Quantitative Economics 14 (2023), 1295–1335 1759-7331/20231295

Spatial interactions

Jun Sung Kim
Department of Economics, Sungkyunkwan University

Eleonora Patacchini
Department of Economics, Bocconi University and Department of Economics, Cornell University

Pierre M. Picard
DEM, University of Luxembourg and LIDAM/CORE, Université catholique de Louvain

Yves Zenou
Department of Economics, Monash University, CEPR, and IZA

This paper studies how the strength of social ties is affected by the geographical location of other individuals and their social capital. We characterize the equilibrium in terms of both social interactions and social capital. We show that lower travel costs increase not only the interaction frequency but also the social capital for all agents. We also show that the equilibrium frequency of interactions is lower than the efficient one. Using a unique geocoded data set of friendship networks among adolescents in the United States, we structurally estimate the model and show that indeed agents socially interact less than that at the first best optimum. Our policy analysis suggests that, at the same cost, subsidizing social interactions yield a higher total welfare than subsidizing transportation costs.

Keywords. Social networks, location, structural estimation, policies.

JEL classification. D85, R23, Z13.

1. Introduction

Over the past two decades, the economics literature has increasingly utilized network analysis to understand decision-making.[1] Surprisingly, however, the importance of spa-

Jun Sung Kim: jskim1221@skku.edu
Eleonora Patacchini: eleonora.patacchini@unibocconi.it
Pierre M. Picard: pierre.picard@uni.lu
Yves Zenou: yves.zenou@monash.edu

We are grateful to three anonymous referees, Jan K. Brueckner, Paco Maruhenda, and the participants at the Inaugural Urban and Regional Economics Conference, Singapore, December 2017, the RIETI Workshop on Frontiers in Urban Economics and Trade, Tokyo, June 2019, the 14th Meeting of the Urban Economics Association, Philadelphia, October 2019, for insightful discussions and comments. Pierre M. Picard was funded in part by the Luxembourg National Research Fund (FNR), grant reference [C20/SC/14755507 and Inter/Mobility/2021/LE/16527808]. For the purpose of open access, Picard has applied a Creative Commons Attribution 4.0 International (CCBY4.0) license to any Author Accepted Manuscript version arising from this submission.

[1]For recent overviews, see Jackson (2008), Ioannides (2013), Jackson and Zenou (2015), Bramoullé, Rogers, and Galeotti (2016), and Jackson, Rogers, and Zenou (2017).

 https://doi.org/10.3982/QE1720

tial proximity in the determination and intensity of network exchange remains underexamined. Indeed, most papers from the network economics literature (Jackson (2008)) assume that the existence and intensity of dyadic contacts do not depend on the agents' location.[2]

In this paper, we develop a new theory of social-tie formation where individuals care about the geographical location of other individuals. In our model, a population of students, embedded in a network and residing in different locations, entertains social interactions with each other. Each student decides the number of visits (social interactions) to every other agent in the network and the value of each interaction depends on the social network of the visited agents. We define the value of such interactions as the *social capital* of the agent (Putnam (2000)). Social capital is thus defined in a recursive fashion: it increases with interactions with highly social individuals. When deciding how much to interact with others, students face the following trade-off. Each student can increase her social capital by interacting with highly social students. However, social interactions require costly travel to the other students. We characterize the equilibrium in terms of social interactions and social capital. We show that the equilibrium frequencies of interactions are lower than the efficient ones. We demonstrate that a policy that subsidizes transportation costs can restore the first best but the subsidy should be higher for trips to students who have higher social capital and for trips from individuals whose social capital increases more with additional interactions.

We then structurally estimate this model using data on patterns of social interactions among high school students in the US recorded in the National Longitudinal Survey of Adolescent Health (Add Health). This survey contains information on friendship nominations and the strength of the interactions between friends, and also allows us to calculate the Euclidean distance between the homes of the respondents. Because residential decisions are taken by parents, this spatial distance is predetermined to the friendship decisions of the children. Our main empirical challenges are due to the fact that there is some discrepancy between the theory and the data in terms of measuring the intensity of social interactions and that the interaction value offered by a friend (social capital) is unobserved to the econometrician. We address these challenges by applying an indirect inference estimation method to simulate unobserved social capital. The main idea of this method is to simulate data from the model, which requires solving for the unobserved equilibrium social capital conditional on structural parameters and unobservables, in order to find the parameters for which the simulated data best match the observed data.[3]

The estimation results highlight the importance of the effects discussed in our theory. We find that transportation costs (and hence geographic distance), social distance, and combined levels of sociodemographic characteristics are all important factors in determining the intensity of social interactions. With the estimated model, we compute the planner's first-best solution for the frequency of social interactions and compare it with the observed equilibrium level. Compared to the socially optimal level, our results

[2]Exceptions include Johnson and Gilles (2000) and Jackson and Rogers (2005).

[3]Fu and Gregory (2019) develop an equilibrium model of post-disaster neighborhood rebuilding choices with externalities and estimate the model using indirect inference to implement policy simulations.

show that students interact with each other far less and accumulate less social capital. We find that these inefficiencies can be explained by the geographical distance between students. With the estimated model, we also simulate the level of social interactions after different policy interventions. By subsidizing social interactions or transportation costs, the policymaker can indeed improve the intensity of social interactions. At the same given cost, we find that subsidizing social interactions is more effective than subsidizing transportation costs because it leads to higher total welfare.

1.1 *Related literature*

We contribute to the literature on *network formation* (Jackson (2008)) by showing the importance of geographical distance in the formation of friendship links. There already exist models of endogenous networks with explicit geographical distance (see, e.g., Johnson and Gilles (2000), Jackson and Rogers (2005)). However, these studies consider a framework where network formation is modeled on a link-by-link basis by extending Jackson and Wolinsky (1996). Thus, these models are usually not tractable and the authors can neither characterize all the equilibria nor derive some comparative statics results and policy implications (see Jackson (2008), for a discussion of these issues). Our model is different; in particular, we have a unique equilibrium. We can also derive comparative statics exercises, explicitly determine the first-best equilibrium and implement some policies. There is another strand of the literature (Brueckner and Largey (2008), Helsley and Strange (2007), Zenou (2013), Mossay and Picard (2011, 2019), Helsley and Zenou (2014), Sato and Zenou (2015), Picard and Zenou (2018)) that studies the role of social networks in cities but take the *network as given*. In the current paper, link formation depends on the location of individuals in the geographical space.

There is also a small empirical literature that studies the relevance of geographical location for social interactions in networks (see Ioannides (2013), for a survey). In fact, it is extremely difficult to find detailed data on social contacts as a function of geographical distance between agents together with information on relevant socioeconomic characteristics. Some evidence can be found in Marmaros and Sacerdote (2006). Using data on email communication between Dartmouth college students, this paper shows that being in the same freshman dorm increases the volume of interactions by a factor of three.[4] Büchel and von Ehrlich (2020) measure social connectedness between postcode areas in Switzerland using mobile phone communication patterns between residents in different areas. They find that distance as measured by travel time is detrimental to private mobile phone interactions by exploiting an exogenous change in travel time.[5] Bailey, Cao, Kuchler, Stroebel, and Wong (2018b) and Bailey, Farrell, Kuchler, and Stroebel

[4]See also Fafchamps and Gubert (2007) who show that geographic proximity is a strong correlate of risk-sharing networks and Rosenthal and Strange (2008), Arzaghi and Henderson (2008), Bisztray, Koren, and Szeidl (2018), and List, Momeni, and Zenou (2019) who find that knowledge and productivity spillovers are important but decay sharply with distance.

[5]Another strand of related literature uses geographic proximity as a proxy for social interactions. Most notably, Bayer, Ross, and Topa (2008) assume that agents living in the same census block exchange information about jobs. Their finding that residing in the same block raises the probability of sharing work location by 33% is thus interpreted as a referral effect. Hellerstein, McInerney, and Neumark (2011), Hellerstein, Kutzbach, and Neumark (2014), and Schmutte (2015) build on the same assumption using matched

(2020) reach a similar conclusion by using anonymized and aggregated data from Facebook to explore the spatial structure of social networks in the New York metropolitan area.

The vast literature in the computer science literature and statistical mechanics looking at the role of distance in social interaction uses primarily mobile phone data or online social networks data and is mainly concerned about describing the shape of the statistical relationship between link probability and distance (see, e.g., Liben-Nowell, Novak, Kumar, Raghavan, and Tomkins (2005) Lambiotte et al. (2008), Goldenberg and Levy (2009), Krings, Calabrese, Ratti, and Blondel (2009) and the excellent reviews of Barthélemy (2011) and Kaltenbrunner et al. (2012)).

To the best of our knowledge, this paper is the first to propose a theory for the relationship between geographical distance and social interactions and to test it using the precise geometry of individual social contacts and the geographical distance between them. It is also the first that empirically establishes the degree of inefficiency of social interactions and, by using counterfactual exercises, determines whether it is more efficient to subsidize transportation costs or social interactions.

The rest of the paper unfolds as follows. Section 2 develops the theoretical model and determines the equilibrium while Section 3 studies its efficiency properties and the policy implications of the model. In Section 4, we describe our data and how we construct our different variables. Section 5 is devoted to the empirical strategy. In Section 6, we provide our main empirical results and discuss some robustness checks. In Section 7, we test the different predictions of the model and determine the level of inefficiencies of social interactions and social capital and how they are affected by the size of the network. We also simulate two policies and determine which one leads to the highest social welfare. Finally, Section 8 concludes the paper and discusses our policy results. All proofs of the theoretical model can be found in the Appendix. In Online Appendix A in the Online Supplementary Material (Kim, Patacchini, Picard, and Zenou (2023)), we solve for the social capital fixed point and show under which condition it is unique. In Online Appendix B, we carry out Monte Carlo simulation experiments while, in Online Appendix C, we explain our calibration method in the policy exercises.

2. The model

2.1 *Notation and definitions*

Consider a set of $N \geq 2$ homogeneous individuals embedded in a social network. As in our data set (see Section 4 below), these are students at a given school, so that all social interactions only take place within the school. We consider one network (within a school) of N students who reside in different locations. Each student i lives with her parent at a given geographical location i.[6] Thus, we denote by d_{ij} the geographical distance

employer–employee data with residential information. Using mobile phone data on one entire city in China, Barwick, Liu, Patacchini, and Wu (2023) show that geographical distance is important in spreading information about jobs.

[6]For the sake of the exposition, we denote by the same letter i both an individual and her residential location.

between two students i and j belonging to the same social network. Each student visits *every other student* in the network and benefits from socially interacting with them. The utility from social interactions for student i is given by

$$S_i = \sum_{j \neq i} v(n_{ij}) s_j, \tag{1}$$

where n_{ij} is the *number* of interactions that student i initiates with student j who offers an interaction value s_j.[7] For the sake of tractability, we assume that[8]

$$v(n_{ij}) = n_{ij} - \frac{1}{2} n_{ij}^2. \tag{2}$$

This expression assumes decreasing returns to the frequency of interactions with a given student; it even assumes negative returns (saturation) above $n_{ij} = 1$. Observe that, in (1), we assume that there are decreasing returns in $v(n_{ij})$ but not in s_j. This is mainly for analytical tractability because we need to calculate a fixed point on social interactions and social capital (see equation (7) below).

The interaction value offered by student j is assumed to be equal to

$$s_j = 1 + \frac{\alpha}{N} \sum_{k \neq j} n_{jk} s_k, \tag{3}$$

where N is the number of students in the network. The first constant term (normalized to 1) represents the idiosyncratic interaction value that student j provides to her visitors. The second term, $(\alpha/N) \sum_{k \neq j} n_{jk} s_k$, reflects the value of her social network. It increases with the number (n_{jk}) and value (s_k) of her interaction with each of her network partners. We refer to s_j as the *social capital* of the student who reside in location j. The parameter $\alpha > 0$ measures the importance of others' social capital in an agent's social capital formation. The higher is α, the higher is the impact of the social network of "friends of friends." We divide α by N to control for network size.

Each student i incurs a cost $c(d_{ij})$ of visiting another student j, where d_{ij} is the geographical distance between i and j. We consider a continuous, increasing cost function with $c(0) = 0$, $c(d_{ij}) > 0$, and $c'(d_{ij}) > 0$, $\forall d_{ij} > 0$. The total social interaction cost of student i is given by

$$C_i = \sum_{j \neq i} n_{ij} c(d_{ij}),$$

which increases with the frequency of social interactions.

[7]Here, as in Cabrales, Calvó-Armengol, and Zenou (2011), individuals do not explicitly choose with whom to link with but decide a level of social interactions at each location in the city.

[8]Observe that in (2), for student i, the curvature in v comes from her interactions with j and not from all her interactions. Since student i has only a limited time for interactions, more interactions with student j could lower her utility from interactions with the other students in the network. Observe also that in (2), we assume that $v(n_{ij})$ only depends on n_{ij}, the number of interactions that student i *initiates* with student j, and not on n_{ji}. In other words, we assume that if student i initiates the interaction with j by commuting to j and bearing this commuting cost, student i will get all the benefits of this interaction while j will not. We assume these two simplifications to keep the model tractable.

We now examine the question of how social capital is distributed across space where students are exogenously located.

