Network Effects on Outcomes and Unequal Distribution of Resources

by

Eaman Jahani

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Social and Engineering Systems and Statistics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author .....
Institute for Data, Systems, and Society
September 01, 2021

Certified by .....
Alex 'Sandy' Pentland
Toshiba Professor of Media Arts and Sciences
Thesis Supervisor

Certified by .....
Dean Eckles
Associate Professor in Management
Thesis Supervisor

Certified by .....
Matthew O. Jackson
William D. Eberle Professor of Economics
Thesis Committee Member

Accepted by .....
Fotini Christia
Chair, Social and Engineering Systems Program

Network Effects on Outcomes and Unequal Distribution of Resources

by

Eaman Jahani

Submitted to the Institute for Data, Systems, and Society
on September 01, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Social and Engineering Systems and Statistics

Abstract

We focus on the link between networks and economic outcomes, study how networks affect different groups differently and in the process provide pathways to reinforce existing inequalities. First, we establish the link between the economic well-being and the network structure of US counties. We show that counties that are rich with long ties, those bridging different communities, have better outcomes over a range of economic indicators. Subsequently. we study the determinants of long ties and find that they are more frequent if the individual has experienced disruptions such as mobility, migration or switching schools throughout their life. Our findings suggest that creating and maintaining long ties require special skills that co-occur with the above mentioned life events.

Second, we provide observational evidence for differential network advantages in access to information: higher status individuals receive higher marginal benefit from networking. We attribute this phenomena to unequal diffusion due to network homophily and provide causal evidence for it in the context of a randomized seeding experiment in networks.

Third, we develop a network model that captures the structure of unequal diffusion or access to opportunities. We show that any departure from the uniform distribution of links to information sources in a group has both first order and second order effects. Not only some individuals will have fewer direct links, but also the whole group will have fewer diffusion paths to the information sources.

Finally, we examine the network mechanisms that widen inter-group differences. We study an information sharing game, in which individuals have to compete for a rivalrous resource over repeated rounds. The equilibrium predicts lower cooperation among lower status agents, which leads the whole group to receive a "a smaller share of the pie". We further validate this prediction in an online lab experiment.

We hope that our findings contribute to the growing literature around the network origins of persistent inequality. Our findings suggest that policies that target groups rather than individuals are more successful in combating inequality as the benefits that arise from lifting a whole group out of poverty will be amplified by the existing

social capital and the feedback mechanisms present in the network.

Thesis Supervisor: Alex 'Sandy' Pentland
Title: Toshiba Professor of Media Arts and Sciences

Thesis Supervisor: Dean Eckles
Title: Associate Professor in Management

Thesis Committee Member: Matthew O. Jackson
Title: William D. Eberle Professor of Economics

Acknowledgments

My work and this thesis would not have been possible without the support of many friends and colleagues. I must express my gratitude to my wonderful supervisors, Prof. Alex 'Sandy' Pentland and Prof. Dean Eckles. During my PhD studies, I enjoyed a great level of independence and it was all due to Sandy's support who consistently motivated me to develop my own ideas and research identity. I sincerely thank him for being such a supportive advisor, for great insights, pushing me to see research findings in a general context and providing me with the resources I would have not received anywhere else. Sandy created an exemplary inter-disciplinary research group for his students and this was a constant source of inspiration. Sandy is a passionate researcher with a clear vision who strives for real-world impact. I have been very fortunate to have worked with and learned from him in the past years, and I hope to always enjoy from his deep insights in the future.

I am grateful to Dean for his mentorship and dedication to my growth as an independent researcher, for his willingness to invest valuable time and energy in my work. Dean was always available for any question I had, ranging from details on methodology to general position of our research in the wider context. Dean was involved in all aspects of our research, from the formulation of hypothesis to estimation procedures. At the same time, I always felt a great deal of autonomy. In other words, he maintained a perfect balance between engagement and independence, a truly rare skill which I hope I have learned from him for my own future students. Dean is a great social scientist with wide ranging interests and in-depth knowledge in various fields. He is also a great methodologist, an important skill in the age of big data when seemingly interesting findings can be easily due to methodological shortcomings. I was always impressed with his quick grasp of research methods, dedication to correct inferences using the right methods, and meticulous attention to details while keeping the general picture in mind. I hope to continue working with him and learning these skills from him.

I would also like to thank Prof. Jackson for his support, great insights, valuable

feedback and most importantly the influential references he directed me to. Our meetings were filled with many great suggestions, and I incorporated many of his ideas into my work especially on the impacts of long ties, unequal diffusion and strategic network formation. Matthew has had many impactful studies on networks from an economic perspective and my work is heavily influenced by his contributions to the field. I closely read almost all of his book 'Social and Economic Networks' page by page in the first two years of my PhD and that was a very worthwhile investment. I am grateful for having his advice on my research and hope to learn from and collaborate with him in the future.

My collaborators were a constant source of inspiration and I want to thank them for their support and contribution to my work. I thank the faculty at the Institute for Data, System and Society, Prof. Ali Jadbabaie, Prof. Esteban Moro, Amin Rahimian, Yuan Yuan, my collaborators from Media Lab, Prof. Xiaowen Dong, Prof. Abdullah Almaatouq, Alejandro Noriega, Prof. Peaks Krafft, and Yoshihiko Suhara for their mentorship and valuable collaboration on various projects. I am also grateful to my collaborators outside MIT, Prof. Matthew Baum, Prof. Christopher Bail, Prof. Nicolo Cavalli and Prof. Douglas Guilbeault. Working with all these colleagues has been a rewarding experience. I specially appreciate my collaboration with Michael Bailey from Facebook and must thank him for the opportunity to work with him. He is a very resourceful researcher and provided me with all the support I needed in our projects. I have had a very fruitful collaboration with him so far with two promising papers and I hope to continue our collaboration in the future. I express my gratitude to Elizabeth Milnes from the IDSS administrative office for her constant support and welcoming me in her office whenever I needed help. IDSS community is very lucky to have her. I also appreciate the funding support from NSF graduate research fellowship program which supported me in the first 3 years of my PhD.

Finally, I must express my love and gratitude to my family and friends. My family has been a constant source of encouragement and I am lucky to have them around me. My work and this thesis is dedicated to them: Mehdi, Homa, Roger, Afarin, Franca, Alberto, Stefano, Valentina, Matteo, Eleonora, Jayran, Irene and Nora.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Inequality has been on the rise in the past few decades [67, 131] and as such has become a renewed focus of social scientists. There is ample new evidence that inequality has become increasingly persistent across generations affecting both individuals [58] and neighborhoods as a whole [152, 173]. Extensive studies have documented negative effects of inequality, on how it slows down growth and development [14] or affects institutions [3]. Most researchers on the topic from the economic perspective address macro-level phenomena such as technology and taxation as instruments that affect and exacerbate inequality [7, 144, 150]. These works document how the wealth share of top 0.1% was continuously declining from its peak in 1929 until 1979. But since then the trend has reversed and their wealth share has more than tripled from 7% to 22% in 2012. Comparative works against other industrialized nations which have not experienced such a dramatic wealth inequality (e.g. France, Germany and Japan) attribute this to declining tax rates on top income shares.

Sociologists on the other hand have focused on social processes that occur in social networks or organizations and lead to different outcomes for different groups, hence creating inequality. Perhaps, one of the most influential recent works, albeit by an economist, that emphasized the importance of network effects on inequality is by Chetty and Hendren [57]. They showed that inter-generational mobility, while limited to a great extent, varies greatly across the US and is correlated with less segregation and more social capital [63]. These factors are both network phenomena

and they determine a child's chance of upward mobility and in aggregate have significant consequences in terms of inequality. More recent work by the same authors made the link to peer effects and exposure more explicit [56], in which they found that moving to better neighborhoods at an early age improves the child's chance of upward mobility and increase their income on average.

The sociological literature has made important contributions to the study of networks and outcomes, which acts as the foundation for our argument on how networks affect inequality. At its core, our argument rests upon the fact that economic action is embedded in social settings [90], and therefore it is affected by unequal distribution of social resources. Sociologists such as Granovetter and Lin argued that any analysis of economic outcomes without considering their social context is futile [91]. They showed empirically how network of peers plays an important role in economic outcomes such as access to business opportunities [89] or status attainment [122, 123]. The foundational work by Granovetter [92] demonstrated that job search is embedded in informal social networks and Lin showed that the prestige of the acquired job not only depends on personal resources such as education but also on the social status of the referral contact [120]. Later studies provided more nuance for network effects and showed how specific, and in particular brokerage, positions in the network lead to greater power, higher salary and status, more creativity, and control over flow of information [34, 39]. Granovetter points out many social processes that can affect economic outcomes, such as cultural norms or prevalence of trust (over incentives), but the main channel and perhaps the most important one is that social networks control the flow and quality of information to actors.

If social networks significantly affect the economic outcomes, a natural question is whether they affect inequality, or the unequal provision of economic opportunities, as well? and if networks do impact inequality, what are the mechanisms under which they generate inequality? There have been several case studies documenting how networks lead to unequal outcomes. A recent study in development economics investigated whether information about a new agricultural composting technology diffuses at the same rate to male and female farmers in a village [20]. Seeding the information

with the most connected individuals in the village network led to significant disadvantages for women as they usually occupied the periphery of village networks and had less access to well-connected nodes, hence information never reached them. Here we have a situation in which one gender group is disadvantaged in terms of its network or social resources, and the diffusion of valuable information, which often originates from prominent individuals, in the network reinforces the existing gender inequality. Network effects on gender inequality at workplace has received some attention mainly from the management science. A prominent paper showed that within an advertisement firm, men receive more career benefits from their networks than women, even after fixing their position in the firm's network [101]. The effect is attributed to the fact that there are homophilous groups [129] within the firm's network and men have access to more resources. In this setting, it might be more beneficial career-wise for women to form ties to their male colleagues, as also evidenced by Burt [33].

The differential network benefits and their effect on inequality is not limited to gender groups. Networks can widen existing inequalities along any attribute. In an interesting study, Laschever exploits the random assignment of WWI draftees to circumvent the problem of endogenous group formation and to causally study the effect of a network on the individual employment outcome [117]. Findings suggest that outcomes (employment rate) of one's assigned group but not its characteristics have a statistically significant effect on a veteran's employment status. This finding is in line with our argument above that networks act as the diffusion channel for economic information and peer effects, at least in the economic context, mainly occur through this mechanism rather other processes such as norms or influence that depend on characteristics of peers.

Another recent study documented how children of slave holders were able to immediately overcome the wealth shock experienced by their fathers due to abolition of slavery after the civil war and reach their parents income position prior to the civil war within one generation [5]. The resurgence of the children to wealth status of their parents before the civil war was mainly due to extensive social resources in the form of dense and exclusive network of favor exchanges and economic assistance that

they inherited from their parents. The authors argue that the pattern of recovery was similar to those of social networks in providing job information. This form of elite persistence is also consistent with the argument of Acemoglu and Robinson that the elite mainly rely on their *de facto* political power, in form of social capital and access to networks, which remains intact despite a big shock to *de jure* political structure [4].

Finally, DiMaggio and Garip treated migration as a diffusion process and by doing so explained why certain villages in Thailand experienced mass migration to cities whereas other villages remained intact [71]. Migration is a social process in which the presence of some contacts who have already migrated to the destination serves as an information source and facilitates the migration of potential new migrants. If two otherwise identical villages started with small differences in their rates of initial migration, over time the network effect described above will amplify the initial differences causing them to follow diverging migration trajectories. All the aforementioned papers share a common theme that there are two groups in the society, and one has access to more valuable resources (e.g. economic information) which is shared and retained within the group giving them exclusive access and more power. These examples highlight the importance of formulating a general mechanism how networks contribute to inequality. The outcome of this mechanism is cumulative advantage [73, 118], such that small individual differences get amplified by the network and intensify pre-existing privilege of certain groups.

In search of a general network mechanism in regenerating or exacerbating inequality, three sets of authors heavily influenced our thinking. The following three papers discussed various similar mechanisms for network effects on inequality at a theoretical level. The model proposed by Calvó-Armengol and Jackson [41] and a similar later model by Bolte, Immorlica and Jackson [28] attempt to provide a modeling perspective on why participation in labor force is very different between whites and blacks [45]. Their work builds on the job search in networks literature [92, 123] and explicitly models access to job opportunities through network of social contacts and referrals. The persistent differences in employment status of agents arises from the simple fact

that "the better the employment status of a given agent's connections (e.g. relatives, friends), the more likely it is that those connections will pass information concerning a job opening to the agent" creating extra incentives to stay in the labor force. This happens because if the contacts are employed themselves, the employment information does not benefit them directly so they pass it on to their unemployed contacts. In contrast, if contacts of an unemployed agent drop out of the labor market, the prospect of a future employment for the agent decreases and as a result there will also be less incentive for the agent to stay in the labor market. This information passing process along with the contagion in drop-outs lead to a correlation in employment status of subgroups of densely connected agents. At its root, this network effect rests on the homophily [105, 129] of workers by employment and future referrals, which leads to clusters of fully employed or fully unemployed agents over time.

Furthermore, the model predicts that a small initial difference in employment status of two otherwise identical subgroups translates into persistent advantage in employment status and future income of one subgroup over the other, since higher employment rates within a subgroup means new employment opportunities are generated more frequently within the group which get further amplified due to homophilous nature of the group. The model also exhibits a stickiness behavior in employment rate: if a group is at full unemployment, their state will be sticky due to reinforcements by the network. In retrospect, Laschever findings [117] above suggested that there were strong spill-over effects in economic status of connected veterans from WWI. Groups were either employed or unemployed together indicating that policies that target unemployment have the additional benefit that would propagate through social networks.

In a separate theoretical paper, DiMaggio and Garip discuss a framework under which individual advantages become multiplied by network advantages, hence exacerbating inter-group differences [72]. This mechanism relies on nuanced processes on the network, such as social learning, higher trust or normative pressure but they all boil down to network externalities that influence adoption decisions about a new beneficial behavior. For example, adopting a new productive enhancing technology (e.g. agri-

cultural composting [20]) improves one's utility and its adoption is not only a function individual endowments (e.g. education) but also the number of one's contacts who have already adopted the new technology. However, in most cases the individual characteristics that increase the likelihood of adoption are correlated among connected individuals in the network (e.g. networks are homophilous in terms of education). As a result, individuals with higher initial endowment receive extra benefits from their network of similarly endowed contacts, further increasing their likelihood to adopt. This theoretical framework is mainly discussed in terms of new behavior adoption, but one can replace behavior adoption with economic or employment information in the framework and reach the Calvó-Armengol and Jackson model described above. Both models rely on the assumption that some members start a with slight advantage and such members are clustered in the network. In both models, the network provides extra utility, either through employment information or externalities.

In the framework above, the two groups had different levels of social resources which they could mobilize to their advantage. The differential access to social capital by certain social groups (e.g. gender or race) is what's referred to as "inequality in social capital" by Lin [119, 120] (As opposed to Colmen [63], by social capital we are referring to the resources embedded in the network and accessible to the individual rather than their relationships). As described above, inequality in social capital happens when members of a group belong to the disadvantaged socioeconomic status and they also have the tendency to associate with each other (e.g. homophily by gender or race). Lin describes the *differential opportunities* between various groups in the society as one implication of this deficiency in social capital.

While different in terms of their language and approach, all the works above by Lin, Garip and Dimaggio, Jackson, Calvó-Armengol, Bolte and Immorilca share the same underlying mechanism, the unequal provision or simply the *unequal diffusion of network resources*. By network resources, as a general concept, we refer to any utility that an individual receives from their network, for example economic or employment information or externalities in behavior adoption. Unequal diffusion of network resources happens if the social structure exhibits the following three characteristics:

1. The resource diffuses through the network links.

2. One of the groups (exogenously) generates the resource at a higher rate compared to the other groups (e.g. high social class owns businesses and generates employment opportunities at a higher rate).

3. The social network is homophilous along a characteristic that is correlated with the group attribute (e.g. business community, comprising the high social class, mostly attend certain colleges).

The resource could be rivalrous, shareable with a limited number of contacts, or non-rivalrous. The effect of network inequality will exist no matter, but we hypothesize the network effects will be stronger for rivalrous resources as the homophily will make them more exclusive to one group than non-rivalrous resources.

At its basic form, we can model the total utility of an individual as follows:

$$Utility = Individual\ Utility + Network\ Utility \tag{1.1}$$

Individual utility is derived from individual endowment (e.g. education or ability) whereas network utility is derived from network resources which depend on the endowment of the contacts in the local network. So we can redefine utility as follows:

$$Utility = f(individual\ endowment) + g(endowments\ of\ local\ network) \tag{1.2}$$

We are interested in variation of network utility across different individuals, no matter if they have different individual utility or not. If there were no network effects on inequality, network utility would be identical for everybody in the network. In order for network utility to vary across different individuals from different groups, both conditions (2) and (3) above must be true. If the groups generated resources at different rates but the network was random with no homophily, all individuals would receive the same expected network utility since all have the same local network composition, even though their individual endowments would be different. Similarly, if the network was homophilous but both groups generated resources at similar rates,

there would not be any difference in available network resources between any two members from different groups. If the above 3 conditions are met, *the individual endowment advantage of one group will get amplified by the network effect, leading to inter-group differences that are larger than what's expected by differences in individual endowment.*

An important property of equation (1.2) is that the function corresponding to network utility could be linear or super-linear in terms of network endowments. A linear function indicates that the network utility scales at the same rate as the number of contacts with the valuable resource (e.g. farming coops). However, a super-linear function indicates cumulative advantage in which *the marginal utility* increases as the number of contacts with the resource increases. This pattern of cumulative advantage leads to larger inter-group differences and is more common in social settings. For example, both cases of new technology adoption [72] and the decision to stay in the labor market [41] fall into this category. In the case of new product adoption, the marginal utility of adoption increases as there are more contacts who have already adopted, as also evidenced by others [166]. In the case of diffusion of employment information, the super-linearity stems from the fact that once one contact drops out of the labor market, it becomes more likely that other contacts will drop out too. In general, we expect the network utility to exhibit a super-linear form in cases that resemble the complex contagion setting [53], in which the network resource becomes beneficial only after multiple contacts provide it.

The processes mentioned above lead to differential network advantages, but they are not easily observable or measurable. For example, there have not been any strong empirical evidence on the nature of unequal diffusion in networks or investigation to the existence of differential marginal benefits on the field as noted by both DiMaggio and Lin. In fact, as stated by Garip and DiMaggio, the research priority for network effects on inequality should be 1. Specifying the mechanism rigorously (e.g. in terms of utilities) and identify the network effects through field work. 2. Collecting appropriate data to perform empirical analysis of network effects and investigate the form of the network component in the utility function. We contribute to the literature by

providing empirical support for the network inequality effects through observational data, field experiment with randomized seeding of a new behavior and an online lab experiment where we control the network structure and initial inter-group differences. Furthermore, to understand the network nature of unequal access to opportunities, we develop a random network model that fits the extent of unequal diffusion in a network, in addition to its other structural properties, and relates it to node-level variation in cross-group link formation.

First, we examine the link between social network structure and economic outcomes. Several foundational works have argued that economic action does not happen in a vacuum of fully rational agents, rather it happens in social structures [91, 106, 110]. The economy and social structure affect each other and co-evolve, a phenomena often referred to as the embeddedness of economic action in social settings [90]. Numerous studies have examined this phenomena in the context of managerial success in firms [11, 66, 116, 145, 146], employment status [21], firm performance [168, 169], health outcomes [52], reputation [181] and education [44, 70]. The work by Burt on structural holes is perhaps the most relevant in this line of research [34]. Burt argued that actors create social capital by positioning themselves as brokers among disconnected communities, which in turns provides them with access to novel and diverse sources of information and at the same time control over the flow of information. Burt and subsequent researchers documented that access to these structural holes, also referred to as long ties, leads to better performance in various micro-scale studies on individual actors.

Despite its importance, there has been very little evidence on the link between social structure and economic well-being, and in particular the frequency of long ties in communication networks, at a macro scale of neighborhoods, mainly due to unavailability of data that fully maps out networks of US neighborhoods. We attempt to fill this gap by examining the relationship between the economic well-being and the network structure of US counties. Using Facebook communication data, we show that counties or zip codes that are rich with long ties, those bridging different communities, have better outcomes over a range of economic indicators including income, social

mobility, and employment. We show that this link is robust under a different network measurement scheme or in a different country (Mexico). These results agree with previous work which used aggregated mobility and purchase activity as proxies of communication across neighborhoods [61, 99]. Furthermore, we find that conditioned on their frequency, these bridging long ties are stronger in more prosperous zip codes. This result suggests that the underlying mechanism is indeed the transmission of novel information across long ties: the stronger the long ties are, the more likely information transmission is successful.

Subsequently, we study the determinants of long ties and the individuals who hold them. We find that long ties are more frequent if the individual has experienced major disruptions throughout their life. We examine three case studies of inter-state migration, out-of-state college attendance and switching high schools and show that in each case, the individual has more long ties in their networks many years after the experience than other comparable individuals but without those experiences. Most strikingly, the higher frequency of long ties among these individuals cannot be fully explained due to geographic mobility, as migrants have more long ties in their current states than the locals or individuals who attended multiple high schools have more long ties with friends outside high school. Our findings suggest that creating and maintaining long ties require special skills that co-occur with the above mentioned life events. Other authors have documented the personality component of individuals who span structural holes. Burt et al. found that these actors are predominantly entrepreneurs and less risk-averse [38] while Kalish and Robins emphasized the reduced vulnerability to change and the ability to engage with multiple distinct social categories among these bridging actors [112, 113]. Our results on migration and mobility are in line with these findings as these individuals have a reduced tendency to classify themselves into distinct social categories, thrive on change and have developed the ability to mitigate conflicting demands.

Second and after showing the link between networks and economic outcomes, we provide correlational results from a cross-sectional study that suggests individuals from a high status group receive differential benefits from their network. Using the

call records from about 33,000 individuals in a south Asian country, we operationalize social capital in terms of structural diversity or the frequency of long ties, and find that structural diversity shows a relatively strong association with individual income. Furthermore, the marginal effect of structural diversity is exclusive to the individuals from a high status group. These results provide an initial evidence for the mechanism above that concentrated distribution of economic opportunities among the high status social strata combined with homophily among members of the same group leads to differential network advantages for the high status group and limited diffusion of economic opportunities to the low status social strata. While these results support the network inequality mechanism from a large real-world data, they are not causal.

Next, we attempt to provide causal evidence for differential diffusion of a new behavior in a controlled setting and relate it to network homophily and the initial advantage of one group. We will use the data provided by a randomized experiments [140] that studied diffusion of a new behavior in a social network when a few initial nodes are randomly seeded with the information. In this study, authors seeded an anti-conflict intervention that was randomly assigned to initial seed students and evaluated the causal effect of its diffusion on the rate of conflict, both at the school level and individual student level. In the context of our study, we would like to show that the students that are similar to the initial seed students are more likely to adopt the new behavior. In other words, the effect of intervention in terms of adoption of the new behavior is larger on students that are homophilous with the initial seed students than non-homophilous students. Relating to the mechanism we are interested in, the group to which the initial seed students belong constitutes the advantaged class, with access to the new behavior by the seeds acting as the initial advantage of the group. We will show that this initial advantage by one member of the group (seed student) leads to differential advantages in terms of adoption for other members of the group compared to non-members in the network.

Third, we argue that network homophily alone does not fully explain the extent of unequal diffusion in the network. Instead, we show that any variation in cross-group linking among nodes of a group leads to higher unequal diffusion than expected

simply by measures of homophily. As the first step, we show this phenomena analytically by building a latent space network model [96] with a parameter that controls the degree of heterogeneity in cross-group linking. As we increase this parameter, the cross-group links become more exclusive to a small number of nodes and the network exhibits smaller number of inter-group paths that cannot be explained by the simple assortativity metrics [133]. As the second step, we investigate the extent of this phenomena in real-world networks. We find that most networks exhibit higher susceptibility to unequal diffusion than expected by simple measures of homophily across various grouping variables (e.g. gender or race).

To address this discrepancy, we develop a variant of the Stochastic Block Model [2] that parameterizes the cross-group degree of each node, and by doing so fits not only assortativity on paths of length 1, but also assortativity on paths of length 2 and 3. This model significantly improves the fit over SBM on empirical networks and accurately predicts the susceptibility of a network to unequal diffusion. Our modeling results show that networks that have similar levels of assortativity or homophily could have very different degrees of susceptibility to unequal diffusion. This variation arises from the extent to which the connections between different social classes are controlled by a small group of individuals acting as brokers [34]. Our findings provide an important policy implication: while brokers act as a bridge between communities, they nevertheless control the flow of information. Networks that heavily depends on brokers for connectivity suffer from unequal diffusion of information much greater than networks whose cross-group links are equally distributed. This happens because the existence of brokers not only hampers diffusion to the first degree but also to the second degree, the whole group ends up with less pathways to information sources.

Finally, we focus on one network mechanism that is responsible for widening inter-group differences. This mechanism rests on the scarcity of a valuable resource and the heterogeneity in group ability to access it. We design a stylized repeated game in which agents make strategic decisions to share the valuable resource with each other, if and when they receive it. In each round, some agents randomly (according to a known distribution) receive the *rivalrous resource* access to which increases one's utility.

Since there is a limited stock of the resources in each round, it gets shared equally among agents with access to it. This resource rivalry creates opposing incentives for the sharing decision: in the short term, it is better to withhold the resource while in the long term it is more optimal to share. The repeated nature of the game could amplify individual or group differences in utility over time and our goal is to explore the mechanisms that affect the final inter-group utility differences. In particular, assume there are two groups in the network (rich/poor). Members of group 1 and 2 have p_1 and p_2 probability of receiving the resource on their own in each round ($p_1 > p_2$). Within each round, agents can share information about the resource with their contacts. Adding the individual and network components of the game, members of each group will have an average utility of u_1 and u_2. We will describe the necessary conditions for the game to result in the following equilibrium:

$$\frac{u_1}{u_2} > \frac{p_1}{p_2} \tag{1.3}$$

Given that individual probabilities, p_1 and p_2, capture variation in individual ability, inequality (1.3) indicates that one group takes a larger share of the common pie (the rivalrous resource) than would be expected solely based on individual differences. We find that if the network is homophilous and the initial difference between the two groups is large enough, cooperation (in the form of resource sharing) emerges only in the advantaged group and the equilibrium satisfies the inequality above. In other words, the network component in this resource sharing game amplifies the inter-group advantages to be greater than the exogenous individual differences. The differential in cooperation rates is due to the differences in future prospects of each group: since the resource is too scarce in the disadvantaged group, there is very little incentive to share it as the expected future payoff from contacts is less than the lost utility if the agent was to share it. These individual micro-level motives lead to increasing inequality between the groups at macro-level [155]. We implement the above game in an online randomized lab experiment in a manner similar to [126] and validate that the sharing decisions resemble the equilibrium predicted by the theory.

We hope that our findings contribute to the growing literature around the network origins of inequality and its persistence. Previous evidence suggests that inequality is closely related to its persistence across generations (i.e. immobility) [64]. The persistence of inequality can be explained by the reproduction of disadvantage across generations: children born into impoverished families inherit their parent's social capital and are inherently more constrained in access to opportunities, which in turn leads them to have poor future outcomes. Most current policies to combat inequality focus on its economic roots, while there is inadequate attention on its social origins. When it comes to the issue of immobility, one could even argue that the current economic policies such as redistribution only address the symptoms of inequality rather its socially driven causes [108]. To address the persistence of inequality, our policies should also take into account the unequal access to opportunities and other mechanisms in which social networks lead to unequal outcomes. We hope that our work contributes to this goal by describing these network based mechanisms and providing suggestions to counteract them. Our findings suggest that policies that target groups rather than individuals are more successful in combating inequality as the benefits that arise from lifting a whole group out of poverty will be amplified by the existing social capital and the feedback mechanisms present in the network.

Chapter 2

Formation of Long Ties and their Economic Outcome

1 Preface

In this chapter, we establish the link between network structure and economic outcomes at the aggregate scale of US counties. In particular, we focus on long ties or the structural hole phenomena as often referred to in the literature. We show that long ties are more frequent in US counties that are richer along various economic indicators. The strength of these long ties further predicts economic well-being: the stronger the long ties are on average, the higher the median household income is. Subsequently. we study the determinants of long ties and find that they are more frequent with mobility, migration and other major disruptions throughout the course of one's life. Our findings suggest that creating and maintaining long ties require special skills that co-occur with the above mentioned life events.

2 Introduction

Economic action is embedded in social networks [90]. There are various ways in which networks can affect economic outcomes both at the micro and macro scale [91, 106, 110]. Perhaps the most important mechanism via which networks determine

outcomes is through their structure which controls who gets access to what opportunities [18, 65, 75]. This link between structure and network advantage has been widely studied as it has important implications both within and among firms. Burt introduced the concept of structural holes, perhaps the most important insight in this line of work [34, 39], which states that brokers who connect two otherwise disconnected communities fill a structural hole, thus facilitate the flow of diverse and novel information between the two disparate communities. But, more importantly, structural holes benefit the brokers themselves by giving them the potential for autonomy, opportunity recognition, information arbitrage, and innovation [1, 25, 32, 77, 147]. Broker networks are characterized by their numerous bridges across structural holes, which we refer to as *long ties*, to remote circles in the social network. Long ties contribute to the structural diversity of an ego-network, since the lack of mutual contacts signifies a connection to a structurally distinct community. The information advantage and innovation mechanisms [32, 147] state that these long ties are more likely to provide access to novel, non-redundant and diverse sources of information and resources because they are structurally diverse and connect to many unique communities.

Since the introduction development of the related concepts of weak ties and structural holes, there have been many studies that have linked advantage in terms of network structure to performance, at the scale of individual actors. Most work has examined the issue within the context of a firm and have documented the role of local network structure to managerial performance [11, 94, 145–147]. These works investigate different nuances of local structure and their interaction with tie strength, but they all report that managers whose networks are diverse and connect to multiple communities are promoted faster, paid more and deemed to have higher performance by their peers compared to managers with densely interconnected network. The underlying mechanism for their success is the investment these managers make in diverse groups which later enables them to search for and transfer valuable information from their diverse contacts at the right time. Several other works have made similar conclusions, based the same mechanism, on the link between local network structure and performance in the context of labor market [49, 83], team innovation [69, 130, 138],

technology adoption [12, 16, 115] and peer influence [10].

Despite the extensive literature on local network structure and individual or team performance in specialized contexts, the link between network structure and important economic indicators such as income, social mobility or employment rate is yet unclear. This is mainly due to the difficulty in accessing such data and more importantly to causally determine such effects over long term. However, with the increasing availability of large amount of digital interaction data which can be linked to outcome profiles, there has been more and more studies that have attempted to link networks with economic outcomes at various scales. The units of analysis in most of these studies have been aggregated at a level of counties or neighborhoods [61, 74, 181]. A recent paper shows how patterns of human mobility across different neighborhoods in Istanbul and Beijing predict neighborhood economic growth [61]. Similarly, Eagle et al. showed how the diversity of phone communication across UK regions, measured by Shannon entropy, is a key indicator of economic prosperity [74]. Another interesting study examined the relationship between position in the communication network and economic status, at the level of individuals in a whole country, and discovered that the importance of an individual to network cohesion, similar to k-core, is correlated with individual wealth [125]. The most relevant to our current work is a recent study on social connectedness from Facebook [15]. Bailey et al. constructs the network of counties in the United States (US) in which connections signify the number of unique friendships between residents of counties on Facebook. The network connections of a county, and in particular their geographic distributions, predicted an array of county-level economic characteristics, such as income, education, teenage birth and life expectancy. While those findings were not causal, they nevertheless for the first time provided a very detailed description of the association between the aggregate characteristics of contacts and economic outcomes at the county-level.

The recent studies have advanced our knowledge on the network correlates of economic prosperity at population scale, but the link between economic outcomes and network structure in particular as it pertains to long ties is still unknown. Our first contribution is to fill this gap by providing comprehensive evidence on the relationship

between the frequency of long ties and economic outcomes at the aggregate scale of US zip codes. Using commenting communication among Facebook users in the US, we construct the networks of each zip code and show that zip codes whose residents have more long ties tend to have better outcomes in an array of economic indicators. We replicate the same findings at the US county level and in a different country, Mexico. Furthermore, we find that the strength of long ties predicts the remaining variation in outcomes even after accounting for the frequency of long ties. In other words, US zip codes with numerous strong long ties tend to have the best economic outcomes. Our findings could involve endogeneities between network structure and outcomes, however given the vast literature on how long ties benefit individuals by facilitating the diffusion of novel economic information, we believe the links we discover are significant and have important policy implication regarding local access to opportunity and economic growth.

Given the importance of long ties in empowering people and their economic consequences, it is fair to ask where do long ties come from, who holds them and what factors enable people to create and maintain long ties? Surprisingly, there has been very little focus on this question and the determinants of long ties is very much an open question. There has been a limited number of previous studies on the personality and the characteristics of network brokers [38, 112, 113, 164] but they do not provide any insight on the qualities of brokers and how they got there when brokerage is viewed as a skill to be acquired rather than a fixed personality attribute. Nevertheless, they heavily influenced our thinking on what external factors make a broker, so we briefly discuss them here. Burt et al. used surveys to explore whether personality varies systematically with access to structural holes. They found that respondents with access to structural holes claim the personality of an entrepreneurial outsider (versus conforming and obedient insider), in search of authority (versus security), and thrive on change (versus stability) [38]. If we view brokerage as a skill rather than an innate ability, the most relevant finding from Burt et al. is that people with closed networks are risk-averse and depend on social support of contacts compared to brokers who have had experience with change and disruptions.

Kalish and Robins treated personality as a permanent trait and looked at how different traits predispose people to build certain network structures. In particular, they studied how the big 5 personality traits [87] correlate with access to structural holes [113]. They found that people with closed triads tend to be vulnerable and passive agents to external forces and define themselves more as part of a group with a collective identity rather than an individual. In contrast, brokers are individualistic, believe in their ability to influence their environment and control the course of events in their lives. Most importantly for our study, they reported that keeping one's friends apart is not a state of balance, and has psychological strain on the individual. Most people seek to avoid these dissonant-like structures, but the fact that brokers have these networks suggest they have special skills and experience to create and maintain these open and structurally diverse networks.

The balance associated with closed triads traces its root to the work of Goerg Simmel [158]. Simmel described a process where a third mediator enters into a dyadic relationship between non-homophilous individuals from different communities and potentially with conflicting demands. The mediator can utilize their unique skills to resolve the conflict between these unconnected individuals. Simmel argued that these mediating individuals see themselves responsible for coordination and managing conflicting incentives. A later study empirically validated these theoretical predictions by evaluating the personality traits and motivations of brokers that connect individuals from different backgrounds engaged in a conflict [112]. They discovered that the mediator brokers described by Simmel are less likely to define themselves and others in terms of social categories and are able to integrate with various social groups. These findings are in agreement with a fascinating study on the characteristics of Cosimo Medici who rose to power in Renaissance era Florence [139] due to his unique network advantage and ability to reap the benefits offered by his many long ties. Medici was a master of multivocality, was able to mediate between conflicting actors, and had the skills to negotiate in presence of uncertainty and heterogeneity.

The studies mentioned above treated brokerage as a matter of personality, an inherent quality, but did not not examine the processes that bring about these char-

acteristics and the skills associated with them. This is specially important since many of the personality correlates of brokers, such as proclivity to change and disruptions, are not innate and can often be acquired through life experience. The fact that brokers and structural holes are rare is in itself an evidence of the difficulty and the skills needed to maintain long ties, as mentioned in previous studies above (e.g. ability to maintain an unbalanced network in the case of Medici). Further evidence on the link between long ties and special skills to maintain them was offered in two studies that found stability in presence of long ties and that the best predictor of access to structural holes is the past access to them [35, 154, 180]. Recent evidence has also documented great variation in performance among brokers, with many receiving no benefit from their long ties [40]. This suggests that having access to long ties is not sufficient, and only those with the skills to take advantage of it have better job performance. Within the context of a firm, randomized experiments have shown that little network training to raise awareness of the social capital in brokerage and strategies to maintain these ties leads to positive causal impact on performance [37, 111]. If the training to take advantage of long ties improves outcomes, then events that enhances social capital skills should also improve outcomes.

Given the gap on our understanding of the necessary skills to maintain long ties, our second contribution is on the origins of long ties and the characteristics of individuals who hold them. In particular, we study several major life events and argue that their experience enables people to form long ties later on their lives. We discuss three major disruptions in life that expose individuals to multiple communities and require them to form new ties with individuals who are potentially very different from their previous contacts. We find that previous experience with these disruption is associated with more open and structurally diverse networks many years later.

Inter-State Migration

First, we consider the case of inter-state migration in the US and show that those who have migrated from a different state to their current state many years prior to the network measurement have higher frequency of long ties, which cannot be simply

explained by their geographic mobility itself. A fascinating recent study provided the rationale behind the study of migrants' experiences [36]. Burt and Merluzzi were interested in understanding why there is so much variation in performance of brokers and found that the benefits of long ties come from making deep investments in one community, followed by brokering between past communities, and then again followed by deep investments in another, rather than shallow brokerage between them simultaneously with minimal investment in all communities. They called this movement between periods of deep embeddedness and brokerage as network oscillation and argued that brokers with network oscillation have developed the skill for effective response to new developments in the surrounding environment [36]. Oscillation is the act of moving in and out of groups, much similar to the experience of migrants:

> "Experience with change is preparation for change... [Oscillators] can be expected to be flexible in moving between identities ... and so develop the adaptive self-monitoring associated with network brokers. The image that comes to mind is people who grew up in multiple countries, or in families that frequently moved between cities."

Oscillation and migration both create "adaptive response" skills to flexibly engage new opportunities in the ever changing environment and manage exogenous shocks.

