Who Knows? Information Access and Endogenous Network Formation

Laura Derksen and Pedro CL Souza*

July 24, 2024

Abstract

Networks play a key role in information diffusion. Yet, the impact of information on network formation is 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. Using a complete panel of detailed network data, we show that changes in information access affect network structure, as students form and maintain strategic links. At the endline, treated students are more well-connected than control students. We calibrate and simulate a model of strategic network formation to demonstrate implications for network-based targeting, information diffusion, academic welfare and inequality.

JEL classification: I25 D83 D85 O12 L86 Z13

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

*Derksen: Ragnar Frisch Centre for Economic Research and Department of Management, University of Toronto (laura.derksen@frisch.uio.no); 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, Ethan Sansom and Emira Sopi 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. American Economic Association RCT Registry number AEARCTR-0003824. This paper was previously circulated under the title "Who Knows? The Effect of Information Access on Social Network Centrality." All errors and omissions are our own.

1 Introduction

The spread of information through social networks shapes individual beliefs and decisions across many domains, including finance (Banerjee et al., 2013), policy (Bertrand et al., 2014), reproductive health (Baumgartner et al., 2022) and technology adoption (Beaman et al., 2021). In theory, information diffusion depends on the structure of the network, and on the network position of each informed node (Jackson et al., 2017). If networks remain fixed over time, it may be possible to amplify information diffusion by targeting well-connected people (Banerjee et al., 2019). Yet, gaining direct access to information might also *cause* a person to become well connected, as others form strategic links to obtain access. If networks change in response to policy, network-based targeting may become unnecessary or inefficient.

In this paper, we show that differential access to information has a causal impact on network structure. 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.

We first observe that the intervention was indeed a shock to information access for treatment students. We find that students used the resource intensively to search for many types of information, including information instrumental to their education (22 percent of browsing time), general knowledge (such as politics, health and world news), sexuality (7 percent of browsing time), popular culture 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). We hypothesized that this shock to information access would make treated students valuable information sources, for control students and for each other, and lead to the strategic formation of information links. Because information is non-rival, information networks might respond flexibly to such a shock.

Our main finding is that consistent, exclusive information access causes network structure to change

over the long term, as connections between informed students and their peers form and persist with higher probability. We examine link formation at the dyad level, as measured eight months after the start of the intervention. We find that, in the information-sharing network, treat-control pairs are .735 percentage points more likely to connect compared to pairs of control students ($p = .005$). This change, while concentrated in the information-sharing network, also appears in the full network, which includes both information-sharing links and personal friendships. This implies that students form completely new links for information sharing, as opposed to relying on existing information links or personal friendships. We also observe that pairs of treated students are 1.49 percentage points more likely to form information links, relative to pairs of control students ($p = .011$), despite having access to the same information source. Because the number of treat-control pairs in the network is approximately eight times larger than the number of treat-treat pairs, the majority of new links are of the treat-control type.[1]

These connections appear to be used for broad information sharing, as opposed to simply learning about the new technology itself. When we decompose the treatment effect by past internet experience, we find that the effect is strongest for control students who do have past experience, as opposed to those who would have the most to learn. For pairs of treated students, we register the opposite pattern: the 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. In every case we observe an increase in information-sharing links across a wide range of topics, measured eight months after the start of the intervention. We see no increase in personal connections of any type between pairs of treated students.

At endline, treated students are significantly better-connected than control students according to standard network centrality measures. We focus on five measures of centrality in the information-sharing network. On average, treated students have a higher degree (.964 additional links) and eigenvector centrality (.18 standard deviations) than control students. Treated students also have higher diffusion centrality, with two different sets of parameters, and higher betweenness centrality. These differences are highly significant, with randomization inference p-values less than or equal to .001. Treated students are in fact more likely to be among the top five percent by all centrality measures (with $p < .1$). We do not detect any significant difference in centrality in the personal friendship network or among other types of contacts. These differences in network centrality are driven by differences in the number of links as opposed to differences in link strength; treated students form more new links, and maintain more existing links than control students.

We ended our intervention with a follow-up exercise to document information diffusion between treated and control students directly. To do this, we conducted a two-week long incentivized experiment. This experiment took place *after* all other data had been collected from the field to avoid contaminating our endline network measures. We assigned a unique question about a recent news event to each student, and

[1]Specifically, we estimate that for every treat-treat link induced by the experiment, there are four treat-control links.

tracked correct answers as well as information sources. We indeed observe that control students were able to access the new information source indirectly through social ties. We observe widespread information diffusion from treated students to control students. 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.

We introduce and calibrate a network formation model with a dual purpose. First, it allows us to interpret the reduced-form estimates – which capture a *relative* change between the treated and the control –, as the true changes under a no-intervention counterfactual. Indeed, model simulations confirm that the no-intervention treatment effects are similar to the ones we observe. Second, and importantly, the model allows us to illustrate the implications of our findings for network-based targeting, information diffusion and cost efficiency. The fact that we observe a significant change between baseline and endline in network structure, and in the composition of central nodes, could potentially offset the advantages of network-based targeting.

The model simulations show that while centrality-based targeting does lead to more information diffusion than random targeting, the gap in diffusion is cut by half due to the network response. Baseline centrality measures remain strong predictors of endline centrality, and remain good candidates for network-based targeting. Yet, random targeting can be used to reach just as many people as network-centrality-based targeting by adding a few additional random seeds. Random targeting may therefore be more cost efficient in many contexts. Network-based targeting that relies on centrality measures appears to outperform targeting strategies that rely on precise network positions, as precise targeting strategies are very sensitive to the network change.

Finally, we use an information diffusion model to characterize the direct and indirect effects of the intervention on academic performance, and to illustrate the implications of network-based targeting for academic welfare and inequality. In previous work, the intervention was shown to have directly benefited low-ability students who are, on average, also less central (Derksen et al., 2022). Yet, this reduced-form evidence cannot identify the total effect of the intervention, relative to a no-intervention counterfactual. It also cannot fully characterize the network-based spillovers, as the network changed over time. We introduce and calibrate a model in which academic performance depends on both the direct effect of the intervention, as well as information diffusion through the network. We find that the intervention likely had large spillovers in our context, and that the total effect of the intervention exceeds the direct effect on treated students. Moreover, centrality-based targeting leads to a larger increase in academic performance for both high and lower ability students, due to the large spillover effects. Yet, high ability students benefit the most from network-based targeting, and random targeting narrows the gap between high and lower ability students more effectively.

This study has implications for policies that target individuals based on their network positions. Sev-

eral studies have shown that simple messages diffuse more widely in the network when central messengers are targeted (Banerjee et al., 2013, 2019; Islam et al., 2021). 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 select candidates for a longer term and more intensive intervention or role may be less effective, as the network partially 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 longer-term studies includes a random selection arm. Random selection could lead to more (or less) network adaptation than selection by authority figures such as teachers or extension workers, for example if the candidates chosen by authority figures are generally less accessible to their peers. Akbarpour et al. (2021) show theoretically that random targeting, while suboptimal, can ultimately produce a level of diffusion that is not far from the maximum level. If, in addition, the network adapts, the benefits of network-based targeting will be further reduced. 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). Finally, collecting complete network data can be costly (Breza et al., 2019), though network-based targeting need not be very costly to implement in practice, as simple survey measures can yield useful proxies (Banerjee et al., 2019).

Network-based targeting strategies can also exacerbate inequality, as central individuals are likely privileged in other ways (Jackson, 2019; Alan et al., 2021). 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, and stand to benefit most from network-based targeting. 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 relative 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.

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.

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 microfinance 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.[3]

We also contribute to an emerging empirical literature on the determinants of network structure, by showing that information access can affect network centrality. 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; Alan et al., 2021). Our findings suggest that network centrality is affected 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.

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

[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).

[3]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).

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 models of strategic network formation and information diffusion. 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 link formation. 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 can therefore observe how differences at the node level lead to relative differences in link probabilities and node-level network positions; 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).[4] 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

[4]With a much larger sample size, we could have also assigned some entire networks to a pure control arm. This combination of cluster- and individual-level randomization would have allowed us to estimate the impact on both relative and absolute differences in network centrality.

co-educational district boarding schools. Students progress through four forms (grade levels), and each form is divided into three different classrooms.[5]

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.[6]

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 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.

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

[6]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.

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. We conducted an incentivized information experiment *after* the endline survey had taken place, and networks were measured. The network measures were therefore not contaminated or distorted by these incentives. We then conducted a small follow-up survey after conclusion of the two-week-long experiment to understand the extent of information spread. We discuss those results in Section

2.2 below. 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 the Online Appendix to Derksen et al. (2022).[7]

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.[8] 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. We conducted an incentivized information-seeking experiment to determine to what extent students were able to access information at school. The experiment took place *after endline data had been collected* and, in particular, after endline networks were measured. 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. 60 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

[7] Available at https://ars.els-cdn.com/content/image/1-s2.0-S0304387821001632-mmc1.pdf.

[8] Source: https://en.wikipedia.org/wiki/2017 and https://en.wikipedia.org/wiki/2018. We used Wikipedia itself to gather a list of 64 major events that happened in the world during our study period. We manually classified the events that were specific to Africa.

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 (97 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. The endline survey took place towards the end of the school year but before the information-seeking experiment described in Section 2.2. At this point network link creation had not been incentivized beyond the Wikipedia intervention itself. 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.[9]

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.[10] For this reason, we focus our

[9] 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?*'

[10] 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.[11] 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

[11] 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). [12]

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

[12] 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 node-level network characteristics including the number of links, the number of treated links, and 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. 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.

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").[13] In Panel A of Appendix Figure A2 we see that the Add-Health in-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 link formation and network centrality.

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 estimate dyad-level regressions to examine strategic link formation between individual students. We then use individual-level linear regressions to determine whether the intervention impacted network centrality measures. 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 Link Formation

Undirected links. We first 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:

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

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

[13] 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.

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 compute classical heteroskedasticity-robust standard errors, but rely on randomization inference to construct p-values, as standard errors may be biased when an intervention is randomly assigned to nodes in a social network (Abadie et al., 2016; Fredrickson and Chen, 2019). Indeed, our intervention likely affects network nodes indirectly, resulting in a kind of "fuzzy clustering" that randomization inference can usefully address (Abadie et al., 2016; Blattman et al., 2021). In particular, we can use randomization inference p-values to reject the sharp null hypothesis of no treatment effect under any vector of treatment assignments (see Appendix A.1 for additional discussion).

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

$$100 \times \text{link}^1_{ij} = \beta_0 + \beta_1 \cdot TC_{ij} + \beta_2 \cdot CT_{ij} + \beta_3 \cdot TT_{ij} + \alpha \cdot \text{link}^0_{ij} + \mathbf{x}'_{ij}\boldsymbol{\chi} + \epsilon_{ij} \tag{2}$$

where $\text{link}^1_{ij} = 100$ 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^0_{ij} 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-treat 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 4). The probability of a link between a pair of control students is 8.43 percent. Treat-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 4), we see an increase in both treat-to-control and control-to-treat information links, but the latter effect is larger. That is, treated students are particularly likely to be *named by others* as information contacts.[14] 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. 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. Taken together, these findings also 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. 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, changes to the information network appear to be driven by the *extensive margin*, as opposed to the strengthening of pre-existing connections.

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

We see limited variation in link strength based on the treatment status of students (Panels C and D of Figure A3), and we see that changes to the network are due to both created and maintained links (Panel E of Figure A3). In Table A2, 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.[15]

Second-order effects. In Appendix Table A3 we assess whether link formation depends on number of *other* treated links at baseline. For this purpose, we estimate the directed link specification (Equation 2). We interact the treatment status indicators with the number of *other* treated friends of i (and j) at baseline, as well as with the overall number of *other* friends.[16] The treat-to-control and treat-to-treat are unchanged up to the second decimal. The control-to-treat appears slightly stronger in this version but within the confidence interval. We interpret this as evidence that potential Stable Unit Treatment Value Assumption (SUTVA) violations do not greatly affect the estimates in Table 4.

[14] This rules out the possibility that our results are driven by a particular measurement issue: it is not the case that treated students are simply naming more contacts than control students, for example, due to higher effort during the survey.

[15] Column 1 of Table A2 shows that effects are larger for links that were present at baseline (3.68 percentage points versus .463). 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 treat-control pairs, approximately 115 new links were created, compared to 89 additional maintained links.

[16] When calculating the number of friends for each node, we do not include links between i and j themselves.

The analysis reveals some additional (suggestive) patterns. First, we see no evidence that students preferentially link to others with more treated links. Second, control students who had more treated friends at baseline appear more likely to connect to another treated student. More specifically, the "Treat-to-control Link x Number of Treated Friends of *j* at Baseline" and "Control-to-treat Link x Number of Treated Friends of *i* at Baseline" are both positive with sizeable coefficients, and significant under classical inference. It is possible that students with a higher number of treated friends at baseline value the information source more, to an extent that outweighs the mechanical effect of having more treated friends. This analysis, however, is at best suggestive: the coefficients are *not* significant based on randomization inference *p*-values. We conclude that there is inconclusive evidence to support the importance of second-order link formation.

Learning about information technology. The fact that we observe an increase in both-treated links as well as an increase in treat-control links (see Column 1 of Table 4) suggests that link formation serves a purpose beyond pure access to the information source. 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 treat-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 treat-control links.

In Table 5, 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 treat-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. There is no increase in links among pairs of treated students both of whom have past internet experience. This is consistent with treatment students having a strong motivation to learn how to use 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 6, 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 treat-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-treat links than for treat-to-control links; treated students are frequently named by others as information contacts across a range of topics (Panel B of Table 6).

Turning our attention to the personal network, we find null results for many subcomponents, with some notable exceptions (Table 7). 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 7). 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.

4.2 The Effect of Information Access on Network Centrality

We next estimate the effect of information access on differences in individual-level network centrality measures with the following specification.

$$\text{centrality}_i^1 = \beta \cdot T_i + \alpha \cdot \text{centrality}_i^0 + \mathbf{x}_i'\boldsymbol{\chi} + \gamma_c + \lambda_s + \epsilon_i \tag{3}$$

Here, centrality_i^1 is either an endline centrality measure for student i, or an indicator for being in the top 5 percent of the centrality distribution. 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).[17] 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 again rely on randomization inference to construct p-values.

When the outcome is a centrality measure, the estimates we obtain must be interpreted as *relative* differences in centrality between treated and control students. This estimate is causal in the sense that we

[17] Covariates include Gender and SES, defined as household having electricity and running water, and above-median academic ability at baseline.

can attribute differences in centrality to the randomized intervention and not to unobservable differences or reverse causality. Randomization inference p-values can be used to test the sharp null hypothesis that the centrality of each node is exactly as it would have been, regardless of its treatment assignments of the other nodes in the network (Fredrickson and Chen, 2019). We discuss this test and the interpretation of our parameter estimates in greater detail in Appendix A.2.

Even *relative* differences in centrality within a network are relevant for many applications, and are particularly relevant for network-based targeting. Indeed, targeting policies typically seed the most central nodes in a network based on their relative positions as opposed to their absolute centrality scores (for example, in Banerjee et al. 2013, Banerjee et al. 2019, Baumgartner et al. 2022 and many others). More broadly, relative centrality measures can predict social status and one's sense of belonging (Alan et al., 2021).

Importantly, relative differences in centrality should not be interpreted as average treatment effects under the no-intervention counterfactual.[18] In a network, centrality measures are interdependent, and the Stable Unit Treatment Value Assumption is violated (Rubin, 1974). In fact, when centrality is an outcome, the treatment effect is not well defined even at the level of a particular node as it will depend on the treatment statuses of all other nodes. In particular, we must be careful to avoid making conclusions about *absolute* changes in centrality. For example, if we find $\hat{\beta}$ to be positive, that relative difference could indicate an increase in links for treated students, a decrease in links for control students, or some combination thereof. In Section 5.3 we will attempt to further characterize the average treatment effects on centrality measures using a structural model.

However, we *can* estimate average treatment effects for the probability of being in the tail of the centrality distribution. When we define our outcome to be an indicator for student i being in the top 5 percent of the centrality distribution, we can not only compare treated students to control students, but also to the counterfactual for the treated students themselves in an untreated network. Of course, in an untreated network, for a randomly selected student, the probability of being in the top 5 percent of the distribution is 5 percent.

Main results: effects on the information network. We find large and significant difference in network centrality between treated and control students in the information-sharing network (see Panel A of Table 8). Students randomly selected to gain information access have higher centrality than control students 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 relative difference. This difference is similar in magnitude to the effect of having high socio-economic status (1.33 additional links, see Table 3). Column 2 of Table 8 shows that eigenvector centrality is .18 standard deviations higher

[18] These considerations also apply to the dyadic regressions in Section 4.1, though the interpretation is simpler. We interpret our dyadic estimates as the likelihood of a treat-treat link relative to a control-control link, *not* relative to a link in a counterfactual network with no intervention.

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 differences are significant with randomization inference p-values less than or equal to .001, suggesting that the intervention had a causal impact on network centrality.

These centrality differences 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 rescale 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. Moreover, this difference appears to be evenly distributed across the baseline distribution. It is not, for example, the case that this change in centrality is 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 the centrality distribution for treated students is shifted to the right, relative to the distribution for control students. 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 8 suggest that treated students are better positioned for information diffusion: the number of length-2 walks is approximately 8.6 percent higher, and diffusion centrality is .18 units of standard deviation higher 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, we observe no significant difference in average link strength between treated and control students (Column 6 of Table 8 and Panel A of Figure A3). This outcome captures an intensive margin of the treatment; a positive effect would indicate an increase in interactions with preexisting information links. The changes we observe are in fact more consistent with information-seeking behavior outside of students' pre-existing information networks.

