the model of Table 2). The BIC result gives us more confidence in preferring the model of Table 4 ($p = .0114$) than does the AICc model ($p = .1075$) because BIC tends to favor models that are more parsimonious.

The BIC p-value and confidence interval tell us that the more parsimonious model of Table 4 fits better than the model of Table 2 in a manner that is robust to the sampling variability that was induced by the bootstrapped sampling. If we ourselves admit to a preference for more parsimonious models, and hence the BIC criterion, then we conclude that the better fit of the model of Table 4 is robust to sampling variability, which (as mentioned) buttresses an argument that the improved fit of the Table 4 model was not due to overfitting.

References cited

- Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer.
- Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297–307. <https://doi.org/10.1093/biomet/76.2.297>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307–333. <https://doi.org/10.2307/1912557>

Appendix B: Elaboration of Methods and Findings

- B.1. Tong (2019:247) uses the term “uncertainty” to refer to “a probability claim, such as a confidence level, Type 1 error, p -value, posterior probability, likelihood function, and so on.” Chatfield (1995, cited by Tong, 2019:248) used the term “model selection bias” to describe the distorted inferences that result when using the same data that determines the form of the final model to also produce inferences from that model. Tong (2019:246) points out that “most scientific research” is “inherently exploratory, requiring flexible methods of analysis that inherently risk overfitting,” and Tong further argues provocatively that “formal, probability-based statistical inference should play no role in most scientific research.”
- B.2 Pairs of students connected by a Friend tie (whether or not reciprocated) have an average cosine profile similarity (our measure of homophily) of +.286. Pairs connected by a reciprocated Friend tie have an average profile similarity of +.318, which is .57 standard deviations above the mean similarity. Pairs of students neither of whom has a Friend tie to the other have average similarity of +.112.
- B.3 A descriptive measure is the correlation of the advice matrix with the profile similarity matrix, a correlation of +.148. To assess this correlation, we apply a permutation test (a simple QAP procedure; Dekker et al., 2007). Keeping one of these matrices fixed, we randomly permute (i.e., rearrange) the rows (and simultaneously the columns) of the other matrix, and then compute their correlation. We repeat this a large number of times (50,000). The null hypothesis is that the labels (names) of individuals are randomly assigned. Of 50,000 random permutations, the highest correlation of the Advice matrix and the (permuted) profile similarity matrix was +.040. Our observed correlation (+.148) is 16 standard deviations above the mean correlation (of 0) under the null hypothesis that labels of the two matrices were matched at random. This provides strong evidence for rejecting the null hypothesis. Qualitatively similar results pertain to the friendship network.²
- B.4 Table 3 has another interesting property, in addition to its rows summing (across sets) to the regression coefficients of Table 2. For any column (set) in Table 3, the sum of cross-products of that column and the column-sums of the data matrix, divided by the number of students in that set, equals the mean of the (observed) scores of those students in the midterm exam. In brief: the rows of Table 3 sum to the regression coefficients, and a weighted version of Table 3’s columns yields the observed mean of each set (in our example, the students in a quadrant) on the outcome variable.
- B.5 Dividing each number in Table 5 by the number of students in each cluster, given in Table 4 (or by the number of students minus one, for the focal student’s own cluster; dividing by one for “self”), we obtain the mean contributions by cluster to each student’s predicted midterm score that are shown in Table B.1.

² The observed correlation between the Friendship social network and the profile similarity network is +.162. In 50,000 random permutations, the highest correlation was +.046. The observed correlation is 15 standard deviations above the mean correlation (of 0).

Table B.1

Peer effects on achievement, for four students: mean within cluster

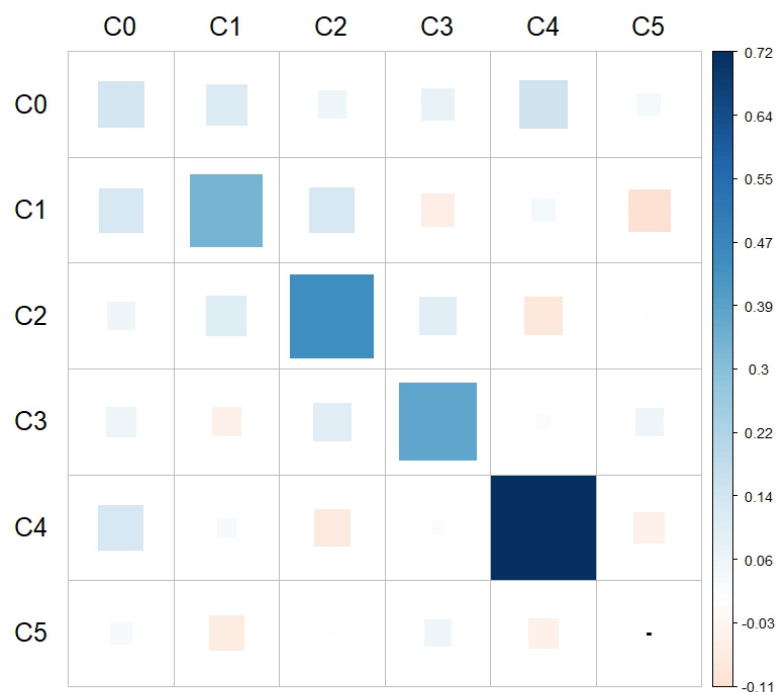
	self or mean within Cluster						
name	self	C0	C1	C2	C3	C4	C5
Anton	0.8885	0.1428	-0.0752	-0.0289	-0.0418	0.7558	0.1733
Bernd	0.8303	0.1617	-0.0604	-0.0669	-0.0394	0.6827	-0.1013
Chiara	0.6428	0.0301	0.1628	0.4186	0.1275	-0.1707	-0.0993
Dario	0.5109	0.0400	0.1708	0.4044	0.1288	-0.2078	-0.2601

