Who knows? The effect of information access on social network centrality

Laura Derksen and Pedro CL Souza*

October 7, 2022

Abstract

Network centrality plays a key role in many economic processes. Yet, the causal determinants of centrality are not well understood. We conducted a randomized experiment in Malawian boarding secondary schools, providing one fifth of students with exclusive access to an online information source throughout the school year. Using a panel of complete and detailed network data, we show that information access leads to a large and sustained increase in network centrality, as students form new strategic links. We calibrate and simulate a model of strategic network formation to demonstrate implications for network-based targeting, information diffusion, inequality and welfare.

JEL classification: I25 D83 D85 O12 L86 Z13

Keywords: Network centrality, Social network formation, Information technology, Development, Randomized experiment

*Derksen: Department of Management, University of Toronto (laura.derksen@rotman.utoronto.ca); Souza: Department of Economics, Queen Mary University of London (p.souza@qmul.ac.uk) We are grateful for insightful comments from Sebastian Axbard, Heski Bar-Isaac, Emily Breza, Nathan Canen, Stefan Dimitriadis, Alan Griffith, Yosh Halberstam, Jonas Hjort, Eliana La Ferrara, Simon Franklin, François Gerard, Matt Jackson, Nicola Lacetera, Matt Mitchell, Imran Rasul and Román Andrés Zárate, and seminar and conference participants. We thank Abdul Chilungo for project management assistance, and Kayla Crowley-Carbery, Catherine Michaud-Leclerc, Dina O'Brien and Ethan Sansom for their excellent research assistance. This study was approved by the University of Toronto Research Ethics Board and the Malawi National Committee on Research Ethics in the Social Sciences and Humanities (P08/17/204). The RCT was funded by a SSHRC Insight Development grant, and was pre-registered in the American Economic Association RCT Registry, number AEARCTR-0003824. All errors and omissions are our own.

1 Introduction

Individual network positions play an important role in many economic behaviors and processes (Jackson et al., 2017), yet the causal determinants of network position are not well understood. In particular, network centrality measures capture an individual's importance and influence across a broad range of domains, including academic performance and teamwork (Calvó-Armengol et al., 2009; Hahn et al., 2015; Breza et al., 2019b), financial risk-sharing (Field and Pande, 2018), political power (Bertrand et al., 2014; Cruz et al., 2017), social norms (Jackson, 2019a), reproductive health (Baumgartner et al., 2022) and technology adoption (Banerjee et al., 2013; Beaman et al., 2021). Centrality has also been shown to directly impact information diffusion (Banerjee et al., 2013; Jackson et al., 2017; Banerjee et al., 2019). The literature often views network position as fixed, or predetermined. Yet, information access might *cause* a person to become well-connected, and even central, as others form strategic links to obtain access.

In this paper we show that access to information has a causal impact on network centrality. We analyze a randomized intervention in Malawian boarding secondary schools, and collect complete panel data on network connections. One fifth of students were given exclusive access to a reliable, wide-ranging information source throughout the school year. Students selected for the treatment group were provided with internet access restricted to Wikipedia. They could access the information resource privately, after school and on weekends, in a digital library on school grounds. Students had no internet access outside of the intervention, no mobile phones, and little access to outside information or social contact.

This isolated setting provides an ideal experimental context in which to manipulate information access and map complete networks over time. Our panel data includes a complete and uncensored set of links across many types of social interactions, classified as either information-sharing links or personal friendships. The richness of this data allows us to conduct a thorough and nuanced analysis of structural changes to the network. We are able to document how particular types of links form or break in response to a node-level shock to information access, and estimate the impact on centrality across various measures and network definitions.

Indeed, the intervention represents a shock to information access for treatment students. Students used the resource intensively to search for many types of information, including information instrumental to their education, general knowledge (such as politics, health and world news), and entertainment. The average treated student spent one hour and twenty minutes per week accessing information, and visited nearly 900 different Wikipedia pages. The impact of this intervention on academic performance has been shown to be positive, and concentrated among lower ability students (Derksen et al., 2022).

Control students were also able to access the new information source indirectly by leveraging social ties; we observe widespread information diffusion from treated students to control students. In a two-week long incentivized experiment, we assigned a unique question about a recent news event to each student, and tracked correct answers as well as information sources. 51 percent of control students were

able to find the information they needed, despite no access to news media. While 93 percent of treated students reported finding the answer on Wikipedia, 65 percent of control students who found the correct answer had asked a treated friend.

Our main finding is that consistent, exclusive information access causes a long-run increase in network centrality. We observe impacts on five different measures of centrality, collected eight months after the start of the intervention. In the information-sharing network, we estimate a 9.5 percent increase in degree and a .18 standard deviation increase in eigenvector centrality for treated students relative to control students. We also observe an increase in diffusion centrality, with two different sets of parameters, and in betweenness centrality. These effects are highly significant, with randomization inference p-values less than or equal to .001. Students with information access are also more likely to be among the most central students by all measures (with $p < .1$). These changes in turn impact the full network, which includes both information-sharing links and personal friendships. Again, the effect is significant across all five centrality measures, with $p < .05$. We do not detect any significant effect on centrality in the personal friendship network or among other types of contacts. In combination, these results show that students form completely new links for information sharing, as opposed to relying on existing information links or personal friendships. We also show that the changes in network structure are driven by changes in the number of links as opposed to changes in link strength; treated students form more new links, and maintain more existing links than control students.

To further characterize these changes in network centrality, we examine strategic link formation at the dyad level. We observe a 9 percent relative increase in information-sharing links between treated and control students ($p = .005$), relative to pairs of control students. Indeed, by forming strategic links, control students can search for information they need but cannot otherwise access. Overall, new links are concentrated in the information-sharing network. Control students are also significantly more likely to name treated students, as opposed to other control students, as sources of personal advice and even as best friends, though these friendships do not appear to be reciprocated. Early on, control students might have formed links to treatment students to learn more about the new technology itself, rather than to access information more generally. Yet, this type of learning does not appear to explain our main results. First, we observe an increase in information-sharing links across a wide range of topics, measured eight months after the start of the intervention. Second, the effect is strongest for control students who do have past internet experience, as opposed to those who would have the most to learn.

We also observe a 18 percent relative increase in information-sharing links between pairs of treatment students ($p = .011$), despite having access to the same information source. This effect is strongest when at least one of the two lacks internet experience, suggesting that treated students with limited internet proficiency might ask other treated students to search on their behalf. Indeed, pairs of treated students share information across a broad range of topics. We see no increase in personal connections of any type between pairs of treated students.

The fact that we observe a significant change in network structure, and a change in the composition of central nodes, has potential implications for information diffusion. Information diffuses more widely from central, well-connected nodes, and policies that target central nodes may therefore amplify learning, behavior change, and technology adoption (Jackson et al., 2017). On the other hand, if a network responds sufficiently to the intervention, network-based targeting may be unnecessary or inefficient. Yet, our reduced form estimates do not allow us to fully characterize the effect of the intervention on network-level information diffusion.

We therefore introduce a model to demonstrate the implications of our findings for targeting and information diffusion, and to illustrate an important trade-off with respect to inequality. We specify and calibrate a dyadic model of strategic network formation. We simulate counterfactual policy experiments to show that information diffusion is less sensitive to network centrality when the information shock is large enough to generate a network response. While targeting information to nodes based on their network centrality does produce more diffusion than random targeting, the gap in diffusion is cut by half due the network response. This occurs because the provision of information itself creates more central nodes, even under random targeting. Moreover, because network centrality is typically correlated with other forms of advantage, network-based targeting can exacerbate inequality. This trade-off has important welfare implications in our setting, as the intervention was shown to directly benefit lower ability students who are, on average, also less central (Derksen et al., 2022).

We contribute to an emerging empirical literature on the determinants of network structure, by providing evidence for a causal determinant of network centrality: information access. Fowler et al. (2009) find that network characteristics are in part genetically determined, and Hasan and Bagde (2015) show that interacting with well-connected peers can affect a person's network position. Other empirical work has identified correlates of network centrality, and has focused on personality traits (Girard et al., 2015; Morelli et al., 2017). Our findings suggest that network centrality is determined not only by intrinsic traits but also by access to information, and is therefore subject to change over time and in response to policy. Recent empirical work has shown that social ties can be formed or strengthened in response to an intervention, without mapping complete networks or analyzing changes in network position. Dar et al. (2020), Fernando (2021), and Bertelli and Fall (2022) show that farmers learn from well-informed contacts outside of existing peer groups, and Berg et al. (2019) show that incentives for information diffusion can overcome the barrier of social distance. Stein (2021) shows that entrepreneurs form strategic links to other entrepreneurs who have access to a formal training program, and Dimitriadis and Koning (2022) find that social skills training can encourage profitable peer connections between entrepreneurs.

Other empirical work has focused on broader network changes at the community level, with networks defined by context-specific social connections and interactions. The fact that different interventions naturally impact different types of links highlights the value of collecting and analyzing comprehensive and detailed network data. Work by Binzel et al. (2017) and Banerjee et al. (2022) has shown that formal mi-

crofinance can crowd out informal lending networks, resulting in changes to the structure of the network at the community level, and Feigenberg et al. (2013) find that social contact increases between members of the same microfinance group. Heß et al. (2021) find that additional public development funding can also cause a decline in informal economic ties, as elite capture leads to an erosion of social capital. Finally, Delavallade et al. (2016) show that randomly selecting students for an after-school program can affect link formation and produce segregation in the friendship network.[1]

Our findings demonstrate that an exogenous change in information access can drive strategic link formation. In random models of network formation, link probabilities might be based on homophily (Girard et al., 2015; Pin and Rogers, 2016), or on endogenous network characteristics such as degree (Barabási and Albert, 1999). On the other hand, economic theory posits that people invest in social ties for strategic reasons, including to increase information access (Jackson and Wolinsky, 1996; Calvó-Armengol et al., 2015; Capozza et al., 2021).[2] Banerjee et al. (2021) find that even when information is publicly available, information sharing across social ties can lead to broader diffusion and greater understanding. Indeed, access to information is a source of advantage in many economic models, and knowledge acquisition is fundamental to human capital formation. A vast empirical literature has shown that access to information technology has important impacts across a wide variety of economic and political domains (Jensen, 2007; Bailard, 2012; Miner, 2015; Galperin and Viecens, 2017; Campante et al., 2018; Chen and Yang, 2019; Hjort and Poulsen, 2019; Derksen et al., 2022). In our study, rather than providing a bundle of information and communication technology, we restrict our intervention to a pure information source.

This study has implications for policies that target individuals based on their network positions. Several studies have shown that simple, one-time messages diffuse more widely in the network when central messengers are targeted (Banerjee et al., 2013, 2019). These studies also show that it is more effective to target based on measures that are grounded in network theory, as opposed to less sophisticated measures of social influence (Kim et al., 2015). Our results suggest that using centrality measures to to select candidates for a longer term intervention or role may be less effective, as the network adapts over time. On the one hand, Baumgartner et al. (2022) find that using centrality measures to select peer educators, rather than teacher recommendations, is more effective for teen pregnancy prevention. Beaman et al. (2021) also find that selecting model farmers based on their network position leads to greater technology diffusion than the status-quo in which extension workers select model farmers. Yet, neither of these studies includes a random selection arm. Indeed, two studies that do compare network-based targeting with random selection both find no difference in the diffusion of agricultural knowledge (Beaman and Dillon, 2018), as farmers frequently seek information outside of their existing networks (Dar et al., 2020). Moreover, network-based targeting can be costly to implement in practice, as it typically involves collecting detailed survey data, if not complete network data (Banerjee et al., 2019; Breza et al., 2019a).

[1] This paper relates to another strand of experimental literature involving direct network manipulation, to measure the effects of having certain types of peers (Sacerdote, 2001; Hasan and Bagde, 2015; Zárate, 2021).

[2] Other applied theoretical work has focused on risk sharing as a motivation for link formation, see for example Bramoullé and Kranton (2007a), Bramoullé and Kranton (2007b) and Ambrus and Elliott (2021).

Network-based targeting strategies can also exacerbate inequality, as central individuals are likely privileged in other ways (Jackson, 2019b). First, many interventions have direct impacts, and targeting decisions can have important welfare implications. In our setting, the intervention had a large direct effect on academic performance and reading ability for lower ability students (Derksen et al., 2022), yet high ability students are much more likely to be central in the network. If we had used network-based targeting, we would have nearly eliminated the first-order impact of the intervention. Second, while targeting improves information diffusion, information that spreads in this way is typically more likely to reach the wealthy and well-connected (Singh et al., 2010; Beaman and Dillon, 2018; Bandiera et al., 2022). Our results also point to a third source of inequality: targeting can make already-influential nodes even more influential by increasing their network centrality. On the other hand, this impact is concentrated primarily in the information-sharing network. Personal friendships appear largely resilient to even a large and sustained information shock.

Finally, our results illustrate an important downside of using baseline network data to estimate peer effects. Both treatment effect and peer effect estimates may be biased if the network responds endogenously to the intervention. Recent work in applied econometrics by Comola and Prina (2021) and Griffith (2022b) has demonstrated this possibility by analyzing network data in the context of an intervention (a financial access intervention and after-school empowerment program, respectively), while offering alternative strategies to estimate treatment effects and peer effects.

This paper proceeds as follows. In Section 2 we describe the setting, experimental design and information seeking behavior among students. In Section 3 we describe our network data. We present our empirical strategy and results in Section 4. In Section 5 we specify and calibrate a theoretical model of strategic network formation. In Section 6 we conclude.

2 Providing Access to Information

We designed an experiment with the goal of obtaining a clean measure of the effect of exclusive information access on network position. We selected a unique, naturally isolated experimental environment with limited baseline information access. This allowed us to provide a significant information shock and map complete social networks. Second, we randomly assigned individuals to obtain information access, and not groups or entire networks. We are therefore able to estimate node-level impacts on network position; this is novel relative to existing literature that primarily uses network-level treatment assignment and analyzes overall network structure (Feigenberg et al., 2013; Banerjee et al., 2022; Heß et al., 2021). Finally, while direct access to information was strictly limited to specific nodes in the network, we allowed information to be freely transmitted from there onward within the bounds of the experimental setting. We now detail the experimental setting and intervention, we describe how the intervention was effectively used for information access, and we document widespread information sharing between peers.

2.1 Experimental design

Setting. Malawi is a low-resource country in southern Africa, where internet access is limited but expanding rapidly. As of 2015, 54 percent of households had a mobile device and 12 percent of individuals had ever used the internet (DHS 2015-16). Data connections though 3G or 4G networks are now available in urban areas and 2G is available in most rural areas (Batzilis et al., 2010).

The experiment took place in four boarding schools, all of which are government secondary schools. Admission is competitive and based on standardized exam scores. Secondary school is not free in Malawi, but bursaries and scholarships are common, and many of the students come from lower socioeconomic backgrounds. Two of the schools are single-sex national boarding schools run by the Catholic church, which accept girls and boys (respectively) from across the country. The other two schools are co-educational district boarding schools. Students progress through four forms (grade levels), and each form is divided into three different classrooms.[3]

Intervention. During term time, students have few sources of outside information. At school, mobile devices are not permitted. Schools have computer rooms and computer classes, but with no internet access. Students can read books they brought from home, or borrow books from a small school library, and can of course speak to teachers and to each other.

At each school, we provided a small subset of randomly-selected students with private access to online information for the duration of the school year. Specifically, we provided access to Wikipedia, an open source of detailed and up-to-date information on a wide range of topics. By restricting internet access to Wikipedia, we deliver access to a pure information source, as opposed to the bundle of information, communication, entertainment and interaction available on the wider internet. While the information available on Wikipedia may be entertaining, it takes the form of information rather than entertainment. For example, a Wikipedia may describe the plot of a novel, but it does not contain full works of fiction. The information on Wikipedia is contributed and edited by volunteers, yet often accurate (Giles, 2005). Wikipedia is the largest and most visited reference site on the internet.[4]

One room at each school was designated for use as a digital library after school and on weekends. It was open for four hours on most weekdays, and for eight hours on Saturday and Sunday. The library was open for between 20 and 22 weeks total in each school; it was occasionally closed due to exams or other events. It was managed by a digital librarian hired by the research team, and equipped with twelve Android devices. We installed an application on each device that allowed us to restrict student access to applications and websites. Students could access online information via Wikipedia and Wiktionary domains on Google Chrome, but were not able to access any other applications or websites. Students were allowed to take notes and to share information outside of the library, but they were not allowed to

[3]See Derksen et al. (2022) for additional detail on the educational setting.

[4]Source: Wikipedia, https://en.wikipedia.org/wiki/Wikipedia, accessed on December 13th, 2021. Wikipedia is free and owned by Wikimedia, a non-profit organization with no advertising.

work in groups inside the digital library. This was intended to prevent network change by means other than through information access; if students had been permitted to socialize in the digital library this could have led to link formation that was not driven by information access. Librarians supervised student use of the digital libraries for the entire duration of the intervention. They did not monitor the specific pages students visited.

Before the study began, the digital librarian visited every classroom in the school to introduce the program to students. The digital librarian informed all students of the nature of the program, including an explanation of Wikipedia, and the randomized study design. The librarian emphasized that while only a few students would be selected for the program, they could freely share information they found online with their friends.

Randomization. We next randomized students to either the treatment arm or the control arm. We stratified the sample based on school, form, whether the student had ever used the internet, and whether the student's baseline exam score was above the median. To construct the baseline score, we used administrative data from the end of the previous school year, and calculated the average of the student's English and Biology scores (core subjects for which we have nearly complete data). Within each stratum, we assigned only one fifth of students to the treatment arm. This process resulted in 301 treated and 1,207 control students across 51 strata. We chose to treat a small fraction of students with the social network in mind; our goal is to investigate the effect of providing certain nodes with rare and exclusive access to information. Indeed, if the resource had been made widely available, students would have been able to access information firsthand or through existing links, which may have dampened the effect on network structure.

Students in the treatment arm were invited to an induction session, where they learned how to use the devices to search for information, and about the privacy protections that were in place. In each session, students picked a username from a hat, which they would use to log into the devices. In this way, students knew that their browsing behavior was anonymous. Yet, the username does allow us to track browsing behavior throughout the year for individual students, and to associate behavior with some coarsened student characteristics; we constructed induction groups (and username codes) stratified by school, gender, above-median baseline exam score, and above-median baseline social network degree.

Treatment students could browse Wikipedia by visiting the digital library during opening hours and signing out a device for use within the library. If a student arrived at the library and all devices were in use, they were placed on a waitlist, and device use was limited to 30 minutes. The digital librarian was responsible for checking the student's identity, recording arrival and departure times for each student, managing the Android devices, and supervising the library. The librarian ensured that students used the devices quietly and individually, and did not remove devices from the room. Control students were not allowed into the digital library, and therefore did not have direct access to the devices.

Data collection. We conducted a baseline survey, and collected administrative data on past exam scores. The baseline survey captured complete social networks for all students in Forms 2, 3 and 4. We excluded Form 1 students from the study, because they had only just arrived at the school, and their baseline social networks would have been noisy or non-existent. We also collected survey data, including social network data, from the full sample of students at endline, as well as supplementary survey data from all treatment students and a random subset of control students. We also collected administrative data on student exam scores throughout the year. Finally, we uploaded browsing data directly from the devices throughout the intervention. We provide further detail on network data in Section 3.

2.2 Information Seeking Behavior

Use of the information source. We collected granular internet browsing data at the individual level throughout the intervention period. We observe timestamped page visits for each username, which we can associate to coarsened student characteristics. We classify these pages to broad topics using the Wikipedia category tree, to specific news events highlighted on Wikipedia, and to school subjects using the Malawi secondary school syllabus. For additional detail on topic classifications, see Derksen et al. (2022).

The students used the online resource frequently, found Wikipedia trustworthy and easy to use (Derksen et al., 2022), and browsed pages across a wide range of topics. Every treatment student used the digital library at least once, and the average student visited 33 times. The average student spent one hour and twenty minutes per week in the digital library and visited nearly 900 different Wikipedia pages. These pages span many different topics, including general interest topics (see Figure 1), topics related to sex and sexuality (7 percent of browsing time) and topics related to the school syllabus (22 percent of browsing time). The students also used Wikipedia to read about news events. At the time of a major event, we observe a significant spike in browsing activity on related pages, especially if the news concerns Africa (see Figure 2). General browsing patterns are explored more fully in Derksen et al. (2022).

Information spread. In the last month of the intervention, we conducted an incentivized information-seeking experiment to determine to what extent students were able to access information at school. The experiment involved all treatment students and a random subset of 298 control students. Each student was given two unique multiple-choice "quiz" questions. The first question was about a recent news event, and the second was about an academic subject. Examples of questions include "*Who won the 2017 Nobel Peace Prize?*" (news) and "*Where is insulin produced?*" (academic). Students had approximately two weeks to find the answers to these questions, and were told that correct answers would be entered into a prize draw.