Beyond average differences in centrality measures, 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 8). 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 8 are subject to some imprecision, the magnitudes are large. For example, treated students have a 6.3 percent chance of being in the top 5 percent by eigenvector centrality, suggesting an average treatment effect of 1.3 percentage points. The difference relative to control students is estimated to be 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 differences are observed across other centrality measures. For the number of length-2 walks, diffusion and betweenness centrality, estimates 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 8, Panel C). These results are entirely consistent with the dyad-level effects presented in the subsection above: 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 D of Table 8). Degree increases by .82 in the full network, compared to .96 in the information network (see Column 1 in Table 8). Point estimates for the other four centrality measures are also similar. There is no impact on average link strength, which again indicates that treated students are becoming more central by forming new links as opposed to simply strengthening existing links (for example, by sharing information with personal links, or by sharing new types of information with existing information links).

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.

We do not find that treatment students simply spend time with a higher number of contacts. In Table A4 we investigate changes to 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 difference in 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.[19] 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 changes in network centrality 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.[20] 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 A5 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

[19] 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.

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

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 A5, Panels B to E). This latter finding is consistent with Panels A and C of Figure 5, in which centrality differences between treated and control students appear broadly homogeneous by baseline degree. This apparent lack of heterogeneity 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 A6 show differences in link creation and destruction between treated and control students. We find that students with access to information both create more new links, and destroy fewer existing links than control students. New link creation appears 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 estimates are smaller, they 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 a difference in 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 is higher among treated students. Across this table, differences 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 A7 we remove all controls except for stratification bin indicators. The estimates remain broadly significant and of similar magnitude, although with lower precision. In Table A8 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 8, and remain broadly significant.

5 Model, Calibration and Simulations

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. In particular, we are not able to say whether these effects are large enough to offset the benefits of network-based target-

ing. 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:

$$u_{ij} = \kappa_{TC} T_i C_j + \kappa_{CT} C_i T_j + \kappa_{TT} T_i T_j + \nu_{ij}, \tag{4}$$

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 serving as an information source for control students and sharing information, and zero otherwise. κ_{CT} is the benefit to a control node of linking to a treatment node. A control students may value linking to a treated student both because this allows them to actively search for information, and because they become the passive recipient of information that treated students decide to share. 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.[21]

We have assumed that link utility depends only on direct access to information, as opposed to indirect access via second-order links. Second-order link utility is important if nodes can access information indirectly by linking to others who are themselves linked to nodes in the treated group. It is difficult to say how allowing for second-order link utility would impact the network, as these types of models do not in general give rise to pairwise-stable equilibria (Jackson, 2010), and simulations would therefore require a more complicated setup and different assumptions. Such complications may not be worthwhile, as we note that, in our setting, most students appear to benefit from direct link formation only, as opposed to

[21] Recall that treated students could not speak to one other in the digital library owing to the design of the intervention. κ_{TT} is therefore unlikely to capture an increase in opportunities to socialize.

gaining utility from second-order link formation (see Table A3). Moreover, our information experiment revealed that 97% of control students who asked for information asked a treated peer directly (see Section 2.2). Finally, we do not see a significant change in the probability of a link between a pair of control students at baseline (0.087) versus at endline (0.084), suggesting that the network response was mostly limited to links that involve treated students directly.

The term ν_{ij} captures a sort of underlying benefit (or, if negative, cost) 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 net 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, but not necessarily having mean zero. In our setup, the utility of a link does not depend on the wider network, nor on the other links of the nodes in question. It only depends on the independently-distributed term ν_{ij} and on the treatment status of each of the two nodes. In particular, this means that nodes consider forming new links independently of their existing links. Again, this simplifies the model and guarantees the existence of a pairwise-stable equilibrium. However, it cannot capture considerations such as *conversation capacity*: the idea that students may simply not be able to maintain a very large number of friendships. Intuitively, if students must choose between friends, we would expect to see the probability of a link between a pair control students to decline over time. Again, because this probability is similar at baseline and endline, and because personal friendship networks were unaffected by the intervention, we believer our independence assumption to be reasonable.

The inclusion of ν_{ij} in equation 4 implies that the utility of a link does depend on underlying characteristics of each node. Yet, we assume that when a link is treated, the marginal increase in link utility does not depend on node characteristics. That is, the information itself is equally valuable, regardless of who it comes from. This assumption may be justified if nodes believe, or learn over time, that the information source (for example, Wikipedia) is reliable, and that nodes (students) are generally capable of transmitting the information across links, regardless of their network position. The assumption is also consistent with our empirical findings: we do not find significant heterogeneous treatment effects by baseline characteristics or baseline network position. However, one might imagine a setting or intervention in which information is only *believed* if obtained from a central student. In such a setting, each term in equation 4 should be interacted with student baseline network characteristics. Such a model could lead to very different conclusions with respect to network response. If initially-central students are more likely to be believed, the intervention may lead to a concentration of the network around these students, rather than a decrease in their relative centrality over time. Our empirics suggest that this type of network concentration is unlikely when the information source is reliable and the network response takes place over an extended period of time.

We allow nodes to form links by mutual consent, and to sever links unilaterally. This results in a

unique pairwise-stable network.[22] 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. In this case, no link will exist.[23] 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,

$$\mathbb{P}(g_{ij} = 1) = \begin{cases} P_{CC} \equiv \mathbb{P}(\nu > 0)^2, & \text{if } T_i = T_j = 0 \\ P_{TT} \equiv \mathbb{P}(\nu > -\kappa_{TT})^2, & \text{if } T_i = T_j = 1 \\ P_{TC} \equiv \mathbb{P}(\nu > -\kappa_{TC})\mathbb{P}(\nu > -\kappa_{CT}) & \text{if } T_i \neq T_j. \end{cases} \tag{5}$$

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 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.3, 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

[22] 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.

[23] Note that the concept of mutual consent is different from the concept of a reciprocated link in the data. If one student borrows money from another, this is captured empirically as an unreciprocated link. Yet, both students consented to the interaction. We therefore include this type of link in the undirected network, both empirically and theoretically.

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 degree, eigenvector centrality and diffusion centrality (again with parameters as in Banerjee et al. (2019)) for simulated networks based on this model. 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 an increase in both eigenvector and diffusion 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 links to persist 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 ability, 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:

$$u_{ij}^0 = \sum_{\theta_1,\theta_2 \in \{L,H\}} \mathbb{1}\{\theta_i = \theta_1, \theta_j = \theta_2\} \kappa_{CC}^{\theta_1\theta_2} + \nu_{ij}^0. \tag{6}$$

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.[24] We model the endline link-utility as

$$u_{ij}^1 = \sum_{\theta_1,\theta_2 \in \{L,H\}} \mathbb{1}\{\theta_i = \theta_1, \theta_j = \theta_2\} (\kappa_{CC}^{\theta_1\theta_2} C_i C_j + \kappa_{TC}^{\theta_1\theta_2} T_i C_j + \kappa_{CT}^{\theta_1\theta_2} C_i T_j + \kappa_{TT}^{\theta_1\theta_2} T_j T_j) + \nu_{ij}^1. \tag{7}$$

We are allowing the value of information to be different for high and lower-ability students, and for the value of information from a high-ability source to differ from the value of information from a lower-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.

This extended model is both general and tractable. It is flexible enough to allow for any degree of link persistence between baseline and endline (as captured by δ). It also allows for baseline centrality patterns to be weakened or amplified over time. For example, if the marginal utility of linking to a treated node, as captured by κ_{CT}, κ_{TC} and κ_{TT}, is much higher for high-ability treated nodes, this will amplify the relationship between ability and centrality, resulting in an even more concentrated network at endline. We could have modeled this potential for amplification in a different way, for example by using a preferential attachment model (Barabási and Albert, 1999) or otherwise allowing link probabilities to depend directly on network positions. However, this would complicate the model considerably, in the sense that it could no longer be represented by a general random graph, and may not provide a better fit. Indeed, our empirics suggest a degree of mean reversion, as opposed to amplification, of centrality over time (see Figure 5, Panels A, C and E), and limited heterogeneous effects by baseline centrality (see Appendix Table A5, Panel E).

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, and calibrated parameters, are detailed in Appendix A.4. Our calibrated parameters suggest a moderate level of persistence between baseline and endline networks. Baseline links persist to endline 35 to 48 percent of the time (depending on ability types). The marginal utility of linking to a treated student is positive across all ability-type combinations, and is highest between high-ability pairs.

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

5.3 Model Fit and Average Treatment Effects

We next assess the quality of fit between our model and our empirical findings. We simulate networks with 117 nodes and 23 treated nodes, chosen to match the average network size and treatment intensity in our data. We then compare simulated moments to their empirical counterparts, not including the moments we used for calibration. Table 9 contains moments from 10,000 simulated networks and matched summary statistics from our data.

The model appears to capture most moments 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 closely predicts the probability that a treatment student will appear among the top 5 percent of students according to all network centrality measures. In both the model and the empirics, this change in centrality appears to be driven by an increase in centrality over time for treated students. For most measures, there is very little difference between the baseline average and the endline control group average. 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, which might be therefore particularly relevant for counterfactual experiments involving network-based targeting and information diffusion.

We can also use these simulations to shed light on the likely magnitude of average treatment effects on centrality outcomes, as we are able to simulate not only treated networks but also counterfactual untreated networks. In Panel C of Table 9, we compare the simulated average treatment effects to our reduced-form estimates. Strictly speaking, our reduced-form estimates should be interpreted as relative differences between treated and control students and not average treatment effects. Yet, the simulations suggest that the average treatment effects and the reduced-form relative differences in centrality are remarkably similar. The simulated average treatment effects are slightly larger for most measures. For example, using the model simulation, we estimate that treated students have approximately 1.02 more links overall than they would have had under no intervention (compared to the reduced-form estimate of 0.964 taken from Panel A of Table 8).

5.4 Counterfactual Policy Simulations

We next simulate sets of networks to conduct counterfactual policy experiments. We compare network-based targeting strategies to random targeting, and we vary the diffusion model and its parameters, as well as the measure of diffusion. We plot the estimates against a benchmark that assumes stable networks over time.

Diffusion models. First, we will examine a variant of the standard Susceptible-Infected-Recovered (SIR) diffusion model. In this model, each treated node shares information with each of its neighbors, independently, with probability q. This process continues for T periods, with all treated and previously informed nodes sharing information in each time period. and then diffusion stops.

Under the SIR model, we will measure the extent of information diffusion in two different ways: the number of nodes in the network that are ever informed, and *total diffusion*, which is equal to the expected total number of times that information is heard. While the number of ever-informed nodes is a natural definition of information diffusion, total diffusion may also be relevant in our setting, where many different pieces of information are shared, and where credence may increase in the number of times a piece of information is heard. Total diffusion can exceed the size of the network, but this does not necessarily imply that everyone has been informed. 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, and diffusion-centrality targeting maximizes expected total diffusion (Banerjee et al., 2019).

Second, following Beaman et al. (2021), we will consider a model in which a node becomes informed when at least two of its neighbors are informed. The process is again repeated for T periods. This is a "threshold model" of diffusion, with threshold $\lambda = 2$ (Granovetter, 1978).

Targeting strategies. One common form of network-based targeting involves targeting central nodes. We will target nodes based on diffusion centrality, with parameters matching the parameters of the model. This targeting strategy maximizes expected total diffusion under the SIR model Banerjee et al. (2019). It does not necessarily maximize the expected number of ever-informed nodes nor the number of informed nodes under the threshold model (Jackson and Storms, 2023). Diffusion centrality also has the advantage of being relatively easy to measure in the field. In the case where $T = 1$, it is equivalent to degree, and Banerjee et al. (2019) show that members of a community may be able to quickly identify diffusion-central nodes. In the appendix, we will show the results of simulations using degree and eigenvector centrality for targeting.

In general, choosing an optimal targeting strategy based on a particular network's structure is computationally intractable (Kempe et al., 2003). Yet, for a very small number of seeds, it is computationally feasible to identify the precise nodes that would maximize information diffusion. This requires a deterministic model of diffusion, such as the threshold model or the SIR model with $q = 1$, and a stable network. In their study of the diffusion of agricultural technology, Beaman et al. (2021) use network data to identify two top seeds based on a threshold model of diffusion. We will use our baseline network data to perform a similar computation, identifying the top two seeds under both the threshold model and the SIR model with $q = 1$.

Simulations. For each model, and each set of parameters, we simulate 1000 networks with 100 nodes. Under network-based targeting, we assign the top N_T nodes to treatment based on their positions in the baseline network.

In Figures 8, 9 and 10 we plot the extent of information diffusion as simulated under different models and measures of diffusion. In each panel of each figure, we plot information diffusion under centrality-based targeting (in red) versus random targeting (in blue).[25] 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 (solid lines). Second, we plot information diffusion in a hypothetical setting where networks remain stable over time (dashed lines). Third, we compare these plots to a hypothetical setting where networks change exogenously over time, but not endogenously in response to the treatment (dotted line). Note that we do not plot this third hypothetical for random targeting, as exogenous link changes do not affect information diffusion under this policy.

In simulations that imitate the conditions of own experimental setting, we find that the gains from centrality-based targeting are cut in half due to the face that the network changes over time. In Panels A and B of Figure 8, we set the number of seeds to 20, and assume a single period of transmission $T = 1$. These parameters were chosen to match our setting, where students appear to obtain information primarily from treated students directly, as opposed to through longer chains of diffusion (see Sections 2.2 and 5.1). For the SIR model, across the full range of $q \in (0, 1)$, we find that the gains from centrality-based targeting are reduced by 44-56 percent in terms of the number of nodes ever informed, and by 48-49 percent in terms of total diffusion. For the threshold model (with $\lambda = 2$), the gains from centrality-based targeting are reduced by 53 percent. In Panels C and D, we simulate a model with more time periods ($T = 4$) and only 2 seeds. Here, we see that when transmission rates are low, it is important to target central nodes to reach as many other nodes as possible (Panel C). As q increases, this becomes less important, and random seeding catches up. Again, even for low values of q, the gains from centrality-based targeting are limited due to the changing network.

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 both targeting strategies, but not necessarily by the same amount. Indeed, we see that under centrality-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.

In Figure 9 and in Figure 10, we also plot information diffusion by number of seeds, under the SIR

[25] Using a different centrality measure for targeting does not perceptibly alter the simulations (see Appendix Figures A5, A6, A7, and A8).

model and threshold model respectively. For the SIR model, we use two different sets of paramaters. In Panels A and B of Figure 9, we simply set $q = 1$ and $T = 1$, that is, each treated node shares information with all of their neighbours, and then diffusion stops. In Panels E and F, we set $q^* = 0.1$ and $T^* = 4$, to equal the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019). Panels A and C plot the total number of nodes ever informed, while Panels B and D plot total diffusion, that is, the number of times the information is heard. For the threshold model (Figure 10), we use $\lambda = 2$ as in Beaman et al. (2021), and two different values for the number of periods: $T = 1$ and $T = 4$ (as in Beaman et al. 2021).

One way of measuring the gains from network-based targeting is to count the number of random seeds that one would need to add in order to achieve the same level of diffusion. In Figures 9 and 10, the number of random seeds one would need to add to match centrality-based targeting can be measured by comparing the horizontal distance between the solid red and blue lines (or, under a stable network, the dashed red and blue lines).

One initial takeaway from Figures 9 and 10 is that, in general, the gains from centrality-based targeting are limited, even on stable networks. Consider the SIR model with $q = T = 1$. Assuming a stable network with ten seeds, the number of random seeds one would need to add to match network-based targeting is only 6. This is consistent with theoretical work by Akbarpour et al. (2021), who show that for SIR models, you typically only need to add a few random seeds in order to achieve the same level of diffusion as under centrality-based targeting. For a large number of seeds, random targeting outperforms centrality-based targeting, even on a stable network, as central nodes may inform overlapping sets of well-connected neighbours. (Panel E of 9).

The gains from centrality-based targeting are reduced by more than half due to the network change. Taking the network change into account, the number of additional random seeds that would be needed to match the level of total diffusion under centrality-based targeting is cut in half: at most 4 seeds under the SIR model (or 5 for total diffusion).

It is worth noting that while targeting nodes based on standard centrality measures is often feasible (Banerjee et al., 2019), and can be optimal for total diffusion under the SIR model, it may be far from optimal for the threshold model (Jackson and Storms, 2023). Indeed, for information to spread under the threshold model, it is important to choose seeds with neighbours in common. While this may be more likely for central nodes, it is possible to choose even better seeds based on precise network data.

In Figure 11, we present results from simulations in which we choose only two seeds, but the two seeds are chosen precisely for maximum diffusion, based on a stable baseline network and a deterministic diffusion process. In the case of total diffusion under the SIR model, this is equivalent to choosing seeds based on diffusion centrality, as in Figure 9. However, it is theoretically possible to choose even better seeds to maximize the number of ever-informed nodes, or to maximize diffusion under the threshold model, as in Beaman et al. (2021).

Consistent with Akbarpour et al. (2021) and Jackson and Storms (2023), we find that under a threshold model, precise targeting vastly outperforms other targeting strategies on a stable network (Figure 11). Precise targeting outperforms centrality-based targeting, and is 3 to 7 times more effective than random targeting.

This advantage shrinks or disappears once the network response is taken into account (Figure 11). Precise targeting relies on specific links remaining intact, and is very sensitive to changes in the network. The gains from precise targeting, relative to random targeting, are reduced by one-half to two-thirds across the range of $T \in \{1, 2, 3, 4\}$. Centrality-based targeting appears more robust, and in fact leads to higher information diffusion across both models. A big part of the reduction in gains is likely due to the fact that links change *exogenously*. That is, two nodes that initially share a neighbor may lose that shared neighbor due simply to the fact that links form and break over time. Indeed, this may explain why Beaman et al. (2021) find that even when using baseline network data to target optimal seeds, the true diffusion of technology over the long run is lower than predicted by their threshold model simulations.

Despite the changing network, targeting based on baseline centrality remains an effective strategy for information diffusion. If one could predict the endline network perfectly, targeting nodes with high endline diffusion centrality would maximize total information diffusion. Our model is not deterministic and therefore does not allow us to make such a prediction, as links form at random. In our data, baseline centrality is a strong predictor of endline centrality, and is a better predictor than other baseline covariates. Appendix Table A9 shows that baseline diffusion centrality is a stronger predictor of endline diffusion centrality than baseline academic ability, SES or gender. Moreover, targeting nodes with high centrality at baseline appears to outperform both random targeting and more precise targeting strategies, and also outperforms targeting by ability (see Appendix Figure A9). In practice, there is a tradeoff between targeting central nodes and adding seeds, and this tradeoff is made sharper by the fact that networks respond to intervention.