Surprisingly, the literature at the intersection of networks and migration is very recent and so far narrow in scope. Most studies have attempted to characterize the changes in migrant networks in a short time scale and how migrants mobilize their networks as social support immediately after the move and during the adaptation period [124, 157, 160]. Since our goal is to evaluate the skills developed during the adaptation period, we must evaluate the long-term impacts of migration on network formation many years after the migration event. Two recent studies using Facebook data have described the long-term patterns of connectivity among immigrants and locals to shed light on assimilation and international connections [60, 95]. While these studies looked at migrant's networks potentially long after the move itself, they only focused on international immigrants who have a drastically different experience and

more assimilation challenges than inter-state migrants. Furthermore, they did not attempt to compare the local networks of migrants against non-migrants as we do here.

Our main finding is that migrants not only have more long ties that are geographically distributed, but also have more long ties in their current new state as compared to the local residents of the state. Perhaps, the closest to our finding is a study that examined the geographic dispersion of migrant connections [171]. Viry found that the farther migrants lived from their home location, the more geographically distant their contacts were from each other. As it relates to our goals, this finding suggests that mobility experiences are linked to individuals skills and resources with regard to the use of technology in maintaining strong long-distance ties in diverse places. While the result of this study generally agrees with our overall conclusion, it does not fully answer our questions because its focus was on dispersion of alters in ego-networks rather than their structural patterns, did not compare migrants versus non-migrants and did not discuss long-term implications of migrations in terms of skills and social capital.

Educational Relocation during Adolescence

The second and third experiences we consider involve major educational disruptions in adolescence years.

- We look for individuals who attended out-of-state college and compare their networks years after college with similar cohorts who attended college in their home state.
- We compare the network of individuals who attended a single high school against those who switched high schools within the same state, ruling out inter-state migration.

In both cases, we find that relocation for education during adolescence leads to structurally more diverse networks years later. The higher frequency of long ties is not simply due to the relocation itself, since the higher frequency of long ties persists even among friends made outside their high school or college. The reasoning why

relocation for schools leads to structurally more diverse networks is similar to the one we mentioned above for migration: both moves involve major disruptions and are preparation for future change. Attending multiple high schools is particularly interesting because it most likely happens due to reasons outside one's choice, hence less likely to suffer from confounders.

There is yet another reason why it is important to examine the experience of network formation during adolescence which has to do with shifting goals over the life-span according to the socioemotional selectivity theory [46–48]. The theory states that changing perspective across different life stages shift social goals and with them the social ties. Motivated by information acquisition goals and gaining independence from the parents, adolescence and young adulthood are the most expansive stages of life in terms of social networks. The focus of young adults is to gather as much information as possible from diverse sources, and as such they start forming large networks with diverse relationships and contacts from different groups. Empirical findings have supported the predictions of the theory and have documented how the personal network size start shrinking and becoming dense after young adulthood. Network formation strategies learned and adopted in this period are likely to persist throughout one's life. In fact, empirical evidence from college students have shown that, despite large differences in networking patterns across different students due to varying cognitive constraints akin to skills, social signatures of tie formation in a single student are very persistent even if there is high turn-over in the network [153] unless there are major (non-normative) life events which demand a change in the social signature [178].

Adolescence is the period individuals learn how to expand their social networks for the first time, so it seems important to understand how disruptions and experiences during this period impact the structure of networks later in life. This is the focus of our study and we will show that adolescents who undergo experiences that require them to form new ties in a new environment are likely to have more long ties later in life. The rest of this article is structured as followed. In section 3 we provide evidence on the economic implications of long ties within US zip codes. Later sections

discuss the determinants of long ties with the three events we mentioned above. In section 4, we show how migration leads to structurally more diverse networks. Subsequently, in sections 4.3 and 5, we show that relocations during adolescence also lead to structurally diverse networks.

3 Economic Implications of Long Ties

In this section, we present our results on the link between structural diversity, measured as the frequency of long ties, and economic outcomes. In this empirical study, we uses each US zip code as a unit of analysis and construct a network for each zip code corresponding to the communication patterns of its residents. We then look for any relationship between the overall network structure of each zip code and an array of economic indicators on that zip code provided by the census bureau. We show that zip code networks with frequent long ties, those without any mutual contacts acting as social bridges, tend to have better economic outcomes.

3.1 Data

There are two data sources in our analysis, one to construct the network corresponding to each zip code and another on its economic indicators.

Network Interactions: To construct the communication networks, we observe the number of times users commented on each other's posts over a 6 month period from December 2020 to June 2021. A commenting interaction involves one user making a comment on another user's original post or mentioning them on a separate thread. We only observe the time and count of the commenting activities and not their contents . Commenting activities are more appropriate for the construction of the network than the actual friendships on Facebook for two reasons: commenting contacts are more likely to be strong and relevant to each user, as opposed to many friends on Facebook who are only acquaintances without any meaningful interaction over months if not years. But more importantly, Facebook users extensively use commenting as a form of communication as there are about 195 million daily active users on the Facebook

website or its mobile application in US and Canada[1].

This would provide us with great coverage and high quality interaction data across many zip codes in the country. For each Facebook user, we also observe the predicted city of residence which relies on the user's information and activity, for example their self-reported city on their profile and their internet connection information. These residential predictions allow us to match users to approximate zip codes, counties, and states and construct the corresponding zip code network, consisting of users residing in the zip code, as we describe later.

Economic Indicators: The zip code outcome variables consists of a set of economic indicators obtained from the census 5-year estimates of the 2018 American Community Survey at the level of zip code tabulation areas (ZCTA). The economic indicators we consider here are mean household income, fraction of households with income below $25K, and the unemployment rate. In addition to these census provided indicators, we also consider measures of social mobility provided by the Atlas of Opportunity project at the ZCTA level [59]. The social mobility indicator we explore here are 1. the probability of a child born in the zip code and from the bottom 25% of the income distribution in the US reaching the top 20% of the income distribution and 2. the percentile income rank of a child born in the zip code from the bottom 25% of the income distribution. Finally, as a robustness check we replicate our analysis in Mexico as a second country. The outcome variable we use in Mexico zip codes is called the "wealth index" which is a summary measure on the intensity of economic, health and education prosperity, provided by the National Council of Population in the Mexican government, which ranges between 0 and 1.

3.2 Methods

Construction of zip code Networks: We generate each zip code network in a manner similar to ego-networks. The nodes in each zip code network consists of all users who reside in that zip code and all their contacts. The network contains all

[1]Figures are from the latest earnings release `https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Second-Quarter-2021-Results/`.

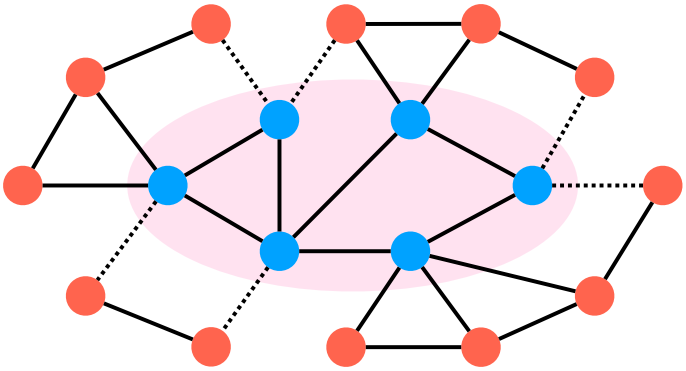


Figure 2.1: A schematic representation of the a zip code network. Blue nodes correspond to individuals residing in the zip code, the purple shaded area indicates the zip code boundary, and the red nodes are contacts who reside in other zip codes. Our analysis only consider ties between a node inside and one outside the zip code. The long ties among such ties have a dotted pattern.

the edges among these nodes, which most certainly includes the edges among nodes who reside outside the zip code but have a connection to at least one user in the zip code. Edges are directed with weights that correspond to the number of directed communication events from December 2020 to June 2021. Even though the edges are directed, we only consider reciprocal ones to filter out one-off communication events that don't signify a meaningful relationship. Thus, the existence of an edge in the zip code network means there has been at least one communication event in each direction. Figure 2.1 shows a schematic view of the zip code networks. In our analysis, we only consider zip codes with at least 100 users who reside in that zip code and for which we have census data available. This leaves us with 27,836 out of about 30,000 zip codes in the US (which are not PO Boxes or businesses) for which we have both network and economic indicator data. Figure 2.1 shows the histogram of some basic characteristics of these networks, along with two economic indicators.

Measuring Structural Diversity in zip code Networks: Structural diversity is a purely topological measure and captures the unique number of communities a node is involved in [32]. It can be operationalized in different ways, but at its core it relies on measuring the extent to which contacts are clustered together. The local clustering coefficient, count of bridges or the connected components among alters in an ego-network are some of the widely used measures [29, 161, 166]. All of these measures

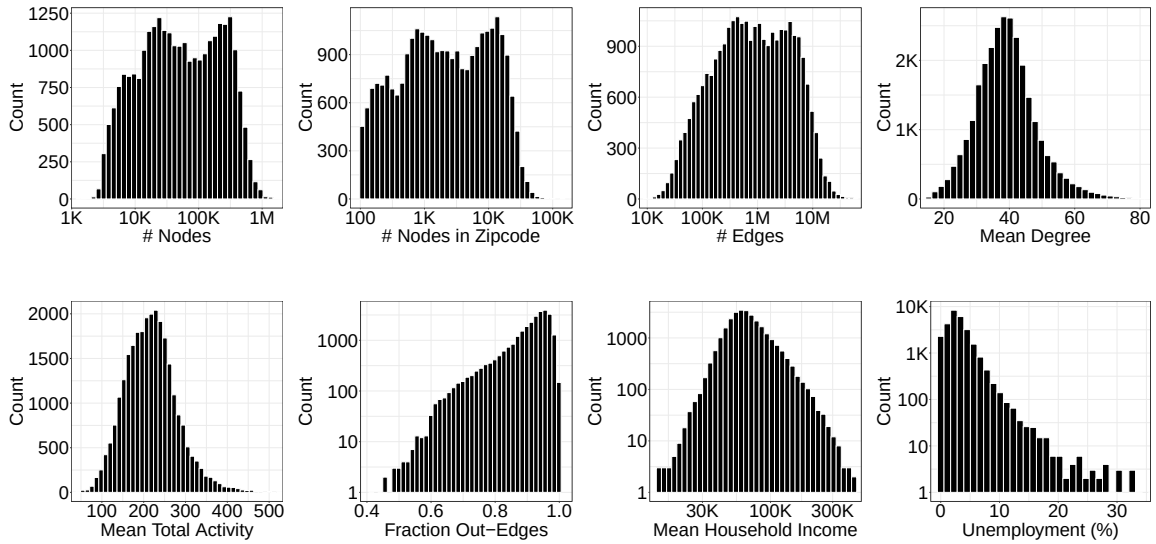


Figure 2.2: Histogram of zip code networks basic metrics, along with their household income and unemployment rate from census. Mean degree/total activity refers to average degree of/number of comments sent by nodes inside the zip code. Fraction out-edges is the fraction of edges originating from a node inside the zip code that are to outside the zip code.

are correlated with each other because they can all be expressed as a function of the number of mutual contacts on each edge of the the network.

The frequency of long ties, those without any mutual contacts, is a very simple such function. The long ties closely resemble the widely used concept of structural holes or social bridges and their overall frequency indicates the extent of novel information received by users residing inside the zip code. We use the fraction of long ties among all edges that involve one node inside and one node outside the zip code as our primary measure of zip code structural diversity as defined below. These ties are shown with dotted lines in figure 2.1.

$$l_z = \frac{\sum_{i \in Z} \sum_{j \in N_i} I(mc_{ij} = 0)}{\sum_{i \in Z} |N_i|} \tag{2.1}$$

where Z is the set of nodes residing inside the zip code, N_i is the set of i's neighbors who reside outside the zip code, $I(.)$ is the indicator function and $mc_{ij} = \sum_{k \in N_i} I_{jk}$ is the number of mutual contacts between i and j among nodes outside the zip code.

The clustering coefficient, the fraction of closed triads in the network, is another

widely used measure for ego-networks. Our zip code networks are very similar to ego-networks, with the exception that they have multiple nodes who reside in the zip code akin to egos. To incorporate multiple egos in the measure, we count all triads that involve one node inside and two nodes outside the zip code and discard triads that involve three nodes outside the zip code or more than one node inside the zip code. This approach is similar to the way triads are counted in ego-networks with an important distinction that we now have triads that involve more than one ego (residing inside the zip code). We don't count these triads for three reasons. First, we must ensure all triads are counted only once because if we were to count triads that involve multiple nodes in the zip code, they would be counted once from the perspective of each node in the zip code, especially in the weighted measures of clustering coefficient. Second, the edges (or triads) that involve more than one node in the zip code are a very small fraction of all edges (triads) as also shown in figure 2.2. Third, only the triads that involve at least one node outside the zip code are relevant considering the acquisition of novel information because the information circulating among triads inside the zip code is most likely redundant. With this explanation, we define our clustering coefficient measure as below.

$$c_z = \frac{\sum_{i \in Z} \sum_{j \in N_i} m c_{ij}}{\sum_{i \in Z} |N_i|(|N_i| - 1)} \tag{2.2}$$

Both l_i and c_i range between 0 and 1, but as opposed to the fraction of long ties, smaller values of clustering coefficient indicate more structural diversity. Our results are very similar using either measures, so here we only present the results showing the link between the fraction of long ties and economic outcomes.

Our second set of results examine the link between economic outcomes and the strength of long ties conditioned on a fixed number of long ties. To do so, first we define the normalized count of exchanged comments as our measure of tie strength that is comparable across users with different levels of activity.

$$w_{ij} = \frac{t_{ij}}{t_i} = \frac{t_{ij}}{\sum_{j \in N_i} t_{ij}} \tag{2.3}$$

In the definition above, w_{ij} represents the normalized tie strength from node i to j, t_{ij} is the number of comments i has sent to j in the 6 month measure period and t_i is the total activity of user i with their contacts N_i. Given our definition of tie strength, a simple and intuitive measure for the strength of long ties is their weighted sum as shown below.

$$l_z^w = \frac{\sum_{i \in Z} \sum_{j \in N_i} w_{ij} I(mc_{ij} = 0)}{|Z|} \tag{2.4}$$

Since the measure of tie strength is normalized, l_z^w ranges between 0 and 1 and can be treated as *weighted fraction of long ties*. Other measures, for example the weighted clustering coefficient, that capture the strength of long ties can be adapted from the unweighted measure [19] but we only focus on the weighted fraction of long ties in the results section.

Control Variables: There is a lot of variation in the size of zip code networks in ways that can influence the range of reasonable values for l_z and affect its statistical relationship with the outcomes. Thus, we ensure to control for the following variables in our analysis: population of the zip code, number of network nodes residing in the zip code and number of edges from a node inside the zip code to one outside. All the economic indicators we consider are highly correlated with racial composition of the zip code. Thus to ensure our network measure explain any outcome variation above and beyond racial composition, we also control for fraction of white race in each zip code. The data from Mexico does not include zip code-level racial composition, so the model for Mexico simply controls for the network covariates mentioned above.

Our second goal is to establish the relationship between the strength of long ties and the economic outcomes given a fixed number of long ties. Since l_z^w is highly correlated with the unweighted measure of structural diversity, l_z, our analysis must control for it. Furthermore, since our normalized tie strength measure is sensitive

to the total level of activity, we attempt to account for it with the average level of activity from users inside the zip code to those outside, $t_z = \sum_{i \in Z} t_i / |Z|$. Thus for our second analysis, we include the the fraction of long ties, l_w, and the average user activity, t_z, as extra control variables in addition to those mentioned above.

Estimation Procedure: Traditional regression based methods require a fixed functional form on the relationship between the zip code outcome and the fraction or the strength of long ties which might not be flexible enough to discover any potential non-linearity in the data. Binscatter methods and in particular, the recent binsreg implementation [51], provide the required flexibility to avoid a fixed functional form in the regression. In addition to visualizing the data with confidence intervals at each bin, binscatter provides a principled way to test for shape restrictions and most importantly adjust for our control variables such that it does not suffer from problems of residual-based approaches [50].

Binscatter regression with covariate adjustment fits the data according to a linear model. In order to avoid any assumption on the functional form of the control variables, they are binned and the model includes each bin as a separate dummy. For example, the model on the relationship between the a zip code outcome and the fraction of long ties is formulated as below.

$$\begin{gathered} y_z = \mu(l_z) + \boldsymbol{\alpha p_z} + \boldsymbol{\beta n_z} + \boldsymbol{\gamma e_z} + \boldsymbol{\theta r_z} + \epsilon_z \\ E[\epsilon_z | l_z, p_z, n_z, e_z, r_z] = 0 \end{gathered} \tag{2.5}$$

where y_z is an outcome measure (e.g. zip code mean household income), l_z is the fraction of long ties, $\boldsymbol{p_z}$ is the zip code population bin, $\boldsymbol{n_z}$ is the binned number of nodes inside the zip code, $\boldsymbol{e_z}$ is the binned number of edges from inside to outside the zip code, and $\boldsymbol{r_z}$ is the fraction of white race in the zip code. Binscatter regression discovers the shape of function $\mu(.)$, and computes the standard errors of each bin. Overall the model above includes 70 dummy controls.

We employ a similar approach for the model on the relationship between zip code outcomes and the strength of long ties, with the additional controls for the frequency

of long ties and the total activity as formulated below.

$$y_z = \mu'(l_z^w) + \boldsymbol{\delta l_z} + \boldsymbol{\eta t_z} + \boldsymbol{\alpha p_z} + \boldsymbol{\beta n_z} + \boldsymbol{\gamma e_z} + \boldsymbol{\theta r_z} + \epsilon_z$$
$$E[\epsilon_z | l_z^w, l_z, t_z, p_z, n_z, e_z, r_z] = 0 \tag{2.6}$$

where l_z^w is the weighted fraction of long ties, $\boldsymbol{l_z}$ is the binned unweighted fraction of long ties and $\boldsymbol{t_z}$ is the average activity level by users inside the zip code to outside. The model above include 100 dummy controls.

Non-parametric Alternatives While our binscatter regression model with binned controls is flexible enough to capture different functional forms, it nevertheless imposes a linear structure on covariates and our binning could be too coarse. Even more importantly, the ideal model as opposed to the ones in equations 2.5 and 2.6 would include the binned control variables in interactions terms. However, our small data size does not provide us with such a flexibility, hence the simple additive models above. Nevertheless as a robustness check, we employ two fully non-parametric approaches from the machine learning literature. First, we train a random forest model on l_z (and a separate model on l_z^w) and the respective control variables as above, but in their original scales, to predict y_z. Given this trained model, we generate the partial dependence plot (PDP) and accumulated local effect (ALE) plot of y_z on l_z (or l_z^w) [9, 93]. Both of these methods provide a visual relationship between two variables after adjusting for other covariate in a fully non-parametric fashion using random forest. The partial dependence plot resembles the stratified mean of y_z at different values of l_z (or l_z^w) but with fine-grained control variables defining the strata. The goal of ALE plot is the same as PDP, and it produces a similar visualization while incorporating the correlation structure of the covariates. Our results based on PDP and ALE on the relationship between economic outcomes and fraction of long ties or their strength both match our findings with the binscatter regression, and they are presented in the appendix.

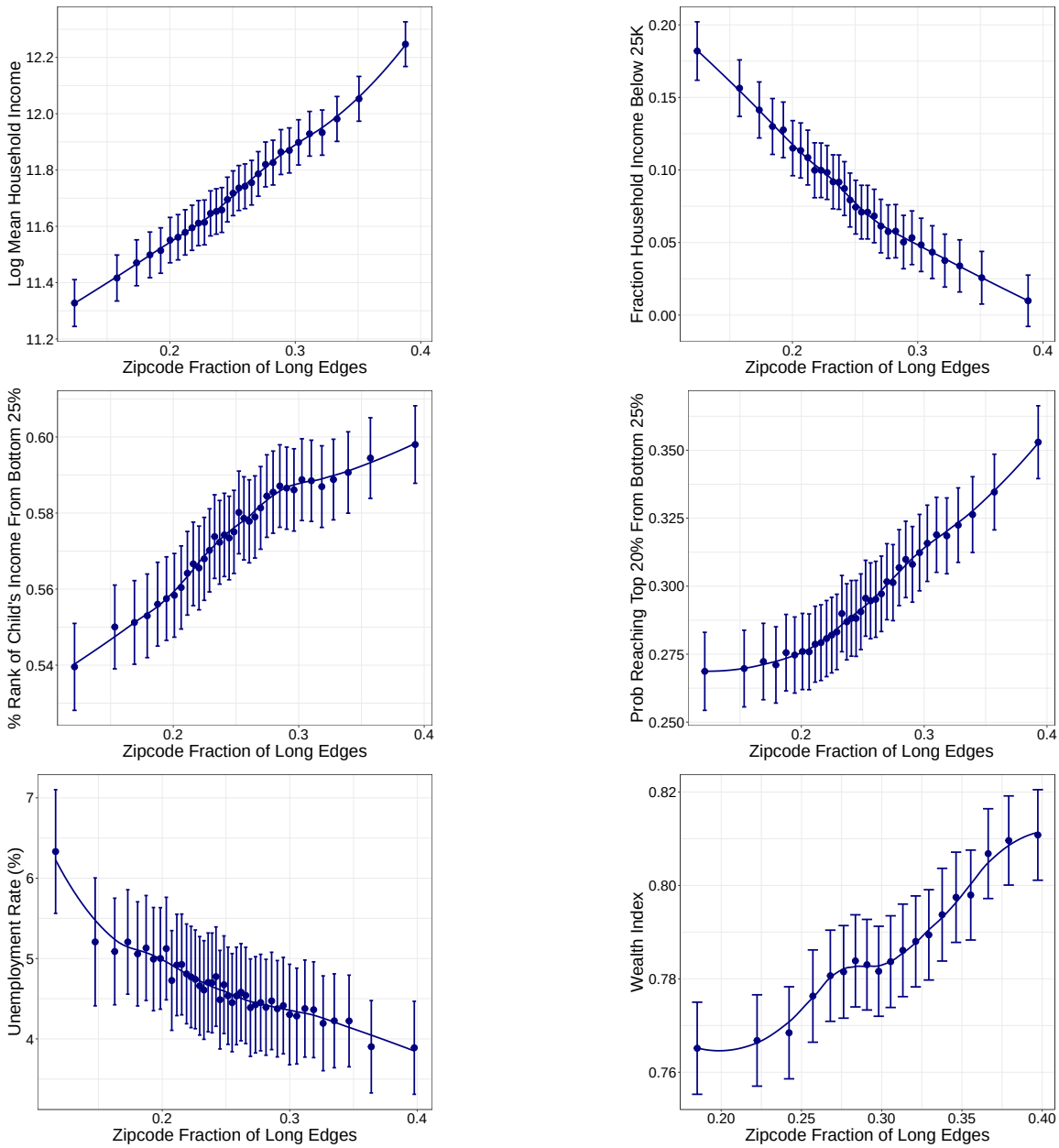


Figure 2.3: Frequency of long ties in the network and zip code outcomes. The binned regression plots are generated according to the model (2.5). With the exception of wealth index from Mexico zip codes, all other plots are based on US zip codes. Solid lines and bars correspond to local smoothers of second degree and 95% confidence intervals respectively.

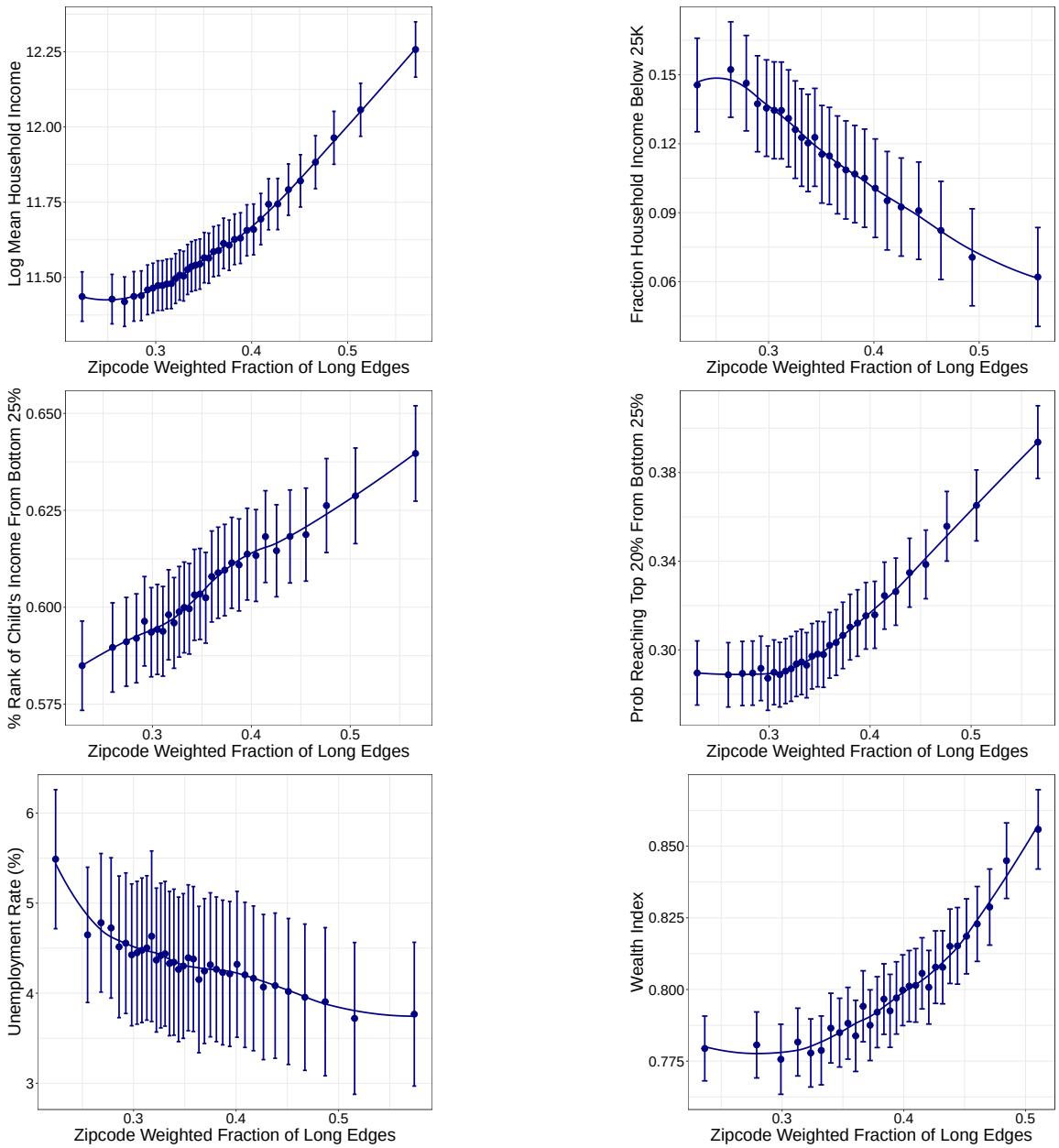


Figure 2.4: The strength of long ties in the network and zip code outcomes. The binned regression plots are generated according to the model (2.6). With the exception of wealth index from Mexico zip codes, all other plots are based on US zip codes. Solid lines and bars correspond to local smoother of second degree and 95% confidence interval respectively.

3.3 Results

How are the networks of zip codes with better outcomes different from those with worse outcomes? Do the zip codes with good outcomes have more long and stronger long ties in their networks? The binned regression plots in figure 2.3 represent the function $\mu(.)$ as defined in the model (2.5) using the fraction of long ties as the independent variable against a number of socioeconomic outcomes. The findings suggest that zip codes with more frequent long ties to outside the zip code tend to have better outcomes along a number of dimensions: household income, unemployment and social mobility.

The second set of results involves the strength of long ties. Conditioned on a fixed frequency of long ties, how is their overall strength linked with economic outcomes? This question is related to the strength of weak tie theory [89] which states that access to weak tie is beneficial due to its high information value. Figure 2.4 shows how the outcomes vary as the long ties in a network become stronger. In contrast to the weak tie theory, we find that the marginal effect of tie strength is positive and toward better outcomes, a finding noted by others [83, 84]. In other words, it is not the weakness of the tie that matters for outcomes, rather it's the structural or informational novelty that determines the outcome and once that is accounted for, stronger ties to structurally diverse contacts tend to have better outcomes. The confidence intervals in figure 2.4 are much larger than those in 2.3. This is due to the high correlation between the independent variable, l_z^w, and its unweighted version, l_z, which has to be controlled for in the model.

Figure 2.5 provides a visual summary of the geographic distribution of strong long ties across the US. In contrast to previous analysis, the networks in figure 2.5 are constructed at the county level and the map shows the color-coded fraction long ties (l_z) among 3076 US counties. A surprising observation from the map is that while most urban areas, especially along the east coast, have high frequency of long ties, the rule is not always true. Several urban counties including Miami-Dade, Wayne (Detroit), Cook (Chicago), Harris (Houston), Bexar (San Antonio), Maricopa (Phoenix) and Los

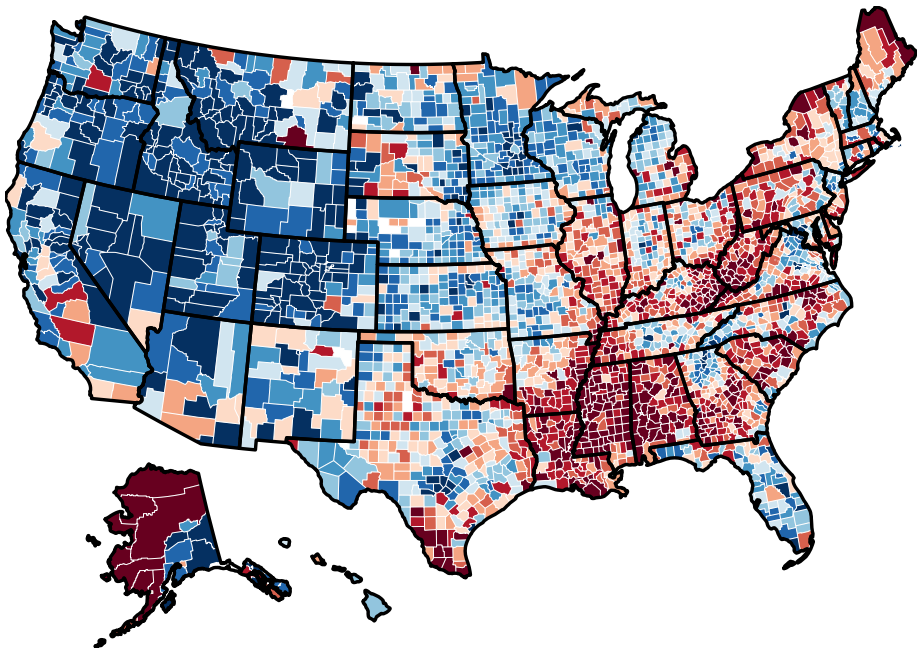


Figure 2.5: County-Level map of strong long ties. Red corresponds to the lowest and blue corresponds to the highest values of fraction of long ties in the county network as defined in equation (2.1). Counties with fewer than 200 active Facebook users or population less than 500 are shown in white.

Angeles county have low structural diversity despite being surrounded with suburban counties with high structural diversity. The second observation is that the counties in the Appalachia and the South seem to consistently have the least structurally diverse networks in the US.

4 The Determinants of Long Ties: Inter-State Migration in US

In the next 3 sections, we discuss 3 major disruptive events, as determinants of long ties and show that individuals who undergo these experiences tend to have structurally more diverse networks many years after the event. In all 3 cases, we compare individuals who have had the experience based on their self-reported profile with those who have not. In this section, we examine US-born individuals who have migrated from their home state to a new state prior to 2012 and since then have

resided in their current state, and compare them against individuals who have lived in their home state since 2012. We show that inter-state migrants not only have more long ties due to their geographic mobility, but also they have more long ties in their current state than the locals of that state.

4.1 Data and Methods

The current state of each user can be easily inferred based on Facebook's internal prediction of residential zip code as we explained in section 3.1. Given this data, we construct a monthly panel of predicted state location for each user for 9 years starting from January 2012. To determine whether a user has experienced inter-state migration, we also need information on their hometown state, for which we rely on the optionally listed hometown on the user's Facebook profile and ensure it can be matched against a known location. Overall, 72% of the active users in our data have provided this self-reported information on their profile. Given the monthly residence and hometown states, we restrict our attention on US-born users aged between 30 and 60 who have not migrated in the last 9 years: their monthly residence state has remained the same since January 2012. We then divide these users to migrants or those have a different hometown state than their current state and non-migrants with the same current and hometown states. Therefore in our analysis, migrants are users who moved to their current state prior to 2012 and have lived in their current state continuously for at least 9 years. Similarly, non-migrants are users who have resided in their home state since 2012.

For each user from either the migrant or the local group, we construct their ego-network from the same commenting communication data spanning a period of 6 months from December 2020 to June 2021, similar to what we used to construct zip code and county level networks in the previous section. Ego-network, often referred to as 1.5 level network, includes all the edges between the user (the ego) and their contacts (alters) and the edges among contacts themselves. The ego-network allows us to compute local measures of structural diversity as we discuss later. With web-based commenting, there is a directed edge between two users if one has replied to

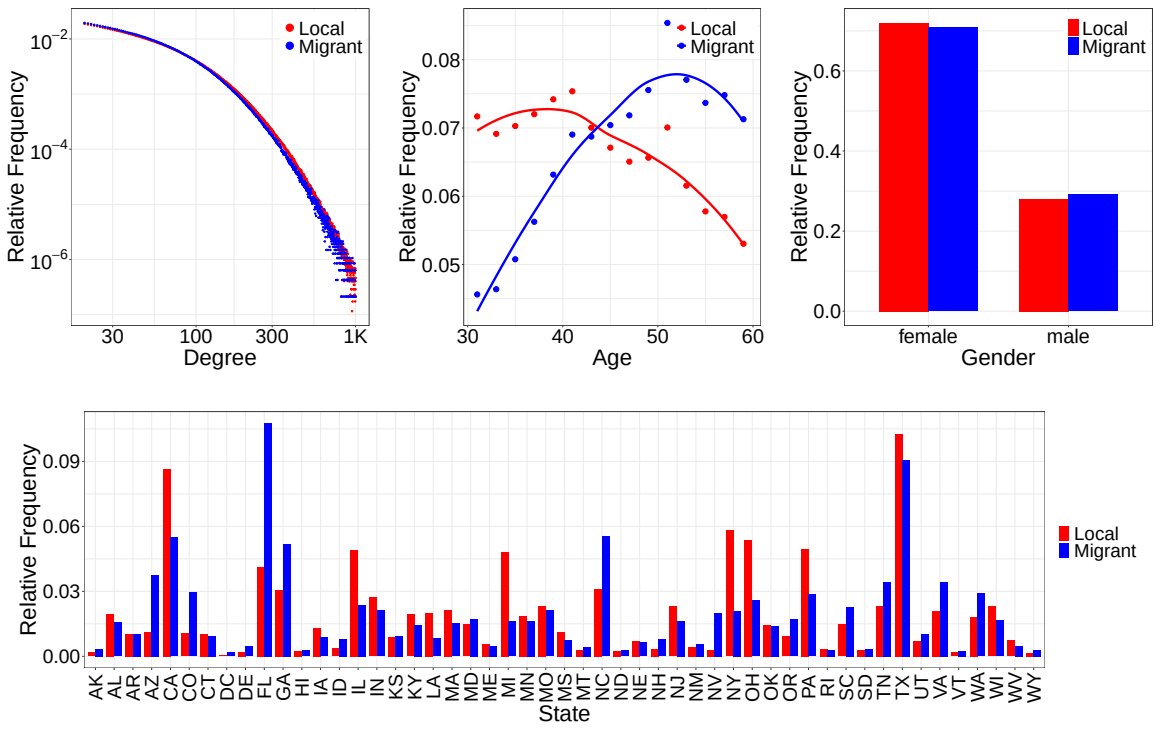


Figure 2.6: The relative histogram of degree, age, gender and current state by migration status. The frequencies sum to 1 within each migration group. The degree distribution is shown with a log-log scale. The solid line in the age distribution correspond to a LOESS smoother.

a post made by the other. In this process, we ensure a minimal level of activity by each user and discard those with degree less than 20 (have communicated with less than 20 people over the 6 month period). The final data consists of 22 million users, out of which 17.3 million users (79%) are local or non-migrant and 4.7 million users (21%) are migrant.

Our general approach is to compare the ego-network structural diversity between these two groups of users: migrants and locals. There are several ways to measure the ego-network structural diversity, but we follow our discussion from section 3.2 and use the fraction of long ties among ego-alter ties to measure structural diversity. The range of common values for the fraction of long ties depends on the degree of the node, so in our analysis we compare the mean fraction of long ties among migrants and locals within each degree bin (with a total of 30 equal-size degree bins). A simple z-test within each degree bin, however, does not account for differences between migrant

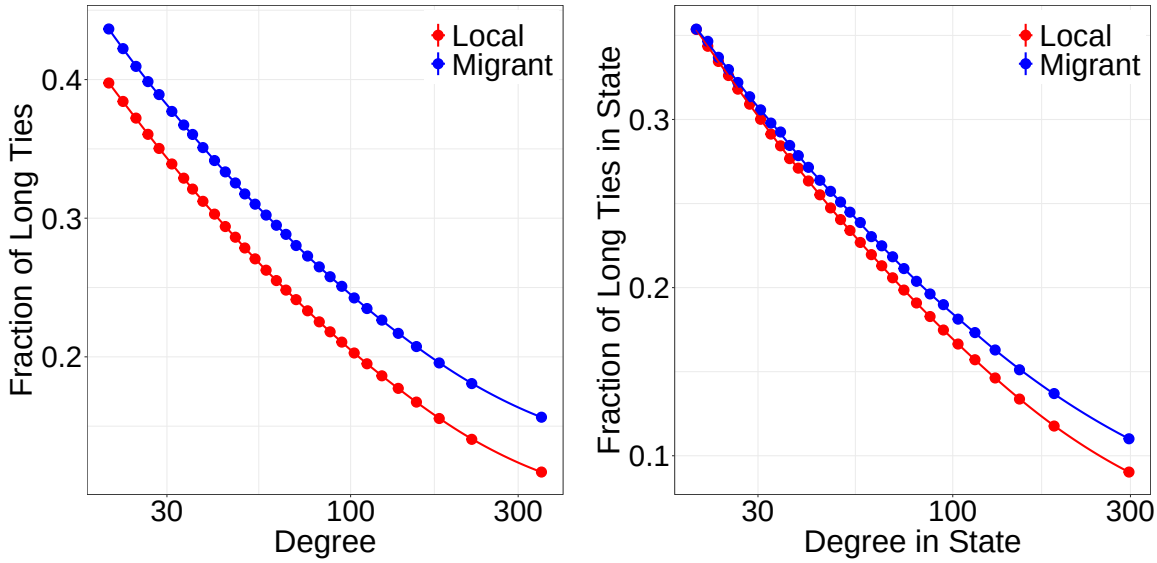


Figure 2.7: The stratified fraction of long ties over all ties (left) and within the ties in the current state (right) by migration status. The combination of gender, age and current state bins constitute the stratification strata. Both plots contain 95% confidence intervals around each point estimate, but the intervals are too small to be visible. In the right panel, the fraction of within state long ties among migrants is larger than locals in all degrees.

and local populations. Thus, in our analysis we attempt to control for some basic user characteristics (age, gender and current state) by computing the post-stratified mean of the fraction of long ties within each degree bin. In our post-stratification, the combination of binned age (3 bins), gender (2 bins), and current state (51 states including the district of Columbia) constitute a stratum (a total of 306 strata). Figure 2.6 shows the distribution of these control variables along with degree within both migrant and local populations. The main observation from figure 2.6 is that while migrant and local populations have the similar degree and gender distribution, they vary significantly by their age and current state, hence the need for stratification.