For the mean contribution (expressed as mean number of points on the midterm exam) of each cluster to each cluster (with clusters defined as in Table 4), across all 138 students, we have this matrix:

	C0	C1	C2	C3	C4	C5
C0	0.1328	0.1054	0.0484	0.0656	0.1448	0.0303
C1	0.1223	0.3333	0.1272	-0.0683	0.0333	-0.1097
C2	0.0475	0.1002	0.4412	0.0892	-0.0890	-0.0059
C3	0.0562	-0.0522	0.0910	0.3795	-0.0122	0.0479
C4	0.1272	0.0224	-0.0826	-0.0107	0.7182	-0.0570
C5	0.0278	-0.0754	-0.0029	0.0447	-0.0564	18.2792

(Recall that cluster C5 consists only of one student, who evidently was influenced essentially not at all by the others.)

What follows is a heat map of the above matrix, calling attention to high mean contributions to midterm scores within clusters. (We have ignored the internal C5 value to aid visualization.)



B.6 We may wish to zoom in on stronger ties, such as ties of reciprocated advice-seeking. For the overall advice network, 2.4% of all pairs of the 138 students seek advice reciprocally. Reciprocated advice among the 59 students not in any of our configurations of interest is slightly lower, at 2.0%. The students who perform on average 2.27 points worse than this comparison group have a slightly higher density on reciprocated advice (5.5%). However, the configurations of students who perform (respectively) 1.40, 1.71, and 2.80 points above the comparison group have densities for reciprocated advice of 7.6%, 11.8%, and 17.6% (respectively), suggesting that, when profile similarity boosts performance, high-density networks of strong ties (reciprocated advice seeking) are discovered. We found qualitatively similar results (not reported) for reciprocated friendship. We also fit simple ERGM models (not reported) with parameters for overall density and density internal to each set in Tables 4 (using option “nodematch” with “diff=TRUE” in the statnet package for the R environment, a model of differential homophily; see Luke 2015: 173-174), finding parameters that varied qualitatively similarly to those reported in Table 6 for the descriptive measures of density by configuration.

Our dataset also features a binary two-mode matrix of size 138 (students) by 143 (companies), as these students checked off a collective total of 143 companies at which they would consider applying for a job after graduation. For the 138 students as a whole, an average pair of students would consider applying for jobs at the same 9.9 companies. With respect to differences by configuration, these are reported in Table B.2 on the following page. We can think of the average shared desire of students within a given configuration to apply to the same companies after graduation as another measure of cohesion among (potential) sets of peers.

Remarkably, the ordering of these means by configuration (Table B.2) is almost identical to the ordering of densities within configurations on the Advice matrix in the previous table. The comparison group (those students not in a configuration of interest to us) and those who performed relatively poorly on the midterm exam shared relatively little desire to apply to the same companies after graduation (7.67 and 7.22 companies respectively). Students in configurations that performed better than the comparison group on average by (respectively) 1.40, 1.71, and 2.80 points shared a desire on average (respectively) to apply to 12.78, 13.76, and 15.62 of the same companies. These data provide yet another indicator that the configurations of cases (students) discovered by our inductive procedure are related both to social network cohesion and shared aspirations after graduation. Those configurations that manifested relatively high academic performance, in particular, were most likely to have high cohesion in their networks and in their aspirations. The conclusion of this part of our analysis is that, in multiple ways, we find that, when similarity in the intersection of multiple categories is associated with a boost in performance of subsets of students, relatively highly cohesive clusters are discovered.

Table B.2

Number of companies at which both members of an average pair of students would consider applying for a job, by configuration and overall

Config. Name	(N)	Configuration	coef.	Mean number of companies
C0	(59)	(Intercept)	13.9831	7.67
C1	(14)	EconMgmt; ~ItalLang; ~Swiss; Female; Older	-2.2688	7.22
C2	(21)	EconMgmt; ItalLang; WorkEx; ~BAQual	1.3979	12.78
C3	(29)	~EconMgmt; ItalLang; ~Swiss	1.7066	13.76
C4	(14)	~EconMgmt; ~ItalLang; Swiss; ~EngLang	2.8027	15.62
C5	(1)	Age=55	5.0169	--
	(138)	Overall mean number of companies		9.92

References cited

- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A*, 158(3), 419. <https://doi.org/10.2307/2983440>
- Dekker, D., Krackhardt, D., & Snijders, T. A. B. (2007). Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika*, 72(4), 563–581. <https://doi.org/10.1007/s11336-007-9016-1>
- Luke, D.A. (2015). *A User's Guide to Network Analysis in R*. Springer.
- Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician*, 73, 246–261. <https://doi.org/10.1080/00031305.2018.1518264>