We find that control students formed or leveraged social ties with treatment students in order to find the information they needed. In Figure 3, we show the percent of students who found correct answers to

their quiz questions, overall and by information source. The majority of students in both treatment and control groups were able to find the correct answers. 59 percent of treatment students and 51 percent of control students were able to correctly answer the news question, despite no access to news media. 68 percent of treatment students and 57 percent of control students found the correct answer to the academic question. After the experiment, we asked students where they had found the answers to their quiz questions. The vast majority of treatment students who found the correct answer report doing so using Wikipedia (93 percent for news, and 83 percent for the academic question). Among control students, the source of academic information was varied. Most asked a friend (57 percent), but others asked a teacher (7 percent) or found the answer in the school library (16 percent). For the news question, 69 percent of those in the control group who found the correct answer had asked a friend, and the vast majority of these friends were in the treatment group (94 percent).

3 Network Data

3.1 Definitions and Measurement

We conducted baseline and endline surveys with the primary goal of collecting detailed and complete measures of social networks. We surveyed all students present at school and measured many different types of links; relying on subsamples of networks can lead to mismeasurement of network characteristics Chandrasekhar and Lewis (2016). We also allowed for an unlimited number of links between students. Indeed, many network surveys limit the number of links a person can report, and this type of censoring introduces bias in centrality estimates as well as in other peer effect estimators Griffith (2022a).

We grouped links into two overlapping networks: information-sharing networks and personal friendships. The information network is composed of five sub-networks. We asked students to list the schoolmates they rely on for information by topic, including music, sports, entertainment, school, news, health, and school activities or topics learned in class. Personal friendships capture a range of interactions at a more individual and intimate nature, and are also constructed from five survey questions. Personal friendships include schoolmates who are "best friends", who have lent the student money or something else, have given the student a gift, or are relied on for advice. We list the full set of social network survey questions in Table 1.[5]

We plot the adjacency matrices for each of these sub-networks for a single school in Figure 4. In this figure, students were ordered first by form, and then by classroom. A dot represents a link between students. It is clear that students are more likely to form links within classrooms, and within forms. We also observe some across-form links, although much less frequently.[6] For this reason, we focus our

[5]We also measured a more general contact network. This network is based on the question "*[Yesterday/Two days ago/Three days ago], did you just hang out, have conversations or play with friends?*'

[6]Approximately 5 percent of information-links are across-form links, at both baseline and endline.

analysis on within-form networks.

While there is substantial overlap between the information and personal networks, they are in fact distinct and differ along several dimensions. At baseline, the average student has 10.8 information links and 6.4 personal links within their school and form, with 13.5 links overall. This implies that 58 percent of pairs who are linked in the personal network are also linked in the information network, and 34 percent of information links are also personal.

We use this data set to construct networks at baseline and endline within each school-form, with on average 117 students per network. A network g^f is a set of $N_f \times (N_f - 1)$ potential links, where form f has N_f nodes. We set $g^f_{ij} = 1$ if there is a link in either direction between students i and j, both of whom are in the same form and school. We consider three distinct sub-networks: the information network g^I, the personal friendship network g^P, and the overall network g which consists of a union of both types of links. We also explore alternative link definitions including directed links and the intersection of directed links. For directed networks, we define a student's *in-degree* as the number of others who nominate a particular student as a source for information, money, gifts or advice. The *out-degree* is the number of others the student nominates.[7] At baseline, 26 percent of directed information links and 40 percent of directed personal links are reciprocated. These patterns are similar at endline: 23 percent of information links and 38 percent of personal links are reciprocated.

3.2 Centrality Measures

We can use our network data to calculate several standard network centrality measures (Bloch et al., 2019). The simplest measure of centrality is the *degree*, defined as the number of other nodes to which i is linked,

$$d_i(g) = \sum_j g_{ij}.$$

Figure A1 shows the degree distribution at baseline and endline, based on the information network. The average degree is 10.1 and 10.3 at baseline and endline respectively, and the distribution has a substantial tail of well-connected students.

Eigenvector centrality captures not only the extent to which a node is connected to other nodes, but also the extent to which those other nodes are themselves highly-connected. This measure is motivated by the premise that the importance of a node depends on the importance of neighboring nodes. The eigenvector centrality of a node i, $C^e_i(g)$, is defined in a recursive way, to equal the sum of the centralities of its neighbors:

$$\lambda C^e_i(g) = \sum_j g_{ij} C^e_j(g).$$

Eigenvector centrality was originally proposed as a measure by Bonacich (1972) and is widely used in the

[7] There is one exception: we invert the direction for the personal subcomponent "Who have you given a gift to at this school?", so that the out-degree always represents a reliance on others, i.e. for advice, loans or gifts.

literature (Jackson, 2010). [8]

We also investigate two versions of *diffusion centrality* (Banerjee et al., 2013, 2019), which capture the extent to which an informational shock reaches other nodes in the network. The diffusion centrality of a node is

$$C^d(g) = \sum_{t=1}^{T} (qg)^t$$

where q is the probability that the information is transmitted among two connected individuals and t are the number of iterations on the network. In the first version of diffusion centrality, we set $q = 1$ and $T = 2$ (labelled as "number of length-2 walks"). This captures the extent of diffusion when information passes along every link for two periods. This is an extension of degree centrality, and represents the number of walks of length two originating from a particular node. In the second, more generalized version of diffusion centrality, we follow Banerjee et al. (2019) and set q equal to the reciprocal of the top eigenvalue, while T is equal to the diameter of the graph. This captures the extent of information diffusion that occurs over a longer time horizon but with lower probability of transmission at the link level. Diffusion centrality and eigenvector centrality are closely related: in networks with high transmission rates, diffusion centrality approaches eigenvector centrality as T tends to ∞ (Banerjee et al., 2019).

Finally, *betweenness centrality* captures the importance of a node as an intermediary along paths between other pairs of nodes in the network. Define $P_i(j,k;g)$ to be the number of shortest paths between j and k that pass through i on network g. Then, betweenness centrality is

$$C_i^b(g) = \sum_{\substack{j,k=\{1,\ldots,n\} \\ j\neq k \\ j,k\neq i}} \frac{P_i(j,k;g)}{(n-1)(n-2)}$$

as there are $(n-1)(n-2)$ potential (j,k) pairs. This measure was first introduced in Freeman (1977) and is widely used as a notion of node-level exposure to information, effectively as an intermediary of its transmission.

Throughout our analysis we normalize eigenvector centrality, diffusion centrality and betweenness centrality for ease of interpretation. We normalize by subtracting the within-form endline control group mean and dividing by the within-form endline control group standard deviation.

In addition to these centrality measures, we compute each node's average link strength. For a linked pair of nodes i and j, the strength of the link is defined as the share of subcomponents that underpin the link, as defined in Table 1. An information link or personal link has five potential subcomponents, and a full-network link has ten. A particular node's average link strength is the average strength calculated

[8] The equation above can be equivalently expressed as $\lambda C^e(g) = gC^e(g)$, and $C^e(g) = [C_1^e(g), \ldots, C_n^e(g)]'$. Typically λ is selected as the largest eigenvalue associated with the adjacency matrix g. By the Perron-Frobenius theorem, the largest eigenvalue is associated with eigenvectors with positive entries, and thus C^e is non-negative.

across that node's existing links.

3.3 Balance and Summary Statistics

Our randomized assignment is balanced on these and other network centrality measures among the 1,508 students surveyed at baseline (Table 2). The randomization is also balanced on non-network student characteristics (Derksen et al., 2022). We attempted to survey all students again at endline, and collected data for 1,402 students; the remaining 106 students were not present and many of these students likely left school. We exclude these students entirely from our network analysis, as their endline network positions cannot be clearly defined or interpreted. That is, the networks we construct at baseline and endline are defined on the same set of 1,402 students present at both times.[9]

At baseline, network centrality is highly correlated with certain student characteristics (Table 3). Both degree and eigenvector centrality are positively correlated with academic ability, female gender, and with socioeconomic status (SES), as measured by the presence of electricity and running water at the student's home. This is particularly relevant in the tail of the distribution, which appears to be dominated by high ability students. They are, for example, 5 percentage points more likely to be in the top 5 percent of the distribution according to eigenvector centrality (Column 7 of Table 3). Student browsing behavior, on the other hand, does not differ significantly by baseline network position (Columns 2, 4, 6 and 8 of Table 3). Taken together, these correlations suggest that networks are not random, and that beyond information access, student characteristics likely play a role.

We next characterize how the networks we computed compare to networks studied elsewhere in the literature. In our setting, students can interact in classrooms, in extracurricular activities, during meal times and, owing to the fact that these are boarding schools, in their residences. These types of interactions are not unique to our setting – in fact they are likely similar across many educational settings. Moreover, the broad range of links we capture, which involve information sharing, gifting and lending, and asking for advice, likely have analogs in many social contexts. The richness of our network data is evident when we compare our network descriptive statistics to those of other networks that have been captured in the literature. In Table A1, we see that when we include all types of links (the *full network*), we observe an average degree of 12.7. This corresponds to a more than a 50 percent increase in links compared to the networks captured in Banerjee et al. (2013) or Coleman (1964). This point is made clearer in a comparison of our networks with the friendship nominations networks obtained from the National Longitudinal Study of Adolescent Health "Add-Health").[10] In Panel A of Appendix Figure A2 we see that the Add-Health in-

[9] While attrition is low in both treatment and control groups, it is significantly higher in the control group (8 percent versus 5 percent, Table 2, Panel C). We mitigate the potential effects of attrition in our main analysis by controlling for baseline centrality measures and other covariates. Attrition is concentrated in the two district boarding schools; in the two national boarding schools the attrition rate is only 3 percent, with no significant difference between treatment and control. As further discussed in Section 4, we are able to replicate our main results on this low-attrition sub-sample.

[10] From the data freely available at https://www.icpsr.umich.edu/web/ICPSR/studies/21600/datasets/0003/variables/ODGX2?archive=icpsr and https://www.icpsr.umich.edu/web/ICPSR/studies/21600/datasets/0003/variables/IDGX2?archive=icpsr.

degree distribution is comparable to either the personal or information network, but we see a substantial shift to the right for the full-network degree distribution. Students interact in ways that are not captured by either of the separate networks in isolation. This highlights the importance of capturing many different types of links. The AddHealth out-degree distribution (Panel B) sharply falls at 10 nominations. This is unsurprising given that friendship nominations were capped at that number. In contrast, we did not cap the nomination process, resulting in the ability to fully observe the right-hand tail of the distribution in the full network and in both sub-networks.

Having described and contextualized our network data, we now proceed to the results section where we estimate the causal impacts of information access on individual network positions and link formation.

4 Results

In Section 2.2, we saw that treated students used the new information source intensively, accessed information on a broad set of topics, and shared information widely. Indeed, we document widespread information diffusion from treated to control students. Most control students were able to find information only available online, despite no internet access, by asking a treated student to find it for them.

In this section, we empirically investigate the causal impact of this exclusive information access on network structure. Our main empirical results are divided in two parts. We first use individual-level linear regressions to estimate the treatment effect on centrality measures. We then move to dyad-level regressions to examine strategic link formation. In the next section we will calibrate a theoretical model to demonstrate the importance and consequences of the network changes we observe for network-based targeting, inequality and welfare.

4.1 The Effect of Information Access on Network Centrality

We estimate the impact of information access on individual-level network centrality measures with the following specification.

$$\text{centrality}_i^1 = \beta \cdot T_i + \alpha \cdot \text{centrality}_i^0 + \mathbf{x}_i'\boldsymbol{\chi} + \gamma_c + \lambda_s + \epsilon_i \tag{1}$$

Here, centrality_i^1 is an endline centrality measure for student i. Centrality measures are computed within school and form. T_i is an indicator for treatment status. We control for the outcome measure at the baseline centrality_i^0, classroom fixed effects γ_c, as well as other individual-level covariates $\mathbf{x}_i$, to increase precision (McKenzie, 2012).[11] We also include stratification-bin indicators λ_s. We use ordinary least squares to obtain an estimate $\hat{\beta}$ for the causal impact of the intervention on network centrality. We compute classical heteroskedasticity-robust standard errors, but rely on randomization inference to construct p-values, as

[11] Covariates include Gender and SES, defined as household having electricity and running water. Academic ability is captured by the stratification bin.

standard errors may be biased when an intervention is randomly assigned to nodes in a social network (Abadie et al., 2016).

This approach to estimation and inference allows us to detect a *relative* change in centrality, in comparison to the control group, as opposed to an *absolute* change. For example, if we find $\hat{\beta}$ to be positive, that could indicate an increase in links for treated students, a decrease in links for control students, or some combination thereof.

Main results: effects on the information network. We find large and significant treatment effects on network centrality in the information-sharing network. Panel A of Table 4 contains estimates of these treatment effects. We find a notable increase in network centrality for students randomly selected to gain information access, across all five measures of centrality. Column 1 shows that treatment students on average have .96 additional links relative to control, from a mean of 10.1 links; this represents a 10 percent increase in links. This effect size is similar in magnitude to the effect of having high socio-economic status (1.33 additional links, see Table 3). Column 2 of Table 4 shows that eigenvector centrality is .18 standard deviations higher for the students that were exposed to the treatment. This suggests that beyond having a higher number of links, treated students are in more prominent network positions. These effects are significant with randomization inference p-values less than or equal to .001.

These treatment effects are illustrated in Figure 5. In Panel A we plot endline degree against baseline degree separately for treated and control students. The corresponding plot for eigenvector centrality is in Panel C. We re-scale the axes in terms of percentiles for ease of interpretation. The plots show a distinct positive correlation between baseline and endline centrality. This is not surprising as central students at baseline tend, on average, to have higher centrality at endline irrespective of treatment status. Importantly, the figure shows a level upward shift of the treated group compared to the control group – which illustrates and is attributable to the treatment effect. Moreover, the increase in centrality appears to be evenly distributed across the baseline distribution. It is not, for example, the case that treatment effects are concentrated among students initially in the upper tail of the distribution. Panels B and D show the distributions of degree and eigenvector centrality at endline, again by treatment status. We again see that treatment increases centrality, shifting the distribution to the right. At endline, we see a high relatively high prevalence of treated students in the tail of the distribution, for both centrality measures.

Columns 3 and 4 of Table 4 suggest that treated students are better positioned for information diffusion: the number of length-2 walks increased approximately 8.6 percent, and diffusion centrality increased by .18 units of standard deviation relative to the control group. Treated students are also more likely to act as intermediaries in the network: betweenness centrality is .24 standard deviations higher in the treated group. All effects are highly statistically significant with randomization inference p-value of at most .001. Finally, there is a small but insignificant positive impact on average link strength (Column 6 of Table 4 and Panel A of Figure A3). This outcome captures an intensive margin of the treatment; a positive

effect indicates an increase in interactions with preexisting information links. The impacts we observe are in fact more consistent with information-seeking behavior outside of students' pre-existing information networks.

Beyond the average treatment effect on centrality, treated students have a higher probability of being *central* at the endline, that is, appearing in the tail of the distribution (Panel B of Table 4). The most central nodes in a network, sometimes referred to as *hubs*, often play a particularly important role in network processes such as information diffusion (Banerjee et al., 2013). While the estimates in Table 4 are subject to some imprecision, the magnitudes are large. For example, the intervention increases the probability of being in the top 5 percent by eigenvector centrality by approximately 2.4 percentage points, significant at the 10 percent level. For context, this coefficient is comparable to the effect of moving from low to high SES (see Column 7 of Table 3). Similar effects are observed across other centrality measures. For the number of length-2 walks, diffusion and betweenness centrality, effects are significant at the 5 percent level.

Effects on the personal and full networks. In the personal network, we find near-zero or slightly negative effects on centrality measures as well as on average link strength, with randomization inference p-values above .620 in all cases (Table 5, Panel A). At the outset it was not clear to us whether the intervention would have an effect on link formation beyond links related to the transmission of information found online. These results suggest that the impact on the network indeed operates through information-sharing links as opposed to other forms of friendship or status.

While only information links appear to be directly affected by the treatment, this has significant implications for the full network. Indeed, treatment effects in the full network are comparable to the effects in the information network (Panel B of Table 5). Degree increases by .82 in the full network, compared to .96 in the information network (see Column 1 in Table 4). Point estimates for the other four centrality measures are also similar, and there is again no impact on average link strength. Taken together, these findings indicate that students are not simply adding information links to existing personal links. Otherwise, we would have expected to see null effects in the full network, as the full network represents the union of personal and information links, along with an increase in full-network link strength. Indeed, the overlap between the personal network and information network actually decreases over time. At endline, 31 percent of information links are also personal friendships, compared to 34 percent at baseline. Taking all types of links into account, the results appear to be driven by the *extensive margin*, as opposed to the strengthening of pre-existing connections.

Mechanisms and robustness. We now provide additional evidence to further rule out competing hypotheses. It could be the case that treated students simply spend more time socializing, for example, with each other. Moreover, the intervention could make treated students more popular for reasons of status or other reasons unrelated to information access. A third possibility will be discussed in Section 4.2: that

students are not forming links to gain access to information in a broad sense, but are interested in learning specifically about information technology. That is, they form links to the treatment group to learn about the new technology and how it is used.

We do not find that treatment students simply spend time with a higher number of contacts. In Table A2 we estimate the impact on the contact network, which is constructed using the three-day recall question *"[Yesterday/Two days ago/Three days ago], did you just hang out, have conversations or play with friends?"* We find that no effects on centrality, suggesting that the treatment is not mechanically promoting other types of interactions that involve simply spending together. Randomization inference p-values are above .497 in all regressions. This is compounded by the evidence presented above showing that the personal networks are largely unaffected.[12] These results are not surprising, as use of the digital library was limited to quiet, individual browsing, and this was enforced by our supervising librarians throughout the intervention.

We next turn our attention to the issue of whether the network changes we observe could be driven by perceptions or changes in status as opposed to real changes in information access. By interacting treatment status with an indicator for high use of the mobile library, we are able to perform a sort of placebo test.[13] This regression must be interpreted with caution as use of the digital library is endogenous. The coefficient cannot be interpreted as a treatment effect, as students with high browsing times form a selected sample, and we are unable to compare to similar students in the control group. Nevertheless, in Panel A of Table A3 we see that those simply belonging to the treatment group but not using the digital library do not have significantly more links than the control group. The difference only materializes for those individuals who also made above-median use of the digital library. This is consistent with the idea that actually accessing online information is driving the treatment effects.

Heterogenous effects by baseline academic ability, SES, gender, and baseline degree are largely absent, though some estimates are imprecise (Table A3, Panels B to E). This latter finding is consistent with Panels A and C of Figure 5, in which treatment effects appear broadly homogeneous by baseline degree. This approximate uniformity of treatment effects across individual-level characteristics will allow us to specify a simple yet informative model in Section 5.

The changes in the network we observe are due to both new and maintained links, and our results are robust to alternative definitions of the network. Columns 1 and 2 of Table 6 show the treatment effects on link creation and destruction. We find that access to information both causes link creation and prevents link destruction between baseline and endline, with the former effect appearing to dominate (.647 versus -.316). In Column 3, we define the network based on the intersection of directed links. That is, for a link to exist in this network, it must be reciprocal. While the effect sizes are smaller, estimates remain significant and the broad conclusions are unchanged. In Columns 4 and 5, we use directed networks to decompose the main effects into in- and out-degrees. We find that the main results are driven by an increase in

[12] This is also consistent with our data on student time use, shown in Table 5 of Derksen et al. (2022). Treated students substitute away from recreational activities in a magnitude comparable to the take-up of the digital library.

[13] "High browsing" is defined as above-median hours of use across the duration of the experiment.

in-degree. Treatment students are more likely to be nominated by others, and also nominate more links themselves, but the latter effect is not significant. Finally, in Column 6, we calculate a student's weighted degree by adding up all of their link strengths. We find that weighted degree increases in line with our other treatment effects. Across this table, effects are broadly present in the information and full networks (Panels A and C) and absent from the personal network (Panel B).

Finally, we show that the main results are robust to reasonable alternative specifications of the empirical models. In Table A4 we remove all controls except for stratification bin indicators. The results remain broadly significant and of similar magnitude, although with lower precision. In Table A5 we explore robustness to attrition. A small number of students were not present at endline and may have left school, and control students were more likely to attrit than treated students (see Panel C of Table 2). We therefore restrict the sample to the two National schools. These higher quality schools had very low attrition (<3 percent) and, importantly, there was no differential attrition between treatment and control. The results are very similar to the full-sample regressions in Table 4, and remain broadly significant.