4.2 Results

Figure 2.7 presents our main results comparing the fraction of long ties between the migrant and local population. The left panel shows that given a fixed degree, migrants on average have about 4.0% more long ties than locals, with a larger difference in

higher degrees. Geographic mobility, rather than a skill, is a simple explanation behind the higher frequency of long ties among migrants than locals. However, as we observe in the right panel, migrants have more long ties than locals even when one only considers the ties to alters who reside in their current state. Migrants on average have about 1.1% more long ties among their in-state contacts than the locals. While the difference in the frequency of in-state long ties is smaller than the same frequency among all ties, migrants nevertheless have consistently more (statistically significant) long ties than locals in all degree bins. This suggests the higher prevalence of long ties among migrants is not simply due to their geographic mobility.

4.3 The Determinants of Long Ties: Out-of-state College Attendance

In this section, we discuss the experience of another disruption, out-of-state college attendance, and its link to network structure years later. In particular, we examine US-born Facebook users who attended a college outside their hometown state and compare them with the population of users who attended a college inside their home state. We show that out-of-state college attendance is associated with higher frequency of long ties years later after the college, and this relationship persists if one looks at friends made out of college or account for the migration effects of out-of-state college attendance.

4.4 Data and Methods

We obtain college attendance data on each user based on their self-reported higher education institution. Overall, more than 47% of the active users in our data have attended a college and have optionally reported the name of the higher education institute attended in their Facebook profile. The self-reported information is generally trustworthy, since we have resolved them against known higher education institutions. Our analysis will examine the subset of users aged between 30 and 60 who have attended at least one college, and divide them into two groups of in-state and out-

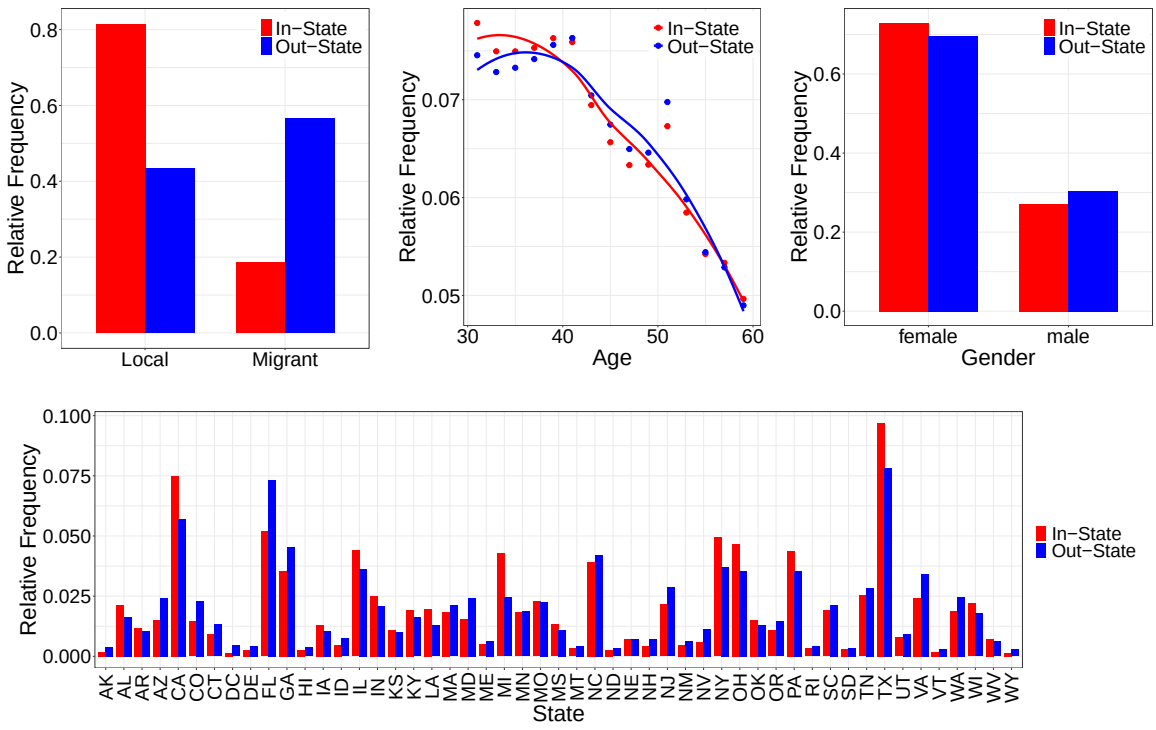


Figure 2.8: The relative histogram of migration status, age, gender and current state by location of college attendance. The frequencies sum to 1 within each college attendance group. The solid line in the age distribution correspond to a LOESS smoother.

of-state college attendance. Comparing the state where the user attended a college with their hometown state, which is also self-reported as explained in section 4.1, we can determine whether at least one attended college is outside the user home state. The users with at least one such college constitute our out-of-state population and the rest are considered in-state. In a manner similar to section 4.1, we match each individual in our population with their 6 month communication ego-network and discard individuals with low levels of activity (degree less than 20 over the 6 month period). The final data consists of 19.0 million users, out of which 12.6 million (66%) attended only colleges within their home state and 6.4 million (34%) attended at least one out-of-state college.

In order to compare users by the location of college attendance, we employ a similar approach as explained in section 4.1. We compute the post-stratified mean of the fraction of long ties in each degree bin and for each college attendance group.

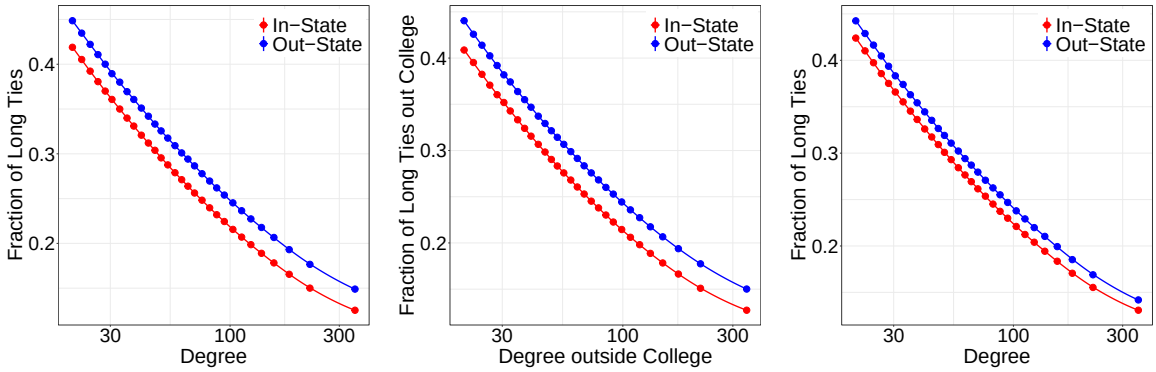


Figure 2.9: The stratified fraction of long ties over all ties (left) and within the ties who did not attend the same college as the user (middle) by location of college attendance. The combination of gender, age, and current state bins constitute the stratification strata for the left and middle plots. The right plot also controls for inter-state migration status by including it in the strata definition. All plots contain 95% confidence intervals around each point estimate, but the intervals are too small to be visible.

All combinations of gender, age and current state constitute the 306 strata of the analysis. Figure 2.8 shows the distribution of these control variables for both groups. The figure does not include the degree distributions since it very similar to the one shown in figure 2.6 without any noticeable difference between the two groups. Any difference in frequency of long ties between in-state and out-of-state user groups could partly be attributed to the effect of migration as users who attended out-of-state college are more likely to permanently move out of their home state, as verified in figure 2.8. However, it is not clear whether one should control for migration status by including it as a stratum, since the two mechanisms of out-of-state college attendance and inter-state migration only partly overlap and even if they were not distinct, it is difficult to separate them. Nevertheless in our secondary analysis, we include the binary migration status as defined in the previous section 4 as another stratum in the post-stratification, which leads to 712 total strata. We will present results comparing the two groups controlling for either stratum definition.

4.5 Results

Figure 2.9 compares the fraction of long ties between users by location of college attendance, using two different stratum definition. The left plot compare users in terms of their overall long ties frequency. Similar to the case of inter-state migration, the higher frequency of long ties might be due to exposure to college friends from a different state, rather than an acquired skill. However, the middle plot shows that users with out-of-state college attendance have more long ties even among contacts who did not attend any of their colleges. Given a fixed degree, the out-of-state group has about 3.0% more long ties than in-state users, in both the left and middle plots, suggesting that their current higher likelihood to have long ties is not directly due to connections with diverse college friends in the past. Out-of-state users are much more likely to be inter-state migrants a shown in figure 2.8, hence the right plot attempts to control for the migration status, even though this might underestimate the effect of out-of-state college attendance. Accounting for migration status reduces the difference between the two groups from 3.0% to 1.9% on average, however users with the experience of out-of-state college have consistently more (statistically significant) long ties in all degree bins. This suggests the higher structural diversity in the out-of-state group is not solely due to inter-state migration.

5 The Determinants of Long Ties: Multiple High schools

In this section, we discuss the experience of attending multiple high schools as yet another major disruption that is associated with higher frequency of long ties later in life. The population is the set of active Facebook users who attended high school in the US and currently live in the US. This case study enjoys more validity as we ensure attending multiple high schools is not accompanied by an inter-state move: we compare individuals who attended multiple high schools in the same state versus those who attended only one. By doing so, we discard individuals who attended multiple

high schools due to inter-state moves and rule out any possibility that migration is the driving force behind any effects we find. We show that attending multiple high schools is associated with higher frequency of long ties years later after high school, and that this relationship is robust even if we focus our analysis only on ties formed outside high school.

5.1 Data and Methods

Similar to our analysis of out-of-state college attendance in section 4.3, we obtain high school attendance data on each user based on the self-reported high school name on their profiles. About 65% of the users we examined have reported their high school information. Our analysis is based on users aged between 30 and 60 who currently live in the US and have attended their high schools in the US and all within a single state. We divide this population of users to two groups of single or multiple high schools and join each user with their ego-network constructed from communication data over a period of 6 months from December 2020 to to June 2021, dropping any user whose degree is less than 20. The final population consists of 27.3 million users, out which 25.2 million (92%) attended a single high school and 2.1 million (8%) attended multiple high schools all within the a single state.

To compare the frequency of long ties across these two groups, we compute the post-stratified mean of the fraction of long ties in each degree bin and high school attendance group. Combinations of gender, age, and current state create 306 strata of the post-stratification. Figure 2.10 illustrates the distribution of these control variables within each high school group. The figure does not include degree distribution as it holds a similar shape between the two groups. The remarkable observation from figure 2.10 is that the two high school groups do not differ in noticeable ways across any of these variables, in particular the current state of residence. This gives us some confidence that any difference between the networks of these two groups is less likely to be confounded by a third variable compared to the previous two case studies. Our analysis does not include migration status in the strata definition for two reasons: first, as we see in figure 2.10, there is no noticeable difference between the two high

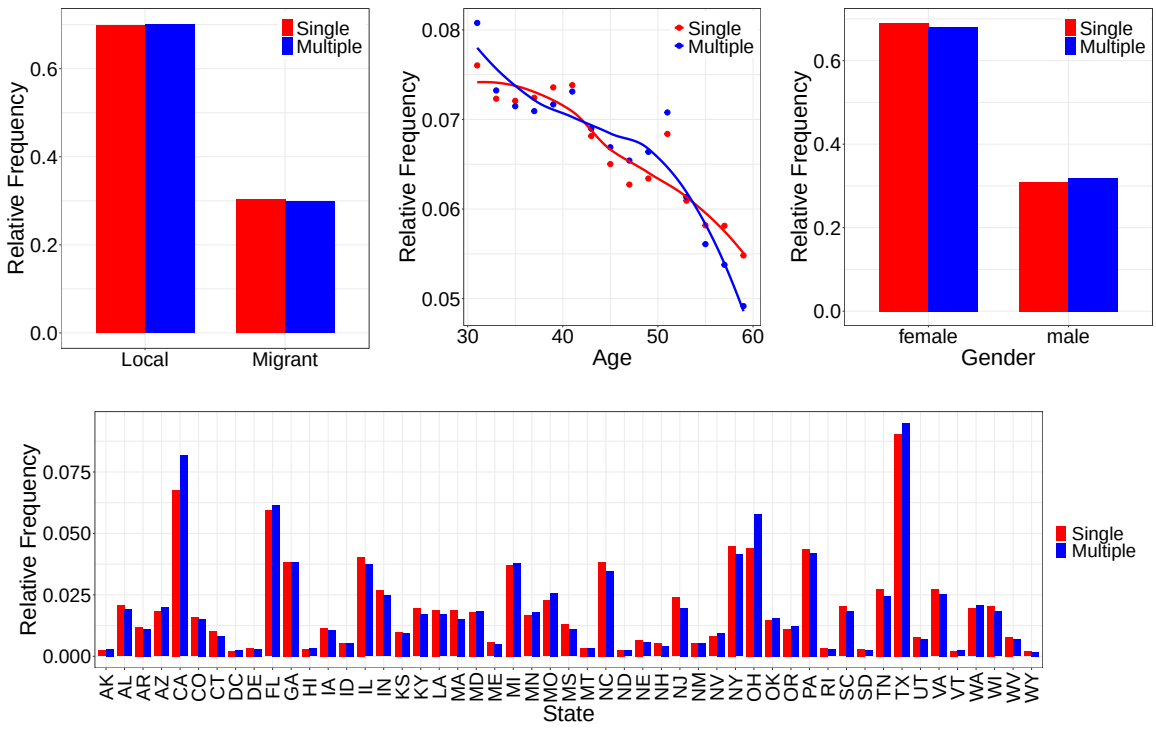


Figure 2.10: The relative histogram of migration status, age, gender and current state by the number of high schools attended. The frequencies sum to 1 within each college attendance group. The solid line in the age distribution correspond to a LOESS smoother.

school groups in terms of migration. Second, our analysis ensures neither group experiences an inter-state migration event during high school. Similar to the two previous case studies, our analysis compares the mean fraction of long ties between the two groups at each degree bin.

5.2 Results

Figure 2.11 compares the fraction of long ties between users who attended a single high school and those who attended multiple but in the same state. The left plot compares the two groups in terms of their overall long ties frequency. We observer that users who attended multiple high schools consistently have more long ties within all degree brackets, which translates to about 2.0% difference between the groups on average. However and just like the previous two case studies, the higher prevalence of long ties in the multiple high schools group might be due to exposure to high

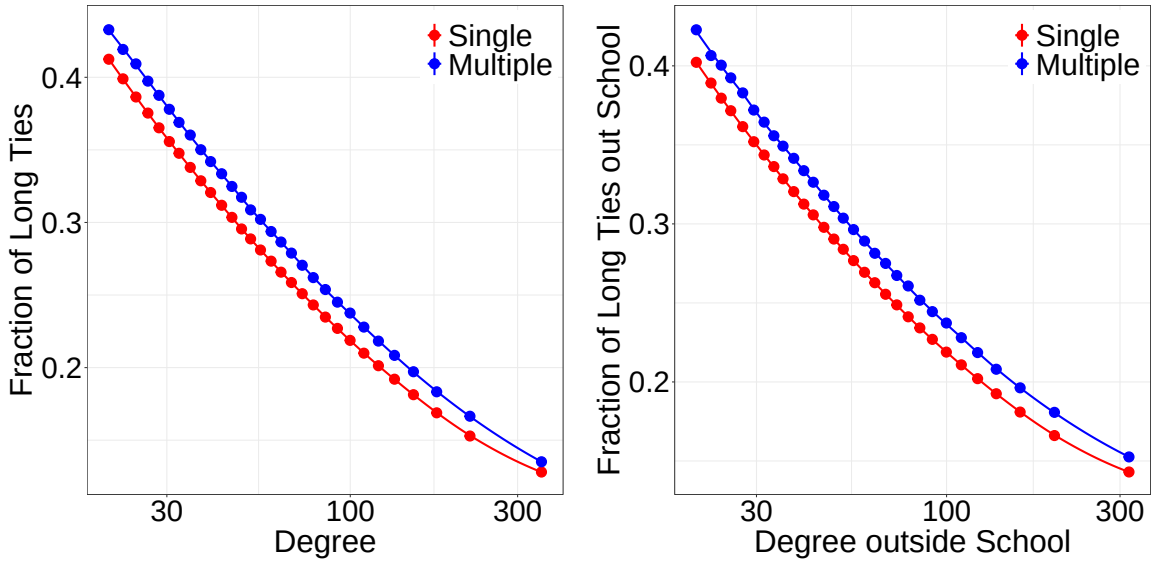


Figure 2.11: The stratified fraction of long ties over all ties (left) and within the ties who did not attend the same high school as the user (right) by the number of high schools attended. The combination of gender, age and current state bins constitute the stratification strata. Both plots contain 95% confidence intervals around each point estimate, but the intervals are too small to be visible.

school friends from different communities, rather than an acquired skill. The right plot addresses this concern by repeating the analysis only among contacts who did not attend any of the high schools the user attended. Given a fixed degree, the multiple high school group still has about 2.0% more long ties among contacts made outside high school. These results indicate that individuals who attended multiple high schools are more likely to develop long ties in ways that is not due to connections made with diverse set of high school contacts.

6 Conclusion

There is extensive literature on how social capital is manifested in the structure of social networks [91, 106, 110]. Many case studies, often with narrow contexts, have documented how structure of ego-networks can affect performance of individuals [39]. Most evidence has focused on managerial performance in firms and shown actors with access to diverse communities through structural holes tend to get paid more and promoted faster [11, 34, 94, 145–147]. The underlying argument behind these findings

is that certain networks, in particular those with access to structural holes, benefit from structural diversity or access to multiple independent communities. Structural diversity leads to better performance through multiple mechanisms such as arbitrage, autonomy and competition [147]. But the main mechanism is one of information advantage and innovation which relies on the observation that each community in a structurally diverse network is a source of novel information and opportunities that is not available through any other community [166]. Information advantage often allows the individual to have the right information at the right time and act fast when a new opportunity such as employment arises [92]. Structural diversity also leads to more innovative teams which can search for and combine information from multiple sources during problem solving or product development [100, 177].

Despite its important implication, the effect of structural diversity on outcomes is rarely studied outside the firm context mainly due to data limitations. The link between structural diversity and important economic indicators such as income or social mobility is yet to be established. Such a link, if causal, would lead to important implications for policy makers as it adds yet another policy tool to combat poverty and generational immbolity. We attempt to fill this gap by providing descriptive evidence on the relationship between network structural diversity and economic outcomes at the aggregate scale of US zip codes. In this study, we focus on a specific operationalizaiton of structural diversity, namely the frequency of long ties in a network. A long tie is a link between two nodes in a network who don't share any mutual contact, making it likely for the edge to be a bridge between two communities. Using Facebook communication data among users residing in the US, we show that zip codes whose networks have an abundance of long ties tend to have better outcomes in an several economic indicators, such as income, social mobility, poverty and unemployment. Furthermore, we show that conditioned on structural diversity, as long ties become stronger, the zip code outcomes get even better. These findings suggest zip codes with the best economic outcomes in the US tend to be those with an abundance of strong long ties in their networks.

Given the importance of long ties on economic prosperity, a natural question is

where do long ties come from and why do some people have more long ties than others? Despite the vast literature on their performance implication, there has been limited focus on this question and the determinants of long ties are largely unknown. Our second contribution attempts to study the origins of long ties and factors that enable people to create and maintain these ties. The seminal work by Simmel suggests that individuals with long ties viewed as mediators have the unique ability to resolve potential conflicts between disconnected communities [158]. The origins of this ability is largely unknown and the empirical evidence is lacking. In contrast, the existing literature on characteristics of long ties has approached this question with a psychological lens and has characterized the personality of individuals with long ties [38, 113, 164]. These studies confirm that people with access to long ties have an entrepreneurial personality, enjoy disruptions and risk-taking, value authority, and thrive on change versus stability.

The studies mentioned above treated the determinants of long ties as a matter of personality, an innate ability, rather than a skill to be acquired. In particular, the existing literature does not examine the processes that lead to these personal characteristics, even though many of the personality correlates of long ties as discovered by existing studies [38, 113, 164], such as risk-taking or proclivity to change, are not innate and can develop through disruptions and experiences. We consider three major life events that nourish such characteristics and examine whether individuals who experienced these events have structurally more diverse networks years later than those who do not. The three disruptions are experience for future change; they expose individuals to diverse communities and require them to form new ties with new contacts who are potentially very different from their previous contacts. These cases are inter-state migration, out-of-state college attendance and attending multiple high schools, excluding international migrants. In all three cases, we find that previous experience with any of these disruptions is associated with less clustered and structurally more diverse networks many years later, in ways that is not simply due to geographic mobility or exposure to a new community during high school or college. Our findings suggest that access to long ties is not only a matter of personality but

also the appropriate skill to see oneself as part of multiple groups rather than strict categorizations, as first argued by Simmel [158].

Our study has some limitations. Our evidence on the link between economic outcomes and frequency of long ties is simply correlational. We attempted to control for variables which could act as potential confounders, nevertheless we cannot make causal claims given our current methodology. However, we believe with the vast literature on the impact of long ties on the firm level outcomes, our findings are less likely to be completely due to endogeneities since our question is the extension of firm-level findings to the general economic outcomes outside of a firm.

Furthermore, our argument on the link between the three life events and the frequency of long ties is not a formally causal argument. Our findings could also be confounded with underlying differences between the comparison groups other than the disruption. However, we hope that our evidence on three different scenarios provide more validity to our argument. More importantly, we believe the case study on high school changes is less likely to be confounded with personal characteristics, thus resembling more like an exogenous shock, due to three reasons. Fist, none of the control variables show noticeable differences in their distribution between the the groups. Second, as opposed to inter-state migration and out-of-state college attendance, attending multiple high schools is not a decision one takes on their own, rather it is due to parents or other changes outside the control of the student. Third which also applies to other two case studies, we examine the networks that form many years after the experience, hence the link cannot be due to endogeneities that happen in a short temporal scale. Nevertheless, it is still very valuable to establish the causal link between these experience and network structure formally through natural experiments. Some potential natural experiments involve closure of high schools, inauguration of a local college or addition of a new major to a local college that reduces out-of-state college attendance. We leave these promising studies as future work.

We hope our study has illustrated the importance of network structure and in particular the social capital in long ties on economic prosperity. Our findings have

important implications for policy makers, if creating and maintaining long ties does indeed benefit from specific set of skills. This link can potentially provide another policy tool to empower people with the right social skills to create these beneficial ties, not only to fuel innovation but also to combat inequality. For this reason, we believe it is very important to continue this line of research and discover other determinants of long ties as it pertains to skills that can be taught.

Chapter 3

Unequal Diffusion in Networks and Differential Network Effects on Outcomes

1 Preface

In the previous chapter, we illustrated how networks are linked to economic outcomes. We argued that networks and in particular their structural diversity measured in terms of long ties frequency affect the level of novel resources that are accessible. In this chapter we argue that these network mechanisms could benefit different groups differently. The same network structure could provide different levels of utility to members of different groups. We provide observational evidence for differential network advantages in access to information: individuals from the low status group receive lower marginal benefits from networking compared to an individual from the high status group. We attribute this phenomena to differential diffusion mainly due to network homophily: high status individuals are mostly connected amongst each other, they hold the most valuable resources and network homophily ensures the resources remain exclusive. Thus, the marginal values of networking are lower than expected for low status individuals. We further provide causal evidence for this unequal diffusion in

the context of a randomized seeding experiment where a new behavior diffuses in the network. We show homophily amplifies an initial advantage in the seeded status of a group and determines which group adopts the new behavior at a higher rate.

2 Introduction

Networks play an important role in access to business opportunities [91, 122]. The foundational work by Granovetter [92] demonstrated that economic activity, and in particular job search, is embedded in informal social networks. Work by Lin [123] showed how the use of social resources in personal networks plays an important role in socioeconomic status attainment. Thus, the composition of the local network, the strength of the ties and in particular the structure of the network influence the access to high quality employment opportunities. It networks influence economic outcomes, it is fair to assume they can also lead to inequality by provision of unequal access to economic opportunities, such as job information [108]. The structure of the network determines how economic information diffuses throughout the network and whether or not it reaches different individuals. Homophily [129], or the tendency of individuals to link with other individuals from their own socio-demographic group, is a structural explanation behind unequal diffusion, since it ensures that a piece of information that originates within a subgroup stays within that subgroup. If members of one group, for example high status individuals, have a slight advantage in terms of economic opportunities, then homophily among members of this group will result in even larger differential advantages in terms of access to the economic information. The homophily among members of different groups will result in unequal diffusion of information among all members of the network, with the high status group having the network advantage.

Unequal diffusion of information through homophilous ties is one possible mechanism through which networks can exacerbate already existing inequality. This simple mechanism, which provides a structural explanation for network inequality, operates through unequal diffusion of economic information to sub-groups that generate them

at a higher rate. In particular, the mechanism assumes the social network, with two subgroups, has the following three characteristics:

1. One of the groups exogenously generates the information at a higher rate compared to the other group (e.g. high social class owns businesses and generates employment opportunities at higher rate).

2. The information diffuses through the network links. The effect of network inequality will exist no matter if the information is rivalrous or non-rivalrous, but should be stronger for rivalrous information.

3. The social network is homophilous along a characteristic that is correlated with the group attribute (e.g. business community, comprising the high social class, mostly attend certain colleges).

If the above 3 conditions are met, one of the groups will start with an initial advantage in terms of access to economic information. But this initial advantage will be exacerbated by the network, through homophilous links to other sources of information. Inspired by models of job search and employment [41, 92], this simple mechanism could result in higher employment rate for members of one group compared to others if they are tightly connected and start with an initial advantage in terms of exogenous job creation. In particular, if each member of the high status group is likely to generate an economic opportunity in each time period independently and identically at a fixed rate higher than the low status group, then this initial advantage grows even larger when we consider the faster rate at which information reaches the high status individuals from other members through the network. Effectively, individual advantages get multiplied by information arrival through the network, increasing the inter-group gap.

The possibility of network effects on inequality have already been discussed in the literature at a theoretical level [72]. These mechanism rely on nuanced processes on the network, such as social learning, higher trust, network externalities or normative pressure that influence adoption decisions about a new behavior. For example, DiMaggio and Garip discuss a model in which adoption decisions with positive

network externalities exhibit inter-group differences if the group characteristics are related to the adoption [71]. Perhaps, the closest work to ours is the model proposed by Calvó-Armengol and Jackson [41] which attempts to provide a modeling perspective on why participation in labor force is very different between whites and blacks [45]. Their work builds on the job search in networks literature and explicitly models access to job opportunities through network of social contacts. The persistent differences in employment status of agents arises from the simple fact that "the better the employment status of a given agent's connections (e.g. relatives, friends), the more likely it is that those connections will pass information concerning a job opening to the agent". This happens because the contacts are employed themselves and the employment information does not benefit them directly. The information passing process leads to correlation in employment status of subgroups of densely connected agents. Furthermore, the model predicts that the initial advantage of a subgroup translates into persistent advantage in employment status and future income for the whole group, since higher employment rates within a subgroup means new employment opportunities are generated more frequently in the subgroup and passed to very few unemployed individuals in the group through homophilous ties.

The processes mentioned above lead to differential network advantages in access to information, but they are not easily observable or measurable. Therefore, there have been very little empirical evidence investigating the existence and the strength of the phenomenon. In contrast to previous proposed mechanisms in [72] that rely on nuanced interactions in the network, the simple mechanism we investigate here is purely structural as it is solely a consequence of unequal diffusion throughout the network. Hence we can easily examine the level of its impact and susceptibility of a network to inequality by analyzing the network structure and the extent to which different subgroups receive the information. This property makes it easy to find causal evidence of the mechanism by tracking the diffusion of information or a behavior within randomized experiments in which random seeds are selected exogenously as the initial nodes with the initial advantage of the information or the behavior.

Our goal is to provide empirical evidence for this mechanism (either in observa-

tional data or in a randomized experimental setting) within a network. To this end, first we provide results from an observational data that suggests individuals from a high status group receive differential benefits from their network, solely through a correlational analysis. Using the call records from about 33,000 individuals in a south Asian country, we find that structural diversity, measured as the fraction of open triads in an ego-network, shows a relatively strong association with individual income. Furthermore, the effect of structural diversity is exclusive to the individuals from a high status group. These results provide suggestive evidence for the mechanism above that concentrated distribution of economic opportunities among the high status social strata combined with homophily among members of the same group leads to differential network advantages for the high status group, similar to the rich club effect. The result also suggest inadequate diffusion of economic opportunities to the low status social strata.

Second, we will use the data provided by a randomized experiments [140] that studied diffusion of a new behavior in a social network when a few initial nodes are randomly seeded with the information. In this study, authors seeded an anti-conflict intervention that was randomly assigned to initial seed students and evaluated the causal effect of its diffusion on conflict rate, both at the school level and individual student level. In the context of our study, we would like to show that the students that are similar to the initial seed students are more likely to adopt the new behavior. In other words, the effect of intervention in terms of adoption of the new behavior on students that are homophilous with the initial seed students is larger than non-homophilous students. In this context, the group to which the initial seed students belong constitutes the high status class, with access to the new behavior by the seeds acting as the initial advantage of the group. We will show that this initial advantage by one member of the group (seed student) leads to differential advantages in terms of adoption for other members of the group compared to non-members in the network.

The rest of this chapter is organized as followed. In section 3, we explain the context of the observational study and provide its results. In section 4, we provide causal evidence for unequal diffusion in the networks, using the data collected from

a previous study of randomized seeding in networks. We finish with some concluding thoughts in section 5.

3 Observational Study

The effect of informal ego-networks on job search can be explained by four mechanisms of employer, worker, and most importantly *contact and relational heterogeneity* [102]. The sociology literature on the effect of networks on economic outcomes has mostly focused on contact heterogeneity: the variation in endowments or the micro characteristics of contacts in the network, such as their education or gender, as different manifestations of social capital. For example, [127] looked at the number of unique occupations and the proportion of white males present among the contacts and its effect of job leads. Similarly, Elliott [76] investigates how race and neighborhood location determine the level of social isolation and consequently how insulated an individual is from the labor market. Lin, Vaughn and Ensel [123] look at outcomes in job referrals and report that the occupational status of the contact, as a measure of social resources, has a strong impact on the prestige of the obtained job.

In contrast to contact heterogeneity, the relational heterogeneity in an ego-network depends on the overall structure of the ego-network and whether individual contacts are themselves connected. The theoretical underpinning for the impact of relational variation on economic outcomes revolves around the Granovetter weak tie theory [89]. The strong ties are associated with dense networks and triadic closures and as a result exhibit high levels of information redundancy. In contrast, weak ties tend to be bridges to diverse communities across structural holes [39], hence have superior information novelty. Thus, a lot of subsequent works on the effect of relational heterogeneity have focused on the strength of the ties to information sources, rather than the local structure of the network. Furthermore, with the exception a recent work [125] that looked at the link between income and the centrality position in the global network, there has not been any studies investigating the association between income and structure of ego-networks. In this study, we examine the association between

income and overall ego-network structure, namely its *structural diversity*, as a measure of relational heterogeneity or structural redundancy, using about 33,000 surveyed individuals. It should be noted that any link we find between income and ego-network structure is simply suggestive as it is based on correlations with the potential of many endogeneities making a causal interpretation difficult.

We have three empirical contributions in this section. First, we examine the link between the structural diversity of ego-networks and economic outcomes. Structural diversity of an ego-network measures the level of information novelty among the contacts purely based on their connections to each other. Second, we use income as our measure of economic outcome instead of the prestige of the jobs obtained through informal referral [123] or other measures limited to labor-market outcomes. Finally, we provide evidence for the differential effects of structural diversity across different educational levels. We show that individuals with high educational status receive larger benefits from the same level of structural diversity when compared to individuals with low educational status. This result is most similar to the argument in [121] in which Lin discusses the deficit in the return to social capital among some groups (e.g. women in contrast to men) as a mechanism that leads to inequality. The deficit in the return to social capital means that the marginal benefits of networking, or structural diversity in our case, for a subgroup is smaller than the other. When considered along with homophily and stratification across social status, the results of [122, 123] suggest that high status individuals receive larger benefits from their social contacts than low status individuals. This observation is in agreement with our findings. However there is an important difference between the findings in [123] and our findings, since in their case the differential effects are due to heterogeneity in contact characteristics. In contrast, we report the same phenomena from a relational perspective: *high status individuals have differential advantages stemming from the structure of their ego-networks, regardless of the characteristics of their contacts.*

3.1 Data

We use an anonymized mobile phone data-set containing one month of standard metadata in a developing country in South Asia. Using the calling records between cell phone users, we can map out the local structural characteristics of the network and link it with income, demographic characteristics and profession we obtained through surveys. In particular we focus on a local view of the network called ego-network. The focal node of interest is called the ego whereas all ego's connections are called alters. In addition to ego-alter edges, the ego-network includes all the edges between the alters, thus enabling us to study structural factors pertaining to redundancy and triadic closure [159]. As our goal is to study the effect of informal networks on economic outcomes measured in terms of income, we excluded those egos who do not hold a valid occupation (student or housewife or retired or unemployed) prior to the analysis.

Income Data: The income categories for a random selection of more than 270,000 individuals across the country were obtained through three sequential large-scale market research household surveys. 101,500 of these surveyed individuals were customers of our phone carrier. Out of these initial 101,500 individual surveys, we restricted our data to those who are employed (no students, housewives, unemployed or retired) and are at least 25 years old. Furthermore, to prevent our results from getting biased by inactive egos without enough communication data, we limited our data only to those individuals who had a phone communication with more than 5 unique individuals over the one month period of the phone data (Approximately 20% of individuals have $degree \leq 5$). This smaller subset of surveys accounted for 32,870 subscribers who we treated as the egos in our analysis. Information about income was directly asked from the respondents, who were requested to place themselves within pre-defined income bins. Several other demographic characteristics such as education, gender, age and occupation were obtained through the same survey. Survey participants were distributed across 220 sales territories proportional to their population so that there were overall about 400 surveyed households in each sales territory. Participant eligi-

Table 3.1: Survey relationship between household income categories and corresponding range in US dollars.

Income Category	**Monthly Household Income ($)**	**Frequency**
1	0-33	1895
2	33-78	9351
3	78-130	29718
4	130-195	28532
5	195-260	17841
6	260-325	9995
7	325-390	4536
8	390-455	3752
9	455-520	2341
10	520-585	929
11	585-651	999
12	651-1301	966
13	1301+	274

bility was defined as individuals with their own phone, between 15 and 65 years of age. The monthly income values were coded as ordinal categories from 1-13. Table 3.1 summarizes the correspondence between the income categories and their actual monetary value after conversion to US dollars. The Pearson correlation between the projected average income per region based on the survey results and their actual values published in official statistics is 0.925.

Social Network Data: We used one month of raw Call Detail Records (CDR) for all carrier subscribers to construct a large-scale undirected call graph, in which two individuals are connected if there is a call between the two in both direction during the observation period. Edges in the call graph can be weighted by the total number of phone calls between the two individuals during the observation period, but for the purpose of this study we only considered unweighted graphs. From the full call graph, we extracted individual undirected ego-networks corresponding to the surveyed individuals for whom we also have income and demographics information. It is important to note that the ego-networks only contain the reciprocal links to avoid spurious one-time contacts (e.g. telemarketing) to influence our results.

3.2 Methods

The general strategy of this correlational study is to define several variables that ultimately come into play in a regression analysis that allows us to investigate the association between structural diversity and income, and the presence of differential effects by social status.

Dependent Variable: Since the income is not observed as a continuous variable, we use the middle income value in each category as representing its actual income value. As confirmed in Table 3.1, the raw income values exhibits a log-normal distribution. Therefore, middle income value of the category in USD converted to the log-scale will serve as our dependent variable.

Independent Variable: Structural diversity serves as our main independent variable. There are multiple operationalizations for structural diversity, all of which measure the extent of structural non-redundancy among alters. We replicated our results using clustering coefficient, density and weighted structural novelty, a measure similar to normalized network effective size discussed in [34], and obtained similar results. Therefore, we only focus on clustering coefficient as the operationalization of structural diversity.

