EUI DEPARTMENT OF ECONOMICS

Essays in the Economics and Econometrics of Networks and Peer Effect

Zheng Wang

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Florence, 23 May 2023

European University Institute
Department of Economics

Essays in the Economics and Econometrics of Networks and Peer Effect

Zheng Wang

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Examining Board

Prof. Andrea Ichino, EUI, Supervisor
Prof. Sule Alan, EUI, Co-Supervisor
Prof. Eric Auerbach, Northwestern University
Prof. Yann Bramoullé, Aix-Marseille School of Economics

Declaration

Researcher declaration to accompany the submission of written work
Department of Economics – Doctoral Programme

I, Zheng Wang, certify that I am the author of the work “Essays in the Economics and Econometrics of Networks and Peer Effect” I have presented for examination for the Ph.D. at the European University Institute. I also certify that this is solely my own original work.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications. I certify that this work complies with the Code of Ethics in Academic Research issued by the European University Institute (IUE 332/2/10 (CA 297).

The copyright of this work rests with its author. Quotation from this thesis is permitted, provided that full acknowledgement is made. This work may not be reproduced without my prior written consent. This authorisation does not, to the best of my knowledge, infringe the rights of any third party. I declare that this work consists of 32069 words.

Signature and date:

Zheng Wang

15.05.2023

Acknowledgments

In this moment of celebration, I want to thank everyone and everything that has happened to my life journey.

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Andrea Ichino and Sule Alan, as well as my mentors, Fabrizia Mealli and Yann Bramoullé, for your unwavering support throughout this journey. Each of you has gone out of your way to help me reach this milestone, and I am deeply grateful for the time and effort you have invested in me. It has truly been a privilege to have had the opportunity to learn from you.

I also want to thank my parents, who have never failed to give me the freedom to make my own choices in life. Your unconditional support and love allowed me to explore and pursue my passions. I am forever indebted to you for making me the person I am today.

Numerous friends and coworkers have supported me through these years. I am deeply appreciative of all the interactions we had, especially the fun times with Comida. I want to thank Dalila and Alaitz, in particular, for always being there for me, be it at work, at home, or at Fiasco.

Finally, my deepest gratitude also goes to Folker for accompanying me on this journey every step of the way, through every obstacle and triumph. I am truly grateful for your companionship and the invaluable support you have provided me with. Your selfless care and unwavering dedication have transformed the challenging task of writing a thesis into an enjoyable experience.

Abstract

This thesis contributes to the understanding of peer effects, both methodologically and empirically.

The endogeneity of network formation has been a major obstacle to the study of peer influence. The first and the second chapters of the thesis propose a causal identification solution in the potential outcome framework. Combining results from multiple causal inference and statistical network analysis, I show that confounding can be addressed by inferring propensity scores of network link formation from the adjacency matrix. This identification strategy imposes minimum restrictions on the data-generating process and, unlike existing econometric solutions, does not rely on any parametric modelling. As an application, I estimate the effect of high school friendships on bachelor's degree attainment. While previous literature finds that exposure to more high-achieving boys makes girls less likely to obtain a bachelor's degree, I show that if the girls consider the boys as friends, their interactions induce a positive impact instead. Since friendship endogeneity has been addressed, the estimated effect is causal.

The third chapter looks at the peer effects generated by group competition. It focuses on the gender differences in preference for competition in a setting where the competition does not involve face-to-face confrontation, and effort is the only determinant of the final ranking. I first develop a model of group competition with heterogeneous preference for ranking. With empirical implications generated from the theoretical model, I then test the gender difference in the preference parameter using web-scraped data from Duolingo, a free online foreign-language learning platform with over 300 million users. Every week, language learners on Duolingo are randomly allocated to groups of 30 people to compete on the number of language lessons completed during that week. The empirical results suggest

in this setting, females have a stronger preference for ranking than males.

Contents

Chapter 1

The linking effect: causal identification and estimation of the effect of peer relationship

1.1 Introduction

Interest in understanding the impact of peer influence within economic and social networks has been growing rapidly in the economics literature, with an increasing emphasis on establishing causality. Knowing how connected agents are affected by each other is important, as welfare can be improved through cultivating certain relationships while discouraging others. However, due to the difficulties in addressing network endogeneity, the causal impact of many important types of relationships, such as friendships, buyer-supplier networks and banking networks, remain understudied.

The difficulty in establishing causal identification partly comes from the lack of a causal framework where treatments and potential outcomes are explicitly defined. In this paper, I propose to treat each potential relationship as a unique treatment. In other words, the existence of each network link is the subject of manipulation or intervention in a hypothetical experiment where we could assign network links at will.[1] This view of what constitutes

[1] In a network with N nodes, each node will have $N-1$ potential network links to form. In other words, the number of potential treatments is $N-1$ for each node.

a treatment contrasts with the existing literature on peer effects, where the treatment is implicitly assumed to be some summary statistics of the entire network, such as the share of one's connected network nodes with certain characteristics.[2] I call the effect of relationships the *linking effect*, emphasising the fact that the treatment is the assignment of links. Explicitly viewing every pairwise relationship as a treatment opens the door to building upon existing causal inference tools for the study of the linking effect. In particular, due to the multiplicity of possible relationships for any network node, we are able to embed the analysis of the network linking effect in the multiple causal inference framework.

This newly discovered connection between these two previously disassociated literature turns out to be highly consequential for the causal identification of the linking effect in endogenous networks. By combining a recent finding in the multiple causal inference literature (Wang and Blei, 2019) and theoretical results in the statistical network analysis literature, I prove that the linking effects are identified under a set of general assumptions. The first assumption is the "doubly individualistic assignment mechanism" assumption, which states that there exist some random variables such that after conditioning on these random variables, the distribution of network links is conditionally independent.[3] This assumption rules out the case where a link directly affects the formation of another link, such as in a marriage network where being married to one person rules out marriage links to all the other people. The second assumption is the "no single-link confounder" assumption. It requires that any variable that affects the outcome variable must affect the formation of more than one link. This assumption is likely to hold in networks of non-trivial size because as the number of possible links to form increases, it becomes more and more difficult to conceive an individual-level confounding variable that affects the formation of only one of these links but not any other. The final assumption is the positivity assumption, which requires that for every pair of nodes on the network, the probability of establishing a link is strictly between 0 and 1, a standard assumption in causal inference.

A direct consequence of the first two assumptions is that the propensity scores of pairwise linking can be identified from the distribution of network links. This is because an unob-

[2] In Manski (1993), this treatment is associated with the contextual peer effect.

[3] These random variables are the ones forming graphons, the canonical form of vertex exchangeable graphs.

served *sufficient confounder*, defined as a random variable that captures all the confounding factors, can be identified up to a measure-preserving transformation. In particular, the first assumption rules out the existence of any multi-link confounders other than the sufficient confounder, and the existence of single-link confounders is assumed away by the second assumption (Wang and Blei, 2019). Even though this sufficient confounder is not directly observed in the data, it is nonetheless identified up to a measure-preserving transformation from the distribution of network links as the number of nodes goes to infinity (Diaconis and Janson, 2007). This identification result means that the propensity scores of pairwise linking can be inferred from the adjacency matrix (Zhang et al., 2017; Auerbach, 2022), allowing the use of propensity score-based estimators to address confounding.

Unlike traditional propensity score estimation procedures where the probability of treatment is regressed on a set of observed pre-treatment variables, here the propensity scores are estimated using only the observed network links, that is, the treatments themselves. One way to operationalize the estimation is to use probabilistic factor models to capture the joint distribution of the links (Wang and Blei, 2019). This involves specifying the distributions of the sufficient confounder and the distributions of the network links conditional on the sufficient confounder. It is, however, not important which specific distributions one chooses to use, as long as the overall joint distribution of the network links is well captured. An alternative is to estimate the propensity scores with procedures developed in the network link prediction literature (e.g. Zhang et al., 2017; Olhede and Wolfe, 2014). With the estimated propensity scores, we can then use inverse probability weighting, subclassification, or propensity score matching to estimate the desired causal effect.

Thanks to these identification and estimation results, this paper will conduct one of the first empirical analyses aiming to understand the causal effect of one of the most well-known endogenous networks, friendships. Despite being the main focus of the social network literature, the impact of friendship networks has not been well-understood empirically due to the endogeneity problem. The only few existing papers that attempted to address the endogeneity issue did so by both restricting the way friendships are formed and the variables that affect this formation, subjecting the estimated results to bias when the true network formation process has a different form (e.g. Goldsmith-Pinkham and Imbens, 2013; Gagete-

Miranda, 2020).

Most papers in the empirical peer effect literature circumvent the endogeneity issue by looking at other social networks which are quasi-randomly formed. For example, Cools et al. (2022) investigates how the presence of more high-achieving male and female students in high school affects boys' and girls' bachelor's degree attainment differently. They do so by exploiting the random variations in cohort composition, a strategy commonly employed in the peer effect literature (e.g. Hoxby, 2000; Olivetti et al., 2020, etc.). Cools et al. (2022) finds that being exposed to more male high achievers decreases girls' likelihood of obtaining a bachelor's degree, in part by decreasing their confidence and aspiration. While these studies offer exciting findings on the effect of cohort composition, a common drawback is that the impact of social interactions cannot be separated from the influence of other factors that also vary across cohorts, such as differences in teachers' attitudes. Moreover, some of the most meaningful social interactions with long-term consequences only exist among close friends and not those who simply attend the same school during the same year. As a consequence, the patterns of peer influence among friends have largely remained unknown.

Using high school friendship data from AddHealth,[4] the same dataset used by Cools et al. (2022) and many other studies on social networks (e.g. Goldsmith-Pinkham and Imbens, 2013; Bifulco et al., 2014; Badev, 2021; Olivetti et al., 2020), I test whether the negative impact of high-achieving male students on female students persists when these boys are considered friends by the girls. Interestingly, I find that an additional male high-achieving friend causally increases the probability that a female student obtains a bachelor's degree by 3 p.p. Further analysis suggests that this positive influence results from an increase in their confidence and not in their academic ability measured by GPA. Indeed, having one more male high-achieving friend means the female student becomes 3.75 p.p more likely to self-report being more intelligent than their same-age peers, but no effect is found for their grades in any of the main subjects.[5] Taking these results together with the findings of Cools et al. (2022), it seems that girls are intimidated by high-achieving boys whom they do not

[4] AddHealth, or the National Longitudinal Study of Adolescent to Adult Health, is a dataset of representative US high schools.

[5] Both the self-reported intelligence and the grades are measured one year after the friendship data was collected. The main subjects are math, science, English, and history.

have close relationships with, but are encouraged by those whom they see as friends. This suggests that a possible way to boost the confidence of female students and increase their chances of graduating from college is by fostering friendships with high-achieving boys in their high school.

This paper is closely related to the literature on peer effect, especially the contextual peer effect defined in Manski (1993). Roughly speaking, contextual peer effect refers to the effect of peer characteristics on own outcome and is usually expressed as a parameter in a regression model. In order to give a causal interpretation to the estimated parameter, empirical researchers have taken advantage of settings with either random treatments or random peers. The former is where peer relationships are fixed and characteristics of the network nodes are randomized, while the latter is where nodal characteristics are fixed but peer relationships are randomized. In other words, the former is related to treatment spillover, while the latter is about the linking effect. Because these two cases correspond to two completely different hypothetical interventions, using one parameter to represent their effects can sometimes lead to misleading interpretations of the estimates.[6] My paper avoids the issue of misinterpretation by developing a causal framework tailored for the study of linking effects where random peers are a special case.[7] Since in the linking effect framework the only type of treatment is the existence of the links, the interpretation of the estimates is clear.