2.2 *Social capital and space*

Each student i chooses the profile of interactions n_{ij} that maximizes her utility

$$U_i = S_i - C_i = \sum_{j \neq i} v(n_{ij}) s_j - \sum_{j \neq i} n_{ij} c(d_{ij}).$$

Note that her utility depends on the profile of other student's social capital levels (s_j, $j \neq i$). It also depends on her own social capital s_i, since s_j is a function of s_i (see (3)). We assume that each student takes the social capital levels of all other students as given and is not strategic with respect to the effect of her own social interactions on her utility.

Define the *access cost measure* as

$$g_j \equiv \sum_{k \neq j} c(d_{jk}), \tag{4}$$

which is the total traveling cost of social interactions for student j. Denote by $\overline{d}$ the maximum geographical distance between two agents in the network.

PROPOSITION 1. *Assume $c(\overline{d}) < N$ and $\alpha < 1$. Then, for all i, j, there exists a unique equilibrium (n_{ij}^*, s_j^*) such that*

$$n_{ij}^* = 1 - \frac{c(d_{ij})}{s_j^*} > 0 \tag{5}$$

and

$$s_j^* = s_0 - \frac{\alpha/N}{1 + \alpha/N} g_j > 1, \tag{6}$$

where

$$s_0 = \frac{1 + \alpha/N - (\alpha/N)^2 \sum_j g_j}{(1 + \alpha/N)\big(1 - \alpha(N-1)/N\big)}. \tag{7}$$

Under the conditions $c(\overline{d}) < N$ and $\alpha < 1$, the optimal frequency of interactions n_{ij}^* is always strictly positive and social capital s_j^* is always larger than one. Intuitively, travel costs should not be too high to entice agents to interact. Also, the importance of others' social capital in an agent's social capital formation should not be too high to avoid that each individual's social capital reinforces each others' social capital and ultimately blows up to infinity.

Consider now (5). For student i, n_{ij}^*, the optimal number of interactions with a student j, increases with student j's social capital and decreases with the geographical distance between i and j. Hence, there is complementarity between the frequency, n_{ij}^*, and the quality of social interactions, s_j.

Let us now discuss the properties of the equilibrium social capital s_j^* defined in (6).[9] First, lower travel costs increase social capital for all agents. Indeed, a downward shift in the travel cost function $c(\cdot)$ reduces the access cost measure g_j, which has a positive effect on both terms in (6), since higher access cost increases s_0. As a result, travel costs can be seen as a *barrier to social capital formation*. Improvements in transportation infrastructure should therefore enhance social capital.

Second, a rise in the importance of peers' social links in the creation of their own social capital α, has the following effects. By using the proof of Proposition 1, we can differentiate each side of (42) with respect to α to obtain

$$\frac{\mathrm{d}s_j^*}{\mathrm{d}\alpha} = \frac{1}{N}\sum_{k\neq j} s_k^* + \frac{\alpha}{N}\sum_{k\neq j}\frac{\mathrm{d}s_k^*}{\mathrm{d}\alpha} - \frac{1}{N}g_j.$$

Thus, an agent's social capital increases with higher α because she places greater value on the social capital of her interaction partners (first term), because her partners themselves have higher social capital (second term), and finally because she is physically closer to her partners, and thus has higher incentives to meet them (third term). By differentiating (6) with respect to α, we obtain the total effect as a function of exogenous variables:

$$\frac{\mathrm{d}s_j^*}{\mathrm{d}\alpha} = \frac{\mathrm{d}s_0}{\mathrm{d}\alpha} - \frac{1}{N(1-\alpha/N)^2}g_j.$$

This expression is always positive for low enough travel costs $c(d_{ij})$, since the terms in g_j are in this case close to zero. Otherwise, geographically distance agents may get lower social capital.

We summarize these findings in the following proposition.

PROPOSITION 2. *Lower travel costs increase social capital for all agents. An increase in α, the importance of peers' social links, increases each agent's social capital for small enough travel cost.*

We now study the optimal levels of social interaction and capital.

3. EFFICIENT SOCIAL INTERACTIONS

We now study the planner's allocation of interaction frequency for each individual i. The planner chooses the profiles of social interactions n_{ij} and social capital s_j that maximize the aggregate utility

$$W = \sum_i U_i = \sum_i (S_i - C_i)$$

subject to the social capital constraint

$$s_i \leq 1 + \frac{\alpha}{N}\sum_{k\neq i} n_{ik}s_k. \tag{8}$$

[9] Once we know the comparative statics results with respect to s_j^*, then it is straightforward to deduce those of n_{ij}^*.

This inequality allows us to define and interpret the (positive) sign of the Kuhn–Tucker multiplier χ_i (which measures the welfare value of a marginal increase of the social capital of agent i) of the social capital formation constraint. The interpretation of this inequality is that the planner cannot give more social capital to a student than what her interactions with her partners can give. Conversely, the planner can erase some of the social capital of an individual but it has no incentives to do so, since welfare increases with social capital.

LEMMA 3. *The efficient frequency of interactions n_{ij}^o and level of social capital s_j^o satisfy the following necessary conditions*:

$$v'\left(n_{ij}^o\right)s_j^o - c(d_{ij}) + \frac{\alpha}{N}\chi_i s_j^o = 0, \tag{9}$$

$$\sum_i \left[v\left(n_{ij}^o\right) + \frac{\alpha}{N}\chi_i n_{ij}^o\right] - \chi_j = 0. \tag{10}$$

Equations (9) *and* (10) *together with the constraint* (8) *solve for n_{ij}^o, s_j^o, and χ_i.*

Condition (9) captures the main externality at work in the process of social interaction. When the planner chooses the interaction frequency n_{ij}, she considers both the benefit and cost experienced by agent i and the fact that an increase in i's social capital increases j's social capital. In the decentralized equilibrium, this last effect is not considered by agent i. One can indeed see that condition (9) is equal to the first-order condition of the individual's choice of interactions if $\chi_i = 0$. The weight that the planner puts on raising another agent's social capital increases with the importance of interactions, α, and with the social benefit of relaxing the social capital constraint, χ_i. Then, because $\chi_i > 0$ and $v'' > 0$, the equilibrium number of interactions n_{ij} is *smaller* than the ones chosen by the planner. In other words, there are too few interactions in equilibrium.

The second condition (10) can be interpreted as follows. When the planner increases the social capital of agent j, she raises the utility of all agents who interact with this agent (first term in brackets) and indirectly increases the social capital of these agents (second term in brackets). In the efficient allocation, this combined effect should be equal to χ_j, the welfare value of a marginal increase of the social capital of an agent at j.

PROPOSITION 4. *The equilibrium frequency of interactions and level of social capital are lower than the efficient ones, that is, $n_{ij}^* \leq n_{ij}^o$ and $s_i^* \leq s_i^o$.*

Intuitively, the planner internalizes the effect that each agent has on others' social capital when she entertains more intense social interactions. As a result, the planner imposes to the agents to increase their frequency of social interactions above the equilibrium level. This welfare result confirms Brueckner and Largey's (2008) and extends their analysis to the case where agents are distributed across space.

Can the efficient allocation of social interactions be restored with a subsidy σ_{ij} on social interactions (for students i and j) or with a subsidy τ_{ij} on travel costs? Let

$$\tau_{ij}^o = \frac{\alpha}{N}\chi_i^o s_j^o \quad \text{and} \quad \sigma_{ij}^o = \frac{s_j^o}{\dfrac{Nc(d_{ij})}{\alpha\chi_i^o s_j^o} - 1}. \tag{11}$$

PROPOSITION 5. *The first-best solutions n_{ij}^o and s_j^o can be restored by either setting a subsidy on travel costs $\tau_{ij} = \tau_{ij}^o$ or a subsidy on social interactions $\sigma_{ij} = \sigma_{ij}^o$. The subsidy τ_{ij}^o on travel costs should be higher for recipient students who have higher social capital and for trips to students whose social capital increases more with additional interactions. The (positive) subsidy σ_{ij}^o on social interactions increases for recipient students with more social capital, from initiator students who are closer and who have higher welfare value of a marginal social capital increase.*

The optimal subsidies τ_{ij}^o and σ_{ij}^o have no direct relation to distance between students, since it is very unlikely that τ_{ij}^o and σ_{ij}^o reduce to a simple function of the geographical distance d_{ij} between students i and j. This result contrasts with Helsley and Zenou (2014), who advocate that the planner should subsidize the most central agents. Their model, which has only two locations, however, imperfectly captures the full picture of spatial interactions. In the present model, we observe that the planner does not subsidize the agents with high social capital but only subsidizes the trips of these agents.

Note that the subsidies τ_{ij}^o and σ_{ij}^o defined in (11) are *not* uniform. This suggests that decentralization is going to be difficult to implement, since subsidies depend on both the originator and recipient of each social interaction. Consequently, in the counterfactual (subsidy) policies in Section 7.2, we will investigate the effect of subsidies that are *uniform* across individuals, and thus easier to implement.

4. DATA

In this section, we describe our data and the fit between our theoretical framework and them. First, we explain the data source and highlight the key features of the data that are relevant to spatial interactions. Second, we describe how the data measures social interaction intensity among individuals. Third, we discuss the geographic space and the residential distance among students. Fourth, we explain how we construct networks from friendship nominations. Fifth, we describe the final sample after deleting missing variables/observations. In each part, we discuss and address the issues related to the discrepancies between our theory and data.

4.1 *Data source*

We use a data set on friendship networks from the National Longitudinal Survey of Adolescent Health (Add Health) to test our theoretical findings and compute the effects of several policies.[10]

[10]This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Car-

The Add Health data set has been designed to study the impact of the social environment (i.e., friends, family, neighborhood, and school) on adolescents' behavior in the United States. It is a school-based survey that contains extensive information on a representative sample of students who were in grades 7–12 in 1995. More than 100 schools were sampled. Three features of the Add Health data are unique and key to our analysis: (i) the nomination-based friendship information, which allows us to reconstruct the precise geometry of social contacts, (ii) the detailed information about the intensity of social interactions between each of two friends in the network, and (iii) the geocoded information on residential locations, which allows us to measure the geographical distance between students.

4.2 *Construction of n_{ij}, the social-interaction intensity*

All students who were present at school in the interview day were asked to identify their best school friends from a school roster (up to five males and five females).[11] For each individual i, the friendship nomination file also contains detailed information on the frequency and nature of interaction with each nominated friend j. The precise questions are as follows:

- Did you go to {NAME}'s house during the past 7 days?
- Did you meet {NAME} after school to hang out or go somewhere during the past 7 days?
- Did you spend time with {NAME} during the past weekend?
- Did you talk to {NAME} about a problem during the past 7 days?
- Did you talk to {NAME} on the telephone during the past 7 days?

Students can answer these questions with a yes or a no; thus, these answers are coded by one (for yes) and zero (for no). From their answers, we are able to measure the intensity of social interactions n_{ij} between students i and j by summing all these items, so that the maximum value of the social interaction intensity is five and the minimum is zero.

4.3 *Geographical space*

A random sample of students in each school (about 20,000 students) is also interviewed at home where a longer list of questions are asked both to the child and to his/her parents. Most notably for this study, the geographical locations of their residential location

olina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due to Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

[11]The limit in the number of nominations is not binding (even by gender). Less than 1% of the students in our sample show a list of ten best friends.

are also recorded. Latitude and longitude coordinates are calculated for each home address and then translated into X- and Y-coordinates in an artificial space. We use this information to derive the *spatial distance* d_{ij} between any two students i and j by computing the Euclidean distance between their homes. The maximum geographical distance between two students in a network is about 47 kilometers. The average distance is 6.75 kilometers, and its standard deviation is 6.71 kilometers.

4.4 *Discrepancy between theory and data*

There are some discrepancies between our theoretical model and the Add Health data. First, recall that the theoretical model assumes that each student visits, and thus socially interacts with every other student in the network. That is, n_{ij} is always positive (see Proposition 1). By contrast, in the Add Health data set, students can only answer the social-interaction questionnaires for their nominated friends (see Section 4.2). Indeed, if student i does not nominate student j as a friend, then clearly i will not be asked about how many times she interacted with j. In addition, students in the Add Health data set may have zero interactions with their (nominated) friends, which is related to the sparsity of networks, a common problem in network data.

Finally, the measurement of social interactions is different in the theoretical model and the data. Indeed, in the model, n_{ij} takes continuous values. By contrast, the social-interaction intensity in the Add Health data takes a discrete value, which is in $\{0, 1, 2, 3, 4, 5\}$, because there are five different survey questions about students' social interactions and students can answer only by yes or no (Section 4.2). In the next subsections, we explain how we construct our networks and how we tackle these three different issues.

4.5 *Construction of networks*

In Add Health, 15,837 students have geocoded data and other socioeconomic characteristics. Information on nominated friends, types of interactions, and geographical location is only available for 5711 students. This large reduction in sample size is common in the papers that use the Add Health data set with friendship information, and it is mainly due to the network construction procedure—roughly 20% of the students do not nominate any friend and another 20% cannot be correctly linked.

Using the friendship nomination data and the corresponding social interaction responses, we construct two 5711 by 5711 adjacency matrices, indicating the status of (directed) friendship and the intensity of (directed) social interaction. Each element in the former matrix is a binary indicator of whether two students are friends or not, while each element in the latter takes one of the values in $\{0, 1, 2, 3, 4, 5\}$.