Clustering Coefficient measures the fraction of closed triads in an ego network. Sparsity and in particular open triads in an ego network is an indication of structural holes and that the ego acts as a bridge between the alters, who belong to different communities. Low clustering coefficient also means that there is little redundancy in the ego network and most alters act as novel sources of information. Since lower values of clustering coefficient correspond to sparsity, we use (1 - clustering coefficient) as our measure of structural diversity. This would effectively measure the fraction of alter pairs between whom ego acts as a bridge which indicates the extent to which ego acts a information broker in their network.

In addition to structural diversity, we use the level of education to demonstrate the differential effect of diversity across different social strata. The country of our interest experiences an excessive level of hereditary stratification and for this reason

we believe education serves as a sufficient proxy for social status. Education will be coded as a binary variable, with high corresponding to high school, Bachelors or Masters and low corresponding to illiterate, primary school or middle school.

Control Variables: In order to control for possible confounders with income, we will include profession (a categorical variable), gender, level of education (a binary variable), age (an interval variable) and the home location of each ego on a 5x5 grid over the country (a categorical variable) as control variables in our regression analysis. Including age ensures we compare income values along the same career phases and allows us to control for long-term changes in communication patterns that are associated with variation in income. By controlling for location fixed effects, we obtain a more justified comparison of income opportunities between individuals across vastly different geographical areas (e.g. urban vs. rural). The log degree of the ego must be present as another control variable in the model, because various measures of structural diversity (e.g. clustering coefficient) are correlated with degree and have different scales or reasonable ranges as the ego network grows larger. For example, as the degree of the ego increases, a fully connected ego-network, corresponding to a clustering coefficient of 1, becomes more unlikely since the edges between alters are not independent and in particular depend on the size of the ego-network. The clustering coefficient and log degree have a correlation of 0.4 in our data.

3.3 Results

The equation below demonstrates the model we will use in the regression analysis:

$$Y = \beta_0 + \beta_1 * I_{high} + \beta_2 * SD + \beta_3 * SD * I_{high} + C + \epsilon \tag{3.1}$$

where Y and SD corresponds to income and structural diversity respectively and I_{high} is an indicator variable taking a value of 1 when ego belongs to the high education group and 0 otherwise and C corresponds to the control variables. We have three main hypothesis:

1. More structurally diverse networks are associated with higher income: $\beta_2 > 0$

Table 3.2: Full Regression Results. Each column successively adds more controls variables to the model. High Education and Gender are binary indicator variables. Age, Profession and Location are all categorical variables and not shown among the control variables, but their corresponding rows indicate in which models they are included. Structural diversity is measured by clustering coefficient, but the results for other operationalizations of structural diversity are similar. Degree exhibits a power law distribution, thus it is transformed to log scale. Both Structural diversity and degree are standardized.

	Log Income				
	(1)	(2)	(3)	(4)	(5)
Structural Diversity	0.019^{***}	0.019^{***}	0.019^{***}	0.008^{**}	-0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Structural Diversity:	0.033^{***}	0.035^{***}	0.034^{***}	0.033^{***}	0.025^{***}
High Education	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)
High Education	0.489^{***}	0.482^{***}	0.487^{***}	0.305^{***}	0.291^{***}
	(0.007)	(0.007)	(0.007)	(0.009)	(0.008)
Degree	0.047^{***}	0.048^{***}	0.048^{***}	0.033^{***}	0.036^{***}
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Gender Female		0.081^{***}	0.094^{***}	0.095^{***}	0.072^{***}
		(0.012)	(0.012)	(0.012)	(0.011)
Age Included	No	No	Yes	Yes	Yes
Profession Included	No	No	No	Yes	Yes
Location Included	No	No	No	No	Yes
Number of Variables	5	6	13	29	44
Observations	32,870	32,870	32,870	32,870	32,870
R^2	0.158	0.159	0.168	0.243	0.301
Adjusted R^2	0.158	0.159	0.167	0.242	0.300
Residual Std. Error	0.581	0.581	0.578	0.551	0.530
F Statistic	$1{,}537.8^{***}$	$1{,}241.0^{***}$	551.9^{***}	376.2^{***}	329.3^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

2. Everything being equal, individuals with low education level have a *deficit* in their economic outcomes: $\beta_1 > 0$.

3. Individuals from a high education level have a larger *return* to the structural diversity of their networks: $\beta_3 > 0$. Note that the *return* refers to the marginal effect of structural diversity.

Table 3.2 shows our regression results where structural diversity is measured as clustering coefficient and in each column we successively add more control variables. We make two main observations from the results. First, all three hypothesis are validated in all models, with the exception of hypothesis 1 in model 6. Second, in model 6 which includes location fixed effects, structural diversity provides no return on income for the group with low education. Effectively, only individuals with high education benefit from access to structurally diverse sources of information. We obtained similar results using the other two operationalizations of structural diversity, which points to the robustness of these observed effects.

It should be noted that females in our data tend to have higher income than men; because while women in our country of study are generally housewives, those women who are employed disproportionately hold better paying jobs such as teacher or government worker. Nevertheless, when we define social status in terms of gender (I_{high} would be an indicator variable taking a value of 1 when ego is male and 0 when ego is female), we obtain similar results. We again observe that men have larger marginal benefits to structural diversity than women, suggesting that women have a deficit in returns to networking, a well-documented phenomenon in various context [30, 128, 132].

These results suggest that information and economic opportunities are not distributed uniformly across the social network and high status individuals have an advantage in terms of their returns to networking. A potential mechanism that explains the differential returns to structural diversity relies on homophily. When the concentration of information about economic opportunities within high status and influential social strata is combined with strongly homophilous ties among individuals

from the same social strata, the result is differential benefits of high status individuals from networking.

We again note that these claims are in no way causal since for establishing the causality we either need an appropriate instrumental variable in place of structural diversity and social status or have to conduct a randomized experiment. Nevertheless, we believe the observation of such differential effects, matching our theoretical expectation, provide more credibility to our proposed mechanism of network inequality. In the next section, we discuss a randomized seeding experiment that provides causal evidence for our proposed mechanism.

4 Causal Study

In the previous section, we provided suggestive evidence that networks provide a mechanism for increasing inter-group differences in terms of access to information by amplifying any initial advantages by members of one subgroup. We can think of structural diversity as the amount of novel information each individual receives from its network and the interaction term in equation 3.1 captures the extra information high status group receives by amplifying the individual advantages of each member of the group. However, our regression analysis might suffer from endogeneities between income and the initial economic advantages of the high status group. In order to address this problem and obtain a causal estimate for the network effects on unequal diffusion of information, we need an exogenous variation in the initial advantage of a subgroup in the network. Furthermore, the attribute(s) that defines the subgroups should exhibit a high degree of assortativity in the group. Random seeding of members in the network with a new behavior or a piece of information which diffuses throughout the network provides an ideal setup to identify the causal effects of the network. In this context, the random selection of a seed individual constitutes an initial advantage for the group to which the seed belongs to. Thus, the seeded subgroup will represent the high-status group and the exogenous variation in status of a subgroup enables us to identify the causal effect of the network on unequal diffusion.

For this study, we reuse the data already collected from a randomized experiment that aimed to estimate the direct and indirect (through diffusion) effects of an anti-conflict program in New Jersey middle schools [140]. This study aimed to measure the efficacy of anti-conflict interventions in changing various outcomes of interest, such as school-level prescriptive norms about conflict or total number of reported conflict incidents. The researchers argued that encouraging and training a small number of students to take a public stance against conflict will have school-wide consequences since the new anti-conflict behavior will diffuse throughout the network, especially if the selected seed students are well-connected or social referents in the school network. The authors carefully measured the full social network in all schools prior to the intervention. At the beginning of the school year, between 20 and 32 students depending on the school size and blocked by grade and gender were randomly seeded with the anti-conflict intervention. The intervention had various components such as assisting seed students to identify conflict behaviors, creating online and offline social campaigns against conflict or distributing orange wristband representing a public stance against conflict. The diffusion of these practices from the seed students did in fact change the norms of conflict and reduced the overall level of conflict by about 30%.

Our goal is to uncover the differential diffusion of the anti-conflict behavior, if any, to different subgroups depending on the random composition of the seed students. For this reason, we only focus on the randomly treated schools that actually did receive the intervention and evaluate whether the choice of initial seed students increases the adoption probability of other students from the same subgroup as the seeds. As GPA exhibits high levels of homophily, we define our two subgroups of interest as students with high or low GPA and evaluate if a seed group of predominantly high (low) GPA causes other students with high (low) GPA to wear the anti-conflict wristband. The schools show very high levels of assortativity at both gender and grade level, hence we believe diffusion from one gender-grade group to another is very unlikely. Therefore, we investigate the differential adoption rate in grade-gender subgroups depending on the composition of seed students within that grade-gender subgroup. The treatment

to each unseeded student takes a linear form as the number of seeds within the same grade-gender group that have the same GPA as the student. Using both randomization inference and inverse probability weighting, we find significant linear effects for the number of same GPA students in the seed set on the average adoption of unseeded students. These results provide a causal evidence for the mechanism we have posited for unequal diffusion in a network. An initial advantage by the members of a subgroup (exogenous seeding of a GPA subgroup) introduces differential advantages for other members of the same subgroup compared to the rest of the network (as the number of seed students within a GPA group increases, the unseeded students from the same GPA group are more likely to adopt the new behavior).

4.1 Data

The original study conducted treatment randomization at two levels of school and within school. At the school level, out of the total 60 middle schools, half were controls and did not receive the anti-conflict treatment. Out of 24191 students in all schools, 11,938 are in treated schools. Within schools, randomization was conducted on the basis of seeding a limited set of students, referred to as seed-eligibles representing 15% of school population, with the anti-conflict program. Out of 24191 students in all 60 schools, 2943 were eligible to be selected as a seed. Out of all the seed-eligible students, 1456 students eventually became seed-eligibles in the 30 treated schools, and among these 1456 seed-eligible student, half of them were randomly selected to be actual seeds. The seed-eligible students and the selection of actual seeds were blocked by grade and gender. There were on average 26 seeds in each treated school. Out of 11,938 students in the treated schools, 5,754 were not directly connected to any seed student. These students will be the focus of our study as their exposure to the new behavior initiated by the seeds are through diffusion in the network.

Prior to the start of the intervention, the researchers mapped out the full network of all schools by asking each student to nominate up to 10 students they are close friends with and look up to. This network structure serves as the backbone of our analysis on the diffusion of the new anti-conflict behavior, originating from the seeds.

The research also collected various attributes, such as gender, grade, age, GPA, home language and race per each student. We will use these attributes to define subgroups within the school network and measure the extent of homophily along each subgroup characterization. We will also use some of these variables as controls in the regression model we develop to estimate the causal effect of network on adoption. The survey also included several question about prescriptive (descriptive) norms about the conflict at the school. Repeating the same survey questions about the conflict norms after the intervention allowed the researchers to evaluate the effectiveness of the intervention program.

As part of the anti-conflict treatment seed students received, they were encouraged to give an orange wristband with the anti-conflict intervention logo to any student who engaged in conflict-mitigating behavior. The goal of this activity was to increase awareness about conflict and promote higher diffusion of the new anti-conflict behavior at the school. Since our goal is to study the effect of seed group composition on differential adoption of a behavior, we focus our attention on the 30 schools that received the treatment. Furthermore, we use the adoption of the wristbands by each student as the new behavior (dependent variable) that diffuses through the network. We make this choice because the wristband adoption seems to be the only outcome variable with high rate of diffusion throughout the network, which makes detection of unequal diffusion, if any, easier.

4.2 Methods

In simplest terms, we will attempt to find evidence for the following hypothesis. Members of a group that have more students belonging to that group in the seed set are more likely to adopt the behavior. As the composition of the seed set is random, the differential adoption by members of a specific group that is well-represented in the seed set compared to the other groups will be a causal effect. We will show this by computing the linear effect of number of seed students from a group on the adoption rate of other students from the same group. We further will show that the observed effect is positive and significant through randomization inference and inverse

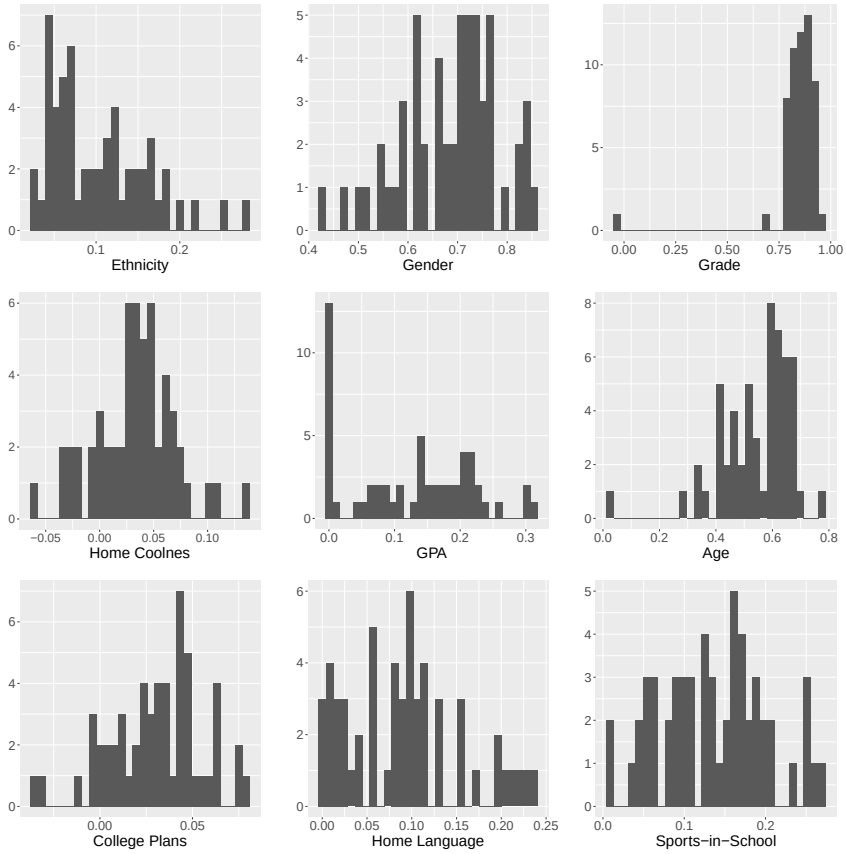


Figure 3.1: The Histograms of assortativity measures among all treated schools. Each plot shows the extent of homophily in schools by a different assortativity attribute.

probability weighting. As we are interested in unequal diffusion as our mechanism, we will estimate the effect of seed composition on the limited set of students that are not seeded and are not directly connected to a seed student, excluding students that are directly connected to a seed as any effect on them cannot be characterized as diffusion. We will refer to these student as control, as opposed to directly treated (seed) or indirectly treated (connected to a seed).

In the original study, adoption of anti-conflict wristbands is the outcome that achieved the highest level of adoption (most wide-spread diffusion in all schools). For this reason, we will focus on the diffusion of this behavior and attempt to provide causal evidence for its unequal diffusion depending on the composition of original seed students. In order to find evidence for differential diffusion of wristband adoption, we first need to characterize the subgroups of students whose links are homophilous.

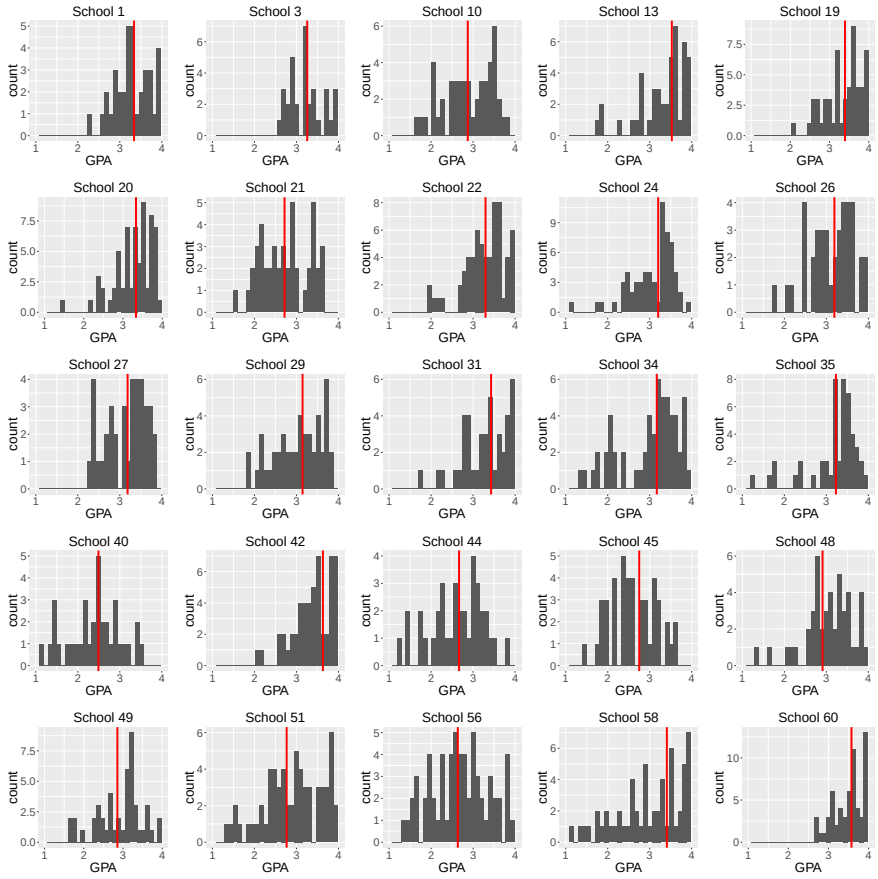


Figure 3.2: The histogram of seed eligible students' GPA in treated schools. The red line indicates the median GPA within that school. Any student to right (left) of the red line belongs to high (low) GPA subgroup.

Figure 3.1 shows different attributes for which we could define the distinct groups in the network and the level of homophily in the school networks for that attribute.

As illustrated in Figure 3.1, gender and grade exhibit the highest levels of homophily in the school networks. But the initial seed students are blocked at the gender and grade level. As a result, the randomization process will not generate any variation on the number of seed students from any grade or gender subgroup. This makes both grade and gender unsuitable as the attribute defining the homophilous subgroups. We cannot use age as the group attribute either since it is highly correlated with grade and randomization inference won't generate enough variation in the age composition of the seed set. The only remaining attribute that exhibits high levels of homophily in the school networks is GPA. We have converted GPA to a binary

variable for which high (low) signify a GPA above (below) the school median. Figure 3.2 shows the extent of GPA variation among seed eligible students. We observe that the randomization can generate considerable level of variation in the number of seed students that belong to each GPA group. As a result, we will use binary GPA as the group attribute in the schools and show that homophily in terms of GPA leads to unequal diffusion of wristband to students of a specific GPA group depending on the (random) composition of the seed set.

Furthermore, as the schools exhibit very high degrees of homophily in terms of grade and gender, we believe the composition of the seed set in one gender-grade group should not have a considerable effect on the adoption of students in a different gender-grade group since the diffusion across gender-groups is likely to be minimal. For this reason, we define the seed set GPA composition within grade-gender groups rather than at the school level. This also fits well with the experimental design of the original study since randomization was blocked by grade and gender. By defining the seed set composition at the grade-gender level, we also reduce the maximum number of seed-eligibles from a GPA group to 15 rather than 56 at the school level. This smaller number also makes the analysis easier as the effect of seed composition on adoption is likely to be concave.

The treatment effect on a control student is a linear effect, rather than a binary one, since treatment effect depends on the number of seed students within the GPA category of the control student. The effect is likely to increase as the number of same-GPA seed students increases. Hence, we fit a model to the data that includes the number of same-GPA students as a linear term. In other words, there will be a maximum of 15 treatment conditions corresponding to the maximum number of seed-eligible students that belong to a specific GPA category at the grade-gender level. In this context, our estimated quantity will be the marginal effect of having one more seed student from a GPA category on the adoption of a control student from the same GPA category and grade-gender group as the seed student. We will evaluate the significance of the estimate using randomization inference with 1000 permutations that exactly follow the randomization process used in the original study.

The Fisherian null hypothesis will be that the GPA composition of the randomly selected seed students has no effect on the adoption rate of all control students. The model we fit on the observed data and the randomized permutations of seed students is as followed:

$$Y_{ij} = \alpha_{ij} + \tau N_{ij} + \gamma C_{ij} + u_{ij} \tag{3.2}$$

where Y_{ij} is the observed adoption outcome of student i in school j, N_{ij} is the number of seed students in the same grade-gender group and with the same GPA as the student i. C_{ij} is the list of control variables. We include home language and perception of student's home by its friends in the control variables since they correspond to economic and minority status of the student. These two variables were also included as control variables in the fitted model of the original study. We also use the population in the student's grade-gender group as another control variable in the model, since the adoption rates and in general the efficacy of the intervention seems to be smaller in larger schools. Finally, grade and gender fixed effects and the population of students at the grade-gender level from the GPA group of student i are the additional control variables in the model.

A more interesting alternative hypothesis is regarding the interaction of GPA assortativity and number of same-GPA students in the seed group. We would expect that as the assortativity in terms of GPA at the school (or school-grade) level increases, the adoption rate of students with the same GPA as a seed student will increase. We can estimate this effect as an interaction term in the linear model of adoption above. We will analyze this effect in future work.

In addition to randomization inference, we will also estimate the quantity above using inverse probability weighting of the treatment condition and make inferences with cluster-robust standard errors at the school level. We choose the clusters to be at the school level, rather grade or gender levels, to be conservative about independence between clusters. Inverse probability weighting (IPW) requires each observation to have a non-zero probability of receiving all possible treatment conditions. There

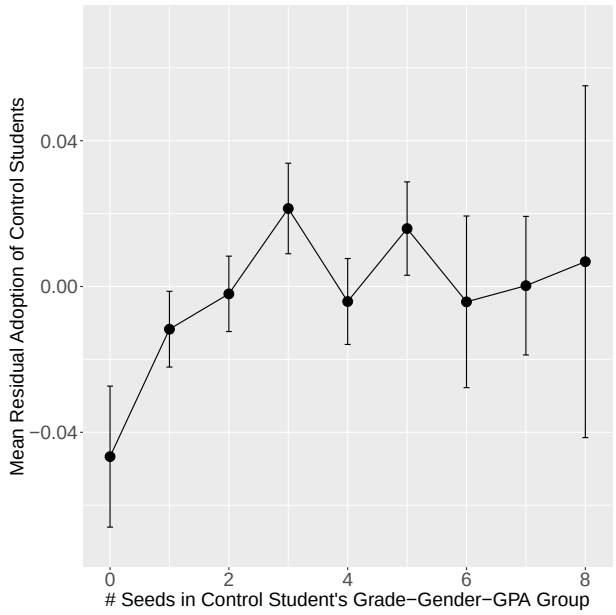


Figure 3.3: Mean residual adoption rate of control students versus the number of seeds from the GPA group of the control student. Grade-gender population, home language, house quality, gender, grade and control student's GPA group population at the grade-gender level are controlled for in the adoption rate. The bars correspond to standard error of the mean. The effect of the seed composition seems to taper off at a threshold of 2 seed students from the same GPA group as the control student.

are 15 treatment conditions in our data corresponding to the maximum number of seed-eligibles from a grade-gender-GPA group. However, many of these treatment conditions are impossible for most control students since they are in smaller schools and don't have that many seed-eligible students within their GPA group. Therefore, if we estimate the marginal effect of the seed composition using IPW with all possible treatment conditions, our data will become extremely restricted. One way to address this problem is to estimate a threshold for the number of seed students within the grade-gender-GPA group of the control student beyond which there is no further marginal effect on the adoption of the control student. This approach is reasonable since the effect of the seed composition on adoption of control students has decreasing marginal returns. Figure 3.3 shows the mean residual adoption rate grouped by the number of seed students from the GPA group of the control students after removing the effect of control variables in equation 3.2. We observe that there is almost no extra effect by the seed composition beyond 2 seed students from the same GPA

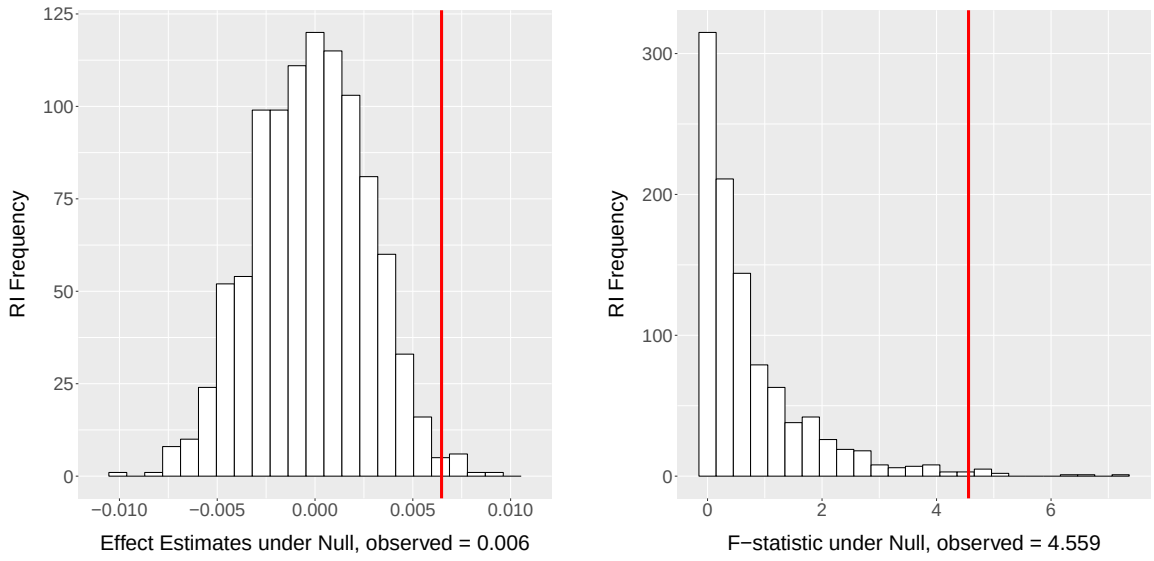


Figure 3.4: The observed effect of the number of same-GPA seeds as the control student vs its distribution under Null obtained through randomization inference at grade-gender level. Left column shows the distribution of regression coefficient obtained through randomization inference and right column shows the RI distribution of the F-statistic from comparison of the model with seed GPA composition vs a model without it. Red lines correspond to the estimate from the observation. The two-tailed p-value for the coefficient (left) is 0.019 and the one-tailed p-value for the F-statistic (right) is 0.011.

group as the control student. Hence, we used 2 as the cap for the number of same-GPA students and converted all values above that to 2. This allowed us to retain 47% of the control students in the inverse probability weighting regression analysis.

4.3 Results

Figure 3.4 shows our main results from randomization inference. We observe that the marginal effect of same-GPA seed student is positive and significant at the grade-gender level, in the direction our hypothesis predicted. The estimated effect of 0.006 means that having one more seed student in the grade-gender group from the same GPA category increases the adoption probability of a control student by 0.6%. This effect size is nontrivial as the total adoption rate among all treated schools is only 11% and the seed group at the grade-gender level can easily contain 4 students from a specific GPA group.

Table 3.3: The inverse probability weighted regression of control student's wristband adoption on the number of seed students from same GPA and grade-gender group as the control student. Model 1 only includes the population of the grade-gender group as a control variable, whereas model 2 includes all the control variables mentioned in equation 3.2. The standard errors are clustered at the school level.

	Dependent variable:	
	Wristband Adoption	
	(1)	(2)
Constant	0.210^{***}	0.187^{**}
	(0.038)	(0.057)
Number of Same GPA Seeds in Grade-Gender Group	0.030^{*}	0.037^{***}
	(0.012)	(0.009)
Grade-Gender Group Population	-0.002^{***}	-0.001^{*}
	(0.0004)	(0.0005)
GPA Group Population in Grade-Gender Group		-0.001
		(0.001)
Home Language Not English		0.017
		(0.054)
House Quality		0.031^{*}
		(0.015)
Gender Male		-0.053^{*}
		(0.022)
Grade Fixed Effects Included	No	Yes
F Statistic	8.7331^{***}	4.18^{***}
	(df = 2; 2101)	(df = 9; 2094)
Observations	2,104	2,104
R^2	0.031	0.044
Adjusted R^2	0.030	0.040
Residual Std. Error	0.595	0.592
	(df = 2101)	(df = 2094)
Note:	*p<0.05; **p<0.01; ***p<0.001	

We observed that excluding the grade-gender population as a control variable made the distribution obtained through randomization inference biased toward negative values and not symmetric around zero (not shown here). The randomization inference distribution becomes centered around zero when we control for grade-gender group population. This suggests that the adoption rate decreases by population size of the grade and gender groups and population is positively correlated with number of same GPA seeds. This observation is further confirmed when we note that effect of grade-gender population is negative and highly significant (in table 3.3).

Table 3.3 shows our results from inverse probability weighted regression of control student's wristband adoption on the number of seed students who belong to the same GPA group as the control student. The probability distribution of treatment levels is hypergeometric since the number of seed-eligible students and selected seeds are fixed in each grade-gender group. The estimated effect of seed composition on adoption is significant and in the direction predicted by our hypothesis. We should note that the estimate in table 3.3 is different from the one shown in figure 3.4 for multiple reasons. First, the estimate and the distribution computed in figure 3.4 do not use inverse probability weights, hence they are computed on a much larger sample. Second and more importantly, since the randomization inference does not use IPW, there is no need to cap the number of seeds from the control student's GPA group. In other words, the computation of the estimate and the distribution in figure 3.4 uses all possible treatment levels, ranging from 0 to 15 seed students from a GPA group.

5 Conclusion

While there has been many studies on negative impacts of inequality or how to address it through macro-level instruments, there has been very little focus on the micro-mechanisms that generate unequal outcomes. Previous studies have discussed how networks play an important role in economic outcomes such as access to employment opportunities and others have even discussed network effects on inequality at a theoretical level. One way networks can regenerate inequality is through unequal

diffusion of economic information to members of the network. Here, we discussed a simple diffusion-based mechanism which predicts that if members of a group start with a slight initial advantage in terms of access to the economic information, the diffusion process will amplify this advantage further leading to more unequal outcomes between different groups in the network. This simple mechanism requires a group to start with initial advantage and the network to be homophilous in the attribute that defines the groups. The main driving force behind the network mechanism is homophily, hence the mechanism purely depends on the structure of the network. The structural nature of this mechanism makes it easy to investigate whether it occurs and measure the impact it has in generating unequal outcomes in networks with known structure.

In this chapter, we provide both suggestive observational evidence and a causal confirmation for this mechanism. The observational evidence, while purely correlational, lends some credibility to a consequence of our proposed mechanism for network inequality that high status individuals receive larger marginal benefits in terms of wealth from their networks compared to low status individuals. The causal study, conducted as a randomized experiment of seeding in networks, evaluates the extent of diffusion to different parts of the network. Using the data from a previous experiment, we show that depending on the composition of initial seeds that receive a treatment to adopt a new behavior, other network members not directly connected to but belonging to the same group as the seeds are more likely to adopt the behavior compared to other network members. The information that diffuses in the network is the adoption of an anti-conflict stance in New Jersey high schools.

We show that the students who belong to the same GPA category as the initial seed students are more likely to adopt the anti-conflict behavior. In other words, the initial advantage by one member of a GPA group (being selected as a seed student) leads to differential advantages in terms of adoption for other members of the same GPA group compared to non-members in the network. In particular, we find causal evidence that the existence of one more seed student within a grade-gender block and with the same GPA category as an untreated control student, increases the adoption

probability of the control student by at least 0.6%.

While the outcome we have studied in the causal study is not economical in nature, it nevertheless serves as a self-contained and easy to measure prototype to illustrate the potency of the proposed mechanism behind unequal diffusion. We expect the same mechanism to play an even more important role in diffusion of economic information that are often rivalrous. As future work, we plan to evaluate the unequal diffusion of information about a desirable but limited insurance product among members of a village where some farmers are randomly selected to be the initial seeds that receive the information about the insurance product. In this context, we will illustrate the existence and measure the extent of unequal diffusion among homophilous gender groups in the village.

Chapter 4

Unequal Diffusion: A Stochastic Network Model with Brokerage

1 Preface

In the previous chapter, we provided observational evidence of differential network effects across different groups. We attributed this effect to unequal diffusion of valuable resources such as information. We argued that unequal diffusion happens due to homophily and provided causal evidence for it within a randomized seeding experiment in networks where a new behavior diffused at a higher rate to groups to which the initial seeds belong. In this chapter, we introduce a random network model to examine how network structure leads to unequal diffusion. We discuss the model properties pertaining to diffusion and present some empirical results. Our results indicate that homophily does not fully explain the extent of unequal diffusion in a network and one should also consider the extent of cross-group brokerage when considering unequal diffusion.

2 Brokerage and Unequal Diffusion in Latent Space Networks

Before discussing the main model, we quickly introduce a simulation based on the latent space network approach [96] that illustrates how brokerage in network structure leads to unequal diffusion or higher assortativity [133] on diffusion paths. The nodes are distributed on the real line which constitutes the one dimensional latent space. We consider nodes with positive and negative latent variable to belong to two different groups and have the probability of an edge between them to decrease with the distance between two nodes in the latent space. The various distributions one can employ for the latent variable allows the model to capture homophily and varying degrees of brokerage. Figure 4.1a depicts how a change in distribution of the latent variable leads to few nodes of high brokerage whose latent variable is close to zero. When the nodes are distributed according to the standard normal, there will be many positive and negative nodes close to zero with an edge. However, when the nodes are distributed according to a normal mixture centered symmetrically at positive and negative values, very few nodes from positive and negative groups close to zero will account for the majority of cross-group links. Figure 4.1b shows while assortativity along paths of length 1 decreases as cross-type edges appear, assortativity along paths of length 2, akin to diffusion assortativity, can remain high if the network exhibits high brokerage. Results suggest that even networks that are extremely homophilous on paths of length 1 can have high levels of mixing when one considers paths of length 2 or 3. However, networks that exhibit a high level of brokerage can still have high levels of assortativity on diffusion paths, hence be more susceptible to unequal diffusion.

3 Introduction

Diffusion of information in social networks determines who gets access to a valuable piece of information, such as a new investment opportunity. The structure of the network plays an important role in which individuals or groups receive the valuable

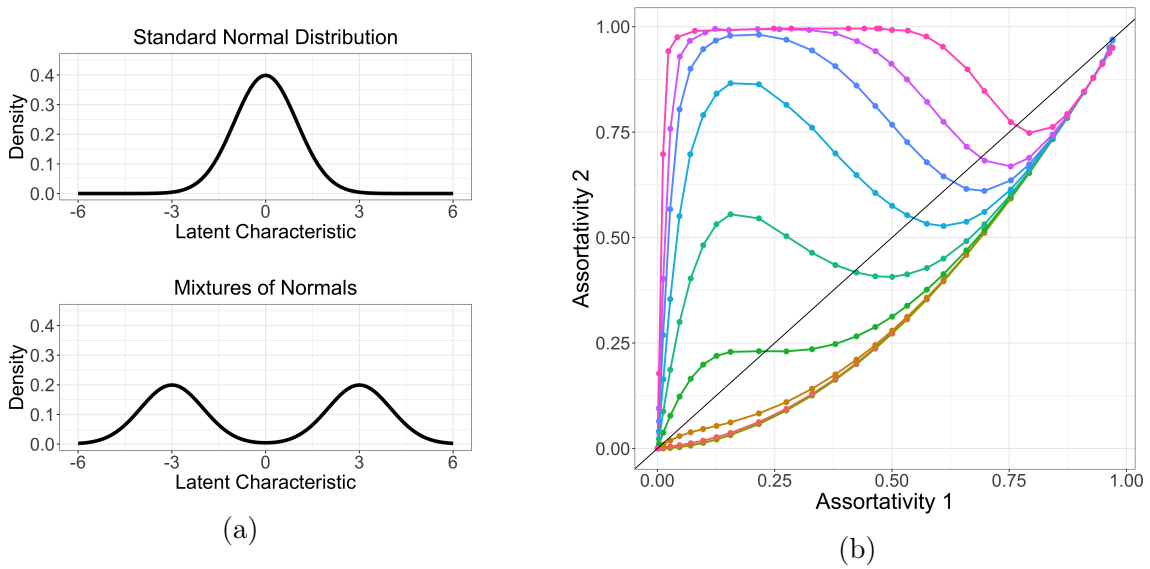


Figure 4.1: Nodes with negative or positive latent variable belong to distinct groups. Top distribution in panel (a) corresponds to a network where cross-group edges are almost equally distributed whereas the bottom distribution corresponds to a network with few nodes of high brokerage close to zero. Panel (b) shows how assortativities along paths of length 1 and 2 vary as the distribution of nodes on the latent space varies. Each point corresponds to a network model as we vary the variance of normal distributions and the distance between them in the mixture. Different colors correspond to varying levels of brokerage. For a fixed level of assortativity on paths of length 1, as brokerage increases, assortativity on paths of length 2 also increases.

information. Certain network structures are more likely to keep a piece of information exclusive to one group, thus leading to unequal diffusion. For example, if there are very few social links between people of different races, the information about a new employment opportunity that is generated among one race might never reach individuals of the other race [43]. Many existing network models aim to explain the absence of diffusion from one group to another through assortative mixing [134]. Assortative mixing, or simply assortativity, captures the bias in forming edges with similar characteristics. It is also referred to as homophily which simply means that attributes of nodes are correlated across the edges. For example, in social networks individuals have a strong tendency to form links with other people who are similar to them in terms of age, language, socioeconomic status or race.

Stochastic Block Model (SBM) along with its variants such as degree-correction [114] are an important class of these models that explicitly account for assortative

mixing in networks. SBM is a generative random network model for modeling blocks or groups in networks. It has been widely used in computer science and social sciences to model community structure in networks [8, 78, 98, 148, 174, 175]. In its original form, vertices in a network exclusively belong to one of the K groups (or blocks) in the network. Each pair of vertices form an edge independently of other edges or vertices. Edge formations between any pairs of two groups are independent, identical and solely determined by the group membership of the pair of vertices. If $g_i \in \{1, 2, ..., K\}$ corresponds to the group of vertex i, then a $K \times K$ matrix, P, determines the edge formation probabilities between any pair of vertices. The probability of an edge between any pairs i and j is the (g_i, g_j) element in the matrix, P_{g_i,g_j}.