To the best of my knowledge, Li et al. (2019) and Basse et al. (2019) are the only papers to have made the distinction between randomized treatments and randomized peers using a formalized causal framework. However, the focus of their papers is on inference issues rather than identification, as they only consider cases where agents are assigned to groups randomly. They also focus their analysis on peer networks with a non-overlapping group structure, such as roommate networks. My framework, in contrast, allows the networks to have arbitrary structures and is suitable for analyzing both experimental and observational data.

[6] See Bramoullé et al. (2020) for more analysis on the problem of misinterpretation.

[7] If peer relationships are randomized, there will be no need to address the confounding (endogeneity) problem. The causal framework of the linking effect can still be used; the only difference is that there will be no need to infer the unobserved confounders and use them to correct for confounding, as randomization guarantees no confounding exists.

In terms of identification, several econometrics solutions have been proposed to tackle the network endogeneity issue for Manski (1993)'s linear-in-means model. The majority do so by jointly modeling the outcome equation and the network formation equation. From Goldsmith-Pinkham and Imbens (2013) and Hsieh and Lee (2016), to Arduini et al. (2015) and Johnsson and Moon (2021), then to Auerbach (2022), assumptions used to achieve identification have been progressively relaxed. My paper takes a step further and reveals the minimum set of assumptions needed for identification.[8] Even though the assumptions of my paper are formulated in the potential outcome framework, they can be translated into modeling restrictions in the linear-in-means regression context. This translation exercise leads to three important observations. First, all aforementioned papers impose the assumptions that form this minimum set. Second, some assumptions made in the aforementioned papers are unnecessary because they are implied by the minimum set of assumptions. Third, neither outcome modeling nor network formation modeling is necessary for identification. This means we do not need to know which observed and unobserved variables enter the outcome equation and network formation equation or how they enter the equations, be it additive, multiplicative, or interactive. In fact, not only is it unnecessary, but it could be harmful because incorrectly specifying these equations could lead to biased estimates.

The rest of the paper is organised as follows. Section 1.2 gives the formal definitions of the treatment and the potential outcome, based on which several linking effect estimands to study peer influence are proposed. Section 1.3 provides the identification conditions and Section 1.4 discusses how existing propensity score-based estimators can be adapted for estimation. Section 1.5 gives simulation evidence on the bias reduction performance of the proposed identification and estimation strategy. Finally, Section 1.6 applies these estimators to real data to study the effect of high school friendship on students' bachelor's degree attainment. Section 1.7 concludes.

[8] More specifically, this is the minimum set of assumptions needed when identification achieved with a single network. Jochmans (2020) proposes and identification strategy based on multiple networks.

1.2 Treatments, potential outcomes, and estimands

Suppose we are interested in a certain peer relationship network with N nodes and directed links among these nodes. A link from one node to another represents the existence of a directed peer relationship. The adjacency matrix $\mathbf{D}$ of the network is a N by N matrix where each entry represents the existence of a link:

$$\mathbf{D} = \begin{bmatrix} 0 & D_1^2 & ... & D_1^N \\ D_2^1 & 0 & ... & D_2^N \\ ... & ... & ... & ... \\ D_N^1 & D_N^2 & ... & 0 \end{bmatrix},$$

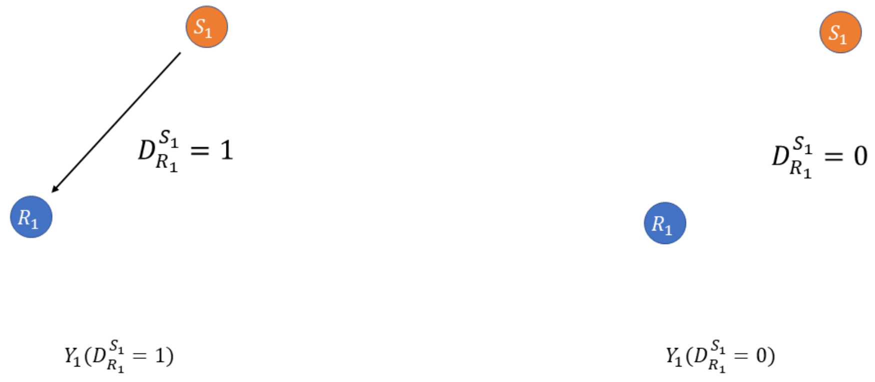
In this paper, I will write $D_i^j = 1$ if there is a directed link from node j to node i. The diagonal of the adjacency matrix is 0 because we do not allow one to be their own peer. When a node is on the receiving end of the link, I call it the *link receiver.* When a node is on the sending side of the link, I call it a *link sender.* A node can act as a link receiver in one link while acting as a link sender in another and vice versa. In this paper, the outcomes of interest are measured on the link receivers, but we could just as easily measure outcomes on the link senders. When I write a pair of nodes (i, j), the first component is the link receiver, and the second component is the link sender. Whenever suitable, I also use subscripts to indicate the link receiver and superscripts as the link sender.

1.2.1 Treatments and potential outcomes

The treatment of interest is the (directed) linking status among pairs of network nodes. For example, for a friendship network, the treatment of interest would be the directed friendship from one person to another.[9] With two hypothetical pairwise relationships, Figure 1.1 highlights the hypothetical intervention, aka the treatment, that is the focus of this paper. Each relationship has three components: the receiver (R), the sender (S), and the linking status (D). In this example, the two relationships have the same receiver and sender but have

[9] Friendship doesn't need to be a reciprocal relationship, as one person consider another person as a friend doesn't necessarily mean the other way holds. This is evidenced by the friendship nominations of high school students in the Add Health data.

Figure 1.1: Hypothetical intervention: two counterfactuals



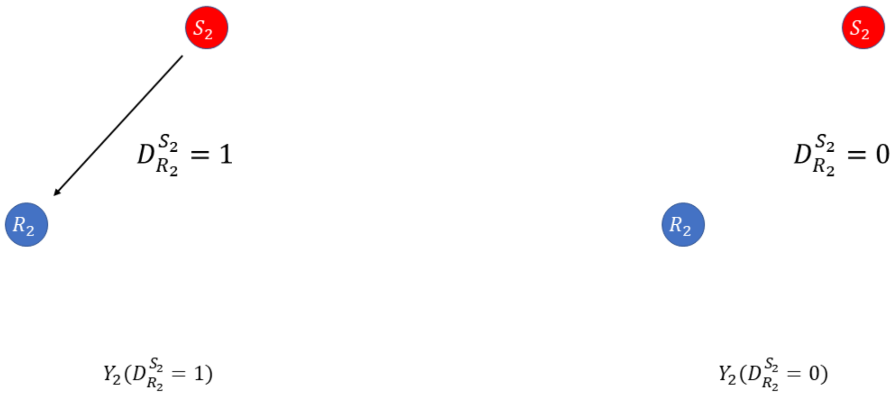
different linking statuses. On the left, the link from the sender to the receiver exists, but on the right, the link doesn't exist. The type of causal question this paper asks is " What would the receiver R_1's potential outcome be if it were "treated" with a link from sender S_1 (left panel of Figure 1.1), and what would the potential outcome be if it weren't " treated" with this link (right panel of Figure 1.1), and the difference between the two potential outcomes? ". In other words, what is the difference between $Y_1(D_{R_1}^{S_1} = 1)$ and $Y_1(D_{R_1}^{S_1} = 0)$? The only difference between the two hypothetical cases is the existence of the directed link from the sender to the receiver. This is why we call the linking status the "treatment".

It is important to emphasize that the hypothetical intervention this paper studies is *not* the change in the sender characteristics.[10] In this paper, link sender nodal characteristics define the multiple versions of the treatment. As an example, consider color as the nodal characteristic.[11] Figure 1.2 shows two hypothetical relationships between R_1 and a different sender S_2, where S_2 is red while S_1 is orange. This means the effect of $D_{R_1}^{S_1}$ on R_1 could be different from the effect of $D_{R_1}^{S_2}$ on R_1, therefore a link from S_2 should be viewed as a different treatment than a link from S_1. In the most general case, we could allow linking effects to

[10] It is, however, the focus of the treatment spillover literature.

[11] For instance, Li et al. (2019) and Basse et al. (2019) assume the effect of linking only depends on some observed characteristic of the node chosen by the researcher ex-ante.

Figure 1.2: A different link sender



differ in arbitrary observed and unobserved sender nodal characteristics. This is the stance taken by this paper. As a result, links from senders with different *identities* are viewed as different treatments. Since sender identity and the link itself has a one-to-one relationship in this paper, I sometimes also refer the link sender as the treatment. However, it should be clear that the hypothetical intervention is on the relationship instead of the sender.

Given that any link receiver could potentially receive a link from $N - 1$ different link senders, and each of these links is considered a unique treatment with a unique effect on the receiver, we are in the case of *multiple treatments*, or multi-cause, causal inference. In other words, for any link receiver i, its *treatment* is a vector of $N - 1$ linking status $\mathbf{D}_i := (D_i^1, ..., D_i^{i-1}, D_i^{i+1}, ..., D_i^N)$.

In traditional treatment causal inference, the potential outcome of any subject, the entity that bears the treatment and whose outcome is measured, could depend on the treatment status of all subjects in the population if no further assumption is made. The Stable Unit Treatment Assumption (SUTVA) restricts the potential outcome to depend only on the subject's own treatment status. Here I will make a similar assumption to allow potential outcomes to only depend on the receiver's own treatment status. As just discussed, for any receiver i', because her treatment is a vector of all pairwise linking status with the senders, this means i's potential outcome can be a function of all pairwise linking status where i is the receiver, but couldn't depend on the linking status where i is not the receiver. I call this assumption the Linking-effect Stable Unit Treatment Unit Assumption (L-SUTVA) to differentiate it from the usual SUTVA.

Assumption 1 (L-SUTVA)**.**

$$Y_i(\mathbf{D}_i, \mathbf{D}_{-i}) = Y_i(\mathbf{D}_i, \tilde{\mathbf{D}}_{-i})$$

for any $(\mathbf{D}_{-i}, \tilde{\mathbf{D}}_{-i})$ and any i, where $\mathbf{D}_{-i} = (\mathbf{D}_1, ..., \mathbf{D}_{i-1}, \mathbf{D}_{i+1}, ..., \mathbf{D}_N)$.

Under L-SUTVA, the potential outcome can be written as $Y_i(\mathbf{D}_i)$ or $Y_i(D_i^1, D_i^2, ..., D_i^N)$. In traditional causal inference, SUTVA is sometimes called the no-interference assumption. However, this paper studies the effect of relationships, which suggests agents must interact or interfere in some way. At first sight, the two may seem to be at odds. The reason why L-SUTVA is perfectly compatible with the study of linking effect lies in the definition of treatment. Recall what SUTVA says is that the *treatment assignment* of one subject does not interfere with another subject's *potential outcome*. In particular, it doesn't require the non-existence of network structure among the units. Whether SUTVA is likely to hold depends on the definition of treatment and potential outcome. In this paper, since the treatment is the relationship, the no-interference assumption implied by L-SUTVA means that one's potential outcome is only affected by one's own relationships. L-SUTVA helps reduce the space of possible potential outcomes and makes it easier to identify and estimate causal estimands. In this paper, I will always assume that L-SUTVA holds.[12]

1.2.2 Estimands

With the perspective that relationships are multiple treatments, causal estimands could be flexibly defined by contrasting different types of potential outcomes. In this section, I will focus on the most straightforward set of estimands, which, loosely speaking, looks at the

[12] L-SUTVA might not be realistic in some situations. In the future, I will extend the analysis by relaxing L-SUTVA to allow some interference.

effect of an additional link. Several other possible estimands, including the commonly used linear-in-means estimands, are outlined in Section 2.1.