4.2 Link Formation

Undirected links. We next estimate the impact of treatment status on the probability of an endline link between students i and j. We use the following dyadic regression specification:

$$100 \times \text{link}_{ij}^{1} = \beta_0 + \beta_1 \cdot TC_{ij} + \beta_2 \cdot TT_{ij} + \alpha \cdot \text{link}_{ij}^{0} + \mathbf{x}_{ij}'\boldsymbol{\chi} + \epsilon_{ij} \tag{2}$$

where $\text{link}_{ij}^{1} = 100$ if there is a link between i and j at the endline, and zero otherwise. We scale the outcome by 100 to easily interpret coefficients as percentage point increases. TC_{ij} is an indicator that is equal to one if one student is in the treatment group and the other is in the control group, and TT_{ij} is an indicator for both students being treated. link_{ij}^{0} is an indicator for a link at baseline. Covariates $\mathbf{x}_{ij}$ are the indicators for same gender, same classroom and form fixed effects. We only consider links within the same school and form. For each school-form f, the number of observations is the number of potential undirected links: $N_f(N_f - 1)/2$, as we do not include both ij and ji. We estimate this equation separately for the information network, the personal friendship network, and the union network that includes both types of links.

Our parameters of interest are β_1 and β_2. β_1 is the increased probability, in percentage points, of a link between a treatment student and a control student, relative to the likelihood of a link between two control students. β_2 is interpreted as the increased likelihood, in percentage points, of a link between two treatment students, relative to the likelihood of a link between two control students. We use linear regression to estimate these parameters. TT_{ij} and TC_{ij} are randomly assigned and independent of ϵ_{ij}, suggesting we might interpret these parameters as causal, relative to pairs of control students. We again report p-values based on randomization inference.

Directed links. We estimate a similar specification for directed links:

$$100 \times \text{link}_{ij}^{1} = \beta_0 + \beta_1 \cdot TC_{ij} + \beta_2 \cdot CT_{ij} + \beta_3 \cdot TT_{ij} + \alpha \cdot \text{link}_{ij}^{0} + \mathbf{x}_{ij}'\boldsymbol{\chi} + \epsilon_{ij} \tag{3}$$

where $\text{link}_{ij}^{1} = 100$ if there is a link from i to j at the endline, that is, if i nominates j, and zero otherwise. In this specification, TC_{ij} is an indicator that is equal to one if i is treated and j is in the control arm, and CT_{ij} is an indicator that is equal to one if i is in the control arm and j is treated. link_{ij}^{0} is an indicator for a directed link at baseline, and other covariates $\mathbf{x}_{ij}$ are as above. For each school-form f, the number of observations is $N_f(N_f - 1)$.

In the directed specification, we interpret β_1 as the increased probability, in percentage points, of a link from a treated student to a control student, relative to the likelihood of a control-to-control link. β_2 is the increased probability of a control-to-treated link, relative to the likelihood of a control-to-control link.

Main results: link formation. We find that, at endline, links involving at least one treatment student become significantly more common than links between two control students (Table 7). The probability of a link between a pair of control students is 8.43 percent. Treated-control pairs are .735 percentage points more likely to connect, a relative 9 percent increase. The effects are even higher for pairs of treatment students (1.49 percentage points, or 18 percent), though estimated with less precision. In the directed specification (Panel B of Table 7), we see an increase in both treated-to-control and control-to-treated information links, but the latter effect is larger. That is, treated students are particularly likely to be *named by others* as information contacts. The dyadic results are entirely consistent with the individual-level effects presented in the subsection above: the effects are driven by changes to the information network, with no significant effect on the personal network, and these changes result in significant differences in link probabilities in the full network.

The vast majority of the links causally induced by the treatment are between treated and control students. In our experiment, the mass of treated-control pairs is substantially larger than treated-treated (27k versus 3.3k links, respectively), overpowering the difference in the estimates in Column 1 of Table 7. Overall this suggests that approximately 50 both-treated links were induced by the experiment, compared to 199 treated-control links, a ratio of approximately 1 to 4. In short, the most important dynamics in the network seem to originate from treated-control links, despite the relatively lower point estimates for treated-control pairs as compared to both-treated.

Consistent with our individual-level results, we see limited variation in link strength based on the treatment status of students (Panels C and D of Figure A3), and that the changes to the network are due to both created and maintained links (Panel E of Figure A3). In Table A6, we estimate heterogeneous effects for links that were present or absent at baseline, as well as the direct effect on the number of created and broken links.[14]

[14]Column 1 of Table A6 shows that effects are larger for links that were present at baseline (3.68 percentage points versus .463).

Mechanisms. The fact that we observe an increase in both-treated links as well as an increase in treated-control links (see Column 1 of Table 7) suggests that link formation serves a purpose beyond pure information access. Pairs of treated students both have access to the same information source, yet do appear to benefit from links. It could be that some treated students are more proficient than others, and search on others' behalf or teach others how to use the new technology. Even between treated-control pairs, is is possible that link formation is driven by a desire to learn about information technology as opposed to a desire to gain information access more broadly. Next, we will show that while internet-proficiency differences do explain the increase in links between treated students, the desire to learn about a new technology does not appear to play a major role in treated-control links.

Learning about information technology. In Table 8, we interact control and treatment status with an indicator for whether the student had ever used the internet at baseline (this applies to approximately half of students). For treated-control pairs we find significant effects when at least one of them had prior internet use. The effects are strongest when the control student had used internet in the past. This suggests that control students are not primarily motivated by curiosity about information technology. Instead, these results suggest that students who already know the value of information technology seek out links for indirect access. On the other hand, pairs of treated students appear to form links when at least one student had *not* used internet before. This is consistent with treatment students learning about the new technology, or relying on friends who are more proficient to find information on their behalf. Indeed, next we will see that pairs of treatment students form links to discuss many different types of information, beyond discussing the technology itself.

Subcomponents of links: information topics and personal friendships. Information links between pairs of students appear to involve information sharing across a broad range of topics. To see this, we note that the information and personal networks are each composed of five subcomponents, based on the survey questions in Table 1. We can therefore explore link formation along each of those subcomponents. In Table 9, we see that students are creating links to discuss many different topics. When it comes to discussing entertainment, news and school activities, effects are large and significant for both treated-control pairs and pairs of treatment students. Both types of pairs also appear to discuss school subjects at similar rates, though the estimate is imprecisely estimated for pairs of treatment students. Health topics appear to be discussed somewhat less frequently; we observe lower point estimates that are insignificant at the 10 percent level. Health-related information is often sensitive and may therefore be less likely to circulate. We again find that the results are larger for control-to-treated links than for treated-to-control links; treated students are frequently named by others as information contacts across a range of topics (Panel B of Table 9).

Yet, 92 percent of pairs are *not* connected at baseline. Most of network change is in fact due to new links; the estimates imply that, for treated-control pairs, approximately 115 new links were created, compared to 89 additional maintained links.

Turning our attention to the personal network, we find null results for many subcomponents, with some notable exceptions (Table 10). We find an increase in undirected links between treated and control students formed for the purpose of discussing personal topics and offering advice. This is driven by an increase in directed links from control students to treated students, who are also more likely to name treated students as their best friends (Panel B of Table 10). This is perhaps unsurprising since, from the previous table and Figure 1, we know that students seek information regarding a vast range of topics, many of which may be classified as personal or form the basis of personal advice. Interestingly, control students appear to seek out treated students for this type of advice and friendship, while treated students do not seek each other out. Indeed, when information resources are available, students may prefer to learn about personal topics directly, in private.

5 Model and Calibration

In this section we characterize the importance of network-based targeting for information diffusion with and without endogenous network response, and explore implications for inequality and welfare. While our reduced form estimates indicate large and significant effects on network centrality, they do not allow us to fully explore the implications of these effects for network-level information diffusion. We therefore specify and calibrate a dyadic model of network formation, adopting the general setup in Jackson and Wolinsky (1996), and allowing strategic link formation based on information access. We generate basic theoretical predictions about link formation and centrality, extend and calibrate the model, and simulate policy counterfactuals. We find that under the endogenous network response, network-based targeting retains an advantage over random targeting in terms of information diffusion, but this comes at the cost of greater inequality and lower academic welfare. Importantly, the diffusion-advantage is reduced by half when the endogenous network response is taken into account.

5.1 A Model of Strategic Link Formation

Consider a set of nodes $\{1, 2, ..., N\}$. Each pair of nodes obtains utility 0 if there is no link between them, and u_{ij} if there is a link. This utility depends on the treatment status of each node. In particular, a node with exclusive information access might be more attractive, and this might also depend on the other node's information access. We model the utility of a link as follows:

$$u_{ij} = \kappa_{TC} T_i C_j + \kappa_{CT} C_i T_j + \kappa_{TT} T_i T_j + \nu_{ij},$$

where $T_i = 1$ is an indicator for the treatment group, and $C_i = 1 - T_i$ is an indicator for the control group, which we assume to be larger than the treatment group. Parameter κ_{TC} captures the benefit to *treatment* node of linking to a *control* node. This will be positive if the treatment node gets utility from sharing

information, and zero otherwise. κ_{CT} is the benefit to a control node of linking to a treatment node. We hypothesize that this parameter is positive, and larger than κ_{TC} as control nodes value increased access to information more than treated nodes like to share information. Finally, κ_{TT} captures the benefit to a treatment node of linking to another treatment node. We again hypothesize that this parameter is positive; a person with information access gets positive utility from talking to their informed friends, because they might search for and share different information.

The term ν_{ij} captures a sort of underlying benefit to node i of forming a link to node j, ignoring information access. If both nodes are in the control group, this captures the entire benefit of the link. This benefit may have some symmetry, for example, two people that share common interests. But it is not necessarily symmetric. For example, one person might be particularly kind, or generous, or intelligent. For simplicity and ease of exposition, we model ν_{ij} as independently and identically distributed. In our setup, the utility of a link does not depend on the wider network. It only depends on the independently-distributed term ν_{ij} and on the treatment status of each of the two nodes.

We allow nodes to form links by mutual consent, and to sever links unilaterally. This results in a unique pairwise-stable network.[15] In the case where $u_{ij} > 0$ for both nodes, a link will be formed. If either node has $u_{ij} < 0$, no link will be formed. Note that situations may arise where one node wants to link to another who does not reciprocate. In this case, no link will exist. The resulting network will take the form

$$g = \{ij : u_{ij} \geq 0, u_{ji} \geq 0\}.$$

This model of network formation simply corresponds to a general random graph (Erdos et al., 1960; Söderberg, 2002). The probability of a link between nodes i and j is independent across links, and takes one of three possible values which depend on the distribution of ν,

$$\mathbb{P}(g_{ij} = 1) = \begin{cases} P_{CC} \equiv \mathbb{P}(\nu > 0)^2, & \text{if } T_i = T_j = 0 \\ P_{TT} \equiv \mathbb{P}(\nu > -\kappa_{TT})^2, & \text{if } T_i = T_j = 1 \\ P_{TC} \equiv \mathbb{P}(\nu > -\kappa_{TC})\mathbb{P}(\nu > -\kappa_{CT}) & \text{if } T_i \neq T_j. \end{cases} \tag{4}$$

We illustrate some simple theoretical predictions about link formation in Figure 6, with ν uniformly distributed on $(-1, 1)$. First, if $\kappa_{TT} > 0$, we expect to see a higher probability of a link between treatment nodes, relative to a link between control nodes. This captures the utility people with information access get from talking to each other. If $(1 + \kappa_{TC})(1 + \kappa_{CT}) > 1$, we expect an increase in the probability of a link between treatment and control nodes. Finally, if $(1 + \kappa_{TC})(1 + \kappa_{CT}) > (1 + \kappa_{TT})^2$, we expect that the increase in links between treatment and control nodes will be larger than the increase in links between two treated nodes, as the desire to seek new information dominates the desire to discuss information two

[15] Pairwise stability, as defined by Jackson and Wolinsky (1996), applies to networks in which no player would benefit from severing a link, and no two players would both benefit from forming a new link.

nodes both have access to.

Next, we will demonstrate some theoretical predictions related to degree and eigenvector centrality. At this point, we will assume that $\kappa_{TC} > 0$, $\kappa_{CT} \geq 0$ and $\kappa_{TT} \geq 0$, so that $P_{TC} > P_{CC}$. That is, links between control and treated nodes strictly increase in response to the intervention. We first demonstrate that the expected degree of a treated node is larger than that of a control node.

Theorem 5.1. *Let $P_{TT} \geq P_{CC}$ and $P_{TC} > P_{CC}$. Suppose that the number of nodes in the control group, N_C is larger than the number of nodes in the treatment group, N_T. Then, treatment nodes have a larger expected degree than control nodes.*

The proof is in Appendix A.1, and the intuition is as follows. Links between treatment and control nodes are on average more beneficial than links between pairs of control nodes. Because there are few treatment nodes, it is not possible for control nodes to increase their degree by much, whereas treatment nodes have many potential control nodes to choose from. If $P_{TT} > P_{CC}$ this effect is amplified, as treatment nodes also form additional links with each other.

Treatment nodes with information access differ from control nodes not only in terms of expected degree, but also in terms of composition of links. In particular, the probability that a linked node is treated, given the treatment status of the node itself, is a function of P_{CC}, P_{TT} and P_{TC}:

$$\begin{aligned} \mathbb{P}(T_j = 1 | g_{ij} = 1, T_i = 1) &= \frac{P_{TT}(N_T - 1)}{P_{TT}(N_T - 1) + P_{TC}N_C} \\ \mathbb{P}(T_j = 1 | g_{ij} = 1, C_i = 1) &= \frac{P_{TC}(N_T)}{P_{TC}(N_T) + P_{CC}(N_C - 1)}. \end{aligned}$$

This implies broader potential impacts on network structure and network centrality. Whether a person has information access affects not only the number of links they form, but the characteristics of those links. For example, if P_{TT} is relatively large, treatment nodes will have a higher number of links, and also a higher proportion of treated links, who are themselves more likely to be well-connected. So, even if link decisions are made without taking the wider network into account, these decisions could affect centrality measures that depend on the wider network, such as eigenvector centrality.

Simulations of the model allow us to illustrate how access to information can cause not only an increase in direct links, but also an increase in eigenvector centrality. Figure 7 plots average eigenvector centrality and degree centrality for simulated networks based on this model. Degree centrality is defined as the total number of links divided by the number of potential links. We simulate 1000 100-node networks, with 20 treatment nodes and 80 control nodes, and fix $P_{CC} = .1$. This loosely approximates our experimental setting for illustrative purposes; we will calibrate the model precisely in the next subsection. We vary P_{TC} and P_{TT}. Holding the other parameter fixed, we see that increasing either P_{TC} or P_{TT} results not only in an increase in degree for treated nodes, but also a sharp increase in eigenvector centrality.

5.2 Extension and Calibration

We now use method-of-moments estimates to calibrate the model in Section 5. We then simulate the calibrated model to compare network-based targeting to random targeting in terms of information diffusion, inequality and academic welfare.

To do this, we must extend our model in two ways. First, to examine the effect of targeting based on baseline network position, we must model the baseline network explicitly, and allow for correlation in links over time. Second, to examine implications for inequality and welfare, we incorporate the possibility that students have other sources of privilege or advantage that persistently attract links. In our data, the largest predictor of baseline centrality is academic achievement, and baseline degree increases sharply at the top of the distribution (Figure A4). Academic ability is also arguably the most important source of individual advantage in our context. By including this variable, we generate a better fit for the degree distribution as well as the persistence of centrality over time, and we facilitate welfare calculations.

We model baseline link utilities as follows:

$$u_{ij}^0 = \sum_{\theta_1,\theta_2 \in \{L,H\}} \mathbb{1}\{\theta_i = \theta_1, \theta_j = \theta_2\} \kappa_{CC}^{\theta_1\theta_2} + \nu_{ij}^0.$$

Here, θ_i represents the academic type for i. H refers to high academic ability, and L refers to low ability. We define a student to be high ability if they belong to the top decile of the exam-score distribution, using the baseline normalized average of English and Biology exam scores.[16] We model the endline link-utility as

$$u_{ij}^1 = \sum_{\theta_1,\theta_2 \in \{L,H\}} \mathbb{1}\{\theta_i = \theta_1, \theta_j = \theta_2\} (\kappa_{CC}^{\theta_1\theta_2} C_i C_j + \kappa_{TC}^{\theta_1\theta_2} T_i C_j + \kappa_{CT}^{\theta_1\theta_2} C_i T_j + \kappa_{TT}^{\theta_1\theta_2} T_j T_j) + \nu_{ij}^1.$$

We are allowing the value of information to be different for high and low ability students, and for the value of information from a high-ability source to differ from the value of information from a low-ability source. To capture correlation in links over time, we assume that $\nu_{ij}^0 = \nu_{ij}^1$ with some probability $(1 - \delta)$, and that otherwise these error terms are independent and identically distributed.

To calibrate the model, we match moments from the model to moments in our empirical information network. The empirical moments we use include the probability of an endline link between students according to treatment status and ability type, and the persistence of control-pair links over time. We simulate a baseline network and an appropriately-correlated endline network. The precise steps involved are detailed in Appendix A.2.

[16] We have nearly complete academic score data for these two core subjects, and we assume students with missing scores are low ability.

5.3 Simulations

We begin by assessing the quality of fit between our model and our empirical findings. We simulate networks with 117 nodes, chosen to match the average network size in our data. We then compare simulated moments to their empirical counterparts, not including the moments we used for calibration. Table 11 contains moments from 10,000 simulated networks and matched summary statistics from our data.

The model appears to capture most moments and treatment effects with reasonable accuracy. In the simulated networks, treated students are, on average, more central than control students according to all centrality measures. They are also more likely to be in the top 5 percent of the distribution. The model does not take into account some covariates that likely affect network formation in our particular setting, such as classroom structure. For this reason, the match is imperfect, especially for betweenness centrality. Yet, it closely predicts the probability that a treatment student will appear among the top 5 percent of students according to all network centrality measures. It also closely predicts a strong correlation between centrality at baseline and endline. Indeed, in both simulations and empirics, nearly half those who were most-central at baseline remain most-central at endline. The match between the model and the data is particularly strong for measures of diffusion centrality. It might be therefore particularly useful for counterfactual experiments involving network-based targeting and information diffusion.

We next simulate sets of networks to conduct counterfactual policy experiments. We compare network-centrality-based targeting strategies to random-targeting strategies, and we vary the size of the target group, the transmission probability, and the centrality measure used for targeting. We plot the estimates against a benchmark that assumes stable networks over time.

We explore implications for total information diffusion. Following Banerjee et al. (2013), we focus on a simple measure of information diffusion, where information is transmitted along each link with probability q for T time periods. We define total diffusion to be the number of times that information is heard, as opposed to the number of nodes that are ever informed. This implies that total diffusion can exceed the size of the network. This may be a particularly suitable measure in our setting, where many different pieces of information are shared, and where students might benefit from hearing the same information from multiple sources. This measure is also closed related to the concept of diffusion centrality; total diffusion is obtained by simply taking the sum of the diffusion centralities of treated nodes.

We consider three different sets of model parameters. We first examine the extent of diffusion in one period ($T = 1$) under complete transmission ($q = 1$), that is, every treated node transmits information to all of their neighbours. Second, we allow information to travel one step further: from a treated node to their neighbours, and on to those nodes' neighbours ($q = 1$, $T = 2$). Finally, we examine the extent of information diffusion under imperfect transmission and many time periods, using the parameters (q^*, T^*) suggested by Banerjee et al. (2019).[17] For each counterfactual, and each set of parameters, we simulate

[17] We set q equal to the reciprocal of the top eigenvalue, and T equal to the diameter of the graph.

1000 networks with 100 nodes.[18]

We compare policy counterfactuals involving network-based targeting and random targeting. Network-based targeting involves identifying nodes that are central at baseline and assigning them to the treatment group. In Figure 8, we target the top N nodes based on initial degree, in Figure 9 we target based on eigenvector centrality, and in Figure 10 we target based on diffusion centrality, with parameters (q^*, T^*) as above.

In each panel of each figure we plot total information diffusion under network-based targeting (in red) versus random targeting (in blue). We consider three different hypothetical settings with respect to network structure. First, we plot information diffusion on networks that change over time, both due to exogenous link changes and endogenous link formation, as estimated in our empirical setting. Second, we plot information diffusion in a hypothetical setting where networks remain stable over time. Third, we compare these plots to a hypothetical setting where networks change exogenously over time, but not endogenously in response to the treatment. Note that we do not plot this third hypothetical for random targeting, as exogenous link changes do not affect information diffusion under this policy.

The simulations show that while network-based targeting does outperform random targeting, the gap in total information diffusion is greatly reduced by the endogenous network change. This pattern is highly similar across all three sets of diffusion parameters, and all three network-based targeting strategies (see Panels A, B and C of Figures 8, 9 and 10). Under stable networks (dashed lines), network-based targeting (in red) vastly outperforms random targeting (in blue). This gap remains but shrinks considerably under networks that respond endogenously (solid lines). Consider a model of perfect information transmission over one period (Panel A of 8). If networks remain stable over time, by targeting the top 20 percent of nodes by degree, as opposed to at random, we increase total information diffusion by 54 percent. Taking the endogenous network change into account, this gain in diffusion is reduced by half; degree-based targeting only increases diffusion by 24 percent. If the treatment group is larger, network-based targeting is less important. When half of nodes are treated, network-based targeting increases information diffusion by only 10 percent, as opposed to 30 percent under stable networks.