This simple model can produce a variety of interesting network structures. For example, an edge probability matrix in which diagonal entries are much larger than off-diagonal entries produces networks with densely connected groups and sparse connections across groups. The ability to model such community structure is the main reason SBM can capture assortative mixing in a network. This has led to the popularity of SBM as one of the main methods for community detection. SBM does so by generating random networks that match the observed network in terms of the frequency of within-group and cross-group edges. The fitted model matches the observed assortativity or homophily in expectation.

SBM or its degree corrected version assume that within-group and cross-group edges are distributed "uniformly" across all pairs: the existence of an edge between any two pairs is identical to other similar pairs. In the case of degree-corrected SBM (DC-SBM), after conditioning on degree two nodes are similar in terms of their cross-group edge formation. In reality, many real networks have heterogeneous propensities in edge formation to various groups. In most cases, social networks exhibit a pattern of brokerage which means cross-group edges are not distributed uniformly, instead a small subgroup of nodes hold a disproportionate level of cross-group edges. Simmel was the first to introduce the concept of network brokerage in triadic relations [158]. Burt later advanced our understanding of brokerage by introducing the concept of "structural holes" between two unconnected communities, across which brokers act as

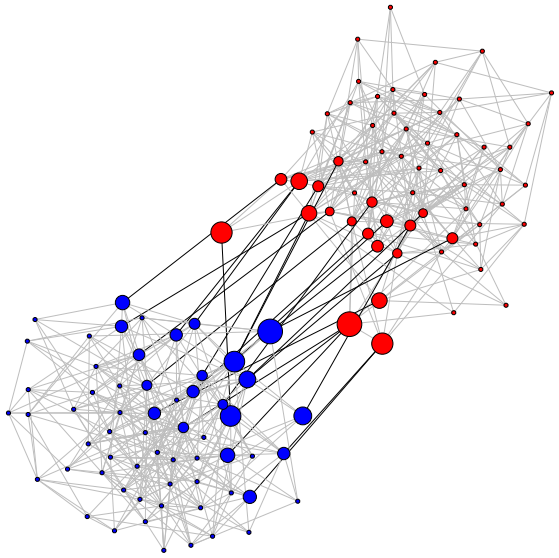


Figure 4.2: A random network with brokerage, in which a disproportionate fraction of cross-type edges are held by a small number of nodes. All nodes have the same expected degree so comparison with baseline SBM is appropriate. The size of each node corresponds to the number of its out-group edges. Majority of the nodes have a small probability of forming out-group links, but a small number of broker nodes have a much higher probability of forming links to out-group brokers.

intermediary [34]. These broker nodes play an important role in connecting otherwise disconnected communities, moving information between them, and acting as an intermediary for resource exchanges. Due to their unique position in the network, brokers benefit from various types of advantages, for example access to diverse information or opportunities for arbitrage in exchanges. However, these advantages to brokers might lead to some costs to other actors in the network or the network as a whole.

In the context of SBM, network brokerage occurs when a few nodes in the network have higher propensity to connect with an out-group than other in-group nodes. Figure 4.2 provides a visual example of a network with brokerage in which a small number of broker nodes have a higher probability of forming links with brokers of the out-group, hence maintaining majority of cross-group edges. While brokers play an integral role in connecting otherwise disconnected communities, they can nevertheless act a bottleneck by reducing the number of possible paths between any two groups when compared to a similar network with cross-group ties uniformly distributed across the network. Because brokers hold a disproportionate number of cross-group ties, they can constrain diffusion of information from one group to another. In this paper, we argue that one needs to not only look at homophily or assortativity on paths of length

1, but also on the extent of assortativity of all possible diffusion paths of varying lengths to completely account for unequal diffusion in networks. We then attempt to incorporate the heterogeneity in edge propensities and in particular brokerage into class of Stochastic Block Models and show that by doing so the model better explains unequal diffusion of information.

We show that while directly fitting for assortativity on paths of length 1, SBM fails to accurately capture (diffusion) assortativity on longer paths in real world networks. In section 4, we discuss SBM and some variant models and show that they consistently under-estimate the observed assortativity on paths of length 2 in 56 school networks, even though these models explicitly accounts for assortativity on paths of length 1. In sections 5, we develop out model which account for node heterogeneity in brokerage and by doing so match assortativity on paths of length 1 and 2 in expectation. In section 6, we provide the results from fitting the school networks to our model and show that even though not explicitly modeled for, it closely matches assortativity on paths of length 3. In the remainder of this document, we mostly focus on Assortativity of path length 2 as opposed to longer paths and refer to it as **Diffusion Assortativity**.

4 Background

4.1 Assortativity

Before discussing the Stochastic Block Model and its properties regarding diffusion, we need to explain the assortativity coefficient, a common way to quantify the level of assortative mixing in a network. The assortativity coefficient in a directed network, which quantifies the bias in favor of edges between in-group nodes, is defined as below [134].

$$r^{(1)} = \frac{\sum_r e_{rr} - \sum_r a_r b_r}{1 - \sum_r a_r b_r} \tag{4.1}$$

where the quantity e_{rs} is the fraction of total (directed) edges from a node in group r to a node in group s, a_r is the fraction of total edges from a node in group r and b_r is the fraction of total edges to a node in group r. Below we denote the adjacency matrix as $\mathbf{A}$ and the group of node i as g_i.

$$e_{rs} = \frac{\sum_{i,j} A_{ij}\delta_{g_i,r}\delta_{g_j,s}}{\sum_{i,j} A_{ij}} \qquad a_r = \sum_s e_{rs} \qquad b_r = \sum_s e_{sr} \tag{4.2}$$

The numerator in equation 4.1 is simply the *modularity* of the network, another quantity for the strength of community structure in networks [86, 135, 136] that measures the fraction of in-group edges minus its expected value if the stubs were randomly rewired. The assortativity coefficient is effectively the scaled modularity such that $-1 \leq r^{(1)} \leq 1$. The (1) superscript in equation 4.1 indicates assortativity is measured on paths of length 1.

4.2 Assortativity on Longer Paths

We can define higher order measures of assortativity to quantify the level of assortative mixing along diffusion paths. For example, to compute assortativity on paths of length 2 on a (directed) network, we first construct its corresponding network along paths of length 2 forbidding the traversal of the same edge multiple times and call it the second order network. In this network, there is a (directed) edge from node i to j for every path of length 2 from i to j in the original network. The assortativity of the second order network corresponds to assortativity along paths of length 2 in the original network denoted by $r^{(2)}$. The second order network will be a multi-graph with potential self-loops, both of which are compatible with the definition of assortativity in equation 4.1. A similar measure to our diffusion assortativity, but in terms of degree assortativity, is discussed in [13].

4.3 Stochastic Block Model

The Stochastic Block Model (SBM) [97] in the most basic form of network models that allow for communities and heterogeneous edge formation between them. It assumes edge formation between a pair of nodes solely depends on their observed block membership and is independent of other pairs. Consequently, all nodes within a block in SBM have the same binomial distribution for their in-group and out-group degree. Often, the SBM is characterized with an matrix whose elements determine the probability of an edge between any pair of blocks. For example, if we assume two groups in the network, the probability matrix for the undirected SBM has the following form.

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \tag{4.3}$$

An appealing property of SBM is that it accurately captures the strength of community structure or assortative mixing in a network. In particular, if we let $\hat{r}$ denote the assortativity coefficient of a sampled network from the maximum likelihood fit, $\widehat{P}$, we have the following convergence in probability as network size grows.

$$\hat{r}^{(1)} \xrightarrow{p} r^{(1)} \tag{4.4}$$

In fact, if the network is large enough it can be shown that assortativity from the fitted MLE model approximately matches the observed assortativity in expectation, with exact equality in the case of microcanonical SBM [142].

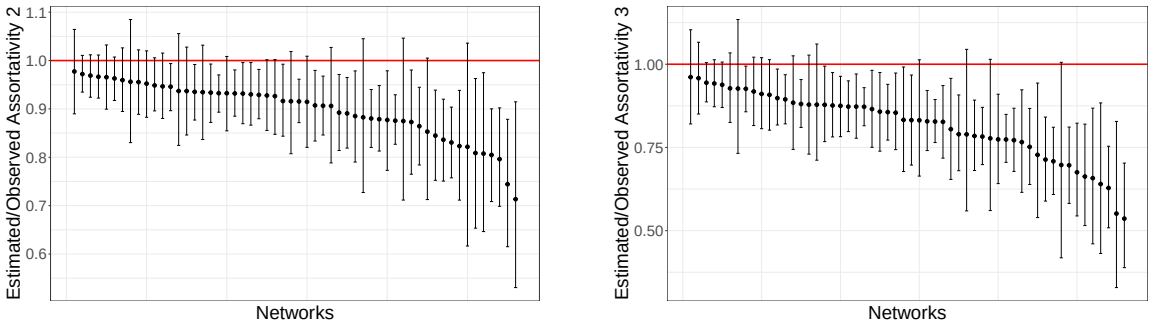
$$E[\,\hat{r}^{(1)}\,] \approx r^{(1)} \tag{4.5}$$

Despite its simplicity and its wide-spread use to model community structure, SBM has serious drawbacks when it is used to model real-world networks. The main problem with SBM is its inability to allow for degree heterogeneity within a block. This makes SBM an unreasonable model in real world networks which exhibit high levels of degree heterogeneity [141]. A maximum likelihood fitting procedure as described

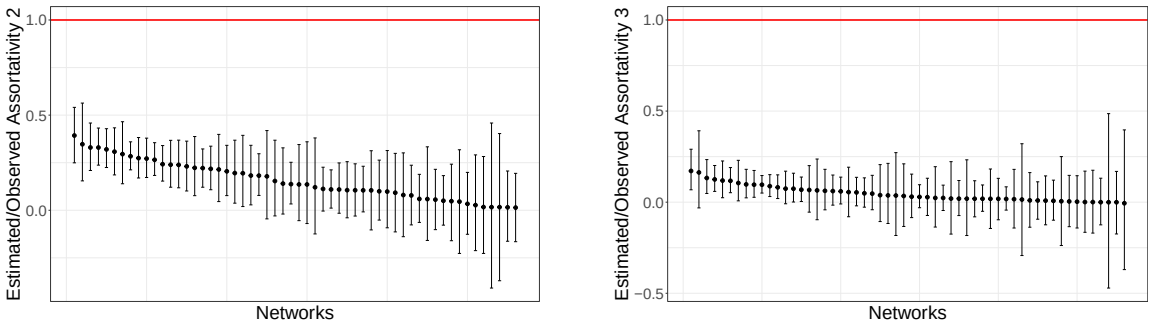
above, in the presence of degree heterogeneity, results in communities of high and low degree nodes. In particular, the MLE captures degree heterogeneity rather than actual community structure since it splits nodes from the same block into distinct blocks differentiated by their degree. For example, Bickel showed that SBM splits nodes in the famous Karate club network according to their degree rather than extracting the actual communities [23].

To avoid this problem, the degree-corrected SBM (DCSBM) modifies the generative model such that nodes can have different degrees in each block. It does so by introducing a degree-correction parameters for each node that simulates the node's propensity to form edges, hence controlling for the expected degree of each node separately. A node with a larger value of degree-correction parameter is expected to have larger degree than a node with smaller value and in the same block. Furthermore similar to SBM, the degree-corrected SBM has additional parameters that control for the propensity of any two groups to form links independent of each node's individual degree propensity. The degree-corrected SBM as a class of models is more general than SBM as it encapsulates SBM. SBM is a special cases of its degree-corrected variant when all node degree parameters within a single block are equal. Similar to SBM, the fitted maximum likelihood model for DCSBM also matches the observed assortativity as expressed in equations 4.4 and 4.5.

Despite its ability to model for degree heterogeneity and its success in real world problems, DCSBM is unable to model heterogeneity in in-group and out-group propensities or brokerage since it uses a single parameter per pair of blocks as their edge propensity. In other words, all nodes within a block have the same in-group and out-group degree distribution conditional on total degree. This makes it difficult for DCSBM to accurately capture assortativity on longer paths if the network exhibits brokerage, as discussed above and shown below empirically. The DCSBM maximum likelihood estimates underestimate diffusion assortativity, even though the expected assortativity on paths of length 1 from DCSBM fitted ML model matches its observed value.



(a) Gender



(b) Race

Figure 4.3: The distribution of predicted over observed ratio of assortativities on paths of length 2 (left column), $\frac{\hat{r}^{(2)}}{r^{(2)}}$, and 3 (right column), $\frac{\hat{r}^{(3)}}{r^{(3)}}$, from DCSBM along gender (top row) and racial (bottom row) groups. Bars correspond to 95% confidence interval and each bar corresponds to one school network. Networks are sorted in descending order of the point estimate.

4.4 Empirical Study of Diffusion Assortativity with DCSBM

In this section, we analyze a collection of real-world social networks and show that many have assortativity on paths of length 2 that is not predicted by SBM which explicitly fits assortativity on paths of length 1. We reuse the data already collected from a previous study that fully mapped out the social network in 56 middle schools [140]. These networks are directed and as such we fit them to a directed DCSBM model. We use these networks to study how and whether DCSBM models mixing structure and in particular diffusion assortativity accurately. The data also contains various attributes, such as gender, grade, age and GPA per each student. We will use these attributes to define subgroups within the school network and measure the extent of homophily and diffusion assortativity along several subgroup characterization.

Given the Maximum likelihood fit to an observed network, we can generate the distribution of diffusion assortativity in a Monte Carlo fashion through repeated sampling of networks from the fitted model, $\widehat{P}$, and computing their assortativity along paths of length 2. This re-sampling procedure to compare other statistical and topological properties of the simulated network not explicitly accounted for in the model with the observed network has been used in previous works [62, 81, 82, 176]. This procedure, similar to posterior predictive checks in the Bayesian context [85], can be used to evaluate the fitness of a model beyond the scope it was designed for. In our case, this process reveals that the observed assortativity on paths of length 2 among the 56 schools is consistently higher than the distribution of diffusion assortativity expected by DCSBM, among all grouping attributes. For example, fitting the DCSBM based on gender fits assortativity on paths of length 1 perfectly, but 31 out of 56 schools (55%) exhibit higher assortativity on paths of length 2 than predicted by the fitted model, with two-tailed p-values less than 0.05. Similarly, DCSBM fit based on gender-grade groups (up to 6 groups) leads to 43 schools (76%) with significantly higher ($p < 0.05$) diffusion assortativity than predicted by the model.

Figure 4.3 compares the observed assortativity on paths of length 2 and 3 based on both gender and race (encoded as majority or other) groups with the estimated value from the maximum likelihood model. Even though the observed assortativity on paths of length 1 is always covered by its 95% confidence interval and very close to the point estimate, the fitted models consistently underestimate diffusion assortativities. The model under-estimates assortativity on paths of length 3 even more than paths of length 2. Furthermore, DCSBM becomes more inaccurate at predicting diffusion assortativity for smaller values of assortativity. For example, racial assortativity (on paths of length 1) in the schools ranges from 0.02 to 0.32 as opposed to gender which ranges from 0.43 to 0.83, and figure 4.3 shows that the scale of underestimation is larger for race than gender.

A possible explanation for these discrepancies is the unequal distribution of cross-group edges in the observed networks, while SBM assumes uniform distribution of cross-group edges among all pairs. High brokerage in a network would suggest that

a small fraction of nodes in each group hold a large fraction of out-group edges. Controlling for degree, a more equal distribution of cross-type edges would create extra paths of length 2, thus reducing diffusion assortativity.

5 Model

In this section, we describe our model that accounts for heterogeneity in out-group edge formation or brokerage and by doing so provides a more accurate estimate of diffusion assortativity. Before explaining the model, we restate important concepts and assumptions made in the model.

Directed Networks: We assume the network is directed as most social networks do have a notion of direction in edges. As we see later, this assumption is necessary for the estimate on the number of diffusion paths to be unbiased. The same model also applies to undirected networks, although it introduces a positive bias in the number of in-group diffusion paths, which vanishes as the size of network grows.

Higher Order Networks: Given a network G, its k^{th} order network $G^{(k)}$ determines the presence or lack of paths of length k (of unique edges) between any pair of nodes in G. For example, the second order network is a multi-graph which has as many edges between a pair of nodes as there are number of paths of length 2 between them in the original network. Since the original network is directed, its diffusion paths and its higher order networks will be directed too.

Self-Loops: We assume that the observed networks do not have self-loops, even though our model allows for it and can certainly generate networks with self-loops. While we assume the first-order observed network does not have self-loops, its higher order networks do (imagine paths of length 2 that start with and end in the same node) and counting them is necessary to obtain an unbiased estimate of diffusion paths. The presence of self-loops in the observed network leads to a positive bias in the number of in-group diffusion paths and consequently the estimated diffusion assortativity. However, this bias vanishes as the size of network grows.

Adjacency Matrix: The (i, j) element contains the number of outgoing stubs from

node i to node j. In contrast to undirected DCSBM [114], diagonal elements contain the number of self-loops, not twice their value, since self-edges are directed and each has only one outgoing stub.

Higher Order Assortativities: Higher order assortativities measure the extent of unequal diffusion in the network. The k$^{\text{th}}$ order assortativity of network G is simply the assortativity of its k$^{\text{th}}$ order network $G^{(k)}$. For example, if we denote the directed adjacency matrix of the second order network as $\mathbf{A}^{(2)}$, then we can define the second order assortativity, $r^{(2)}$, in a manner similar to equations 4.1 and 4.2.

$$r^{(2)} = \frac{\sum_r e_{rr}^{(2)} - \sum_r a_r^{(2)} b_r^{(2)}}{1 - \sum_r a_r^{(2)} b_r^{(2)}} \tag{4.6}$$

where the quantity $e_{rs}^{(2)}$ is the fraction of total (directed) paths of length 2 from a node in group r to a node in group s, $a_r^{(2)}$ is the fraction of total directed paths of length 2 from a node in group r and $b_r^{(2)}$ is the fraction of total paths of length 2 to a node in group r in the original network.

$$e_{rs}^{(2)} = \frac{\sum_{i,j} A_{ij}^{(2)} \delta_{g_i,r} \delta_{g_j,s}}{\sum_{i,j} A_{ij}^{(2)}} \qquad a_r^{(2)} = \sum_s e_{rs}^{(2)} \qquad b_r^{(2)} = \sum_s e_{sr}^{(2)} \tag{4.7}$$

5.1 Setup

Our random graph model is based on the degree-corrected Stochastic Block Model [114]. In contrast to DCSBM and instead of correcting for the total degree of each node, we correct for its degree to each group. By correcting for the out-group degree of each node, we can differentiate between networks whose cross-group links are exclusive to a small number of brokers versus those with an equal distribution of cross-group links. We show that by including extra parameters for this correction, the model not only corrects for the degree of each node, but also fits the number of in-group and out-group paths of length 1 and 2 in expectation and as a result the estimated assortativity on paths of length 2 is approximately equal to its observed value.

The main difference with DCSBM and our model is that after conditioning on degree, cross-group links are not distributed equally among all nodes of a group. Instead, each node will have a separate parameter for propensity of linking with each group and the combination of these cross-group propensity parameters determines how cross-group edges are distributed among nodes of a group. Furthermore, as the network is directed, we introduce one such node-level parameter and one group-level baseline linking parameter for each incoming and outgoing direction. Given these parameters, the number of edges from a node i from group r to a node j from group s is modeled as a Poisson random variable with mean $\theta^o_{i,s}\theta^i_{j,r}\omega_{rs}$ where $\theta^o_{i,s}$ is the outgoing propensity parameter for node i to group s, $\theta^i_{j,r}$ is the incoming propensity parameter for node j from group r and ω_{rs} parameter controls the baseline number of edges from group r to s. Thus, the expected value of A_{ij} element in the adjacency matrix is $\theta^o_{i,g_j}\theta^i_{j,g_i}\omega_{g_ig_j}$. A nice property of this model over DCSBM is that the expected number of self-loops match the expected value of their corresponding diagonal elements without an extra $\frac{1}{2}$ factor since we only count the number of outgoing stubs in the adjacency matrix of a directed network.

We can now express the likelihood function in this model with node degree variation in cross-group linking:

$$L(\Theta, \Omega; \mathbf{A}) = \prod_{i,j} \frac{(\theta^o_{i,g_j}\theta^i_{j,g_i}\omega_{g_ig_j})^{A_{ij}}}{A_{ij}!} \exp(-\theta^o_{i,g_j}\theta^i_{j,g_i}\omega_{g_ig_j}) \tag{4.8}$$

where Θ is the set of node-level outgoing and incoming degree propensity parameters, Ω is the group-level edge formation propensity parameters, g_i denotes the group of node i and $\mathbf{A}$ is the (directed) adjacency matrix where A_{ij} is the number of outgoing edges from node i to j. Given this setup, the MLE for Ω is as followed:

$$\widehat{\omega}_{rs} = \frac{m_{rs}}{\sum_{i\in r, j\in s} \widehat{\theta}^o_{i,s}\widehat{\theta}^i_{j,r}} \tag{4.9}$$

where m_{rs} is the number of outgoing edges from group r to group s. The denominator resembles the effective number of pairs for such links. To derive the MLE for Θ, we

note that θ parameters can be arbitrary to within a constant, therefore we must impose additional structure on the model. These constraints can take different forms and one of our contributions is to show that different constraints lead to different models. Below we briefly discuss two constraints and derive their resulting MLE. Throughout, we refer to set of all groups as G, the set of all nodes as N, the group to which node i belongs as g_i, total number of edges from group r to s as m_{rs}, total out-degree (in-degree) of node i as d_i^o (d_i^i) and the out-degree (in-degree) of node i to group r as $d_{i,r}^o$ ($d_{i,r}^i$).

5.2 Node Level Constraint

One alternative for model structure is to impose a constraint on total propensity of each node, as shown below.

$$\forall i \in N: \quad \sum_{g \in G} \theta_{i,g}^o = 1, \quad \sum_{g \in G} \theta_{i,g}^i = 1 \tag{4.10}$$

This constraint imposes the same fixed value on total propensity of linking to and from all groups for each node. It still allows for cross-group linking variation within each group, as each node can distribute its linking propensity differently. However, the constraint limits the degree variation of all nodes, in a manner similar to regular SBM without any degree correction. The MLE of this model simplifies to a system of equations, as shown below.

$$\begin{aligned}
\forall i \in N \quad \forall g_1, g_2 \in G: \quad & \sum_{j \in g_1} (\widehat{\omega}_{g_1 g_i} \widehat{\theta}_{j,g_i}^o - \frac{A_{ji}}{\widehat{\theta}_{i,g_1}^i}) = \sum_{j \in g_2} (\widehat{\omega}_{g_2 g_i} \widehat{\theta}_{j,g_i}^o - \frac{A_{ji}}{\widehat{\theta}_{i,g_2}^i}) \\
\forall i \in N \quad \forall g_1, g_2 \in G: \quad & \sum_{j \in g_1} (\widehat{\omega}_{g_i g_1} \widehat{\theta}_{j,g_i}^i - \frac{A_{ij}}{\widehat{\theta}_{i,g_1}^o}) = \sum_{j \in g_2} (\widehat{\omega}_{g_i g_2} \widehat{\theta}_{j,g_i}^i - \frac{A_{ij}}{\widehat{\theta}_{i,g_2}^o})
\end{aligned} \tag{4.11}$$

Combining the MLE equations 4.9 and 4.11 with the constraint equations 4.10, one can numerically compute the MLE. In general, the maximum likelihood estimates don't have a closed-form solution, but if the observed out-degree and in-degree of all nodes within each group are identical, the MLE takes the following convenient and

intuitive form.

$$\begin{aligned} &if \quad \forall r \in G \quad \forall i \in r \quad d_i^o = d_r^o : \quad \widehat{\theta}_{i,s}^o = \frac{d_{i,s}^o}{d_i^o} \\ &if \quad \forall r \in G \quad \forall i \in r \quad d_i^i = d_r^i : \quad \widehat{\theta}_{i,s}^i = \frac{d_{i,s}^i}{d_i^i} \end{aligned} \tag{4.12}$$

Where d_r^o (d_r^i) denotes the total out-degree (in-degree) of any node in group r. This result implies that the propensity of linking to a group s is simply the observed fraction of the node's total degree to that group.

5.3 Group Level Constraint

Another alternative for model structure is to impose a constraint on total propensity of all nodes within a group, as shown below. This model will be the main focus of our work and has close resemblance to DC-SBM but with extra desirable properties.

$$\forall r, s \in G : \quad \sum_{i \in r} \theta_{i,s}^o = 1, \quad \sum_{i \in r} \theta_{i,s}^i = 1 \tag{4.13}$$

The constraint states that the total propensity of linking to and from group s is fixed among all nodes of group r. Variation in cross-group linking among nodes of a group can still exist. Naturally, a good model will distribute the propensity supply of each group according to cross-group degree of the nodes within that group. The MLE of the model simplifies to the following intuitive forms:

$$\begin{aligned} &\forall r, s \in G \quad \forall i \in r : \quad \widehat{\theta}_{i,s}^o = \frac{d_{i,s}^o}{m_{rs}} \\ &\forall r, s \in G \quad \forall i \in r : \quad \widehat{\theta}_{i,s}^i = \frac{d_{i,s}^i}{m_{rs}} \end{aligned} \tag{4.14}$$

In contrast to the previous constraint at the node-level which led to within-node fractions, the MLE for linking propensity to a group s with the group-level constraint becomes the within-group fraction: the observed fraction of total cross-group degree that originates from the focal node. This estimate closely resembles that of the propensity parameter in regular DC-SBM with the exception that MLE fractions in

DC-SBM did not differentiate between the degrees to each group. Given the estimates above for propensity parameters, the MLE for group-level parameters from equation 4.9 simplifies to the number of cross-group edges:

$$\widehat{\omega}_{rs} = m_{rs} \tag{4.15}$$

5.4 Frequency of Diffusion Paths Under MLE Model

Before deriving the expected number of cross-group edges from the model fit, we compute a few useful parameters that result from the fitted model: the expected number of edges between any two nodes and the expected out-degree (in-degree) of a node to a group. The variables with a hat are generated by the model and refer to the corresponding observed quantity with same symbol.

$$E[\widehat{A}_{ij}] = \frac{d^o_{i,g_j} d^i_{j,g_i}}{m_{rs}} \tag{4.16}$$

$$E[\widehat{d}^o_{i,s}] = E[\sum_{j \in s} \widehat{A}_{ij}] = d^o_{i,s} \tag{4.17}$$

$$E[\widehat{d}^i_{i,s}] = E[\sum_{j \in s} \widehat{A}_{ji}] = d^i_{i,s} \tag{4.18}$$

We now show that the fitted model matches not only the observed number of paths of length 1 but also the observed number of paths of length 2 between any two groups in expectation, even though the model does not explicitly account for it. Throughout, we assume that traversing the same edge twice is not permissible (e.g. paths cannot use a self-loop twice). However, traversing from a node to its neighbor and back to itself is allowed as long as there is a directed edge in each direction. This is possible under our analysis since edges with different directions between any pair are drawn independently and considered different.

As the first step, we show that that expected number of edges between any two groups in the fitted model matches that of the observed network. Below, we denote the observed and (random) model-generated number of paths of length k from group r to group s by $P^{(k)}_{rs}$ and $\widehat{P}^{(k)}_{rs}$ respectively.

$$
\begin{aligned}
P^{(1)}_{rs} &= \sum_{i \in r} d^o_{i,s} = m_{rs} \\
\widehat{P}^{(1)}_{rs} &= \sum_{i \in r} \sum_{j \in s} \widehat{A}_{ij} \\
E[\widehat{P}^{(1)}_{rs}] &= \sum_{i \in r} \sum_{j \in s} \frac{d^o_{i,s} d^i_{j,r}}{m_{rs}} = m_{rs} \\
E[\widehat{P}^{(1)}_{rs}] &= P^{(1)}_{rs}
\end{aligned} \tag{4.19}
$$

We used equation (4.16) in the third line above. Therefore, the expected number of cross-group edges from the MLE model matches its observed value. We now show a similar result for paths of length 2. First, we show that the expected number of paths of length 2 between two different groups matches the observed network.

$$
\begin{aligned}
P^{(2)}_{rs} &= \sum_{j} d^i_{j,r} d^o_{j,s} \\
\widehat{P}^{(2)}_{rs} &= \sum_{j} \sum_{i \in r} \widehat{A}_{ij} \sum_{k \in s} \widehat{A}_{jk} \\
E[\widehat{P}^{(2)}_{rs}] &= \sum_{j} \sum_{i \in r, k \in s} E\big[\widehat{A}_{ij} \widehat{A}_{jk}\big] \\
&= \sum_{j} \sum_{i \in r, k \in s} E\big[\widehat{A}_{ij}\big] E\big[\widehat{A}_{jk}\big] \\
&= \sum_{j} E\big[\widehat{d}^i_{j,r}\big] E\big[\widehat{d}^o_{j,s}\big] \\
&= \sum_{j} d^i_{j,r} d^o_{j,s} \\
E[\widehat{P}^{(2)}_{rs}] &= P^{(2)}_{rs}
\end{aligned} \tag{4.20}
$$

In the fifth line above, we used the fact that edges are independent and in the seventh line, we relied on equations (4.17) and (4.18). We now show that the expected number of paths of length 2 within a single group is also the same as the observed value, keeping in mind that traversing any edge, including self-loops, is allowed only once.

$$
\begin{aligned}
P_{rr}^{(2)} &= \sum_j d_{j,r}^i d_{j,r}^o \\
\widehat{P}_{rr}^{(2)} &= \sum_j \sum_{\substack{i,k\in r \\ i\neq k}} \widehat{A}_{ij}\widehat{A}_{jk} + \sum_j \sum_{\substack{i\in r \\ i\neq j}} \widehat{A}_{ij}\widehat{A}_{ji} + \sum_{j\in r} \widehat{A}_{jj}(\widehat{A}_{jj}-1) \\
E\big[\widehat{P}_{rr}^{(2)}\big] &= \sum_j \sum_{\substack{i,k\in r \\ i\neq k}} E\big[\widehat{A}_{ij}\big]E\big[\widehat{A}_{jk}\big] + \sum_j \sum_{\substack{i\in r \\ i\neq j}} E\big[\widehat{A}_{ij}\big]E\big[\widehat{A}_{ji}\big] + \sum_{j\in r}\big(E\big[\widehat{A}_{jj}^2\big] - E\big[\widehat{A}_{jj}\big]\big) \\
&= \sum_j \sum_{\substack{i,k\in r \\ i\neq k}} E\big[\widehat{A}_{ij}\big]E\big[\widehat{A}_{jk}\big] + \sum_j \sum_{\substack{i\in r \\ i\neq j}} E\big[\widehat{A}_{ij}\big]E\big[\widehat{A}_{ji}\big] + \sum_{j\in r} E\big[\widehat{A}_{jj}\big]^2 \\
&= \sum_j \sum_{i,k\in r} E\big[\widehat{A}_{ij}\big]E\big[\widehat{A}_{jk}\big] \\
&= \sum_j E\big[\widehat{d}_{j,r}^i\big]E\big[\widehat{d}_{j,r}^o\big] \\
&= \sum_j d_{j,r}^i d_{j,r}^o \\
E[\widehat{P}_{rr}^{(2)}] &= P_{rr}^{(2)}
\end{aligned}
\tag{4.21}
$$

The first line above uses our assumption that the observed network does not have any self-loops. The second line uses the fact that traversing the same edge is not allowed twice. In the the third line, we relied on the independence of edges, with the exception of the last term which refers to the number of different self-loops that are determined from a single Poisson draw. We note that if the observed network has self-loops, then the estimated number of within-group paths of length 2 would be biased positively by the total number of self-loops within the group.

5.5 Asymptotic Behavior of Diffusion Assortativity Under MLE Model

First, we quickly prove a simple extension of weak law of large numbers which we will use in our proof of diffusion assortativity consistency.

Proposition 1. *Let* $\{X_i\}_1^\infty$ *be a sequence of independent random variables with*

$E[X_i] = \mu_i$ *and* $Var(X_i) = \sigma_i^2$. *If the sequence of variances* $\{\sigma_i^2\}_1^\infty$ *is bounded, then* $\frac{\sum_i^n X_i}{n} \to \frac{\sum_i^n \mu_i}{n}$ *in probability.*

Proof. Let $S_n = \frac{\sum_i^n X_i}{n}$ and $\mu = \frac{\sum_i^n \mu_i}{n}$, then $Var(S_n) = \frac{\sum_i^n \sigma_i^2}{n^2} \to 0$. This follows from simple application of Chebychev's inequality.

$$P(|S_n - \mu| \geq \epsilon) \leq \frac{Var(S_n)}{\epsilon^2} \to 0$$

□

Proposition 2. *Let* n_r *be the size of nodes in group* r *and* $n = \sum_r n_r$ *be the size of all nodes in the network. If* $\hat{e}_{rs}^{(2)}$ *is determined from the sampled network according to equation 4.7 and* $\mathbf{A} \neq 0$*, then* $\hat{e}_{rs}^{(2)} \xrightarrow{p} e_{rs}^{(2)}$ *as* $n \to \infty$.

Proof. Below we denote the adjacency matrix of the second order network of the sampled network as $\hat{\mathbf{A}}^{(2)}$. We allow for traversing the same edge multiple times and show that prohibiting them does not affect the result.

$$\begin{aligned}
\hat{e}_{rs}^{(2)} &= \frac{\sum\limits_{i\in r, j\in s} \hat{A}_{ij}^{(2)}}{\sum\limits_{i,j} \hat{A}_{ij}^{(2)}} \\
&= \frac{\sum\limits_{i\in r, j\in s, k} \hat{A}_{ik}\hat{A}_{kj}}{\sum\limits_{i,j,k} \hat{A}_{ik}\hat{A}_{kj}} \\
&= \frac{n_r n_s}{n^2} \frac{\frac{\sum\limits_{i\in r, j\in s, k} \hat{A}_{ik}\hat{A}_{kj}}{n_r n_s n}}{\frac{\sum\limits_{i,j,k} \hat{A}_{ik}\hat{A}_{kj}}{n^3}}
\end{aligned} \tag{4.22}$$

In the second line above, we allowed for traversing the same edge twice. This can happen only if $i = j = k$ (self-loops). The terms $\hat{A}_{ik}$ and $\hat{A}_{kj}$ are two Poisson random variables with finite mean and variance, thus their product also has finite mean and variance. By applying proposition 1, we get

$$\frac{\sum\limits_{i\in r, j\in s, k} \hat{A}_{ik}\hat{A}_{kj}}{n_r n_s n} \xrightarrow{p} \frac{\sum\limits_{i\in r, j\in s, k} E[\hat{A}_{ik}\hat{A}_{kj}]}{n_r n_s n} \tag{4.23}$$

The terms $\hat{A}_{ik}$ and $\hat{A}_{kj}$ are independent unless $i = j = k$ which can only happen if $r = s$. Below we assume this is the case but the results remain the same even if $r \neq s$.

$$\begin{aligned}\lim_{n\to\infty} \frac{\sum_{i\in r, j\in s, k} E[\hat{A}_{ik}\hat{A}_{kj}]}{n_r n_s n} &= \lim_{n\to\infty} \frac{\sum_{i\in r, j\in s, k} E[\hat{A}_{ik}]E[\hat{A}_{kj}]}{n_r n_s n} + \lim_{n\to\infty} \frac{\sum_{i\in r} E[\hat{A}_{ii}]}{n_r n_s n} \\ &= \lim_{n\to\infty} \frac{\sum_{i\in r, j\in s} d_i^o d_j^i}{n_r n_s n}\end{aligned} \tag{4.24}$$

In the second line we used the fact that expected number of self-loops is finite. Combining the result above with equation 4.23 and performing the same analysis for the denominator in equation 4.22, we get the following convergences.

$$\begin{aligned}\frac{\sum_{i\in r, j\in s, k} \hat{A}_{ik}\hat{A}_{kj}}{n_r n_s n} &\xrightarrow{p} \frac{\sum_{i\in r, j\in s} d_i^o d_j^i}{n_r n_s n} \\ \frac{\sum_{i,j,k} \hat{A}_{ik}\hat{A}_{kj}}{n^3} &\xrightarrow{p} \frac{\sum_{i,j} d_i^o d_j^i}{n^3}\end{aligned} \tag{4.25}$$

Combining equations 4.22 and 4.25 and the fact that $\sum_{i,j} d_i^o d_j^i \neq 0$, we get the result using the continuous mapping theorem:

$$\hat{e}_{rs}^{(2)} \xrightarrow{p} e_{rs}^{(2)} \tag{4.26}$$

□

Remark 1. If we had not allowed for traversing the same edge multiple times, the second term in equation 4.24 would not be present and the final limit would be the same.

Remark 2. In case of an undirected network, we would have the same convergence results as long as $n_s \to \infty$ for all s when $n \to \infty$. In this case, the second term in equation 4.24 would be replaced by $\frac{\sum_{i\in r, k} E[\hat{A}_{ik}]}{n_r n_s n}$ which still tends to zero as $n \to \infty$.