As a fist step, I define the pairwise estimand τ_i^j as the following contrast of i's potential outcomes:

$$\tau_i^j = Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}) - Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})$$

where $\mathbf{D}_i^{-j} = (D_i^1, ..., D_i^{j-1}, D_i^{j+1}, ..., D_i^N)$, and $\bar{\mathbf{d}}_i^{-j}$ is the corresponding vector of the *realised* or *observed* treatment assignment for i after taking out d_i^j. τ_i^j contrasts link receiver i's potential outcome when it receives treatment (a link) from link sender j with its potential outcome when it doesn't receive the link from j, while keeping the linking status from other link senders fixed at their observed value.

As in traditional causal inference where individual causal effect is not identifiable, here the pairwise linking effect is also not identifiable due to the fact that only one potential outcome is ever observed for a given node. However, an average causal effect is potentially identifiable. Next, I define the average treatment effect of a link from link senders with some attributes $A = a$ to link receivers with certain attributes $R = r$.

$$\tau_r^a := \mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}) - Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}) | A^j = a, R_i = r]$$

where $\mathbb{E}_{(i,j)}$ represents the the expectation over the distribution generated by random sampling of pairs of nodes from the super-population. The interpretation of these estimands deserves some special attention. Under L-SUTVA, these estimands are well-defined and can be interpreted as the all-or-nothing effect in the following sense. Take τ^a as an example, it can be interpreted as the *expected* contrast between the *average* potential outcome of assigning a sender-j link to *everyone* in the node set and the *average* potential outcome of assigning a sender-j link to *no one* in the node set, where this j is *chosen randomly* (hence the expected contrast) with equal probability from the set of link senders with attribute $A^j = a$. The interpretation of τ_r^a is similar to that of τ^a, except that instead of looking at all link receivers in the node set, now we only look at link receivers with $R_i = r$. However, similar estimands can also be defined without the assumption of L-SUTVA. In this case, we could simply modify the potential outcome function to include the entire adjacency matrix $(D_i^j, \mathbf{D}_{-(i,j)})$. But we can no longer interpret the estimands as the all-or-nothing effect. This is because when we simultaneously change $(D_1^j, D_2^j, ..., D_N^j)$ for a given sender j, $\mathbf{D}_{-(i,j)}$ is no longer at its observed value. Instead, the estimands need to be interpreted as the *expected* treatment effect of j on a *randomly chosen* link receiver i, again keeping the other links at their realized value. The difference is that in the second interpretation, in every hypothetical experiment, intervention is only done on one link, and the average linking effect τ^j is the average from repeated experiments where a different link is modified each time. This is similar to the $EATE$ in Sävje et al. (2021) and the τ defined in Forastiere et al. (2021).

1.2.3 Relationship between the linking effect and the contextual peer effect

The linking effect is related with the contextual peer effect defined in Manski (1993). Contextual peer effect is the effect of peer characteristics on own outcome and is expressed as a parameter in a regression model. However, this parameter does not have a clear causal interpretation because it is associated with two distinct types of treatments. To see this, note that peer characteristic is defined over two variables: the network adjacency matrix and the vector of the characteristics of all network nodes. Intervention on the adjacency matrix and intervention on the nodal characteristics correspond to two completely different causal effects, namely the linking effect and the spillover effect.

The linking effect is typically related to studies where random peers are used to establish causality. For example, many empirical papers utilise the random formation of dorm rooms, classrooms, and cohorts (e.g. Sacerdote, 2001; Carrell et al., 2013; Cools et al., 2019; Olivetti et al., 2020, etc.). It is the effect of forming groups (network links) in a certain way while keeping the characteristics of the people involved fixed. Such effects could inform policymakers of the benefit and cost of forming groups in certain ways but cannot reveal the exact mechanism behind such effects: whether it's due to differences in gender, race, social economic status, GPA, some unobserved characteristics, or a certain combination of all of the above.

In contrast, the treatment spillover effect literature deals with the case where the network structure is fixed, but some treatment, e.g. vaccine, is assigned to everyone in the network. It studies the effect of other nodes' treatment status on one's own outcome. This effect can tell us how someone is affected by certain characteristics (vaccinated or not) of the others when these characteristics are manipulated. But it cannot inform policymakers how outcomes will change if they were to manipulate the network structure so that one is connected to someone with or without those characteristics, simply because the effect was not estimated from an experiment where the network structure is manipulated. Indeed, any network structure manipulation would not only result in changes in a single characteristic of one's peers but many other, possibly unlimited number of characteristics of one's peers. After all, the identities of their peers have been changed. Therefore, the causal framework established by this paper to study the linking effect complements the existing treatment spillover literature and completes the mission of giving a clear causal interpretation to the contextual peer effect parameter in any context.

1.3 Identification

At the center of causal identification is the treatment assignment mechanism. In experimental studies where network links are randomized, the assignment mechanism is known, such as the cases studied in Sacerdote (2001); Carrell et al. (2013); Li et al. (2019); Basse et al. (2019); Olivetti et al. (2020). In this case causal identification does not pose any challenge. However, in non-experimental studies with observational data, the assignment mechanism is unknown, and assumptions must be imposed on it to make causal discoveries. This is the case with endogenously formed peer networks. These assumptions on the links assignment mechanism are variations of the three assumptions used in traditional causal inference: individualistic assignment mechanism, unconfoundeness, and positivity.

1.3.1 Doubly individualistic assignment mechanism

In the study of linking effects, the individualistic assignment mechanism assumption takes the form of conditionally independent links across both the link receivers and the link senders. In other words, conditional independence is required both across the subjects and the treatments.

Assumption 2 (Doubly individualistic assignment mechanism)**.** There exists sequences of random variables (vectors) $\{\mathbf{U}_i\}_{1\leq i\leq N}$ and $\{\mathbf{V}_i\}_{1\leq i\leq N}$ such that equation (1.1) holds.

$$Pr_d(\mathbf{D} = \mathbf{d}|\mathbf{U}_1, ..., \mathbf{U}_N, \mathbf{V}_1, ..., \mathbf{V}_N) = \prod_{i=1}^{N}\prod_{j\neq i}^{N} Pr_d(D_i^j = d_i^j|\mathbf{U}_i, \mathbf{V}_j) \tag{1.1}$$

where Pr_d indicates the probability distribution over random treatment (link) assignment. We can think of $\mathbf{U}_i$ as link receiver specific variables and $\mathbf{V}_j$ as link sender specific variables. For any node i, $\mathbf{U}_i$ and $\mathbf{V}_i$ could share some common components. For example, for a high school friendship network, the ambition of student i could affect both from whom they receive links through $\mathbf{U}_i$ and to whom they send links through $\mathbf{V}_i$.

To compare the linking effect doubly individualistic assignment mechanism assumption with the individualistic assignment mechanism sssumption of traditional causal inference, it is useful to rewrite the "Doubly" assumption as follows.

$$\begin{aligned}
&Pr_d(\mathbf{D} = \mathbf{d}|\mathbf{U}_1, ..., \mathbf{U}_N, \mathbf{V}_1, ..., \mathbf{V}_N) \\
&= \prod_{i=1}^{N} Pr_d(\mathbf{D}_i = \mathbf{d}_i|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N) \\
&= \prod_{i=1}^{N} Pr_d(D_i^1 = d_i^1, ..., D_i^N = d_i^N|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N) \\
&= \prod_{i=1}^{N}\prod_{j\neq i}^{N} Pr_d(D_i^j = d_i^j|\mathbf{U}_i, \mathbf{V}_j)
\end{aligned}$$

Recall that the vector $\mathbf{D}_i$ represents the treatment assignment vector of link receiver i, therefore the first equation, and equivalently, the second equation, is exactly the individualistic assignment mechanism sssumption in traditional causal inference, which states that conditional on some random variables, the treatment assignments across *subjects*, in our case, the link receivers, are independent.[13] This assumption always holds if we view the subjects as randomly sampled from some superpopulation, as a result of the De Finetti's theorem (Imbens and Rubin, 2015).[14] On top of that, the linking effect doubly individualistic assignment

[13] $\mathbf{V}_1, ..., \mathbf{V}_N$ can be thought of as the parameters related to each treatment.

[14] Superpopulation sampling is a perspective commonly adopted in traditional causal inference. See Imbens and Rubin (2015) and Hernán and Robins (2020) for more discussions on this.

mechanism assumption also assumes that for each link receiver i, the assignment of each individual link across all link senders are independent conditional on some link sender specific variable: as the name "doubly" suggests. However, if the network nodes are randomly sampled from a superpopulation, the linking effect doubly individualistic assignment mechanism assumption will be satisfied as a direct result of the Aldous-Hoover Theorem (Crane, 2018), the equivalence of the De Finetti's Theorem for network data.[15] This means with the super population perspective, both doubly individualistic assignment mechanism assumption and the usual individualistic assignment mechanism sssumption will automatically hold.

When the data is *not* sampled from a superpopulation, the assignment mechanism is usually modeled as a stochastic process. For example, we might view the choice of a binary treatment as the result of a random utility model. In traditional causal inference, for the given treatment, the individualistic assignment mechanism sssumption restricts this stochastic process of treatment assignment to be conditionally independent across subjects. In the network case, the observed nodes may also be regarded as the finite population itself. But because nodes are both link receivers and link senders, they are both subjects and treatments. This is why modeling of two (or double) stochastic processes is needed. The first part of the doubly individualistic assignment mechanism assumption requires that in the modeled stochastic process, each link receiver is independently assigned the vector of all links, conditional on their own receiver specific variables. Here, "individualistic", or "independence", is with regard to the subjects, or the link receivers. The second part of the assumption requires that in the modelled stochastic process, for each link receiver, their link assignment from each sender is independent across all link senders, conditional on the sender specific variables. Here "individualistic", or "independence", is with regard to the treatments, or the link senders. This means with the finite population perspective, the doubly individualistic assignment mechanism assumption requires more restrictions than the usual individualistic assignment aechanism assumption.

The second layer of the doubly individualistic assignment mechanism assumption requires that for any given link receiver, when they decide which links to form, the linking decisions must be mutually independent to some extent. That means even though the decisions might

[15] More details of this are provided in Section 2.3.1.

not be unconditionally mutually independent, they must be conditionally mutually independent. This excludes some networks, such as those with a non-overlapping group structure by construction. For example, a roommate network cannot be conditionally mutually independent. This is because if i and j are roommates, and j and k are not roommates, that means i and k are not roommates, no matter what variables are conditioned on. However, the assumption does accommodate cases where networks are formed with strategic considerations, as long as the equilibrium linking decisions are not direct functions of each other. An example where the assumption could be satisfied is the case analyzed by Leung (2015). In that paper, the network formation game is characterized by strategic interactions with incomplete information, where utility depends on the entire network structure. The idea is that when the agents' objective is to maximize their *expected* utility, i's linking decisions will be a function of equilibrium beliefs about others' linking decisions, which is a function of the observed attributes of all agents in the network. This means for each agent i, her linking decisions are not directly dependent of each other. If we allow independent utility shocks for all her linking decisions, the doubly individualistic assignment mechanism assumption will be satisfied. More details of this example are given in Section 2.3.2.