The same pattern holds true if we vary the rate of information transmission across links. In Panel D, we consider a two-period model of transmission, and vary the probability of transmission, q, between 0 and 1. Again, network-based targeting outperforms random targeting, with a gap in information transmission that is greatly reduced by the endogenous network response. The advantage of network-based targeting is larger when q is higher, that is, when information flows easily across links.

The gap in information diffusion decreases under endogenous network change for two reasons. First, networks change over time, independent of the information intervention. These exogenous changes, on their own, reduces the gains from network-based targeting. Second, there is a treatment effect: the information intervention increases the centrality of treated nodes. This increases information diffusion under

[18] We vary the target group size in increments of 5.

both targeting strategies, but not necessarily by the same amount. Indeed, we see that under network-based targeting, total information diffusion would be lower if the network were to change exogenously but not in response to the information treatment. Under random targeting, total information diffusion is not affected by exogenous network change, but increases under endogenous network response.

Targeting based on network centrality also has implications for inequality, as centrality is correlated with other forms of privilege in many settings (Jackson, 2019b). In our model, this privilege applies to the top decile of nodes according to ability-type θ, and we allow link formation probabilities to depend on type. In our setting, θ corresponds to academic ability, but θ could theoretically represent any source of privilege that is highly correlated with network centrality.

In Figure 11, we plot the share of targeted nodes that are in the top decile of ability. In Panel A, this is estimated by simulating the model, and in Panel C, we make direct use of our baseline network data. When targeting students at random, we expect approximately 10 percent of targets to belong to the privileged group. This is much higher in the network-based targeting counterfactual, especially when the treatment group is small. For example, if we select the top 10 percent of nodes by degree centrality, approximately half will be in the high-ability group.

Inequality concerns are made worse in our context by the fact that our intervention has a significant impact on academic performance only for lower ability students. If we estimate the main specification from Derksen et al. (2022) with heterogeneous treatment effects, we find that the intervention caused a significant .14 standard deviation increase in English scores for those in the bottom 90 percent of the distribution, and an insignificant .04 standard deviation increase for those in the top 10 percent.

We use these direct impact estimates to show that academic welfare is uniformly lower under network-based targeting (Panels B and D of Figure 11). We plot total English score gains in standard deviations under network-based and random targeting policies, and vary the size of the treatment group. The relative loss in welfare is again largest when the treatment group is small. If we were to target 10 percent of students, random targeting would approximately double total academic welfare from .7 standard deviations to 1.3 standard deviations per 100 students. We can again estimate these effects using model simulations (Panel B) or baseline data (Panel D), with similar results.

Random targeting leads to larger aggregate academic gains in our setting, but likely lowers total information diffusion. Consider again full information transmission in a single period, that is, information can only be obtained directly by or from the person with direct access. If we target 10 percent of students at random, as opposed to based on network-centrality, this decreases total information transmission by approximately 30 percent. In a network of 100 students, the cost of a 0.6 standard deviation total increase in academic scores is a decrease in information transmission from 100 to 70.

6 Conclusion

This paper identifies a causal determinant of network centrality: information access. We show that providing individuals with exclusive, long-term access to a high quality information source causally affects their network positions. We conducted a randomized trial in Malawian secondary schools, and provided a small subset of students with exclusive access to online information. Over the course of the school year, this caused students to form new information-sharing links, which led to a signficant increase in centrality for treated students according to many different centrality measures. After eight months, treated students were more likely than control students to be among the highest centrality students.

The impact of information access on network structure is likely to vary based on the nature, scale, usefulness and importance of the information provided, the degree of exclusivity, the duration of access, and the level of trust in the community. For example, effects may be larger in a setting where the information provided is highly instrumental, such as agricultural information, as opposed to a mix of instrumental, general knowledge and entertainment. On the other hand, the networks we observe appear comparable to networks captured in other real world settings, and interactions between students likely include many of the dynamics present in any close-knit community.

Because our reduced form results yield a straightforward interpretation, we chose to specify a simple and transparent model. The model abstracts from many potential determinants of network formation, such as student characteristics and classroom structure, to focus on the role of information access and underlying advantage. We also assume only first-order information transmission, that is, that students cannot obtain information second-hand from a treated student. This allows us to describe a simple equilibrium network, but may not deliver some of nuanced predictions a richer theoretical model would provide.

This study has implications for policies that target an intervention to participants based on their network positions. Information interventions, especially when implemented at scale and over the longer term, can make initially ordinary members of a social network central and influential. Expensive network mapping exercises undertaken with the goal of targeting influential people may therefore be an inefficient and suboptimal use of resources. Moreover, centrality-based targeting can amplify existing inequalities, as influence is typically correlated with privilege. To maximize aggregate welfare and limit inequality, policymakers should consider not only diffusion but also the direct impact of the intervention, and its potential to close outcome gaps.

Our findings also highlight the potential pitfalls of using network data to estimate spillovers. Networks can change over time, both exogenously and in response to an experiment. Standard specifications may produce biased spillover estimates. For example, estimates may be biased towards zero if spillovers occur along links that form in response to the intervention itself.

Information access is a natural source of advantage in a social network. Yet, other resources likely also impact network position. Moreover, information likely has a different effect on network structure when

provided at the network-level rather than to individual nodes. If network formation is purely strategic, we would expect our effects to fade after the end of the intervention. Whether endogenous network changes persist is an open empirical question. Developing a broader understanding of the determinants of social network position and overall network structure is an important direction for future work.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2016). Clustering as a design problem. *Working Paper*.

Ambrus, A. and Elliott, M. (2021). Investments in social ties, risk sharing, and inequality. *The Review of Economic Studies*, 88(4):1624–1664.

Bailard, C. S. (2012). A field experiment on the Internet's effect in an African election: Savvier citizens, disaffected voters, or both? *Journal of Communication*, 62(2):330–344.

Bandiera, O., Burgess, R., Deserranno, E., Morel, R., Rasul, I., and Sulaiman, M. (2022). Social incentives, delivery agents and the effectiveness of development interventions. *Journal of Political Economy Microeconomics*.

Banerjee, A., Breza, E., Chandrasekhar, A. G., Duflo, E., Jackson, M. O., and Kinnan, C. (2022). Changes in social network structure in response to exposure to formal credit markets. *The Review of Economic Studies*.

Banerjee, A., Breza, E., Chandrasekhar, A. G., and Golub, B. (2021). When less is more: Experimental evidence on information delivery during India's demonetization. *Working Paper*.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144).

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2019). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, 86(6):2453–2490.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Batzilis, D., Dinkelman, T., Oster, E., Thornton, R., and Zanera, D. (2010). New cellular networks in Malawi: Correlates of service rollout and network performance. *Working Paper*.

Baumgartner, E., Breza, E., La Ferrara, E., Orozco, V., and Rosa Dias, P. (2022). The nerds, the cool and the central: Peer education and teen pregnancy in Brazil. *Working Paper*.

Beaman, L., BenYishay, A., Magruder, J., and Mobarak, A. M. (2021). Can network theory-based targeting increase technology adoption? *American Economic Review*, 111(6):1918–43.

Beaman, L. and Dillon, A. (2018). Diffusion of agricultural information within social networks: Evidence on gender inequalities from Mali. *Journal of Development Economics*, 133:147–161.

Berg, E., Ghatak, M., Manjula, R., Rajasekhar, D., and Roy, S. (2019). Motivating knowledge agents: Can incentive pay overcome social distance? *The Economic Journal*, 129(617):110–142.

Bertelli, O. and Fall, F. (2022). Mobilizing farmer trainers: Experimental evidence from rural uganda. *Working paper*.

Bertrand, M., Bombardini, M., and Trebbi, F. (2014). Is it whom you know or what you know? An empirical assessment of the lobbying process. *American Economic Review*, 104(12):3885–3920.

Binzel, C., Field, E., and Pande, R. (2017). Does the arrival of a formal financial institution alter informal sharing arrangements? Experimental evidence from village India. *Working Paper*.

Bloch, F., Jackson, M. O., and Tebaldi, P. (2019). Centrality measures in networks. *Working Paper*.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120.

Bramoullé, Y. and Kranton, R. (2007a). Risk sharing across communities. *American Economic Review*, 97(2):70–74.

Bramoullé, Y. and Kranton, R. (2007b). Risk-sharing networks. *Journal of Economic Behavior & Organization*, 64(3-4):275–294.

Breza, E., Chandrasekhar, A., Golub, B., and Parvathaneni, A. (2019a). Networks in economic development. *Oxford Review of Economic Policy*, 35(4):678–721.

Breza, E., Chandrasekhar, A. G., and Larreguy, H. (2019b). Network centrality and institutional design: evidence from a lab experiment in the field. *Working Paper*.

Calvó-Armengol, A., De Martí, J., and Prat, A. (2015). Communication and influence. *Theoretical Economics*, 10(2):649–690.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267.

Campante, F., Durante, R., and Sobbrio, F. (2018). Politics 2.0: The multifaceted effect of broadband Internet on political participation. *Journal of the European Economic Association*, 16(4):1094–1136.

Capozza, F., Haaland, I., Roth, C., and Wohlfart, J. (2021). Studying information acquisition in the field: A practical guide and review. *Working Paper*.

Chandrasekhar, A. and Lewis, R. (2016). Econometrics of sampled networks. *Working Paper*.

Chen, Y. and Yang, D. Y. (2019). The impact of media censorship: 1984 or brave new world? *American Economic Review*, 109(6):2294–2332.

Coleman, J. S. (1964). *Introduction to Mathematical Sociology*. London Free Press Glencoe.

Comola, M. and Prina, S. (2021). Treatment effect accounting for network changes. *Review of Economics and Statistics*, 103(3):597–604.

Cruz, C., Labonne, J., and Querubin, P. (2017). Politician family networks and electoral outcomes: Evidence from the Philippines. *American Economic Review*, 107(10):3006–37.

Dar, M., de Janvry, A., Emerick, K., Kelley, E., and Sadoulet, E. (2020). Casting a wider net: Sharing information beyond social networks. *Working paper*.

Delavallade, C., Griffith, A., and Thornton, R. (2016). Network partitioning and social exclusion under different selection regimes. *Working Paper*.

Derksen, L., Michaud-Leclerc, C., and Souza, P. C. (2022). Restricted access: How the internet can be used to promote reading and learning. *Journal of Development Economics*.

Dimitriadis, S. and Koning, R. (2022). Social skills improve business performance: evidence from a randomized control trial with entrepreneurs in Togo. *Management Science*.

Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Feigenberg, B., Field, E., and Pande, R. (2013). The economic returns to social interaction: Experimental evidence from microfinance. *Review of Economic Studies*, 80(4):1459–1483.

Fernando, A. N. (2021). Seeking the treated: The impact of mobile extension on farmer information exchange in India. *Journal of Development Economics*, 153:102713.

Field, E. and Pande, R. (2018). An experimental test of the association between network centrality and cross-village risk-sharing links. *Working Paper*.

Fowler, J. H., Dawes, C. T., and Christakis, N. A. (2009). Model of genetic variation in human social networks. *Proceedings of the National Academy of Sciences*, 106(6):1720–1724.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Galperin, H. and Viecens, M. F. (2017). Connected for Development? Theory and evidence about the impact of Internet technologies on poverty alleviation. *Development Policy Review*, 35(3):315–336.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438:900–901.

Girard, Y., Hett, F., and Schunk, D. (2015). How individual characteristics shape the structure of social networks. *Journal of Economic Behavior & Organization*, 115:197–216.

Griffith, A. (2022a). Name your friends, but only five? The importance of censoring in peer effects estimates using social network data. *Journal of Labor Economics*, 40(4):000–000.

Griffith, A. (2022b). Random assignment with non-random peers: A structural approach to counterfactual treatment assessment. *The Review of Economics and Statistics*, pages 1–40.

Hahn, Y., Islam, A., Patacchini, E., and Zenou, Y. (2015). Teams, organization and education outcomes: Evidence from a field experiment in Bangladesh. *Working Paper*.

Hasan, S. and Bagde, S. (2015). Peers and network growth: Evidence from a natural experiment. *Management Science*, 61(10):2536–2547.

Heß, S., Jaimovich, D., and Schündeln, M. (2021). Development projects and economic networks: Lessons from rural Gambia. *The Review of Economic Studies*, 88(3):1347–1384.

Hjort, J. and Poulsen, J. (2019). The arrival of fast Internet and employment in Africa. *American Economic Review*, 109(3):1032–1079.

Jackson, M. O. (2010). *Social and Economic Networks*. Princeton University Press.

Jackson, M. O. (2019a). The friendship paradox and systematic biases in perceptions and social norms. *Journal of Political Economy*, 127(2):777–818.

Jackson, M. O. (2019b). *The Human Network: How we're connected and why it matters*. Atlantic Books.

Jackson, M. O., Rogers, B. W., and Zenou, Y. (2017). The economic consequences of social-network structure. *Journal of Economic Literature*, 55(1):49–95.

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

Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the south Indian fisheries sector. *The Quarterly Journal of Economics*, 122(3):879–924.

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., and Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2):210–221.

Miner, L. (2015). The unintended consequences of Internet diffusion: Evidence from Malaysia. *Journal of Public Economics*, 132:66–78.

Morelli, S. A., Ong, D. C., Makati, R., Jackson, M. O., and Zaki, J. (2017). Empathy and well-being correlate with centrality in different social networks. *Proceedings of the National Academy of Sciences*, 114(37):9843–9847.

Pin, P. and Rogers, B. W. (2016). Stochastic network formation and homophily. In Bramoullé, Y., Galeotti, A., and Rogers, B. W., editors, *The Oxford Handbook of the Economics of Networks*, chapter 7. Oxford University Press.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704.

Singh, J., Hansen, M. T., and Podolny, J. M. (2010). The world is not small for everyone: Inequity in searching for knowledge in organizations. *Management Science*, 56(9):1415–1438.

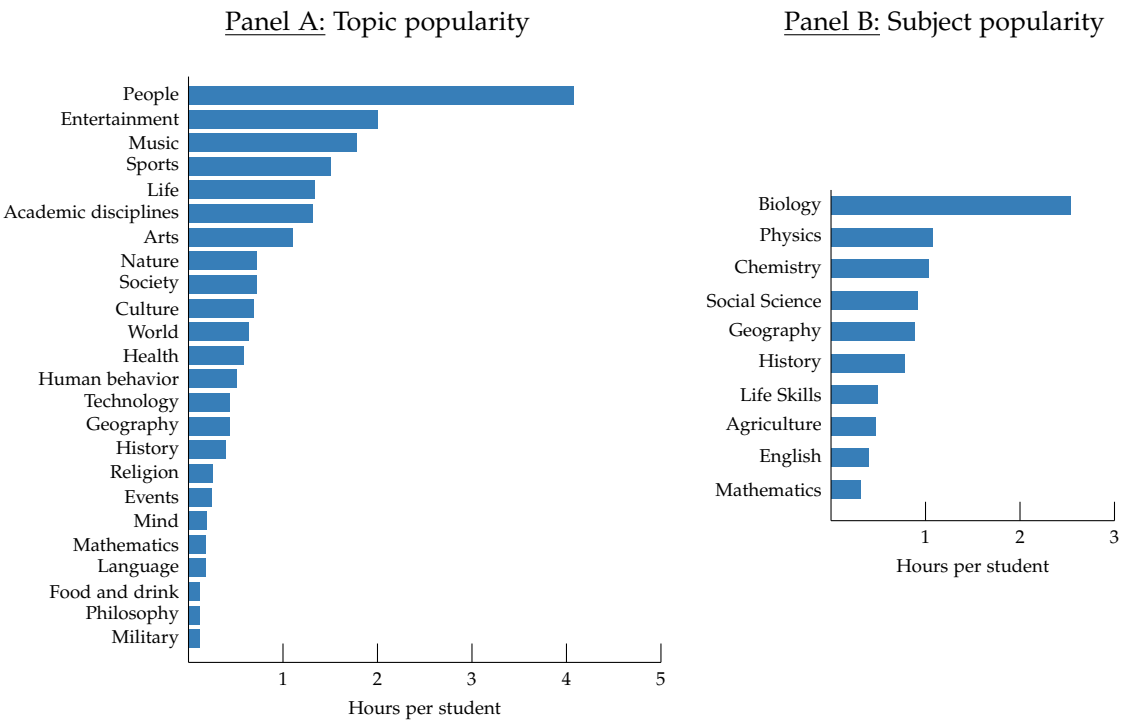
Söderberg, B. (2002). General formalism for inhomogeneous random graphs. *Physical Review E*, 66(6):066121.

Stein, M. (2021). Know-how and know-who: Effects of a randomized training on network changes between small urban entrepreneurs. *CSEF Working Paper*, (622).

Zárate, R. A. (2021). Uncovering peer effects in social and academic skills. *Working Paper*.

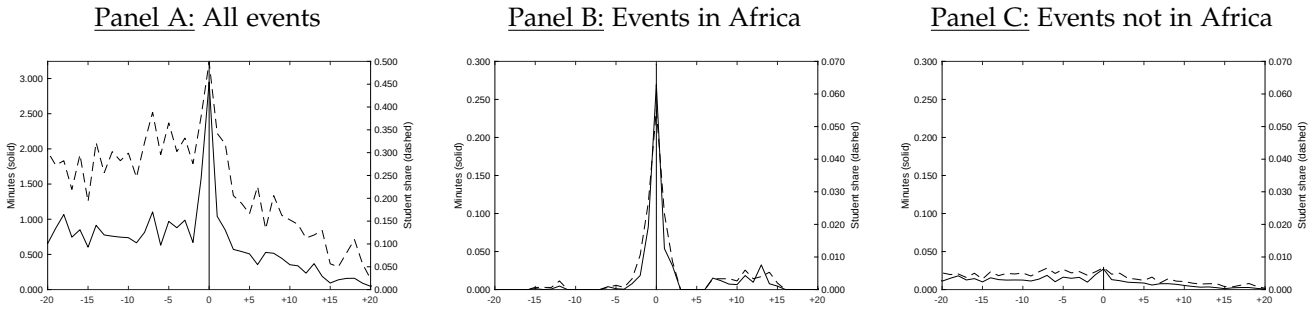
Figures and Tables

Figure 1: Hours Spent Browsing Wikipedia by Topic and School Subject



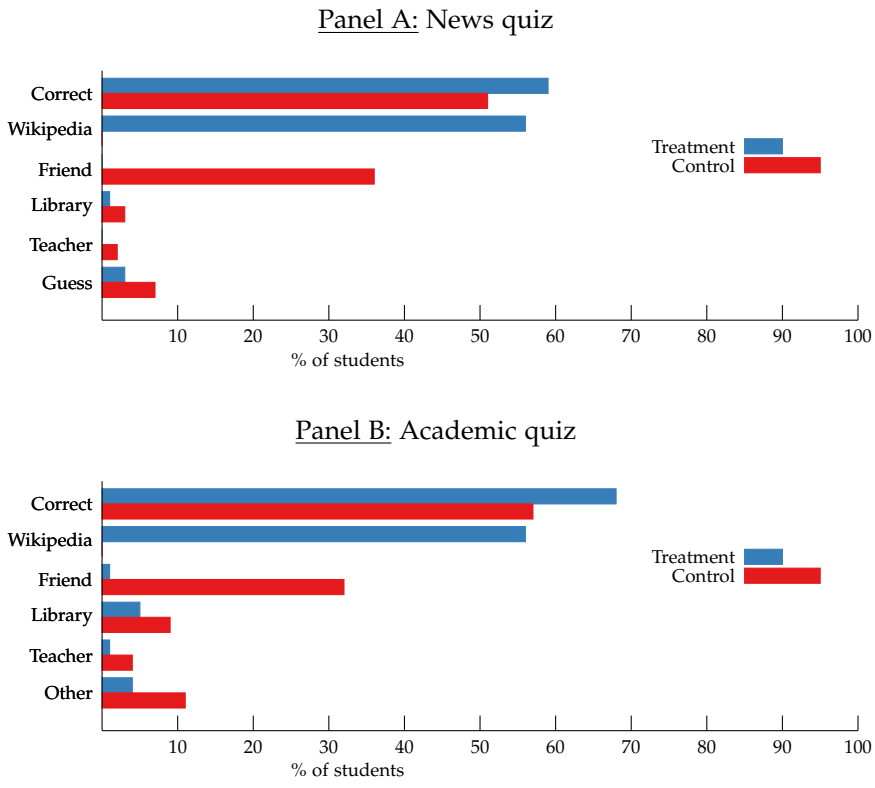
Notes: Figure reproduced from Derksen et al. (2022). Panel A: Browsing hours per topic, per student, aggregated over one academic year. The topics Business, Concepts, Crime, Economy, Education, Energy, Government, Humanities, Knowledge, Law, Objects, Organizations, Politics, Science, and Universe are excluded from the figure and are less than 0.12 hours. Panel B: Browsing hours per school subject, per student, aggregated over one academic year. See Derksen et al. (2022) for details on topic classification.