Proposition 3. *The sampled assortativity on paths of length 2 from the MLE model converges in probability to the observed assortativity on paths of length 2.*

Proof. The assortativity on paths of length 2 from a sampled network is defined as below.

$$\hat{r}^{(2)} = \frac{\sum_r \hat{e}_{rr}^{(2)} - \sum_r \hat{a}_r^{(2)} \hat{b}_r^{(2)}}{1 - \sum_r \hat{a}_r^{(2)} \hat{b}_r^{(2)}} \qquad \hat{a}_r^{(2)} = \sum_s \hat{e}_{rs}^{(2)} \qquad \hat{b}_r^{(2)} = \sum_s \hat{e}_{sr}^{(2)}$$

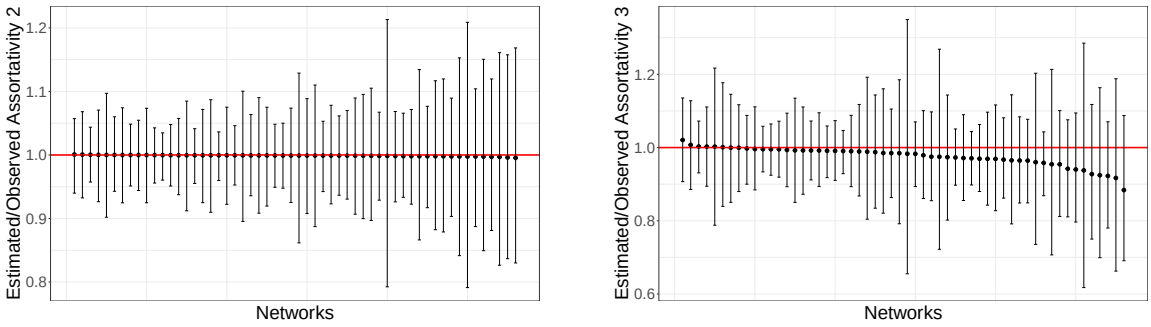
where all quantities $\hat{e}_{rs}^{(2)}, \hat{a}_r^{(2)}, \hat{b}_r^{(2)}$ are determined from the sampled network. The result follows using proposition 2 on each individual term of $\hat{r}^{(2)}$ and the continuous mapping theorem. In the application of continuous mapping theorem we rely on $\sum_r a_r^{(2)} b_r^{(2)} < 1$ since $\sum_{r,s} e_{rs}^{(2)} = 1$. □

Remark. In case of an undirected network, the same result holds as long as $n_s \to \infty$ for all s when $n \to \infty$.

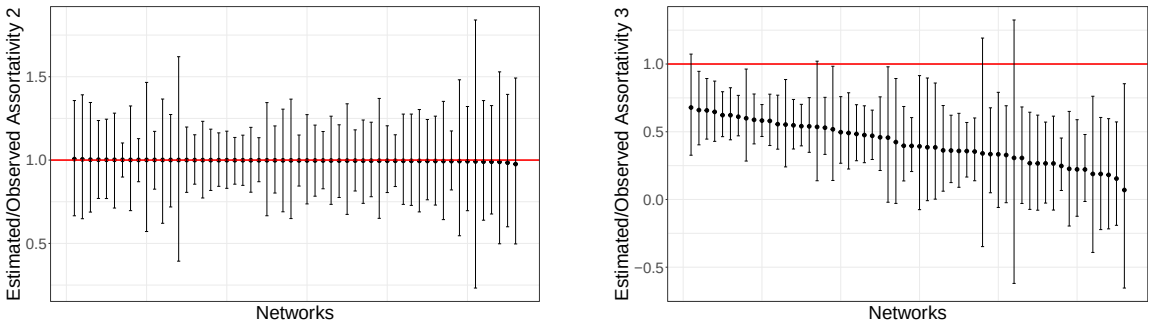
6 Results

In this section, we show the same results as in section 4.4 but using our (directed) model instead of the (directed) DCSBM. In particular, we consider the same networks as before and compare their observed assortativity on paths of length 2 and 3 versus the distribution of those quantities generated by the MLE model. As shown above, we would expect the observed assortativity on paths of length 2 to be close to the predicted value by the fitted model, since the networks are large enough. Even if the networks are not large enough for proposition 3 to be valid, the bias in model-generated assortativity should be small since the number of in-group and out-group paths of length 2 from the model match the corresponding observed values in expectation, as shown in section 5.4.

Figure 4.4 compares the distribution of assortativites generated by the model against the observed values along gender and racial groups. The figure is produced exactly as figure 4.3, with the same networks and attributes, except that the fitted model accounts for brokerage. First, we observe that in contrast to regular DCSBM, our model accurately captures assortativity along paths of length 2 for both attributes.



(a) Gender



(b) Race

Figure 4.4: The distribution of predicted over observed ratio of assortativities on paths of length 2 (left column), $\frac{\hat{r}^{(2)}}{r^{(2)}}$, and 3 (right column), $\frac{\hat{r}^{(3)}}{r^{(3)}}$, from our model along gender (top row) and racial (bottom row) groups. Bars correspond to 95% confidence interval and each bar corresponds to one school network. Networks are sorted in descending order of the point estimate.

This is expected since our model fitted through MLE matches the observed frequency of paths of length 2 in expectation. Second, Even though our model does not make any guarantees about assortativity on longer paths, it nevertheless provides a close match with observed assorativity on paths of length 3, at least along gender groups. The observed gender assortativity on paths of length 3 is not significantly different from the generated distribution by the model in any of the network. However, the model consistently underestimates racial assortativity along paths of length 3. This is mainly because the the absolute level of assortativity along race is much smaller than gender, with values that are often close to zero (only 12 out 56 networks have racial assortativity greater than 0.1). At such small values the model requires higher precision and small absolute differences can make the its predictions significantly

different relative to the observations. Nevertheless, comparing the distribution of generated racial assortativities along paths of length 3 in figure 4.4 with figure 4.3, we observe that our model's predictions are at least an order of magnitude closer to observations than DCSBM.

7 Model Validation

In this section, we attempt to provide some evidence that our model indeed captures salient patterns in the networks, in ways that lead to statistically significant improvements in its goodness of fit over DCSBM as the null model. To compare the model against the regular DCSBM, we use the likelihood ratio test with our model as the alternative. The challenge with this approach is the difficulty to determine the exact distribution of the likelihood ratio statistic. We could appeal to the Wilks theorem [24] which provides a convenient form for the asymptotic distribution of the likelihood ratio, so long as DCSBM (null) is nested within our model. DCSBM is in fact a special case of our model since it imposes a homogeneity constraint on cross-group linking propensities of our model. In particular, we obtain the regular directed DCSBM if we make the following assumption in our model:

$$\forall j \in N,\ \forall r, s \in G: \quad \theta^o_{j,r} = \theta^o_{j,s}, \quad \theta^i_{j,r} = \theta^i_{j,s} \tag{4.27}$$

Given the nested models and under Wilks theorem, the test statistic $-2\log(\hat{\lambda})$ (with $\hat{\lambda}$ as the likelihood ratio) is asymptotically distributed as chi-squared with the number of constraints that we must impose on our model to obtain DCSBM as its degrees of freedom. This type of hypothesis testing that uses approximate likelihood ratio chi-squared statistic for network models has been used before [174].

The χ^2 distribution of Wilks theorem assumes that the log-likelihood of both the null and alternative models are well-behaved and resemble a quadratic function close to their maximum [24]. The justification for this assumption is the central limit theorem together with a growing "effective large sample size". This would be true in

the case of dense graphs (i.e. individual degree tends to infinity as number of nodes increases), since the effective data size in dense graphs is of order $O(n^2)$ and even though there are $O(n)$ parameters in DCSBM, the model fit would still be in the large data limit [179]. However, in the sparse regime, the Wilks theorem is no longer valid on DCSBM due to the combination of growing parameters and network sparsity [79]. In sparse graphs, the effective samples size and number of parameters grow at the same rate of $O(n)$. As there is only $O(1)$ observations per each parameter, we never satisfy the large sample assumptions behind Wilks theorem and as it has been shown before the usual χ^2 distribution often underestimates the likelihood ratio in sparse graphs [179].

As most social networks fall into the sparse regime, we need alternative methods to approximate the distribution of our log-likelihood ratio shown below.

$$\hat{\lambda} = \log \frac{\sup_{(\Theta,\Omega)\in\mathcal{P}} L(\Theta, \Omega; \mathbf{A})}{\sup_{(\Theta,\Omega)\in\mathcal{P}_0} L(\Theta, \Omega; \mathbf{A})} \tag{4.28}$$

where $\mathcal{P}_0$ and $\mathcal{P}$ denote the restricted and full model parameter spaces respectively. Large values of $\hat{\lambda}$ test statistic indicates support for the full model that it provides statistically significant improvements over DCSBM. One approach to find the null distribution of $\hat{\lambda}$ is through parametric bootstrap [68] where we first fit the observed network against the null (DCSBM), then repeatedly draw new networks from the fitted model and compute their $\hat{\lambda}$ to generate its distribution under the null.

Before conducting the log-likelihood ratio (LLR) test on our model, we first test the validity of the parametric bootstrap procedure explained above. In particular, we draw 5 random networks from each of 56 DCSBM model fits against our actual networks and generate their LLR distribution via bootstrap (total of 280 networks). These randomly drawn networks should resemble our actual networks if they were generated by a DCSBM. In this test, these synthetic networks act as our observed networks and are truly from the null model. Hence, we expect their observed LLR to fall within the bootstrapped distribution if the procedure generates a valid LLR distribution. Figure 4.5 illustrates this comparison for these synthetic networks. For

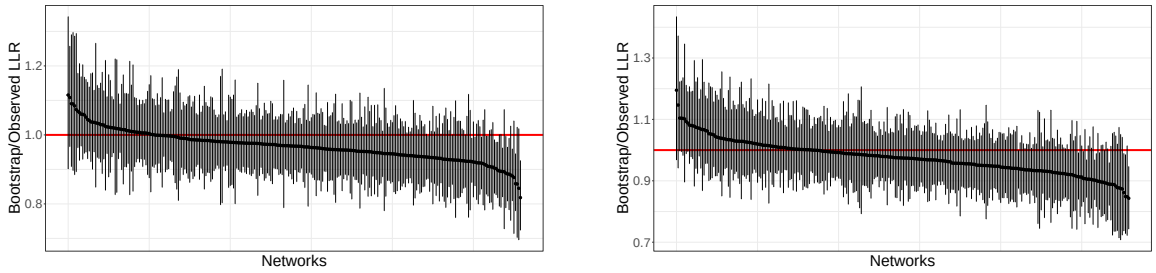


Figure 4.5: The bootstrapped distribution of log-likelihood ratio scaled by the actual log-likelihood ratio (LLR) when the true model is actually a DCSBM fit. Each of 56 DCSBM models corresponds to the MLE of a school network based on gender (left) and ethnicity (right) groups. For each fitted DCSBM, we draw 5 random networks and generate the bootstrap distribution of the LLR from that realization. The 280 synthetic networks are sorted by the mean of their LLR bootstrap distribution. Each bar corresponds to 95% confidence interval of one network.

easier comparison across all networks, the bootstrapped distribution is divided by the observed LLR, so that the value of 1 corresponds to observed LLR. The observed LLR is often within the 95% confidence region constructed by the bootstrap, however we can observe that the bootstrapped distribution systematically under-estimates the observed LLR. This indicates a potential for over-rejection of the null that is again induced by the network sparsity.

Even though there is a slight under-estimation of LLR under parametric bootstrap, we don't expect it would change our conclusions from testing our model against the DCSBM. This is because the log-likelihood ratios from our observed networks are so far outside the bootstrapped distribution that the p-value in almost all networks is practically zero and the slight under-estimation by the bootstrap procedure should not make a substantial difference on our inference. Figure 4.6 compares the observed LLR from our 56 networks against its bootstrapped distribution assuming DCSBM as the null model, in a manner similar to figure 4.5 except that here we use actual instead of synthetic networks. The observed value is considerably greater than the null distribution and we can safely conclude our model provides significant improvements on the goodness of fit over DCSBM in a way that a slight under-estimation observed in figure 4.5 won't substantially change our conclusion.

In conclusion, the likelihood ratio test provides strong support for our model over

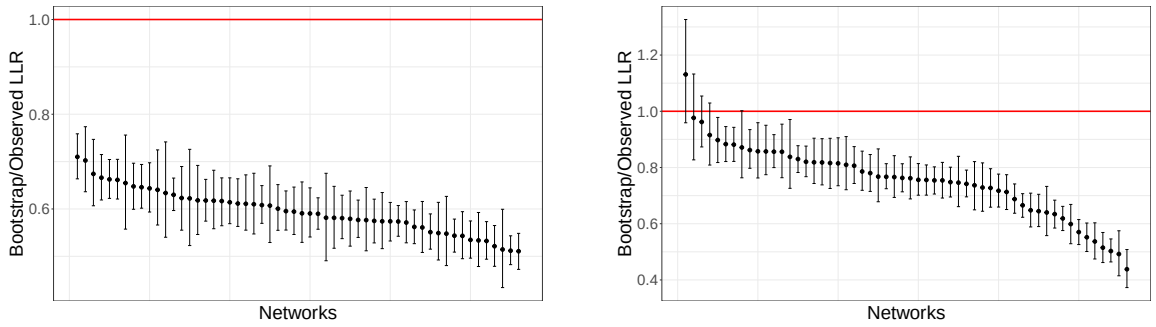


Figure 4.6: The bootstrapped distribution of log-likelihood ratio (LLR) scaled by the actual LLR from each of the 56 networks using gender (left) and ethnicity (right) as the blocking variable in DCSBM. Each bar corresponds to the (scaled) LLR 95% confidence interval obtained through (parametric) bootstrap of one network assuming DCSBM as its null model. The 56 school networks are sorted by the mean of their LLR bootstrap distribution.

regular DCSBM and we conclude it significantly improves the fit in all 56 school networks no matter if the grouping of nodes is conducting along low (race) or high (gender) homophilous attributes. This finding suggests that variation in cross-group linking is not simply a spurious pattern in social networks and network models would provide a more precise description of the network structure if they account for this pattern.

Chapter 5

Network Processes can Exacerbate Existing Inequalities

1 Preface

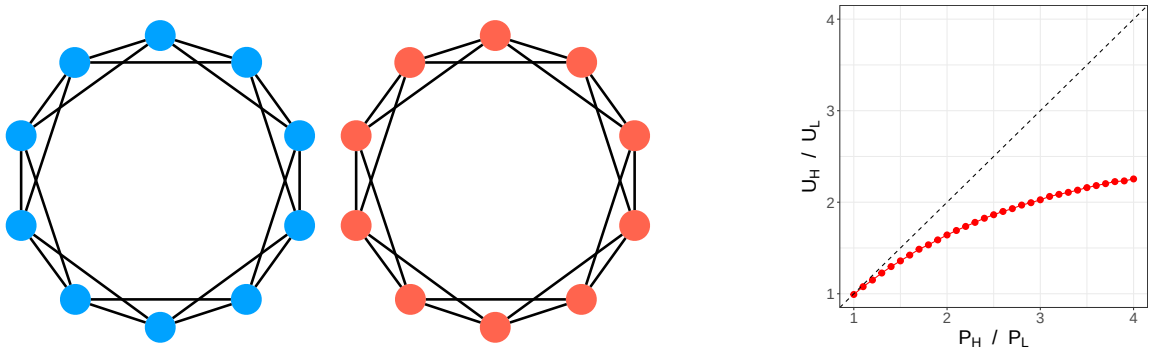
In the previous chapter, we developed a model of network structure with unequal distribution of cross-group links. The goal of the model was to show how network structure and specifically brokerage can lead to unequal diffusion of information and resources in the network. Our efforts up to know have only focused on the structure of the network and we have assumed individuals in the networks are myopic: they simply transfer their resources to their neighbors whenever they have them. But actors in social networks are often strategic and make decisions on who to help based on different criteria such as reciprocity or self-interest from future interactions. In this chapter, we explore the interaction of resource sharing and strategic decision making in networks. We present one such mechanism and show that strategic decision making exacerbates existing inequalities. We further verify the predictions of our model in a multi-player online platform we developed to study how micro decisions in networks have macro consequences in terms of inequality.

2 Introduction

Inequality has been consistently rising over the past 40 years in the US [143]. Its destabilizing forces and negative impact on economic growth have motivated academics and policy makers to address the roots of its persistence [104, 167]. Policy makers and the economics literature has mostly focused on the economic forces behind inequality such as market imperfections, tax policy or monopoly rents. However, there is increasing recognition that the roots of inequality trace back to social structures and how they interact with economic institutions [108]. Segregation or status homophily in network is believed to be main social driver behind inequality [71, 108, 121, 163, 170]. The link between social network homophily and inequality is based on unequal access to information. In basic terms, sorting in social networks by status and access to economic information leads to concentration of opportunities in a small part of the society, widening existing gaps over time.

The unequal diffusion of resources or opportunities is the basis of several studies which have provided a theoretical account of how small differences in individual advantage can translate into large and persistent differences over time [43, 71, 121]. The unequal diffusion is an informational account of network effects on inequality and occurs when valuable information is generated by different people at different times and the network exhibits three characteristics 1. Information diffuses across network ties, 2. One group generates the information or opportunities at a higher rate, 3. The network is homophilous in the group attribute. These three conditions make the networks of the advantaged group richer in resources, a phenomenon referred to as "inequality in social capital" [121]. Homophily is the main driver of the differential access to information and it implies that opportunities remain exclusive to one group, exacerbating existing differences over time.

Given the significant implications of networks in exacerbating inequalities, it is important to go beyond the general theoretical framework and determine the process in detail. Only then we can prescribe interventions to combat the forces that regenerate inequality. For example, a simple structural explanation based on information



(a) (b)

Figure 5.1: **Diffusion in homophilous networks does not necessarily widen inter-group difference.** Two circular lattices comprising the full (left). Nodes with the same color, a lattice, all have the same probability of generating valuable information in each round. One group has higher probability than the other. Lack of cross-group connections implies extreme homophily. The ratio of group utilities after accounting for one-hop diffusion for various levels of initial differences (right). Estimates are from 20 simulations each with 10,000 rounds. 95% confidence intervals are too small to be visible.

diffusion does not lead to larger than initial levels of inequality. Figure 5.1 shows the results of a simple simulation that confirms this point in a network with two disconnected circular lattices, the most extreme case of homophily. In each round, each node in the network independently generates valuable information with fixed probability and passes it along to its neighbors. The network has two disconnected components, each representing a group with fixed probabilities P_H and P_L, akin to status. Access to the information increases a node's utility by a unit in each round. If we denote the total utility of all nodes in each group after 10000 rounds by U_H and U_L, then the ratio $\frac{U_H}{U_L}$ signifies the level of inter-group difference accounting for the network effects. This ratio is compared against the initial inter-group difference $\frac{P_H}{P_L}$ in figure 5.1b, which indicate that the diffusion in the network actually reduces the inter-group difference by about 54% from the initial exogenous differences when $\frac{P_H}{P_L} = 4$.

The problem with this simple approach is that it is purely structural and ignores the fact that inequality arises from the incentive structure of processes that occur in the network and that inequality is durable because it's indeed the equilibrium state

of that process [162]. In other words, individuals in a social network are not myopic, instead they strategically form links and decide who to cooperate with. Our measures of social capital, in particular those capturing information or favor exchange capital [107], mostly focus on the network structure, and ignore its interaction with strategic behavior and human capital. But in reality, it is possible that the same expansive network structure for the poor does not provide the same informational benefits as it does for the rich, despite predicting similar level of informational social capital. Thus, it is imperative to account for the nuanced processes that occur within the network and lead to unequal access to valuable resources.

Another problem with the simple structural account above is the assumption that the valuable resource is non-rivalrous and individuals don't compete for accessing it. However in reality, many resources shared in social networks, such as employment information or rations or new business innovation opportunities, are rivalrous (their utility goes down as more people share it). This rivalry introduces strategic behavior in resource sharing because sharing will reduce own utility but might encourage anticipated reciprocity with contacts that improves utility in the future. Thus if an individual believes that the future gains it receives from its contacts do not compensate for the losses it incurs currently by sharing the valuable resources, then it might decide to withhold the resource from its contacts. This sort of strategic behavior resembles the conditionally cooperative behavior that has been illustrated in lab experiments [80]: people cooperate if they know others will also cooperate. This is very relevant in the case of inequality in endowments since if high type players anticipate their low type contacts cannot cooperate, they will in turn reduce their cooperation. Beyond the effects of rivalry on individual decision making, its macro effects at the group level outcomes are more intriguing. Does rivalry affect different groups differently? In this paper, we attempt to study information sharing processes in networks pertaining to rival resources and its implications on inequality.

This type of rivalry in resources and its effect on cooperation has been studied before. It is usually introduced by changing the number of individuals who compete for the same fixed resources. For example when the rivalrous resource is employment

information, Beaman finds that refugees who get resettled in locations with a larger community tend to have worse employment outcomes and wages because a larger pool of refugees will compete for the same fixed employment opportunities. Similarly in the context of cooperation and information sharing, there is evidence of crowding effects in public good games. Increasing the number of players sharing the same rivalrous common good decreases individual contributions and leads to worse welfare outcomes [103]. Similar patterns have been reported when individuals share valuable information with low-degree contacts fearing that sharing with high-degree contacts might lead to over-crowding on the rivalrous resource [17]. A similar force affects the choice of migration as migrant move to places where their contacts don't have too many friends they have to compete with [27]. Our work was specifically inspired by the derivation of pairwise stability in job contact networks which showed the positive correlation between employment outcome and the number of contacts, due to increasing information sources, but negative correlation with number of two-links-ways contacts, due to increased competition [42]. Similar to [42], we derive the pairwise stable subgame perfect equilibrium in our (rivalrous) information sharing model. But in addition we examine the inequality implications of rivalry by introducing heterogeneous agents which are present simultaneously in the network and adapt different strategies.

The effect of heterogeneity among agents on individual decisions and total welfare has also been studied. There have been mixed results mainly due to different setups. Some evidence suggests that heterogeneity does not affect cooperation rate in public good games and could sometimes even increase it [31, 54, 149]. In a context with collective risk such as climate change, Wang et al. show that heterogeneity leads to higher cooperation because the stakes are higher for richer agents and their cooperation incentivizes poorer agents to also cooperate [172]. The main question in all these studies is stated in terms of cooperation rate in non-rivalrous game with heterogeneity and the inequality implications are not considered. Our work examines the effects of agent heterogeneity and rivalry simultaneously in a network game and while our main goal concerns the inequality implications, we also derive how these factors affect the

cooperation.

In this paper, we develop a model that introduces a strategic sharing process among agents with heterogeneous levels of initial endowments in a network where the resources are rivalrous. The agents play a repeated game with their neighbors and in each round if they receive the rivalrous resource, they decide whether to share it with their contacts or not. If the heterogeneity is observable, then in equilibrium the agents will follow a conditional cooperation strategy: they will share the resource, if they know their contact will reciprocate in the future. We show that if the initial differences are large enough, the low type has no incentive to share information with their contact whereas the high type will, essentially leading to homophily in type. These micro-scale decisions made by individual have macro-scale implication at the group level, such that they will exacerbate inter-group differences if the initial differences are large enough. In terms of theoretical contributions, this simple model brings to light the importance of network processes involving complex decisions and the interaction of social capital with human capital in the study of inequality.

Furthermore, we implement a randomized multi-player online lab experiment closely resembling the model to validate its theoretical prediction. The advent of crowd-sourcing platforms such as Amazon Mechanical Turk have enabled researchers to develop and test their hypothesis in large-scale online lab experiment [126, 137, 156]. We conducted our experiment similar to these past works by recruiting participants from Amazon Mechanical Turk, and randomly assigning them to a binary type and a position in a homophilous network. Multiple participants played a game simultaneously over multiple rounds and made decisions whether to share rivalrous monetary rewards with each other. We find strong evidence of conditional cooperation as the low type cooperates at a much lower rate than the high type. The adoption of different strategies employed by different groups leads the high type to take a larger share of the rivalrous resources than expected by its initial endowment.

3 Model

We study a network process that exacerbates inter-group differences beyond what's expected by exogenous variation in individual ability. Our setup considers a game in which a rivalrous resource repeatedly diffuses in the network, access to which increases one's utility.

3.1 Game Setup

The game has infinite number of rounds with discount factor c. In each round, there exists a rivalrous resource with total value of 1 and all players that have access to the resource will equally share its utility. There are two types of players: there are n_H agents of high type and n_L agents of low type. There are more low type than high type agents: $n_L > n_H$. In each round, exactly one player of each type receives the resource with uniform within-type probability. A high type player will receive the resource with probability $p_H = \frac{1}{n_H}$ while a low type player will receive it with probability $p_L = \frac{1}{n_L}$. Given that there are more low type than high type agents, a high type agent is more likely to independently receive the resource: $p_H > p_L$. If an agent receives the resource, it has the option to share it with any of its network contacts. Since the resource is rivalrous, sharing it will reduce potential utility from the current round, but the agent still has incentive to share if it believes the contact has a high enough probability to receive the resource in the future and reciprocate. The combined strategy of all players leads to an undirected network structure that is endogenous to the game. A link appears in the network when both players' strategies are to share with each other. In the following we assume that each player can have at most a degree of d.

3.2 Pairwise Nash Stable Network

We will now derive the subgame perfect equilibrium (SPE) based on grim trigger strategies on each agent. The outcome will effectively describe a pairwise Nash stable network [26, 109] where the existence of links indicate sharing by both agents and

their absence indicates no sharing by either. For simplicity, we assume only ties within the same type are possible, hence the network will be maximally homophilous. After describing the SPE, we will argue that the same conclusions holds if we were to allow cross-type edges as well. Furthermore, we assume $n_L c$ and $n_H c$ are not integers to avoid a few uninteresting edge cases that are easy to solve for but greatly expand the set of possibilities to enumerate. We will briefly remark how the equilibrium looks like when these conditions are not true. In the following, we denote the equilibrium degree of each player by d_H^* and d_L^* for either type.

Theorem 3.1. *Assuming $n_L c$ and $n_H c$ are not integers, agents employ grim trigger strategies and only within-group sharing is possible, then*

1. *if $n_L > \frac{d+1}{c-1}$ and $n_H < \frac{1}{c-1}$, then $d_H^* = d$ and $d_L^* = 0$ in the pairwise Nash stable network. A circular lattice within each type is such a network.*
2. *if $n_L > \frac{d+1}{c-1}$ and $n_H > \frac{1}{c-1}$, then $d_H^* = d_L^* = 0$ in the pairwise Nash stable network.*
3. *if $\frac{1}{c-1} < n_L < \frac{d+1}{c-1}$ and $n_H > \frac{1}{c-1}$, then either $d_H^* = d_L^* = 0$ or $d_H^* = d_L^* = d$ in the pairwise Nash stable network.*
4. *if $\frac{1}{c-1} < n_L < \frac{d+1}{c-1}$ and $n_H < \frac{1}{c-1}$, then $d_H^* = d_L^* = d$ in the pairwise Nash stable network.*
5. *if $n_L < \frac{1}{c-1}$, then $d_H^* = d_L^* = d$ in the pairwise Nash stable network.*

Proof. The grim trigger strategy of each player is a binary vector corresponding to sharing decisions with all other players if having received the resource. Since players are exchangeable, we can simplify the notation and express the strategy of each player as the number of other players it is sharing with: d_H and d_L for either type. To derive SPE, we express the expected utility of a player from either type starting from the current round if the player has received the resource (the player does not take any

action if it does not receive the resource).

$$
\begin{aligned}
U_H(d_H, d_L) &= \frac{1}{d_H + d_L + 2} + \sum_{i=1}^{\infty} \frac{1}{c^i} \frac{1}{d_H + d_L + 2} \frac{d_H + 1}{n_H} \\
&= \frac{1}{d_H + d_L + 2}(1 + \frac{d_H + 1}{n_H(c-1)}) \\
U_L(d_H, d_L) &= \frac{1}{d_H + d_L + 2}(1 + \frac{d_L + 1}{n_L(c-1)})
\end{aligned} \tag{5.1}
$$

The term $\frac{1}{d_H+d_L+2}$ in equation 5.1 corresponds to utility from the shared resource in a round and the constant 2 refers to the original receivers of the resource, one from either type. The term $\frac{d_H+1}{n_H}$ in equation 5.1 denotes the probability of receiving the resource in future rounds by either the player itself or one of its neighbors. In SPE, each player maximizes utility starting in each round conditioned on receiving the resource:

$$
\begin{aligned}
d_H^* &= \underset{d_H \in \{0,1,\ldots,d\}}{\arg\max} \ \frac{1}{d_H + d_L^* + 2}(1 + \frac{d_H + 1}{n_H(c-1)}) \\
d_L^* &= \underset{d_L \in \{0,1,\ldots,d\}}{\arg\max} \ \frac{1}{d_H^* + d_L + 2}(1 + \frac{d_L + 1}{n_L(c-1)})
\end{aligned}
$$

The marginal utilities are:

$$
\begin{aligned}
u'_H(d_H) &= \frac{d_L^* - n_H(c-1) + 1}{n_H(c-1)(d_H + d_L^* + 2)^2} \\
u'_L(d_L) &= \frac{d_H^* - n_L(c-1) + 1}{n_L(c-1)(d_H^* + d_L + 2)^2}
\end{aligned}
$$

Depending on the sign of the numerator in the marginal utilities, we characterize the Nash stable network with different cases:

1. if $n_L > \frac{d+1}{c-1}$ then $u'_L(.) < 0$ for any value of d_H^*. Thus, the optimal choice for d_L is the lower corner point: $d_L^* = 0$. If $n_H < \frac{1}{c-1}$, then $u'_H(.) > 0$ and the optimal choice for d_H is the upper corner point: $d_H^* = d$.

2. if $n_L > \frac{d+1}{c-1}$ and $n_H > \frac{1}{c-1}$, we have $d_L^* = 0$ from the previous case. But now

$u'_H(.) < 0$, thus the optimal choice for d_H is the lower corner point: $d^*_H = 0$.

3. if $\frac{1}{c-1} < n_L < \frac{d+1}{c-1}$ and $n_H > \frac{1}{c-1}$, then $u'_L(.) < 0$ if $d^*_H = 0$, hence $d^*_L = 0$. Similarly, $u'_H(.) < 0$ if $d^*_L = 0$, hence $d^*_H = 0$. So $d^*_L = d^*_H = 0$ is one SPE, but there is another possible SPE. $u'_L(.) > 0$ if $d^*_H = d$, hence $d^*_L = d$ and similarly $d^*_H = d$ if $d^*_L = d$. So $d^*_L = d^*_H = d$ is another SPE. It is easy to see that mid-values for either d^*_H or d^*_L cannot be SPE, because we have assumed $n_L c$ and $n_H c$ are not integers, so the marginal utilities cannot be zero requiring the optimal choices to be corner points.

4. if $\frac{1}{c-1} < n_L < \frac{d+1}{c-1}$ and $n_H < \frac{1}{c-1}$, then $u'_H(.) > 0$ and the optimal choice for d_H is the upper corner point: $d^*_H = d$. Given $d^*_H = d$, then $u'_L(.) > 0$ and $d^*_L = d$.

5. if $n_L < \frac{1}{c-1}$, then $u'_L(.) > 0$ and subsequently $u'_L(.) > 0$ since $n_H < n_L$. Thus, $d^*_H = d^*_L = d$.

All equilibrium choices above have positive or negative marginal utility at the equilibrium depending on the corner point they occur in. Thus adding or severing links only reduce utility, which implies the solutions concepts above are also pairwise stable. Therefore, all equilibrium solutions above correspond to pairwise Nash stable networks. □

Remark 3.1.1. *As mentioned earlier, allowing for $n_L c$ or $n_H c$ to be integers do not lead to interesting predictions, but greatly expand the possible cases. For example, if we allow $n_H c$ to be an integer, then in addition to cases (1) and (2) in theorem 3.1, we will have yet another case as following: if $n_L > \frac{d+1}{c-1}$ and $n_H = \frac{1}{c-1}$, then $d^*_L = 0$ but now $d^*_H \in \{0, 1, ..., d\}$ since the high type player will always have zero marginal utility regardless of its choice. These edge cases are not interesting and we don't explore them further.*

Corollary 3.1.1. *If the game allows for cross-type edges, it is easy to see that the network formation would exactly follow theorem 3.1. Because a high type player has more incentive to share with another high type than a low type. Thus when $d^*_H > 0$ in theorem 3.1, the connections will all be to the high type and when $d^*_H = 0$ there won't*

be any connection to the low type either. Since the high type does not share with the low type, a low type player will not share with the high type either resulting in two disconnected components in equilibrium even if cross-type edges were possible.

Corollary 3.1.1 states that no sharing will occur from the high type to the low type. The conditional cooperation argument [80], which is supported in lab experiments [55], provides a mechanism behind this result. A high type player anticipates that a low type cannot sufficiently reciprocate in the future, thus it reduces its cooperation with the low type.

Corollary 3.1.2. *If* $n_L > \frac{d+1}{c-1}$ *and* $n_H < \frac{1}{c-1}$*, then the expected utility of high type in each round is* $E[u^*_{H,r}] = \frac{d+1}{(d+2)n_H}$*. The total share of the high type as a group from the rivalrous resource will be* $U^*_H = \frac{d+1}{d+2}$*.*

If the rivalrous resource was shared equally or there was no network, then we would expect $E[u_{H,r}] = \frac{1}{2n_H}$ and the total share of high type to be $U_H = \frac{1}{2}$. Comparing this equality baseline versus the Nash stable equilibrium outcome from corollary 3.1.2, we conclude that *if there is sufficiently high rivalry among the low type and sufficiently low rivalry among the high type, the intergroup differences will be exacerbated in the network game. The same conclusion would hold even if cross-type edges were possible.*

In summary, the exogenous variation in the level of access to a rivalrous resource leads to different strategies adopted by the low and high types such that information sharing only occurs among the high type. This results from large differences in future prospects of network benefits between the low and high type. The macro implication of the adopted strategies is that the high type as a group will receive a larger share of the common resource than expected simply by the exogenous differences.

4 Experimental Design

We now discuss a randomized experiment we developed using the Empirica platform [6] to test the predictions of our model in a multi-player online game. The goal here is not to exactly replicate the model predictions above as satisfying the assumptions

of theorem 3.1 is very challenging (e.g. it will require a large low type population and even if so not all players will be strategic). Rather, our goal is to experimentally verify that high type players cooperate at a higher rate than low type players and as a result collectively receive a larger share of the common resource than expected simply by their exogenous advantage over the low type players.

In this game, players are recruited and randomly assigned to either high or low type and placed into different positions in a fixed network. The network is homophilous by type. The game has multiple rounds and in each round one player from each type receives valuable information about a rivalrous resource, in this case the location of a gold mine on a map, and decides whether they want to share this information with their neighbors in the network. Because there are less high type than low type players, a high type player receives the information about the location of the gold mine more often than a low type player. Players try to maximize their reward by finding and collecting the gold over all rounds as it translates to their final compensation in dollars. The gold mine is a rivalrous resource, as sharing it with others reduces one's reward in the current round, but sharing might still be a good idea for potential reciprocated benefits in the future.

4.1 Status Structure and Randomized Resource Allocation

Each game has 9 players, 3 of which are randomly selected to be of high status (type) and the remaining 6 will become the low status (type). In each round, the game reveals the location of the gold mine to one randomly selected player from each type. The game instruction ensures players are aware of the status structure and states that the high type players receive the location of the gold on average in twice as many rounds as the low type players. The instructions is purposefully vague on the exact process and it could be interpreted as independent gold assignment in each round, but to ensure a level of fairness so that players within each group potentially receive equal payoffs, players within each group receive the location of the gold in equal number of rounds and the game randomly shuffles the order they receive it.

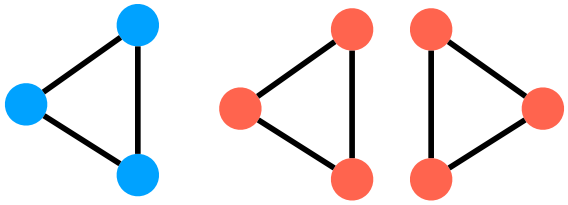


Figure 5.2: The structure of the network. Red (blue) nodes correspond to low (high) status players. Each player has two neighbors with whom they can share information about the location of the gold mine and is randomly assigned to a node in the network.

4.2 Reward Structure

The game needs to repeat over many rounds for player strategies to resemble an equilibrium state. However, we are limited by the time each game can take and use 12 rounds since it also ensures high and low type players each receive the gold 4 and 2 times respectively. A gold mine in each round has \$2.4 total value which will be distributed equally among all players digging it. For example, if none of the players to whom the gold is revealed originally share the information with their contacts, each will receive \$1.2 in that round. If each of them shares it with one neighbor, then each of the four players digging will receive \$0.60.

4.3 Network Structure

In contrast to our model which treats the network formation as an endogenous process, the experiment simply uses a fixed network structure which corresponds to the model prediction when the maximum degree is $d = 2$. The network will effectively have three disconnected triangles, one with the high status and two with the low status players. Figure 5.2 illustrates the network structure. The choice of two disconnected triangles among the low status players rather than a single connected hexagon is made intentionally to first avoid leakage or interference between pairs of users not directly connected and second to make comparison with the high status network and inference using resampling easier. Players upon arrival to the experiment platform will be randomly assigned to a node in the network, which will also determine their status.

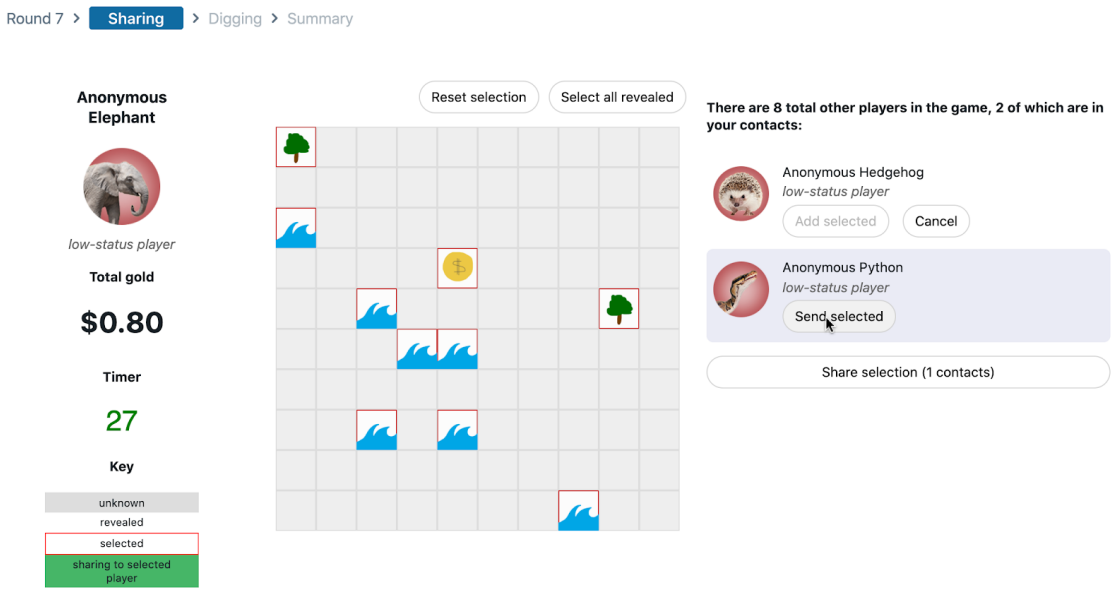


Figure 5.3: The snapshot of the first (sharing) stage in round 7. The player profile is shown on the left and the neighbors list is on the right. In this case, the player and their neighbors are all from the low status group with a red background. High status profiles have a blue background. In this stage, the player has received the location of the gold and is sharing it with the python.