It is important to point out that Assumption 2 is different from, and in fact, less restrictive than the assumption underlying dyadic regressions. Dyadic regressions, such as those analyzed in Graham (2020), usually assumes that linking decisions are independent conditional on some observed attributes X and unobserved latent attributes ϵ satisfying $\mathbb{E}[\epsilon|X = 0]$:

$$Pr_d(\mathbf{D} = \mathbf{d}|X_1, X_2, ..., X_N, \epsilon_1, \epsilon_2, ..., \epsilon_N) = \prod_{i=1}^{N} \prod_{j \neq i}^{N} Pr_d(D_i^j = d_i^j|X_i, X_j, \epsilon_i, \epsilon_j) \tag{1.2}$$

Running a dyadic regression requires one to impose a functional form for the pairwise linking probability: $Pr(D_i^j = 1|X_i, X_j, \epsilon_i, \epsilon_j) = f(X_i, X_j, \epsilon_i, \epsilon_j)$ for some known f. This functional form differentiates dyadic regressions and the assumption of doubly individualistic assignment mechanism. When the functional form restriction does not reflect the true data-generating process, the parameters in dyadic regressions are biased for the true effect of X

on the pairwise linking probabilities and inference is invalid.[16] But why is the functional form assumption necessary for dyadic regressions but not for this paper? This is due to different objectives of the two cases. The goal of dyadic regressions is usually to estimate the parameters associated with the observed covariates to understand the role of these covariates in determining linking probabilities, such as those in estimating the gravity models studying the association between GDP and trade flow. In contrast, this paper aims to identify and estimate the causal parameters of the outcome equation. Assumptions on link formation, are only used to correct for confounding. Identifying such causal effects does not require knowing the functional form of the linking equation. Therefore, there is no need to estimate parameters associated with the observed attributes.

1.3.2 Unconfounded Assignment Mechanism

Definition 1.3.1 (Confounder)**.** A *confounder* is a pre-treatment variable that is associated with both the treatment and the outcome.

Definition 1.3.2 (Cause)**.** A random variable X is a *cause* of another random variable Y if X is realised before Y and is associated with Y.

Assumption 3 (No single-link confounder)**.** Any confounder must be a cause of more than one link.

Proposition 1 (Unconfoundedness)**.** Let $\mathbf{Y}_i^{pot}$ be the vector of i's potential oucomes. Under Assumption 2 and Assumption 3, the following holds:

$$Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j, \mathbf{Y}_i^{pot}) = Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j)$$

for $\mathbf{U}_i, \mathbf{V}_j$ defined in equation (1.1).

Proof. First, suppose there exists another variable $\mathbf{U}_i'$ that is a cause of $\mathbf{Y}_i^{pot}$ and a cause of more than one of the links that i potentially receives, say D_i^1 and D_i^2. Then if we omit $\mathbf{U}_i'$

[16] To see why inference is invalid, note that mis-specifying the functional form will make the linking probabilities dependent across pairs, while pairwise independence is crucial for likelihood based inference.

in the conditioning set of

$$
\begin{aligned}
&Pr_d(\mathbf{D}=\mathbf{d}|\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N)\\
=&Pr_d(\mathbf{D}_1=\mathbf{d}_1,...,D_i^1=d_i^1,D_i^2=d_i^2,...,D_i^N=d_i^N,...,\mathbf{D}=\mathbf{d}|\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N)
\end{aligned}
$$

D_i^1 and D_i^2 couldn't be conditionally independent (without conditioning on $\mathbf{U}_i'$). This is a contradiction to our starting point, which is that conditioning on $\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N$ makes all links independent (equation (2)). In other words, the variables $\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N$ by definition make all links independent, and the existence of $\mathbf{U}_i'$ is in contradiction of that definition.

With similar logic, suppose there exists a variable $\mathbf{V}_j'$ that is a cause of $\mathbf{Y}_i^{pot}$ and is a cause of more than one of the links that j potentially sends, say D_1^j and D_2^j. Then if we omit $\mathbf{V}_j'$ in the conditioning set of

$$
\begin{aligned}
&Pr_d(\mathbf{D}=\mathbf{d}|\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N)\\
=&Pr_d(D_1^2=d_1^2,...,D_1^j=d_1^j,...,D_1^N=d_1^N,...,D_2^1=d_2^1,...,D_2^j=d_2^j,...,D_2^N=d_2^N,...,\\
&D_N^1=d_N^1,...,D_N^j=d_N^1,...,D_N^{N-1}=d_N^{N-1}|\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N)
\end{aligned}
$$

D_1^j and D_2^j couldn't be conditionally independent (without conditioning on $\mathbf{V}_i'$). This again is a contradiction to our starting point, which is that conditioning on $\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N$ makes all links independent (equation 2).

By Assumption 3, which states confounders that only affect one link don't exist, we have effectively ruled out the existence of any confounders that affect the formation of any link. This means

$$
\begin{aligned}
&Pr_d(D_i^j=1|\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N,\mathbf{Y}_i^{pot})\\
=&Pr_d(D_i^j=1|\mathbf{U}_1,...,\mathbf{U}_N,\mathbf{V}_1,...,\mathbf{V}_N)\\
=&Pr_d(D_i^j=1|\mathbf{U}_i,\mathbf{V}_j)
\end{aligned}
$$

The last equation comes from equation (2). This argument is a direct adaption of the identification proof in Wang and Blei (2019)

□

The intuition is that $\mathbf{U}_i, \mathbf{V}_j$ must include all the multiple-link causes of link formation. Otherwise, the doubly individualistic assignment mechanism assumption wouldn't have held. Some of these causes will confound the outcome variable; some will not. In theory, we only need to condition on the confounders, but the insight is that since we don't know which are the confounders, conditioning on all of these causes will for sure address confounding. It is for this reason that I call $\mathbf{U}_i, \mathbf{V}_j$ the *sufficient confounders*.

Next, I prove that the unconfoundedness condition also holds conditional on the propensity score based on $\mathbf{U}_i, \mathbf{V}_j$.

Definition 1.3.3 (Pairwise propensity score)**.** A pairwise propensity score $e(u, v)$ is defined as

$$e(u, v) := \mathbb{E}_{(i,j)}[Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j)|\mathbf{U}_i = u, \mathbf{V}_j = v]$$

Because $\mathbb{E}_{(i,j)}[Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j)|\mathbf{U}_i = u, \mathbf{V}_j = v] = Pr_d(D_i^j = 1|\mathbf{U}_i = u, \mathbf{V}_j = v)$, the propensity score $e(u, v)$ is equal to the link assignment probability.

Lemma 1. $e(\mathbf{U}_i, \mathbf{V}_j)$ is a balancing score, that is:

$$Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j, e(\mathbf{U}_i, \mathbf{V}_j)) = Pr_d(D_i^j = 1|e(\mathbf{U}_i, \mathbf{V}_j))$$

Lemma 2 (Unconfoundedness given $e(\mathbf{U}_i, \mathbf{V}_j)$)**.**

$$Pr_d(D_i^j = 1|Y_i^{pot}, e(\mathbf{U}_i, \mathbf{V}_j)) = Pr_d(D_i^j = 1|e(\mathbf{U}_i, \mathbf{V}_j))$$

This result is similar to the propensity score property result in the traditional causal inference, where unconfoundedness holds given the propensity score. The proof of Lemma 2 is given in Section 2.4.2.

1.3.3 From Unconfoundedness to the Identification of Estimands

Assumption 4 (Pairwise Positivity)**.** The link assignment probability satisfies

$$0 < Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j) < 1$$

for each possible $\mathbf{U}_i, \mathbf{V}_j$.

Proposition 2. Under Assumptions 2-4, the direct linking effect is identified given the true pairwise propensity scores, in the sense that the estimand can be expressed in terms of the observed outcome. For link receivers with characteristics r and link senders with characteristics a, this means

$$\begin{aligned}\tau_r^a = &\mathbb{E}\left[\mathbb{E}_{(i,j)}[Y_i^{obs}|e(\mathbf{U}_i, \mathbf{V}_j), D_i^j = 1, A^j = a, R_i = r]|A^j = a, R_i = r\right] \\ &- \mathbb{E}\left[\mathbb{E}_{(i,j)}[Y_i^{obs}|e(\mathbf{U}_i, \mathbf{V}_j), D_i^j = 0, A^j = a, R_i = r]|A^j = a, R_i = r\right]\end{aligned}$$

This is proved in Section 2.4.3. Note that here we only need pairwise positivity because the estimand is defined through pairwise contrasts where all non-focal pairwise links are kept at their realised value. If we want to define an estimand where all of one's links are manipulated simultaneously, we will run into a problem where the positivity condition will fail. This is discussed more in detail in Section 2.1.1. Similar discussions can be found in Imai and Jiang (2019); Johnsson and Moon (2021); Auerbach (2022).

1.3.4 Relationship with previous literature

Assumptions 2-4 are the minimum set of assumptions for identification. Without any one of these assumptions, the linking effect cannot be identified through the unconfoundedness condition. To see how the identification strategy relates to the existing econometric methods, I will first express my assumptions in terms of modeling assumptions. A general outcome model is given by equation (1.3), and a general link formation model that satisfies Assumption 2 can be expressed by equation (1.4). T is the treatment of interest. For example, T_i is the mean characteristic of one's peers in linear-in-means models and is D_i^j for the estimand considered by this paper. κ, w, ϵ, and η are all unobserved variables. For simplicity, I do not include any observed covariates in the model, but including them doesn't pose additional complication.

$$Y_i = g(T_i, \kappa_i, \epsilon_i) \tag{1.3}$$

$$D_i^j = \mathbb{1}\{f(w_i, w_j) \geq \eta_i^j\} \quad i \neq j \tag{1.4}$$

where $T_i \not\perp\!\!\!\perp \kappa_i$, $T_i \perp\!\!\!\perp \epsilon_i | \kappa_i$, $\eta_i^j \perp\!\!\!\perp w_i, w_j$ and elements of $\{\eta_i^j\}_{i,j=1}^N$ are independently distributed. The difference in κ_i and ϵ_i in equation is that κ_i confounds the outcome Y_i while ϵ_i does not. Note that since we can always assign $w = (U, V)$ and find a f' such that $f(w_i, w_j) = f((U_i, V_i), (U_j, V_j)) = f'(U_i, V_j)$, equation (1.4) is not restrictive. To reiterate, the doubly individualistic assignment mechanism assumption is equivalent to $\eta_i^j \perp\!\!\!\perp w_i, w_j$ and elements of $\{\eta_i^j\}_{i,j=1}^N$ being independently distributed. Second, the no single-link confounder assumption is equivalent to requiring η_i^j and ϵ_i being independent for all i, j. Recall that the no single-link confounder assumption says that any random variable that is a cause of only one link cannot be a cause of the outcome. Here in the link formation equation, the only single-link cause is η_i^j because w_i (w_j) enters all link formation equitation where i (j) is a node. Therefore, η_i^j not being a confounder is equivalent to η_i^j being independent of κ_i, ϵ_i. Finally, positivity says that $0 < Pr(T_i = t|w) < 1$ for all $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of all treatment values of interest.

Comparing these assumptions with the assumptions made in previous econometric papers (Goldsmith-Pinkham and Imbens, 2013; Hsieh and Lee, 2016; Arduini et al., 2015; Johnsson and Moon, 2021; Auerbach, 2022), the first obvious difference is on the modelling of the outcome equation. All previous papers are concerned with a linear outcome model. In particular, they assume $Y_i = T_i\beta + \lambda(w_i) + \epsilon_i$. That is, w_i is assumed to be the same as κ_i, meaning w_i are exactly the confounders, and the marginal effect of T_i is constant (β) and w_i enters the equation separately through some function $\lambda(\cdot)$. A limitation of the identification strategies based on this model is that they cannot deal with the case where the effect of T_i is heterogeneous in w_i. In contrast, I do not make such restrictions on the outcome model. The identification result from Wang and Blei (2019) says w_i is able to capture κ_i, but they don't need to be equal. The second major difference is in the modeling of the network formation equation. Except for Auerbach (2022), all the other papers achieve identification through decomposing w into observed and unobserved components and specifying how these components enter the equation, e.g., through homophily or additive. In contrast, just like this paper, Auerbach (2022) does not make any functional assumptions on equation (1.4). Intuitively, both papers address confounding by using only the information provided by the adjacency matrix.