Because the students and their schools in the data are geographically dispersed all around the United States, we cannot assume that all those students know each other. Hence, we consider *each school as a network*. From the data, we have a total of 128 *directed* networks (schools), where the values of n_{ij} are not constrained to be equal to those of n_{ji}. Then we exclude networks that have less than five students and two large

networks of 652 and 857 students, and focus on the other 122 networks. Among these 122 networks, the largest one has 134 students, and the smallest one has six students for whom the information on social interaction and geographic location is available. We exclude the two large networks for the following reasons. First, the upper tails of the distribution of networks by network size is commonly trimmed since the strength of peer effects may be too different in too large networks (see Calvó-Armengol, Patacchini, and Zenou (2009)). Second, and most importantly, because the large networks include students who are not likely to know many others, its exclusion reduces the discrepancy between the theoretical model, where all students interact with each other (i.e., $n_{ij} > 0$), and the data, where many students are not nominated as friends by other students (see Section 4.4). By doing so, the number of students in the final sample is 4036.

While we use almost all schools up to a size of 134 students in the structural estimation, we also run the estimation with 100 schools of size up to 50 students. We find that the coefficients estimated in the structural estimation are similar between these two data sets. In the policy section, we use the later data set (up to 50 students) for computational feasibility. Indeed, in our policy simulations, we need to solve optimization problems where the choice variable has a dimension equal to the square of student size; the policy simulations become infeasible for the large data set.

4.6 *Final sample*

Our final sample consists of 4036 individuals distributed over 122 schools. Table 1 describes our data and details our sample selection procedure. We report the characteristics of four different samples, which correspond to the three steps of our selection procedure. In column (1), we consider the original sample of students who have valid individual characteristics. In columns (2)–(3), we further restrict the sample to those with geocode information and intensity of social interactions. Finally, in column (4), we report our final sample where we exclude students in the two large schools.

Table 1 shows that the differences in means between these different samples are mostly statistically insignificant. This strongly suggests no specific bias in the selection of the sample. Among the adolescents in the final sample, 52% are female and 22% are blacks. Approximately 71% of the students live in a household with two parents. The average parental education is high-school graduate. The performance at school, as measured by the grade point average or GPA, exhibits a mean of 2.88, meaning slightly less than a grade of "B." The average family income is 44,050 in 1994 dollars, although 10% of parents chose not to report such information.

In Table 2, we document the number of social interactions. There are a total of 10,582 social interactions over 122 schools, which are mainly between white students; there are fewer interethnic interactions. Further, on average, each pair has 2.533 social interactions; black pairs socially interact with each other slightly more than white pairs do.

5. Empirical strategy

5.1 *Incorporating agents' heterogeneity*

To bring the model to the data, we introduce agents' heterogeneity. We assume that the benefits of the intensity of interactions between individuals at i and j also depend on

TABLE 1. Data description: Individual characteristics.

Variable	Variable Definition	(1) Mean (Std. Dev.)	(2) Mean (Std. Dev.)	Difference [P-Value]	(3) Mean (Std. Dev.)	Difference [P-Value]	(4) Mean (Std. Dev.)	Difference [P-Value]
Female	Dummy variable taking value one if the respondent is female	0.51 (0.50)	0.50 (0.50)	[0.55]	0.50 (0.50)	[0.51]	0.52 (0.50)	[0.48]
Black	Dummy variable taking value one if the respondent is Black or African American. "White" is the reference category	0.23 (0.42)	0.23 (0.42)	[0.45]	0.20 (0.40)	[0.35]	0.22 (0.41)	[0.44]
Student grade	Grade of student in the current year, range 7 to 12	9.67 (1.63)	9.56 (1.62)	[0.41]	9.65 (1.61)	[0.37]	9.30 (1.64)	[0.03]
Grade Point Average	Grades defined from "A" = 4 to "D and lower" = 0. Average of grades in English, math, science, and history is taken	2.75 (0.77)	2.77 (0.77)	[0.47]	2.80 (0.77)	[0.39]	2.88 (0.75)	[0.19]
Physical development	Answer to the question "How advanced is your physical development compared to other boys your age?" Coded as 1 = "I look younger than most," 2 = "I look younger than some," 3 = "I look average," 4 = "I look older than some," 5 = "I look older than most"	3.19 (1.13)	3.21 (1.12)	[0.52]	3.24 (1.11)	[0.54]	3.28 (1.11)	[0.47]
Religion practice	Answer to the question "In the past 12 months, how often did you attend religious services?" Coded as 1 = "once a week or more," 2 = "once a month or more, but less than once a week," 3 = "once a month," 4 = "never"	2.44 (1.44)	2.40 (1.42)	[0.51]	2.34 (1.39)	[0.41]	2.32 (1.40)	[0.48]
Family size	Number of people living in the household	3.61 (1.66)	3.60 (1.57)	[0.51]	3.58 (1.53)	[0.54]	3.45 (1.42)	[0.19]

(*Continues*)

TABLE 1. *Continued.*

Variable	Variable Definition	(1) Mean (Std. Dev.)	(2) Mean (Std. Dev.)	Difference [*P*-Value]	(3) Mean (Std. Dev.)	Difference [*P*-Value]	(4) Mean (Std. Dev.)	Difference [*P*-Value]
Two parents	Dummy variable taking value one if the respondent lives in a household with two parents (both biological and nonbiological) that are married Two parent	0.66 (0.47)	0.69 (0.46)	[0.55]	0.72 (0.45)	[0.24]	0.71 (0.45)	[0.49]
Parental education	Schooling level of the (biological or nonbiological) parent who is living with the child, coded as 1 = "never went to school," 2 = "some school" and "less than high school," 3 = "high school graduate," "GED," "went to a business, trade or vocational school," "some college," 4 = "graduated from college or a university," 5 = "professional training beyond a 4-year college" If both parents are in the household, the maximum level of schooling is considered	3.09 (0.97)	3.12 (0.95)	[0.51]	3.16 (0.94)	[0.44]	3.18 (0.96)	[0.48]
Family income	Family income in thousands of dollars	40.72 (50.76)	40.84 (49.00)	[0.52]	43.16 (53.15)	[0.42]	44.05 (59.51)	[0.50]
Family income refused	Dummy variable taking value one 1 if family income of the respondent is missing	0.09 (0.29)	0.11 (0.31)	[0.49]	0.11 (0.31)	[0.51]	0.10 (0.31)	[0.51]
N. obs		20,745	15,837		5711		4036	

Note: (1) Original sample, (2) sample with geocoded information and no-missing individual characteristics, (3) sample with social-interaction information, (4) sample from all schools except the two large school. T-tests for differences in means are performed. *P*-values are reported in squared brackets. Differences are computed with respect to the larger sample in the previous column.

TABLE 2. Number of social interactions per pair.

	Pair Types			
	Black–Black	Black–White	White–White	All
Number of total social interactions	1688	480	8414	10,582
Number of friendship pairs	659	218	3300	4177
Average social interactions per friendship pair	2.561	2.202	2.550	2.533

Note: The statistics are computed with 122 networks (schools).

their social distances, that is, on their distances in terms of sociodemographic characteristics:

$$v(n_{ij}) = (n_0 + \theta_{ij})n_{ij} - \frac{1}{2}(n_{ij})^2,$$

where θ_{ij} denotes the social distance between individuals i and j and n_0 a constant that captures the baseline level of social interactions. We include θ_{ij} in this equation because students are heterogeneous in the data; θ_{ij} captures the effects of social distance in terms of observable and unobservable characteristics (different from geographic proximity) on the intensity of interactions. The difference between students in the observable and unobsevable characteristics of θ_{ij} is exogenous in the sense that it is not a choice variable of students. We further assume linear travel cost such that $c(d_{ij}) = c \times d_{ij}$ where $c > 0$ is a constant.

In the model, we consider one social network of N students at school who reside in different residential locations. In the data, we have $R = 122$ networks ($r = 1, \ldots, R$) or $R = 100$ networks when we restrict the size of networks to be up to 50 students. Since networks are defined as schools that are independent from each other, we can use our theoretical results by adding the subscript r. In other words, all the results of our theoretical model are valid for each network.

Consequently, the optimal frequency of interaction can be written as follows:

$$n^*_{ij,r} = n_0 - \frac{cd_{ij,r}}{s^*_{j,r}} + \theta_{ij,r}, \tag{12}$$

and the social capital is equal to

$$s^*_{j,r} = 1 + \frac{\alpha}{N_r} \sum_{k=1, k\neq j}^{N_r} n^*_{jk,r} s^*_{k,r}. \tag{13}$$

We allow the social distance to depend on observed (pair-level) individual characteristics $x_{ij,r}$ and on unobserved factors $\varepsilon_{ij,r}$. For simplicity, we assume that $\varepsilon_{ij,r}$ is independent and identically distributed across pairs and networks with mean zero and variance σ^2_ε, but the i.i.d. assumption within a network can be relaxed.

To capture *homophily*, that is, the tendency of individuals to associate and bond with similar others (McPherson, Smith-Lovin, and Cook (2001), Currarini, Jackson, and

Pin (2009), Graham (2017)), we employ the following *undirectional* specification:

$$\theta_{ij,r} = \sum_{m=1}^{M} \beta_m |x_{i,m,r} - x_{j,m,r}| + \sum_{m=1}^{M} \beta_{M+m}(x_{i,m,r} + x_{j,m,r}) + \varepsilon_{ij,r}, \tag{14}$$

where negative values in the vector $(\beta_1, \ldots, \beta_M)$ capture homophily effects (associated with smaller socioeconomic distance $|x_{i,m,r} - x_{j,m,r}|$), and $(\beta_{M+1}, \ldots, \beta_{2M})$ measures the effect of the combined level of x_i and x_j, where M is the number of individual-level covariates. Indeed, under homophily behavior, individuals with similar characteristics (same race, same gender, etc.) will tend to interact more than less similar individuals (thus β_m should be negative under homophily). Similar specifications have been used in the literature; see, for example, Fafchamps and Gubert (2007). Note that having an undirectional specification on for $\theta_{ij,r}$ does not necessarily mean that $n_{ij,r}$ and $n_{ji,r}$ are the same. Because of the presence of social capital $s^*_{j,r}$ in equation (12), the social interaction intensity can be asymmetric between ij and ji. The Add Health data set also exhibits asymmetry between $n_{ij,r}$ and $n_{ji,r}$.

By plugging the value of $n^*_{ij,r}$ from (12) into (13), in Online Appendix A, we solve for the social capital fixed point and show under which condition it is unique. The social capital fixed point is given by (see equation (A.4) in Online Appendix A):

$$\mathbf{s}^*_r = \left[\mathbf{I}_r - \frac{\alpha}{N_r}(\mathbf{N}_{0,r} + \mathbf{\Theta}_r)\right]^{-1} \left(\mathbf{I}_r - \frac{\alpha}{N_r} c\mathbf{D}_r\right) \mathbf{1}_r, \tag{15}$$

where $\mathbf{s}_r = (s_{i,r})$ is a $(N_r \times 1)$ vector; $\mathbf{1}_r$ is the $(N_r \times 1)$ vector of 1; $\mathbf{N}_{0,r}$ is an $(N_r \times N_r)$ matrix in which the off-diagonal elements are n_0 and the diagonal elements are all zero; $\mathbf{\Theta}_r = (\theta_{ij,r}) = (x^{\mathrm{T}}_{ij,r}\beta + \varepsilon_{ij,r})$ is an $(N_r \times N_r)$ matrix; $\mathbf{D}_r = (d_{ij,r})$ is an $(N_r \times N_r)$ matrix (see equation (A.3) in Online Appendix A).

5.2 *Estimation strategy*

5.2.0.1 *Indirect inference* For each network r, our data set provides us with $x_{ij,r}$, the agents' characteristics, $n^*_{ij,r}$, the intensity of social interactions between agents i and j, $d_{ij,r}$, the geographical distance between agents i and j, and N_r, the number of agents in the network. Using this information, we can recover the parameters α, β (or $\beta_1, \ldots, \beta_{2M}$), c, n_0, σ_ε, and the equilibrium social capital, $s^*_{j,r}$. For that, we employ the indirect inference (I-I) estimation method, proposed by Gourieroux, Monfort, and Renault (1993), which recovers the true parameters from the data by attempting to closely match simulated and observed levels of social interactions.[12] The estimator is indirect in the sense that, rather than directly estimating the structural model, it estimates an *auxiliary* model with (computationally) easier methods such as the ordinary least squares (OLS). We run the auxiliary model with the observed data and the simulated ones. The estimates for the structural parameters are the ones that best match the two sets of auxiliary

[12]The I-I method was introduced by Smith (1993) and Gourieroux, Monfort, and Renault (1993), and later extended by Gallant and Tauchen (1996). For overviews on I-I, see Gourieroux and Monfort (1996) and Smith (2008).

parameters, based on an injectivity assumption (i.e., one-to-one mapping between the structural parameters and the auxiliary parameters).