4.4 Game Setup

The game has 12 rounds, however the instructions on the game does not specify the number of rounds, as such the players do know when the game will end. Each round has the following 3 stages.

1. **Sharing Stage:** In the first stage, the experiment platform reveals 10 random squares of the map to each player. If a player is assigned to receive the location of the gold, it will be revealed among these 10 squares. Each player then decides which squares to share or not share with which of the two neighbors. This decision is probably informed by the interactions with the neighbor in the previous rounds. Figure 5.3 shows a screen snapshot of the sharing stage.

2. **Digging Stage:** In the second stage, the experiment platform reveals the squares that were shared by the neighbors. If the gold mine was originally revealed to the player or one of their neighbors shared its location with them, the player can dig the location and is guaranteed to receive some reward. Oth-

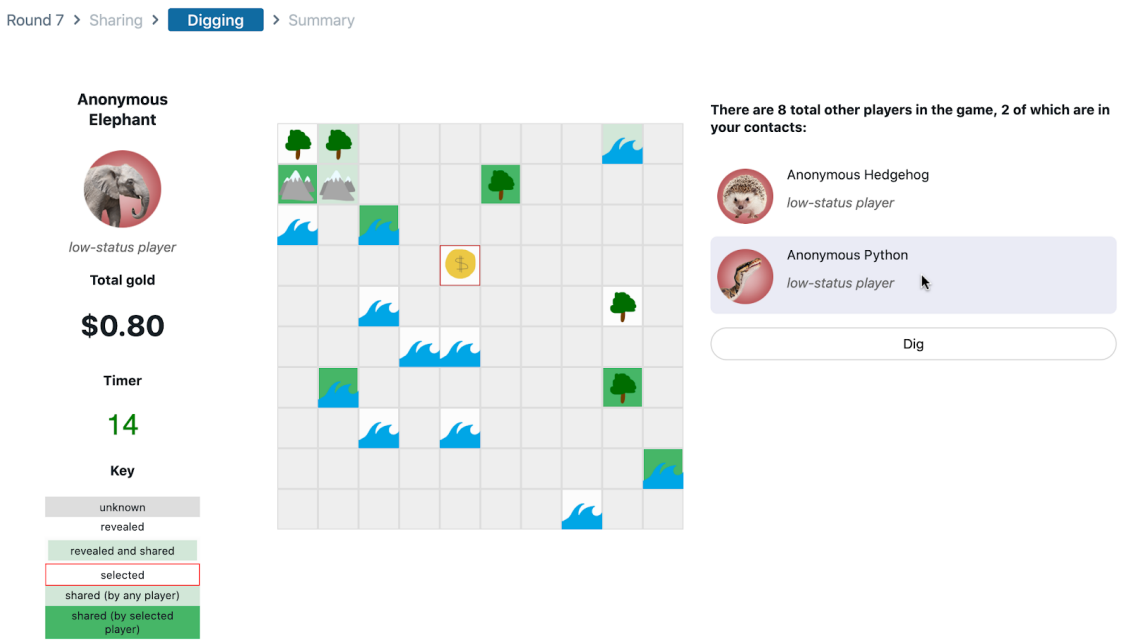


Figure 5.4: The snapshot of the second (digging) stage in round 7. This stage immediately follows the snapshot shown in Figure 5.3. The squares that were shared by one of the neighbors (python) are highlighted in green by hovering over the neighbor. The player originally received the location of the mine and has selected its square to dig.

erwise, the player can choose another square as a best guess to dig. Figure 5.4, shows a screen snapshot of the digging stage.

3. **Summary Stage:** In the third stage, the full map is revealed and if the player successfully dug at a gold mine, they will receive information about their reward, which depends on how many other players were also digging. This stage also summarizes the sharing decision of all neighbors. In particular, it shows the player which squares (potentially including the gold mine) the neighbors decided to share and which ones they decided to hide. Figure 5.5 shows a screen snapshot of the summary stage.

5 Data

We collected data for 38 games that successfully finished with all players present. Games were advertised on MTurk in batches of maximum 3 games so that no more

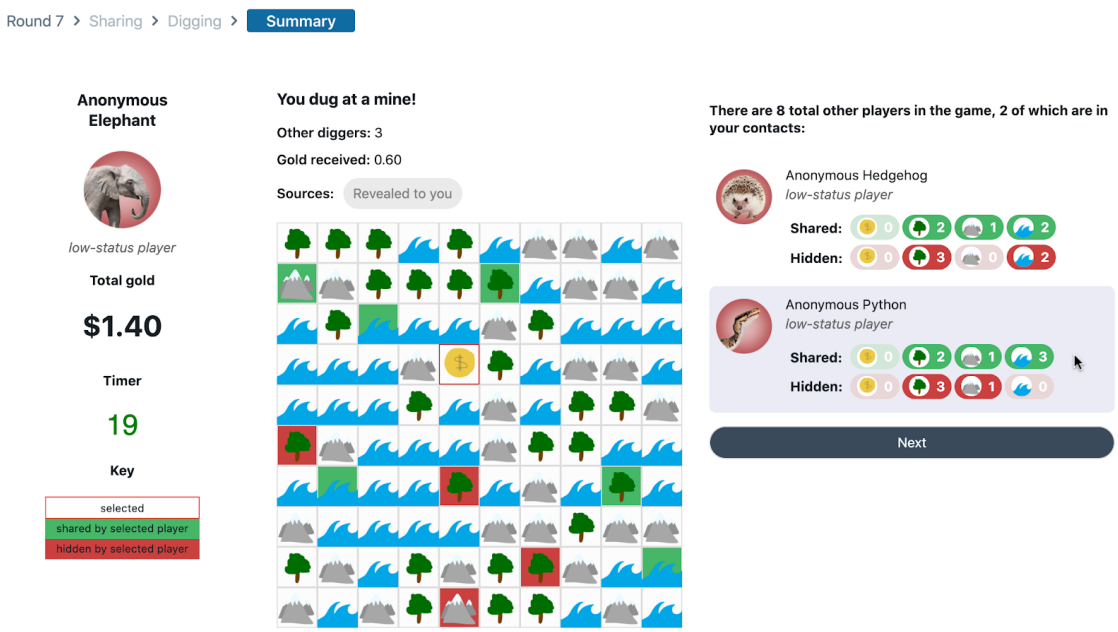


Figure 5.5: The snapshot of the third (summary) stage in round 7. This stage immediately follows the snapshot shown in figure 5.4. By hovering over each neighbor, the player can see their sharing decision in this round. The squares that python decided to share are highlighted in green and the squares they decided to hide are highlighted in red. Since the player dug at a mine and there were 3 other players digging too, the player receives \$0.6 (\$2.4/4).

than 30 players were connected to the platform at the same time. MTurk workers who signed up for a batch received an email 15 minutes before the game started and those who joined the platform were randomly assigned to a position in the network. Each game took about 15 minutes and MTurk workers were not allowed to play more than once. Data collection took a period of 2 weeks from 2021-03-22 to 2021-04-05.

Out of the 38 games, there were 10 games in which players of a single type missed digging the gold more than once even if they knew its location. This can happen either due to connection problems or player inattentiveness. As we also mentioned in the pre-registration document, analyzing such games and comparing them against a null model is challenging, because not only group level rewards will be lower due to the missed opportunities but also inattentiveness might affect cooperation. As outlined in the exclusion criteria of our pre-registration document, the final data excluded these games and had 28 games with 252 unique players. Comparing the treatment groups (high or low) along three basic demographic variables does not

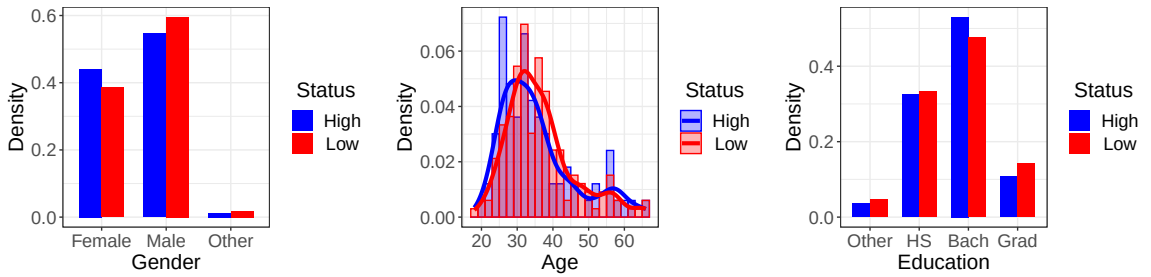


Figure 5.6: Distribution of player gender, age and education by treatment (status) condition. In the education plot, HS, Bach, Grad refer to High school, Bachelor's degree and post-graduate degree respectively.

reveal a significant difference. The p-value of the two-sample chi-square test on gender and education level between the high and low treatment groups are 0.69 and 0.79 respectively. Similarly, the p-value of the two-sample Kolmogorov-Smirnov test on age is 0.35. Figure 5.6 compares the distribution of these variables among participants across the status treatment.

6 Methods

Our main hypothesis is that high status players share the location of gold more frequently than low status players. This leads to high status players as a group receiving a larger share of total available gold than would be expected without network sharing, which would have been about 50% given that exactly one high-status and one low-status player receive the information in each round. Similarly, the mean reward or mean fraction of total rewards that goes to a high type player is larger than the value predicted without network effects. These hypothesis involve quantities at the individual and group levels. Hence, we compare the experimental data against a null model in two ways. In the first analysis, described in section 6.2, the dependent variable is the within-status dyadic sharing rate and the null model indicates no difference in sharing rate by status treatment. In the second analysis in section 6.3, the dependent variable is the fraction of rewards to each status and the null model predicts equal distribution of rewards to the status groups.

6.1 Notation

In what follows, we let $I_{gold}(g, r, i)$ represent an indicator variable which takes the value of 1 when the gold is revealed to player i in round r of game g. Similarly, $I_{shared}(g, r, i, j)$ is a binary indicator that takes the value of 1 when player i shares the gold with player j in round r of game g. $U_{g,i}$ is the utility or total reward of player i at the end of game g. G corresponds to the set of all games, H_g is a set that contains the 6 directional edges in the form of (i, j) between high status players game g and L_g contains the set of 12 directional edges between low status players. Given the notations above, we can express the average sharing rate from player i to player j in game g as followed.

$$S_{g,i,j} = \frac{\sum_{r\in\{1,\ldots,12\}} I_{shared}(g, r, i, j)}{\sum_{r\in\{1,\ldots,12\}} I_{gold}(g, r, i)} \tag{5.2}$$

Similarly, we can define the average sharing rate within each status group as followed.

$$S_{g,H} = \frac{\sum_{(i,j)\in H_g} S_{g,i,j}}{|H_g|} \tag{5.3}$$

$$S_{g,L} = \frac{\sum_{(i,j)\in L_g} S_{g,i,j}}{|L_g|} \tag{5.4}$$

6.2 Dyadic Sharing Rate

Our main hypothesis examines the difference in sharing rate, or *P(sharing | gold is revealed)*, at the level of each dyad across status groups. In this analysis, a unit of observation is the sharing rate on a single directed edge over all 12 rounds or $S_{g,i,j}$. Since sharing is directional, there will be two observations for each dyad corresponding to each direction. We compare the sharing-rate of high-status and low-status players in two ways.

Fisher Exact Test: The sharp null here implies that status has no effect at all on sharing decisions of a player. Since a unit of observation involves each directed edge, we can use the difference in the mean sharing rate of high status group and low status

group as the test statistic.

$$t = \frac{\sum_{g \in G} S_{g,H}}{|G|} - \frac{\sum_{g \in G} S_{g,L}}{|G|} \tag{5.5}$$

The test statistic is effectively the estimated average treatment effect on the sharing rate along a dyad where the treatment is the assignment of the dyad to high or low status. Given the sharp null, we can conduct the usual Fisher randomization technique to compute the exact p-value of our observed statistic. However, it is important to note that not all randomizations are valid. A valid randomization should generate three disconnected components with one as a high status clique similar to figure 5.2. But more importantly, the randomization must maintain the same neighbors for each player because the sharing rate of each player is dependent on sharing decisions of their neighbors. If we had allowed randomizations that create different pairings of players than the actual realized network, each player would be exposed to a different neighbor history which could have changed their sharing rate. In other words, the sharp null does not imply the sharing rate is independent of neighbor actions, rather it only assumes independence from status labeling. Hence, we are comparing against a *conditional sharp null: conditioned on the realized assignment of players to network positions, the status has no effect on sharing rate.*

There are only 3 randomizations per game that keep positions and neighbors in the network fixed but flip the status. In each randomization, the status of one of the three triads in figure 5.2 is set to be high and the remaining two triads are low status. Given 28 collected games, there are 3^{28} possible permutations, so we use sampling from these permutations to generate the distribution of the statistic under the sharp null.

Average Treatment Effect: We could also test the effect of status against the Neyman null of zero average treatment effect or ATE. The challenge is that the Stable Unit Treatment Value Assumption or SUTVA is violated: the outcome of an edge not

only depends on the status assignment of the players on each side of the edge but also on the assignment of their neighbors in the triad. This implies that there is potential spillover from one dyad to another. However, we are not interested in the effect of individual status assignments, rather the status assignment in groups. We can denote the potential outcome of a dyad as $S(t_1, t_2, t_3)$ where t_1, t_2 and t_3 correspond to the status or type assignment of the three players in the triad that contains the dyad and $S(t_1, t_2, t_3)$ is the sharing rate from player with status t_1 to player with status t_2. We are not interested in causal quantities such as $E[S(H, L, L) - S(L, L, L)]$, instead we are after a causal quantity such as $E[S(H, H, H) - S(L, L, L)]$. This is because our theory is about how groups of high status cooperate differently than low status and not about the effect of individual status changes.

Using $E[S(H, H, H) - S(L, L, L)]$ as the estimand addresses the SUTVA violation within each triad as the treatment now explicitly accounts for the full triad assignment. Nevertheless, there is still the possibility of spillovers from disconnected triads since there is information flow between triads when sharing the common resource. Therefore, we expand the potential outcome function on a dyad to $S(t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9)$ where the first 3 arguments correspond to the status assignment in the triad that contains the dyad, the remaining 6 arguments correspond to assignment of players in other triads and $S(.)$ is the sharing rate from player with status t_1 to player with status t_2. With this definition, the main estimand incorporates the status assignment of players in other triads as shown below.

$$ATE = E[S(H, H, H, L, L, L, L, L, L) - S(L, L, L, L, L, L, H, H, H)] \tag{5.6}$$

One could use a difference-in-means estimator similar to the one shown in equation 5.2 for the ATE above. If we were to assume sharing decisions of players are independent of each other, we could conduct inference using a two independent sample t-test with unequal variances. However, one might expect that the potential outcomes in triads might be correlated. For example, it is possible that players engage in tit-for-tat or grim trigger strategies, in which case their sharing rates might be correlated.

In the terminology of linear regression, we would say that the error terms are correlated in triads. Therefore, we need to account for this clustering in our inference. We could account for this clustering by defining each triad in each game as a cluster (75 total clusters) and use cluster-robust standard errors. But since there is information flow from one cluster to another during the game summary stages, one needs to be conservative and use the games or the coarsest level possible as the clusters. The only concern with this choice is the small number of clusters (28) which might adversely affect our standard error estimate. However simulations with 28 games that use a probabilistic grim trigger strategy among players of each triad suggest that the inference with cluster robust standard errors and the game as the clustering unit has a correct type I error (type I error = 0.03 when α=0.05) whereas the regular standard error without clustering greatly over-rejects when the null is true (type I error = 0.12 when α=0.05).

In summary to conduct inference on the ATE in equation 5.6, we use the following regression model with cluster robust standard errors and each game as a cluster.

$$S_{g,i,j} = \beta_0 + \beta_1 t_{g,i,j} + \gamma \mathbf{X_g} + \epsilon_{g,i,j} \tag{5.7}$$

where $S_{g,i,j}$ as defined in equation 5.2 is the sharing rate in the dyad from player i to player j in game g, $t_{g,i,j} \in \{H, L\}$ is the randomized status treatment on the tiad that includes i and j and $\mathbf{X_g}$'s are the game fixed effects.

6.3 Fraction of Group Rewards

Any difference in sharing rates will directly lead to unequal shares of total rewards collected by the status groups. We conduct tests to evaluate whether the high status group receives a larger fraction of the total gold than would be expected under the null model. This analysis alleviates any concerns of dependence within games when using sharing rates as the dependent variable since the unit of analysis is a game which is clearly independent of other units.

Null Model: We refer to the model of each player acting individually without any network effects as the null model. Under the null model, utility solely derives from the exogenous individual ability, captured by status in our experiment, and does not have a network component. Without network effects, each group will receive about half of the total gold available, but not exactly 50% since players can still guess the location of the gold if it is not revealed to them. In particular, the low type will receive slightly more than 50% since there are 5 players guessing the location in each round as opposed to 2 in the case of high type. If we denote a binomial process by $Binom(n, p)$ where n corresponds to the number of trials and p is the success probability, then the expected fraction of total gold earned by each group and their ratio under the null model, denoted by μ_H, μ_L and ρ, take the binomial forms below.

$$\mu_H = E\big[\frac{1 + Binom(2, 1/90)}{2 + Binom(2, 1/90) + Binom(5, 1/90)}\big] = 0.498 \tag{5.8}$$

$$\mu_L = E\big[\frac{1 + Binom(5, 1/90)}{2 + Binom(2, 1/90) + Binom(5, 1/90)}\big] = 0.514 \tag{5.9}$$

$$\rho = E\big[\frac{1 + Binom(2, 1/90)}{1 + Binom(5, 1/90)}\big] = 0.994 \tag{5.10}$$

where $Binom(2, 1/90)$ corresponds to a binomial process in which 2 high type players without the gold guess its location among the 90 unrevealed squares and similarly $Binom(5, 1/90)$ corresponds to the same process for the 5 low type players without the gold.

Non-Parametric Test: This is be our primary analysis at the game-level. The analysis involves the following two measures.

1. Mean fraction of total reward collected by the high status group.

$$\widehat{\mu}_H = \sum_{g \in G} \big[\frac{\sum_{i \in H_g} U_{g,i}}{\sum_{i \in H_g \cup L_g} U_{g,i}}\big] / |G| \tag{5.11}$$

2. Mean ratio of total reward collected by the high status group over the low status

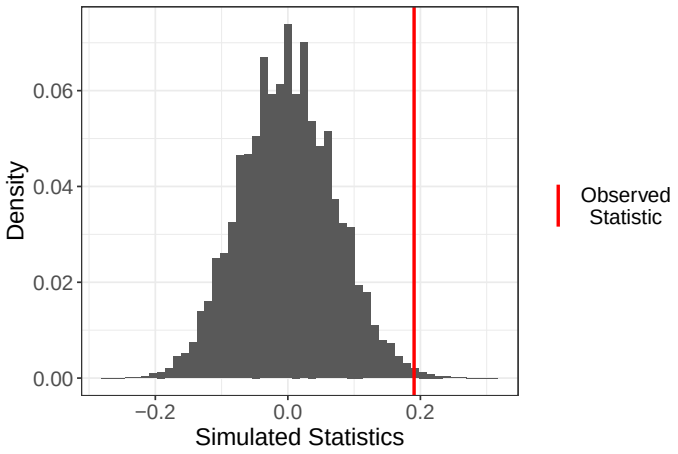


Figure 5.7: The distribution of the test statistic with the randomization inference versus the observed statistic.

group.

$$\widehat{\rho} = \sum_{g \in G} [\frac{\sum_{i \in H_g} U_{g,i}}{\sum_{i \in L_g} U_{g,i}}] / |G| \tag{5.12}$$

We compare the above measures against their corresponding values from the null model in equations 5.8 and 5.10 using the one-sample Wilcoxon signed rank test. With 28 games, we have $\binom{56}{28}$ possible permutations so we need to appeal to its normal approximation to compute the p-value. The null hypothesis in Wilcoxon signed rank test assumes symmetry around the median of paired differences. Since this might not be appropriate, we also report the results from the weaker sign test whose null hypothesis simply assumes the median is a given value.

Parametric Test: We compare the above measures, $\widehat{\mu}_H$ and $\widehat{\rho}$, against the null model predictions using the one-sample t-test. This will be a secondary game-level analysis since we don't expect that the distribution of $\widehat{\mu}_H$ and $\widehat{\rho}$ would be close to their asymptotic normal under the null given only 28 games. In fact, our simulations suggest that this test has a higher type I error rate than the significance level with $n = 28$ (e.g. type I error=0.018 when α=0.01).

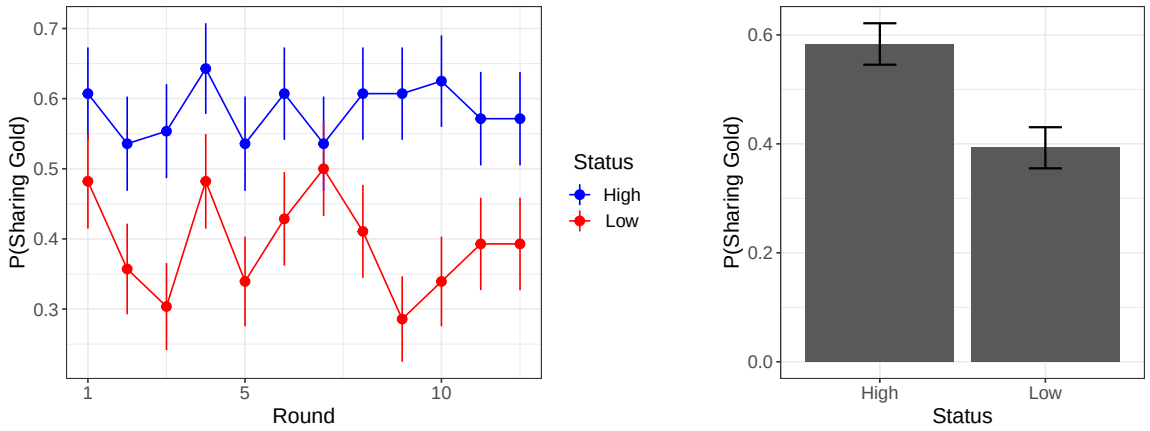


Figure 5.8: Probability of sharing conditioned on receiving the gold per round (left) and over all rounds (right) by status of players. Bars in the left plot correspond to standard error while they correspond to 95% confidence interval on the right

7 Results

We present the results of different analysis methods in the same order as described in section 6.

7.1 Dyadic Sharing Rate

Fisher Exact Test: Figure 5.7 shows the result of this analysis. The test statistic, in equation 5.5, is effectively the average treatment effect at the dyad level. The statistic is positive and significant (two-tailed $p = 0.0067$ with 50000 simulated random assignments) indicating that the high status treatment has a higher sharing rate than the low status treatment.

Average Treatment Effect: Figure 5.8 compares the mean sharing rate along high status and low status dyads as defined in equations 5.3 and 5.4. The results clearly indicate that the high status players share the rivalrous resource with each other at a significantly higher rate in all rounds and overall. This suggests that the rivalry in resource sharing promotes strategic behavior, especially among the low status players. This can be further validated by examining the number of non-gold squares shared by low status players. The game revealed 10 squares to each player in each round,

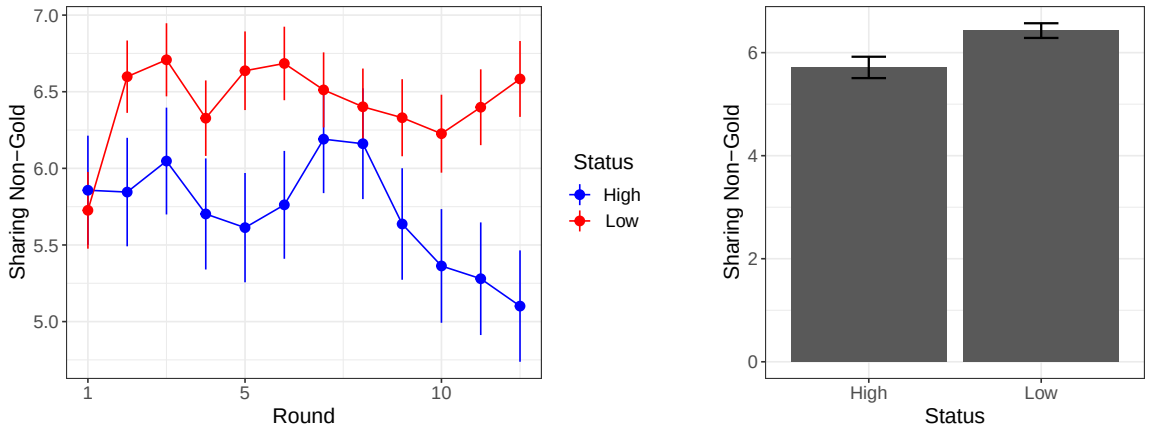


Figure 5.9: Mean number of non-gold squared shared per round (left) and over all rounds (right) by status of players. Bars in the left plot correspond to standard error while they correspond to 95% confidence interval on the right

one of which could be a gold mine, and the players could share any number of these squares with any of their neighbors. Sharing non-gold squares might still be a form of cooperation since it helps the other players to find the gold mine through the process of elimination. Figure 5.9 compares the mean number of non-gold squares shared along high status and low status dyads. As opposed to the results for sharing the gold mine itself, we observe that low status players share more squares on average than high status players. This finding suggests that low status players act very strategically as they tend to keep the rivalrous resource exclusively, but nevertheless share other valuable information with their neighbors hoping to keep a cooperative relationship in the future.

Table 5.1 shows the formal inference results on the average treatment effect. The model explained in section 6.2 with cluster robust standard errors at the game level is included in the second column. According to this model, random assignment to a high status triad causes the sharing rate to increase by about 19%. The cluster robust p-values from models with and without game fixed effects (columns 2 and 3) are $p = 0.011$ and $p = 0.009$ respectively.

Table 5.1: Estimated Average Treatment Effect under different models. First column includes the game fixed effects but uses regular standard errors. Second column includes game fixed effects along with cluster robust standard errors. Third column does not include the fixed effects but uses cluster robust standard errors. Fixed effect estimates are not shown.

	Dependent variable:		
	Sharing Rate		
	(1)	(2)	(3)
High Status	0.190***	0.190**	0.190***
	(0.038)	(0.074)	(0.072)
Constant	0.187**	0.187***	0.393***
	(0.095)	(0.025)	(0.040)
Cluster Robust SE	No	Yes	Yes
Game Fixed Effects	Yes	Yes	No
Observations	504	504	504
R^2	0.216	0.216	0.042
Adjusted R^2	0.170	0.170	0.040
Residual Std. Error	0.400 (df = 475)	0.400 (df = 475)	0.430 (df = 502)
Note:			*p<0.1; **p<0.05; ***p<0.01

7.2 Fraction of Group Rewards

In this section, we present the results of hypothesis tests that compare the observed fraction of rewards earned at the end of the game by the high status group ($\widehat{\mu}_H$ from equation 5.11) and the ratio of high and low status rewards ($\widehat{\rho}$ from equation 5.12) versus their respective null model predictions in equations 5.8 and 5.10.

Non-Parametric Test: Given the small sample size ($n = 28$), non-parametric tests that don't make any assumption on the distribution of the test statistic seem to be more appropriate. The null model in the following non-parametric tests assumes that the median of the distribution, from which we observe $\widehat{\mu}_H$ values, is equal to μ_H. The one-sample Wilcoxon signed rank test rejects the null model with $p = 0.021$ and $(0.507, 0.580)$ as the 95% confidence interval for the fraction of rewards collected by

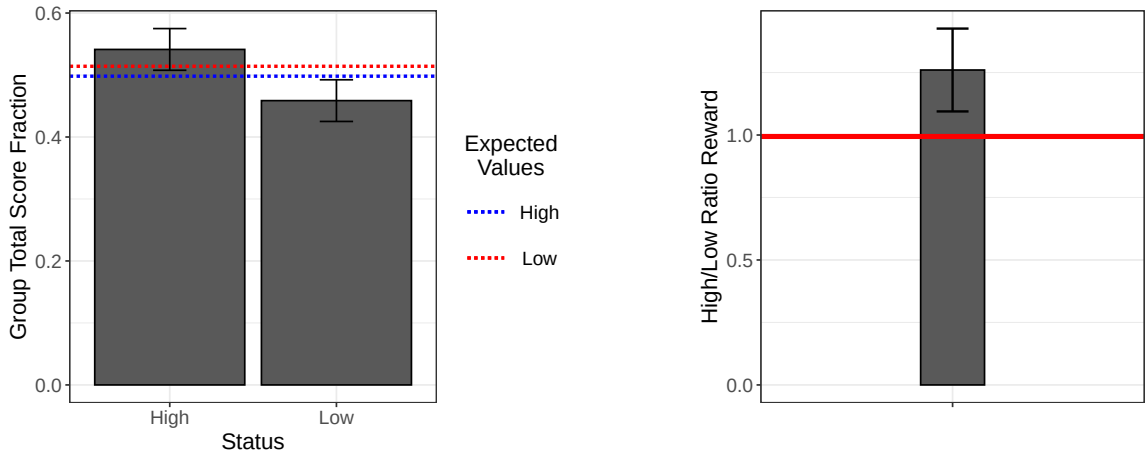


Figure 5.10: Share of each group from total gold distributed over the game versus their expected share (left). The ratio of high status group versus low status group total gold distributed over the game compared with the expected ratio (right). Bars correspond to 95% confidence interval.

the high group. The weaker sign test also rejects the null model with $p = 0.013$.

We could conduct the same tests using a different statistic and compare the observed ratio of rewards $\widehat{\rho}$ against its null model prediction ρ. The one-sample Wilcoxon signed rank test rejects this null model with $p = 0.010$ and $(1.064, 1.426)$ as the 95% confidence interval on the true ratio of rewards ρ. The sign test also rejects the null with $p = 0.012$. The direction of the observed statistic relative to the null in all the tests above indicate that the high status group collects a larger share of the rivalrous resource than expected under the null model without network effects.

Parametric Test: Figure 5.10 compares the mean fraction of total rewards collected by each group and their ratio against the null model predictions in equations 5.8, 5.9 and 5.10. The results indicate that the high status assignment has a positive and statistically significant impact on the share of rewards collected by the group. The one sample t-test on the fraction of rewards by the high group rejects the null with $p = 0.016$ and $(0.507, 0.576)$ as the 95% confidence interval. Similarly, the t-test rejects the null on the ratio of rewards with $p = 0.003$ and $(1.090, 1.430)$ as the 95% confidence interval. However, as we mentioned in section 6.3, our simulations suggest that the type I error rate of the parametric t-test with $n = 28$ is slightly higher than

the α, so the results above might not fully satisfy the asymptotic assumptions. Nevertheless, our non-parametric tests provide clear evidence rejecting the null, offering more credibility to the findings from these parametric tests.

8 Conclusion

The persistence of inequality has been linked to social networks [108]. The most common account of network effects on inequality takes a purely structural perspective since it considers homophily and network segregation as the drivers of unequal access to opportunities. While this is largely true, a purely structural view misses the nuanced processes that occur in networks [162]. With a simple example of diffusion in networks, we showed that network structure and homophily does not explain how one group can take a "larger share of the pie" than expected solely based on heterogeneity in individual ability. In this paper, we go beyond the simple structural perspective and examine one potential process that affects the unequal distribution of resources. In particular, we assume that agents in a network share information about a common rivalrous resource, and are heterogeneous in terms of their individual ability in accessing the resource. As opposed to the structural perspective which treats agents as myopic, we assume they are strategic and forward-looking. This makes sense since the rivalry in a valuable resource necessitate competition and strategic cooperation. Our experimental results further validate that individuals engage in strategic behavior.

We develop this process into a repeated game of information sharing in networks where network formation is endogenous to the model. There are two types of agents, one with a higher probability of accessing the resource in each round than the other, and each agent decides whether to share information about the rivalrous resource with any other agent. If the differences between agents type are sufficiently large, the model predicts that information sharing or cooperation in the pairwise Nash stable network exists only among the high type. As a result, the high type as a group will receive a larger share of the rivalrous resource than expected solely based on the exogenous probabilities which can be thought of individual ability without network

sharing. Furthermore, a randomized multi-player lab experiment that closely mimics the game validates the model predictions. We observe that players who have a higher probability of accessing a valuable resource, which directly translates to monetary reward, are about 19% more likely to share it with their other high status neighbors. Both the theoretical and experimental results indicate the importance of network processes other than simple diffusion in generation of inequality. One can think of the status differences among players in terms of accessing the valuable resource as differences in human capital and the number of neighbors with reciprocal cooperation as social capital. Thus, our results suggest that the interaction of human capital and social capital play an important role in unequal access to opportunities. It also invites further experimental and modeling studies to fully characterize how differences in human capital lead to inequality in social capital [63, 88].

We hope our study of network effects on group-level outcomes contributes to the larger discussion around why a small minority can take an exceptionally large share of common resources. The model predictions demonstrate how the differences among agents and the relative scarcity of the resource in each group lead to cooperation among the high status but competition among the low status to benefit from the limited stock of valuable resource. The extreme scarcity in the low status group prohibits the formation of social capital and promotes a form of elite capture. The model also implies that the low status type is stuck in a durable poverty trap because inequality is the equilibrium of the incentive structure in the network, an argument that is inline with views of inequality as a process [162]. The formation of inequality in this process is an example on how micromotives (e.g. cooperation) lead to macrobehavior (e.g. larger share of the pie by one exclusive group) [155].

Many recent studies have looked at urban segregation, status homophily and lack of mixing in networks and have reported a link, albeit correlational, between inequality and the extent of clustering and homophily [163, 165]. These findings have been the basis of a policy recommendation to promote cross-group linking in social structures as a means of combating inequality. While interventions that bring diverse people together and facilitate inter-group linking could be helpful, without changing

the underlying incentive structure of inter-group linking, their impact would not be as large as one hopes. Our model suggests that, at least when it comes to sharing limited common resources, inter-group links are difficult to create and persist and even if they were to form due to policies that encourage mixing, valuable information would rarely be shared across them. This happens due to the incentive structure of cooperation between low and high status groups in social networks which reduces the extent of possible reciprocity and consequently the motivation for high status individuals to share valuable information with the disadvantaged group. A potential remedy to this problem is to limit the visibility of individual endowments, as recent studies have shown that visibility of advantage increases inequality [151] and negatively impacts cooperation [137]. In the context of our model, if it is not easy to observe the agent type, we would expect more inter-group linking and cooperation. However, the efficacy of this approach over long-term might not be stable as agent types can eventually be inferred.

The lack of cooperation in the low status group originates from the rare ability of accessing valuable resources by the group. Thus, it might be helpful to target individuals in the low status group by investing in their human capital. While this will improve the chances of accessing resources and opportunities by those individuals, it does not make them more likely to cooperate with the rest of the low status group because this interventions does not change the underlying incentive structure among the disadvantaged population. The barrier to cooperation in the low status group is a group phenomena and thus requires an intervention at the group level rather than individuals. The ideal intervention is one that empowers the whole group by investing in their human capital just enough to tip the balance in favor of cooperation rather than competition. This intervention not only helps the individuals to access valuable resources independently (e.g. find jobs and be productive), but also increases the incentives for others in both high and low status group to share valuable information with them. A broad investment in the disadvantaged population just enough to overcome the barriers to cooperation has multiplicative effect on their outcome as it also enables the formation of social capital, improves the outcome of the whole

group and reduces inter-group differences even beyond what's expected by individual differences. When information sharing becomes the dominant strategy of the low status group, network effects in fact alleviate inequality in a manner similar to what we described at the beginning of the paper and illustrated in figure 5.1b.

Chapter 6

Conclusion

The findings presented in this dissertation contribute to the growing literature around the network origins of economic outcomes, inequality and its persistence in multiple and original ways.

For the first time, we used Facebook communication data and economic indicators in US counties to establish an empirical link between the residents' economic well-being and the counties network structure. We found that US counties rich in long ties (i.e., those bridging different communities, representing structural diversity) report better economic outcomes along various indicators such as income, unemployment and social mobility. This finding is novel in at least two ways. While the relation between long ties and economic outcomes has been theorized in the past, it has been empirically studied solely within the corporate context. The link between structural diversity and general economic outcomes outside of a firm hasn't been demonstrated empirically using large-scale data from the wider US population. In addition, we found that long ties are more frequent in individuals of older age who are relatively educated, but more importantly those who have experienced major disruptive events such as mobility and migration in the course of their lives. Critically, our findings suggest that long ties are not created simply as a consequence of these life events (e.g., people migrate, hence they have more long ties), but because of the specific behavioral and social skills that people develop when they go through such major changes. For example, individuals learn how to avoid social categorization, connect

with different communities and adapt to changes during the migration experience, and going forward, will retain such skills that are crucial in developing and maintaining structurally diverse networks which in turn lead to better economic outcomes. We explored three different case studies all involving disruptive events and observed that individuals with those experiences tend to have more long ties many years after the event than those without them. In our analysis, we attempted to control for potential confounders and designed the analysis in ways that minimize the risk of unobserved endogeneities, nevertheless the exploration of the causal link between these major life changes, behavioral skills and better economic outcome will be the subject of our future investigations.