Figure 2: Wikipedia Browsing for News about World Events in 2017-18



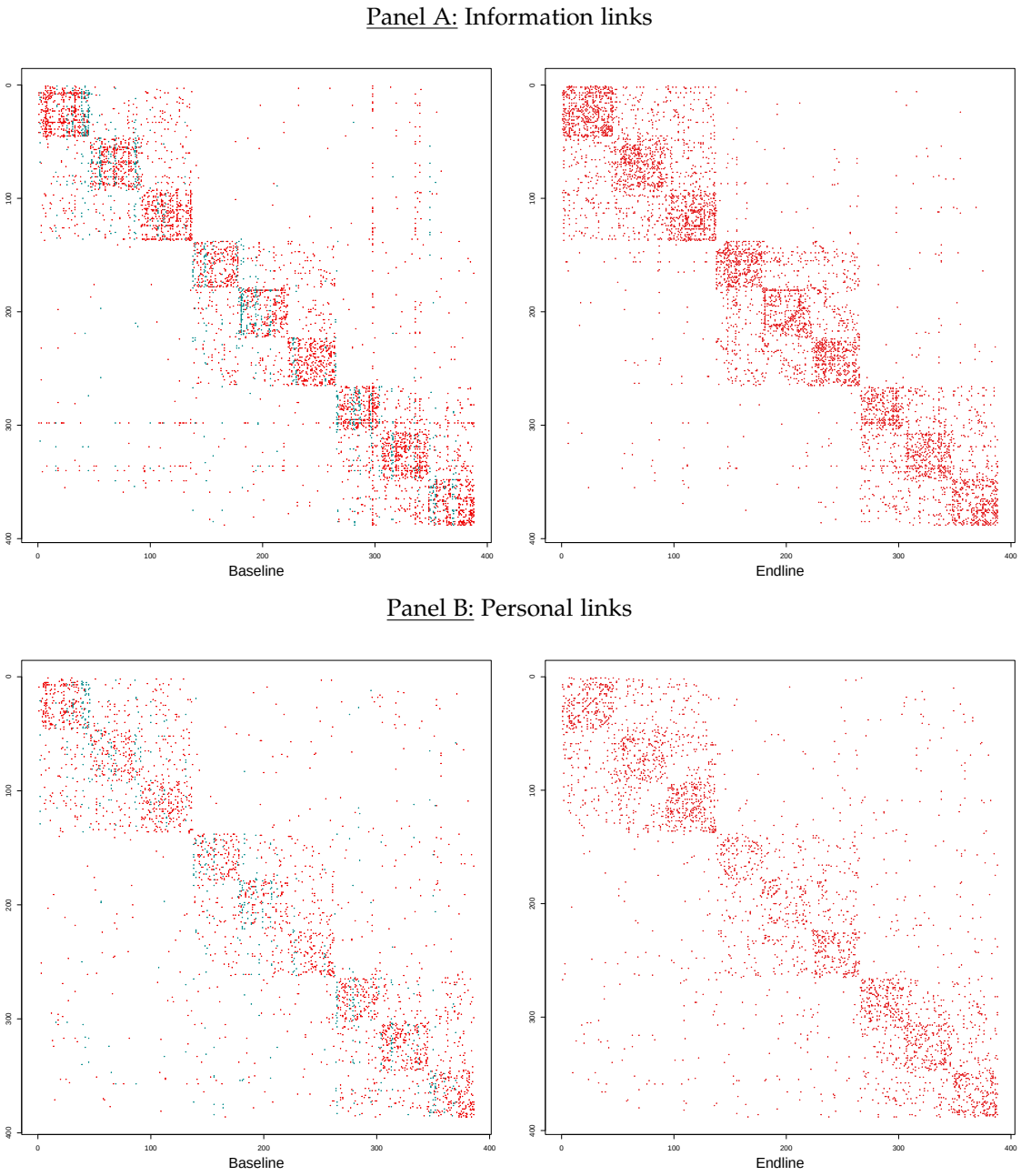
Notes: Figure reproduced from Derksen et al. (2022). Panel A: Left axis (solid line) shows total average browsing minutes per student on pages related to full set of worldwide events. Right axis (dashed line) shows share of students that visited pages associated to at least one event. Panels B and C: Left axis (solid line) shows average number of minutes per student and event. Right axis (dashed line) shows average share of students that visited pages associated to a single event. All events from November 2nd 2017 to May 9th 2018 as reported in `https://en.wikipedia.org/wiki/2017` and `https://en.wikipedia.org/wiki/2018` are included, with the 20 weeks before and after they occurred. See Derksen et al. (2022) for details on classification of news events. Week of the event is set at zero. Negative (positive) numbers on the x-axis are weeks before (after) the event.

Figure 3: Information Sources for Correct Quiz Answers



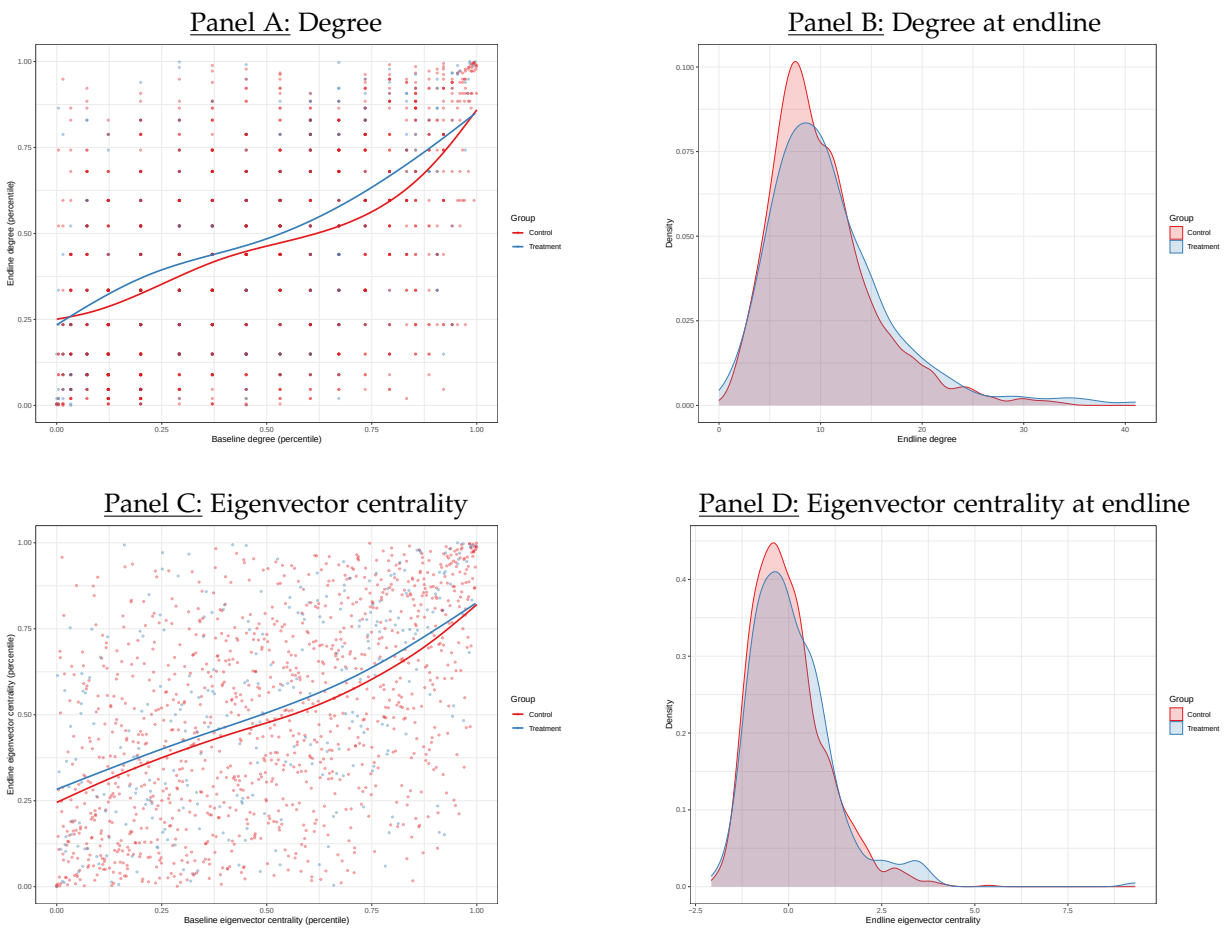
Notes: Percent of students that found correct answers to their unique quiz questions, overall and by information source. Among friends who provided correct answers, 94% were in the treatment group.

Figure 4: Networks at a Single School



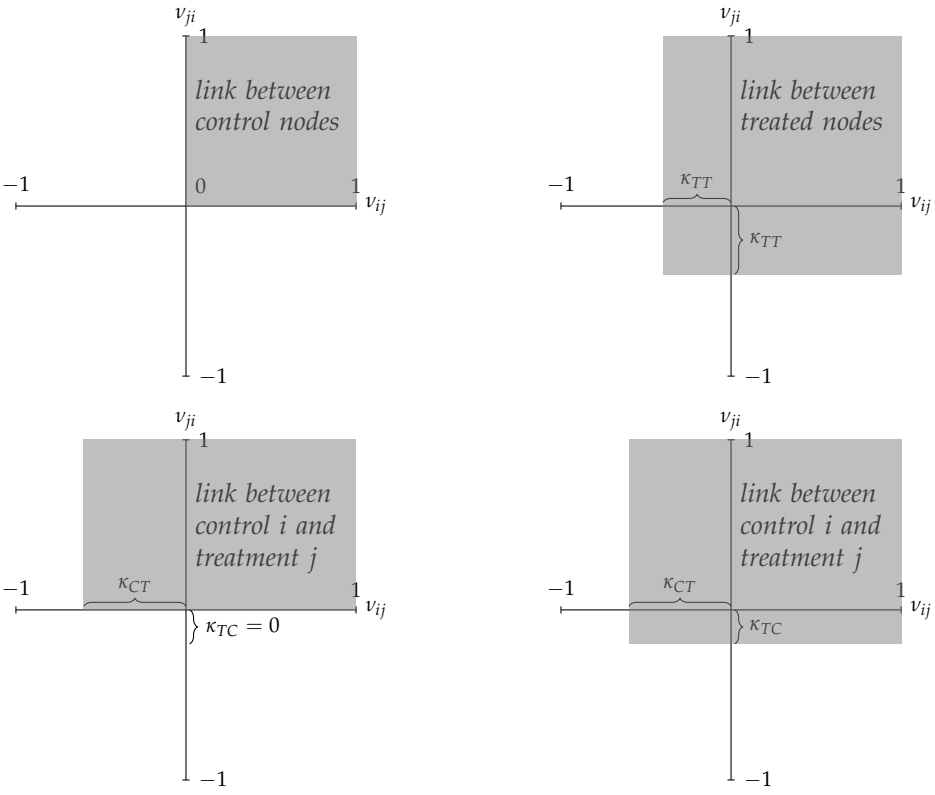
Notes: Baseline network adjacency matrices for a single school, including links across forms. Nodes are ordered by form and classroom. A dot represents an undirected link between nodes. On the horizontal axis, blue nodes are treated and red nodes are control. Panel A: Information links. Panel B: Personal friendship links.

Figure 5: Centrality in the Information Network



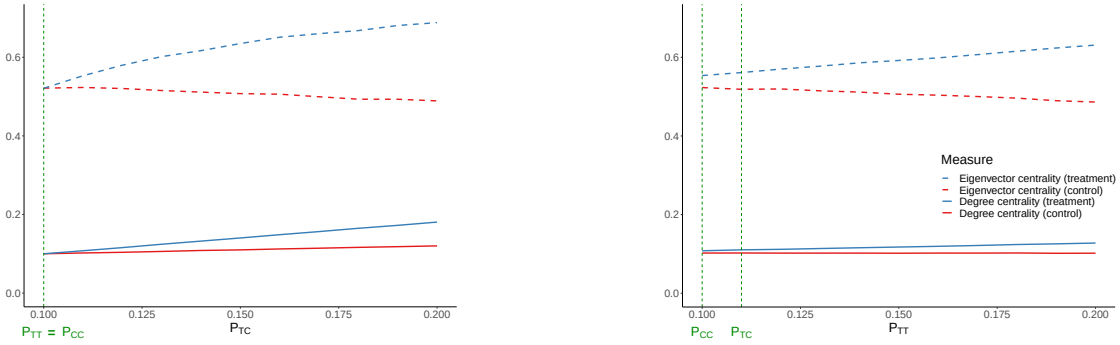
Notes: Centrality measure distributions in the information network for treatment and control groups.

Figure 6: The Model of Link Formation



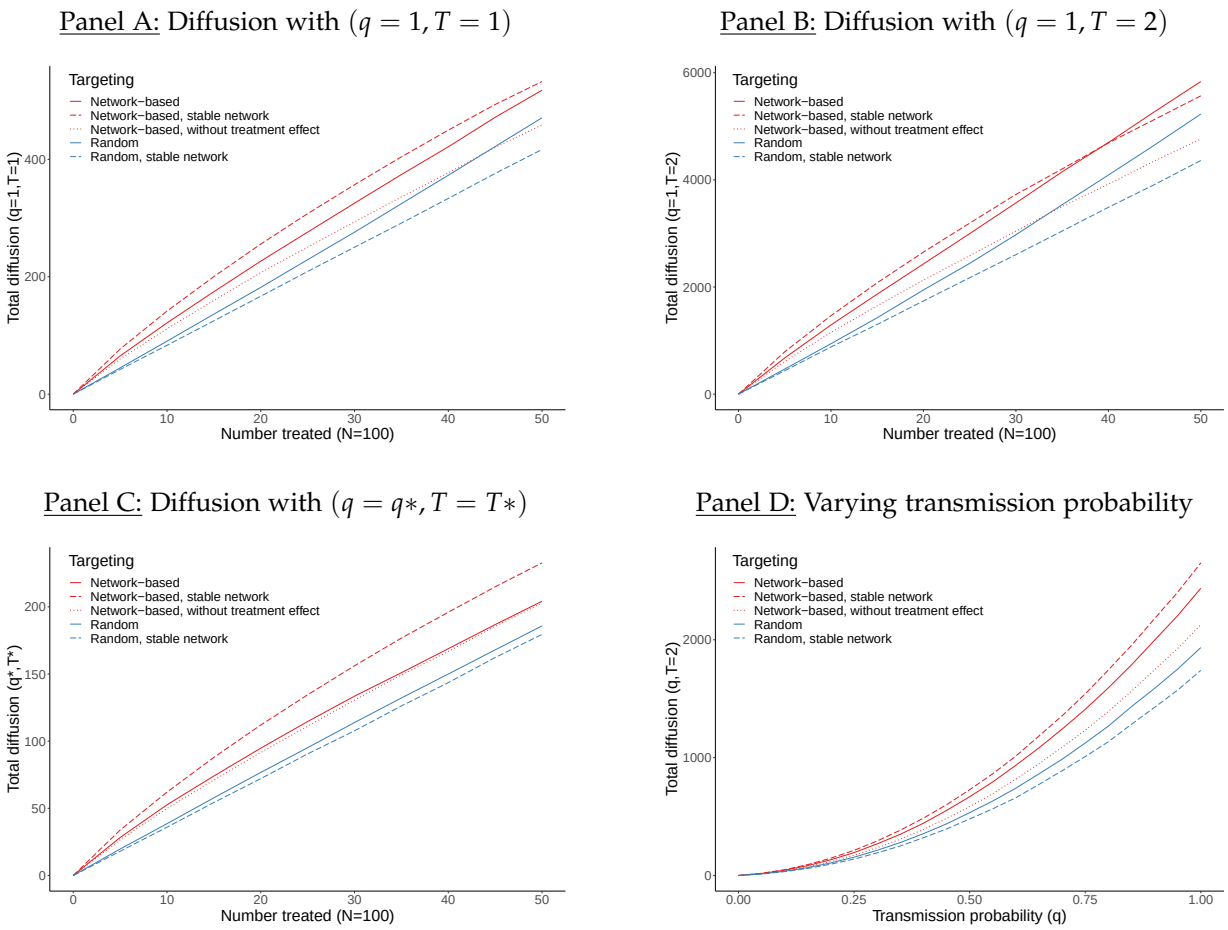
Notes: Pairwise-stable equilibria for link formation. The (relative) area of the shaded rectangle represents the probability of forming a link.

Figure 7: Model Simulations



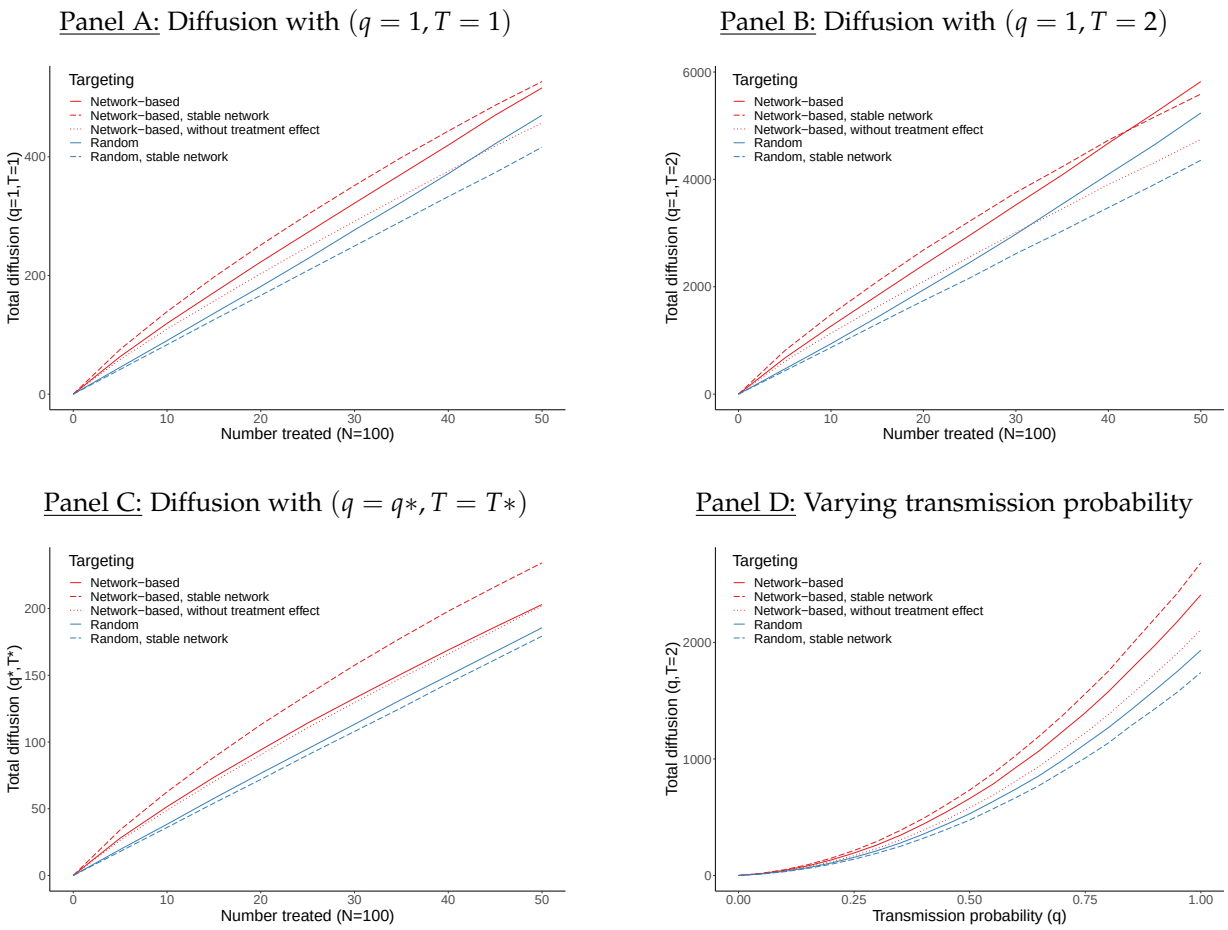
Notes: Simulated average centrality measures for treatment and control groups based on the model in Section 5.1. Degree centrality is the total number of links divided by the number of potential links. Left: P_{TC} varies while other parameters are fixed with $P_{CC} = P_{TT} = .10$. Right: P_{TT} varies while other parameters are fixed with $P_{CC} = .10$ and $P_{TC} = .11$. 1000 simulated networks for each set of parameters. Each network has 20 treated and 80 control nodes.