5.2.0.2 *Structural model* For the sake of exposition, we denote the vector of structural parameters by $\mu \equiv (n_0, \alpha, c, \beta, \sigma_\varepsilon)$ and we group the unobserved information into the vector $\mathcal{E}_r \equiv (\varepsilon_{ij,r})$ and the observed information into the vector $\mathbf{Y}_r \equiv (\mathbf{X}_r, \mathbf{D}_r, N_r)$ where $\mathbf{X}_r$ and $\mathbf{D}_r$ capture the individuals characteristics $x_{i,r}$ and the distances $d_{ij,r}$, respectively. The structural models (12) and (15) can now be written as the following system of equations:

$$n^*_{ij,r}(\mathbf{Y}_r, \mathcal{E}_r; \mu) = n_0 - \frac{c d_{ij,r}}{s^*_j(\mathbf{Y}_r, \mathcal{E}_r; \mu)} + x_{ij}^{\mathrm{T}}\beta + \varepsilon_{ij,r}, \tag{16}$$

$$\mathbf{s}^*(\mathbf{Y}_r, \mathcal{E}_r; \mu) = \left[\mathbf{I}_r - \frac{\alpha}{N_r}(\mathbf{N}_{0,r} + \mathbf{\Theta}_r)\right]^{-1}\left(\mathbf{I}_r - \frac{\alpha}{N_r} c\mathbf{D}_r\right)\mathbf{1}_r. \tag{17}$$

As explained in Section 4.2, the observed $n^{\mathrm{obs}}_{ij,r}$ in the data takes one of the six integer values $\{0, 1, 2, 3, 4, 5\}$ while $n^*_{ij,r}$ in the theoretical model can take all values in the set of all nonnegative real numbers. Hence, to fill the gap between $n^*_{ij,r}$ and $n^{\mathrm{obs}}_{ij,r}$, we apply the following mapping to calculate the final $n^{\mathrm{sim}}_{ij,r}$, which we will be the counterpart to $n^{\mathrm{obs}}_{ij,r}$ in the I-I procedure:

$$n^{\mathrm{sim}}_{ij,r} = \begin{cases} 0 & \text{if } -\infty < n^*_{ij,r} < 1; \\ 1 & \text{if } 1 < n^*_{ij,r} \le 2; \\ 2 & \text{if } 2 < n^*_{ij,r} \le 3; \\ 3 & \text{if } 3 < n^*_{ij,r} \le 4; \\ 4 & \text{if } 4 < n^*_{ij,r} \le 5; \\ 5 & \text{if } n^*_{ij,r} \ge 5. \end{cases} \tag{18}$$

More precisely, we set the social interaction intensity between two students as the closest integer value that is lower than the simulated intensity. Then, if the value is less than zero, we make it zero. If the value is greater than five, we set it as five.

5.2.0.3 *Auxiliary model* The main advantage of the I-I method is that researchers can use a simple model to match the simulated data and the observed ones. Specifically, we use simple linear regression equations as auxiliary models. We propose a first auxiliary model equation that expresses the relationship between social interaction intensities, individual characteristics, and distance between interaction partners as follows:

$$n_{ij,r} = \gamma_{10} + x_{ij,r}^{\mathrm{T}}\gamma_{11} + \gamma_{12} d_{ij,r} + \epsilon_{1,ij,r}. \tag{19}$$

We propose a second auxiliary model equation expressing a similar relationship with respect to indirect interactions. Let us denote by $\mathbf{N}_r = (n_{ij,r})$ the $(N_r \times N_r)$ matrix of social interaction intensities for network r. We further define the matrix of second degree interaction as the square matrix $\mathbf{N}_r^2 \equiv \mathbf{N}_r\mathbf{N}_r$. We denote by $[\mathbf{N}_r^2]_{ij}$ the ith row and jth column

element of this matrix. The second auxiliary equation can then be written as

$$\left[\mathbf{N}_r^2\right]_{ij} = \gamma_{20} + x_{ij,r}^{\mathrm{T}}\gamma_{21} + \gamma_{22}d_{ij,r} + \epsilon_{2,ij,r}. \tag{20}$$

We denote by $\boldsymbol{\gamma}$ the vector of the above auxiliary model coefficients and R^2.

5.2.0.4 *Algorithm* We draw T sets of simulation errors, $\mathcal{E}^t \equiv (\varepsilon_{ij,r}^t)$, $t = 1, \ldots, T$, for all pairs i and j and all networks r. These sets of errors are fixed for the entire estimation process.[13] First, we compute social capital $\mathbf{s}_r^t$ and predict the intensity of social interactions $\widehat{n}_{ij,r}^t$ for each set of errors using equations (16) and (17). To match the data, we constrain $\widehat{n}_{ij,r}^t$ to lie between zero and five (included). This process yields the first degree interaction matrix $\widehat{\mathbf{N}}_r(\mathbf{Y}_r, \mathcal{E}_r^t; \mu)$ and the second degree interaction matrix as the square of the latter. Let $\mathbf{Y}$, $\mathcal{E}^t$, $\mathbf{N}$, and $\widehat{\mathbf{N}}(\mathbf{Y}, \mathcal{E}^t; \mu)$ collect the observed data, the nonobserved data, the observed interactions and the predicted interactions in all networks. We then run OLS regressions on the auxiliary models (19) and (20) separately with the observed and simulated interaction values. As a result, we obtain a set of the OLS estimates $\widehat{\boldsymbol{\gamma}}(\mathbf{N}, \mathbf{Y})$, including R^2's, with the observed interactions and a set of estimates $\widehat{\boldsymbol{\gamma}}[\widehat{\mathbf{N}}(\mathbf{Y}, \mathcal{E}^t; \mu), \mathbf{Y}]$, $t = 1, \ldots, T$ with the simulated interactions. Finally, since OLS estimates using the simulated data are functions of the structural parameter vector μ, we choose μ that leads the closest difference between $\widehat{\boldsymbol{\gamma}}(\mathbf{N}, \mathbf{Y})$ and $\widehat{\boldsymbol{\gamma}}[\widehat{\mathbf{N}}(\mathbf{Y}, \mathcal{E}^t; \mu), \mathbf{Y}]$. Formally, the I-I estimator $\widehat{\mu}_{\mathrm{II}}$ is constructed such that

$$\widehat{\mu}_{\mathrm{II}} = \arg\min_{\mu} \left\| \widehat{\boldsymbol{\gamma}}(\mathbf{N}, \mathbf{Y}) - \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{\gamma}}\left[\widehat{\mathbf{N}}\left(\mathbf{Y}, \mathcal{E}^t; \mu\right), \mathbf{Y}\right] \right\|, \tag{21}$$

where the norm $\|\cdot\|$ is defined by a (positive-definite) weight matrix, $\mathbf{A}$, with dimension equal to the number of the auxiliary model parameters. Gourieroux, Monfort, and Renault (1993) show that the efficient weight matrix is given by the inverse of the variance of the moment conditions in (21), evaluated at the true parameter value μ_0. Hence, we use

$$\mathbf{A} = \left[\left(1 + \frac{1}{T}\right)\mathrm{var}\left(\widehat{\boldsymbol{\gamma}}(\mathbf{N}, \mathbf{Y})\right)\right]^{-1} \tag{22}$$

as our weight matrix. We estimate $\mathbf{A}$ using a bootstrap (e.g., Ackerberg and Gowrisankaran (2006)). Given the complex dependence structure of dyadic observations within each network, we also use a bootstrap to calculate the standard errors of our estimated structural parameters, where we resample networks instead of individuals to address clustering at the network level.

5.3 *The advantages of indirect inference in filling the gap between theory and data*

In Section 4.4, we highlighted the discrepancy between the model and data; in particular, the fact that n_{ij} was a continuous variable while being discrete in the data and that

[13] Gourieroux, Monfort, and Renault (1993) show that the I-I estimator is consistent for a fixed number of simulation draws.

n_{ij} could be equal to zero and mismeasured in the data. The I-I method, which incorporates a simulation procedure, can help us deal with these limitations. First, when we run the simulations to calculate the social capital fixed points and the corresponding social interaction intensity matrix $\mathbf{N}$, we restrict and discretize the values of n_{ij} so that they belong to $\{0, 1, 2, 3, 4, 5\}$; see equation (18).

Second, in the data, students can only answer about their social interactions with their nominated friends. That is, we do not observe student i's social interaction intensity with j if i does not nominate j as a friend. In this unobserved case, we do not impute any arbitrary values and leave it unobserved. In contrast, during the simulation process, as long as i and j are in the same school (network), quite naturally, the simulation method will generate their social interaction intensity; it can take any value. Then we discretize the social interaction intensity of all pairs of students and, following their censusing strategy, we let the value be one of $\{0, 1, 2, 3, 4, 5\}$.

More importantly, when we match the simulated data with the observed ones, we only consider those (directed) pairs of students ij, such that student i nominates j as a friend. In other words, the pair-level sample for the auxiliary dyadic regressions consists of (directed) friendship pairs. We believe that excluding the nonfriend pairs in the auxiliary regressions is appropriate given that we do not observe their social interactions. The reasons are as follows. First, in a large network, no one can be sure of the fact that if two students do not nominate each other as friends, it is because of their preferential choices or a consequence of not knowing each other. Second, if we consider friendship networks as given and apply the same census strategy to both observed and simulated data, the most reasonable strategy is to assign a value of zero to the social interaction intensity of nonfriend pairs both in observed and simulated data sets. In this case, including or excluding those pairs will make no difference to the auxiliary regression results. Third, the friendship network is usually very sparse. For example, when we consider 122 networks (schools), approximately 2% of the directed pairs are friends, and 98% are not. We choose not to include the latter pairs in the auxiliary regression rather than assigning potentially arbitrary values of social interaction intensity between them.

5.4 *Identification*

Our model consists of four main parameters: the baseline social interaction intensity n_0, the social capital accumulation parameter α, the transportation cost c, and the effect of social distance β. Understanding the separate identification of each of these parameters is challenging because our model is nonlinear and our error terms are not additively separable, which is more complicated than a typical model of network externalities, such as the linear-in-means network model (Manski (1993)). Although matching the OLS estimates of the auxiliary model between the observed and simulated data yields reasonable estimates of the structural parameters, it is important to discuss the identification of our model.

To illustrate the separate identification of these four parameters, let us focus on the sources of identification. First, consider β. In the first equation (19) of the auxiliary model, it is straightforward to assume that there is a one-to-one relationship between

γ_{11} and β, as equation (19) closely mimics equation (16) in x_{ij} term. The intercept, or the baseline intensity level, n_0, is similarly identified from its one-to-one relationship with γ_{10}. Next, the cost parameter c is identified given that both equations (19) and (20) contain the term $d_{ij,r}$. Given that the cost parameter is a coefficient on $d_{ij,r}/s_j^*$ in equation (16), having γ_{22} in addition to γ_{12} helps the identification of c.

The most challenging (structural) parameter to identify is α in equation (17). To obtain α, consider the social capital equation (13). Social capital is recursively defined, and hence, it is a function of not only the first degree network connections (or social interactions) but also of higher-degree indirect connections. Therefore, we use the additional equation (20), which uses $[\mathbf{N}_r^2]_{ij}$, the number of second-degree interactions between i and j as a dependent variable, to identify the importance of others' social capital in an agent's social capital formation. The overall fit of two auxiliary equations, measured by two R^2 values will help the identification of the social capital parameter α. The R^2 in equation (19) is closely related to the sum of squared residuals, and hence, helps identify the variance of the idiosyncratic error σ_ε. Since the identification is based on a rather heuristic consideration, we run Monte Carlo simulations to evaluate whether the parameter values are precisely identified and estimated using our proposed empirical method. Online Appendix B shows the details of the Monte Carlo simulations and results and confirm that our method can capture the true parameter values precisely.

6. Structural estimation results

Table 3 reports the estimation results of the key structural parameters and other parameters in the undirectional specification. We also include all sociodemographic characteristics that are related to the intensity of social interactions and social capital. In column (1), we present the results with schools of size up to 150 students, and in column (2) we show the results with schools of size up to 50 students; we also compare whether the results are different when we reduce the network size. For the sake of the exposition, we report the estimates related to social distances and combined levels in two different subcolumns. Note that those two subcolumns of estimates are from the same estimation.

Let us start with the sociodemographic characteristics of the students in column (1) with school size of up to 150 students. Students' preferences exhibit homophily in their own characteristics if the coefficient β_m is negative and significantly different from zero. Table 3 shows that this is the case for most individual characteristics: female, ethnicity, GPA, and religion practice. The estimates are all negative and significant, which supports homophily behavior. When it comes to family background, we find strong homophily behavior in family size, having two parents, family income, and whether they refuse to answer family income. The degree of homophily is the largest in gender.

The estimated coefficients on the $(x_i + x_j)$ variables exhibit mixed signs. Indeed, the intensity of social interactions is increasing if two students are older (i.e., higher grade), if they are female, and nonblack students, if they are more physically developed or practicing religion, or if they are from families from higher parents' education, bigger family size, and more family income. By contrast, the intensity of social interactions is decreasing if the students have a higher GPA or if they have two parents.

TABLE 3. Structural estimation results.

