ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Kim, Jun Sung; Patacchini, Eleonora; Picard, Pierre M.; Zenou, Yves

Working Paper

Spatial Interactions

IZA Discussion Papers, No. 15376

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Kim, Jun Sung; Patacchini, Eleonora; Picard, Pierre M.; Zenou, Yves (2022) : Spatial Interactions, IZA Discussion Papers, No. 15376, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/263592

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15376

Spatial Interactions

Jun Sung Kim
Eleonora Patacchini
Pierre M. Picard
Yves Zenou

JUNE 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15376

Spatial Interactions

Jun Sung Kim
Kyung Hee University

Eleonora Patacchini
Cornell University, EIEF, CEPR and IZA

Pierre M. Picard
University of Luxembourg, Université catholique de Louvain and IZA

Yves Zenou
Monash University, CEPR and IZA

JUNE 2022

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Spatial Interactions*

This paper studies how the strength of social ties are 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 geo-coded dataset 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 yields a higher total welfare than subsidizing transportation costs.

JEL Classification: D85, R1, Z13

Keywords: social networks, location, structural estimation, policies

Corresponding author:
Yves Zenou
Department of Economics
Monash University
Caulfield East VIC 3145
Australia

E-mail: yves.zenou@monash.edu

* We are grateful to the editor Christopher Taber, three anonymous referees as well as 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.

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 spatial proximity in the determination and intensity of network exchange remains under-examined. Indeed, most papers from the network economics literature (Jackson, 2008) assume that the existence and intensity of dyadic contacts do not depend on 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 requires 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, 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 pre-determined to the friendship decisions of the children. Our main empirical challenges are due to the fact that there is some discrepancy between

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

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

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 having small networks (and conducting different robustness checks) and 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 socio-demographic 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 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) and 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

[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.

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 socio-economic 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 et al. (2018b) and Bailey et al. (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 et al. (2005); Lambiotte et al. (2008); Goldenberg and Levy (2009); Krings et al. (2009) and the excellent reviews of

[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 employer-employee data with residential information. Using mobile phone data on one entire city in China, Barwick et al. (2019) show that geographical distance is important in spreading information about jobs.

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 empirical 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 in the theoretical model can be found in Appendix A. In Appendix B, we solve for the social capital fixed point and show under which condition it is unique. In Appendix C, we perform some robustness checks. In Section D, we carry out Monte Carlo simulation experiments while we explain our calibration in the policy exercises in Appendix E.

2 The model

2.1 Notations and definitions

Consider a set of $N \geq 2$ homogeneous individuals embedded in a social network. As in our dataset (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 between two students i and j belonging to the same social network. Each student visits *every other*

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

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.

[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.

Each student i incurs a cost $c(d_{ij})$ of visiting another student residing at j, where d_{ij} is the geographical distance between i and j. We consider 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.

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.

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

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

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 residing in 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 own social capital α, has the following effects. By using the proof of Proposition 1, we can differentiate each side of (A.4) 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)

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

with respect to α, we obtain the total effect as a function of exogenous variables:

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

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

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

subject to the social capital constraint

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

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

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

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 7.3, 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 how they fit with our theoretical framework. 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 the issues related to the discrepancy between the theoretical model and the data and how we address them.

4.1 Data source

We use a dataset on friendship networks from the National Longitudinal Survey of Adolescent Health (Add Health) to test our theoretical findings and run some counterfactual policies.[10]

The Add Health dataset 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 geo-coded 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 seven days?

[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 Carolina 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 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.

- Did you meet {NAME} after school to hang out or go somewhere during the past seven days?
- Did you spend time with {NAME} during the past weekend?
- Did you talk to {NAME} about a problem during the past seven days?
- Did you talk to {NAME} on the telephone during the past seven 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 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, $n_{ij} > 0$ is always positive (see Proposition 1). By contrast, in the Add Health dataset, 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 dataset may have zero interactions

with their (nominated) friends. The latter issue is a missing-data problem because social interaction between some pairs of students are unobserved in the data. The former is related to the sparsity of networks, a common problem in network data.

Finally, the measurement of social interactions is different in the 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 equal to $\{0, 1, 2, 3, 4, 5\}$, 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, 12,761 students have geo-coded data (and other socioeconomic characteristics). Information on nominated friends, types of interactions and geographical location is only available for 4,449 students. This large reduction in sample size is common in the Add Health when mapping friendships and it is mainly due to the network construction procedure —roughly 20 percent of the students do not nominate any friend and another 20 percent cannot be correctly linked.

Using the friendship nomination data and the corresponding social interaction responses, we construct a 4,449 by 4,449 adjacency matrix of directed friendship network and another 4,449 by 4,449 adjacency matrix of directed social interaction intensity. 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 in the data are geographically dispersed all around the United States, we cannot assume that they all know each other. Hence, we partition the data into small *connected components*. A connected component of a network is a maximal set of nodes such that each pair of nodes is connected by a path. We obtain a total of 1,120 *directed* network components, that is, networks for which n_{ij} is not necessary equal to n_{ji}. Then, we focus on network component sizes between four and ten members and define each component as a network. We do this for the following two reasons. First, the upper and lower tails of the distribution of networks by network size are commonly trimmed since the strength of peer effects may be too different in too small or too large networks (see Calvó-Armengol, Patacchini, and Zenou, 2009). Second, and most importantly, to reduce

the discrepancy between the theoretical model, which assumes that all individuals interact with each other (i.e., $n_{ij} > 0$), and the data, for which many students are not friends with each other (i.e., $n_{ij} = 0$) (see Section 4.4), we keep the sample size of the networks (i.e., connected components) relatively small (4 to 10 students), so that students are more likely to be friends with each other. To be more precise, for each school, we remove both very small networks that have less than 4 students and larger networks that have more than 10 students. By doing so, the number of students reduces from 4,449 to 739.

In Appendix C, we provide different robustness checks of our estimation results by using larger network size (up to 50 students) and different definitions of the network. In particular, as an alternative definition, we choose the school instead of the component to define the network.

4.6 Final sample

Our final sample consists of 739 individuals distributed over 139 networks. 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 geo-coded information. In columns (2)–(3), we further restrict the sample to those with friendship information and intensity of interactions. Finally, in column (4), we report our sample where we only keep students in networks of 4–10 agents.

Table 1 also shows that the differences in means between these different samples are almost never statistically significant, which strongly suggests no specific bias in the selection of the sample. Among the adolescents selected in our sample of students, 53% are female and 18% are blacks. Slightly more than 70% live in a household with two married 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 2.86, meaning slightly less than a grade of "B." The average family income is 48,410 in 1994 dollars, although 10% of parents chose not to report such information. The average number of social interactions is 1.05.