Figure 8: Total Diffusion Under Degree-Based Random Versus Random Targeting



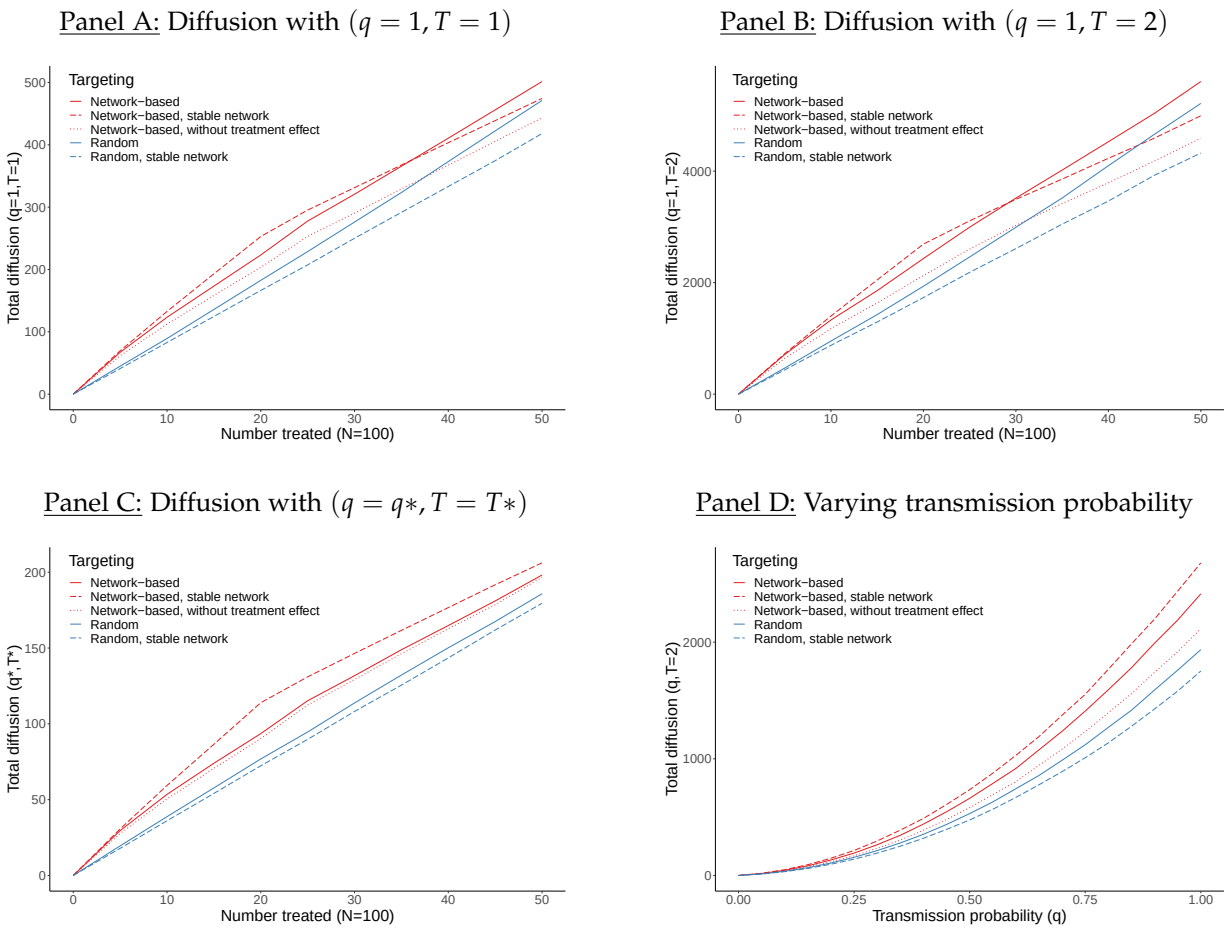
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values (number treated varied in intervals of 5, q varied in intervals of 0.05). Total diffusion is defined as the sum of the diffusion centralities of treated nodes. Network-based targeting involves targeting the top nodes by degree. Parameters q^* and T^* are set to equal the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019).

Figure 9: Total Diffusion Under Eigenvector Centrality-Based Versus Random Targeting



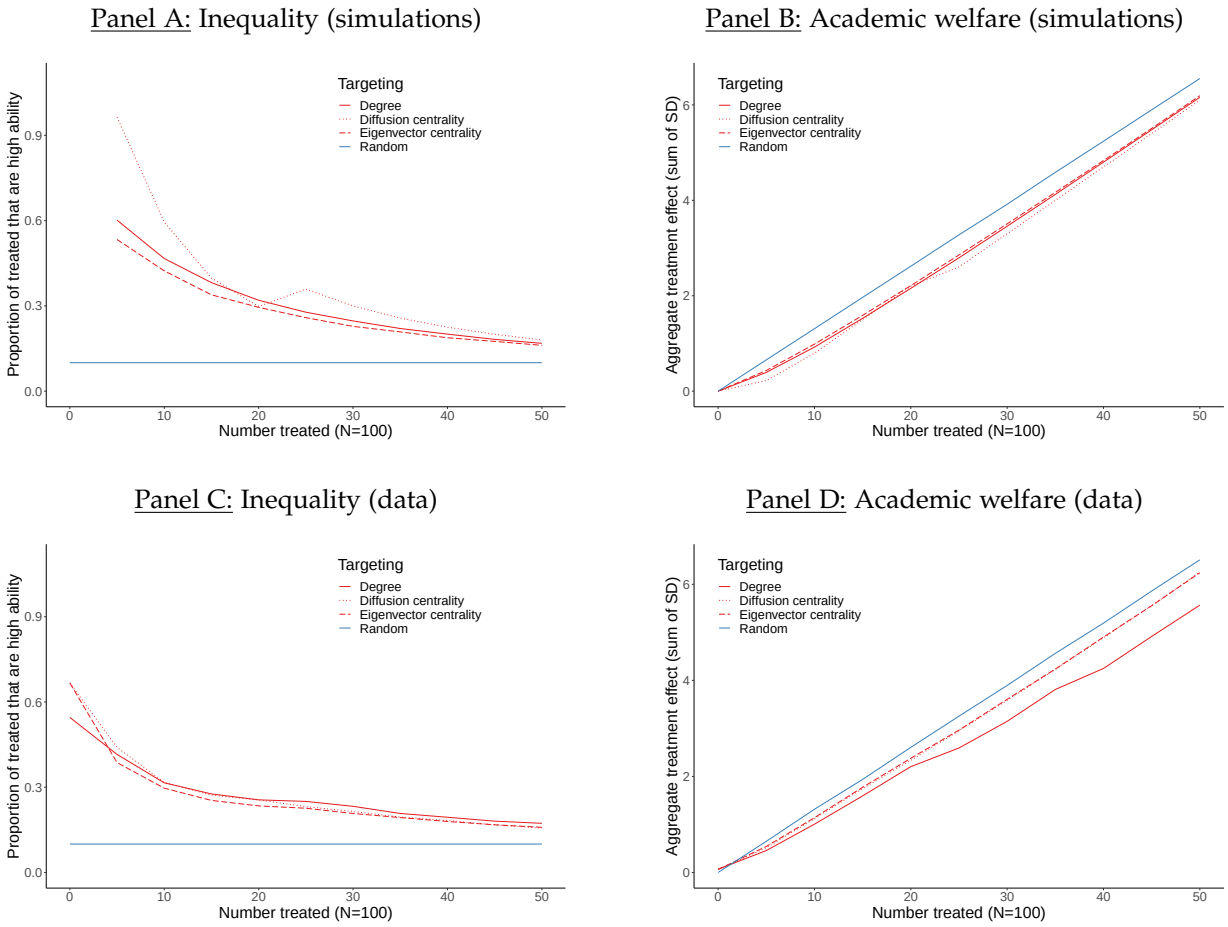
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values (number treated varied in intervals of 5, q varied in intervals of 0.05). Total diffusion is defined as the sum of the diffusion centralities of treated nodes. Network-based targeting involves targeting the top nodes by eigenvector centrality. Parameters q^* and T^* are set to equal the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019).

Figure 10: Total Diffusion Under Diffusion Centrality-Based Versus Random Targeting



Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values (number treated varied in intervals of 5, q varied in intervals of 0.05). Total diffusion is defined as the sum of the diffusion centralities of treated nodes. Network-based targeting involves targeting the top nodes by diffusion centrality, with parameters q^* and T^*. Parameters q^* and T^* are set to equal the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019).

Figure 11: Inequality and Academic Welfare



Notes: Panels A and B: simulations of 100-node networks, with 1000 replications for each set of parameter values (number treated varied in intervals of 5). Panels C and D: overlap between high ability and high centrality students is calculated directly from baseline data. Academic welfare is calculated by aggregating average treatment effects, measured in SD, estimated using a similar specification to Derksen et al. (2022). Diffusion centrality calculated using q^* and T^* as in Banerjee et al. (2019).

Table 1: Survey Measures of Network Links

Information Network
Who do you talk to about movies, music, sports and entertainment?
Who do you ask for information that might be useful when researching a topic learned in class?
Who do you ask for information about the news?
Who do you ask for information about health?
Who do you ask for information about school activities?
Personal Friendship Network
Who is your best friend at school?
Who have you borrowed money from at this school?
Who have you borrowed things from at this school?
Who have you given a gift to at this school?
Who do you talk to about personal topics or ask for advice?

Notes: The same survey measures were collected at baseline and endline from the full sample.

Table 2: Balance Table and Attrition

	Control (N=1207)		Treatment (N=301)			
	Mean	SD	Mean	SD	Difference	p-value
Panel A. Information Network						
Degree	10.792	6.391	10.764	6.497	-0.028	0.947
Eigenvector centrality	0.275	0.178	0.274	0.177	-0.000	0.972
Number of length-2 walks	167.781	108.081	169.880	113.573	2.099	0.772
Diffusion	3.481	2.026	3.416	1.901	-0.065	0.599
Betweenness	0.011	0.019	0.011	0.013	-0.000	0.743
Treated links	2.124	1.806	2.246	1.867	0.122	0.310
Average link strength	0.307	0.088	0.307	0.096	0.000	0.959
Has treated links	0.842	0.365	0.857	0.351	0.015	0.500
Panel B. Personal Network						
Degree	6.418	3.373	6.502	3.471	0.083	0.708
Eigenvector centrality	0.274	0.213	0.271	0.214	-0.003	0.828
Number of length-2 walks	59.203	36.364	59.900	37.874	0.697	0.773
Diffusion	4.306	2.798	4.225	2.726	-0.080	0.649
Betweenness	0.016	0.019	0.016	0.016	-0.001	0.608
Treated links	1.312	1.224	1.243	1.213	-0.069	0.379
Average link strength	0.337	0.097	0.340	0.110	0.003	0.699
Has treated links	0.734	0.442	0.694	0.461	-0.040	0.179
Panel C. Full Network						
Degree	13.452	6.973	13.528	7.236	0.077	0.868
Eigenvector centrality	0.322	0.182	0.323	0.184	0.001	0.941
Number of length-2 walks	243.304	143.087	247.429	151.069	4.125	0.669
Diffusion	3.536	1.826	3.487	1.742	-0.049	0.667
Betweenness centrality	0.010	0.015	0.010	0.011	-0.000	0.786
Treated links	2.678	1.999	2.791	2.113	0.113	0.402
Average link strength	0.206	0.064	0.206	0.066	0.001	0.891
Has treated links	0.900	0.300	0.894	0.309	-0.006	0.759
Attrition	0.076	0.265	0.047	0.211	-0.030	0.039

Notes: Baseline balance between treatment (N=301) and control students (N=1207) across the node-level centrality measures. "SD" stands for standard deviation. "Difference" is the difference of means between control and treatment. *p*-value tests the null hypothesis that the difference in means of control and treatment are equal to zero. Attrition is included in the last row of Panel C (Full Network).

Table 3: Correlates of Centrality at Baseline

	Degree		Eigenvector Centrality		Top 5% Degree		Top 5% Eigenvector Centrality	
Panel A. Information Network								
High Achiever	2.43***		0.400***		0.053***		0.043***	
	(0.293)		(0.051)		(0.013)		(0.012)	
SES	1.33***		0.354***		0.028**		0.031**	
	(0.310)		(0.056)		(0.014)		(0.013)	
Male	-1.40***		-0.511***		-0.017		-0.042**	
	(0.456)		(0.081)		(0.020)		(0.019)	
High Browsing		-0.068		-0.087		-0.009		0.013
		(0.723)		(0.135)		(0.032)		(0.032)
Observations	1,402	287	1,402	287	1,402	287	1,402	287
R^2	0.1684	0.1865	0.0862	0.0326	0.0154	0.0219	0.0163	0.0247
Panel B. Personal Network								
High Achiever	1.08***		0.295***		0.055***		0.038***	
	(0.159)		(0.051)		(0.014)		(0.012)	
SES	0.649***		0.282***		0.029**		0.043***	
	(0.169)		(0.054)		(0.015)		(0.012)	
Male	-1.59***		-1.02***		-0.046**		-0.101***	
	(0.256)		(0.079)		(0.023)		(0.022)	
High Browsing		0.066		-0.121		0.007		-0.018
		(0.386)		(0.125)		(0.033)		(0.030)
Observations	1,402	287	1,402	287	1,402	287	1,402	287
R^2	0.1792	0.1684	0.1409	0.0450	0.0225	0.0361	0.0364	0.0303
Panel C. Full Network								
High Achiever	2.54***		0.390***		0.047***		0.053***	
	(0.316)		(0.052)		(0.013)		(0.012)	
SES	1.33***		0.310***		0.015		0.024*	
	(0.334)		(0.055)		(0.013)		(0.013)	
Male	-2.02***		-0.577***		-0.026		-0.059***	
	(0.494)		(0.081)		(0.021)		(0.020)	
High Browsing		0.226		-0.031		-0.0006		-0.008
		(0.780)		(0.132)		(0.028)		(0.029)
Observations	1,402	287	1,402	287	1,402	287	1,402	287
R^2	0.1996	0.2180	0.0852	0.0282	0.0133	0.0174	0.0220	0.0233

Notes: Odd columns: regressions of degree, eigenvector centrality, top 5% degree and top 5% eigenvector centrality on high-ability, high SES (SES), and male (estimating a single regression for each outcome). Even columns: regressions of the same outcomes on high-browsing dummy (treatment students only). High-ability is defined as above-median exam score at baseline. SES is equal to 1 if respondent's house has electricity and running water. High browsing is defined as above-median time in the digital library across the duration of the experiment. Regressions include school-form fixed effects. Heteroskedasticity-robust standard errors in parentheses. *** p<0.01; ** p<0.05; * p<0.1.

Table 4: Effect of Information Access on Centrality in the Information Network

	Degree	Eigenvector	Number of Length-2 Walks	Diffusion	Betweenness	Average Link Strength
Panel A. Effects on centrality measures						
Treatment	0.964***	0.183***	12.3***	0.187***	0.239**	0.007
	(0.299)	(0.065)	(4.03)	(0.065)	(0.094)	(0.004)
	p = 0.000	p = 0.001	p = 0.001	p = 0.001	p = 0.001	p = 0.116
Control Mean	10.1	0.000	142.9	0.000	0.000	0.299
R^2	0.510	0.414	0.630	0.414	0.367	0.187
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel B. Effects on the probability of being in top 5%						
Treatment	0.023	0.024*	0.037**	0.033**	0.029**	0.006
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
	p = 0.086	p = 0.067	p = 0.004	p = 0.008	p = 0.023	p = 0.691
Control Mean	0.051	0.047	0.046	0.045	0.045	0.051
R^2	0.309	0.230	0.249	0.273	0.283	0.061
Observations	1,402	1,402	1,402	1,402	1,402	1,402

Notes: Panel A shows the treatment effects on five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength (equation 1). Eigenvector, diffusion and betweenness centralities are normalized. Panel B shows the probability of being in the top 5% by centrality within form. Regressions have controls for baseline measure of the outcome (and, in Panel B, baseline centrality measure), gender, SES, stratification bins and class fixed effects. "Control Mean" represents the mean of the outcome in the control arm. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Table 5: Effect of Information Access on Centrality in the Full and Personal Networks

	Degree	Eigenvector	Number of Length-2 Walks	Diffusion	Betweenness	Average Link Strength
Panel A. Personal Network						
Treatment	-0.011	-0.028	-0.378	-0.020	0.013	0.002
	(0.169)	(0.051)	(1.42)	(0.056)	(0.065)	(0.006)
	p = 0.949	p = 0.620	p = 0.801	p = 0.721	p = 0.845	p = 0.738
Control Mean	5.91	0.000	49.9	0.000	0.000	0.325
R^2	0.330	0.346	0.495	0.279	0.174	0.158
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel B. Full Networks						
Treatment	0.822**	0.129**	12.3**	0.140**	0.204**	0.003
	(0.320)	(0.060)	(5.28)	(0.061)	(0.080)	(0.003)
	p = 0.005	p = 0.017	p = 0.014	p = 0.011	p = 0.002	p = 0.406
Control Mean	12.9	0.000	218.3	0.000	0.000	0.193
R^2	0.519	0.424	0.677	0.412	0.368	0.191
Observations	1,402	1,402	1,402	1,402	1,402	1,402

Notes: Treatment effects on personal network (Panel A) and full network (Panel B) on five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength (equation 1). Eigenvector, diffusion and betweenness centralities are normalized. Regressions have controls for baseline measure of the outcome, gender, SES, stratification bins and class fixed effects. "Control Mean" represents the mean of the outcome in the control arm. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Table 6: Alternative Network Definitions

	Links Created	Links Broken	Intersection Degree	In-Degree	Out-Degree	Weighted Degree
Panel A. Information Network						
Treatment	0.647***	-0.316**	0.281***	0.699***	0.247	0.392***
	(0.240)	(0.126)	(0.081)	(0.258)	(0.213)	(0.097)
	p = 0.003	p = 0.009	p = 0.000	p = 0.002	p = 0.218	p = 0.000
Control Mean	6.28	6.33	1.30	5.65	5.77	2.96
R^2	0.247	0.803	0.290	0.599	0.206	0.485
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel B. Personal Network						
Treatment	-0.011	0.0002	-0.051	0.043	-0.205	0.009
	(0.147)	(0.078)	(0.068)	(0.134)	(0.129)	(0.051)
	p = 0.948	p = 0.998	p = 0.448	p = 0.751	p = 0.107	p = 0.860
Control Mean	3.81	3.93	1.41	3.62	3.69	1.86
R^2	0.198	0.810	0.264	0.323	0.228	0.358
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel C. Full Network						
Treatment	0.497*	-0.325**	0.218**	0.608**	0.121	0.200***
	(0.262)	(0.142)	(0.103)	(0.270)	(0.248)	(0.063)
	p = 0.041	p = 0.015	p = 0.034	p = 0.012	p = 0.618	p = 0.001
Control Mean	7.60	7.33	2.59	7.68	7.82	2.41
R^2	0.250	0.779	0.371	0.586	0.248	0.491
Observations	1,402	1,402	1,402	1,402	1,402	1,402

Notes: Treatment effects considering alternative definitions of the network. First and second columns decompose the main effects into links that were created and broken, respectively. Third, fourth and fifth columns alternatively use the intersection, in- and out- degrees. Sixth column computes the weighted degree by the number of interactions within the subcomponents of each network. Panel A considers the information network, followed by the personal network (Panel B) and the full network (Panel C). Regressions have controls for baseline degree, gender, SES, stratification bins and class fixed effects. "Control Mean" represents the mean of the outcome in the control arm. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 7: Dyadic Regressions

	Information	Full	Personal
Panel A. Undirected links			
Treat-Control Link	0.735***	0.694***	0.065
	(0.198)	(0.217)	(0.155)
	p = 0.005	p = 0.018	p = 0.710
Treat-Treat Link	1.49***	1.02*	-0.335
	(0.497)	(0.530)	(0.355)
	p = 0.011	p = 0.121	p = 0.409
R^2	0.1339	0.1152	0.1541
Observations	82,711	82,711	82,711
Panel B. Directed links			
Treat-to-Control Link	0.344**	0.259	-0.070
	(0.140)	(0.158)	(0.111)
	p = 0.053	p = 0.232	p = 0.567
Control-to-Treat Link	0.646***	0.609***	0.124
	(0.145)	(0.163)	(0.115)
	p = 0.004	p = 0.010	p = 0.341
Treat-to-Treat Link	1.00***	0.691**	-0.255
	(0.279)	(0.306)	(0.202)
	p = 0.004	p = 0.084	p = 0.308
R^2	0.0987	0.1231	0.0969
Observations	165,422	165,422	165,422