Second, we studied the determinants and the impacts of unequal diffusion in networks. We provided observational evidence that individuals from low status groups receive lower marginal benefit from networking compared to individuals from high status groups. We attributed this phenomenon to network homophily which is the tendency for high status individuals with high levels of novel information to link with each other and not to low status individuals. Homophily leads to heterogeneity in potential social resources available in two ego-networks with the same structure but from two distinct (high vs low) social status groups. At its core, the differential in marginal benefits arises from homophily which causes the unequal diffusion of information and resources in social networks: valuable resources originate from high status group and remain exclusively there. We provided causal evidence for unequal diffusion in the context of a randomized seeding experiment where a new piece of information diffuses in the network. In the experiment, the diffusion exhibited a highly unequal pattern as the information was more likely to reach the social group to which the randomly assigned seeds belonged to. These findings might have important implications for economic policies of social mobility. For example, it suggests that if a new economic opportunity comes along, it will have higher chance of making it to those who already have a certain economic advantage, rather than to those who are economically marginalized. We found this result particularly powerful since unequal diffusion has serious implications in terms of unequal access to opportunities.

Third, we examined how network structure leads to unequal diffusion in more detail. We showed that common network models fail to capture unequal diffusion and almost always underestimate it because these models ignore brokerage patterns in networks. Gate keeping and brokerage are very common in social structures and occur when the links from one community to another have to go through a small number of individual brokers. Thus we developed a random network model that accounts for brokerage and by doing so explains the network structure of unequal diffusion and significantly improves the predictions on the extent of diffusion to various groups in the network over the existing models. This model showed that any departure from the uniform distribution of links to information sources – among members of a group – limits the diffusion of information to the group as a whole. The unequal distribution of cross-group links presents effects of both first and second order: not only some individuals will have fewer direct links to information sources than others, but also the whole group will have fewer diffusion paths to the information sources. From a structural point of view, distributing the few existing cross group links equally among all individuals in a group will lead to both higher levels of access to economic information and more equal distribution of such opportunities than the common pattern of centralization and accumulation in the hands of few individuals. Most importantly, the argument implies that while brokerage-like patterns are beneficial for the individual brokers in a group, they at the same time actively contribute to unequal opportunities and worse outcomes for the whole group. This work also offered some methodological advances on the evaluation of network models. Models are often fit to the data using a single insight or criteria, for example the prevalence of cross-group links in the Stochastic Block Model, and evaluated solely based on that criteria but the fitted models are often used to examine other network characteristics such as diffusion structure. By re-sampling from the fitted model and comparing the distribution of other network statistics, that are not explicitly accounted for in the model, with the observed network, we showed how one could obtain a better understanding of model validity and generalization.

Finally and after giving a purely structural perspective on networks and inequality,

we focused on how more nuanced processes such as strategic behavior in networks exacerbate existing inequalities. We focused on information sharing in networks and discussed how a mechanism that is closely related to scarcity of resources widens inter-group differences and yields different returns on social capital to different group. In an information sharing game, individuals compete for a rivalrous resource over repeated rounds. The subgame perfect equilibrium predicts lower cooperation among lower status agents, which in turn leads the whole group to receive a small share of the common resource. The main insight behind the equilibrium is that more intense competition over limited resources in the low status group leads to lower levels of cooperation and information sharing than the high status group. We also validated this prediction in an online lab experiment by recruiting Amazon Mechanical Turk workers to play an online collaborative game in which they had to find and dig gold mines and in the process could pass information to their neighbors according to their local network structure. We found that the players which had a lower chance of receiving the gold in each round were less likely to share it with their contacts when they received the gold. As a result, these low status players took "a smaller share of the pie" than what was expected by the individual differences in "ability". The results indicate that individuals in low status groups who have limited access to rivalrous opportunities do not have enough incentives to share such resources with other low status individuals as they would not get relevant resources in exchange. On the contrary, individuals in high status groups help each other out by sharing relevant information, as they know they can expect a return on their investment. As a result, low status individuals who have limited access to valuable resources persist in being low status, while the high status individuals who already have better access to resources gain even more advantage, widening any initial gaps over time.

While being conscious in transferring results from simple models, controlled experiments and empirical analysis to real world scenarios, we conclude by highlighting that – taken at once – our work suggests that:

- In the US, neighborhoods with structurally diverse connections to other geo-locations are economically more prosperous and better positioned to obtain

valuable information. Analysis of communication ties among 120 million individuals over 6 months suggest a close link between network structure and an array of important economic indicators such as social mobility and income. These findings suggest a potential network pathway through which inequalities could become reinforced since economic opportunities seems to be geographically concentrated;

- Individuals who are able to consciously establish and maintain structurally diverse networks with many long ties and at the same time turning these connections into intimate, trustworthy links (strong ties) have more social capital and resources which improve their economic outcomes.

- These individuals are likely to have faced constructive challenges – such as migration – during the course of their lives and a valuable set of skills developed out of these experiences. In the words of Nietzsche, "Out of life's school of war — What does not kill me makes me stronger".

- The importance of skills in forming valuable networks suggests an array of policies to first fully characterize the nature of these skills and second equip individuals with these skills as a means of combating inequality;

- When we look at the diffusion of economic opportunities among low status and high status individuals, we observe that social networks by nature of homophily and brokerage have mechanisms that keep the opportunities exclusive to the high status group and limit the social capital of low status individuals;

- The incentive structure for cooperation in networks is yet another barrier in combating inequality. Given that valuable resources are often limited in stock and rare among low status individuals, there will be adverse competition over these resources in disadvantaged groups in ways that inhibit formation of social capital and trust. To counteract inequality, one needs to change the social structure in ways that incentivize sharing opportunities among low status fellows and thus generate social capital.

The main takeaway of this work is that policies aimed at counteracting inequality that target individuals might be a good starting point but not sufficient in themselves to fight inequality efficiently and on a large scale. Because of the network mechanisms we tested and explained in this dissertation, such policies risk to lift up a few selected individuals while at the same time perpetuating inequality among those who are left behind. One way to counteract this network effect might be to make sure that those low status individuals who get access to opportunities receive incentives to share resources with other low status individuals once they become high status. Our models suggest that, indeed, this does not normally occur when incentives are not in place. Potentially, the best policies involve distributing assistance and resources equally and lifting a whole group out of poverty since the individual benefits will be amplified by the existing social capital and the feedback mechanisms present in social networks.

Bibliography

[1] *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition)*. Princeton University Press, 2007. ISBN 9780691138541. URL http://www.jstor.org/stable/j.ctt7sp9c.

[2] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[3] Daron Acemoglu and James A Robinson. *Economic origins of dictatorship and democracy*. Cambridge University Press, 2005.

[4] Daron Acemoglu and James A Robinson. Persistence of power, elites, and institutions. *American Economic Review*, 98(1):267–93, 2008.

[5] Philipp Ager, Leah Platt Boustan, and Katherine Eriksson. The intergenerational effects of a large wealth shock: White southerners after the civil war. Technical report, National Bureau of Economic Research, 2019.

[6] Abdullah Almaatouq, Joshua Becker, James P Houghton, Nicolas Paton, Duncan J Watts, and Mark E Whiting. Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, pages 1–14, 2021.

[7] Facundo Alvaredo, Anthony B Atkinson, Thomas Piketty, and Emmanuel Saez. The top 1 percent in international and historical perspective. *Journal of Economic perspectives*, 27(3):3–20, 2013.

[8] Carolyn J Anderson, Stanley Wasserman, and Katherine Faust. Building stochastic blockmodels. *Social networks*, 14(1-2):137–161, 1992.

[9] Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020. doi: https://doi.org/10.1111/rssb.12377. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12377.

[10] Sinan Aral and Christos Nicolaides. Exercise contagion in a global social network. *Nature communications*, 8(1):1–8, 2017.

[11] Sinan Aral and Marshall Van Alstyne. The diversity-bandwidth trade-off. *American journal of sociology*, 117(1):90–171, 2011.

[12] Sinan Aral and Dylan Walker. Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6):1352–1370, 2014. doi: 10.1287/mnsc.2014.1936.

[13] Alberto Arcagni, Rosanna Grassi, Silvana Stefani, and Anna Torriero. Higher order assortativity in complex networks. *European Journal of Operational Research*, 262(2):708–719, 2017.

[14] Anthony B Atkinson and François Bourguignon. *Handbook of income distribution*, volume 2. Elsevier, 2014.

[15] Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80, August 2018. doi: 10.1257/jep.32.3.259.

[16] Michael Bailey, Drew M Johnston, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Peer effects in product adoption. Technical report, National Bureau of Economic Research, 2019.

[17] Abhijit Banerjee, Emily Breza, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Come play with me: Experimental evidence of information diffusion about rival goods. *Work in Progress*, 2012.

[18] Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. The diffusion of microfinance. *Science*, 341(6144), 2013. ISSN 0036-8075. doi: 10.1126/science.1236498. URL `https://science.sciencemag.org/content/341/6144/1236498`.

[19] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0400087101. URL `https://www.pnas.org/content/101/11/3747`.

[20] Lori Beaman and Andrew Dillon. Diffusion of agricultural information within social networks: Evidence on gender inequalities from mali. *Journal of Development Economics*, 133:147–161, 2018.

[21] Lori Beaman and Jeremy Magruder. Who gets the job referral? evidence from a social networks experiment. *American Economic Review*, 102(7):3574–93, 2012.

[22] Lori A. Beaman. Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S. *The Review of Economic Studies*, 79(1):128–161, 08 2011. ISSN 0034-6527. doi: 10.1093/restud/rdr017. URL `https://doi.org/10.1093/restud/rdr017`.

[23] Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0907096106. URL https://www.pnas.org/content/106/50/21068.

[24] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press, 2015.

[25] Peter M Blau. *Exchange and power in social life*. Routledge, 2017.

[26] Francis Bloch and Matthew O Jackson. Definitions of equilibrium in network formation games. *International Journal of Game Theory*, 34(3):305–318, 2006.

[27] Joshua Evan Blumenstock, Guanghua Chi, and Xu Tan. Migration and the value of social networks. 2019.

[28] Lukas Bolte, Nicole Immorlica, and Matthew O Jackson. The role of referrals in immobility, inequality, and inefficiency in labor markets. *Inequality, and Inefficiency in Labor Markets (January 1, 2020)*, 2020.

[29] Stephen P Borgatti. Structural holes: Unpacking burt's redundancy measures. *Connections*, 20(1):35–38, 1997.

[30] Daniel J. Brass. Men's and women's networks: A study of interaction patterns and influence in an organization. *The Academy of Management Journal*, 28(2):327–343, 1985. ISSN 00014273. URL http://www.jstor.org/stable/256204.

[31] Edward Buckley and Rachel Croson. Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics*, 90 (4-5):935–955, 2006.

[32] R. S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, Cambridge, MA, 1992.

[33] Ronald S Burt. The gender of social capital. *Rationality and society*, 10(1):5–46, 1998.

[34] Ronald S Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.

[35] Ronald S Burt. Network-related personality and the agency question: Multirole evidence from a virtual world. *American Journal of Sociology*, 118(3):543–591, 2012.

[36] Ronald S. Burt and Jennifer Merluzzi. Network oscillation. *Academy of Management Discoveries*, 2(4):368–391, 2016. doi: 10.5465/amd.2015.0108.

[37] Ronald S. Burt and Don Ronchi. Teaching executives to see social capital: Results from a field experiment. *Social Science Research*, 36(3):1156–1183, 2007. ISSN 0049-089X. doi: https://doi.org/10.1016/j.ssresearch.2006.09.005. URL https://www.sciencedirect.com/science/article/pii/S0049089X06000767.

[38] Ronald S. Burt, Joseph E. Jannotta, and James T. Mahoney. Personality correlates of structural holes. *Social Networks*, 20(1):63–87, 1998. ISSN 0378-8733. doi: https://doi.org/10.1016/S0378-8733(97)00005-1. URL https://www.sciencedirect.com/science/article/pii/S0378873397000051.

[39] Ronald S Burt, William Barnett, James Baron, Jon-Athan Bendor, Jack Birner, Matthew Bothner, Frank Dobbin, Chip Heath, Rachel Kranton, Rakesh Khurana, Jeffrey Pfeffer, Joel Podolny, Holly Raider, James Rauch, and Ron Burt. Structural Holes and Good Ideas. *Ajs*, 110(2):349–99, 2004. ISSN 00029602. doi: 10.1086/421787.

[40] Ronald S. Burt, Martin Kilduff, and Stefano Tasselli. Social network analysis: Foundations and frontiers on advantage. *Annual Review of Psychology*, 64(1):527–547, 2013. doi: 10.1146/annurev-psych-113011-143828.

[41] Antoni Calvo-Armengol and Matthew O Jackson. The effects of social networks on employment and inequality. *American economic review*, 94(3):426–454, 2004.

[42] Antoni Calvó-Armengol. Job contact networks. *Journal of Economic Theory*, 115(1):191–206, 2004. ISSN 0022-0531. doi: https://doi.org/10.1016/S0022-0531(03)00250-3. URL https://www.sciencedirect.com/science/article/pii/S0022053103002503.

[43] Antoni Calvó-Armengol and Matthew O. Jackson. The effects of social networks on employment and inequality. *American Economic Review*, 94(3):426–454, June 2004. doi: 10.1257/0002828041464542. URL https://www.aeaweb.org/articles?id=10.1257/0002828041464542.

[44] Antoni Calvó-Armengol, Eleonora Patacchini, and Yves Zenou. Peer Effects and Social Networks in Education. *The Review of Economic Studies*, 76(4):1239–1267, 10 2009. ISSN 0034-6527. doi: 10.1111/j.1467-937X.2009.00550.x. URL https://doi.org/10.1111/j.1467-937X.2009.00550.x.

[45] David Card and Alan B. Krueger. School quality and black-white relative earnings: A direct assessment. *The Quarterly Journal of Economics*, 107(1):151–200, 1992. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/2118326.

[46] Laura L Carstensen. Selectivity theory: Social activity in life-span context. *Annual review of gerontology and geriatrics*, 11(1):195–217, 1991.

[47] Laura L Carstensen. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. *Psychology and aging*, 7(3):331, 1992.

[48] Laura L Carstensen, Derek M Isaacowitz, and Susan T Charles. Taking time seriously: a theory of socioemotional selectivity. *American psychologist*, 54(3): 165, 1999.

[49] Emilio J Castilla, George J Lan, and Ben A Rissing. Social networks and employment: Mechanisms (part 1). *Sociology Compass*, 7(12):999–1012, 2013.

[50] Matias D Cattaneo, Richard K Crump, Max H Farrell, and Yingjie Feng. On binscatter. *arXiv preprint arXiv:1902.09608*, 2019.

[51] Matias D Cattaneo, Richard K Crump, Max H Farrell, and Yingjie Feng. Binscatter regressions. *arXiv preprint arXiv:1902.09615*, 2019.

[52] Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011. ISSN 0036-8075. doi: 10.1126/science.1207055. URL `https://science.sciencemag.org/content/334/6060/1269`.

[53] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[54] Kenneth S Chan, Stuart Mestelman, Robert Moir, and R Andrew Muller. Heterogeneity and the voluntary provision of public goods. *Experimental Economics*, 2(1):5–30, 1999.

[55] Todd L. Cherry, Stephan Kroll, and Jason F. Shogren. The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. *Journal of Economic Behavior & Organization*, 57(3):357–365, 2005. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2003.11.010. URL `https://www.sciencedirect.com/science/article/pii/S0167268104001167`.

[56] Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162, 2018.

[57] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.

[58] Raj Chetty, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. The fading american dream: Trends in absolute income mobility since 1940. *Science*, 356(6336):398–406, 2017.

[59] Raj Chetty, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. The opportunity atlas: Mapping the childhood roots of social mobility. Working Paper 25147, National Bureau of Economic Research, October 2018. URL `http://www.nber.org/papers/w25147`.

[60] Guanghua Chi, Joshua E Blumenstock, Lada Adamic, et al. Who ties the world together? evidence from a large online social network. In *International Conference on Complex Networks and Their Applications*, pages 451–465. Springer, 2019.

[61] Chong, Shi Kai, Bahrami, Mohsen, Chen, Hao, Balcisoy, Selim, Bozkaya, Burcin, and Pentland, Alex 'Sandy´. Economic outcomes predicted by diversity in cities. *EPJ Data Sci.*, 9(1):17, 2020. doi: 10.1140/epjds/s13688-020-00234-x. URL `https://doi.org/10.1140/epjds/s13688-020-00234-x`.

[62] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[63] James S. Coleman. Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120, 1988. ISSN 00029602, 15375390. URL `http://www.jstor.org/stable/2780243`.

[64] Miles Corak. Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives*, 27(3):79–102, September 2013. doi: 10.1257/jep.27.3.79. URL `https://www.aeaweb.org/articles?id=10.1257/jep.27.3.79`.

[65] Robin Cowan and Nicolas Jonard. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, 28(8):1557–1575, 2004. ISSN 0165-1889. doi: https://doi.org/10.1016/j.jedc.2003.04.002. URL `https://www.sciencedirect.com/science/article/pii/S0165188903001520`.

[66] Jonathon N. Cummings. Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science*, 50(3):352–364, 2004. doi: 10.1287/mnsc.1030.0134. URL `http://dx.doi.org/10.1287/mnsc.1030.0134`.

[67] Sheldon H Danziger and Peter Gottschalk. *Uneven tides: Rising inequality in America*. Russell Sage Foundation, 1992.

[68] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

[69] Mathijs de Vaan, David Stark, and Balazs Vedres. Game changer: The topology of creativity. *American Journal of Sociology*, 120(4):1144–1194, 2015. ISSN 00029602, 15375390. URL `http://www.jstor.org/stable/10.1086/681213`.

[70] Carlo L Del Bello, Eleonora Patacchini, and Yves Zenou. Neighborhood effects in education. 2015.

[71] Paul DiMaggio and Filiz Garip. How network externalities can exacerbate intergroup inequality. *American Journal of Sociology*, 116(6):1887–1933, 2011. ISSN 00029602, 15375390. URL `http://www.jstor.org/stable/10.1086/659653`.

[72] Paul DiMaggio and Filiz Garip. Network effects and social inequality. *Annual review of sociology*, 38:93–118, 2012.

[73] Thomas A. DiPrete and Gregory M. Eirich. Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, 32(1):271–297, 2006. doi: 10.1146/annurev.soc.32.061604.123127. URL https://doi.org/10.1146/annurev.soc.32.061604.123127.

[74] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010. ISSN 0036-8075. doi: 10.1126/science.1186605. URL http://science.sciencemag.org/content/328/5981/1029.

[75] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

[76] James R Elliott. Social Isolation and Labor Market Insulation: Network and Neighborhood Effects on Less-Educated Urban Workers. *The Sociological Quarterly*, 40(2):199–216, 1999. ISSN 0038-0253. doi: 10.1111/j.1533-8525.1999.tb00545.x. URL http://www.jstor.org/stable/4121231http://www.jstor.org/stable/http://doi.wiley.com/10.1111/j.1533-8525.1999.tb00545.x.

[77] Richard M Emerson. Power-dependence relations. *American sociological review*, pages 31–41, 1962.

[78] Katherine Faust and Stanley Wasserman. Blockmodels: Interpretation and evaluation. *Social networks*, 14(1-2):5–61, 1992.

[79] Stephen E. Fienberg and Stanley Wasserman. An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association*, 76(373):54–57, 1981. ISSN 01621459. URL http://www.jstor.org/stable/2287039.

[80] Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404, 2001. ISSN 0165-1765. doi: https://doi.org/10.1016/S0165-1765(01)00394-9. URL https://www.sciencedirect.com/science/article/pii/S0165176501003949.

[81] Rico Fischer, Jorge C Leitao, Tiago P Peixoto, and Eduardo G Altmann. Sampling motif-constrained ensembles of networks. *Physical review letters*, 115(18):188701, 2015.

[82] David V Foster, Jacob G Foster, Peter Grassberger, and Maya Paczuski. Clustering drives assortativity and community structure in ensembles of networks. *Physical Review E*, 84(6):066117, 2011.

[83] Laura K. Gee, Jason Jones, and Moira Burke. Social networks and labor markets: How strong ties relate to job finding on facebook's social network. *Journal of Labor Economics*, 35(2):485–518, 2017. doi: 10.1086/686225. URL https://doi.org/10.1086/686225.

[84] Laura K. Gee, Jason J. Jones, Christopher J. Fariss, Moira Burke, and James H. Fowler. The paradox of weak ties in 55 countries. *Journal of Economic Behavior & Organization*, 133:362–372, 2017. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2016.12.004. URL https://www.sciencedirect.com/science/article/pii/S0167268116302864.

[85] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[86] Junxian Geng, Anirban Bhattacharya, and Debdeep Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526):893–905, 2019. doi: 10.1080/01621459.2018.1458618. URL https://doi.org/10.1080/01621459.2018.1458618.

[87] Lewis R Goldberg. The structure of phenotypic personality traits. *American psychologist*, 48(1):26, 1993.

[88] Claudia Goldin and Lawrence F. Katz. Human capital and social capital: The rise of secondary schooling in america, 1910-1940. *The Journal of Interdisciplinary History*, 29(4):683–723, 1999. ISSN 00221953, 15309169. URL http://www.jstor.org/stable/206979.

[89] Mark Granovetter. The strength of weak ties. *Journal of Sociology*, 78 (6):1360–1380, 1973. URL http://www.jstor.org/stable/2776392http://about.jstor.org/terms.

[90] Mark Granovetter. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3):481–510, 1985. doi: 10.1086/228311.

[91] Mark Granovetter. The Impact of Social Structure on Economic Outcomes Social Networks and Economic Outcomes: Core Principles. *The Journal of Economic Perspectives*, 19(1):33–50, 2005. URL http://www.jstor.org/stable/4134991{%}0Ahttp://about.jstor.org/terms.

[92] Mark S. Granovetter. *Getting a Job: A Study of Contacts and Careers*, volume 25. University of Chicago Press, 1996. ISBN 0226305813. doi: 10.2307/2077488. URL https://books.google.com/books/about/Getting{_}a{_}Job.html?id=R7-w4BLg7dAC.

[93] Patrick Hall, Navdeep Gill, Megan Kurka, and Wen Phan. Machine learning interpretability with h2o driverless ai. *H2O. ai*, 2017.

[94] Morten T. Hansen. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, 44(1):82–111, 1999. ISSN 00018392. URL http://www.jstor.org/stable/2667032.

[95] Amaç Herdağdelen, Bogdan State, Lada Adamic, and Winter Mason. The social ties of immigrant communities in the united states. In *Proceedings of the 8th ACM Conference on Web Science*, pages 78–84, 2016.

[96] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97 (460):1090–1098, 2002. doi: 10.1198/016214502388618906. URL https://doi.org/10.1198/016214502388618906.

[97] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137, 1983. ISSN 0378-8733. doi: https://doi.org/10.1016/0378-8733(83)90021-7. URL http://www.sciencedirect.com/science/article/pii/0378873383900217.

[98] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[99] Buster O. Holzbauer, Boleslaw K. Szymanski, Tommy Nguyen, and Alex Pentland. Social ties as predictors of economic development. In Adam Wierzbicki, Ulrik Brandes, Frank Schweitzer, and Dino Pedreschi, editors, *Advances in Network Science*, pages 178–185, Cham, 2016. Springer International Publishing. ISBN 978-3-319-28361-6.

[100] Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389, 2004. doi: 10.1073/pnas.0403723101. URL http://www.pnas.org/content/101/46/16385.abstract.

[101] Herminia Ibarra. Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative science quarterly*, pages 422–447, 1992.

[102] Yannis M Ioannides and Linda Datcher. Job Information Networks, Neighborhood Effects, and Inequality. *Journal of Economic Literature*, 42 (4):1056–1093, 2004. ISSN 0022-0515. doi: 10.1257/0022051043004595. URL http://www.jstor.org/stable/3594916http://about.jstor.org/termshttp://www.jstor.org/stable/3594916{%}5Cnhttp://www.jstor.org/stable/3594916.

[103] R. Mark Isaac and James M. Walker. Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, 103(1):179–199, 1988. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/1882648.

[104] Matthew Jackson. Policy cocktails: Attacking the roots of persistent inequality. *Policy*, 2021.

[105] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[106] Matthew O Jackson. An overview of social networks and economic applications. *Handbook of social economics*, 1:511–585, 2011.

[107] Matthew O Jackson. A typology of social capital and associated network measures. *Social Choice and Welfare*, 54(2):311–336, 2020.

[108] Matthew O Jackson. Inequality's economic and social roots: The role of social networks and homophily. *Available at SSRN 3795626*, 2021.

[109] Matthew O. Jackson and Asher Wolinsky. A strategic model of social and economic networks. *Journal of Economic Theory*, 71(1):44–74, 1996. ISSN 0022-0531. doi: https://doi.org/10.1006/jeth.1996.0108. URL `https://www.sciencedirect.com/science/article/pii/S0022053196901088`.

[110] Matthew O Jackson, Brian W Rogers, and Yves Zenou. The economic consequences of social-network structure. *Journal of Economic Literature*, 55(1): 49–95, 2017.

[111] Gregory A Janicik and Richard P Larrick. Social network schemas and the learning of incomplete networks. *Journal of personality and social psychology*, 88(2):348, 2005.

[112] Yuval Kalish. Bridging in social networks: Who are the people in structural holes and why are they there? *Asian Journal of Social Psychology*, 11(1):53–66, 2008. doi: https://doi.org/10.1111/j.1467-839X.2007.00243.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-839X.2007.00243.x`.

[113] Yuval Kalish and Garry Robins. Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks*, 28(1):56–84, 2006. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2005.04.004. URL `https://www.sciencedirect.com/science/article/pii/S037887330500033X`.

[114] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1), Jan 2011. ISSN 1550-2376. doi: 10.1103/physreve.83.016107. URL `http://dx.doi.org/10.1103/PhysRevE.83.016107`.

[115] Isabel Kloumann, Lada Adamic, Jon Kleinberg, and Shaomei Wu. The lifecycles of apps in a social ecosystem. In *Proceedings of the 24th International*

Conference on World Wide Web, WWW '15, page 581–591, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741684. URL https://doi.org/10.1145/2736277.2741684.

[116] Marie Lalanne and Paul Seabright. The old boy network: The impact of professional networks on remuneration in top executive jobs. 2016.

[117] Ron Laschever. The doughboys network: Social interactions and the employment of world war i veterans. *Available at SSRN 1205543*, 2013.

[118] Sandro Claudio Lera, Alex Pentland, and Didier Sornette. Prediction and prevention of disproportionally dominant agents in complex networks. *Proceedings of the National Academy of Sciences*, 117(44):27090–27095, 2020. ISSN 0027-8424. doi: 10.1073/pnas.2003632117. URL https://www.pnas.org/content/117/44/27090.

[119] Nan Lin. Building a Network Theory of Social Capital. *Connections*, 22 (1):28–51, 1999. ISSN 14691930. doi: 10.1108/14691930410550381. URL http://www.insna.org/PDF/Connections/v22/1999{_}I-1-4.pdfhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.3792{&}rep=rep1{&}type=pdf.

[120] Nan Lin. Social networks and status attainment. *Annual review of sociology*, 25(1):467–487, 1999.

[121] Nan Lin. Inequality in social capital. *Contemporary Sociology*, 29(6):785–795, 2000. ISSN 00943061, 19398638. URL http://www.jstor.org/stable/2654086.

[122] Nan Lin, Walter M . Ensel, and John C . Vaughn. Social Resources and Strength of Ties : Structural Factors in Occupational Status Attainment. *American sociological review*, 46(4):393–405, 1981. ISSN 00031224. doi: 10.2307/2095260. URL http://www.jstor.org/stable/2095260http://about.jstor.org/terms.

[123] Nan Lin, Walter M Ensel, and John C Vaughn. Social resources and occupational status attainment. *American sociological review*, 59(46):393–405, 1981. ISSN 0011-1384. doi: 10.1111/j.1745-9125.2001.tb00924.x. URL http://www.jstor.org/stable/2577987http://www.jstor.org/stable/2577987?seq=1{&}cid=pdf-reference{#}references{_}tab{_}contentshttp://about.jstor.org/termshttp://jaar.oxfordjournals.org/content/XXXVI/1/88.full.pdf.

[124] Miranda J Lubbers, José Luis Molina, Jürgen Lerner, Ulrik Brandes, Javier Ávila, and Christopher McCarty. Longitudinal analysis of personal networks. the case of argentinean migrants in spain. *Social Networks*, 32(1):91–104, 2010.

[125] Shaojun Luo, Flaviano Morone, Carlos Sarraute, Matías Travizano, and Hernán A Makse. Inferring Personal Economic Status from Social Network Location. *Nature Communications*, 8, 2017. ISSN 2041-1723. doi: 10.1038/ncomms15227. URL https://www.nature.com/articles/ncomms15227.pdfhttp://arxiv.org/abs/1704.01572.

[126] Winter Mason and Duncan J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1110069108. URL https://www.pnas.org/content/109/3/764.

[127] Steve McDonald, Nan Lin, and Dan Ao. Networks of Opportunity: Gender, Race, and Job Leads. *Social Problems*, 56(3):385–402, 2009. ISSN 00377791. doi: 10.1525/sp.2009.56.3.385.

[128] J. Miller McPherson and Lynn Smith-Lovin. Women and weak ties: Differences by sex in the size of voluntary organizations. *American Journal of Sociology*, 87 (4):883–904, 1982. ISSN 00029602, 15375390. URL http://www.jstor.org/stable/2778782.

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

[130] Ajay Mehra, Andrea L. Dixon, Daniel J. Brass, and Bruce Robertson. The social network ties of group leaders: Implications for group performance and leader reputation. *Organization Science*, 17(1):64–79, 2006. doi: 10.1287/orsc.1050.0158.

[131] Branko Milanovic. *Global income inequality: What it is and why it matters*. The World Bank, 2006.

[132] Allison Munch, J. Miller McPherson, and Lynn Smith-Lovin. Gender, children, and social contact: The effects of childrearing for men and women. *American Sociological Review*, 62(4):509–520, 1997. ISSN 00031224. URL http://www.jstor.org/stable/2657423.

[133] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2): 026126, 2003.

[134] Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2): 026126, 2003.

[135] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[136] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[137] Akihiro Nishi, Hirokazu Shirado, David G Rand, and Nicholas A Christakis. Inequality and visibility of wealth in experimental social networks. *Nature*, 526 (7573):426–429, 2015.

[138] David Obstfeld. Social networks, the tertius iungens orientation, and involvement in innovation. *Administrative Science Quarterly*, 50(1):100–130, 2005. doi: 10.2189/asqu.2005.50.1.100.

[139] John F. Padgett and Christopher K. Ansell. Robust action and the rise of the medici, 1400-1434. *American Journal of Sociology*, 98(6):1259–1319, 1993. ISSN 00029602, 15375390. URL http://www.jstor.org/stable/2781822.

[140] Elizabeth Levy Paluck, Hana Shepherd, and Peter M. Aronow. Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3):566–571, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1514483113. URL https://www.pnas.org/content/113/3/566.

[141] Tiago P. Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5(1), Mar 2015. ISSN 2160-3308. doi: 10.1103/physrevx.5.011033. URL http://dx.doi.org/10.1103/PhysRevX.5.011033.

[142] Tiago P Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317, 2017.

[143] Thomas Piketty. Capital in the 21st century. In *Inequality in the 21st Century*, pages 43–48. Routledge, 2018.

[144] Thomas Piketty and Emmanuel Saez. Income inequality in the united states, 1913–1998. *The Quarterly journal of economics*, 118(1):1–41, 2003.

[145] Ray Reagans and Bill McEvily. Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly*, 48(2):240–267, 2003. ISSN 00018392. URL http://www.jstor.org/stable/3556658.

[146] Ray Reagans and Ezra W. Zuckerman. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization Science*, 12(4):502–517, 2001. ISSN 10477039, 15265455. URL http://www.jstor.org/stable/3085985.

[147] Simon Rodan. Structural holes and managerial performance: Identifying the underlying mechanisms. *Social Networks*, 32(3):168–179, 2010. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2009.11.002. URL https://www.sciencedirect.com/science/article/pii/S0378873309000549.

[148] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[149] Abdolkarim Sadrieh and Harrie A.A. Verbon. Inequality, cooperation, and growth: An experimental study. *European Economic Review*, 50(5):1197–1222, 2006. ISSN 0014-2921. doi: https://doi.org/10.1016/j.euroecorev.2005.01.009. URL https://www.sciencedirect.com/science/article/pii/S0014292105000528.

[150] Emmanuel Saez and Gabriel Zucman. Wealth inequality in the united states since 1913: Evidence from capitalized income tax data. *The Quarterly Journal of Economics*, 131(2):519–578, 2016.

[151] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006. ISSN 0036-8075. doi: 10.1126/science.1121066. URL https://science.sciencemag.org/content/311/5762/854.

[152] Robert J Sampson. Individual and community economic mobility in the great recession era: The spatial foundations of persistent inequality. *Economic Mobility: Research and Ideas on Strengthening Families, Communities and the Economy*, pages 261–287, 2016.

[153] Jari Saramäki, E. A. Leicht, Eduardo López, Sam G. B. Roberts, Felix Reed-Tsochas, and Robin I. M. Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1308540110. URL https://www.pnas.org/content/111/3/942.

[154] Zuzana Sasovova, Ajay Mehra, Stephen P Borgatti, and Michaéla C Schippers. Network churn: The effects of self-monitoring personality on brokerage dynamics. *Administrative Science Quarterly*, 55(4):639–670, 2010.

[155] Thomas C Schelling. *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[156] Hirokazu Shirado, George Iosifidis, and Nicholas A. Christakis. Assortative mixing and resource inequality enhance collective welfare in sharing networks. *Proceedings of the National Academy of Sciences*, 116(45):22442–22444, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1911606116. URL https://www.pnas.org/content/116/45/22442.

[157] Irina Shklovski. *Residential Mobility, Technology &; Social Ties*, page 1787–1790. Association for Computing Machinery, New York, NY, USA, 2006. ISBN 1595932984. URL https://doi.org/10.1145/1125451.1125789.

[158] Georg Simmel. *The sociology of georg simmel*, volume 92892. Simon and Schuster, 1950.

[159] Georg Simmel. *Soziologie: Untersuchungen über die formen der vergesellschaftung*. BoD–Books on Demand, 2015.

[160] Joan Ellen Starker. *The development of a social network following geographic relocation.* PhD thesis, Portland State University, 1988.

[161] Jessica Su, Krishna Kamath, Aneesh Sharma, Johan Ugander, and Sharad Goel. An experimental study of structural diversity in social networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):661–670, May 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7332.

[162] Charles Tilly. *Durable inequality.* University of California Press, 1998.

[163] Gergő Tóth, Johannes Wachs, Riccardo Di Clemente, Ákos Jakobi, Bence Ságvári, János Kertész, and Balázs Lengyel. Inequality is rising where social network segregation interacts with urban topology. *Nature communications*, 12 (1):1–9, 2021.

[164] Peter Totterdell, David Holman, and Amy Hukin. Social networkers: Measuring and examining individual differences in propensity to connect with others. *Social Networks*, 30(4):283–296, 2008. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2008.04.003. URL https://www.sciencedirect.com/science/article/pii/S0378873308000269.

[165] Milena Tsvetkova, Claudia Wagner, and Andrew Mao. The emergence of inequality in social groups: Network structure and institutions affect the distribution of earnings in cooperation games. *PLOS ONE*, 13(7):1–16, 07 2018. doi: 10.1371/journal.pone.0200965. URL https://doi.org/10.1371/journal.pone.0200965.

[166] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012. doi: 10.1073/pnas.1116502109. URL http://www.pnas.org/content/109/16/5962.abstract.

[167] United Nations. Transforming our world: The 2030 agenda for sustainable development, 2020. URL https://sdgs.un.org/goals/goal10.

[168] Brian Uzzi. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review*, 61(4):674–698, 1996. ISSN 00031224. URL http://www.jstor.org/stable/2096399.

[169] Brian Uzzi. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, 42(1):35–67, 1997. ISSN 00018392. URL http://www.jstor.org/stable/2393808.

[170] Gwen Van Eijk. *Unequal networks: Spatial segregation, relationships and inequality in the city*, volume 32. Gwen van Eijk, 2010.

[171] Gil Viry. Residential mobility and the spatial dispersion of personal networks: Effects on social support. *Social Networks*, 34(1):59–72, 2012. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2011.07.003. URL https://www.sciencedirect.com/science/article/pii/S0378873311000475. Capturing Context: Integrating Spatial and Social Network Analyses.

[172] Jing Wang, Feng Fu, and Long Wang. Effects of heterogeneous wealth distribution on public cooperation with collective risk. *Physical Review E*, 82(1): 016102, 2010.

[173] Qi Wang, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. Urban mobility and neighborhood isolation in america's 50 largest cities. *Proceedings of the National Academy of Sciences*, 115(30):7735–7740, 2018.

[174] Yuchung J. Wang and George Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987. doi: 10.1080/01621459.1987.10478385. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1987.10478385.

[175] Stanley Wasserman and Katherine Faust. Canonical analysis of the composition and structure of social networks. *Sociological Methodology*, pages 1–42, 1989.

[176] Richard J Williams and Neo D Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–183, 2000.

[177] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010. ISSN 0036-8075. doi: 10.1126/science.1193147. URL http://science.sciencemag.org/content/330/6004/686.

[178] Cornelia Wrzus, Martha Hänel, Jenny Wagner, and Franz J Neyer. Social network changes and life events across the life span: a meta-analysis. *Psychological bulletin*, 139(1):53, 2013.

[179] Xiaoran Yan, Cosma Shalizi, Jacob E Jensen, Florent Krzakala, Cristopher Moore, Lenka Zdeborová, Pan Zhang, and Yaojia Zhu. Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(5):P05007, 2014.

[180] Akbar Zaheer and Giuseppe Soda. Network evolution: The origins of structural holes. *Administrative Science Quarterly*, 54(1):1–31, 2009.

[181] Yafei Zhang, Lin Wang, Jonathan JH Zhu, Xiaofan Wang, and Alex 'Sandy' Pentland. The strength of structural diversity in online social networks. *Research*, 2021, 2021.