To understand how the two papers differ in terms of identification assumptions, I list five of Auerbach (2022)'s main assumptions and discuss how they relate to the assumptions of this paper. Note that because Auerbach (2022) assumes $Y_i = T_i\beta + \lambda(w_i) + \epsilon_i$ for some unknown λ, I will continue the discussion on the premises that it is the true outcome data-generating process.

1. The random sequence $\{T_i, w_i, \epsilon_i\}_{i=1}^N$ is independent and identically distributed with entries mutually independent of $\{\eta_i^j\}_{i,j=1}^N$. This is the assumption of no single-link confounder.

2. $\{\eta_i^j\}_{i,j=1}^N$ are i.i.d and $\eta_i^j \perp\!\!\!\perp w_i, w_j$. I also assume this. As discussed above, this is related to the assumption of doubly individualistic assignment mechanism.

3. $\mathbb{E}[\epsilon_i|T_i, w_i] = 0$, that is, the treatment T_i is unconfounded after conditioning on w_i. Proposition 1 of this paper proves that unconfounedness is implied, not assumed, by the doubly individualistic assignment mechanism assumption and the no single-link confounder assumption.

4. There is variation in T_i after conditioning on w_i. This is related to the positivity condition that $0 < Pr(T_i = x|w_i) < 1$ for all $x \in \mathcal{X}$, because if there is no variation in T_i after conditioning on w_i, $Pr(T_i = x|w_i)$ must be either 0 or 1, violating the positivity assumption.

5. The function $f(w_i, \cdot)$ is enough for controlling for the confounding from $\lambda(w_i)$. This assumption is not needed when T_i is peer characteristics because $f(w_i, \cdot)$ is actually the generalized propensity score. As all propensity scores, it has the property that

unconfoundedness holds by either conditioning on w_i or conditioning on the propensity score of w_i

1.4 Estimation

The estimation of the linking effect involves two steps. The first step is to estimate the propensity scores. Unlike in traditional causal inference, the propensity scores estimated in the first step are functions of unobserved latent variables. Therefore, the traditional propensity score estimation methods won't apply here. In Section 1.4.1, I show how techniques developed in the graphon estimation literature in network analysis and the multiple treatment literature in causal inference can be used for propensity score estimation. The second step is to use the estimated propensity scores to estimate the linking effects. Here many established methods from traditional causal inference can be used, such as inverse probability weighting (IPW), propensity score matching, and propensity score subclassification. In Section 1.4.2, I will illustrate how the inverse probability weighting method can be used to estimate the linking effects. Propensity score matching and subclassification can be adapted similarly as shown in Section 2.2.

1.4.1 1st-step estimation: propensity scores

Graphon Estimation

As discussed in Section 2.3.1, the linking probability in a graphon and the propensity score e_{ij} are, in fact, exactly the same when nodes are randomly sampled from superpopulation. This means we could use the many statistical methods in graphon estimation to estimate the propensity scores. Here I briefly discuss how the neighborhood smoothing method proposed by Zhang et al. (2017) works. Compared to other graphon estimation methods, such as stochastic block models (Olhede and Wolfe, 2014), it has the advantage of not making restrictive assumptions on how links are formed.

First let's define a probability slice as $e(\mathbf{U}_i, \cdot) = (e(\mathbf{U}_i, \mathbf{V}_1), e(\mathbf{U}_i, \mathbf{V}_2), ..., e(\mathbf{U}_i, \mathbf{V}_N))$. The main idea is that for any link receiver i, if we could find other link receivers with similar probability slices as i, we could then use the realized treatment assignment of these link receivers to estimate $(e(\mathbf{U}_i, \mathbf{V}_1), e(\mathbf{U}_i, \mathbf{V}_2), ..., e(\mathbf{U}_i, \mathbf{V}_N))$. Specifically, let $\mathcal{N}_i := \{i' : e(\mathbf{U}_{i'}, \cdot) \approx e(\mathbf{U}_i, \cdot)\}$ be the neighbourhood of link receiver i. Then an estimator for $e_i^j := e(\mathbf{U}_i, \mathbf{V}_j)$ would be

$$\tilde{e}_i^j = \frac{\sum_{i' \in \mathcal{N}_i} D_{i'}^j}{|\mathcal{N}_i|}$$

To define the neighborhood, we first need a definition of similarity, or equivalently the distance, between probability slices. Zhang et al. (2017) uses the d^2 distance:

$$d(i, i') = ||e(\mathbf{U}_i, \cdot) - e(\mathbf{U}_{i'}, \cdot)||_2 = \left\{ \int_v |e(\mathbf{U}_i,) - e(\mathbf{U}_{i'}, v)|^2) \right\}^{1/2}$$

Then

$$\begin{aligned} d(i, i')^2 &= \int_v e(u_i, v)e(u_i, v) + \int_v e(u_{i'}, v)e(u_{i'}, v) - 2 \int_v e(u_i, v)e(u_{i'}, v) \\ &= \int_v (e(u_i, v) - e(u_{i'}, v))e(u_i, v) + \int_v (e(u_{i'}, v) - e(u_i, v))e(u_{i'}, v) \\ &\leq \left| \int_v (e(u_i, v) - e(u_{i'}, v))e(u_{\tilde{i}}, v) \right| + \left| \int_v (e(u_i, v) - e(u_{i'}, v))e(u_{\tilde{i}'}, v) \right| + 2e_N \\ &\leq \max_{k \neq i, i'} 2 \left| \int_v (e(u_i, v) - e(u_{i'}, v))e(u_k, v) \right| + 2e_N \end{aligned}$$

where $\tilde{i}$ and $\tilde{i}'$ are such that $|u_{\tilde{i}} - u_i| \leq e_N$ and $|u_{\tilde{i}'} - u_{i'}| \leq e_N$, and e_N depends on n and is the error rate. Zhang et al. (2017) shows that such $\tilde{i}$ and $\tilde{i}'$ can be found with high probability.

The first part of $\max_{k \neq i, i'} 2 \left| \int_v (e(u_i, v) - e(u_{i'}, v))e(u_k, v) \right|$ can be estimated by

$$\tilde{d}(i, i') = \max_{k \neq i, i'} \frac{|(\mathbf{D}_i - \mathbf{D}_{i'})\mathbf{D}'_k|}{n}.$$

Intuitively, neighbourhood $\mathcal{N}_i$ should include i' with small $\tilde{d}(i, i')$. Zhang et al. (2017) defines $\mathcal{N}_i$ as

$$\mathcal{N}_i = \{i' \neq i : \tilde{d}(i, i') \leq q_i(m)\}$$

where $q_i(m)$ is the m'th quantile of $\{i' \neq i : \tilde{d}(i, i')\}$. Zhang et al. (2017) showed that with $m = C(n^{-1}logn)^{1/2}$ for any constant $C \in (0, 1]$, if the propensity score function $e(\cdot, \cdot)$ is Piecewise-Lipschitz, then $\tilde{e}_i^j$ is consistent for $e(\mathbf{U}_i, \mathbf{V}_j)$.[17] [18]

Factor models

Propensity scores $e(\mathbf{U}_i, \mathbf{V}_j)$ can also be estimated with factor models. This method requires us to specify the distribution of $\mathbf{U}_i$, $\mathbf{V}_j$, and $Pr(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j)$ for all $i, j = 1, ...N$. For exposition purposes, let $\mathbf{U}_i = (U_{1i}, U_{2i})$ and $\mathbf{V}_j = (V_{1j}, V_{2j})$ be vectors of length 2. A simple factor model could be

$$\begin{aligned} &\alpha, U_{1i}, U_{2i}, V_{1j}, V_{2j} \sim \mathcal{N}(0, 1), \quad i, j = 1, ..., N \\ &e(\mathbf{U}_i, \mathbf{V}_j) = logit(\alpha + U_{1i}V_{1j} + U_{2i}V_{2j}), \quad i, j = 1, ..., N \end{aligned} \tag{1.5}$$

Even though estimating the propensity scores with factor models impose additional functional form assumptions on the network formation, they are very flexible and versatile. Every aspect of the model can be modified, including the length of unobserved sufficient confounders, their distributions and how these confounders enter the probability distribution of the propensity scores, be it additive or multiplicative, be it linear or quadratic.[19]

To operationalize the use of factor models, I follow the deconfounding procedure proposed by Wang and Blei (2019). The deconfounder is a procedure proposed by Wang and Blei (2019) to address confounding in the setting of multiple treatments. It can be used in our setting because each link can be viewed as a treatment. Applying the deconfounder to estimate the propensity scores involves three steps. In the first step, we need to randomly

[17] Definition of Piecewise-Lipschitz: For any $\delta, L > 0$, let $\mathcal{F}_{\delta;L}$ denote a family of piecewise-Lipschitz functions m: $[0, 1]^2 \to [0, 1]$ such that (i) there exists an integer $K \geq 1$ and a sequence $0 = x_0 < \cdots < x_K$ satisfying $min_{0 \leq s \leq K-1}(x_{s+1} - x_s) \geq \delta$, and (ii) both $|e(u_1, v) - e(u_2, v)| \leq L|u_1 - u_2|$ and $|e(u, v_1) - e(u, v_2)| \leq L|u_1 - u_2|$ hold for all $u, u_1, u_2 \in [x_s, x_{s+1}]$, $v, v_1, v_2 \in [x_t, x_{t+1}]$ and $0 \leq s, t \leq K - 1$.

[18] Auerbach (2022) uses a similar idea but bounds the distance $d(i, i')$ with a different metric.

[19] Not only can we choose another family of distributions, it is also possible to allow dependence among $\mathbf{U}$ and $\mathbf{V}$ by adding another layer of factorization. For example,

$$\begin{aligned} w_i, w_j &\sim \mathcal{N}(0, 1), \quad i, j = 1, ..., N \\ U_{1i}, U_{2i}, V_{1j}, V_{2j}|w_i, w_j &\sim \mathcal{N}(w_i + w_j, 1), \quad i, j = 1, ..., N \end{aligned}$$

select a portion of links in the adjacency matrix and set them to 0, effectively partitioning it into training data and validation data. In the second step, we need to pick a factor model and fit the factor model with the training data. In the third step, validation data is used to compute a test statistics to decide whether the factor model fits the data well enough. If the test is passed, then we proceed to the estimation with the estimated propensity scores. If the test fails, then another factor model could be used and step two repeated until we find a factor model that passes the test.[20]

1.4.2 2nd-step estimation: treatment effect

Once the propensity scores of link formation are estimated, we could use the many propensity score-based methods commonly used in the treatment effect estimation literature to estimate the linking effects of interest. In this section, I will use an inverse probability weighting estimator to illustrate how these propensity score-based methods can be adapted to the current setting. The most basic IPW estimator is the Horvitz–Thompson estimator. The augmented inverse probability weighting (AIPW) could be used to include covariates in the outcome model. AIPW is commonly referred to as the doubly robust estimator because it is consistent if either the propensity score is correctly estimated or the outcome model is correctly specified.