[*Insert Table 1 here*]

In Table 2, we document the number of social interactions. On average, there are 2,111

social interactions, which are mainly between white students; there are fewer inter-ethnic interactions. Further, on average, each pair has 2.849 social interactions; white pairs socially interact more with each other than black pairs do.

[*Insert Table 2 here*]

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 their social distances, that is, on their distances in terms of socio-demographic characteristics:

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

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 further assume linear travel cost such that $c\left(d_{ij}\right) = c \times d_{ij}$ 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 $R = 139$ networks ($r = 1, ..., R$). Since networks are defined as connected components 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 of the $R = 139$ networks.

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

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

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 socio-economic 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 also exhibits asymmetry between $n_{ij,r}$ and $n_{ji,r}$.

By plugging the value of $n^*_{ij,r}$ from (12) into (13), in Appendix B, 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 (B.17) in Appendix B):

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

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 (B.16) in Appendix B).

5.2 Estimation strategy

Indirect inference For each network r, our dataset 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 will recover the parameters α, β (that is, all β_ms and all β_Ms), 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 parameters, based on an injectivity assumption (i.e., one-to-one mapping between the structural parameters and the auxiliary parameters).

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 model (12) and (15) can now be written as the following system of equations:

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

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

As explained in Section 4.2, the observed $n^{obs}_{ij,r}$ 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 non-negative real numbers. Hence, to fill the gap between $n^*_{ij,r}$ and $n^{obs}_{ij,r}$, we apply the following mapping to calculate the final $n^{sim}_{ij,r}$, which we will be the counterpart to $n^{obs}_{ij,r}$ in

[12] Indirect inference was introduced by Smith (1993) and Gourieroux, Monfort, and Renault (1993) and later extended by Gallant and Tauchen (1996). For overviews on indirect inference, see Gourieroux and Monfort (1996) and Smith (2008).

the I-I procedure.

$$n_{ij,r}^{sim} = \begin{cases} 0, \text{ if } -\infty < n_{ij,r}^{*} < 1; \\ 1, \text{ if } 1 < n_{ij,r}^{*} \leq 2; \\ 2, \text{ if } 2 < n_{ij,r}^{*} \leq 3; \\ 3, \text{ if } 3 < n_{ij,r}^{*} \leq 4; \\ 4, \text{ if } 4 < n_{ij,r}^{*} \leq 5; \\ 5, \text{ if } n_{ij,r}^{*} \geq 5. \end{cases} \tag{18}$$

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.

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:

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

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

Algorithm We draw T sets of simulation errors, $\mathcal{E}^t \equiv (\varepsilon^t_{ij,r})$, $t = 1, \ldots, T$, 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 $\mathbf{s}^t_r$ and predict the intensity of social interactions $\widehat{n}^t_{ij,r}$ for each set of errors using equations (16) and (17). To match the data, we constrain $\widehat{n}^t_{ij,r}$ to lie between zero and five (included). This process yields the first degree interaction matrix $\widehat{\mathbf{N}}_r(\mathcal{E}^t_r, \mathbf{Y_r}; \mu)$ and the second degree interaction matrix as the square of the latter. Let $\mathbf{Y}$, $\mathcal{E}^t$, $\mathbf{N}$ and $\widehat{\mathbf{N}}(\mathcal{E}^t, \mathbf{Y}; \mu)$ collect the observed data, the non-observed data, the observed interactions and the predicted interactions in all networks. We then run OLS regressions on the auxiliary model (19) and (20) separately with the observed and simulated interaction values. As a result, we obtain a set of the OLS estimates $\widehat{\boldsymbol{\gamma}}(\mathbf{N}, \mathbf{Y})$ with the observed interactions and a set of estimates $\widehat{\boldsymbol{\gamma}}[\widehat{\mathbf{N}}(\mathcal{E}^t, \mathbf{Y}; \mu), \mathbf{Y}]$, $t = 1, \ldots, T$ 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 $\widehat{\boldsymbol{\gamma}}[\widehat{\mathbf{N}}(\mathcal{E}^t, \mathbf{Y}; \mu), \mathbf{Y}]$. Formally, the I-I estimator $\widehat{\mu}_{\text{II}}$ is constructed such that

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

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

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

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.

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

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

Second, for the unobserved social interactions of students pairs who are not friends, we allow for any integer value of social interactions to be between zero and five. Then, we do not include these pairs when we estimate the auxiliary model of the dyadic regressions. In other words, in the auxiliary models, when we compare the observed values in the data and the simulated ones, we only use the pairs of students who are friends. Note that by only focusing on the small-sized networks (4 to 10 students), we minimize the inclusion of these non-friends pairs. In Table C1 in C, we do a robustness check ("Pairs with positive social interactions") where we run the auxiliary regressions by only using the pairs of students who are friends *and* who have positive social interactions.

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 (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 R^2 will help the identification of the social capital parameter α. 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. Appendix D shows the details about the Monte Carlo simulations and results, which confirms that our method can capture the true parameter values precisely.

6 Empirical results

6.1 Main results

Table 3 reports the estimation results of the key structural parameters and other parameters in the undirectional specification. We also include all socio-demographic characteristics that are related to the intensity of social interactions and social capital. We display the estimates related to social distances and combined levels in two different columns for the sake of the exposition. Note that those two columns of estimates are from the estimation of the same model.

[*Insert Table 3 here*]

Let us start with the socio-demographic characteristics of the 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 non-black students, if they are more physically developed or practicing religion, or if they have two parents with higher education and more family income. By contrast, the intensity of social interactions is decreasing if the students have a higher GPA or if they are from larger families.

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 approximately 1.59. The estimated cost of transportation c is 0.21. After multiplying this cost to the average pairwise distance (equal to 6.71 kilometers), the average estimated transportation cost is roughly 1.41. Finally, α, which measures the importance of others' social capital on an agent's social capital formation, has an estimated value of 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 positive externalities in education (see e.g., Calvó-Armengol, Patacchini, and Zenou, 2009; Boucher et al., 2022).[14]

6.2 Robustness checks

We conduct different robustness checks. The detailed results can be found in Appendix C.

6.2.1 Different sizes of network components and different types of friendship pairs

In our main specification, to reduce the gap between the model and the data, we only considered network components of size 4-10. Also, in the auxiliary model of the dyadic regressions, we did not include the pairs of students who were not friends (that is, we only

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

included friendship pairs). In Section C.1, as a robustness check, we increase the size of the network up to 20, 30, 40, and 50 students. Furthermore, we run the auxiliary regressions by only using the pairs of students who are friends *and* who have positive social interactions (that is, we include friendship pairs with positive social interactions). The results can be found in columns (1)–(5) of Table C1. We can see that our estimation results are very similar to the one obtained in Table 3.