	Undirected Model With Directed n_{ij}			
	(1) Network Size: Up to 150 Students		(2) Network Size: Up to 50 Students	
n_0	1.3504		1.5900	
	(0.0082)		(0.0127)	
α	0.1287		0.1281	
	(0.0014)		(0.0005)	
c	0.2101		0.2088	
	(0.0019)		(0.0016)	
β	$\lvert x_i - x_j \rvert$	$(x_i + x_j)$	$\lvert x_i - x_j \rvert$	$(x_i + x_j)$
Female	−0.9660	0.1833	−0.9943	0.1932
	(0.0068)	(0.0015)	(0.0104)	(0.0013)
Black	−0.2697	−0.0161	−0.6641	−0.0772
	(0.0022)	(0.0001)	(0.0083)	(0.0007)
Grade	0.1081	0.0888	0.2967	0.0829
	(0.0012)	(0.0004)	(0.0021)	(0.0006)
GPA	−0.1636	−0.1406	−0.1114	−0.0703
	(0.0016)	(0.0009)	(0.0009)	(0.0005)
Physical development	0.0317	0.0552	0.0046	0.0641
	(0.0002)	(0.0002)	(0.0001)	(0.0003)
Religious practice	−0.0232	0.0119	−0.1186	0.0334
	(0.0001)	(0.0001)	(0.0007)	(0.0003)
Family size	−0.0659	0.0250	−0.0671	−0.0153
	(0.0003)	(0.0002)	(0.0004)	(0.0003)
Two parents	−0.0770	−0.0675	−0.0076	0.0305
	(0.0006)	(0.0005)	(0.0001)	(0.0003)
Parental education	−0.0050	0.0254	−0.0463	0.0145
	(0.00002)	(0.0001)	(0.0003)	(0.0001)
Family income	−0.0017	0.0024	−0.0016	0.0016
	(0.00002)	(0.00002)	(0.00002)	(0.00001)
Family income refused	−0.1923	0.1532	−0.1224	0.1346
	(0.0023)	(0.0008)	(0.0012)	(0.0010)
σ_ε	1.3488		1.3484	
	(0.0121)		(0.0047)	
Number of networks	122		100	
Number of pupils	4036		3538	
Number of directed pairs	199,892		183,080	
Objective function	16,421.7		8401.7	

Note: We estimate parameters $(n_0, \alpha, c, \beta^{\mathrm{T}})^{\mathrm{T}}$ from equations (12)–(14). We try many starting values to ascertain that a global minimum is attained. Bootstrap standard errors (clustered by networks) in parentheses.

Turning our attention to the structural parameters of the model, we see that they are all statistically significant and have reasonable values. Indeed, the estimated baseline level of social interactions n_0 is 1.35 and α, which measures the importance of others' social capital on an agent's social capital formation, has an estimated value of approximately 0.13. This means that there are positive externalities from peers' social capital. This estimated value of α is in line with standard estimation of network models with pos-

itive externalities in education (see, e.g., Calvó-Armengol, Patacchini, and Zenou (2009), Boucher, Del Bello, Panebianco, Verdier, and Zenou (2023)).[14]

In column (2) of Table 3, we further present the structural estimation results by restricting the size of networks to up to 50 students. We do this for the following reasons. First, to fill the gap between our theoretical model (where all students interact with each other) and the data (where students nominate up to ten friends), we restrict the sample of students in smaller networks. By doing so, we can reduce the number of no-friendship pairs in the observed data. Second, for policy simulations, we have to numerically solve for the social planner's optimization problem. The larger the number of students in a network, the more severe is the curse of dimensionality. Hence, for computational feasibility, we check the robustness of our results to difference choices of network size. By comparing columns (1) and (2) in Table 3, we see that the estimates of our structural parameters as well as the homophily parameters are almost identical across the two samples.

7. Policy analysis

7.1 *Welfare*

We now use the estimated parameters of the model provided in Table 3, that is, α, c, and n_0, to calculate the welfare loss and to perform policy analyses. By extending Lemma 3 to agents' heterogeneity and linear travel cost, we get the following conditions for the optimal choice of interaction and social capital:

$$n^o_{ij,r} = n_0 - \frac{cd_{ij,r}}{s^o_{j,r}} + \frac{\alpha}{N_r}\chi_{i,r}s^o_{j,r} + \theta_{ij,r}, \tag{23}$$

$$\chi_{j,r} = \sum_{i=1,i\neq j}^{N_r}\left\{(n_0+\theta_{ij,r})n^o_{ij,r} - \frac{1}{2}\left(n^o_{ij,r}\right)^2 + \frac{\alpha}{N_r}\chi_{i,r}n^o_{ij,r}\right\}, \tag{24}$$

$$s^o_{j,r} = 1 + \frac{\alpha}{N_r}\sum_{k=1,k\neq j}^{N} n^o_{jk,r}s^o_{k,r}. \tag{25}$$

From the previous estimations of the equilibrium model, we have the estimated values of n_0, α, c, and $\theta_{ij,r}$ (Table 3). From the data, we know $d_{ij,r}$. By plugging these values into (23), (24), and (25), we can solve *numerically* these equations and determine the interaction frequency $n^o_{ij,r}$, for each pair i, j, $s^o_{j,r}$ for all j, $\chi_{i,r}$ for all i, and ultimately the first-best welfare level W^o_r for each network r. For each network r, we have $2N_r + L_r$ unknowns, where L_r is the number of links in network r, and we have $2N_r + L_r$ equations since there are L_r equations for (23), N_r equations for (24), and N_r equations for (25). We then compare the observed equilibrium values of $n^*_{ij,r}$ and $s^*_{j,r}$ with the social optimum values $n^o_{ij,r}$ and $s^o_{j,r}$ (using equations (17) and (25) evaluated at our parameter estimates). According to Proposition 4, we should find that students socially interact too little compared to the social optimal outcome, such that $n^o_{ij,r} > n^*_{ij,r}$, $\forall i$, j, and $s^o_{i,r} > s^*_{i,r}$, $\forall i$.

[14] For an overview of this literature, see Sacerdote (2011).

TABLE 4. Social interactions and social capital: optimal level vs. observed level.

Social Interactions

Optimal Level (SD)	Observed Level	Average Difference (SD)	Minimum Difference [95% CI]	Maximum Difference [95% CI]
4.060 (0.030)	2.818 -	1.242 (0.030)	1.227 [1.279, 1.371]	1.390

Social Capital

Optimal Level (SD)	Observed Level	Average Difference (SD)	Minimum Difference [95% CI]	Maximum Difference [95% CI]
2.048 (0.007)	1.014 -	1.034 (0.007)	1.025 [1.031, 1.050]	1.059

Note: The statistics are computed using the network-level average social interactions and social capital from 100 schools (networks) over 100 simulations. Standard deviations over 100 simulations are in parentheses, and 95% confidence interval (CI) for the differences are in brackets. Note that these statistics differ from pair-level averages.
The observed level of social capital is augmented using equation (15).

We numerically solve the optimal level of social interactions and social capital, using the sample of 100 schools of size up to 50 students and the I-I parameter estimates displayed in column (2) in Table 3, by running a total of 100 simulations. Table 4 displays the results. Note that, in this table, we first take the average of social interactions in each network, and then take the average again over all networks. We find that, on average, each pair interacts 1.24 fewer times than what is socially optimal. The difference between the socially optimal and the observed levels of social interactions varies from 1.23 to 1.39 across networks. Students also have less social capital than the optimal one; they have, on average, 51% less social capital.

7.1.0.0.1 *Network size and social interactions* We would now like to find which variables are closely associated with the discrepancy between the optimal level and the observed level.[15] For that, we regress the differences $\overline{n}_r^o - \overline{n}_r^*$ and $\overline{s}_r^o - \overline{s}_r^*$ on the network size, network measures, and average characteristics (e.g., average family income) of students in each network r:

$$\overline{n}_r^o - \overline{n}_r^* = \gamma_0 + \gamma_1 N_r + \gamma_2 (N_r)^2 + \gamma_3 \overline{d}_r + \gamma_z z_r + \gamma_x x_r + \epsilon_r, \tag{26}$$

$$\overline{s}_r^o - \overline{s}_r^* = \delta_0 + \delta_1 N_r + \delta_2 (N_r)^2 + \delta_3 \overline{d}_r + \delta_z z_r + \delta_x x_r + \zeta_r. \tag{27}$$

Tables 5 and 6 display the results. Consider, first, *social interactions* (Table 5) and let us examine if the difference between the optimal and the observed levels of social interactions, $(\overline{n}_r^o - \overline{n}_r^*)$, is associated with network size N_r or other variables. The coefficients on the network size and its square are insignificant in column (4), but the average

[15]In this subsection and the next one, we do not use any structural estimation methods. We just document some interesting correlations.

TABLE 5. Difference between optimal level and observed level of social interactions.

	Optimal-Observed (Social Interactions)			
	(1)	(2)	(3)	(4)
Network population	0.038	0.034	0.033	−0.003
	(0.027)	(0.028)	(0.029)	(0.030)
Network population squared	−0.0007	−0.0004	−0.0004	0.00001
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Avg. geographic distance		−0.069	−0.069	−0.066
		(0.024)	(0.024)	(0.019)
Avg. degree centrality			−0.042	−0.133
			(0.221)	(0.214)
Std. dev. of degree centrality			0.167	0.650
			(0.375)	(0.349)
Female fraction				0.067
				(0.517)
Black fraction				−0.008
				(0.376)
Avg. student grade				0.230
				(0.039)
Avg. GPA				−0.091
				(0.254)
Avg. level of physical development				−0.061
				(0.190)
Avg. level of religion practice				0.222
				(0.114)
Avg. family size				−0.084
				(0.150)
Fraction of students with two parents				0.546
				(0.562)
Avg. level of parent education				−0.017
				(0.250)
Avg. family income				−0.005
				(0.005)
Fraction family income refused				−0.388
				(0.681)
Constant	0.869	1.206	1.207	−0.343
	(0.342)	(0.398)	(0.443)	(1.399)
Observations (networks)	100	100	100	100
R-squared	0.019	0.156	0.161	0.426

Note: The outcome variable is the average difference between optimal level and observed level of social interactions ($n^o - n^*$) over 100 simulations for each network. Robust standard errors in parentheses.

geographic distance is significantly associated with the inefficiency. A one-kilometer increase in the average pairwise distance lead to a 0.066 decrease in the inefficiency. Only a few average characteristics of the students are associated with the optimal-observed difference in social interactions. In particular, networks that consist of students with a higher average grade (and hence age) or religion practice level are more likely to have high inefficiencies in terms of social interactions.

TABLE 6. Difference between optimal level and observed level of social capital.

	Optimal-Observed (Social Capital)			
	(1)	(2)	(3)	(4)
Network population	0.058	0.055	0.057	0.028
	(0.014)	(0.012)	(0.012)	(0.006)
Network population squared	−0.0009	−0.0007	−0.0007	−0.0003
	(0.0003)	(0.0002)	(0.0002)	(0.0001)
Avg. geographic distance		−0.050	−0.050	−0.048
		(0.010)	(0.010)	(0.005)
Avg. degree centrality			0.063	−0.047
			(0.105)	(0.041)
Std. dev. of degree centrality			−0.365	0.037
			(0.169)	(0.068)
Female fraction				0.279
				(0.088)
Black fraction				−0.163
				(0.064)
Avg. student grade				0.182
				(0.008)
Avg. GPA				−0.151
				(0.053)
Avg. level of physical development				0.116
				(0.037)
Avg. level of religion practice				0.082
				(0.025)
Avg. family size				0.002
				(0.024)
Fraction of students with two parents				0.074
				(0.083)
Avg. level of parent education				0.143
				(0.047)
Avg. family income				0.000
				(0.001)
Fraction family income refused				−0.109
				(0.161)
Constant	0.291	0.536	0.575	−1.450
	(0.151)	(0.153)	(0.170)	(0.247)
Observations (networks)	100	100	100	100
R-squared	0.150	0.349	0.431	0.935

Note: The outcome variable is the difference between optimal level and observed level of social capital ($s^o - s^*$) over 100 simulations for each network. The observed level of social capital is augmented using equation (15). Robust standard errors are in parentheses.

7.1.0.0.2 *Network size and social capital* Let us now turn to the inefficiencies in terms of social capital (Table 6). From the estimates in column (4), we have

$$\frac{\partial\left(\overline{s}_r^o - \overline{s}_r^*\right)}{\partial N_r} = \delta_1 + 2\delta_2 N_r = 0.028 - 2(0.0003)N_r = 0. \tag{28}$$

Solving this equation leads to $N_r = \frac{0.028}{2(0.0003)} = 46.67$. This means that the difference between the optimal and the observed level of social interactions is increasing until the network size reaches (approximately) 46 students and then decreases. As a result, there is a nonmonotonic relationship between $\overline{s}_r^o - \overline{s}_r^*$ and N_r where an increase in the network size increases $\overline{s}_r^o - \overline{s}_r^*$ up to $N_r = 46$ and, above this size, an increase in the network size decreases $\overline{s}_r^o - \overline{s}_r^*$. Thus, $N_r = 46$ is the size of the network that maximizes these inefficiencies.

We also find that the average geographic distance is significantly associated with the inefficiency in social capital. A one-kilometer increase in the average pairwise distance leads to a 0.048 decrease in the inefficiency.

Although these regressions do not have a formal identification strategy, the results, partly based on the structural estimation of the model (that determine $\overline{n}_r^o - \overline{n}_r^*$ and $\overline{s}_r^o - \overline{s}_r^*$), provide some interesting explanations on what drives the size of inefficiency of the intensity of social interactions and social capital accumulation.