5.5 Information Diffusion and Academic Performance

Previous work analyzing data from the same experiment has shown that Wikipedia access had a direct impact on students' English and Biology scores (Derksen et al., 2022). These two particular exam scores are pre-registered as primary outcomes, as almost all students complete the exams. English is compulsory for graduation, and Biology is the most popular subject, as it is required for entry into the most popular post-secondary programs including nursing and other medical programs. Students in the treated group had significantly higher scores than those in the control group, with effects concentrated among students with below-median scores at baseline (Appendix Table A10). For this subgroup, treated students scored 0.2 standard deviations higher in English, and 0.14 standard deviations higher in Biology, compared to below-median students in the control group. There was no significant effect for students with above-above median scores at baseline, for whom point estimates are zero or negative.

Reduced-form analysis cannot, however, identify the total effect of the intervention nor the extent of spillovers. First, we do not have any pure control schools, so we cannot directly estimate effects relative to a counterfactual in which no student was treated. Second, attempts to estimate spillovers using baseline network data will be biased due to the endogenous network response, and will likely understate their importance.

In this section, we calibrate our model to explore how the intervention affected students both directly and indirectly due to information diffusion, and to characterize the total effect of the intervention. In our setting, access to information may affect a student's academic performance and potentially even later life outcomes. As information diffuses through the network, academic outcomes for non-treated students might also be affected, and the effect on treated students might be amplified.

We model a student's academic score y_i as follows.

$$y_i = s_i + \tau(a_i)T_i + \tau(a_i) \sum_{j:g^1_{ij}=1} T_j Q_{ij} \tag{8}$$

Here, s_i is the student's counterfactual score, if the intervention had not taken place in their school. T_i represents treatment status. τ is the effect of becoming informed, which varies by the student's ability $a_i \in [0,1]$. Q_{ij} are independent Bernoulli random variables with identical probability parameters q, each indicating information transmission between a particular pair of nodes.

In this model, academic scores depend on total information diffusion under an SIR process with one time period, and probability of information transmission q.[26] Here, by measuring total diffusion, we assume that students receive a direct effect τ each time they become informed. This measure is particularly relevant to our setting, where students receive different information from different peers. This also captures the fact that even treated nodes may benefit from links to other treated nodes, in line with our empirical findings.

We can again calibrate this model by matching moments to empirical moments in our data (see Appendix A.4 for details). To match moments, we will make use of final exam scores in both the year prior to and the year of the intervention. We did not collect form 4 exam scores for the year before the intervention, as this cohort did not include our study participants. We therefore restrict our dataset to include only form 2 and 3 exam scores. We again define y_i to be the average of the student's English and Biology scores. We normalize with respect to the distribution of grades from the previous year's cohort, within the same school and form, by subtracting the mean and dividing by the standard deviation.[27]

Our measure of student ability is, as above, their percentile in the year prior to the intervention. Then, we can map ability a_i to ability type θ_i as follows.

[26] Again, we assume $T = 1$ not only for simplicity, but also because control students report obtaining information primarily from treated students directly, and do not appear to form new links amongst themselves (see Sections 2.2 and 5.1).

[27] This normalization differs from that in Derksen et al. (2022), because for the purposes of this exercise we need to capture changes in both the control and treatment arms, relative to a fixed benchmark.

$$\theta_i = \begin{cases} \theta_H & \text{if } a_i >= 0.9, \\ \theta_L & \text{if } a_i < 0.9. \end{cases}$$

Abusing notation to write $\tau(\theta_i) = \mathbb{E}(\tau(a_i)|\theta_i)$, we obtain the following parameter estimates.

$$q = 0.67 \tag{9}$$

$$\tau(\theta_H) = 0.05 \tag{10}$$

$$\tau(\theta_L) = 0.11 \tag{11}$$

This exercise sheds some light on the likely magnitudes of the direct effects, spillovers, and total effect of the intervention. Taking a weighted average of $\tau(\theta_H)$ and $\tau(\theta_L)$, we obtain an average direct treatment effect of $\tau = 0.10$. These direct effects are not very different from the estimates we obtain by simply comparing treatment and control students at endline, as in equation 19 (Appendix A.4). However, this exercise does suggest that the total effect of the intervention, including spillovers, is larger. Indeed, equation 21 (Appendix A.4) indicates that control group students' end-of-year scores are 0.26 standard deviations higher than those of the previous cohort. The total effect of the intervention, based on the average score overall in the year of the intervention, could be as high as 0.29 standard deviations. This conclusion rests on the assumption that the previous year's cohort, in the same form, is a good comparison group for the cohort that received the intervention. Reassuringly, there does not appear to be significant grade inflation between the two cohorts: the average grade in Chichewa, a subject that should not be impacted by Wikipedia, is unchanged.

5.6 Inequality and Academic Welfare

Finally, we can use the estimates from Section 5.5 to the explore implications of network-based targeting for inequality and academic welfare. In our setting, targeting central nodes could affect academic welfare through two direct channels. First, central students are more likely to be high-ability students. By targeting them, we would provide a direct benefit to students who are already advantaged. Even the indirect benefits may accrue primarily to high-ability students, due to network homophily. Second, the intervention appears to have a larger average treatment effect on lower-ability students, as shown in the previous subsection and in (Derksen et al., 2022). By targeting high-ability students, we therefore would expect smaller direct effects overall. The fact that network-based targeting increases information diffusion has the potential to offset both of these considerations, as lower-ability students who are not treated nevertheless benefit indirectly.

In general, network-based targeting carries implications for inequality, as centrality is correlated with

other forms of privilege in many settings (Jackson, 2019). In our model, this privilege applies to the top decile of nodes according to ability-type θ, and we allow link formation to depend on type. In our setting, θ corresponds to academic ability, but θ could theoretically represent any source of privilege that is correlated with network centrality. In our data, when targeting students at random, we expect approximately 10 percent of targets to belong to the privileged group. This is much higher when targeting by baseline centrality. For example, if we select the top 10 percent of nodes by degree centrality, approximately half will be in the high-ability group (see Appendix Figure A10).

To explore the implications of this imbalance in targeting for academic welfare, we simulate expected total treatment effects at the node level. Here, we define the *total treatment effect* for student i to be the effect on their exam score relative to (and normalized with respect to) a counterfactual in which the intervention did not take place. Importantly, control students may have non-zero total treatment effects.

$$\begin{aligned}\mathbb{E}(y_i - s_i) &= \tau(\theta_i)T_i + q\tau(\theta_i)\sum_{j:g^1_{ij}=1} T_j \\ &= \begin{cases} 0.05T_i + 0.67 * 0.05\sum_{j:g^1_{ij}=1} T_j & \text{if } \theta_i = \theta_H, \\ 0.11T_i + 0.67 * 0.11\sum_{j:g^1_{ij}=1} T_j & \text{if } \theta_i = \theta_L. \end{cases}\end{aligned}$$

In Figure 12, we plot total treatment effects under network-based targeting (in red) versus random targeting (in blue), again simulating 1000 100-node networks under each set of parameters. We target nodes with top diffusion centralities; because $T = 1$ this is equivalent to degree-centrality targeting. We again consider networks that change over time (solid lines), hypothetical networks that remain stable over time (dashed lines), and hypothetical networks that change exogenously but not endogenously in response to the treatment (dotted line).

Network-based targeting increases academic welfare, as captured by average total treatment effects, with any number of seeds, though the endogenous network response lessens this increase considerably (Panel A of Figure 12). The number of additional random seeds needed to match the effect achieved under network-based targeting is at most 5, compared to 12 under stable networks. The fact that network-based targeting continues to outperform random targeting despite larger direct treatment effects for lower-ability students is due to the fact that spillovers are large in our context. That is, many students benefit indirectly even if treated students are primarily high-ability. Our estimate of $q = 0.67$ relies heavily on the assumption that the previous year's cohort serves as a good counterfactual, yet, even with smaller values of q network-based targeting dominates (Panel B). Random targeting appears to only outperform network-based targeting in networks where information transmission is rare ($q \leq 0.15$).

However, the gains from network-based targeting are attributed disproportionately to high-ability students, and therefore lead to a relatively wider achievement gap (Panels C and D of Figure 12). With 20

seeds, the average total treatment effect for high-ability students jumps from 0.09 standard deviations under random targeting to 0.14 under network-based targeting. This jump is much smaller for lower-ability students, whose total treatment effects are large under both random (0.17 standard deviations) and network-based targeting (0.19 standard deviations). The intervention narrows the achievement gap under both targeting strategies, but this effect is much stronger under random targeting (Panels C and D, solid lines).

6 Conclusion

This paper demonstrates that providing exclusive, long-term access to a high quality information source can causally affect networks, as nodes form strategic links to informed peers to gain access. 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 difference in network centrality between treated and control students. After eight months, treated students were more likely than control students to be among the highest centrality students, according to many different measures. By calibrating models of network formation and information diffusion, we show that this has important implications for network-based targeting and academic performance.

By randomizing at the individual level, we are able to examine how node-level changes in information access causally affect strategic link formation and relative network positions. Yet, we cannot directly measure average treatment effects in absolute terms as we do not observe any pure control networks. This is a sample size limitation: with data from many more schools, we would have been able to compare students not only within but across clusters with varying treatment intensity. Instead, we must interpret our reduced-form estimates as relative differences, between treated and control students, caused by the intervention. To shed light on absolute treatment effects, we rely on estimates from a calibrated model.

The impact of information access on network structure is likely to vary based on the benefits and costs of interaction, for both treated and control students. The benefits of forming new links likely depend 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. Benefits to the treated students specifically might depend on whether information sharing is enjoyable, or whether they are able to gain status or other forms of favor by sharing information. The costs of interaction likely depend on whether a community is geographically and socially well-connected, and whether norms allow for communication with a diverse set of peers. We may expect the network response to 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. We may also expect smaller effects in settings where existing information-sharing networks are rigidly determined by norms. Yet, 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 suggest 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. We rely on this model to characterize the average treatment effect of the intervention on network centrality relative to a no-intervention counterfactual; we cannot estimate these average treatment effects directly without a pure control arm.

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*.

Akbarpour, M., Malladi, S., and Saberi, A. (2021). Just a few seeds more: The inflated value of network data for diffusion. *Working Paper*.

Alan, S., Kubilay, E., Bodur, E., and Mumcu, I. (2021). Social status in student networks and implications for perceived social climate in schools. *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.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

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*.

Blattman, C., Green, D. P., Ortega, D., and Tobón, S. (2021). Place-based interventions at scale: The direct and spillover effects of policing and city services on crime. *Journal of the European Economic Association*, 19(4):2022–2051.

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. (2019). Networks in economic development. *Oxford Review of Economic Policy*, 35(4):678–721.

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

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.

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*.

Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962.

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.

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.

Fredrickson, M. M. and Chen, Y. (2019). Permutation and randomization tests for network analysis. *Social Networks*, 59:171–183.

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.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443.

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.

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.

Islam, A., Vlassopoulos, M., Zenou, Y., and Zhang, X. (2021). Centrality-based spillover effects.

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

Jackson, M. O. (2019). *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 Storms, E. C. (2023). Behavioral communities and the atomic structure of networks. *Working Paper*.

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.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146.

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.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

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.

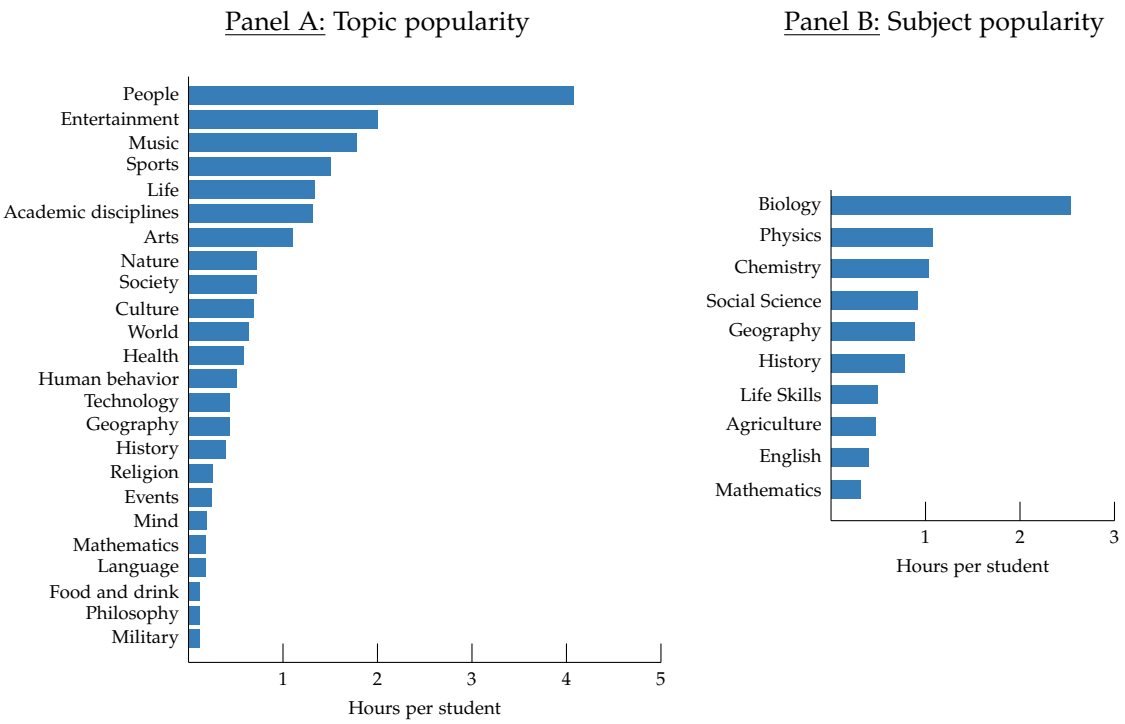
Söderberg, B. (2002). General formalism for inhomogeneous random graphs. *Physical Review E*, 66(6):066121.

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*.

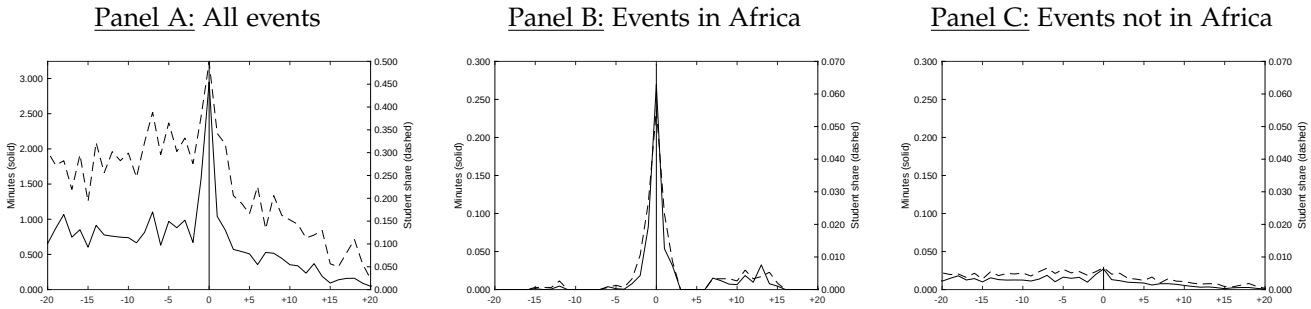
Figures and Tables

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



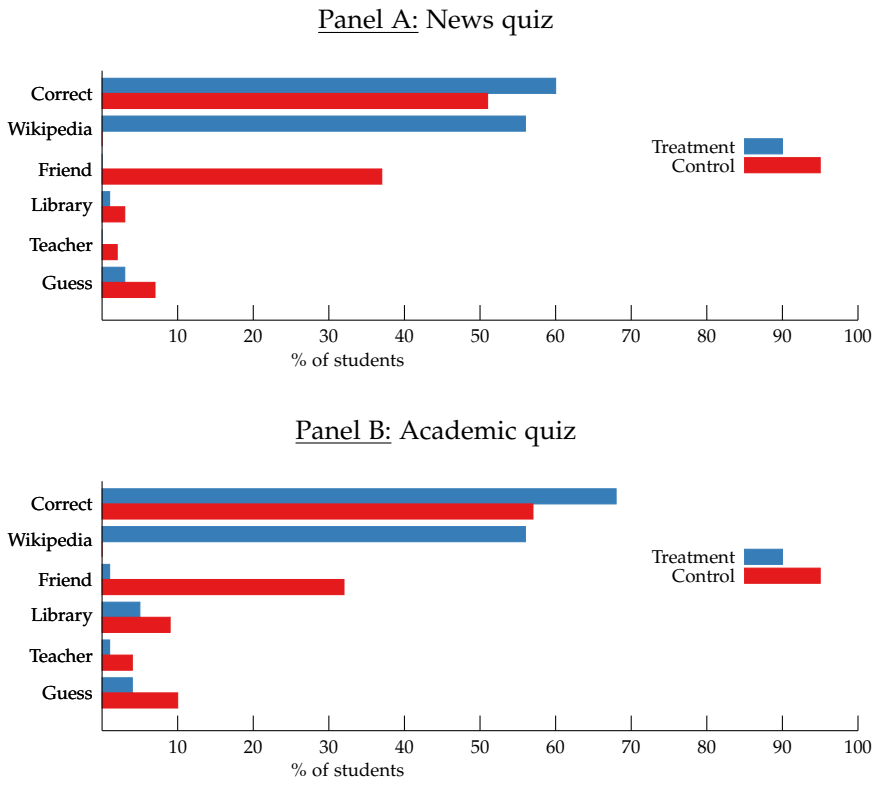
Notes: Figure reproduced from Derksen et al. (2022). Panel A: Browsing hours per topic, per student, aggregated over one academic year. The topics Business, Concepts, Crime, Economy, Education, Energy, Government, Humanities, Knowledge, Law, Objects, Organizations, Politics, Science, and Universe are 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