6.2.2 Networks as schools

So far, we only measured networks by connected components. In Section C.2, we have another definition in which each network is a school; that is, we only consider schools in which there is one connected-component network of a given size. The results can be found in columns $(6)-(10)$ in Table C1. The results are almost identical to those based on network components, regardless of school size.

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 some simulations. 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:

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

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

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

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 $s^o_{i,r} > s^*_{i,r}$, $\forall i$.

We numerically solve the optimal level of social interactions and social capital with 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 0.76 fewer times than what is socially optimal. The difference between the socially optimal and the observed levels of social interactions varies from -2.69 to 3.47 across networks. Although there are a few networks where the observed level is larger than the optimal level, many social interactions fall short of the optimum. Students also have less social capital than the optimal one; they have, on average, 34% less social capital.

[*Insert Table 4 here*]

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 $\overline{n}^o_r - \overline{n}^*_r$ and $\overline{s}^o_r - \overline{s}^*_r$ on the network size, network measures, and average characteristics (e.g., average family income) of students in each network r:

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

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

Tables 5 and 6 display the results. Consider, first, *social interactions* (Table 5) and let us

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

examine if the difference between the optimal and the observed levels of social interactions, $(\overline{n}_r^o - \overline{n}_r^*)$, is increasing or decreasing with network size N_r. Although the coefficients on the network size and its square are insignificant in column (5), we have:

$$\frac{\partial(\overline{n}_r^o - \overline{n}_r^*)}{\partial N_r} = \gamma_1 + 2\gamma_2 N_r = 2.261 - 2(0.106)N_r = 0 \tag{28}$$

Solving this equation leads to: $N_r = \frac{2.261}{2(0.106)} = 10.67$. This means that the difference between the optimal and the observed level of social interactions is increasing until the network size reaches (approximately) 10 students and then decreases. As a result, there is a non-monotonic relationship between $\overline{n}_r^o - \overline{n}_r^*$ and N_r where an increase in the network size increases $\overline{n}_r^o - \overline{n}_r^*$ up to $N_r = 10$ and, above this size, an increase in the network size decreases $\overline{n}_r^o - \overline{n}_r^*$. Thus, $N_r = 10$ is the size of the network that *maximizes* these inefficiencies. Given that the median size of the networks is $N_r^{med} = 5$, in terms of magnitude, an increase by one person in the network from $N_r^{med} = 5$, raises these inefficiencies by $2.261 - 2(0.106)N_r^{med} = 1.201$.

[*Insert Tables 5 and 6 here*]

The average geographic distance is significantly associated with the inefficiency. A one-kilometer increase in the average pairwise distance lead to a 0.113 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), physical development level, or family income are more likely to have high inefficiencies in terms of social interactions.

Let us now turn to the inefficiencies in terms of social capital (Table 6). We find that the network population is not strongly associated with the inefficiency in social capital, but the average geographic distance is substantially related to 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.

Network size and average welfare Another interesting exercise, for which we do not have a theory, is to determine the optimal network, i.e., 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 139 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:

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

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, clustering coefficient, and diameter.[17] We include the network measures (such as average degree and average eigenvector centrality of a network) to see how the shape of a network is associated with the welfare.

Table 7 reports the results. We can first calculate the network size that maximizes the average welfare per network AW_r^*. Using column (5), we have:

$$\frac{\partial AW_r^*}{\partial N_r} = \delta_1 + 2\delta_2 N_r = -2.006 + 2(0.082) N_r = 0 \tag{30}$$

This means the network that comprises (approximately) 12 students is the one that minimizes the average welfare per network. Although the coefficients are insignificant, this is consistent

[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.

[17] We compute the clustering coefficient as the ratio of the number of triangle loops to the number of connected triples.

with our previous calculation on the size of networks that maximizes the inefficiency in social interactions, which is approximately 11 students.

[*Insert Table 7 here*]

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 0.007 to 0.507, we find that the average geographic distance explains a significant proportion of the average welfare in a network.

7.2 Counterfactual analysis

Next, we modify the geographic distribution of students in a way that we reduce the extent to the geographical segregation of students from different ethnic groups. For example, if a white student in a network lives relatively far from black students in the network, we switch the residential location of the white student with that of a black one in the same network. Then, we recalculate the social interaction intensity and social capital in the network. W

7.3 Policies

We have seen in Proposition 5 that the social optimal allocation can be restored if social interactions are not subsidized while commuting trips are subsidized as a function of the locations of the destination and origin partners. Because the latter policy requires detailed information about every interaction pair, it is unlikely to be implemented. 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 policy simulations. 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 pre-commit 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.[18]

For each individual residing in i and interacting with someone in 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.

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

7.3.1 Subsidizing social interactions

We consider a *uniform* subsidy σ_r for each network. We use the following discrete version of the equilibrium identities:

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

and

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

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

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

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 Appendix E for technical details.

The first three columns in Table 8 display the results. On average, a subsidy level of 2.896 (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.

[*Insert Table 8 here*]

7.3.2 Subsidizing transportation costs

In the case of subsidies on transport cost, we consider the following equilibrium conditions:

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

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

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 $\tau = 0.729$ (72.9%) 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.

7.3.3 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 ($R = 139$ in our dataset). 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 $N_r(N_r - 1)$, i.e., $B_r = \frac{N_r(N_r-1)}{\sum_{r'} N_r'(N_{r'}-1)} B$.

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, $\bar{\sigma}^o = 2.896$, and $\bar{n}^o$ is the average optimal social interaction level, as obtained in Table 4, that is, $\bar{n}^o = 3.608$.

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, $\bar{\tau}^o = 0.729$ (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 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)).[19] 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.

[*Insert Table 9 here*]

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

8 Concluding remarks

In this paper, we presented a behavioral microfoundation for the relationship between geographical distance and social interactions. We characterized the equilibrium in terms of optimal level 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. Social capital tended to be more concentrated than agents themselves. 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 is was non-monotonic 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 is composed of ten students. We then performed two different subsidy policies. Our results suggested that the individuals interacted at optimal levels when either social interactions or transportation costs were subsidized. However, subsidies on social interactions were more effective than subsidies on transportation costs.

Our analysis thus suggests that encouraging social interactions in cities are 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) and 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 et al. (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.

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.

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

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.

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.

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

Barthélemy, Marc. 2011. "Spatial networks." *Physics Reports* 499 (1-3):1–101.

Barwick, Panle Jia, Yanyan Liu, Eleonora Patacchini, and Qi Wu. 2019. "Information, mobile communication, and referral effects." CEPR Discussion Paper No. 13786 .