The IPW estimator for the linking effect

$$\tau_r^a := \mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}) - Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})]$$

is

$$\frac{1}{\sum_{i=1}^N R_i = r} \cdot \frac{1}{\sum_{j=1}^N A^j = a} \Big(\sum_{i:R_i=r} \sum_{j:A^j=a} \frac{D_i^j \cdot Y_i^{obs}}{e(\mathbf{U}_i, \mathbf{V}_j)} - \sum_{i:R_i=r} \sum_{j:A^j=a} \frac{(1-D_i^j) \cdot Y_i^{obs}}{1 - e(\mathbf{U}_i, \mathbf{V}_j)} \Big) \tag{1.6}$$

where $e(\mathbf{U}_i, \mathbf{V}_j)$ is substituted with its estimate since the true propensity score is unknown. Same as the conventional IPW estimator, the IPW estimator in equation 1.6 is unbiased

[20] The idea of using a statistical test on validation data to see if propensity scores are accurately estimated can also be used for the neighborhood smoothing estimator, or any other graphon estimator. In fact, a similar idea was used in Zhang et al. (2017) to compare the performance of different graphon estimators.

for the linking effect τ_r^a. The proof is detailed in Section 2.4.4. A regression model (1.7) can be used to incorporate additional pre-treatment control variables, where each pairwise observation is weighted based on their propensity score. The additional control variables help reduce finite sample biases just as in the traditional augmented inverse probability weighting estimator.

$$Y_i = \alpha + \beta D_i^j + \theta Controls + \epsilon_i^j \tag{1.7}$$

1.5 Simulation

In this section, I conduct simulation exercises with synthetic data to assess the performance of the proposed linking effect estimators. I will generate the synthetic data according to the data generation model (1.8); one is a version of the homophile model, and the other is a statistical block model. Then I use a factor model to estimate the propensity scores. These propensity scores are then used in the second stage estimation with three different estimators: the inverse probability weighting (IPW) estimator, the nearest matching estimator, and the subclassification estimator. Finally, I will compute the bias and the mean absolute error (MAE) of the estimates relative to the true effect and compare them with the bias and MAE of the naive OLS estimator that ignores confounding.

$$
\begin{aligned}
\epsilon_i^c &\sim \mathcal{N}(0,1) \\
\epsilon_i^b &\sim U[0,1] \\
X_i &\sim Bernoulli(0.6) \\
C_i &\sim U[0,1] \\
\eta_{ij} &\sim U[0,1] \\
D_{ij} &= \mathbb{1}\{g(C_i, C_j) \geq \eta_{ij}\}, \quad g = g1, g2 \\
Y_i^c &= \alpha^c + \mathbf{D}_i\beta^c + \gamma^c C_i + \delta^c X_i + \epsilon_i^c \\
Y_i^b &= \mathbb{1}\{logit(\alpha^b + \mathbf{D}_i\beta^b + \gamma^b C_i + \delta^b X_i) \geq \epsilon_i^b\} \\
&\text{where } logit(s) = \frac{1}{1+exp(-s)}
\end{aligned} \tag{1.8}
$$

where $(\alpha^c, \gamma^c, \delta^c) = (0.5, 4, 1)$, $(\alpha^b, \gamma^b, \delta^b) = (-4, 4, 1)$. $\beta^{c(b)} = (\beta_1^{c(b)}, ...\beta_j^{c(b)}, ..., \beta_N^{c(b)})$ is a vector of parameters relating to the causal effect of a link from sender j. I set $\beta_j^c = X_j/2$ for all j. $\beta_j^b = X_j/2$ for all j. g_1 is specified in equation (1.9) and g_2 is specified in Section 2.5.1, equation (2.5).

$$
g_1 : P_i^j = 1/5\big(1 + exp(-(-6 + 2.5C_1 + 1.5C_j + |C_i - C_j|))\big) \tag{1.9}
$$

In this simulation exercise, I consider both continuous and binary outcome variables, which are denoted by Y^c and Y^b, respectively. The network links are generated through a binomial process with success probability specified according to two different link generation processes, g_1 as in equation (1.9) and g_2 as in equation (2.5). g_1 incorporates both degree heterogeneity and homophily. On the one hand, it is an increasing function in C_i and C_j. On the other hand, the probability of linking increases as the difference in C_i and C_j becomes smaller between the link receiver and the link sender. g_2 corresponds to a stochastic block model. The details of g_2 and its corresponding simulation result is detailed in Section 2.5.1. Both g_1 and g_2 generate directed networks. In our setup C_i is the confounder. It enters both the outcome and link formation equations and is unobserved to the econometrician. C_i, X_i,

ϵ_i^c,ϵ_i^b and η_{ij} are independent of each other, for all $i, j = 1, ...N$.

The mean degree distribution of g_1 from the simulated datasets is given in Table 1.1. As the network size increases, the degree increases. This is because the linking probability doesn't change as the network grows in our link generation model. This means the more nodes there are in the network, the more link senders there are, and thus the more links a link receiver will have.

Table 1.1: Mean degree distribution for simulated g1 networks

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
N=100	0	0	0	0	0	0.1	0.8	1	1.1	1.9	4.2
N=300	0	0	0.2	1	1	1.6	2	2.6	3.3	4.7	10.2
N=500	0	0.2	1	1.5	2	2.7	3.2	4.1	5.5	7.4	15.7

Note: This table reports the mean degree distribution of the simulated networks. For each size N=100,300,500, and for each simulated network of that size, I caculate the deciles of the number of links each link receiver receives, and average over all the 500 simulated networks of that size.

For the continuous outcome, I estimate the linking effects with the linear OLS regression (1.10), and the binary outcome is estimated with the logistic regression (1.11). I run these regressions separately for link senders with $X_j = 0$ and link senders with $X_j = 1$ to study the effects of these link senders separately. For the propensity score-based methods, the regressions are weighted with weights based on propensity scores that correct for confounding. For the naive OLS, the regressions are unweighted, thus not correcting for any confounding. The target estimand in this simulation exercise is ATT. This choice is reflected in the regression weights.

$$Y_i^c = \mu_i = \rho_0 + \rho_1 D_i^j + v_i \tag{1.10}$$

$$Pr(Y_i^b = 1) = \frac{1}{1 + exp\big(-(\rho_2 + \rho_3 D_i^j)\big)} \tag{1.11}$$

Table 1.2 compares the bias and MAE for the three propensity score-based estimators and the naive ols estimator. The propensity score used in this table is estimated using the factor model specified in equations (1.12)-(1.14). Comparing this factor model to the one in Section 1.4.1, Z_i can be seen as $\gamma_i \mathbf{U}_i$ and $\mathbf{V}_j$ can be seen as $\beta_j \mathbf{V}_j$ where $\mathbf{U}_i$ for $i = 1, ..., N$ are vectors of length two. The number of matches for the matching estimator is 1, and the number of subclasses for the subclassification estimator is 8. The rows under X_0 are the estimates for the linking effect of a link from a sender with $X_j = 0$, whose true effects are 0 on both the binary and the continuous outcomes. The rows under X_1 are the estimates for the linking effect of a link from a sender with $X_j = 1$, whose true effect is 0.5 on the continuous outcome. The true effect of an additional link from a sender with $X_j = 1$ on the binary outcome depends on the number of other links from senders with $X_j = 1$ because the true data generation process is non-linear. It is therefore calculated from the data generation process for each observation and then averaged over all observations.

$$Z_i = (z_{1i}, z_{2i}) \sim \mathcal{N}(0,1) \times \mathcal{N}(0,1), \quad i = 1, ..., N \tag{1.12}$$

$$K_j = (k_{1j}, k_{2j}) \sim \mathcal{N}(0,1) \times \mathcal{N}(0,1), \quad j = 1, ..., N \tag{1.13}$$

$$D_i^j | Z_i, K_j \sim Bernoulli\big(logit(Z_i + K_j)\big), \quad i, j = 1, ..., N \tag{1.14}$$

From Table 1.2, we can see that the estimators based on the propensity scores estimated by the factor model offer significant bias reduction compared to the naive ols estimator. The inverse probability weighting estimator performs the best among the three propensity score-based estimators. Compared to the naive ols estimator, the inverse probability weighting estimator reduces 90% - 97% of the biases for the binary outcome and 51% - 83% of the biases for the continuous outcome. As the network becomes larger, the bias reduction increases. An interesting observation from the table is that the bias from the naive ols estimator increases as the network becomes larger. This is because as the network becomes larger, the number of links for link receivers increases. This will lead to increasingly larger accumulated linking effects from all the other links being attributed to the effect of the link under consideration as in equation (1.6). This phenomenon doesn't happen if confounding is corrected because, in this case, the other links are independent of the link under consideration. As we see from the first three columns, the bias from the propensity score-based estimators continues to decrease as N increases despite the increasing bias from the naive ols estimator. Table (2.5) in Section 2.5 shows similar results for the statistical block model for network formation.

In Section 2.5, I also show the biases and MAEs of propensity score-based estimators using the factor model estimated propensity scores concerning the estimators using the true propensity scores (Table 2.8 and Table 2.9). Finally, I show simulation results when I increase the number of matches (from 1 to 3 to 5) and the number of subclasses (from 8 to 10 to 12) as the size of the network increases. The results from these different comparisons stay similar to the ones shown in Table 1.2.

1.6 Empirical Application

Almost everyone would agree that friendship is one of the most important social networks in a person's life. After all, one does not simply spend time with their friends; they also share information, receive their help, value their opinions, mimic their actions, and learn from their experiences. But it would be much more difficult to get everyone to agree on the direction and extent to which a person would be affected by their friends. The social network literature has long been interested in understanding the pattern of peer influence among friends for outcomes including risky behavior, smoking habits, obesity, education level, labor outcomes, fertility, etc. However, due to the obstacle posed by endogenous friendship formation, these questions remain largely unanswered, at least not in ways where the endogeneity issue is adequately accounted for.

Thanks to the theoretical results developed in this paper, I am able to make one of the first steps toward uncovering the true impacts of friendship. With the AddHealth data, I will be investigating the patterns of peer influence among high school friends in the U.S. Specifically, I look at how students' probability of graduating from college is affected by having more high-achieving friends, and whether this effect differs by both the gender of themselves and the gender of the high-achieving friend.[21] The analysis is inspired by the recent paper by Cools et al. (2022), which also uses the AddHealth data and finds that being exposed to more high-achieving males in one's high school decreases the likelihood that a

[21] A high-achieving student is defined as a student who has at least one residential parent with a postgraduate degree. This is the same definition used in Cools et al. (2022)

Table 1.2: Simulation results for g_1

				IPW	Matching	Sub	Naive ols
Yb	Bias	X_0					
			N=100	0.077445	0.096851	0.093895	0.132864
			N=300	0.051917	0.086736	0.091974	0.1705
			N=500	0.033176	0.084199	0.087752	0.184117
		X_1					
			N=100	0.078718	0.094838	0.092844	0.132779
			N=300	0.04753	0.083476	0.087602	0.166086
			N=500	0.034532	0.085371	0.089037	0.185265
	MAE	X_0					
			N=100	0.102707	0.137418	0.111374	0.142808
			N=300	0.054298	0.087305	0.091974	0.1705
			N=500	0.03435	0.084199	0.087752	0.184117
		X_1					
			N=100	0.09447	0.11907	0.103271	0.137679
			N=300	0.050589	0.0838	0.087611	0.166086
			N=500	0.036061	0.085371	0.089037	0.185265
Yc	Bias	X_0					
			N=100	0.494439	0.591209	0.583515	0.802809
			N=300	0.261106	0.483215	0.498769	0.9596
			N=500	0.173155	0.512221	0.533149	1.144263
		X_1					
			N=100	0.454354	0.534451	0.539381	0.765181
			N=300	0.257676	0.47056	0.493571	0.95376
			N=500	0.174887	0.507023	0.533092	1.142917
	MAE	X_0					
			N=100	0.518549	0.62927	0.595459	0.806389
			N=300	0.26409	0.483215	0.498769	0.9596
			N=500	0.176396	0.512221	0.533149	1.144263
		X_1					
			N=100	0.466298	0.558012	0.542142	0.765513
			N=300	0.258652	0.47056	0.493571	0.95376
			N=500	0.176375	0.507023	0.533092	1.142917

Note: This table reports for the g_1 model the bias and the mean absolute error (MAE) of the inverse probability weighting estimator, the nearest neighbour matching estimator with replacement, the subclassification estimator and the narive ols estimator, compared to the true linking effects, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1, the number of subclasses for the subclassification estimator is 8. All the estimates are for the average treatment effect for the treated.

female student obtaining a bachelor's degree. It also finds that this negative effect could be partly explained by a decrease in the girls' confidence and aspirations, as well as their grades in math and science. But do high-achieving male *friends* also have this negative impact on girls? At the end of the day, interactions and social influence among close friends could be very different from those among students who simply attend the same school and might not have close and friendly interactions.