7.1.0.0.3 *Network size and average welfare* Another interesting exercise, for which we do not have a theory, is to determine the optimal network, that is, the one that maximizes total welfare.[16] For that, without any policy, we compare the average welfare (to avoid size effects, the welfare is not defined as the sum of utilities but as the average utility) in each of the 100 networks. Remember that the welfare in network r is given by

$$W_r^* = \sum_{i=1}^{N_r} \sum_{j=1, j\neq i}^{N_r} \left[\left((n_0 + \theta_{ij,r}) n_{ij,r}^* - \frac{1}{2} \left(n_{ij,r}^* \right)^2 \right) s_{j,r}^* - n_{ij,r}^* c d_{ij,r} \right]. \tag{29}$$

As a result, the average welfare per network is

$$AW_r^* = \frac{W_r^*}{N_r}.$$

We would like know which network size N_r yields the largest AW_r^*.

For that, we run the following regression:

$$AW_r^* = \delta_0 + \delta_1 N_r + \delta_2 (N_r)^2 + \delta_z z_r + \delta_x x_r + \epsilon_r$$

to investigate the relationship between average welfare and network size. In addition, as controls, we include the average geographical distance and network measures, such as mean and standard deviation of the degree distribution, average eigenvector centrality.

Table 7 reports the results. We can first calculate the network size that maximizes the average welfare per network AW_r^*. Using column (4), we have

$$\frac{\partial AW_r^*}{\partial N_r} = \delta_1 + 2\delta_2 N_r = 0.113 - 2(0.0021) N_r = 0. \tag{30}$$

[16] Determining the optimal network is a very difficult exercise; see König, Tessone, and Zenou (2014), Belhaj, Bervoets, and Deroïan (2016), and Chen, Zenou, and Zhou (2022) for such attempts when the network is given. Jackson and Wolinsky (1996) provide a similar exercise for endogenous network formation. Because this exercise is complicated, only extreme structures emerge such as the complete network, the star network or nested split graphs. This is why we do it here by numerical simulations based on the estimated parameters.

TABLE 7. Optimal network design: average welfare and number of students.

	(1) Welfare	(2) Welfare	(3) Welfare	(4) Welfare
Network population	0.126	0.105	0.163	0.113
	(0.064)	(0.040)	(0.033)	(0.030)
Network population squared	−0.0027	−0.0013	−0.0026	−0.0021
	(0.0013)	(0.0008)	(0.0006)	(0.0005)
Avg. geographic distance		−0.352	−0.346	−0.343
		(0.034)	(0.032)	(0.024)
Avg. degree centrality			1.366	1.080
			(0.291)	(0.294)
Std. dev. of degree centrality			−1.220	−0.393
			(0.457)	(0.424)
Sociodemographic controls	No	No	No	Yes
Observations (networks)	100	100	100	100
R-squared	0.070	0.653	0.725	0.841

Note: The outcome variable is the simulated average welfare (AW_r), averaged over 100 simulations for each network. Sociodemographic control variables: female fraction, black fraction, average student grade, average GPA, the average level of physical development, the average level of religion practice, average family size, the fraction of students with two parents, average level of parent education, average family income, and the fraction of students who refuse to answer family income. Robust standard errors are in parentheses.

This means that the network that comprises (approximately) 27 students is the one that maximizes the average welfare per network.

In Table 7, we also find that the average pairwise geographic distance is an important factor for designing an optimal network. The longer is the distance between two students, the lower is the average welfare. In addition, from the changes in R^2 across columns (1) and (2), from 0.070 to 0.653, we find that the average geographic distance explains a significant proportion of the average welfare in a network.

7.2 *Policies*

We have seen in Proposition 5 that social optimal allocations can be restored with appropriate subsidies on student's travel cost or interactions. However, such subsidies are unlikely to be implemented because they depend on detailed information about every interaction pair such as destination and origin of interaction partners, which does not only imply a strong problem of information collection but also an issue of equity between the recipients of unequal subsidies.

In this section, we consider the more realistic case of *uniform subsidies* on social interactions and/or travel costs that only target each individual irrespective of their personal characteristics but not a pair of individuals. We evaluate their impact on the frequency of interactions, n_{ij} by running a total of 100 simulation exercises using the sample of 100 schools of size up to 50 students. We provide the average, the sample standard deviations, and/or 95% confidence intervals for each policy question from these 100 simulations. Which policy is more effective at moving the observed interactions/social capital closer to the optimal levels?

Assume that each individual receives a common subsidy σ for each interaction made with a friend and a (percentage) subsidy τ on her transport cost c. The total amount of each subsidy received by an individual i is therefore given by $\sum_j \sigma n_{ij}$ for social interactions and $\sum_j n_{ij}\tau c d_{ij}$ for transportation costs.

Note that the government (or the planner) is here introduced as an agent that can set subsidies on social-interaction efforts before the individuals decide upon their efforts. The assumption that the government can precommit itself to such subsidies, and thus can act in this leadership role is fairly natural. As a result, this subsidy will affect the levels of social interaction efforts of all individuals.[17]

For each individual i interacting with j, when subsidies are included, the equilibrium conditions lead to the following level of social interactions:

$$n_{ij}^* = \left(n_0 - \frac{cd_{ij}}{s_j^*} + \theta_{ij}\right) + \frac{\sigma}{s_j^*} + \frac{\tau c d_{ij}}{s_j^*},$$

while the social capital is still given by

$$s_j^* = 1 + \frac{\alpha}{N}\sum_{l\neq j} n_{jl}^* s_l^*.$$

Holding social capital constant, quite naturally, the subsidies increase the number of social interactions. Subsidies can entice interactions with new partners as the number of interactions to a partner may rise from zero to a positive value in the presence of the subsidy. The total welfare is now defined as

$$W = \sum_i \sum_{j\neq i}\left((n_0+\theta_{ij})n_{ij}^* - \frac{1}{2}\left(n_{ij}^*\right)^2\right)s_j^* - n_{ij}^* c d_{ij} + \sum_i \sum_{j\neq i} n_{ij}^*(\sigma + \tau c d_{ij}).$$

We now implement two uniform-subsidy policies (first, we subsidize social interactions and then transportation costs) whose aim is to find the subsidy that achieves the same welfare level as the level obtained at the first best.

7.2.1 *Subsidizing social interactions* We consider a *uniform* subsidy σ_r for each social interaction in network r. We use the following discrete version of the equilibrium identities:

$$n_{ij,r}^{\sigma} = n_0 + \frac{\sigma_r - c d_{ij,r}}{s_{j,r}^{\sigma}} + \theta_{ij,r} \tag{31}$$

and

$$s_{j,r}^{\sigma} = 1 + \frac{\alpha}{N_r}\sum_{k=1}^{N_r} n_{jk,r}^{\sigma} s_{k,r}^{\sigma}, \tag{32}$$

[17]This is is similar to the standard policy of firms' subsidies on R&D efforts; see, for example, Spencer and Brander (1983) and König, Liu, and Zenou (2019).

TABLE 8. Policy levels for optimal outcomes.

(1) Subsidizing Social Interactions: σ			(2) Subsidizing Transportation Costs: τ		
Average	Minimum	Maximum	Average	Minimum	Maximum
(SD)	[95% CI]		(SD)	[95% CI]	
3.009	1.325	7.224	0.635	0.532	0.744
(0.901)	[2.099, 4.660]		(0.046)	[0.553, 0.702]	

Note: The subsidy level for each network is computed for students in each network to obtain the optimal level of social interactions and social capital in (23)–(25).
We report the average results over 100 simulations and confidence intervals among 100 schools.

where the superscript σ denotes the subsidy policy outcome. For the estimation, the total welfare per network is equal to

$$W_r^\sigma = \sum_{i=1}^{N_r} \sum_{j=1, j\neq i}^{N_r} \left[\left((n_0 + \theta_{ij,r}) n_{ij,r}^\sigma - \frac{1}{2} \left(n_{ij,r}^\sigma \right)^2 \right) s_{j,r}^\sigma - (c d_{ij,r} - \sigma_r) n_{ij,r}^\sigma \right]. \tag{33}$$

In this exercise, we determine the subsidy σ_r^* that gives network r the same aggregate welfare W_r^σ as its first best level W_r^o. From the estimated value of the equilibrium model, we have α, c, and n_0; from the data we have $d_{ij,r}$ and N_r. We then numerically solve equations (31) and (32) and find the subsidy such that $W_r^\sigma = W_r^o$; see Online Appendix C for technical details.

The first three columns in Table 8 display the results. On average, a subsidy level of 3.009 (units of utility) for each social interaction is required for a network to achieve the first-best aggregate level of social interactions and social capital.

7.2.2 *Subsidizing transportation costs* In the case of subsidies on transport cost, we consider the following equilibrium conditions:

$$n_{ij,r}^\tau = n_0 - \frac{(1 - \tau_r) c d_{ij,r}}{s_{j,r}^\tau} + \theta_{ij,r}, \tag{34}$$

$$s_{j,r}^\tau = 1 + \frac{\alpha}{N_r} \sum_{k=1, k\neq j}^{N_r} n_{jk,r}^\tau s_{k,r}^\tau. \tag{35}$$

The total welfare per network is defined as

$$W_r^\tau = \sum_{i=1}^{N_r} \sum_{j=1, j\neq i}^{N_r} \left[\left((n_0 + \theta_{ij,r}) n_{ij,r}^\tau - \frac{1}{2} \left(n_{ij,r}^\tau \right)^2 \right) s_{j,r}^\tau - n_{ij,r}^\tau (1 - \tau_r) c d_{ij,r} \right]. \tag{36}$$

As for the social interaction subsidy, we find the subsidy τ_r^* that gives the same aggregate utility W_r^τ in network r as the first-best W_r^0. From the estimated value of the equilibrium model, we have α, c, and $n_{0,r}$, and from the data $d_{ij,r}$ and b_r. We can then numerically solve equations (34) and (35) and find the subsidy such that $W_r^\tau = W_r^0$.

The last three columns in Table 8 display the results. On average, a subsidy level of $\tau = 0.635$ (63.5% of travel cost) is required for a network to achieve the first best aggregate level of social interactions and social capital. From this result, we can also infer that a decrease in a geographical distance between two students with different socioeconomic backgrounds would increase their levels of social interactions and social capital.

7.2.3 *Comparing the two policies* In the two above exercises, subsidy policies are given at no social cost by the planner. It is then interesting to compare these two policies at the *same given cost*. The question is then as follows: Given that the planner has a budget of B to spend, which policy should she choose? In order to distribute a total amount of subsidy B to each network, we consider three different schemes. First, we distribute the same amount $B_r = B/R$ for each network r (uniform subsidy), where R is the total number of networks. The second scheme gives an amount proportional to network population N_r. Hence, $B_r = \frac{N_r}{\sum_{r'} N_{r'}} B$. The last subsidy scheme provides an amount proportional to the number of pairs $N_r(N_r - 1)$, that is, $B_r = \frac{N_r(N_r-1)}{\sum_{r'} N_{r'}(N_{r'}-1)} B$.

We also need to set the total budget B to a level that is comparable to the subsidy budget spent in the two above exercises. We consider two ways of setting this budget. First, we choose the amount of budget that corresponds to the average social interaction subsidy level that achieves the first-best level of social interactions:

$$B := B^{\sigma} = \bar{\sigma}^{o} \bar{n}^{o} \sum_{r=1}^{R} N_r(N_r - 1), \tag{37}$$

where $\bar{\sigma}^{o}$ is the average optimal social interaction subsidy level, as obtained in Table 8, that is, $\bar{\sigma}^{o} = 3.009$, and $\bar{n}^{o}$ is the average optimal social interaction level, as obtained in Table 4, that is, $\bar{n}^{o} = 4.060$.

Second, we use the amount of budget that corresponds to the average transportation subsidy level to achieve the first-best level of social interactions:

$$B := B^{\tau} = \bar{\tau}^{o} c \bar{n}^{o} \sum_{r=1}^{R} N_r(N_r - 1), \tag{38}$$

where $\bar{\tau}^{o}$ is the average transportation subsidy rate, that is, $\bar{\tau}^{o} = 0.635$ (Table 8).

We proceed as follows. First, we consider the *social-interaction subsidy policy*. We observe $d_{ij,r}$ and N_r in the data and have estimated α, c, and n_0. Then we solve simultaneously equations (31), (32), and (37). We get the different endogenous variables, in particular, the different subsidies σ_r. Then, for each value of σ_r, we calculate the total welfare W_r^{σ} given by (33). Second, we consider the *transportation subsidy policy*. We observe $d_{ij,r}$ and N_r in the data and have estimated α, c, and n_0. Then we solve simultaneously equations (34), (35), and again (37). We obtain the endogenous variables, in particular, the different subsidies τ_r. Then, for each value of τ_r, we calculate the total welfare W_r^{τ} given by (36). We finally repeat these two steps with the budget B^{τ} given by (38).

Our key question is then about which subsidy on travel costs or social interactions yields the highest welfare in each network for either budget B^{σ} or B^{τ}. That is, we examine

TABLE 9. Comparison of two policies.