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.

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

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

Boucher, Vincent, Carlo Del Bello, Fabrizio Panebianco, Thierry Verdier, and Yves Zenou. 2022. "Education transmission and network formation." *Journal of Labor Economics,* forthcoming .

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

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

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.

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

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.

Chen, Ying-Ju., Yves Zenou, and Junjie Zhou. 2022. "The impact of network topology and market structure on pricing." *Journal of Economic Theory,* forthcoming .

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.

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

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

Fu, Chao and Jesse Gregory. 2019. "Estimation of an Equilibrium Model With Externalities: Post-Disaster Neighborhood Rebuilding." *Econometrica* 87 (2):387–421.

Gallant, A. Ronald and George Tauchen. 1996. "Which moments to match?" *Econometric Theory* 12:657–681.

Glaeser, Edward L. and Bruce Sacerdote. 2000. "The social consequences of housing." *Journal of Housing Economics* 9:1–23.

Goldenberg, Jacob and Moshe Levy. 2009. "Distance is not dead: Social interaction and geographical distance in the internet era." *arXiv preprint arXiv:0906.3202* .

Gourieroux, Christian and Alain Monfort. 1996. *Simulation-Based Econometric Methods*. Oxford: Oxford University Press.

Gourieroux, Christian, Alain Monfort, and Eric Renault. 1993. "Indirect inference." *Journal of Applied Econometrics* 8:S85–S118.

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

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.

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.

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

Helsley, Robert W. and Yves Zenou. 2014. "Social networks and interactions in cities." *Journal of Economic Theory* 150:426–466.

Ioannides, Yannis M. 2013. *From Neighborhoods to Nations: The Economics of Social Interactions*. Princeton: Princeton University Press.

Jackson, Matthew O. 2008. *Social and Economic Networks*. Princeton: Princeton University Press.

Jackson, Matthew O. and Brian W. Rogers. 2005. "The economics of small worlds." *Journal of the European Economic Association* 3:617–627.

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.

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

Jackson, Matthew O and Yves Zenou. 2015. "Games on networks." In *Handbook of Game Theory, Volume 4*, edited by P. Young and S. Zamir. Amsterdam: Elsevier, 91–157.

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

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.

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.

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

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.

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

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.

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.

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.

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 .

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

Marmaros, David and Bruce Sacerdote. 2006. "How Do Friendships Form?" *The Quarterly Journal of Economics* 121 (1):79–119.

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.

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

———. 2019. "Spatial segregation and urban structure." *Journal of Regional Science* 59:480–507.

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

Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community.* Simon and Schuster.

Rosenthal, Stuart S. and William C. Strange. 2008. "The attenuation of human capital spillovers." *Journal of Urban Economics* 46 (2):373—-389.

Sacerdote, Bruce. 2011. "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" In: E. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education, Volume 3,* Amsterdam: Elsevier :249–277.

Sato, Yasuhiro and Yves Zenou. 2015. "How urbanization affect employment and social interactions." *European Economic Review* 75:131–155.

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

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

———. 2008. "Indirect inference." In: S. Durlauf and L. Blume (Eds.), *The New Palgrave Dictionary of Economics, 2nd edition,* London: Palgrave Macmillan .

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

Zenou, Yves. 2013. "Spatial versus social mismatch." *Journal of Urban Economics* 74:113–132.

Appendix

A Proofs

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

$$v'\left(n_{ij}\right) s_j - c(d_{ij}) = 0, \qquad j = 1, ..., N. \tag{A.1}$$

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

$$n_{ij}^* = 1 - \frac{c(d_{ij})}{s_j}, \qquad j = 1, ..., N. \tag{A.2}$$

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,

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

A sufficient condition for this inequality to hold is

$$\min_j s_j > c(\overline{d}), \tag{A.3}$$

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

By plugging (A.2) into (3) and using (4), we obtain the equilibrium level of social capital s_j^*. It is given by

$$s_j^* = 1 + \frac{\alpha}{N} \sum_{k \neq j} s_k^* - \frac{\alpha}{N} g_j. \tag{A.4}$$

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

$$\sum_j s_j^* = \frac{1}{1 - \alpha\left(\frac{N-1}{N}\right)} \left[N - \frac{\alpha}{N} \sum_j g_j \right], \tag{A.5}$$

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 (A.5) into (A.4) yields the

following closed-form solution for the equilibrium social capital:

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

Let us show that the global interaction condition (A.3) 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

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

where $f\,(\alpha) = \left(1+\frac{\alpha}{N}\right)^2 + N\left[1 - 2\left(\frac{\alpha}{N}\right) - \left(\frac{\alpha}{N}\right)^2\right]$. 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

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

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

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

Note that $\sum_i \chi_i\,(s_i - 1)$ evaluates to the same value as $\sum_i \sum_{j \neq i} \chi_j\,(s_j - 1)\,/\,(N-1)$. Sub-

stituting the latter for the former, we re-write the Lagrangian function as

$$\mathcal{L} = \sum_i \sum_{j \neq i} v\left(n_{ij}\right) s_j - n_{ij} c(d_{ij}) + \left(\alpha/N\right) \chi_i n_{ij} s_j - \chi_j \left(s_j - 1\right) / \left(N - 1\right) \tag{A.7}$$

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

$$\begin{aligned} v'\left(n_{ij}\right) s_j - c(d_{ij}) + \left(\alpha/N\right) \chi_i s_j &= 0 \\ \sum_{i \neq j} \left[v\left(n_{ij}\right) + \left(\alpha/N\right) \chi_i n_{ij} - \chi_j / \left(N - 1\right) \right] &= 0 \end{aligned}$$

The last equality is equivalent to

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

This gives (9) and (10). ∎

Proof of Proposition 4: Condition (9) yields

$$v'\left(n_{ij}\right) = \frac{c(d_{ij})}{s_j} - \frac{\alpha}{N} \chi_i, \tag{A.8}$$

which gives

$$n_{ij}^o = 1 - \frac{c(d_{ij})}{s_j^o} + \frac{\alpha}{N} \chi_i^o, \tag{A.9}$$

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

$$s_i^o = 1 + \frac{\alpha}{N} \sum_{l \neq i} s_l^o - \frac{\alpha}{N} g_i + \left(\frac{\alpha}{N}\right)^2 \chi_i^o \sum_{l \neq i} s_l^o. \tag{A.10}$$

Observe that, for $\chi_i^o = 0$, (A.9) and (A.10) are identical to the equilibrium conditions and therefore yield the equilibrium values n_{ij}^* and s_i^*. The RHS of (A.9) and (A.10) are increasing functions of χ_i^o and/or s_i^o. From (A.10), we see that an increase in χ_i^o above zero raises s_i^o.