The results indicate that the effect of friendship could indeed be very different from the effect of cohort peers. Noticeably, an additional male high-achieving friend increases the probability of a female student obtaining a bachelor's degree by 3 percentage points. Heterogeneity analysis reveals that this positive effect of male high flyer friendship is mainly driven by female students with below median ability as measured by their PVT score. Evidence also suggests that the effect mainly comes from a confidence boost instead of a tangible influence on their GPA.

1.6.1 Data

The data used by this analysis is from the National Longitudinal Study of Adolescent to Adult Health (Add Health).[22] It is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-95 school year (Wave I). In total, 172 schools were sampled. The Wave I data consists of an in-school questionnaire for all students in the sampled schools, followed by an in-home interview conducted for only a sample of these students. Out of the 172 schools, 16 are the so-called saturated schools, where all students who answered the in-school questionnaire were selected for the in-home interview. The sample of students who answered the Wave I in-home interview was interviewed again during the 1995-1996 school year (Wave II), another time in 2001-2002

[22] This research uses data from Add Health, funded by grant P01 HD31921 (Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health is currently directed by Robert A. Hummer and funded by the National Institute on Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Aiello and Hummer) at the University of North Carolina at Chapel Hill. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill. The Add Health Parent Study/Parents (2015-2017) data collection was funded by a grant from the National Institute on Aging (RO1AG042794) to Duke University, V. Joseph Hotz (PI) and the Carolina Population Center at the University of North Carolina at Chapel Hill, Kathleen Mullan Harris (PI).

(Wave III), again in 2007-2008 (Wave IV), and most recently in 2016-2018 (Wave V).

For my empirical analysis, information on educational attainment is taken from the Wave IV data, when respondents were between 26-32 years old. They were asked to give their highest level of education achieved by the time of the interview. As in Cools et al. (2022), I define a dummy variable for bachelor's degree attainment equal to 1 if the respondent had obtained a four-year college degree or more and 0 otherwise. Some other secondary outcome variables are also used in this analysis. These include Wave II information on students' grades, willingness and confidence in going to college, and self-assessment of their intelligence compared to their peers.

Friendship information comes from the Wave I in-home interviews. During the interview, students were asked to nominate at most five of their female friends and five of their male friends from their school's and the sister school's roaster. Students' pre-treatment information comes from Wave I. This includes background information on the students and their parents. On the students' side, I use data on their gender, age, race, whether they were born in the US, and their PVT score.[23] On the parents' side, I use data on the residential mother and father's education level, whether they worked for pay for more than 10 hours per week at the time interview was conducted, whether they were born in the U.S., and the annual family income. The exact definitions of all variables are detailed in Table 2.12, along with the definitions used in Cools et al. (2022). In order to compare the results with the CFP paper, I further restrict the data following their procedure, keeping only those in grades 7-12 during Wave I, except those with less than 20 students.

1.6.2 Estimation of propensity scores and the linking effects

The first step of estimating the linking effect is to estimate the propensity scores from the adjacency matrix. When students were interviewed for the AddHealth data, they were only allowed to nominate their friends within the same school. This means that for each school s, we have a network represented by an adjacency matrix $\mathbf{D}_s$ with N_s nodes. The N_s nodes include every student on the school roaster. In each school, a sample of n_s students who

[23] A Picture Vocabulary Test (PVT) was administered by the interviewer during the Wave I in-home interview. PVT measures an individual's verbal ability.

were also in the school roaster was selected for the in-home interview and therefore asked to nominate their friends from the N_S students listed on the roaster. For each i of the sampled students and each student j on the school roaster, $D_{s,i}^{j}$ is recorded as 1 if i nominates j as their friend and is recorded as 0 if j is nominated by i as a friend. I remove any column j of the adjacency matrix $\mathbf{D}_s$ if j was not nominated by any sampled student i. For the $N_s - n_s$ students who were not sampled for the in-home interview, their adjacency matrix entries are missing, which prevents us from estimating their propensity scores of linking. This is not a problem for our analysis for two reasons. First, since they were not selected for the in-home interviews, their information on outcome variables would also be missing, meaning they wouldn't have been included in the analysis anyway. Second, the propensity scores of linking of the sampled students can still be estimated through factor models, even though they can no longer be estimated by graphon estimators. The factor model I use for this empirical analysis is the same as the one specified in (1.5).

After the propensity scores of linking are estimated for all the sampled students in each school, we are ready to estimate the linking effects of interest. In this empirical analysis, I use the augmented inverse probability weighting estimator (AIPW). Specifically, I run the propensity score re-weighted pairwise regression specified in (1.15) for the characterization of the link receivers and the link senders of interest, for example, female link receivers and male high-achieving senders.

$$Y_{s,i} = \beta_{s,0} + \beta_{s,1} D_{s,i}^{s,j} + \rho_s \mathbf{X}_{s,i} + \epsilon_{s,i}^{s,j} \tag{1.15}$$

where $Y_{s,i}$ and $\mathbf{X}_{s,i}$ are respectively the outcome and covariates of student i in school s. $D_{s,i}^{j}$ is a dummy variable that equals to 1 if student i nominates j as their friend where both i and j are from school s. Each pairwise observation is weighted according to its propensity of linking and its linking status. Here I estimate the treatment effect of treated (ATT), which means the weights are generated according to (1.16).

$$w_{s,i}^{s,j} = \begin{cases} 1 & \text{if } D_{s,i}^{s,j} = 1 \\ \frac{p_{s,i}^{s,j}}{1-p_{s,i}^{s,j}} & \text{if } D_{s,i}^{s,j} = 0 \end{cases} \tag{1.16}$$

where $w_{s,i}^{s,j}$ is the pairwise weight and $p_{s,i}^{s,j}$ is the estimated propensity of linking from j to i. Note that using the propensity score weighted regressions to estimate the linking effects does not mean we assume the true effect is linear and additive with respect to the covariates. Just like in traditional causal inference, regressions are only used as a way of estimation. Finally, I get the overall linking effect across schools β_1 by weighting the school linking effect by the number of observed links in that school.[24]

Wang and Blei (2019) suggested using a test statistics to assess the adequacy of propensity score estimation. This test statistics is based on the idea that well-estimated propensity scores should have good predictive power for the validation data. Following their procedure, our estimated propensity scores for each school network pass the test and perform well.

Traditionally, the adequacy of the estimated propensity scores is assessed by balance tests, where the difference in pre-treatment variables between the treated group and the control group is calculated using the propensity score-adjusted sample. This method is not directly applicable to our context. First of all, since each link sender is associated with a unique treatment, ideally, we would compare for each link sender the pre-treatment characteristics of the students who were treated by this link sender and the students who were not treated by this link sender. Because in our networks of finite size, each link sender only has a few treated students, this comparison suffers from finite sample bias. We could, however, average the differences in pre-treatment variables between treated and control students over all link senders. The second issue is that our propensity scores are based on the unobserved sufficient confounders that do not correspond directly to any observed variables. Since the propensity scores were not estimated using any observed pre-treatment variables, there is no guarantee that any selected pre-treatment variable will be balanced across the treated and the control groups. Nonetheless, we could still evaluate the balance for some variables we believe are part of the confounders.

Balance tests could be conducted by running a pairwise regression similar to (1.15), except that the covariates will become the outcome variables. Table 1.3 shows the result of a balance test for some pre-treatment variables. According to Currarini et al. (2009) race is

[24] I weight it by the number of observed links because the estimand is ATT. If we are interested in ATE, the weight should be the number of all potential links.

a strong predictor of friendship formation, and (Carrell et al., 2013) suggests the same for ability. Table 1.3 shows that the balance for the ability variables (column 3 and column 4) and the race variable of being black are improved.

Table 1.3: Balance test

	Pre-treatment variable:											
	Male	US born	PVT	PVT +	M C+	F C +	Income	Age	M nHH	F nHH	Black	Hispanic
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original	0.002	−0.002***	0.688***	0.025***	0.023***	0.022***	0.037***	−0.993***	−0.004***	−0.006***	−0.014***	−0.001
	(0.002)	(0.001)	(0.041)	(0.002)	(0.001)	(0.001)	(0.003)	(0.043)	(0.001)	(0.001)	(0.001)	(0.001)
AIPW	0.001	−0.001	0.449***	0.016***	0.017***	0.012***	0.036***	−0.637***	−0.002**	−0.006***	−0.008***	0.001
	(0.002)	(0.001)	(0.044)	(0.002)	(0.001)	(0.002)	(0.003)	(0.048)	(0.001)	(0.001)	(0.001)	(0.001)

Note: This table reports the average differences between the treated and the control across all link senders. The first row is the balance test for the origianl sample. The second row is the balance test for the sample re-weighted by the propensity scores according to inverse probability weighting method. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. The pre-treatment variables from column 1 to column 12 are: whether the ego is male, born in US, their PVT score, whether their PVT score is above the population median, whether their mother has colllege degree or above, whether their father has colllege degree or above, their annual family income (log), their age in months, whether their mother is not in the household, whether their father is not in the household, whether the respondent is black, and whether the respondent is hispanic. *p<0.1; **p<0.05; ***p<0.01

1.6.3 Results

Table 1.4 reports the estimated effects of friendships from different types of link sender on bachelor's degree attainment (column 1) and some intermediate outcomes recorded during Wave II interviews. Each row corresponds to a characterization of the friendship based on the character of the receiver and the sender. The receiver characteristic is shown before the underbar "_", and the sender characteristic is shown after. "F" and "M" refer to the gender female and male, respectively. "H" and "L" refer to whether the individual is a high achiever or non-high achiever (low achiever), respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high achiever link senders.