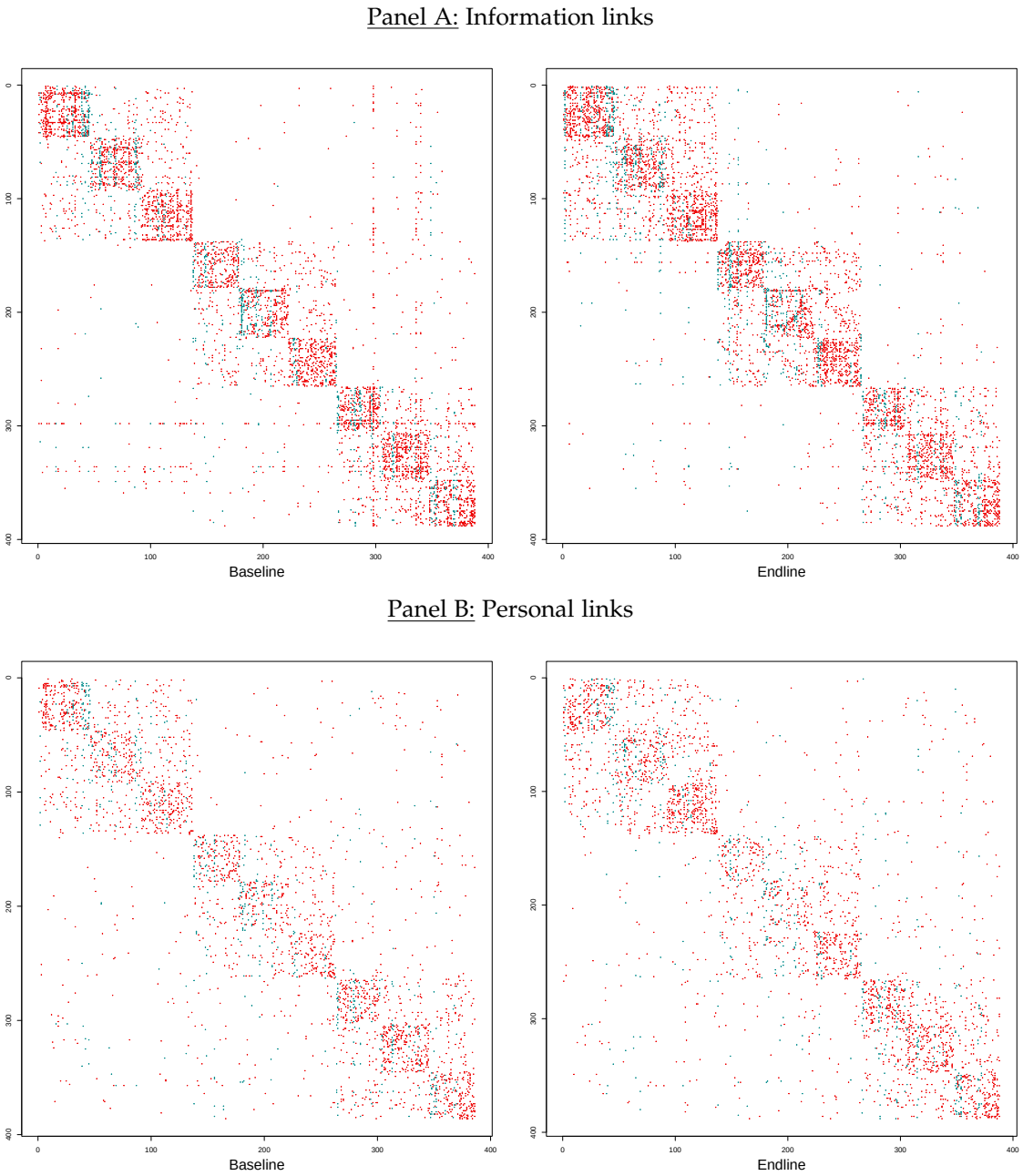
Notes: Figure reproduced from Derksen et al. (2022). Panel A: Left axis (solid line) shows total average browsing minutes per student on pages related to full set of worldwide events. Right axis (dashed line) shows share of students that visited pages associated to at least one event. Panels B and C: 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



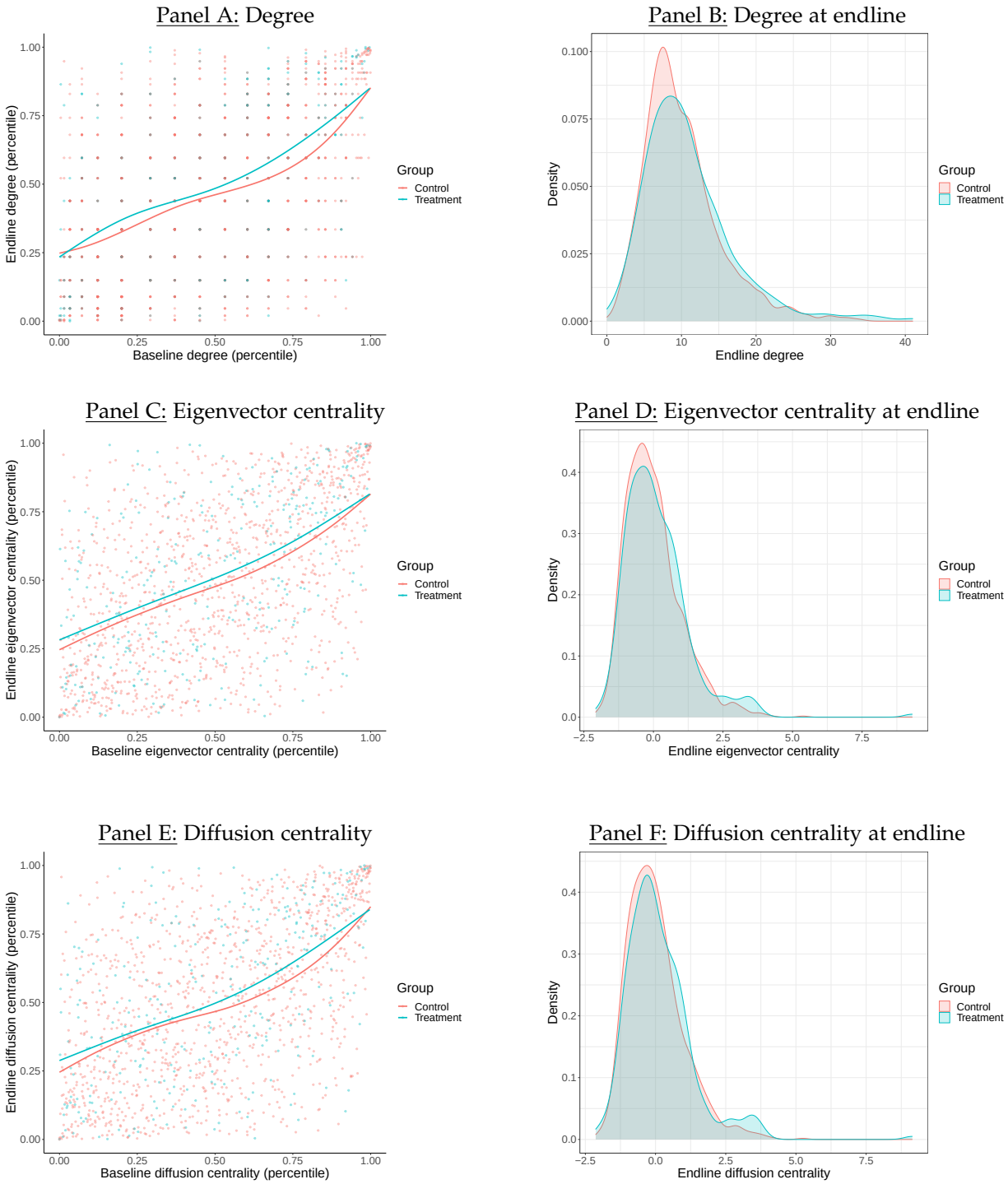
Notes: Percent of students that found correct answers to their unique quiz questions, overall and by information source. Among control students who received correct news quiz answers from friends, 97% were in the treatment group.

Figure 4: Networks at a Single School



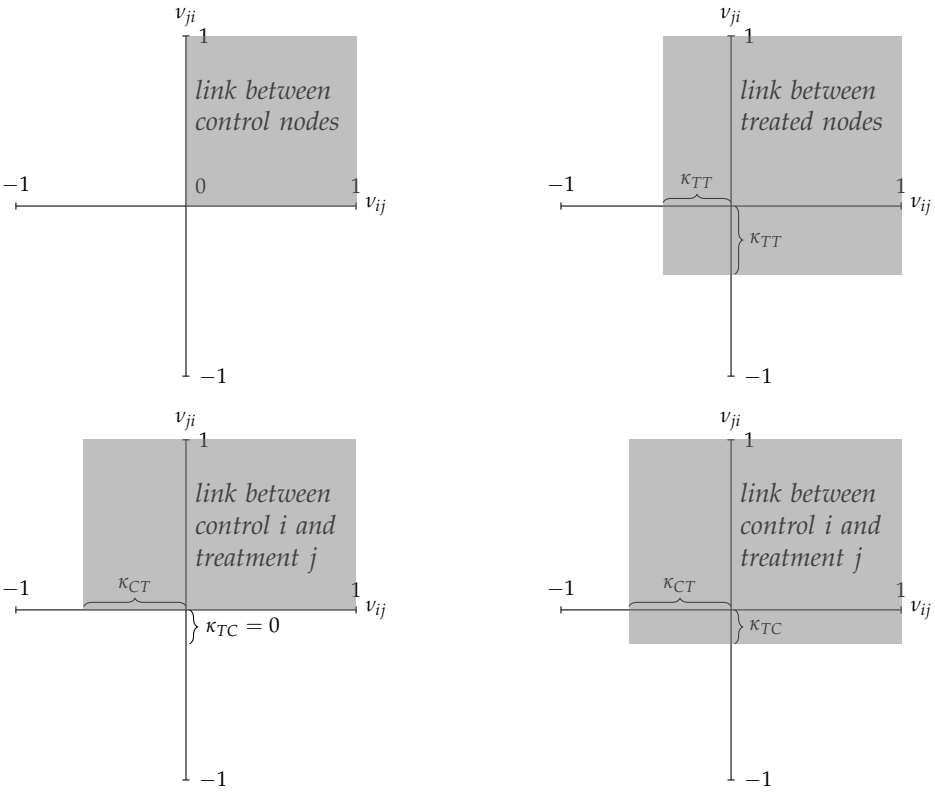
Notes: Baseline network adjacency matrices for a single school, including links across forms. Nodes are ordered by form and classroom. A dot represents an undirected link between nodes. On the horizontal axis, blue nodes are treated and red nodes are control. Panel A: Information links. Panel B: Personal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

Figure 5: Centrality in the Information Network



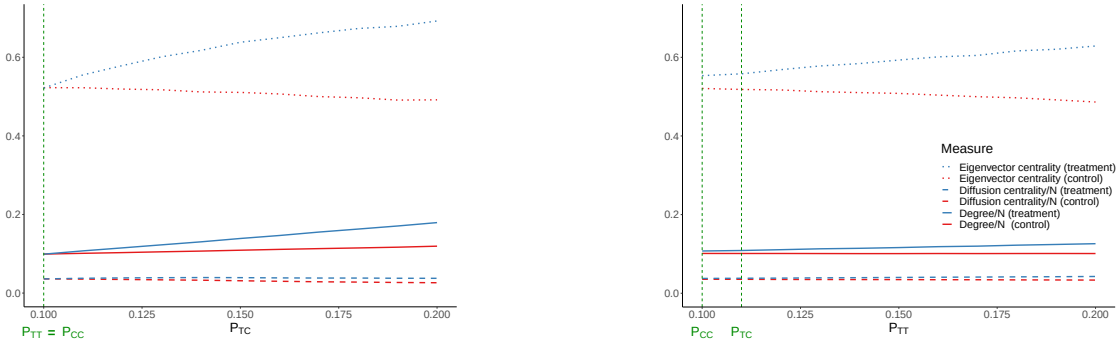
Notes: Centrality measure distributions in the information network for treatment and control groups.

Figure 6: The Model of Link Formation



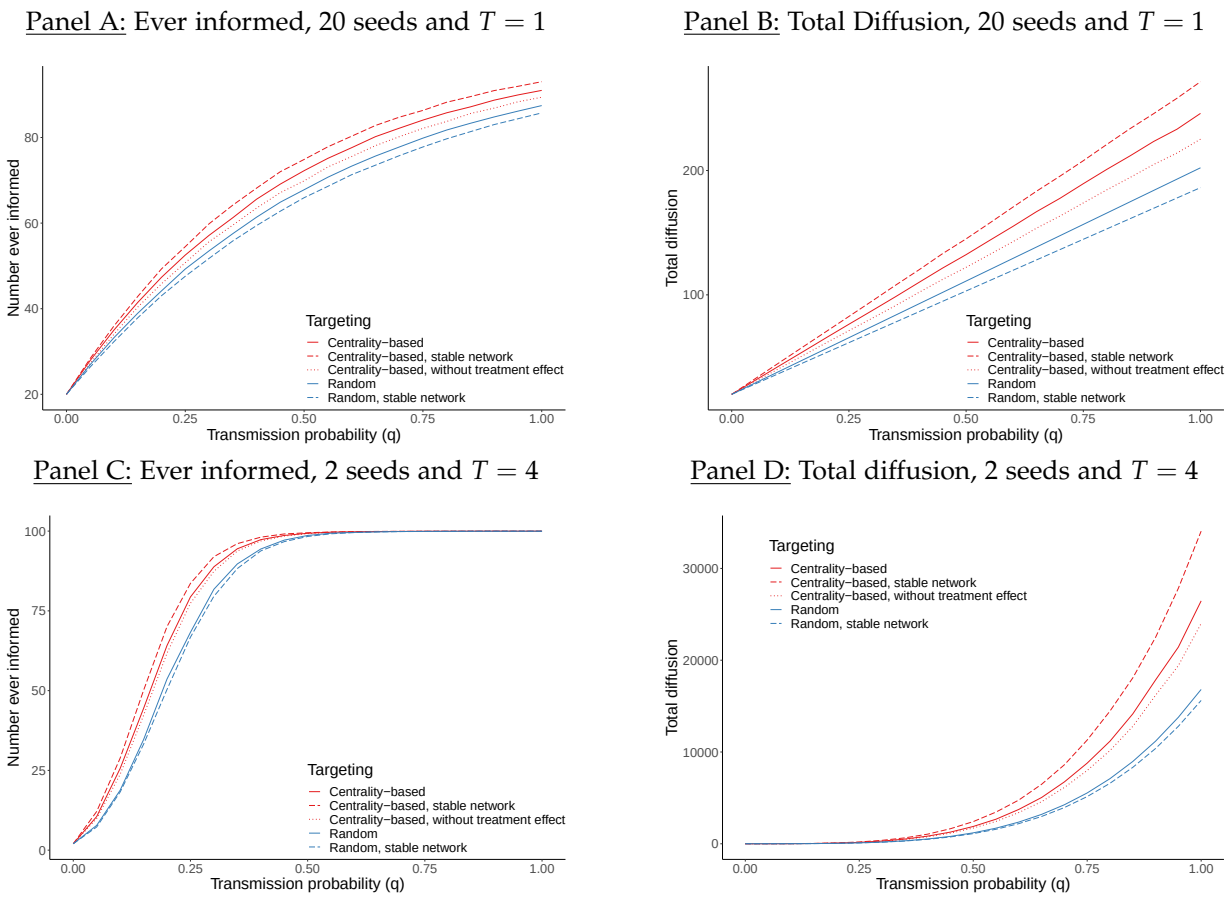
Notes: Pairwise-stable equilibria for link formation. The (relative) area of the shaded rectangle represents the probability of forming a link.

Figure 7: Model Simulations



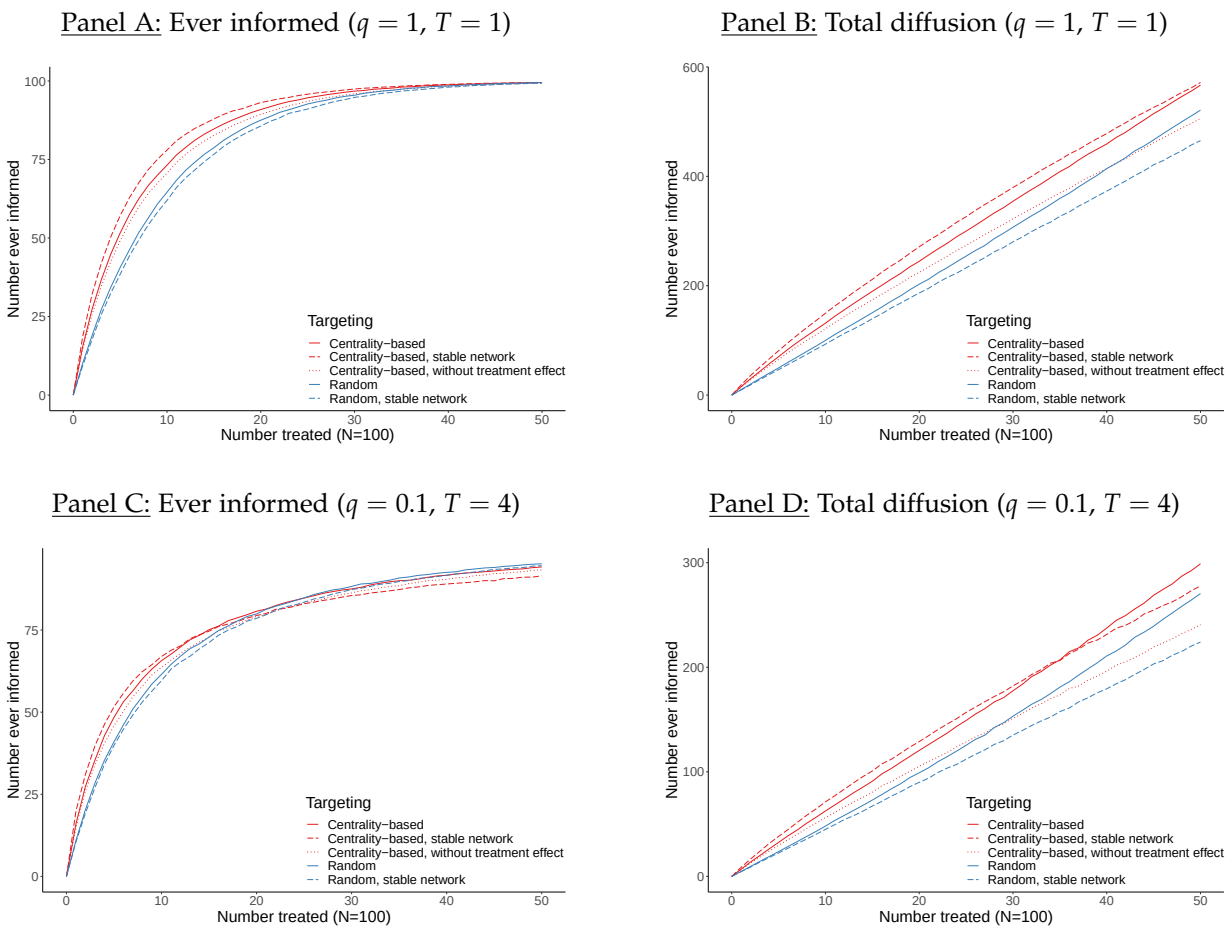
Notes: Simulated average centrality measures based on the model in Section 5.1. Diffusion centrality parameters are set equal to the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019). Left: P_{TC} varies while other parameters are fixed with $P_{CC} = P_{TT} = .10$. 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 $N = 100$ with 20 treated nodes.