From (A.9), 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

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

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

$$\begin{aligned} s_j^* &= 1 + \frac{\alpha}{N} \sum_{k \neq j} n_{jk}^* s_k^* \\ &= 1 + \frac{\alpha}{N} \sum_{k \neq j} \left(1 - \frac{c(d_{jk}) - \tau_{jk}}{s_k^* + \sigma_{jk}}\right) s_k^*. \end{aligned} \tag{A.11}$$

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

$$n_{ij}^* = n_{ij}^o \iff \frac{c(d_{ij}) - \tau_{ij}}{s_j^* + \sigma_{ij}} = \frac{c(d_{ij})}{s_j^o} - \frac{\alpha}{N} \chi_i^o, \tag{A.12}$$

and $s_j^* = s_j^o$ given by (A.11) and (A.10).

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

$$\sigma_{ij} = \frac{s_j^o}{\frac{Nc(d_{ij})}{\alpha \chi_i^o s_j^o} - 1} \tag{A.13}$$

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.

B Social capital fixed point

The fixed point in social capital can be computed by rewriting equation (12) as $n_{ij,r}s_{j,r} = (n_0 + \theta_{ij,r})\, s_{j,r} - cd_{ij,r}$, so that (13) becomes

$$s_{j,r} = 1 + \frac{\alpha}{N_r} \sum_{k=1,k\neq j}^{N_r} \left[(n_0 + \theta_{jk,r})\, s_{k,r}\right] - \frac{\alpha}{N_r} c \sum_{k=1,k\neq j}^{N_r} d_{jk,r}, \tag{B.14}$$

where the last term is $g_{j,r} = \sum_{k=1,k\neq j}^{N_r} c(d_{jk,r}) = c \sum_{k=1,k\neq j}^{N_r} d_{jk,r}$, the linear-cost equivalent of the access cost measure defined in (4) in the model. The system of linear equations (B.14) can be written in vector-matrix form as

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

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; $\mathbf{\Theta}_r = (\theta_{ij,r}) = (x_{ij,r}^{\mathrm{T}}\beta + \varepsilon_{ij,r})$ is an $(N_r \times N_r)$ matrix; $\mathbf{D}_r = (d_{ij,r})$ is an $(N_r \times N_r)$ matrix. Namely,

$$\mathbf{D}_r = \begin{pmatrix} d_{11,r} & \dots & d_{1i,r} & \dots & d_{1N_r,r} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i1,r} & \dots & d_{ii,r} & \dots & d_{iN_r,r} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{N_r1,r} & \dots & d_{N_ri,r} & \dots & d_{N_rN_r,r} \end{pmatrix} \text{ and } \mathbf{\Theta}_r = \begin{pmatrix} \theta_{11,r} & \dots & \theta_{1i,r} & \dots & \theta_{1N_r,r} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_{i1,r} & \dots & \theta_{ii,r} & \dots & \theta_{iN_r,r} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_{N_r1,r} & \dots & \theta_{N_ri,r} & \dots & \theta_{N_rN_r,r} \end{pmatrix}. \tag{B.16}$$

Solving the system of linear equations (B.14) leads to

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

where $\mathbf{I}_r$ is the $(N_r \times N_r)$ identity matrix. The matrix $\mathbf{I}_r - \alpha\left(\mathbf{N}_{0,r} + \mathbf{\Theta}_r\right)$ is invertible if $\alpha < \frac{1}{\rho(\mathbf{N}_{0,r} + \mathbf{\Theta}_r)}$, where $\rho\left(\mathbf{N}_{0,r} + \mathbf{\Theta}_r\right)$ is the spectral radius of the matrix $\mathbf{N}_0 + \mathbf{\Theta}_r$. When this condition is satisfied, there is a unique solution to the system of linear equations (B.14).

C Robustness checks

C.1 Different sizes of network components

C.1.1 Friendship pairs

We first check the robustness of our estimation results when we increase the size of the network components. For that, we expand the data to include all connected components within a school of size 20, 30, 40, and 50. When doing so, the numbers of students and networks increase from 739 and 139, respectively in the our benchmark case (networks of size 4 to 10), to 877 and 149 (networks of size 4 to 20), to 961 and 152 (networks of size 4 to 30), to 1,032 and 154 (networks of size 4 to 40), and to 1,212 and 158 (networks of size 4 to 50). Columns (1)–(5) in the upper left quadrant of Table C1 display the results.[20] We find that all the structural parameter estimates are almost identical regardless of component size.

C.1.2 Pairs with positive social interactions

In Section 5.3, for our main estimation results in Table 3, we did not include the pairs of students who were not friends when we estimated the auxiliary model of the dyadic regressions. Here, as a robustness check, we do not include the pairs of students who are not friends *and* the pairs of students who are friends but report having no social interactions with each other. In other words, in this robustness check ("Pairs with positive social interactions"), we run the auxiliary regressions by only using the pairs of students who are friends *and* who have positive social interactions. Columns (1)–(5) in the lower left quadrant of Table C1 display the results. We can see that our estimation results are not sensitive to this choice of different types of pairs.

C.2 Networks as schools

In our main specification (Table 3) and in our robustness checks (Section C.1), we defined networks as connected components of a certain size in each school. However, there may be more than one connected component in a school if there are many disconnected cliques.

[20] Note that column (1) in the upper left quadrant of Table C1 corresponds to our main results displayed in Table 3.

Consequently, in this robustness check, we propose another definition in which each network is a school. In other words, we only consider schools in which there is one connected-component network of a given size while, in our main specification, we included schools that had many connected-component networks of a given size.

As in our main specification, consider networks of size $4-10$. When networks are defined as schools, they need to satisfy two conditions: (i) the school has to have one (connected) component and (ii) this component has to be of size $4-10$. Column (6) in Table C1 displays the results of this estimation by differentiating between friendship pairs (right upper panel) and pairs with positive social interactions (right lower panel). Quite naturally, we observe that the numbers of students and networks decrease from 739 and 139 (column (1)) to 76 and 9 (column (6)), respectively. When we look at the estimated coefficients of the main variables, we see that they remain roughly the same. Then, as above, in columns $(7)-(10)$, we increase the size of the networks up to 50 students and we observe that the results are almost identical to those based on network components, regardless of school size.