Notes: Dyadic regressions (equation 2). Unit of observation is a pair of students in the same form and school, *i* and *j*. Panel A: The outcome is coded as 100 if either student named the other as a contact and 0 otherwise. "Treat-Control" is a dummy equal to 1 if *i* is treated and *j* is control, or vice-versa. "Treat-Treat" is a dummy equal to one if both *i* and *j* are in the treatment group. Panel B: The outcome is coded as 100 if *i* named *j* as a contact and 0 otherwise. "Treat-to-Control" is a dummy equal to 1 if *i* is treated and *j* is control, and other covariates are defined similarly. Column "Information" refers to information network, followed by the personal and full networks. Specifications have baseline link, same-class and same-gender controls, and include form fixed effects. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Table 8: Link Formation by Baseline Internet Use

	Undirected link at endline
Control, no internet - Treated, no internet	0.062
	(0.387)
	p = 0.896
Control, no internet - Treated, internet	0.909**
	(0.369)
	p = 0.036
Control, internet - Treated, no internet	0.874**
	(0.365)
	p = 0.064
Control, internet - Treated, internet	1.11***
	(0.391)
	p = 0.023
Treated, no internet - Treated, no internet	2.41**
	(1.03)
	p = 0.028
Treated, no internet - Treated, internet	2.15***
	(0.713)
	p = 0.004
Treated, internet - Treated, internet	-0.557
	(0.919)
	p = 0.616
No internet - internet	-0.798***
	(0.288)
Internet - internet	0.626*
	(0.346)
R^2	0.1345
Observations	82,711

Notes: Dyadic regressions (equation 2). Unit of observation is a pair of students in the same form and school, *i* and *j*. The outcome is coded as 100 if there is a connection and 0 otherwise. "Control, no internet - Treated, no internet" is equal to 1 if *i* is in the control group and had no access to internet at the baseline and *j* is treated group and had no access to internet at the baseline, or vice-versa. "Control, no internet - Treated, internet" is equal to 1 if *i* is in the control group and had no access to internet at the baseline and *j* is treated group and had access to internet at the baseline, or vice-versa. Remaining covariates are defined similarly. Specifications have baseline link, same-class and same-gender controls, and include form fixed effects. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Table 9: Information Network Subcomponents

	Entertainment	Topic learned in class	News	Health	School activities
Panel A. Undirected links					
Treat-Control Link	0.287**	0.473***	0.466***	0.147	0.258**
	(0.131)	(0.120)	(0.111)	(0.101)	(0.120)
	p = 0.045	p = 0.003	p = 0.000	p = 0.234	p = 0.069
Treat-Treat Link	0.688**	0.451	0.673**	0.211	0.661**
	(0.327)	(0.294)	(0.284)	(0.246)	(0.308)
	p = 0.042	p = 0.207	p = 0.019	p = 0.454	p = 0.045
R^2	0.0760	0.0984	0.0380	0.0393	0.0367
Observations	82,711	82,711	82,711	82,711	82,711
Panel B. Directed links					
Treat-to-Control Link	0.158*	0.163**	0.184**	0.065	0.057
	(0.090)	(0.078)	(0.073)	(0.067)	(0.079)
	p = 0.083	p =0.017	p = 0.014	p = 0.336	p = 0.467
Control-to-Treat Link	0.227**	0.360***	0.365***	0.061	0.248***
	(0.092)	(0.085)	(0.078)	(0.067)	(0.083)
	p = 0.042	p = 0.012	p = 0.000	p = 0.549	p = 0.035
Treat-to-Treat Link	0.393**	0.235	0.382***	0.140	0.289*
	(0.176)	(0.153)	(0.148)	(0.130)	(0.158)
	p = 0.038	p =0.219	p = 0.014	p = 0.352	p = 0.105
R^2	0.0522	0.0778	0.0219	0.0254	0.0216
Observations	165,422	165,422	165,422	165,422	165,422

Notes: Dyadic regressions (equation 2). Unit of observation is a pair of students in the same form and school, i and j. Panel A: The outcome is coded as 100 if either student named the other as a contact and 0 otherwise. "Treat-Control" is a dummy equal to 1 if i is treated and j is control, or vice-versa. "Treat-Treat" is a dummy equal to one if both i and j are in the treatment group. Panel B: The outcome is coded as 100 if i named j as a contact and 0 otherwise. "Treat-to-Control" is a dummy equal to 1 if i is treated and j is control, and other covariates are defined similarly. The "entertainment" subcomponent refers to the survey question "*Who do you talk to about movies, music, sports and entertainment?*". "Topic learned in class" refers to the question "*Who do you ask for information that might be useful when researching for a topic learned in class?*". News/health/school activities refers to the question "*Who do you ask for information about the news/health/school activities?*" Specifications have baseline link, same-class and same-gender controls, and include form fixed effects. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Table 10: Personal Network Subcomponents

	Best friend	Borrowed money	Borrowed things	Gift	Personal topics or advice
Panel A. Undirected links					
Treat-Control Link	0.068	0.006	0.024	-0.085	0.267***
	(0.069)	(0.088)	(0.104)	(0.084)	(0.100)
	p = 0.240	p = 0.951	p = 0.847	p = 0.360	p = 0.007
Treat-Treat Link	-0.182	-0.215	0.035	-0.131	-0.298
	(0.150)	(0.193)	(0.249)	(0.195)	(0.220)
	p = 0.218	p = 0.316	p = 0.901	p = 0.559	p = 0.203
R^2	0.2127	0.0517	0.0304	0.0484	0.1339
Observations	82,711	82,711	82,711	82,711	82,711
Panel B. Directed links					
Treat-to-Control Link	0.020	-0.047	0.028	-0.016	0.089
	(0.050)	(0.057)	(0.069)	(0.058)	(0.071)
	p = 0.544	p = 0.423	p = 0.752	p = 0.806	p = 0.155
Control-to-Treat Link	0.114**	0.114*	0.004	-0.071	0.221***
	(0.054)	(0.062)	(0.069)	(0.056)	(0.074)
	p = 0.021	p = 0.082	p = 0.957	p = 0.300	p = 0.003
Treat-to-Treat Link	-0.091	-0.099	0.010	-0.087	-0.117
	(0.084)	(0.101)	(0.128)	(0.103)	(0.122)
	p = 0.340	p = 0.399	p = 0.949	p = 0.502	p = 0.413
R^2	0.1805	0.0344	0.0158	0.0344	0.1038
Observations	165,422	165,422	165,422	165,422	165,422

Notes: Dyadic regressions (equation 2). Unit of observation is a pair of students in the same form and school, i and j. Panel A: The outcome is coded as 100 if either student named the other as a contact and 0 otherwise. "Treat-Control" is a dummy equal to 1 if i is treated and j is control, or vice-versa. "Treat-Treat" is a dummy equal to one if both i and j are in the treatment group. Panel B: The outcome is coded as 100 if i named j as a contact and 0 otherwise. "Treat-to-Control" is a dummy equal to 1 if i is treated and j is control, and other covariates are defined similarly. "Best friend" refers to the survey question "*Who is your best friend?*". "Borrowed money/things" refers to the question "*Who have you borrowed money/things from?*". "Gift" refers to the question "*Who have you given a gift to?*"; the direction of this link is inverted for consistency of interpretation. "Personal topic or advice" refers to "*Who do you talk to about personal topics or ask for advice?*" Specifications have baseline link, same-class and same-gender controls, and include form fixed effects. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Table 11: Calibration

Moment	Simulated	Empirical
Panel A. Average centrality		
Degree (control)	9.85	10.1
Degree (treatment)	10.7	10.9
Eigenvector centrality (control)	0.447	0.324
Eigenvector centrality (treatment)	0.488	0.345
Number of length-2 walks (control)	120	143
Number of length-2 walks (treatment)	130	153
Diffusion centrality (control)	3.49	3.55
Diffusion centrality (treatment)	3.8	3.76
Betweenness centrality (control)	0.447	0.0117
Betweenness centrality (treatment)	0.488	0.0138
Panel B. Probability of being in top 5% by centrality		
Top 5% by degree (treatment)	0.0908	0.0697
Top 5% by eigenvector centrality (treatment)	0.0738	0.0627
Top 5% by number of length-2 walks (treatment)	0.0749	0.0732
Top 5% by diffusion centrality (treatment)	0.0739	0.0697
Top 5% by betweenness centrality (treatment)	0.0732	0.0697

Notes: Comparing moments from simulated endline networks to empirical moments. Excludes moments used for calibration. 10,000 117-node networks simulated based on the model and calibration in Section 5.2. Empirical moments based on the information network. Centrality measures are not normalized. Panel A: averages over simulated nodes and networks. Panel B: averages over simulated networks.

Online Appendix

"Who knows? The effect of information access on social network position"

For Online Publication

Laura Derksen and Pedro Souza

November 7, 2022

A Appendix

A.1 Proof of Theorem 5.1

Proof. For a node in the treatment group, the expected degree is

$$\begin{aligned}\mathbb{E}(d_i|T_i=1) &= (N_T-1)P_{TT}+N_CP_{TC}\\ &= (N_T-1)P_{TT}+N_CP_{CC}+N_C(P_{TC}-P_{CC}).\end{aligned}$$

For a node in the control group, the expected degree is

$$\begin{aligned}\mathbb{E}(d_i|C_i=1) &= (N_C-1)P_{CC}+N_TP_{TC}\\ &= (N_C-1)P_{CC}+N_TP_{CC}+N_T(P_{TC}-P_{CC})\\ &= (N-1)P_{CC}+N_T(P_{TC}-P_{CC}).\end{aligned}$$

Because $N_C > N_T$ and $P_{TC} > P_{CC}$,

$$\mathbb{E}(d_i|T_i=1)-\mathbb{E}(d_i|C_i=1) > (N_T-1)P_{TT}+N_CP_{CC}-(N-1)P_{CC} \geq 0.$$

□

A.2 Calibration Details

To calibrate the model, we match parameters to moments in our empirical information network as follows. First, we note that under this model both the baseline network and endline network are still general random graphs, with unconditional link probabilities represented similarly to those in equation (4), but with with link probabilities that depend on the academic types of the nodes.

Between pairs of control nodes, these link probabilities are the same at baseline and at endline, and are symmetric in θ_1 and θ_2.

$$\mathbb{P}(g_{ij}^0=1|T_i=T_j)=\mathbb{P}(g_{ij}^1=1|T_i=T_j)=P_{CC}^{\theta_1\theta_2}=P_{CC}^{\theta_2\theta_1}\equiv\mathbb{P}(\nu>-\kappa_{CC}^{\theta_1\theta_2})\mathbb{P}(\nu>-\kappa_{CC}^{\theta_2\theta_1})$$

In our data, the probability of an endline information-link between two control students in the same school and form is used to estimate these probabilities as follows:

$$\hat{P}_{CC}^{LL}=0.08 \qquad \hat{P}_{CC}^{HL}=0.11 \qquad \hat{P}_{CC}^{HH}=0.21$$

These estimates allow us to simulate a simple baseline network. To simulate a corresponding endline network, we start by constructing a "shadow" network $\tilde{g}^1$. This is the network of links that would exist at endline absent the intervention, but allowing for residual network changes to occur over time. In order to simulate shadow network that is suitably correlated with the baseline network, we must estimate the probability of a shadow link (or equivalently, an endline link) between control nodes conditional on a baseline link

$$\mathbb{P}(g^1_{ij}=1|g^0_{ij}=1,T_i=T_j=0,\theta_i=\theta_1,\theta_j=\theta_2)=(1-\delta)^2+2\delta(1-\delta)\sqrt{P^{\theta_1\theta_2}_{CC}}+\delta^2P^{\theta_1\theta_2}_{CC}\equiv P^{\theta_1\theta_2}_{CC|CC} \quad (5)$$

$$\mathbb{P}(\tilde{g}^1_{ij}=1|g^0_{ij}=1,\theta_i=\theta_1,\theta_j=\theta_2)=P^{\theta_1\theta_2}_{CC|CC}$$

Note that this probability depends on the types $\{\theta_i,\theta_j\}$ but is symmetric in these types. If a link exists in the baseline network, the probability it should appear in the shadow network, $\hat{P}^{\theta_1\theta_2}_{CC|CC}$ is estimated directly from the moment (5) in the data. That is, we take the probability that a control-pair with types θ_1 and θ_2 is linked at endline, conditional on a link existing at baseline:

$$\hat{P}^{LL}_{CC|CC}=0.35 \qquad \hat{P}^{HL}_{CC|CC}=0.41 \qquad \hat{P}^{HH}_{CC|CC}=0.48$$

Conversely, if a link does not exist in the baseline network, the probability that it should appear in the shadow network can be calculated using Bayes' rule.

$$\mathbb{P}(\tilde{g}^1_{ij}=1|g^0_{ij}=0,\theta_i=\theta_1,\theta_j=\theta_2)=\frac{P^{\theta_1\theta_2}_{CC}}{1-P^{\theta_1\theta_2}_{CC}}\left(1-P^{\theta_1\theta_2}_{CC|CC}\right)$$

This probability, that a pair of control students is linked at endline given there is no link at baseline, is also estimated directly from the corresponding moment in the data.

$$\hat{P}^{LL}_{CC|!CC}=0.05 \qquad \hat{P}^{HL}_{CC|!CC}=0.07 \qquad \hat{P}^{HH}_{CC|!CC}=0.13$$

Next, we simulate an endline network by adding links to to the shadow network. We assume that for fixed θ_1,θ_2, $\kappa^{\theta_1\theta_2}_{CC}$ is weakly smaller than $\kappa^{\theta_1\theta_2}_{TC}$, $\kappa^{\theta_1\theta_2}_{CT}$ and $\kappa^{\theta_1\theta_2}_{TT}$. That is, information is valuable and not costly to spread. This implies that $P^{\theta_1\theta_2}_{CC}$ is weakly smaller than $P^{\theta_1\theta_2}_{TC}$ and $P^{\theta_1\theta_2}_{TT}$, consistent with our reduced form empirical results (see Table 7). Then, conditional on having a link in the shadow network, the probability of having a link in the endline network is one.

$$\mathbb{P}(g^1_{ij}=1|\tilde{g}^1_{ij}=1,\theta_i=\theta_1,\theta_j=\theta_2)=1$$

Conditional on having no link in the shadow network, the probability of a link in the endline network depends on the treatment statuses and academic types of the nodes involved.

$$\mathbb{P}(g^1_{ij}=1|\tilde{g}^1_{ij}=0,T_i=T_j=0,\theta_i=\theta_1,\theta_j=\theta_2)=0$$

$$\mathbb{P}(g^1_{ij}=1|\tilde{g}^1_{ij}=0,T_i=T_j=1,\theta_i=\theta_1,\theta_j=\theta_2)=1-\frac{\mathbb{P}(\nu<-\kappa^{\theta_1\theta_2}_{TT})\mathbb{P}(\nu<-\kappa^{\theta_2\theta_1}_{TT})}{\mathbb{P}(\nu<-\kappa^{\theta_1\theta_2}_{CC})\mathbb{P}(\nu<-\kappa^{\theta_2\theta_1}_{CC})}=\frac{P^{\theta_1\theta_2}_{TT}-P^{\theta_1\theta_2}_{CC}}{1-P^{\theta_1\theta_2}_{CC}}$$

$$\mathbb{P}(g^1_{ij}=1|\tilde{g}^1_{ij}=0,T_i=1,T_j=0,\theta_i=\theta_1,\theta_j=\theta_2)=1-\frac{\mathbb{P}(\nu<-\kappa^{\theta_1\theta_2}_{TC})\mathbb{P}(\nu<-\kappa^{\theta_2\theta_1}_{CT})}{\mathbb{P}(\nu<-\kappa^{\theta_1\theta_2}_{CC})\mathbb{P}(\nu<-\kappa^{\theta_2\theta_1}_{CC})}=\frac{P^{\theta_1\theta_2}_{TC}-P^{\theta_1\theta_2}_{CC}}{1-P^{\theta_1\theta_2}_{CC}}$$

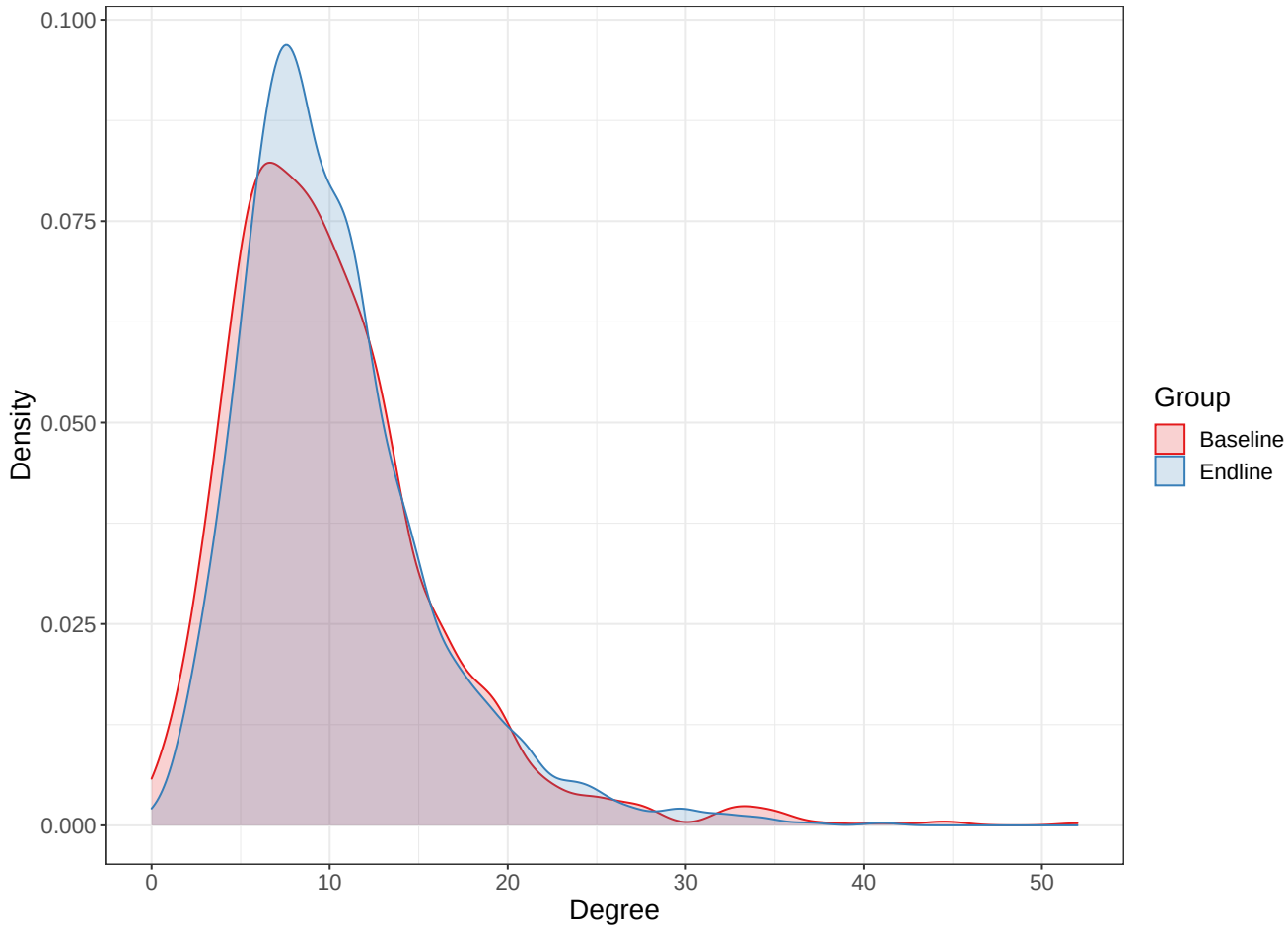
We estimate the relevant moments from our endline data as follows:

$\hat{P}_{TT}^{LL} = 0.09$	$\hat{P}_{TT}^{HL} = 0.13$		$\hat{P}_{TT}^{HH} = 0.29$
$\hat{P}_{TC}^{LL} = 0.08$	$\hat{P}_{TC}^{HL} = 0.12$	$\hat{P}_{TC}^{LH} = 0.12$	$\hat{P}_{TC}^{HH} = 0.25$

These calculations allow us to simulate an endline network based on the shadow network. We now have all the required ingredients to simulate an baseline network, a shadow network, and an endline network with appropriately correlated links.