Figure 8: SIR Model, Centrality-Based Versus Random Targeting



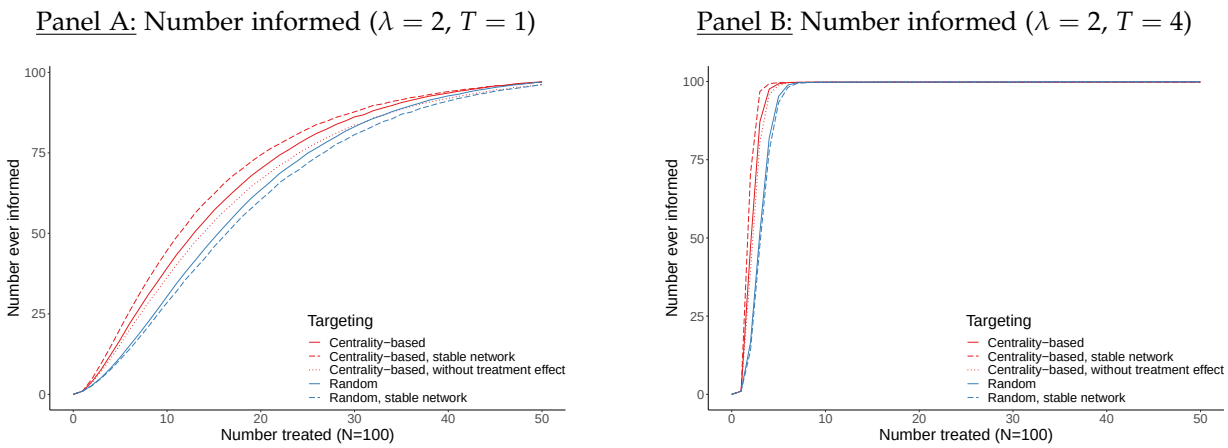
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by diffusion centrality, with parameters q and T matching the parameters of the SIR diffusion model.

Figure 9: SIR Model, Centrality-Based Versus Random Targeting



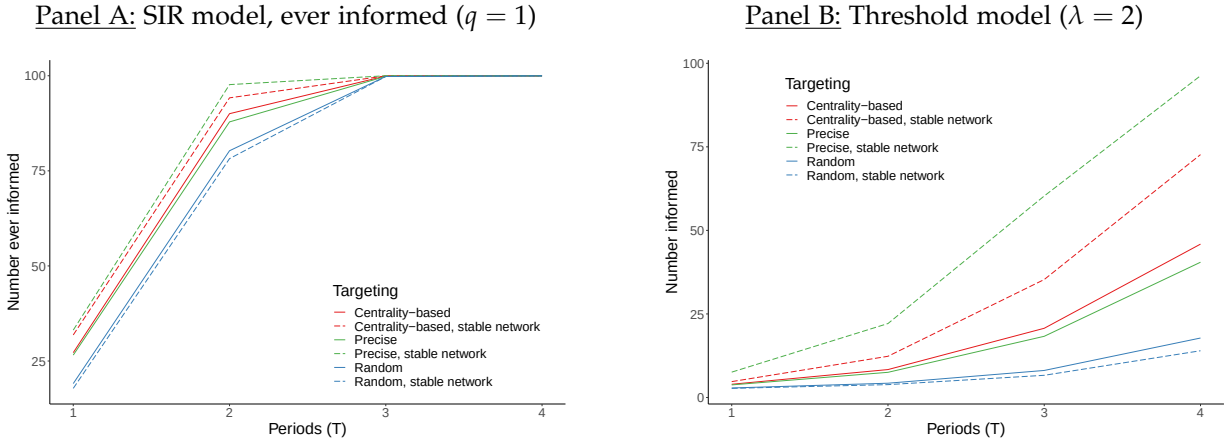
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by diffusion centrality, with parameters q and T matching the parameters of the SIR diffusion model. In Panels C and D, $q^* = 0.1$ and $T^* = 4$ are set to equal the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019).

Figure 10: Threshold Model, Centrality-Based Versus Random Targeting



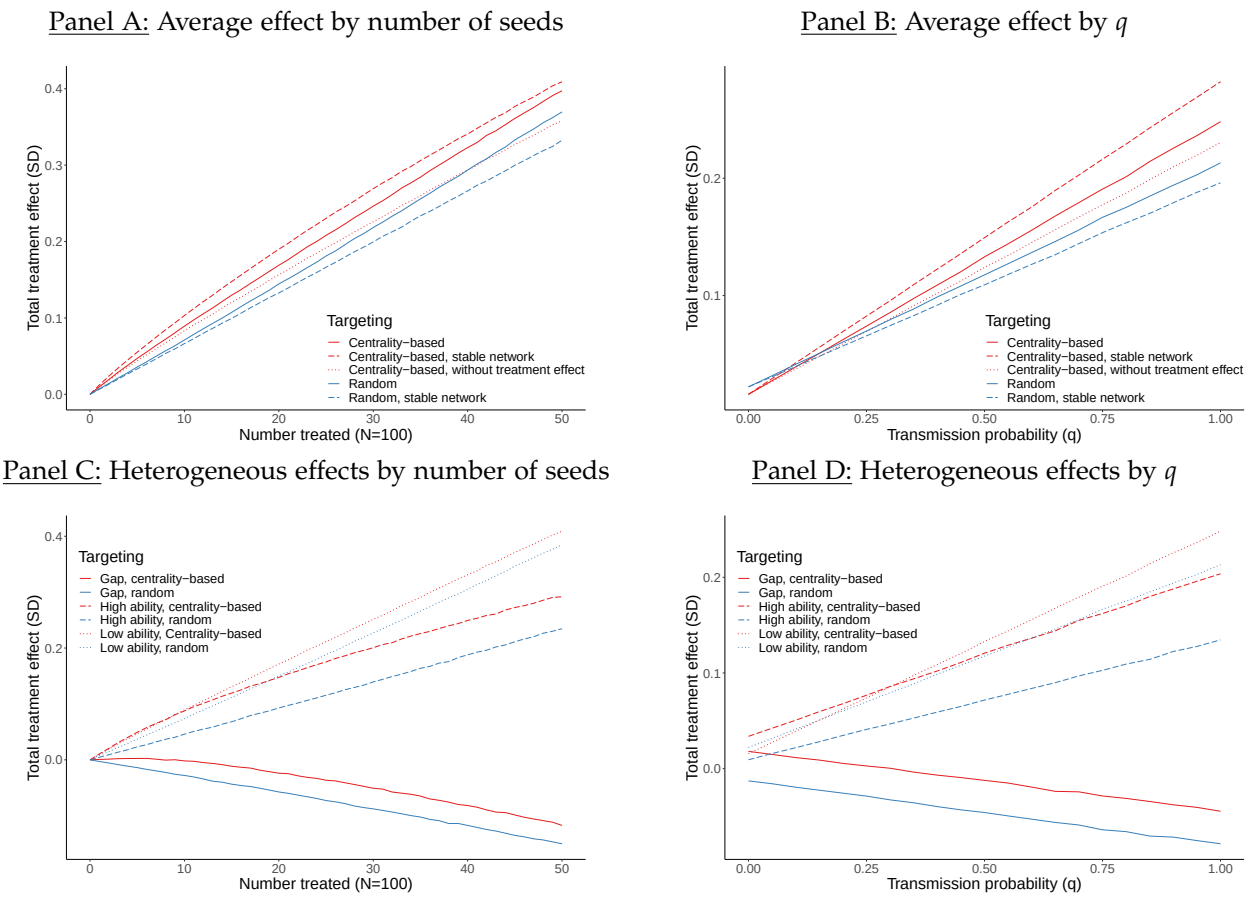
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by diffusion centrality, with parameters $q = 1$ and T matching the parameter of the diffusion model.

Figure 11: Centrality-Based Targeting Versus Precise Targeting (2 seeds)



Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Optimal targeting involves targeting the two nodes that maximize diffusion on the baseline network. Network-based targeting involves targeting the top nodes by diffusion centrality, with parameters $q = 1$ and T matching the parameter of the diffusion model.

Figure 12: Simulated Total Treatment Effects on Academic Performance



Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by diffusion centrality with $T = 1$, which is equivalent to targeting by degree.

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, on the entire baseline sample (including those who are absent from the endline network). "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	Diffusion Centrality	Top 5% Degree	Top 5% Eigenvector Centrality	Top 5% Diffusion Centrality
Panel A. Information Network						
Academic Ability	2.43***	0.400***	0.427***	0.053***	0.043***	0.051***
	(0.293)	(0.051)	(0.052)	(0.013)	(0.012)	(0.012)
SES	1.33***	0.354***	0.321***	0.028**	0.031**	0.033**
	(0.310)	(0.056)	(0.056)	(0.014)	(0.013)	(0.013)
Male	-1.40***	-0.511***	-0.404***	-0.017	-0.042**	-0.027
	(0.456)	(0.081)	(0.082)	(0.020)	(0.019)	(0.019)
Observations	1,402	1,402	1,402	1,402	1,402	1,402
R^2	0.1684	0.0862	0.0765	0.0154	0.0163	0.0181
Panel B. Personal Network						
Academic Ability	1.08***	0.295***	0.334***	0.055***	0.038***	0.038***
	(0.159)	(0.051)	(0.052)	(0.014)	(0.012)	(0.012)
SES	0.649***	0.282***	0.261***	0.029**	0.043***	0.036***
	(0.169)	(0.054)	(0.056)	(0.015)	(0.012)	(0.013)
Male	-1.59***	-1.02***	-0.814***	-0.046**	-0.101***	-0.082***
	(0.256)	(0.079)	(0.081)	(0.023)	(0.022)	(0.021)
Observations	1,402	1,402	1,402	1,402	1,402	1,402
R^2	0.1792	0.1409	0.1038	0.0225	0.0364	0.0257
Panel C. Full Network						
Academic Ability	2.54***	0.390***	0.417***	0.047***	0.053***	0.050***
	(0.316)	(0.052)	(0.052)	(0.013)	(0.012)	(0.012)
SES	1.33***	0.310***	0.284***	0.015	0.024*	0.023*
	(0.334)	(0.055)	(0.055)	(0.013)	(0.013)	(0.013)
Male	-2.02***	-0.577***	-0.480***	-0.026	-0.059***	-0.044**
	(0.494)	(0.081)	(0.082)	(0.021)	(0.020)	(0.020)
Observations	1,402	1,402	1,402	1,402	1,402	1,402
R^2	0.1996	0.0852	0.0763	0.0133	0.0220	0.0177

Notes: Regressions of degree, eigenvector centrality, diffusion centrality, top 5% degree and top 5% eigenvector centrality and top 5% diffusion centrality on academic ability, high SES (SES), and male (estimating a single regression for each outcome). Eigenvector and diffusion centralities are normalized with respect to the control arm mean and standard deviation. Diffusion centrality parameters follow Banerjee et al. (2019) with q equal to the reciprocal of the top eigenvalue, and T equal to the diameter of the graph. Academic Ability is defined as above-median exam score at baseline. SES is equal to 1 if respondent's house has electricity and running water. The sample consists of students present at both baseline and endline (N=1,402). Heteroskedasticity-robust standard errors in parentheses. *** $p<0.01$; ** $p<0.05$; * $p<0.1$.

Table 4: 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 1). 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. The sample consists of students present at both baseline and endline. 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: 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 1). 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. The sample consists of students present at both baseline and endline. 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: Information Network Subcomponents

	Entertain-ment	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 1). 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. The sample consists of students present at both baseline and endline. 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: 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 1). 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. The sample consists of students present at both baseline and endline. 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: Information Access and Centrality

	Degree	Eigenvector	Number of Length-2 Walks	Diffusion	Betweenness	Average Link Strength
Panel A. Information Network						
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. Information Network: top 5% by centrality						
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
Panel C. 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 D. Full Network						
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: Panels A, C and D show the estimated differences between treated and control students for five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength (equation 3) in the Information Network, Personal Network and Full Network. Eigenvector, diffusion and betweenness centralities are normalized with respect to the control arm mean and standard deviation. Diffusion centrality parameters follow Banerjee et al. (2019) with q equal to the reciprocal of the top eigenvalue, and T equal to the diameter of the graph. Panel B shows the probability of being in the top 5% by centrality in the Information Network 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. The sample consists of students present at both baseline and endline (N=1,402). 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: Model Calibration and Treatment Effect Simulation

Moment	Simulated	Empirical
Panel A. Average centrality		
Degree (baseline)	9.79	10.1
Degree (control)	9.85	10.1
Degree (treatment)	10.7	10.9
Eigenvector centrality (baseline)	0.449	0.272
Eigenvector centrality (control)	0.447	0.324
Eigenvector centrality (treatment)	0.489	0.345
Number of length-2 walks (baseline)	117	148
Number of length-2 walks (control)	120	143
Number of length-2 walks (treatment)	130	153
Diffusion centrality (baseline)	3.54	3.5
Diffusion centrality (control)	3.49	3.55
Diffusion centrality (treatment)	3.8	3.76
Betweenness centrality (baseline)	0.0115	0.0123
Betweenness centrality (control)	0.011	0.0117
Betweenness centrality (treatment)	0.0129	0.0138
Panel B. Probability of being in top 5% by centrality		
Top 5% by degree (treatment)	0.0917	0.0697
Top 5% by eigenvector centrality (treatment)	0.0741	0.0627
Top 5% by number of length-2 walks (treatment)	0.0754	0.0732
Top 5% by diffusion centrality (treatment)	0.0743	0.0697
Top 5% by betweenness centrality (treatment)	0.0738	0.0697
Panel C. Simulated average treatment effects and reduced-form estimates		
Degree	1.02	0.964
Eigenvector centrality	0.0344	0.0277
Number of length-2 walks	15.8	12.3
Diffusion centrality	0.248	0.286
Betweenness centrality	0.00129	0.00245

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. Panel C: simulated average treatment effects against a counterfactual untreated network, compared to the reduced-form estimates of difference in centrality between treated and control students at endline. These reduced-form estimates are based on the specification in equation 3 (as in Table 8), but in this table the centrality measures are not normalized with respect to the control arm.

Online Appendix

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

For Online Publication

Laura Derksen and Pedro Souza

July 24, 2024

A Appendix

A.1 Randomization Inference for Dyadic Regressions

In Section 4.1, we estimate regressions of the following form:

$$100 \times \text{link}_{ij}^1 = \beta_0 + \beta_1 \cdot TC_{ij} + \beta_2 \cdot TT_{ij} + \alpha \cdot \text{link}_{ij}^0 + \mathbf{x}_{ij}'\boldsymbol{\chi} + \epsilon_{ij} \tag{12}$$

where $\text{link}_{ij}^1 = 1$ if a link is formed between i and j at the endline, $TC_{ij} = 1$ if i is treated and j is control, or vice-versa, $TT_{ij} = 1$ is i and j are treated, $\text{link}_{ij}^0 = 1$ if a link exists between i and j at the baseline, $\mathbf{x}_{ij}$ is a set of controls, and ϵ_{ij} is the error term. For this appendix, we focus on the case of the undirected networks in Table 4. Other specifications follow with minor modifications.

Consider a null hypothesis under which the intervention does not affect link formation at all. In particular, the sharp null hypothesis is that every dyad in the network would have the same relationship (linked or not linked) regardless of the treatment status of the two nodes involved, and regardless of the treatment assignments of the nodes in the wider network. Under this null hypothesis, the variable link_{ij}^1 is equal to its potential outcome under any treatment assignment. Thus, under the null hypothesis, the estimates of the effects of the intervention (β_1 and β_2) are statistically indistinguishable under various different treatment assignments.

More specifically, the randomization inference procedure in this case recovers the p-values the following way.

Step 1. Estimate Equation (12) under the original treatment allocation, and store $\hat{\beta}_1$ and $\hat{\beta}_2$.

Step 2. Reshuffle the treatment vector respecting the original stratification bins, and recompute the TC_{ij} and TT_{ij} variables that are consistent with the new treatment allocation, referred to as TC_{ij}^s and TT_{ij}^s.

Step 3. Reestimate (12) with the reshuffled treatment status,

$$100 \times \text{link}_{ij}^1 = \beta_0^s + \beta_1^s \cdot TC_{ij}^s + \beta_2^s \cdot TT_{ij}^s + \alpha^s \cdot \text{link}_{ij}^0 + \mathbf{x}_{ij}'\boldsymbol{\chi} + \epsilon_{ij} \tag{13}$$

and save the estimates $\hat{\beta}_1^s$ and $\hat{\beta}_2^s$.