Table C1: Robustness checks: network definition and network size

	Network components					Schools				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Network size	4 to 10	4 to 20	4 to 30	4 to 40	4 to 50	4 to 10	4 to 20	4 to 30	4 to 40	4 to 50
Friendship pairs						***Friendship pairs***				
n_0	1.5895	1.5825	1.5917	1.5961	1.5942	1.5950	1.5923	1.5931	1.5891	1.5892
	(0.0197)	(0.0119)	(0.0155)	(0.0230)	(0.0087)	(0.0125)	(0.0196)	(0.0128)	(0.0120)	(0.0038)
α	0.1286	0.1286	0.1285	0.1303	0.1288	0.1281	0.1287	0.1292	0.1287	0.1302
	(0.0010)	(0.0010)	(0.0010)	(0.0014)	(0.0024)	(0.0012)	(0.0009)	(0.0012)	(0.0015)	(0.0014)
c	0.2099	0.2095	0.2098	0.2088	0.2097	0.2075	0.2103	0.2086	0.2094	0.2099
	(0.0019)	(0.0012)	(0.0032)	(0.0029)	(0.0018)	(0.0022)	(0.0037)	(0.0048)	(0.0022)	(0.0014)
Pairs with positive social interactions						***Pairs with positive social interactions***				
n_0	1.5895	1.5904	1.5893	1.5895	1.5952	1.5951	1.6074	1.5886	1.5871	1.5857
	(0.0273)	(0.0254)	(0.0181)	(0.0157)	(0.0104)	(0.0085)	(0.0272)	(0.0251)	(0.0181)	(0.0205)
α	0.1286	0.1286	0.1286	0.1286	0.1278	0.1290	0.1292	0.1284	0.1285	0.1296
	(0.0013)	(0.0009)	(0.0019)	(0.0011)	(0.0012)	(0.0016)	(0.0005)	(0.0010)	(0.0009)	(0.0011)
c	0.2099	0.2097	0.2071	0.2099	0.2087	0.1983	0.2102	0.2100	0.2101	0.2099
	(0.0028)	(0.0031)	(0.0014)	(0.0018)	(0.0028)	(0.0023)	(0.0025)	(0.0034)	(0.0020)	(0.0017)
# networks	139	149	152	154	158	9	35	62	78	88
# students	739	877	961	1032	1212	76	479	1138	1714	2162
# pairs	3512	5320	7590	10064	17994	582	6602	22270	42502	62190

Note: Bootstrap standard errors are in parentheses. All the coefficients estimates are statistically significant at the 1% level of significance.

D Monte Carlo simulations

We carry out Monte Carlo simulation experiments to demonstrate that our structural estimation method can precisely capture the value of parameters in a complicated data generating process of social interactions among students. Each experiment is concerned with estimating the parameters in the model that we discussed in Section 6. That is,

$$n^*_{ij,r} = n_0 - \frac{cd_{ij,r}}{s^*_{j,r}} + \theta_{ij,r}, \tag{D.18}$$

and

$$s^*_{j,r} = 1 + \frac{\alpha}{N_r}\sum_{k=1}^{N_r} n^*_{jk,r} s^*_{k,r}, \tag{D.19}$$

where

$$\theta_{ij,r} = \beta_1 |x_{i,r} - x_{j,r}| + \beta_2 (x_{i,r} + x_{j,r}) + \varepsilon_{ij,r}, \tag{D.20}$$

We set the values of structural parameters as the ones we have estimated in our structural estimation. That is, $n_0 = 1.5$, $\alpha = 0.12$, and $c = 0.2$. We assign -0.3 for the parameter β_1 to assume homophily and 0.2 for β_2 to have positive the effect of combined levels on social interactions. The data generating processes for x_i and ε are the uniform distribution from the interval of $(0, 5)$ and the normal distribution with mean zero and standard deviation $\sigma_\varepsilon = 1.3$.

We generate $R = 50$, 100, and 150 networks, which correspond to connected components as in our empirical setup. Each network has four to ten individuals. Using the social interaction and social capital fixed points, that is, equations (16) and (17), we generate $n^*_{ij,r}$ for all networks and all pairs.

We generate $H = 100$ sets of generated sample of R networks. For each set of generated data, we run the I-I estimation method. Each hth estimation requires the estimation of the weight matrix A in equation (22) using a bootstrap method and the generation of additional $T = 100$ sets of simulation errors. Although the dimension of the parameter vector is smaller than that in the empirical analysis, this Monte Carlo simulation is also computationally heavy. Hence, to facilitate the computation, we reduce the size of the bootstrap sample for the weight matrix estimation from 3,000 in the empirical analysis to 100.

The results of the Monte Carlo simulations are displayed in Table C2. We report the

Table C2: Monte Carlo simulation results

		Number of networks (R)		
		50 networks	100 networks	150 networks
n_0	Average	1.5269	1.5264	1.5275
(True value= 1.5)	Bias	0.0269	0.0264	0.0275
	RMSE	0.0448	0.042	0.0388
α	Average	0.1204	0.1208	0.1205
(True value= 0.12)	Bias	0.0004	0.0008	0.0005
	RMSE	0.0026	0.0023	0.0026
c	Average	0.2006	0.1994	0.20005
(True value= 0.2)	Bias	0.0006	0.0006	0.00005
	RMSE	0.0045	0.005	0.0041
β_1	Average	-0.2987	-0.2991	-0.2995
(True value= -0.3)	Bias	0.0013	0.0009	0.0005
	RMSE	0.0053	0.0065	0.0063
β_2	Average	0.2022	0.2027	0.2024
(True value= 0.2)	Bias	0.0022	0.0027	0.0024
	RMSE	0.0052	0.0053	0.0043
σ_ε	Average	1.3150	1.3100	1.3134
(True value= 1.3)	Bias	0.0150	0.0100	0.0134
	RMSE	0.0298	0.0218	0.0283

Note: A total of 100 simulations for each experiment.

averages of the estimate, bias, and the Root Mean Squared Error (RMSE) for each method. In general, regardless of the number of networks, our structural estimation method that employs indirect inference captures accurately the value of true parameters in the data generating process. In particular, we succeed to estimate the most important structural parameters, α and c, very precisely.