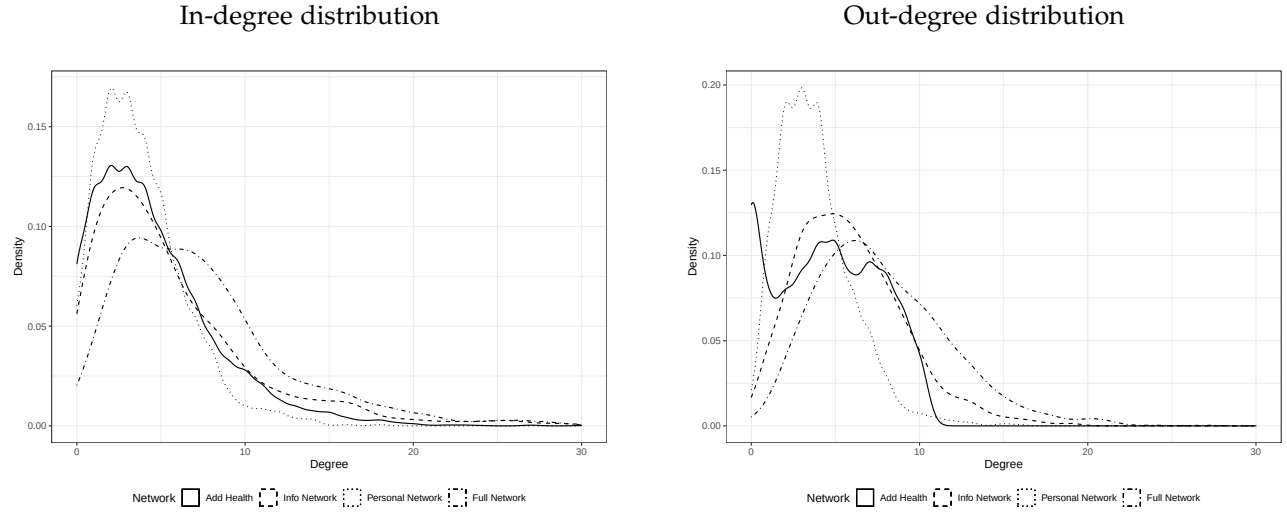
Appendix Figures and Tables

Appendix Figure A1: Degree Distribution at Baseline and Endline, Information Network



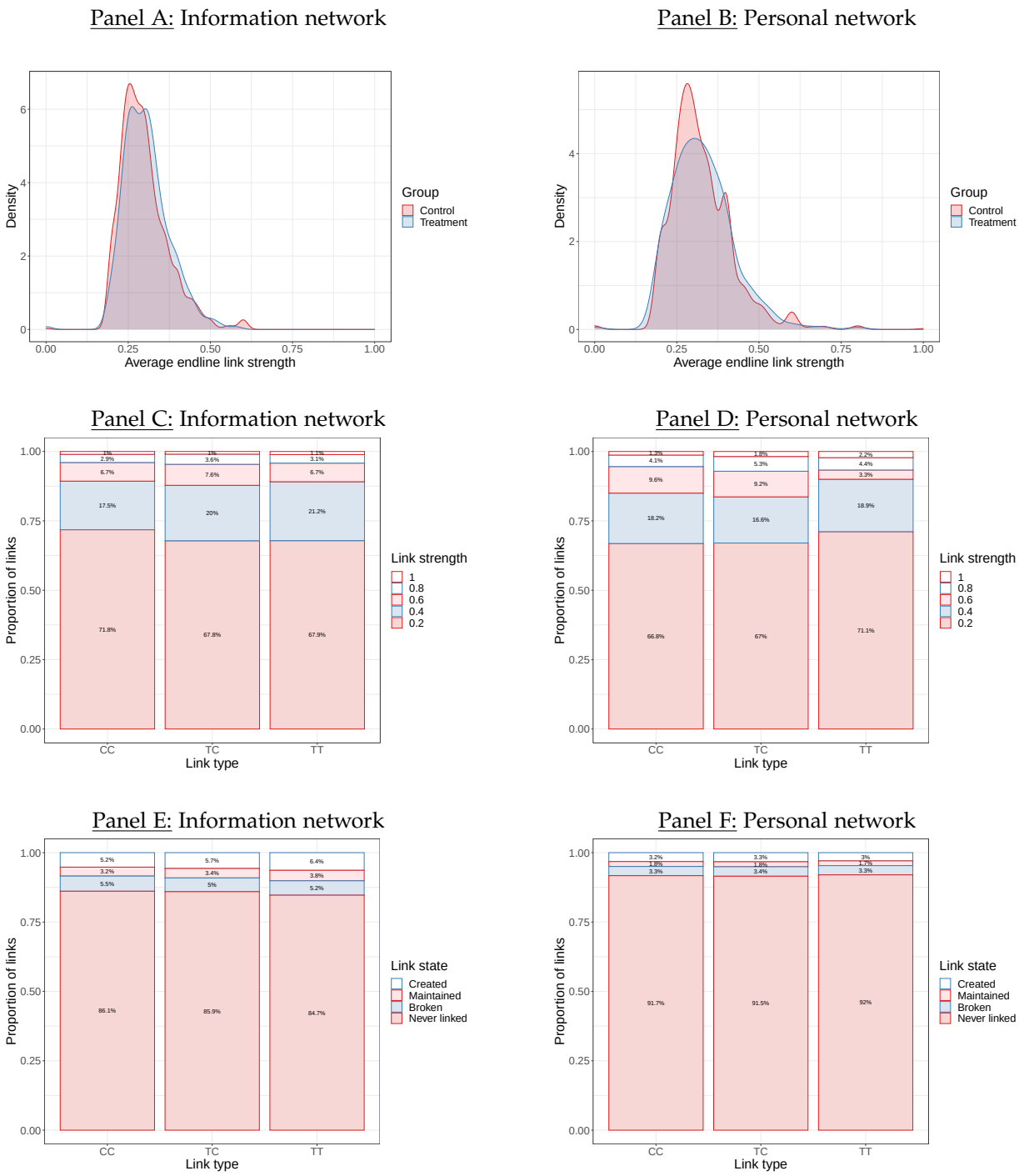
Notes: Degree distribution at baseline and endline. Information network. Sample restricted to nodes observed at both times.

Appendix Figure A2: Degree Distribution in Our Data vs AddHealth



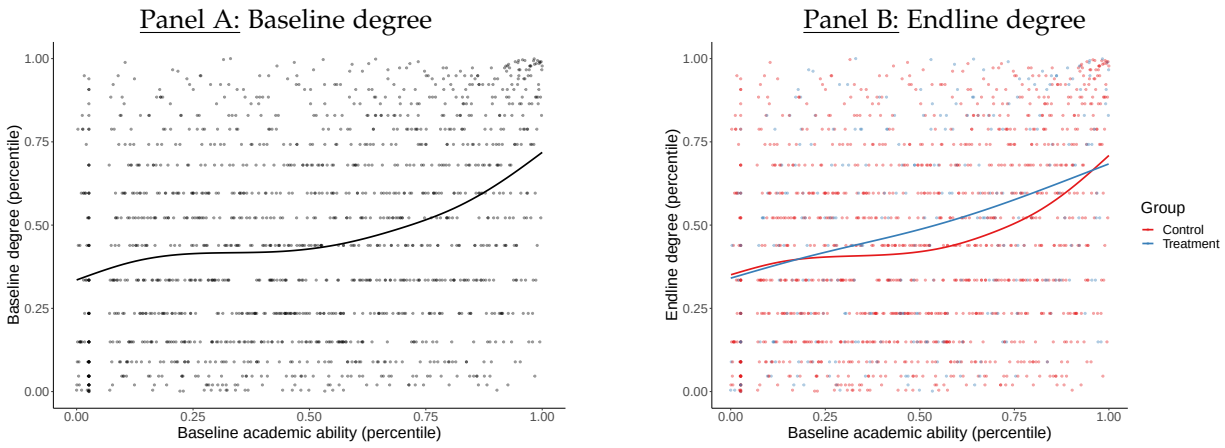
Notes: Degree distributions computed over our data for the information, personal and full networks; and the National Longitudinal Study of Adolescent Health ("AddHealth") obtained from https://www.icpsr.umich.edu/web/ICPSR/studies/21600/datasets/0003/variables/ODGX2?archive=icpsr and https://www.icpsr.umich.edu/web/ICPSR/studies/21600/datasets/0003/variables/IDGX2?archive=icpsr

Appendix Figure A3: Link Strength and Link Dynamics



Notes: Panels A-D: link strength between pairs of nodes. Each link consists of five sublinks (see Table 1), strength is defined as the fraction of sublinks present. In all four panels, strength is calculated conditional on the presence of a link. Panels E-F: link dynamics between baseline and endline.

Appendix Figure A4: Baseline Academic Ability and Network Degree



Notes: Correlation between baseline academic ability (percentile) and network degree (percentile) at baseline and endline.

Appendix Table A1: Comparison to Alternative Networks

	Full Network		Coleman High School		Diffusion of Microfinance	
	Mean	SD	Mean	SD	Mean	SD
Degree	12.71	5.98	7.83	3.43	8.43	5.92
Eigenvector Centrality	0.33	0.18	0.29	0.22	0.07	0.13
Number of length-2 walks	216.45	98.32	80.74	39.62	115.79	113.33
Diffusion	3.52	1.75	4.24	2.40	2.60	3.02
Betweenness	0.01	0.01	0.02	0.03	0.00	0.01

Notes: Comparison to alternative networks. "Coleman High School" is the high-school network from Coleman (1964). "Diffusion of Microfinance" from Banerjee et al. (2013). "SD" refers to the standard deviation across observations.

Appendix Table A2: Effect of Information Access on Centrality in the Contact Network

	Degree	Eigenvector	Number of Length-2 Walks	Diffusion	Betweenness	Average Link Strength
Treatment	-0.093	-0.040	-1.41	-0.037	0.013	0.010*
	(0.246)	(0.058)	(3.55)	(0.060)	(0.072)	(0.006)
	p = 0.705	p = 0.497	p = 0.670	p = 0.536	p = 0.849	p = 0.071
Control Mean	10.8	0.000	143.9	0.000	0.000	0.464
R^2	0.285	0.281	0.361	0.237	0.099	0.150
Observations	1,402	1,402	1,402	1,402	1,402	1,402

Notes: Treatment effects on the contact network along five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength (equation 1). Eigenvector, diffusion and betweenness centralities are normalized. Contact links are identified based on the survey question "*[1,2,3] days ago, did you just hang out, have conversations or play with friends?*" Column 6 is calculated based on the fraction of days during which the pair spent time together. Regressions have controls for baseline measure of the outcome, gender, SES, stratification bins and class fixed effects. "Control Mean" represents the mean of the outcome in the control arm. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Appendix Table A3: Heterogeneous Effects

	Degree	Eigenvector	Number of Length-2 Walks	Diffusion	Betweenness	Average Link Strength
Panel A. By use of the digital Library						
Treatment	0.426	0.065	7.28	0.059	0.024	0.005
	(0.339)	(0.071)	(4.98)	(0.071)	(0.080)	(0.006)
Treatment x High Browsing	1.03*	0.227**	9.69	0.247**	0.414**	0.005
	(0.550)	(0.115)	(7.59)	(0.116)	(0.165)	(0.008)
Control Mean	10.1	0.000	142.9	0.000	0.000	0.299
R^2	0.512	0.416	0.630	0.416	0.373	0.187
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel B. By academic ability						
Treatment	0.918**	0.165*	13.4**	0.167*	0.206	0.011*
	(0.396)	(0.095)	(5.26)	(0.093)	(0.144)	(0.006)
	p = 0.008	p = 0.027	p = 0.005	p = 0.028	p = 0.024	p = 0.111
Treatment x High Achiever	0.092	0.036	-2.20	0.042	0.068	-0.007
	(0.601)	(0.130)	(8.10)	(0.131)	(0.190)	(0.009)
	p = 0.865	p = 0.747	p = 0.769	p = 0.714	p = 0.609	p = 0.419
Control Mean	10.1	0.000	142.9	0.000	0.000	0.299
R^2	0.510	0.414	0.630	0.414	0.367	0.187
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel C. By SES						
Treatment	1.10***	0.256**	14.8***	0.260**	0.282*	0.008
	(0.423)	(0.103)	(5.18)	(0.104)	(0.161)	(0.007)
	p = 0.002	p = 0.001	p = 0.002	p = 0.002	p = 0.007	p = 0.220
Treatment × SES	-0.273	-0.145	-4.96	-0.143	-0.086	-0.002
	(0.608)	(0.132)	(8.16)	(0.133)	(0.195)	(0.009)
	p = 0.606	p = 0.196	p = 0.512	p = 0.205	p = 0.513	p = 0.802
Control Mean	10.1	0.000	142.9	0.000	0.000	0.299
R^2	0.510	0.415	0.630	0.415	0.367	0.187
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel D. By gender						
Treatment	0.908**	0.201**	17.1**	0.180**	0.136	0.013**
	(0.444)	(0.082)	(6.63)	(0.083)	(0.106)	(0.006)
	p = 0.023	p = 0.011	p = 0.006	p = 0.022	p = 0.133	p = 0.025
Treatment × Male	0.101	-0.033	-8.69	0.012	0.186	-0.011
	(0.593)	(0.121)	(8.25)	(0.123)	(0.171)	(0.009)
	p = 0.854	p = 0.770	p = 0.260	p = 0.914	p = 0.166	p = 0.225
Control Mean	10.1	0.000	142.9	0.000	0.000	0.299
R^2	0.510	0.414	0.630	0.414	0.368	0.188
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel E. By baseline degree						
Treatment	0.751**	0.134*	10.6**	0.142*	0.082	0.007
	(0.349)	(0.071)	(4.75)	(0.073)	(0.078)	(0.007)
	p = 0.025	p = 0.054	p = 0.025	p = 0.045	p = 0.230	p = 0.311
Treatment × High Degree	0.451	0.104	3.67	0.096	0.338*	9.53×10^{-5}
	(0.617)	(0.134)	(8.39)	(0.135)	(0.200)	(0.009)
	p = 0.403	p = 0.360	p = 0.634	p = 0.400	p = 0.014	p = 0.992
Control Mean	10.1	0.000	142.9	0.000	0.000	0.299
R^2	0.510	0.414	0.630	0.414	0.371	0.189
Observations	1,402	1,402	1,402	1,402	1,402	1,402

Notes: Heterogeneous treatment effects on the information network along five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength. Eigenvector, diffusion and betweenness centralities are normalized. Panel A interacts the treatment variable with above-median hours of the digital library use during the experiment ("High Browsing"); Panel B with above-median exam scores at the baseline; Panel C with SES (SES) defined as respondent's house having access to electricity and running water; Panel D with gender; and Panel E with above-median baseline degree. Regressions have controls for the covariate main effect, baseline degree, gender, SES, stratification bins and class fixed effects. "Control Mean" represents the mean of the outcome in the control arm. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.

Appendix Table A4: Robustness to the Exclusion of Covariates

	Degree	Eigenvector	Number of Length-2 Walks	Diffusion	Betweenness	Average Link Strength
Panel A. Information Networks						
Treatment	0.851**	0.138*	11.4**	0.151**	0.222**	0.007*
	(0.356)	(0.074)	(4.78)	(0.075)	(0.103)	(0.004)
	p = 0.008	p = 0.040	p = 0.011	p = 0.027	p = 0.004	p = 0.116
Control Mean	10.1	0.000	142.9	0.000	0.000	0.299
R^2	0.225	0.097	0.443	0.094	0.071	0.117
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel B. Personal Network						
Treatment	-0.018	-0.055	-0.724	-0.043	-0.002	0.004
	(0.181)	(0.061)	(1.58)	(0.063)	(0.066)	(0.006)
	p = 0.923	p = 0.399	p = 0.669	p = 0.509	p = 0.971	p = 0.561
Control Mean	5.91	0.000	49.9	0.000	0.000	0.325
R^2	0.181	0.053	0.363	0.052	0.045	0.105
Observations	1,402	1,402	1,402	1,402	1,402	1,402
Panel C. Full Network						
Treatment	0.734*	0.089	11.0*	0.106	0.183**	0.003
	(0.383)	(0.071)	(6.37)	(0.071)	(0.091)	(0.003)
	p = 0.039	p = 0.176	p = 0.066	p = 0.110	p = 0.014	p = 0.356
Control Mean	12.9	0.000	218.3	0.000	0.000	0.193
R^2	0.257	0.095	0.505	0.094	0.067	0.108
Observations	1,402	1,402	1,402	1,402	1,402	1,402

Notes: Treatment effects on five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength. Eigenvector, diffusion and betweenness centralities are normalized. Regressions include only stratification bins. Panel A considers the information network, followed by the personal network (Panel B) and the full network (Panel C). "Control Mean" represents the mean of the outcome in the control arm. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Appendix Table A5: Treatment Effects in Low-Attrition Schools

	Degree	Eigenvector	Number of Length- 2 Walks	Diffusion	Betweenness	Average Link Strength
Panel A. Information Network						
Treatment	1.12***	0.155**	14.6**	0.174**	0.228**	0.007
	(0.421)	(0.075)	(5.87)	(0.079)	(0.102)	(0.006)
	p = 0.003	p = 0.029	p = 0.008	p = 0.017	p = 0.004	p = 0.245
Control Mean	10.9	0.000	164.7	0.000	0.000	0.296
R^2	0.515	0.444	0.622	0.428	0.402	0.204
Observations	791	791	791	791	791	791
Panel B. Information Network: probability of being in top 5%						
Treatment	0.024	0.021	0.023	0.022	0.023	0.014
	(0.020)	(0.019)	(0.019)	(0.019)	(0.020)	(0.021)
	p = 0.186	p = 0.243	p = 0.203	p = 0.199	p = 0.189	p = 0.476
Control Mean	0.052	0.047	0.051	0.047	0.047	0.049
R^2	0.305	0.256	0.286	0.304	0.258	0.062
Observations	791	791	791	791	791	791
Panel C. Personal Network						
Treatment	0.192	0.042	1.16	0.033	0.058	-0.004
	(0.236)	(0.071)	(2.05)	(0.077)	(0.085)	(0.007)
	p = 0.417	p = 0.573	p = 0.598	p = 0.677	p = 0.493	p = 0.602
Control Mean	6.27	0.000	56.1	0.000	0.000	0.317
R^2	0.366	0.346	0.527	0.277	0.199	0.160
Observations	791	791	791	791	791	791
Panel D. Full Network						
Treatment	0.984**	0.117	15.4**	0.130*	0.191**	0.004
	(0.458)	(0.076)	(7.77)	(0.079)	(0.095)	(0.004)
	p = 0.020	p = 0.107	p = 0.042	p = 0.079	p = 0.017	p = 0.307
Control Mean	13.9	0.000	253.0	0.000	0.000	0.189
R^2	0.522	0.423	0.679	0.405	0.367	0.211
Observations	791	791	791	791	791	791

Notes: Regression restricting the sample to the two national schools, which have very low attrition. Panel A shows the treatment effects on five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength on the information network (equation 1). Eigenvector, diffusion and betweenness centralities are normalized. Regressions have controls for baseline measure of outcome (and, in Panel B, baseline centrality measure), SES, stratification bins and class fixed effects. Panel B shows the probability of being in the top 5% central within forms on the information network. Panel C observes the effect on personal networks, and Panel D on the full network. "Control Mean" represents the mean of the outcome in the control arm. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Appendix Table A6: Dynamics of Link Formation

	Link at endline	Info link created	Info link broken	Personal link created	Personal link broken	Full network link created	Full network link broken
Treat-Control x No Baseline Link	0.463**						
	(0.183)						
	p = 0.053						
Treat-Control x Baseline Link	3.68***						
	(1.25)						
	p = 0.008						
Treat-Treat x Baseline Link	5.25*						
	(2.90)						
	p = 0.035						
Treat-Treat x No Baseline Link	1.13**						
	(0.465)						
	p = 0.108						
Treat-Control		0.465***	-0.525***	0.069	0.081	0.423**	-0.456***
		(0.169)	(0.162)	(0.132)	(0.133)	(0.184)	(0.174)
		p = 0.027	p = 0.024	p = 0.655	p = 0.626	p = 0.074	p = 0.064
Treat-Treat		1.05**	-0.378	-0.281	-0.045	0.522	-0.366
		(0.428)	(0.390)	(0.301)	(0.316)	(0.447)	(0.419)
		p = 0.030	p = 0.465	p = 0.426	p = 0.906	p = 0.330	p = 0.506
R^2	0.1342	0.0273	0.0388	0.0163	0.0235	0.0285	0.0405
Observations	82,711	82,711	82,711	82,711	82,711	82,711	82,711

Notes: Dyadic regressions. Unit of observation is a pair of students in the same form and school, *i* and *j*. The outcome is coded as 100 if there is a connection and 0 otherwise. "Treat-Control" is equal to 1 if *i* is treated and *j* is control, or vice-versa. "Treat-Treat" is equal to 1 if both *i* and *j* are in the treatment group. Covariates are interacted with the indicator for presence of link at the baseline. Specifications have baseline link, same-class and same-gender controls, and include form fixed effects. Heteroskedasticity-robust standard errors in parentheses and randomization inference p-value with "p = ". Stars represent classical inference p-values with *** p<0.01; ** p<0.05; * p<0.1.