Step 4. Repeat steps 2 and 3 above B =10,000 times, and compute the randomization inference p-values

$$p_{\beta_1} = \frac{1}{B}\sum_{s=1}^{B} I\left[|\hat{\beta}_1| > |\hat{\beta}_1^s|\right]$$

and similarly for β_2. Those estimates are reported in the paper.

Fredrickson and Chen (2019) show that randomization inference can be used to estimate causal effects on both local (e.g. link formation) and global (e.g. centrality measures) network outcomes with individual-level randomization. Blattman et al. (2021) show how randomization inference can be used to measure treatment effects and spillovers on a geographical network. For further applications and uses of randomization inference, see Duflo et al. (2007), Athey and Imbens (2017) and many others.

A.2 Interpreting Reduced-Form Estimates of Centrality Differences

In equation 3, we regress a node's centrality on its treatment status. These estimates, as discussed in Section 4.2, must be interpreted as relative differences between treatment and control nodes, as opposed to treatment effects. The estimates are, nevertheless, causal, in the sense that these relative differences are due to the intervention. They are unbiased estimates of the expected average difference in centrality between treated and control nodes.

Importantly, our estimates cannot be interpreted as an *average treatment effect*. In fact, in the context of network centrality, it is not straightforward to define an *average treatment effect*, and such effects are often not the parameter of primary interest. Moreover, centrality measures are highly interdependent, and the Stable Unit Treatment Value Assumption (SUTVA, Rubin 1974) will be violated. While treatment effects bounds can be estimated even when SUTVA violations are present (Manski, 2013), the required assumptions are likely too strong for a setting in which the outcomes themselves are non-localized network measures.

Let us demonstrate the appropriate interpretation of our estimates with a simple example. Consider a small network of five nodes, one of which was treated at random, with realized treatment vector

$$\mathbf{T} = \{T_1, T_2, T_3, T_4, T_5\}$$

with $T_i \in \{0, 1\}$ and $|\mathbf{T}| = 1$. We observe realized centrality measures for all nodes:

$$\mathbf{c} = \{c_1, c_2, c_3, c_4, c_5\}$$

One important parameter of interest in this context is the expected difference in centrality between treated and control nodes. This parameter allows us to shed light on the determinants of relative (as opposed to absolute) centrality within a given network. This is particularly relevant for network-based targeting, where policies typically target the most central nodes in a network based on their relative positions as opposed to their raw centrality scores. Within the context of our simple example with a single randomly-treated node, we can define this parameter as follows:

$$\rho = \sum_{i=1}^{5} \left(c_i^{T_i=1, T_{-i}=0} - \frac{1}{4} \sum_{j \neq i} c_j^{T_i=1, T_{-i}=0} \right) \mathbb{P}(T_i = 1)$$

where $c_j^{T_i=1, T_{-i}=0}$ is the potential outcome for node j when only node i is treated. We can produce an unbiased estimate of ρ by taking a simple difference of means. Suppose that in our realized sample, node k was treated. Then, we obtain the estimate

$$\hat{\rho} = c_k - \frac{1}{4} \sum_{j \neq k} c_j.$$

Taking the expectation over different realizations of $\mathbf{T}$, this estimator is unbiased:

$$\mathbb{E}(\hat{\rho}) = \frac{1}{5} \sum_{i=1}^{5} c_i^{T_i=1, T_{-i}=0} - \frac{1}{5} \sum_{i=1}^{5} \frac{1}{4} \sum_{j \neq i} c_j^{T_i=1, T_{-i}=0} = \rho \tag{14}$$

While the parameter we estimate is relevant and has a simple causal interpretation, it does not correspond to any parameter that would typically be thought of as an *average treatment effect*. First, because node centrality measures are interdependent, and depend on the entire vector of treatment statuses, defining a "treatment effect" for node i is not straightforward. For example, we could compare the treated node's centrality in a network in which only that node is treated to its potential outcome under no treatment $c_i^{\mathbf{T}=\mathbf{0}}$. Or, we could compare its potential outcome in a fully treated network $c_i^{\mathbf{T}=\mathbf{1}}$ to the untreated network $c_i^{\mathbf{T}=\mathbf{0}}$. Second, even if we settle on a "treatment effect" definition, we cannot produce an unbiased estimate of the average treatment effect across nodes with our data.

For example, if we define the average treatment effect as the expected effect on the treated node's centrality relative to the node's potential outcome in the untreated network, we then seek to estimate the following parameter:

$$\beta^{ATE} = \sum_{i=1}^{5} \left(c_i^{T_i=1, T_{-i}=0} - c_i^{\mathbf{T}=\mathbf{0}} \right) \mathbb{P}(T_i = 1) = \frac{1}{5} \sum_{i=1}^{5} c_i^{T_i=1, T_{-i}=0} - \frac{1}{5} \sum_{i=1}^{5} c_i^{\mathbf{T}=\mathbf{0}}.$$

Referring to equation 14, the problem becomes clear. Because $c_j^{T_i=1, T_{-i}=0} \neq c_j^{\mathbf{T}=\mathbf{0}}$, in general this expected value will not be equal to β^{ATE}.

Nevertheless, with our data we are able to test sharp null hypothesis of no treatment effect, for any node, under any treatment vector (Fredrickson and Chen, 2019). Under this null hypothesis, the potential outcomes are equal to the realized outcomes, $c_i^{\mathbf{T}} = c_i$, for any treatment vector $\mathbf{T}$. This null hypothesis implies zero expected centrality difference between treated and control students,

$$\mathbb{E}(\hat{\rho}) = \frac{1}{5} \sum_{i=1}^{5} \left(c_i - \frac{1}{4} \sum_{j \neq i} c_j \right) = 0,$$

and we can test the null hypothesis using randomization inference p-values for the parameter estimate $\hat{\rho}$.

A.3 Proof of Theorem 5.1

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

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

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

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

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

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

□

A.4 Calibration Details

Network formation. 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 (5), 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.

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

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 a 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

$$\mathbb{P}(g_{ij}^1 = 1 | g_{ij}^0 = 1, T_i = T_j = 0, \theta_i = \theta_1, \theta_j = \theta_2) = (1-\delta)^2 + 2\delta(1-\delta)\sqrt{P_{CC}^{\theta_1\theta_2}} + \delta^2 P_{CC}^{\theta_1\theta_2} \equiv P_{CC|CC}^{\theta_1\theta_2} \tag{15}$$

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

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, $\hat{P}_{CC|CC}^{\theta_1\theta_2}$ is estimated directly from the moment (15) 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:

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

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.

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

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.

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

Next, we simulate an endline network by adding links to to the shadow network. We assume that for fixed θ_1, θ_2, $\kappa_{CC}^{\theta_1\theta_2}$ is weakly smaller than $\kappa_{TC}^{\theta_1\theta_2}$, $\kappa_{CT}^{\theta_1\theta_2}$ and $\kappa_{TT}^{\theta_1\theta_2}$. That is, information is valuable and not costly to spread. This implies that $P_{CC}^{\theta_1\theta_2}$ is weakly smaller than $P_{TC}^{\theta_1\theta_2}$ and $P_{TT}^{\theta_1\theta_2}$, consistent with our reduced-form empirical results (see Table 4). Then, conditional on having a link in the shadow network, the probability

of having a link in the endline network is one.

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

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.

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

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

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

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

$$\begin{array}{llll} \hat{P}_{TT}^{LL} = 0.09 & \hat{P}_{TT}^{HL} = 0.13 & & \hat{P}_{TT}^{HH} = 0.29 \\ \hat{P}_{TC}^{LL} = 0.08 & \hat{P}_{TC}^{HL} = 0.12 & \hat{P}_{TC}^{LH} = 0.12 & \hat{P}_{TC}^{HH} = 0.25 \end{array} \tag{16}$$

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.

Academic performance. Next, we calibrate our model of academic performance. We modeled a student's academic score y_i as follows.

$$y_i = s_i + \tau(a_i)T_i + \tau(a_i) \sum_{j:g_{ij}^1=1} T_j Q_{ij} \tag{17}$$

We begin by taking the conditional expectation. We abuse notation to write $\tau(\theta_i) = \mathbb{E}(\tau(a_i)|\theta_i)$. While we will focus on capturing this average effect, we do not explicitly assume treatment effects to be uniform within ability-types. In the model of link formation we do assume that the expected number of treated links is independent of ability a_i given type θ_i.

$$\mathbb{E}(y_i|T_i, \theta_i) = \mathbb{E}(s_i|\theta_i) + \tau(\theta_i)T_i + q\tau(\theta_i)\mathbb{E}(\sum_{j:g_{ij}^1=1} T_j | T_i, \theta_i) \tag{18}$$

For a particular T_i and θ_i, we can use final exam scores in the year of the intervention to match $\mathbb{E}(y_i|T_i, \theta_i)$.

$$
\begin{aligned}
\mathbb{E}(y_i|T_i = 1) &= 0.38 \\
\mathbb{E}(y_i|T_i = 0) &= 0.26 \\
\mathbb{E}(y_i|T_i = 1, \theta_i = \theta_H) &= 1.43 \\
\mathbb{E}(y_i|T_i = 0, \theta_i = \theta_H) &= 1.38 \\
\mathbb{E}(y_i|T_i = 1, \theta_i = \theta_L) &= 0.26 \\
\mathbb{E}(y_i|T_i = 0, \theta_i = \theta_L) &= 0.13
\end{aligned} \tag{19}
$$

The last expectation in equation 18 can also be matched to a corresponding moment in the data, making use of the number of high and lower-ability treated students as well as the link probabilities we computed in equation 16.

$$
\begin{aligned}
\mathbb{E}(\sum_{j:g_{ij}^1=1} T_j|T_i = 1, \theta_i = \theta_H) &= 3.37 \\
\mathbb{E}(\sum_{j:g_{ij}^1=1} T_j|T_i = 0, \theta_i = \theta_H) &= 3.36 \\
\mathbb{E}(\sum_{j:g_{ij}^1=1} T_j|T_i = 1, \theta_i = \theta_L) &= 2.29 \\
\mathbb{E}(\sum_{j:g_{ij}^1=1} T_j|T_i = 0, \theta_i = \theta_L) &= 2.12
\end{aligned} \tag{20}
$$

Next, we subtract the conditional expectation for the control arm from the conditional expectation for the treated arm, as follows.

$$
\begin{aligned}
\mathbb{E}(y_i|T_i = 1, \theta_i = \theta_H) - \mathbb{E}(y_i|T_i = 0, \theta_i = \theta_H) &= \tau(\theta_H) + q\tau(\theta_H)(3.37 - 3.36) \\
0.05 &= \tau(\theta_H)(1 + 0.01q) \\
\mathbb{E}(y_i|T_i = 1, \theta_i = \theta_L) - \mathbb{E}(y_i|T_i = 0, \theta_i = \theta_L) &= \tau(\theta_L) + q\tau(\theta_L)(3.41 - 3.28) \\
0.12 &= \tau(\theta_L)(1 + 0.16q)
\end{aligned}
$$

Then, we take the unconditional expectation for the control arm.

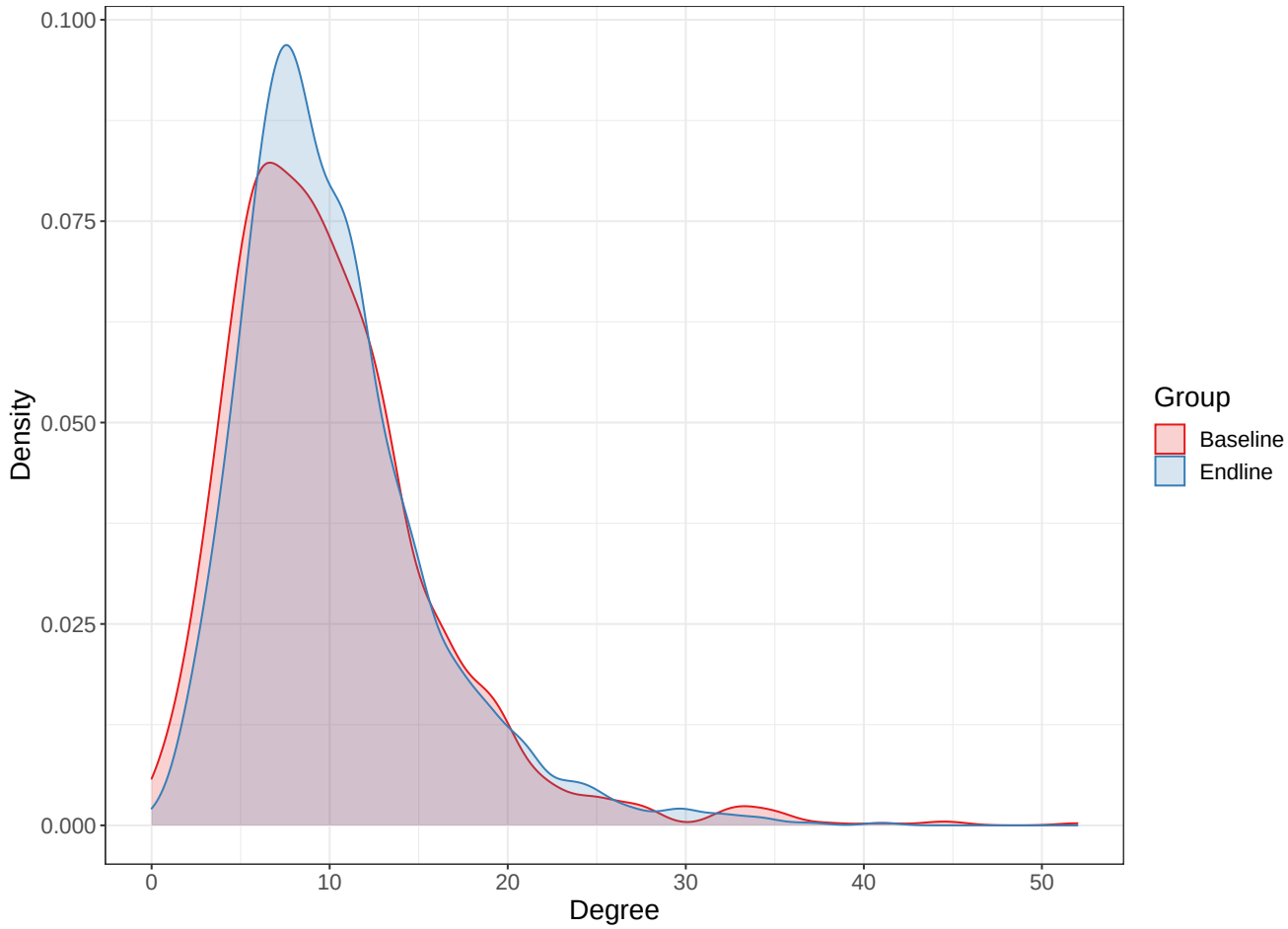
$$
\begin{aligned}
\mathbb{E}(y_i|T_i = 0) &= \mathbb{E}(s_i) + q\tau(\theta_H)\mathbb{E}(\sum_{j:g_{ij}^1=1} T_j|T_i = 0, \theta_i = \theta_H) + q\tau(\theta_L)\mathbb{E}(\sum_{j:g_{ij}^1=1} T_j|T_i = 0, \theta_i = \theta_L) \\
0.26 &= 0 + q(3.36\tau(\theta_H) + 2.12\tau(\theta_L))
\end{aligned} \tag{21}
$$

Here, we matched $\mathbb{E}(y_i|T_i = 0)$ to the endline control-arm mean, and $\mathbb{E}(s_i)$ to the mean from the previous school year, which is zero due to normalization. We now have three equations with three unknowns, which we can solve to obtain:

$$
\begin{aligned}
q &= 0.67 \\
\tau(\theta_H) &= 0.05 \\
\tau(\theta_L) &= 0.11
\end{aligned}
$$