E Calibration in the policy exercises

Consider equations (31) and (35) in Section 7 and denote them as follows:

$$n_{ij,r} = n_0 + \theta_{ij,r} - \frac{\sigma_r - (1-\tau_r)\, cd_{ij,r}}{s_{j,r}}, \tag{E.1}$$

and

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

where we implement together the two policies. The first equation can be written as

$$n_{ij,r} s_{j,r} = (n_0 + \theta_{ij,r})\, s_{j,r} + \sigma_r - (1-\tau_r)\, cd_{ij,r},$$

so that the second equation becomes

$$s_{j,r} = 1 + \frac{\alpha}{N_r} \sum_{k=1,k\neq j}^{N_r} \left[(n_0 + \theta_{jk,r})\, s_{k,r}\right] - \frac{\alpha}{N_r} \sum_{k=1}^{N_r} \left[\sigma_r - (1-\tau_r)\, cd_{jk,r}\right]. \tag{E.2}$$

Denote by $\mathbf{s}_r = (s_{1,r}, ..., s_{n,r})^{\mathrm{T}}$ the $(N_r \times 1)$ vector of social capital. Thus, in vector-matrix form, (E.2) can be written as:

$$\mathbf{s}_r = \mathbf{1}_r + \alpha\left(\mathbf{N}_{0,r} + \mathbf{\Theta}_r\right)\mathbf{s}_r + \alpha\sigma_r N_r \mathbf{1}_r - \alpha\left(1-\tau_r\right) c\mathbf{D}_r \mathbf{1}_r.$$

Solving this equation leads to:

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

or, equivalently,

$$\mathbf{s}_r = \left[\mathbf{I}_r - \alpha\left(\mathbf{N}_{0,r} + \mathbf{\Theta}_r\right)\right]^{-1}\left[\left(1 + \alpha\sigma_r N_r\right)\mathbf{I}_r - \alpha\left(1-\tau_r\right) c\mathbf{D}_r\right]\mathbf{1}_r. \tag{E.3}$$

The matrix $\mathbf{I}_r - \alpha\left(\mathbf{N}_0 + \mathbf{\Theta}_r\right)$ is invertible if $\alpha < \frac{1}{\rho(\mathbf{N}_{0,r} + \mathbf{\Theta}_r)}$, where $\rho\left(\mathbf{N}_{0,r} + \mathbf{\Theta}_r\right)$ is the spectral radius of the matrix $\mathbf{N}_{0,r} + \mathbf{\Theta}_r$. Consequently, we could solve the model using (E.1) and

(E.3). Observe that $n_{ij,r} > 0$ if $(1 + \theta_{ij,r})\, s_{j,r} > (1 - \tau_r)\, cd_{ij,r}$, $\forall i, j$. A sufficient condition is

$$s_{j,r} > \max_i \frac{(1 - \tau_r)\, cd_{ij,r} - \sigma_r}{(1 + \theta_{ij,r})}.$$

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)	0.5 (0.50)	[0.52]	0.51 (0.50)	[0.48]	0.53 (0.50)	[0.51]
Black	Dummy variable taking value one if the respondent is Black or African American. "White" is the reference category	0.23 (0.42)	0.24 (0.43)	[0.39]	0.20 (0.40)	[0.22]	0.18 (0.38)	[0.48]
Student grade	Grade of student in the current year, range 7 to 12	9.67 (1.63)	9.49 (1.62)	[0.27]	9.47 (1.61)	[0.51]	9.30 (1.68)	[0.42]
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)	2.78 (0.76)	[0.48]	2.83 (0.75)	[0.30]	2.89 (0.74)	[0.46]
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)	3.23 (1.12)	[0.47]	3.30 (1.10)	[0.36]	3.31 (1.12)	[0.50]
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)	2.38 (1.41)	[0.49]	2.38 (1.41)	[0.51]	2.36 (1.38)	[0.53]
Family size	Number of people living in the household	3.61 (1.66)	3.52 (1.51)	[0.28]	3.42 (1.39)	[0.31]	3.47 (1.37)	[0.52]
Two parents	Dummy variable taking value one if the respondent lives in a household with two parents (both biological and non biological) that are married Two parent	0.66 (0.47)	0.68 (0.47)	[0.48]	0.71 (0.45)	[0.28]	0.73 (0.44)	[0.48]
Parental education	Schooling level of the (biological or non-biological) 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 four-year college" If both parents are in the household, the maximum level of schooling is considered	3.09 (0.97)	3.11 (0.95)	[0.50]	3.19 (0.92)	[0.27]	3.16 (0.98)	[0.51]
Family income	Family income in thousands of dollars	40.72 (50.76)	39.93 (50.32)	[0.41]	43.37 (56.78)	[0.32]	48.80 (67.81)	[0.45]
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)	[0.52]	0.10 (0.31)	[0.53]	0.09 (0.29)	[0.50]
N.obs		20,745	12,761		4,449		739	

Note: (1): original sample, (2): sample with geo-coded information, (3): Sample with social-interaction information, (4) Sample in networks of size 4–10. 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	294	74	1,743	2,111
Number of friendship pairs	109	36	596	741
Average social interactions per pair	2.697	2.056	2.924	2.849

Note: The statistics are computed using the network-level average social interactions from 139 networks.

Table 3: Structural estimation results

	Undirected model with directed n_{ij}	
n_0	1.5895***	
	(0.0197)	
α	0.1286***	
	(0.0010)	
c	0.2099***	
	(0.0019)	
β	$\lvert x_i - x_j \rvert$	$(x_i + x_j)$
Female	-0.9979***	0.1929***
	(0.0109)	(0.0064)
Black	-0.6653***	-0.0764***
	(0.0139)	(0.0026)
Grade	0.2889***	0.0846***
	(0.0071)	(0.0008)
GPA	-0.1106***	-0.0705***
	(0.0011)	(0.0018)
Physical development	0.0046***	0.0640***
	(0.0001)	(0.0024)
Religious practice	-0.1187***	0.0332***
	(0.0016)	(0.0006)
Family size	-0.0670***	-0.0153***
	(0.0013)	(0.0004)
Two parents	-0.0075***	0.0305***
	(0.0004)	(0.0004)
Parental education	-0.0457***	0.0144***
	(0.0008)	(0.0002)
Family income	-0.0016***	0.0016***
	(0.00002)	(0.00005)
Family income refused	-0.1214***	0.1348***
	(0.0020)	(0.0022)
σ_ε	1.3501***	
	(0.0070)	
Number of networks	139	
Number of pupils	739	
Number of directed pairs	3,512	
Objective function	4,967.5	

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. *** p<0.01, ** p<0.05, * p<0.1

Table 4: Social interactions and social capital: Optimal level vs. observed level

Social interactions					Social capital				
Optimal level (SD)	Observed level	Average difference (SD)	Minimum difference [95% CI]	Maximum difference	Optimal level (SD)	Observed level	Average difference (SD)	Minimum difference [95% CI]	Maximum difference
3.612	2.847	0.760	0.616	0.889	1.632	1.078	0.554	0.524	0.577
(0.050)	-	(0.050)	[0.676, 0.837]		(0.011)	-	(0.011)	[0.536, 0.572]	

Note: The statistics are computed using the network-level average social interactions and social capital from 139 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).