Panel A: Budget Corresponding to the Average (Optimal) Social Interaction Subsidy Level

Subsidy Schemes	Number of Networks With Higher Welfare for Each Policy [95% CI]		Difference in Average Welfare [95% CI]
	Policy: σ	Policy: τ	Policy σ − Policy τ
(1) Uniform subsidy amount for each network	97	1	293.65
	[97, 98]	[0, 1]	[270.56, 323.62]
(2) Subsidy proportional to N_r	97	1	271.46
	[97, 97]	[1, 1]	[241.80, 301.02]
(3) Subsidy proportional to $N_r(N_r-1)$	97	1	251.03
	[97, 98]	[0, 1]	[209.66, 325.83]

Panel B: Budget Corresponding to the Average (Optimal) Transportation Subsidy Level

Subsidy Schemes	Number of Networks With Higher Welfare for Each Policy [95% CI]		Difference in Average Welfare [95% CI]
	Policy: σ	Policy: τ	Policy σ − Policy τ
(1) Uniform subsidy amount for each network	96	2	71.70
	[96, 97]	[1, 2]	[56.22, 74.38]
(2) Subsidy proportional to N_r	97	1	67.46
	[97, 97]	[1, 1]	[49.65, 76.47]
(3) Subsidy proportional to $N_r(N_r-1)$	97	1	62.43
	[97, 98]	[0, 1]	[42.70, 78.30]

Note: The median number of networks over 100 simulations, which leads to higher welfare for each policy is reported, along with the 95% confidence interval among 100 schools (networks). There are a couple of schools in which two policies are tied in their total welfare. The term "Policy σ − Policy τ" indicates the average welfare after the social interaction subsidy policy (policy σ) minus the average welfare after the transportation subsidy policy (policy τ).

whether $W_r^\tau \gtreqless W_r^\sigma$. Table 9 shows the results of this analysis by counting the number of networks for which the total welfare is higher under one policy versus the other. In this table, we find that, under the social-interaction subsidy policy, the total welfare is higher for most networks, regardless of the amount of budget we assign (panels A and B) and the type of subsidy scheme (uniform, proportional to N_r and proportional to $N_r(N_r-1)$; rows (1), (2), and (3)).[18] As a result, if a planner has a given amount of money to spend, she should subsidize social interactions and not transportation costs because it yields greater improvements of total welfare.

8. CONCLUDING REMARKS

In this paper, we presented a behavioral microfoundation for the relationship between geographical distance, social interactions, and social capital. We characterized the equi-

[18]We also try different values of the total amount to be spent to check whether there are nonlinear effects, but the results remain the same regardless of the value of the budget.

librium in terms of levels of social interactions and social capital for a general distribution of individuals in the geographical space. An important prediction of the model was that the level of social interactions was inversely related to the geographical distance. Travel costs and spatial dispersion of agents were barriers to the development of social capital formation. Because of the externalities that agents exerted on each other, we demonstrated that the equilibrium levels of social interactions and social capital were lower than the efficient ones.

When we estimated the model using data on adolescents in the United States, we found that, indeed, geographical distance was an hinder to social interactions. Moreover, we determined the exact inefficiencies of the market equilibrium. Interestingly and surprisingly, we found that there was a nonmonotonic relationship between the inefficiencies in terms of social interactions and the network size. In our empirical context, these inefficiencies were the largest when the network was composed of ten students. We then performed two different subsidy policies and show that uniform subsidies on social interactions were more effective than uniform subsidies on transportation costs.

Extrapolating those results to social interactions in city suggests that encouraging social interactions is likely to enhance social welfare. In the real world, there are different ways governments can subsidize social interactions. One natural way is *social mixing* such as the Moving to Opportunity (MTO) programs in the United States where the local government subsidizes housing to allow families to move from poor to richer neighborhoods (see, e.g., Katz, Kling, and Liebman (2001), Kling, Liebman, and Katz (2007), Chetty, Hendren, and Katz (2016)). These programs allow people from different neighborhoods to interact with each other. Other policies that enhance social interactions are those that improve physical environment such as zoning laws and public housing rules (Glaeser and Sacerdote (2000)). For example, Glaeser and Sacerdote (2000) find that individuals in large apartment buildings are more likely to socialize with their neighbors than those living in smaller apartment buildings. Using Facebook data from the United States, Bailey, Cao, Kuchler, and Stroebel (2018a) document that, at the county level, friendship networks are a mechanism that can propagate house price shocks through the economy via housing price expectations.

This paper is a first stab at a complex problem. We hope that most research will be conducted in the future on the interaction between the social and the geographical space.

Appendix: Proofs of the theoretical results

Proof of Proposition 1. The equilibrium number of interactions n_{ij}^* of student i with student j, is found by differentiating U_i with respect to n_{ij} taking s_j as given. We obtain

$$v'(n_{ij})s_j - c(d_{ij}) = 0, \quad j = 1, \ldots, N. \tag{39}$$

Using (2), this is equivalent to $(1 - n_{ij})s_j = c(d_{ij})$. Thus, the equilibrium number of interactions is equal to

$$n_{ij}^* = 1 - \frac{c(d_{ij})}{s_j}, \quad j = 1, \ldots, N. \tag{40}$$

For simplicity, we assume away corner solutions and assume *global interactions*, so that students agents interact with every other student in the network, that is,

$$n_{ij}^* > 0 \quad \Leftrightarrow \quad s_j > c(d_{ij}), \quad \forall i, j.$$

A sufficient condition for this inequality to hold is

$$\min_j s_j > c(\overline{d}), \tag{41}$$

where $\overline{d}$ is the maximum distance between two agents in the network.

By plugging (40) into (3) and using (4), we obtain the equilibrium level of social capital s_j^*. It is given by

$$s_j^* = 1 + \frac{\alpha}{N}\sum_{k\neq j} s_k^* - \frac{\alpha}{N} g_j. \tag{42}$$

To solve for the fixed-point solution of this equation, we sum over j on both sides and simplify as

$$\sum_j s_j^* = \frac{1}{1-\alpha\left(\dfrac{N-1}{N}\right)}\left[N - \frac{\alpha}{N}\sum_j g_j\right], \tag{43}$$

since $\frac{\alpha}{N}\sum_j\sum_{k\neq j} s_k^* = \frac{\alpha}{N}\sum_{k\neq j}\sum_j s_j^* = \frac{\alpha(N-1)}{N}\sum_j s_j^*$. Inserting (43) into (42) yields the following closed-form solution for the equilibrium social capital:

$$s_j^* = s_0 - \frac{\alpha/N}{1+\alpha/N} g_j. \tag{44}$$

Let us show that the global interaction condition (41) is satisfied if $c(\overline{d}) < N$ and $\alpha < 1$. Indeed, using $g_j < (N-1)c(\overline{d})$ and $\alpha < 1$, the global interaction condition $\min_j s_j > c(\overline{d})$ is satisfied if

$$c(\overline{d}) < N\frac{1-\alpha(1-2/N)+\alpha^2(1-1/N)^2}{1-\alpha+2\alpha/N}.$$

It can be shown that the ratio in the right-hand side (RHS) is larger than one. So, a sufficient condition for global interaction is that $c(\overline{d}) < N$. □

Proof of Proposition 2. We demonstrate that the importance of peers' social links, increases each agent's social capital for small enough travel cost. We need to compute

$$\frac{\mathrm{d}s_0}{\mathrm{d}\alpha} = \frac{f(\alpha) - \alpha(2-\alpha)\dfrac{1}{N}\sum_l g_l}{N\left(1-\dfrac{\alpha}{N}\right)^2\left(1-\alpha+\dfrac{\alpha}{N}\right)^2},$$

where $f(\alpha) = (1+\frac{\alpha}{N})^2 + N[1-2(\frac{\alpha}{N}) - (\frac{\alpha}{N})^2]$. It can be shown that $f'(\alpha) = -2(N+\alpha)\frac{N-1}{N^2} < 0$ so that $f(\alpha) \geq f(0) = 1+N \geq 3$. So, when travel costs $c(\cdot)$ tend to zero, g_l

and $\sum_l g_l$ also tend to zero while $\mathrm{d}s_0/\mathrm{d}\alpha$ is bounded above zero. So, $\mathrm{d}s_j^*/\mathrm{d}\alpha > 0$ for small enough travel costs $c(d_{ij})$. □

Proof of Lemma 3. The government chooses the profiles n_{ij} and s_j that maximize the Lagrangian function

$$\mathcal{L} = \sum_i \sum_{j\neq i} \big[\big(v(n_{ij})s_j - n_{ij}c(d_{ij})\big)\big] - \sum_i \chi_i \bigg(s_i - 1 - \frac{\alpha}{N}\sum_{j\neq i} n_{ij}s_j\bigg),$$

where $\chi_i \geq 0$ is the Kuhn–Tucker multiplier of the social capital constraint. Thus, χ_i measures the welfare value of a marginal increase of the social capital of agent i.

We can write the Lagrangian function as

$$\mathcal{L} = \sum_i \sum_{j\neq i} \big[v(n_{ij})s_j - n_{ij}c(d_{ij}) + (\alpha/N)\chi_i n_{ij}s_j\big] - \sum_i \chi_i(s_i - 1).$$

Note that $\sum_i \chi_i(s_i - 1)$ evaluates to the same value as $\sum_i \sum_{j\neq i} \chi_j(s_j - 1)/(N-1)$. Substituting the latter for the former, we rewrite the Lagrangian function as

$$\mathcal{L} = \sum_i \sum_{j\neq i} v(n_{ij})s_j - n_{ij}c(d_{ij}) + (\alpha/N)\chi_i n_{ij}s_j - \chi_j(s_j - 1)/(N-1). \tag{45}$$

First-order conditions with respect to n_{ij} and s_j yield

$$v'(n_{ij})s_j - c(d_{ij}) + (\alpha/N)\chi_i s_j = 0$$
$$\sum_{i\neq j} \big[v(n_{ij}) + (\alpha/N)\chi_i n_{ij} - \chi_j/(N-1)\big] = 0.$$

The last equality is equivalent to

$$\sum_{i\neq j} \big[v(n_{ij}) + (\alpha/N)\chi_i n_{ij}\big] - \chi_j = 0.$$

This gives (9) and (10). □

Proof of Proposition 4. Condition (9) yields

$$v'(n_{ij}) = \frac{c(d_{ij})}{s_j} - \frac{\alpha}{N}\chi_i, \tag{46}$$

which gives

$$n_{ij}^o = 1 - \frac{c(d_{ij})}{s_j^o} + \frac{\alpha}{N}\chi_i^o, \tag{47}$$

under our specification of utility function v. With social capital held fixed at j at the equilibrium level ($s_j^* = s_j^o$), this expression is larger than the equilibrium number of visits n_{ij}^*

because $\chi_i^o \geq 0$. The question thus becomes how social capital changes in this efficient allocation.

By inserting (7) in the binding condition (8), we obtain

$$s_i^o = 1 + \frac{\alpha}{N}\sum_{l\neq i} s_l^o - \frac{\alpha}{N} g_i + \left(\frac{\alpha}{N}\right)^2 \chi_i^o \sum_{l\neq i} s_l^o. \tag{48}$$

Observe that, for $\chi_i^o = 0$, (47) and (48) are identical to the equilibrium conditions and therefore yield the equilibrium values n_{ij}^* and s_i^*. The RHS of (47) and (48) are increasing functions of χ_i^o and/or s_i^o. From (48), we see that an increase in χ_i^o above zero raises s_i^o. From (47), the joint increase in χ_i^o and s_i^o raises n_{ij}^o. So, we conclude that $n_{ij}^o \geq n_{ij}^*$ and $s_i^o \geq s_i^*$. □

Proof of Proposition 5. If we include the subsidies τ_{ij} and σ_{ij}, the utility becomes

$$\begin{aligned} U_i &= S_i - C_i \\ &= \sum_j \big\{ v(n_{ij})(s_j + \sigma_{ij}) - n_{ij}\big[c(d_{ij}) - \tau_{ij}\big]\big\}. \end{aligned}$$

This implies the following equilibrium number of social interactions:

$$n_{ij}^* = 1 - \frac{c(d_{ij}) - \tau_{ij}}{s_j + \sigma_{ij}}.$$

The social capital level is then given by the following fixed point:

$$\begin{aligned} s_j^* &= 1 + \frac{\alpha}{N}\sum_{k\neq j} n_{jk}^* s_k^* \\ &= 1 + \frac{\alpha}{N}\sum_{k\neq j}\left(1 - \frac{c(d_{jk}) - \tau_{jk}}{s_k^* + \sigma_{jk}}\right) s_k^*. \end{aligned} \tag{49}$$

The frequency of social interactions and the level of social capital are the same in equilibrium and in the first best if and only if

$$n_{ij}^* = n_{ij}^o \quad \Longleftrightarrow \quad \frac{c(d_{ij}) - \tau_{ij}}{s_j^* + \sigma_{ij}} = \frac{c(d_{ij})}{s_j^o} - \frac{\alpha}{N}\chi_i^o, \tag{50}$$

and $s_j^* = s_j^o$ given by (49) and (48).

The first best can be decentralized with the subsidies $\sigma_{ij} = 0$ and $\tau_{ij} = (\alpha/N)\chi_i^o s_j^o$. Indeed, in this case, we find

$$n_{ij}^* = 1 - c(d_{ij})/s_j^o + (\alpha/N)\chi_i^o = n_{ij}^o.$$

Given that $n_{ij}^* = n_{ij}^o$, it is straightforward to see that $s_j^* = s_j^o$.