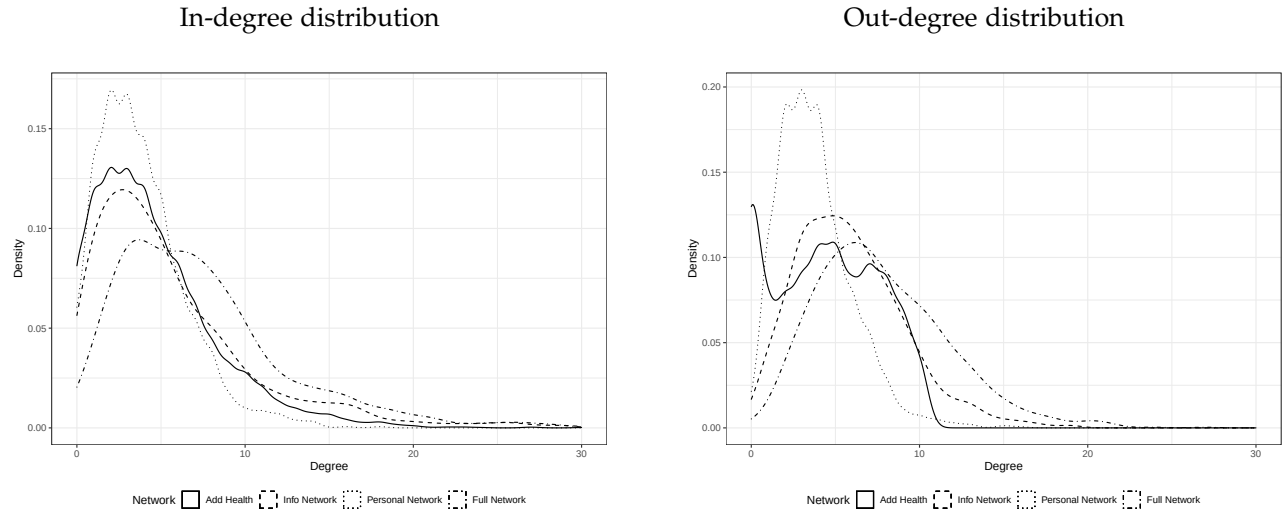
Appendix Figures and Tables

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



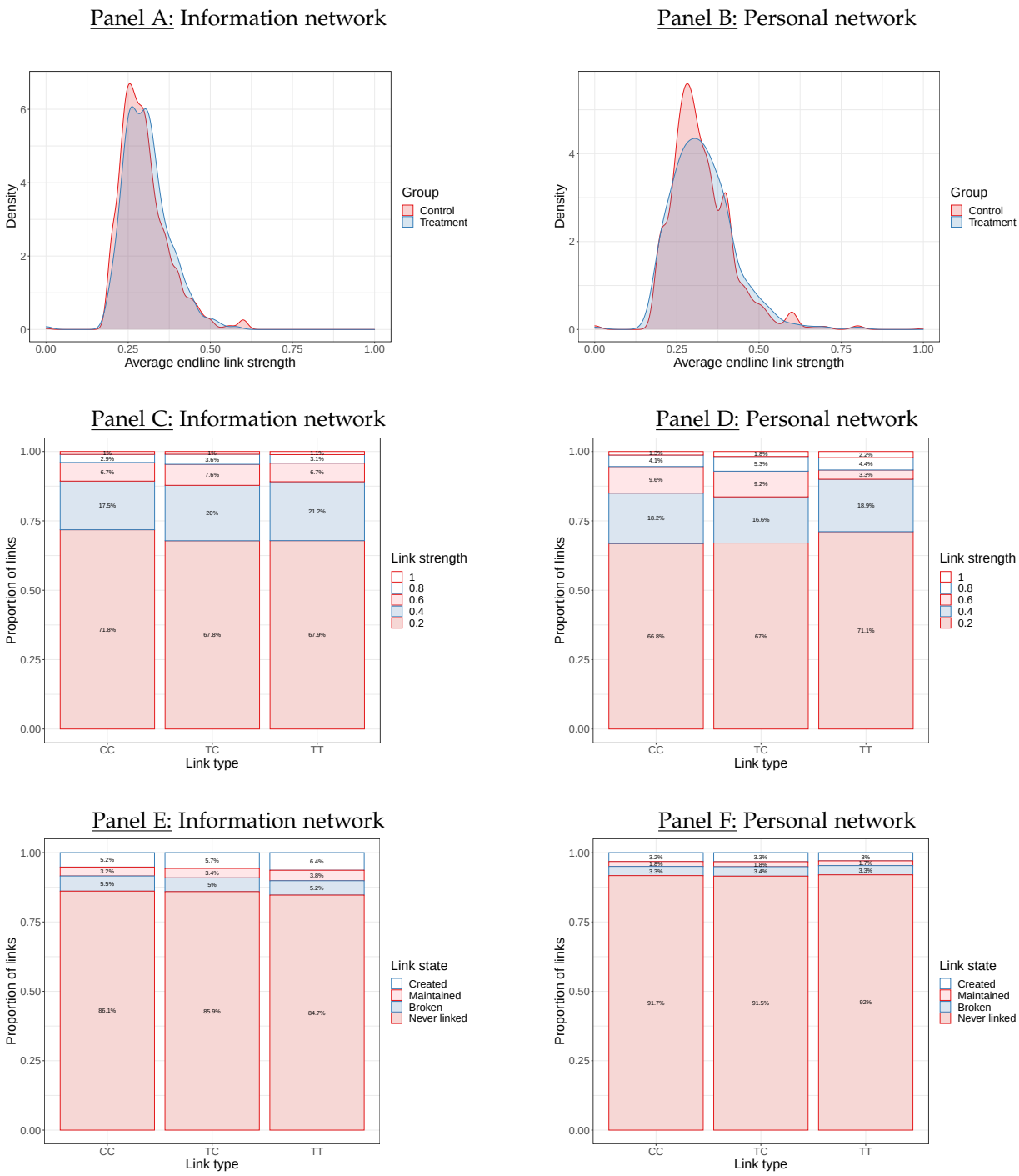
Notes: Degree distribution at baseline and endline. Information network. Sample restricted to nodes observed at both times.

Appendix Figure A2: Degree Distribution in Our Data vs AddHealth



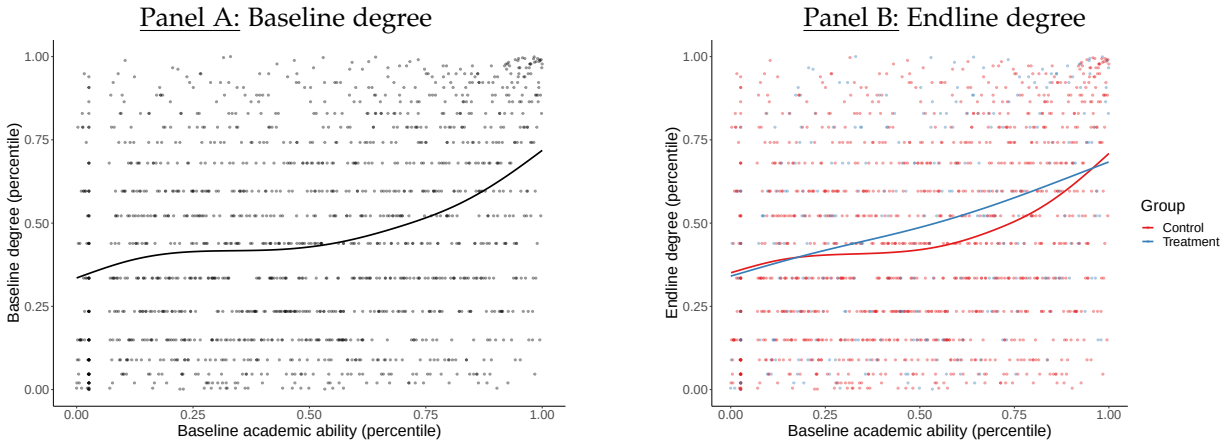
Notes: Degree distributions computed over our data for the information, personal and full networks; and the National Longitudinal Study of Adolescent Health ("AddHealth") obtained from https://www.icpsr.umich.edu/web/ICPSR/studies/21600/datasets/0003/variables/ODGX2?archive=icpsr and https://www.icpsr.umich.edu/web/ICPSR/studies/21600/datasets/0003/variables/IDG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

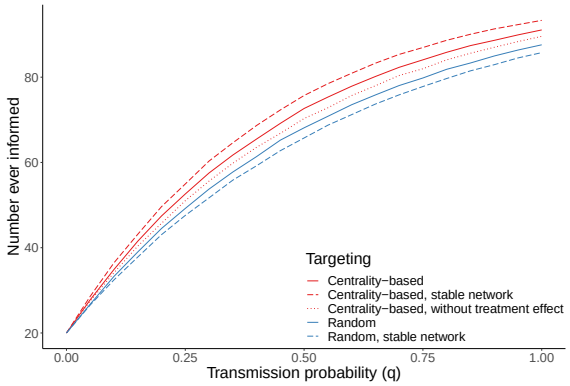
Appendix Figure A4: Baseline Academic Ability and Network Degree



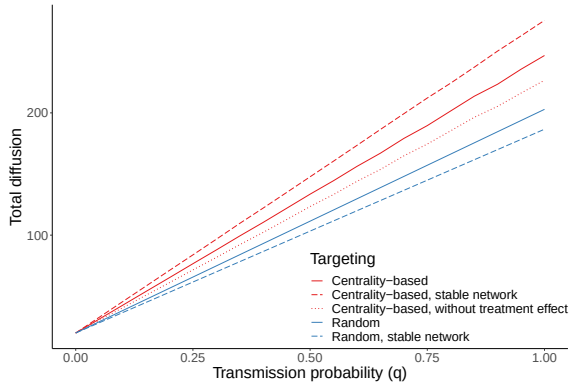
Notes: Correlation between baseline academic ability (percentile) and network degree (percentile) at baseline and endline.

Appendix Figure A5: SIR Model, Degree Centrality Targeting

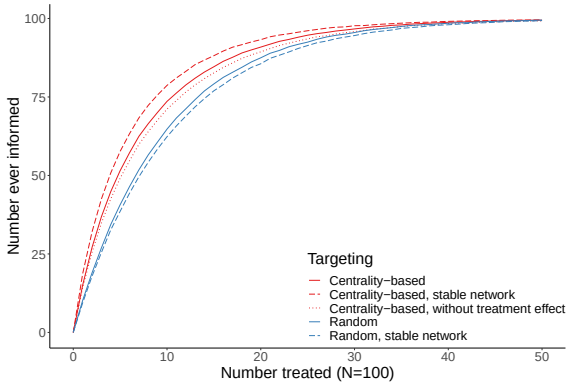
Panel A: Ever informed, 20 seeds and $T = 1$



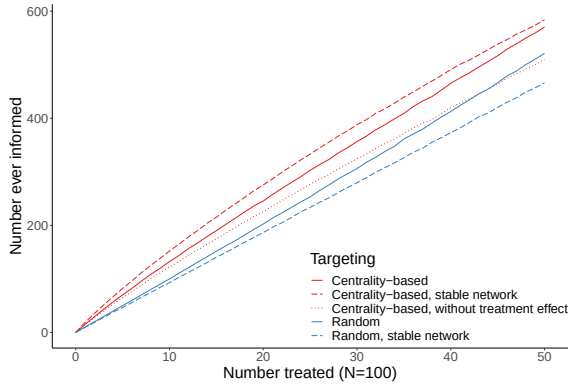
Panel B: Total Diffusion, 20 seeds and $T = 1$



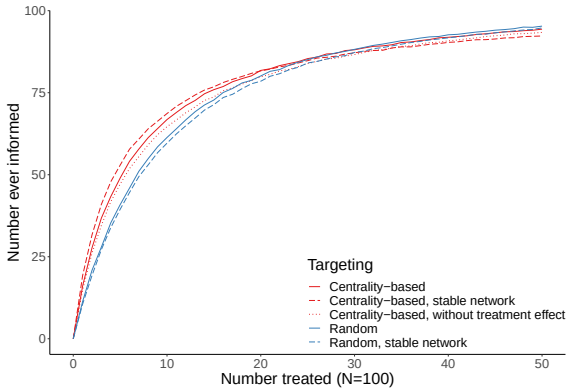
Panel C: Ever informed ($q = 1$, $T = 1$)



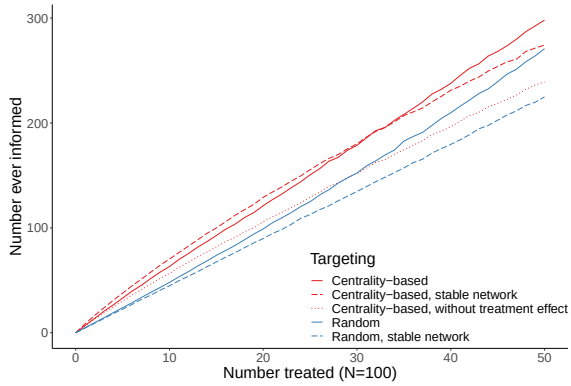
Panel D: Total diffusion ($q = 1$, $T = 1$)



Panel E: Ever informed ($q = 0.1$, $T = 4$)



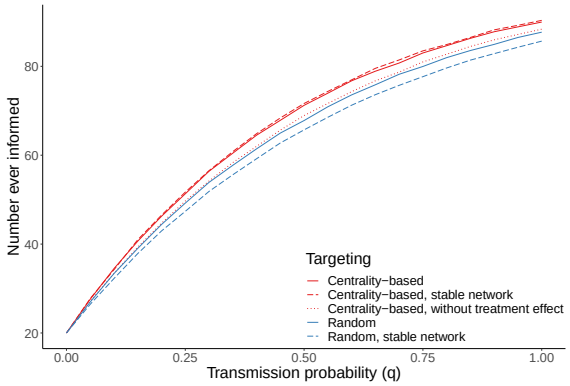
Panel F: Total diffusion ($q = 0.1$, $T = 4$)



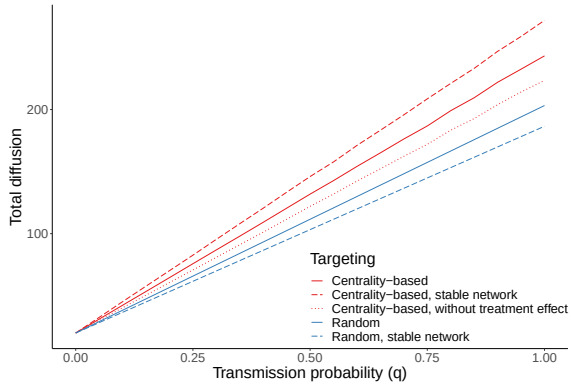
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by degree. $q^* = 0.1$ and $T^* = 4$ are set to equal the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019).

Appendix Figure A6: SIR Model, Eigenvector Centrality Targeting

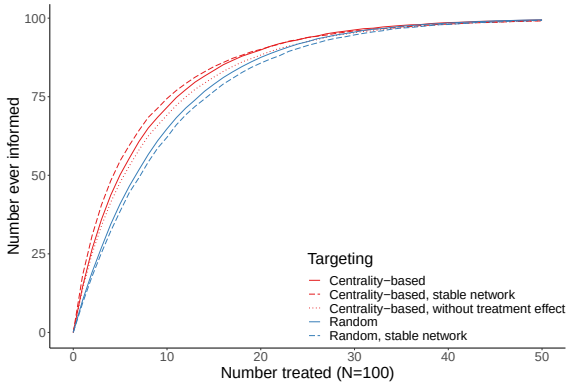
Panel A: Ever informed, 20 seeds and $T = 1$



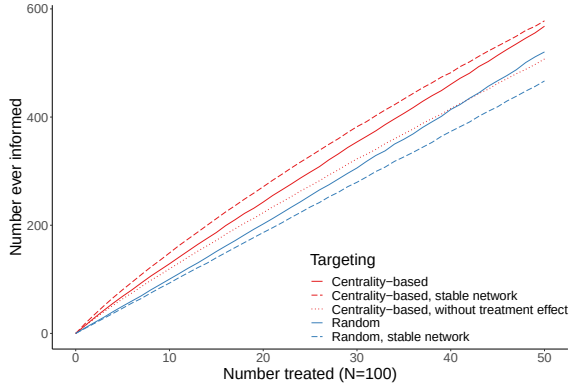
Panel B: Total Diffusion, 20 seeds and $T = 1$



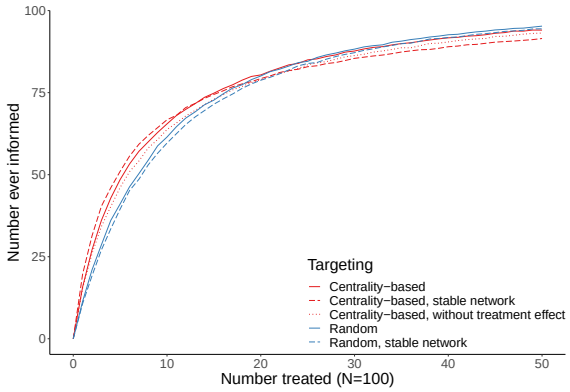
Panel C: Ever informed ($q = 1$, $T = 1$)



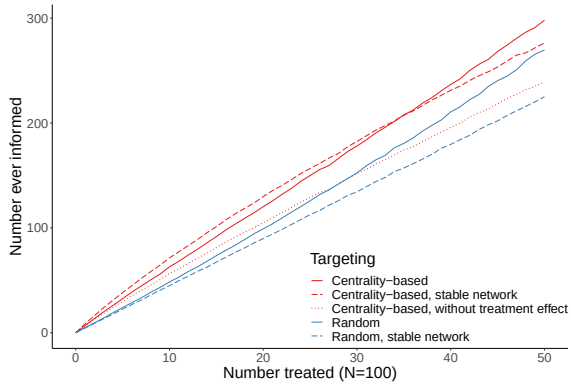
Panel D: Total diffusion ($q = 1$, $T = 1$)



Panel E: Ever informed ($q = 0.1$, $T = 4$)



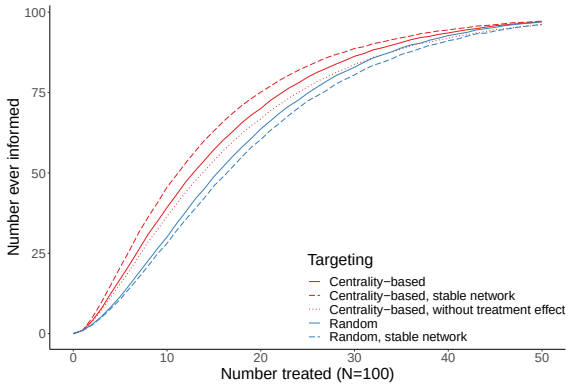
Panel F: Total diffusion ($q = 0.1$, $T = 4$)



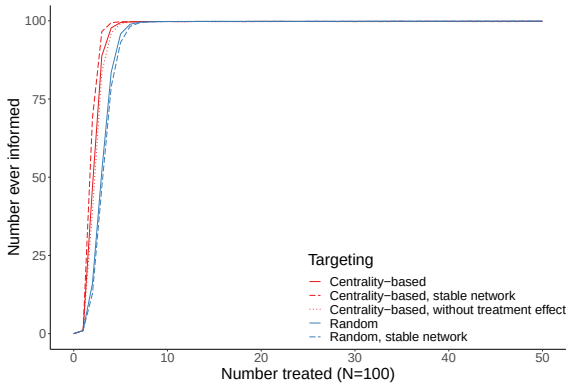
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by eigenvector centrality. $q^* = 0.1$ and $T^* = 4$ are set to equal the reciprocal of the top eigenvalue and diameter of the graph respectively, as in Banerjee et al. (2019).