Table 5: Difference between optimal level and observed level of social interactions

	Optimal−Observed (social interactions)				
	(1)	(2)	(3)	(4)	(5)
Network population	0.554*	0.431	0.491	2.659*	2.261
	(0.322)	(0.287)	(0.306)	(1.414)	(1.438)
Network population squared	-0.031	-0.024	-0.027	-0.124*	-0.106
	(0.023)	(0.021)	(0.022)	(0.065)	(0.066)
Avg. geographic distance		-0.111***	-0.112***	-0.113***	-0.113***
		(0.014)	(0.014)	(0.014)	(0.016)
Avg. degree centrality			0.047	-0.970**	-1.051**
			(0.402)	(0.452)	(0.490)
Std.dev. of degree centrality			-0.257	-0.910	-0.591
			(0.409)	(0.965)	(0.993)
Avg. eigenvector centrality				17.524	14.708
				(14.567)	(15.321)
Clustering coefficient				1.424	1.640*
				(0.875)	(0.917)
Diameter				-0.269	-0.183
				(0.265)	(0.283)
Female fraction					0.596
					(0.398)
Black fraction					-0.101
					(0.247)
Avg. student grade					0.102**
					(0.051)
Avg. GPA					-0.228
					(0.166)
Avg. level of physical development					0.277*
					(0.160)
Avg. level of religion practice					0.044
					(0.104)
Avg. family size					0.099
					(0.100)
Fraction of students with two parents					-0.246
					(0.372)
Avg. level of parent education					0.076
					(0.150)
Avg. family income					0.004*
					(0.002)
Fraction family income refused					0.729
					(0.509)
Constant	-1.221	-0.100	-0.198	-13.213	-13.135
	(1.039)	(0.916)	(0.969)	(11.186)	(11.849)
Observations	139	139	139	139	139
R-squared	0.063	0.327	0.329	0.351	0.454

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. *** p<0.01, ** p<0.05, * p<0.1

Table 6: Difference between optimal level and observed level of social capital

	Optimal−Observed (social capital)				
	(1)	(2)	(3)	(4)	(5)
Network population	0.194**	0.165**	0.189***	-0.010	-0.005
	(0.077)	(0.064)	(0.066)	(0.292)	(0.124)
Network population squared	-0.009	-0.007	-0.008	0.002	0.002
	(0.006)	(0.005)	(0.005)	(0.014)	(0.006)
Avg. geographic distance		-0.026***	-0.026***	-0.026***	-0.027***
		(0.002)	(0.002)	(0.002)	(0.002)
Avg. degree centrality			-0.064	-0.028	-0.068
			(0.056)	(0.094)	(0.047)
Std.dev. of degree centrality			-0.030	-0.302	-0.154
			(0.076)	(0.302)	(0.100)
Avg. eigenvector centrality				-3.345	-2.812**
				(3.140)	(1.361)
Clustering coefficient				-0.088	-0.030
				(0.168)	(0.078)
Diameter				-0.085	-0.052
				(0.097)	(0.034)
Female fraction					0.146***
					(0.029)
Black fraction					-0.082***
					(0.017)
Avg. student grade					0.070***
					(0.004)
Avg. GPA					-0.060***
					(0.012)
Avg. level of physical development					0.060
					(0.011)
Avg. level of religion practice					0.017**
					(0.007)
Avg. family size					-0.015
					(0.009)
Fraction of students with two parents					-0.043*
					(0.023)
Avg. level of parent education					0.026*
					(0.014)
Avg. family income					0.001***
					(0.000)
Fraction family income refused					0.154***
					(0.035)
Constant	-0.206	0.054	0.101	2.691	1.429
	(0.225)	(0.184)	(0.195)	(2.411)	(1.067)
Observations	139	139	139	139	139
R-squared	0.295	0.565	0.570	0.578	0.938

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 in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7: Optimal network design: average welfare and number of students

	(1) Welfare	(2) Welfare	(3) Welfare	(4) Welfare	(5) Welfare
Network population	0.831	0.274	-1.030	-0.864	-2.006
	(1.170)	(0.821)	(0.776)	(2.914)	(3.144)
Network population squared	-0.053	-0.018	0.053	0.035	0.082
	(0.085)	(0.060)	(0.057)	(0.135)	(0.152)
Avg. geographic distance		-0.506***	-0.473***	-0.483***	-0.469***
		(0.051)	(0.045)	(0.046)	(0.047)
Avg. degree centrality			5.088***	9.926***	9.421***
			(1.924)	(1.272)	(1.440)
Std.dev. of degree centrality			0.232	3.072	5.513**
			(0.872)	(2.402)	(2.332)
Avg. eigenvector centrality				26.283	24.638
				(32.059)	(30.964)
Clustering coefficient				-6.586***	-5.234**
				(2.088)	(2.282)
Diameter				1.048	1.731**
				(0.764)	(0.738)
Controls	No	No	No	No	Yes
Observations	139	139	139	139	139
R-squared	0.007	0.507	0.612	0.661	0.757

Note: The outcome variable is the simulated average welfare (AW_r), averaged over 100 simulations for each network.
Control variables include the averages of the social distances and the combined levels used in structural estimation. See Tables 5 and 6.
Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

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]	
1.916	1.516	4.189	0.728	0.633	0.817
(0.379)	[1.614, 2.689]		(0.037)	[0.665, 0.779]	

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.

Table 9: Comparison of two policies

Panel A: Budget corresponding to the average (optimal) social interaction subsidy level

	Number of networks with higher welfare for each policy [95% CI]		Difference in average welfare [95% CI]
Subsidy schemes	Policy: σ	Policy: τ	Policy $\sigma-$ Policy τ
(1) Uniform subsidy amount	135	4	35.84
for each network	[133, 136]	[3, 6]	[35.39, 36.28]
(2) Subsidy proportional	135	4	36.24
to N_r	[133, 136]	[3, 6]	[35.87, 36.61]
(3) Subsidy proportional	135	4	36.61
to $N_r(N_r-1)$	[134, 136]	[3, 5]	[36.28, 36.97]

Panel B: Budget corresponding to the average (optimal) transportation subsidy level

	Number of networks with higher welfare for each policy [95% CI]		Difference in average welfare [95% CI]
Subsidy schemes	Policy: σ	Policy: τ	Policy $\sigma-$ Policy τ
(1) Uniform subsidy amount	135	4	16.49
for each network	[133, 136]	[3, 6]	[16.29, 16.69]
(2) Subsidy proportional	135	4	16.74
to N_r	[133, 136]	[3, 6]	[16.57, 16.89]
(3) Subsidy proportional	135	4	16.93
to $N_r(N_r-1)$	[133, 136]	[3, 6]	[16.78, 17.10]

The median number of networks over 100 simulations, which lead to higher welfare for each policy is reported, along with the 95% confidence interval among 139 networks. The term 'Policy $\sigma-$ 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).