The first best can also be decentralized with the subsidies $\tau_{ij} = 0$ and

$$\sigma_{ij} = \frac{s_j^o}{\frac{Nc(d_{ij})}{\alpha\chi_i^o s_j^o} - 1}. \tag{51}$$

This gives the interaction frequency

$$n_{ij}^* = 1 - \frac{c(d_{ij})}{s_j^* + \frac{1}{\frac{1}{s_j^o} - \frac{\alpha}{N}\frac{\chi_i^o}{c(d_{ij})}} - s_j^o}$$

and the social capital fixed point

$$s_j^* = 1 + \frac{\alpha}{N}\sum_{k\neq j} s_k^* - \frac{\alpha}{N}\sum_{k\neq j} \frac{c(d_{jk})}{s_k^* + \frac{1}{\frac{1}{s_k^o} - \frac{\alpha}{N}\frac{\chi_j^o}{c(d_{jk})}} - s_k^o} s_k^*.$$

Yet, the solution $s_j^* = s_k^o$ is a fixed point of the latter expression as it gives the fixed point for the following first-best social capital formation

$$s_j^o = 1 + \frac{\alpha}{N}\sum_{k\neq j} s_k^o - \frac{\alpha}{N}\sum_{k\neq j} c(d_{jk}) + \left(\frac{\alpha}{N}\right)^2 \sum_{k\neq j} \chi_j^o s_k^o.$$

Importantly, the subsidy τ_{ij} and σ_{ij} are not uniform ones. This suggests that decentralization would be difficult to implement.

How to interpret σ_{ij}? Suppose that the denominator is positive, so that the subsidy is a positive transfer for holding a social partner. We have

$$\sigma_{ij} = \frac{s_j^o}{\frac{Nc(d_{ij})}{\alpha\chi_i^o s_j^o} - 1} > 0.$$

Hence, we need to subsidize more partnership with recipient individuals j with more social capital and initiator individuals i with higher welfare value of a marginal increase of the social capital and smaller distances.

Suppose the above denominator is negative so that σ_{ij} is a tax:

$$\text{tax} = -\sigma_{ij} = \frac{s_j^o}{1 - \frac{Nc(d_{ij})}{\alpha\chi_i^o s_j^o}} > 0.$$

Hence, we need to tax less partnership from initiator individuals i with higher welfare value of a marginal increase of the social capital and smaller distances. □

References

Ackerberg, Daniel A. and Gautam Gowrisankaran (2006), "Quantifying equilibrium network externalities in the ACH banking industry." *The RAND Journal of Economics*, 37 (3), 738–761. [1312]

Arzaghi, Mohammad and J. Vernon Henderson (2008), "Networking off Madison avenue." *The Review of Economic Studies*, 75 (4), 1011–1038. [1297]

Bailey, Michael, Ruiqing Rachel Cao, Theresa Kuchler, and Johannes Stroebel (2018a), "The economic effects of social networks: Evidence from the housing market." *Journal of Political Economy*, 126 (6), 2224–2276. [1326]

Bailey, Michael, Ruiqing Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong (2018b), "Social connectedness: Measurement, determinants, and effects." *Journal of Economic Perspectives*, 32 (3), 259–280. [1297]

Bailey, Michael, Patrick Farrell, Theresa Kuchler, and Johannes Stroebel (2020), "Social connectedness in urban areas." *Journal of Urban Economics*, 118, 103264. [1298]

Barthélemy, Marc (2011), "Spatial networks." *Physics Reports*, 499 (1–3), 1–101. [1298]

Barwick, Panle Jia, Yanyan Liu, Eleonora Patacchini, and Qi Wu (2023), "Information, mobile communication, and referral effects." *American Economic Review*, 113 (5), 1170–1207. [1298]

Bayer, Patrick, Stephen L. Ross, and Giorgio Topa (2008), "Place of work and place of residence: Informal hiring networks and labor market outcomes." *Journal of Political Economy*, 116 (6), 1150–1196. [1297]

Belhaj, Mohamed, Sebastian Bervoets, and Frédéric Deroïan (2016), "Efficient networks in games with local complementarities." *Theoretical Economics*, 11 (1), 357–380. [1320]

Bisztray, Marta, Miklós Koren, and Adam Szeidl (2018), "Learning to import from your peers." *Journal of International Economics*, 115, 242–258. [1297]

Boucher, Vincent, Carlo Del Bello, Fabrizio Panebianco, Thierry Verdier, and Yves Zenou (2023), "Education transmission and network formation." *Journal of Labor Economics*, 41 (1), 129–173. [1316]

Bramoullé, Yann, Brian W. Rogers, and Andrea Galeotti (2016), *The Oxford Handbook of the Economics of Networks*. Oxford University Press. [1295]

Brueckner, Jan K. and Ann G. Largey (2008), "Social interaction and urban sprawl." *Journal of Urban Economics*, 64 (1), 18–34. [1297, 1302]

Büchel, Konstantin and Maximilian von Ehrlich (2020), "Cities and the structure of social interactions: Evidence from mobile phone data." *Journal of Urban Economics*, 119, 103276. [1297]

Cabrales, Antonio, Antoni Calvó-Armengol, and Yves Zenou (2011), "Social interactions and spillovers." *Games and Economic Behavior*, 72 (2), 339–360. [1299]

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou (2009), "Peer effects and social networks in education." *The Review of Economic Studies*, 76 (4), 1239–1267. [1306, 1316]

Chen, Ying-Ju, Yves Zenou, and Junjie Zhou (2022), "The impact of network topology and market structure on pricing." *Journal of Economic Theory*, 204, 105491. [1320]

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz (2016), "The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment." *The American Economic Review*, 106 (4), 855–902. [1326]

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin (2009), "An economic model of friendship: Homophily, minorities, and segregation." *Econometrica*, 77 (4), 1003–1045. [1309, 1310]

Fafchamps, Marcel and Flore Gubert (2007), "Risk sharing and network formation." *The American Economic Review*, 97 (2), 75–79. [1297, 1310]

Fu, Chao and Jesse Gregory (2019), "Estimation of an equilibrium model with externalities: Post-disaster neighborhood rebuilding." *Econometrica*, 87 (2), 387–421. [1296]

Gallant, A. Ronald and George Tauchen (1996), "Which moments to match?" *Econometric Theory*, 12, 657–681. [1310]

Glaeser, Edward L. and Bruce Sacerdote (2000), "The social consequences of housing." *Journal of Housing Economics*, 9, 1–23. [1326]

Goldenberg, Jacob and Moshe Levy (2009), "Distance is not dead: Social interaction and geographical distance in the Internet era." arXiv preprint, arXiv:0906.3202. [1298]

Gourieroux, Christian, Alain Monfort, and Eric Renault (1993), "Indirect inference." *Journal of Applied Econometrics*, 8, S85–S118. [1310, 1312]

Gourieroux, Christian and Alain Monfort (1996), *Simulation-Based Econometric Methods*. Oxford University Press, Oxford. [1310]

Graham, Bryan S. (2017), "An econometric model of network formation with degree heterogeneity." *Econometrica*, 85 (4), 1033–1063. [1310]

Hellerstein, Judith K., Mark J. Kutzbach, and David Neumark (2014), "Do labor market networks have an important spatial dimension?" *Journal of Urban Economics*, 79, 39–58. [1297]

Hellerstein, Judith K., Melissa McInerney, and David Neumark (2011), "Neighbors and coworkers: The importance of residential labor market networks." *Journal of Labor Economics*, 29 (4), 659–695. [1297]

Helsley, Robert W. and William C. Strange (2007), "Urban interactions and spatial structure." *Journal of Economic Geography*, 7 (2), 119–138. [1297]

Helsley, Robert W. and Yves Zenou (2014), "Social networks and interactions in cities." *Journal of Economic Theory*, 150, 426–466. [1297, 1303]

Ioannides, Yannis M. (2013), *From Neighborhoods to Nations: The Economics of Social Interactions*. Princeton University Press, Princeton. [1295, 1297]

Jackson, Matthew O. (2008), *Social and Economic Networks*. Princeton University Press, Princeton. [1295, 1296, 1297]

Jackson, Matthew O. and Brian W. Rogers (2005), "The economics of small worlds." *Journal of the European Economic Association*, 3, 617–627. [1296, 1297]

Jackson, Matthew O., Brian W. Rogers, and Yves Zenou (2017), "The economic consequences of social network structure." *Journal of Economic Literature*, 55 (1), 1–47. [1295]

Jackson, Matthew O. and Asher Wolinsky (1996), "A strategic model of social and economic networks." *Journal of Economic Theory*, 71 (1), 44–74. [1297, 1320]

Jackson, Matthew O. and Yves Zenou (2015), "Games on networks." In *Handbook of Game Theory*, Vol. 4 (Petyon Young and Shmuel Zamir, eds.), 91–157, Elsevier, Amsterdam. [1295]

Johnson, Cathleen and Robert P. Gilles (2000), "Spatial social networks." *Review of Economic Design*, 5 (3), 273–299. [1296, 1297]

Kaltenbrunner, Andreas, Salvatore Scellato, Yana Volkovich, David Laniado, Dave Currie, Erik J. Jutemar, and Cecilia Mascolo (2012), "Far from the eyes, close on the web: Impact of geographic distance on online social interactions." In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, 19–24. [1298]

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman (2001), "Moving to opportunity in Boston: Early results of a randomized mobility experiment." *Quarterly Journal of Economics*, 116, 607–654. [1326]

Kim, Jun Sung, Eleonora Patacchini, Pierre M. Picard, and Yves Zenou (2023), "Supplement to 'Spatial interactions'." *Quantitative Economics Supplemental Material*, 14, https://doi.org/10.3982/QE1720. [1298]

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz (2007), "Experimental analysis of neighborhood effects." *Econometrica*, 75 (1), 83–119. [1326]

König, Michael, Xiaodong Liu, and Yves Zenou (2019), "R&D networks: Theory, empirics and policy implications." *The Review of Economics and Statistics*, 101 (3), 476–491. [1322]

König, Michael, Claudio Tessone, and Yves Zenou (2014), "Nestedness in networks: A theoretical model and some applications." *Theoretical Economics*, 9, 695–752. [1320]

Krings, Gautier, Francesco Calabrese, Carlo Ratti, and Vincent D. Blondel (2009), "Urban gravity: A model for inter-city telecommunication flows." *Journal of Statistical Mechanics: Theory and Experiment*, 2009 (07), L07003. [1298]

Lambiotte, Renaud, Vincent D. Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren (2008), "Geographical dispersal of mobile communication networks." *Physica A: Statistical Mechanics and its Applications*, 387 (21), 5317–5325. [1298]

Liben-Nowell, David, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins (2005), "Geographic routing in social networks." *Proceedings of the National Academy of Sciences*, 102 (33), 11623–11628. [1298]

List, John A., Fatemeh Momeni, and Yves Zenou (2019), "Are estimates of early education programs too pessimistic? Evidence from a large-scale field experiment that causally measures neighbor effects." CEPR Discussion Paper No. 13725. [1297]

Manski, Charles F. (1993), "Identification of endogenous social effects: The reflection problem." *The Review of Economic Studies*, 60 (3), 531. [1313]

Marmaros, David and Bruce Sacerdote (2006), "How do friendships form?" *The Quarterly Journal of Economics*, 121 (1), 79–119. [1297]

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook (2001), "Birds of a feather: Homophily in social networks." *Annual Review of Sociology*, 27 (1), 415–444. [1309]

Mossay, Pascal and Pierre M. Picard (2011), "On spatial equilibria in a social interaction model." *Journal of Economic Theory*, 146 (6), 2455–2477. [1297]

Mossay, Pascal and Pierre M. Picard (2019), "Spatial segregation and urban structure." *Journal of Regional Science*, 59, 480–507. [1297]

Picard, Pierre M. and Yves Zenou (2018), "Urban spatial structure, employment and social ties." *Journal of Urban Economics*, 104, 77–93. [1297]

Putnam, Robert D. (2000), *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster. [1296]

Rosenthal, Stuart S. and William C. Strange (2008), "The attenuation of human capital spillovers." *Journal of Urban Economics*, 46 (2), 373–389. [1297]

Sacerdote, Bruce (2011), "Peer effects in education: How might they work, how big are they and how much do we know thus far?" In *Handbook of the Economics of Education*, Vol. 3 (Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds.), 249–277, Elsevier, Amsterdam. [1316]

Sato, Yasuhiro and Yves Zenou (2015), "How urbanization affect employment and social interactions." *European Economic Review*, 75, 131–155. [1297]

Schmutte, Ian M. (2015), "Job referral networks and the determination of earnings in local labor markets." *Journal of Labor Economics*, 33 (1), 1–32. [1297]

Smith, Anthony A. Jr. (1993), "Estimating nonlinear time–series models using simulated vector autoregressions." *Journal of Applied Econometrics*, 8, S63–S84. [1310]

Smith, Anthony A. Jr. (2008), "Indirect inference." In *The New Palgrave Dictionary of Economics*, second edition (Steven N. Durlauf and Lawrence E. Blume, eds.), Palgrave Macmillan, London. [1310]

Spencer, Barbara J. and James A. Brander (1983), "International R&D rivalry and industrial strategy." *The Review of Economic Studies*, 50 (4), 707–722. [1322]

Zenou, Yves (2013), "Spatial versus social mismatch." *Journal of Urban Economics*, 74, 113–132. [1297]

Co-editor Christopher Taber handled this manuscript.

Manuscript received 12 September, 2020; final version accepted 8 May, 2023; available online 24 May, 2023.