Appendix Figure A7: Threshold Model, Degree Centrality Targeting

Panel A: Number informed ($\lambda = 2$, $T = 1$)



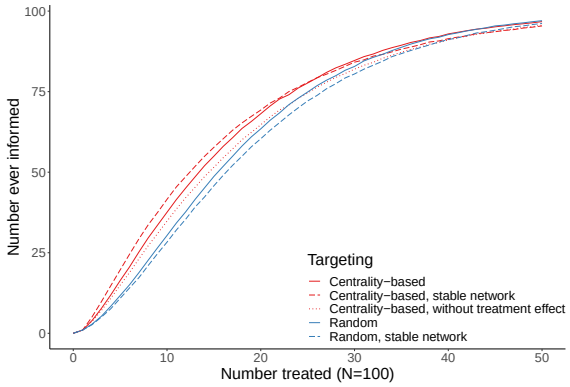
Panel B: Number informed ($\lambda = 2$, $T = 4$)



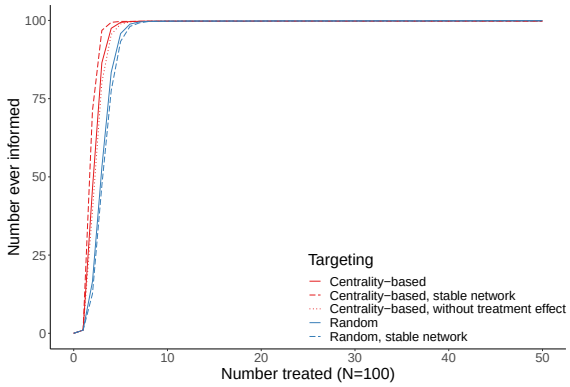
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by degree centrality.

Appendix Figure A8: Threshold Model, Eigenvector Centrality Targeting

Panel A: Number informed ($\lambda = 2$, $T = 1$)



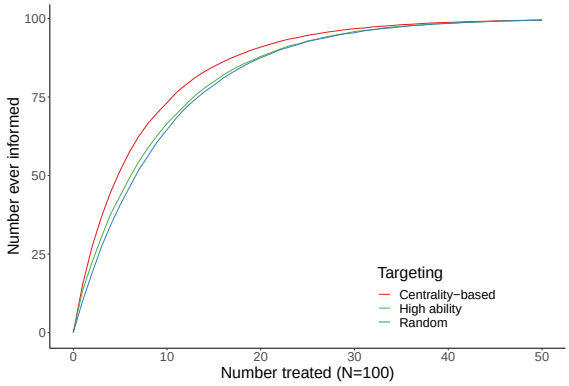
Panel B: Number informed ($\lambda = 2$, $T = 4$)



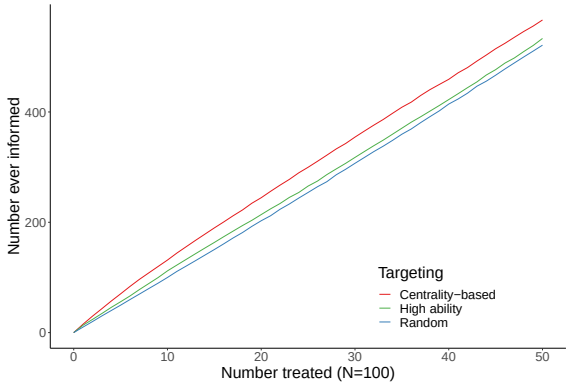
Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by eigenvector centrality.

Appendix Figure A9: SIR Model, Centrality-Based Versus High-Ability Targeting

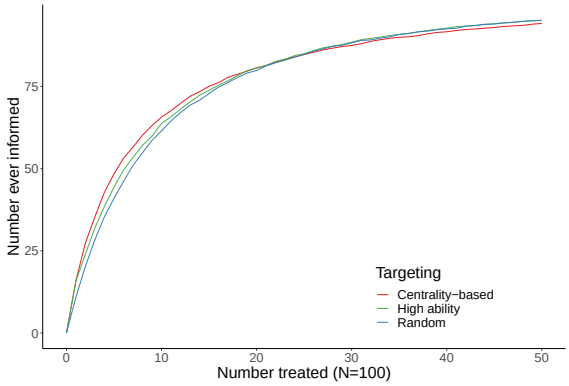
Panel A: SIR, ever informed ($q = 1$, $T = 1$)



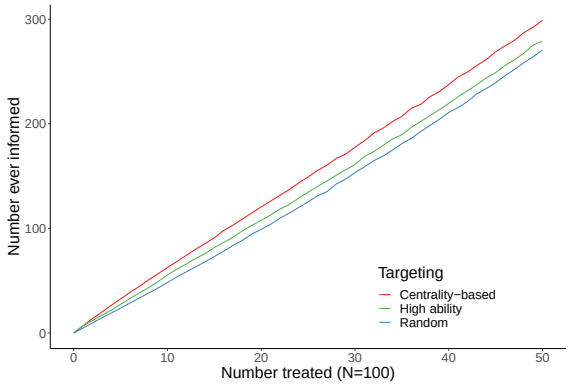
Panel B: SIR, total diffusion ($q = 1$, $T = 1$)



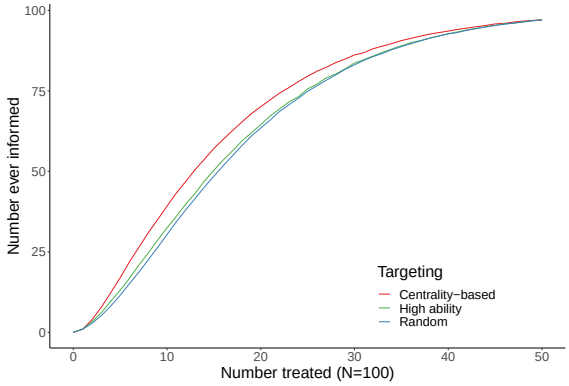
Panel C: SIR, ever informed ($q = 0.1$, $T = 4$)



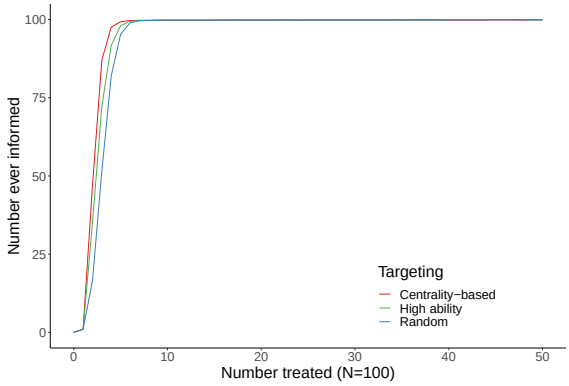
Panel D: SIR, total diffusion ($q = 0.1$, $T = 4$)



Panel E: Threshold model ($\lambda = 2$, $T = 1$)

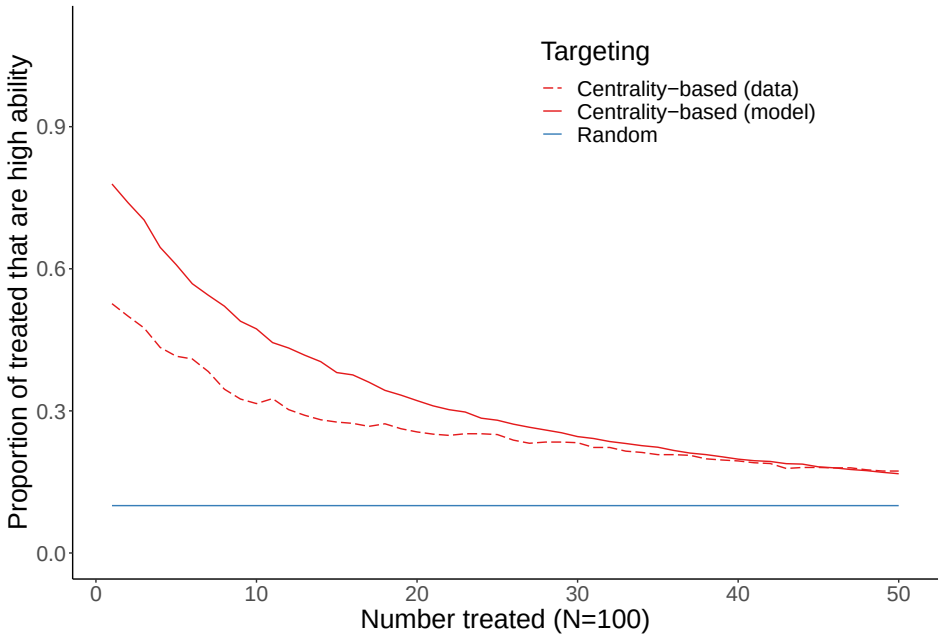


Panel F: Threshold model ($\lambda = 2$, $T = 4$)



Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Network-based targeting involves targeting the top nodes by diffusion centrality, with parameters q and T matching the parameters of the SIR diffusion model.

Appendix Figure A10: Centrality-Based Targeting and Baseline Ability



Notes: Simulations of 100-node networks, with 1000 replications for each set of parameter values. Diffusion centrality calculated using $q = 1$ and $T = 1$ (equivalent to degree). The dashed line corresponds to the share of top-centrality nodes that are high-ability in our baseline data.

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: 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. The sample consists of students present at both baseline and endline. 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: Directed Links with Second-order Effects

	Information	Full	Personal
Treat-to-Control Link	0.346**	0.275*	-0.078
	(0.140)	(0.159)	(0.111)
	p = 0.053	p = 0.207	p = 0.530
Treat-to-Control Link x Number of Treated Friends of i at Baseline	0.056	-0.029	0.039
	(0.098)	(0.090)	(0.087)
	p = 0.665	p = 0.823	p = 0.719
Treat-to-Control Link x Number of Friends of i at Baseline	-0.020	-0.017	-0.050
	(0.031)	(0.029)	(0.034)
	p = 0.588	p = 0.632	p = 0.225
Treat-to-Control Link x Number of Treated Friends of j at Baseline	0.233**	0.257***	0.231***
	(0.102)	(0.093)	(0.088)
	p = 0.265	p = 0.323	p = 0.647
Treat-to-Control Link x Number of Friends of j at Baseline	-0.040	-0.058*	-0.041
	(0.034)	(0.030)	(0.036)
	p = 0.435	p = 0.303	p = 0.632
Control-to-Treat Link	0.738***	0.697***	0.120
	(0.146)	(0.164)	(0.115)
	p = 0.001	p = 0.003	p = 0.351
Control-to-Treat Link x Number of Treated Friends of i at Baseline	0.233**	0.219**	0.332***
	(0.100)	(0.093)	(0.088)
	p = 0.271	p = 0.478	p = 0.256
Control-to-Treat Link x Number of Friends of i at Baseline	-0.047	-0.064**	-0.088**
	(0.029)	(0.028)	(0.035)
	p = 0.331	p = 0.231	p = 0.168
Control-to-Treat Link x Number of Treated Friends of j at Baseline	-0.082	-0.020	0.081
	(0.110)	(0.099)	(0.094)
	p = 0.612	p = 0.893	p = 0.472
Control-to-Treat Link x Number of Friends of j at Baseline	0.065*	0.036	-0.050
	(0.036)	(0.033)	(0.035)
	p = 0.239	p = 0.456	p = 0.243
Treat-to-Treat Link	1.06***	0.742**	-0.235
	(0.279)	(0.305)	(0.205)
	p = 0.002	p = 0.067	p = 0.365
Treat-to-Treat Link x Number of Treated Friends of i at Baseline	0.299	0.123	0.204
	(0.216)	(0.191)	(0.174)
	p = 0.279	p = 0.716	p = 0.646
Treat-to-Treat Link x Number of Friends of i at Baseline	-0.081	-0.081	-0.072
	(0.060)	(0.055)	(0.059)
	p = 0.249	p = 0.277	p = 0.426
Treat-to-Treat Link x Number of Treated Friends of j at Baseline	0.185	0.357*	0.335*
	(0.218)	(0.197)	(0.172)
	p = 0.532	p = 0.260	p = 0.373
Treat-to-Treat Link x Number of Friends of j at Baseline	0.103	0.069	-0.072
	(0.069)	(0.062)	(0.063)
	p = 0.210	p = 0.378	p = 0.421
Number of Treated Friends of i at Baseline	-0.051	-0.044	-0.094**
	(0.043)	(0.042)	(0.040)
	p = 0.474	p = 0.572	p = 0.196
Number of Friends of i at Baseline	0.030**	0.054***	0.060***
	(0.013)	(0.013)	(0.017)
	p = 0.349	p = 0.260	p = 0.072
Number of Treated Friends of j at Baseline	-0.105**	-0.121***	-0.095**
	(0.046)	(0.043)	(0.042)
	p = 0.177	p = 0.107	p = 0.199
Number of Friends of j at Baseline	0.265***	0.228***	0.104***
	(0.016)	(0.014)	(0.017)
	p = 0.392	p = 0.280	p = 0.163
R^2	0.1046	0.1272	0.0974
Observations	165,422	165,422	165,422

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 *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 number of treated friends of *i* and *j* at the baseline, number of friends of *i* and *j* at baseline, 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.

Appendix Table A4: Information Access and 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: Estimated differences between treated and control students in the contact network for five measures of centrality (degree, eigenvector, number of length-2 walks, diffusion, and betweenness centralities) and average link strength (equation 3). 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. The sample consists of students present at both baseline and endline (N=1,402). 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: 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 Academic Ability	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. The sample consists of students present at both baseline and endline (N=1,402). 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: 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: Estimated differences in centrality between treated and control students 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. The sample consists of students present at both baseline and endline (N=1,402). 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 A7: 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: Estimated differences between treated and control students for 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. The sample consists of students present at both baseline and endline (N=1,402). 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 A8: Information Access and Centrality 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 3). 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. The sample consists of students present at both baseline and endline. 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 A9: Predictors of Centrality at Endline

	Degree			Eigenvector Centrality			Diffusion Centrality		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Information Network									
Baseline Degree	0.536***		0.510***						
	(0.026)		(0.028)						
Baseline Eigenvector Centrality				0.607***		0.584***			
				(0.031)		(0.034)			
Baseline Diffusion Centrality							0.614***		0.586***
							(0.030)		(0.034)
Academic Ability		2.38***	1.11***		0.456***	0.233***		0.478***	0.232***
		(0.270)	(0.231)		(0.055)	(0.050)		(0.056)	(0.050)
SES		1.53***	0.867***		0.338***	0.146***		0.334***	0.154***
		(0.273)	(0.229)		(0.056)	(0.049)		(0.056)	(0.049)
Male		-0.568	0.114		-0.129	0.154**		-0.138	0.087
		(0.356)	(0.305)		(0.087)	(0.078)		(0.086)	(0.075)
R^2	0.471	0.227	0.485	0.364	0.127	0.381	0.368	0.112	0.384
Observations	1,402	1,402	1,402	1,402	1,402	1,402	1,402	1,402	1,402
Panel B. Personal Network									
Baseline Degree	0.337***		0.308***						
	(0.027)		(0.027)						
Baseline Eigenvector Centrality				0.364***		0.287***			
				(0.027)		(0.029)			
Baseline Diffusion Centrality							0.366***		0.314***
							(0.028)		(0.030)
Academic Ability		0.810***	0.485***		0.214***	0.138***		0.263***	0.165***
		(0.150)	(0.141)		(0.048)	(0.046)		(0.051)	(0.048)
SES		1.02***	0.822***		0.368***	0.289***		0.374***	0.294***
		(0.156)	(0.148)		(0.051)	(0.050)		(0.054)	(0.052)
Male		-0.768***	-0.284		-0.702***	-0.414***		-0.519***	-0.268***
		(0.224)	(0.216)		(0.077)	(0.080)		(0.080)	(0.081)
R^2	0.287	0.221	0.308	0.286	0.258	0.322	0.226	0.172	0.255
Observations	1,402	1,402	1,402	1,402	1,402	1,402	1,402	1,402	1,402
Panel C. Full Network									
Baseline Degree	0.532***		0.504***						
	(0.025)		(0.026)						
Baseline Eigenvector Centrality				0.599***		0.567***			
				(0.029)		(0.032)			
Baseline Diffusion Centrality							0.594***		0.563***
							(0.029)		(0.032)
Academic Ability		2.58***	1.27***		0.447***	0.233***		0.467***	0.236***
		(0.298)	(0.255)		(0.054)	(0.048)		(0.055)	(0.049)
SES		1.94***	1.28***		0.369***	0.204***		0.368***	0.216***
		(0.301)	(0.254)		(0.055)	(0.047)		(0.055)	(0.048)
Male		-0.749*	0.244		-0.283***	0.034		-0.202**	0.060
		(0.405)	(0.343)		(0.085)	(0.075)		(0.085)	(0.074)
R^2	0.476	0.264	0.493	0.373	0.143	0.391	0.361	0.121	0.381
Observations	1,402	1,402	1,402	1,402	1,402	1,402	1,402	1,402	1,402

Notes: Regressions of endline degree, eigenvector centrality, diffusion centrality on their baseline values (Columns 1, 4, and 7), on academic ability, high SES (SES), and male (Columns 2, 5 and 8) and all (Columns 3, 6 and 9). Eigenvector and diffusion centralities are normalized with respect to the control arm mean and standard deviation. Diffusion centrality parameters follow Banerjee et al. (2019) with q equal to the reciprocal of the top eigenvalue, and T equal to the diameter of the graph. Academic Ability is defined as above-median exam score at baseline. SES is equal to 1 if respondent's house has electricity and running water. All regressions have class fixed effects. Heteroskedasticity-robust standard errors in parentheses. *** p<0.01; ** p<0.05; * p<0.1.

Appendix Table A10: Treatment Effects on Academic Scores

	English	Biology
Panel A. Overall effects		
Treatment	.103**	.063
	(.050)	(.047)
	p = .046	p = .192
Panel B. Heterogeneous treatment effects		
Treatment x Below Median Ability	.195**	.143**
	(.076)	(.067)
	p = .016	p = .043
Treatment x Above Median Ability	.003	-.025
	(.062)	(.064)
	p = .964	p = .707
Control Mean	.000	.000
Observations	1412	1406

Notes: Table reproduced from Derksen et al. (2022). Treatment effects on final exam scores. Ability defined as above (below) median exam scores (average of English and Biology) at the baseline. We include a control for baseline exam score, an indicator for missing baseline score, and strata fixed effects. Randomization was stratified by school, form, above median achievement and past internet use. Robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Randomization inference p-values based on 10,000 replications denoted as "p =".