Table 1.4 shows a nearly 3 p.p increase in female students' likelihood of obtaining a bachelor's degree by having an additional male high-achieving friend. For male students, an extra male high-achieving friend means an increase of 4.6 p.p in the probability of graduating from college. Looking at the last three columns of the table, it appearss that the positive effect of a male high-achieving friend on both female and male students could be attributed to an increase in their confidence. In particular, an additional male high-achieving friend

Table 1.4: Effect of friendship on bachelor's degree attainment and confidence

	Dependent variable:			
	Bachelor's Degree (p.p)	Want (p.p)	Will (p.p)	Intelligence (p.p)
	(1)	(2)	(3)	(4)
F_FL	0.354*	−0.171	−0.819***	−0.389
	(0.191)	(0.241)	(0.227)	(0.242)
F_ML	0.336	−0.361	−0.797**	−0.602
	(0.313)	(0.381)	(0.373)	(0.439)
F_FH	1.877	2.377*	0.737	1.364
	(1.262)	(1.245)	(1.062)	(1.345)
F_MH	2.981***	1.602	2.370***	3.748***
	(0.978)	(1.116)	(0.858)	(1.324)
M_FL	−0.041	0.144	−0.026	−0.623*
	(0.279)	(0.269)	(0.270)	(0.336)
M_ML	−0.068	0.058	−0.553**	−0.816***
	(0.227)	(0.204)	(0.247)	(0.253)
M_FH	2.801	0.930	−1.919	−1.652
	(1.906)	(1.818)	(1.764)	(1.773)
M_MH	4.645***	1.361	0.821	4.539***
	(1.526)	(1.544)	(1.314)	(1.153)

Note: This table reports the estimated effects of high school friendship on students' bachelor's degree attainment (column 1), and their intermediate outcomes (column 2-4). The dependent variable in Column (2) is a dummy variable recording whether the student reported a scale 5 (1 is the lowest and 5 is the highest) on the the extent of how much they want to go to college (Wave II). The dependent variable in Column (3) is a dummy variable recording whether the student reported a scale 5 (1 is the lowest and 5 is the highest) on the likelihood that they will go to college (Wave II). The dependent variable in Column (4) is a dummy variable recording whether the student reported a scale 5 or 6 (1 is the lowest and 6 is the highest) on their intelligence compared to other people of their age (Wave II). The estimands are all ATT. Each row corresponds to a characterisation of the friendship, based on the character of the receiver and the sender. Receiver characteristics is shown before the underbar _, and sender characteristics is shown after. "F" and "M" are used to refer to the gender female and male respectively. "H" and "L" are used to refer to whether the individual is a high flyer or non-high flyer (low flyer) respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high flyer link senders. The regressions reported in all columns include cohort dummies, whether the student was born in the US, their PVT score, whether their PVT score is above the population median PVT score, whether their mother's and father's highest degree is high school, some college, college, or post college, whether their mother's and father's highest education level is missing, the student's log family income, whether family is missing, the age of the student during Wave I, whether the student's mother and father were in the household, dummies for whether the student is black, hispanic, white, asian and indian. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. *p<0.1; **p<0.05; ***p<0.01

Table 1.5: Heterogeneous effects of friendship on bachelor's degree attainment and intelligence

	Dependent variable:			
	Bachelor's degree		Intelligence	
	PVT Median -	PVT Median +	PVT Median -	PVT Median +
	(1)	(2)	(3)	(4)
F_FL	0.510	−0.347	-1.121^{***}	−0.026
	(0.362)	(0.424)	(0.429)	(0.451)
F_ML	0.818	-1.038^{*}	−0.139	-1.823^{**}
	(0.634)	(0.540)	(0.836)	(0.903)
F_FH	5.912^{***}	−0.782	6.552^{**}	−1.115
	(2.294)	(2.656)	(3.054)	(2.531)
F_MH	3.649^{*}	0.492	10.283^{***}	−2.967
	(2.084)	(1.952)	(2.585)	(2.620)
M_FL	−0.780	0.329	-1.916^{**}	1.550^{***}
	(0.649)	(0.581)	(0.802)	(0.568)
M_ML	0.670	−0.260	-2.051^{***}	0.736
	(0.451)	(0.423)	(0.474)	(0.462)
M_FH	0.305	3.377	−5.573	−0.074
	(5.056)	(2.160)	(3.607)	(2.236)
M_MH	4.231	8.267^{***}	−2.164	11.954^{***}
	(3.005)	(2.511)	(2.911)	(2.225)

Note: This table reports the estimated heterogeneous effects of high school friendship on students' bachelor's degree attainment and self-assessed intelligence. Column (1) and (3) reports results for ego whose PVT score is below population median PVT score. Column (2) and (4) reports results for ego whose PVT score is above population median PVT score. The estimands are all ATT. Each row corresponds to a characterisation of the friendship, based on the character of the receiver and the sender. Receiver characteristics is shown before the underbar _, and sender characteristics is shown after. "F" and "M" are used to refer to the gender female and male respectively. "H" and "L" are used to refer to whether the individual is a high flyer or non-high flyer (low flyer) respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high flyer link senders. The regressions reported in all columns include cohort dummies, whether the student was born in the US, their PVT score, whether their PVT score is above the population median PVT score, whether their mother's and father's highest degree is high school, some college, college, or post college, whether their mother's and father's highest education level is missing, the student's log family income, whether family is missing, the age of the student during Wave I, whether the student's mother and father were in the household, dummies for whether the student is black, hispanic, white, asian and indian. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. *p<0.1; **p<0.05; ***p<0.01

increases the probability that a female student reports having a high likelihood of going to college and being more intelligent than their same-age peers during the Wave II interview, one year after friendship information was recorded. As for male students, their self-assessment of being more intelligent than their same-age peers is also increased.

Female egos are also slightly more likely to graduate from college when they have an additional female friend who is not a high achiever. However, this effect disappears if we separately look at the effect on low-ability and high-ability female students. As shown in Table 1.5, estimates for both ability groups of female students are not significantly different from 0. Moreover, the positive effect of male high achiever friends seems to only exist for low-ability female students and high-ability male students, with an increase in the probability of going to college by about 3.6 p.p and 8.3 p.p, respectively. These positive effects are also found in their self-assessment of being more intelligent than their peers. However, is this positive impact on self-assessment of intelligence due to a confidence boost or an increase in academic performance? To answer this question, I look at the effect of friendship on egos' grades during Wave II. Table 1.6 and Table 1.7 show that across all four academic subjects, none of the grades of low-ability female students were increased by having an additional male high-achieving friend. As for male high-ability students, their English grade was improved by 0.196 points on average (lowest 1, highest 5) by having an additional male high-achieving friend, but none of the grades of the other subjects were improved.

1.7 Conclusion

By looking at the problem of peer influence through the causality lense and thereby bridging the multiple causal inference literature and the network analysis literature, this paper shows that the network endogeneity problem tormenting the study of the linking effect can be solved under a set of assumptions that are easy to satisfy for many common networks. However, this is not to say that the solution can be used for any network. In some situations, these assumptions could fail, and alternative solutions must be used. For example, the assumption of doubly individualistic assignment mechanism fails in the case of the marriage network or the roommate network, where some links are direct causes of other links. In these cases, we

Table 1.6: Heterogeneous effects of friendship on English and Math grades

	Dependent variable:			
	English grade		Math grade	
	PVT Median -	PVT Median +	PVT Median -	PVT Median +
	(1)	(2)	(3)	(4)
F_FL	-0.004	-0.002	0.002	-0.007
	(0.007)	(0.008)	(0.007)	(0.008)
F_ML	-0.035^{**}	0.037^{**}	0.002	-0.020
	(0.016)	(0.017)	(0.014)	(0.019)
F_FH	0.050	0.014	0.232^{***}	0.022
	(0.045)	(0.028)	(0.053)	(0.025)
F_MH	-0.025	0.050	0.039	-0.012
	(0.036)	(0.032)	(0.037)	(0.041)
M_FL	0.021	-0.006	-0.044^{**}	-0.008
	(0.015)	(0.010)	(0.019)	(0.008)
M_ML	0.009	0.00001	-0.025^{***}	0.004
	(0.008)	(0.006)	(0.009)	(0.008)
M_FH	0.159^{**}	-0.00003	0.080^{*}	0.061
	(0.079)	(0.067)	(0.043)	(0.045)
M_MH	-0.039	0.196^{***}	0.060	0.002
	(0.055)	(0.053)	(0.075)	(0.043)

Note: This table reports the estimated heterogeneous effects of high school friendship on students English and Math grades (Wave II). Column (1) and (3) reports results for ego whose PVT score is below population median PVT score. Column (2) and (4) reports results for ego whose PVT score is above population median PVT score. The estimands are all ATT. Each row corresponds to a characterisation of the friendship, based on the character of the receiver and the sender. Receiver characteristics is shown before the underbar _, and sender characteristics is shown after. "F" and "M" are used to refer to the gender female and male respectively. "H" and "L" are used to refer to whether the individual is a high flyer or non-high flyer (low flyer) respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high flyer link senders. The regressions reported in all columns include cohort dummies, whether the student was born in the US, their PVT score, whether their PVT score is above the population median PVT score, whether their mother's and father's highest degree is high school, some college, college, or post college, whether their mother's and father's highest education level is missing, the student's log family income, whether family is missing, the age of the student during Wave I, whether the student's mother and father were in the household, dummies for whether the student is black, hispanic, white, asian and indian. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. *p<0.1; **p<0.05; ***p<0.01

Table 1.7: Heterogeneous effects of friendship on History and Science grades

	Dependent variable:			
	History grade		Science grade	
	PVT Median -	PVT Median +	PVT Median -	PVT Median +
	(1)	(2)	(3)	(4)
F_FL	-0.008	-0.013	-0.025^{***}	-0.011
	(0.006)	(0.008)	(0.007)	(0.013)
F_ML	0.046^{***}	0.014	-0.014	0.022
	(0.016)	(0.013)	(0.017)	(0.018)
F_FH	0.081	-0.025	0.142^{**}	-0.030
	(0.058)	(0.032)	(0.062)	(0.026)
F_MH	-0.028	0.104^{***}	0.007	0.010
	(0.041)	(0.037)	(0.046)	(0.026)
M_FL	-0.035	0.030^{***}	-0.042^{**}	0.007
	(0.024)	(0.008)	(0.016)	(0.008)
M_ML	-0.025^{***}	-0.014^{*}	0.010	-0.002
	(0.008)	(0.008)	(0.013)	(0.006)
M_FH	0.091	-0.129^{**}	0.086^{**}	-0.130^{**}
	(0.059)	(0.051)	(0.042)	(0.052)
M_MH	-0.093	0.037	0.004	-0.033
	(0.061)	(0.037)	(0.078)	(0.036)

Note: This table reports the estimated heterogeneous effects of high school friendship on students History and Science grades (Wave II). Column (1) and (3) reports results for ego whose PVT score is below population median PVT score. Column (2) and (4) reports results for ego whose PVT score is above population median PVT score. The estimands are all ATT. Each row corresponds to a characterisation of the friendship, based on the character of the receiver and the sender. Receiver characteristics is shown before the underbar _, and sender characteristics is shown after. "F" and "M" are used to refer to the gender female and male respectively. "H" and "L" are used to refer to whether the individual is a high flyer or non-high flyer (low flyer) respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high flyer link senders. The regressions reported in all columns include cohort dummies, whether the student was born in the US, their PVT score, whether their PVT score is above the population median PVT score, whether their mother's and father's highest degree is high school, some college, college, or post college, whether their mother's and father's highest education level is missing, the student's log family income, whether family is missing, the age of the student during Wave I, whether the student's mother and father were in the household, dummies for whether the student is black, hispanic, white, asian and indian. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. *p<0.1; **p<0.05; ***p<0.01

could resort to explicit network formation modelling.

Moreover, the definition of the linking effect makes it clear that nodal characteristics are not the treatment but the variables that could be used to define effect heterogeneity. This means we could adapt the machine learning techniques developed to study heterogeneous effects to the case of linking effects. Finally, we could extend this paper by relaxing the L-SUTVA assumption and defining more sophisticated estimands.

Chapter 2

Extensions, theoretical proofs, and additional results on the linking effect

2.1 Extensions

2.1.1 Treatment defined over all links

Suppose we are interested in the comparison between two configurations, such as c^1 and c^2. A configuration is a rule C that the treatment vector has to satisfy. For example, c^1 could be 2 female and 1 male and c^2 be 1 female and 2 male. Assume L-SUTVA holds, for any node i let us denote the set of treatments that satisfies configuration c as $\mathcal{D}_i^c = \{\mathbf{D}_i | C(\mathbf{D}_i) = c\}$, where $\mathbf{D}_i = (D_{i1}, ...D_{ij}, ..., D_{iN})$. For any $d^{c_1} \in \mathcal{D}^{c_1}$ and $d^{c_2} \in \mathcal{D}^{c_2}$. we can define an estimand $m_i^{c_1,c_2}$:

$$m_i^{d^{c_1},d^{c_2}} = Y_i(d^{c_1}) - Y_i(d^{c_2})$$

For any configuration c, use $|D^c|$ to denote the number of elements in the set $\mathcal{D}^c$ and the expectation $\mathbb{E}_c$ as the expectation over the set $\mathcal{D}^c$ with uniform probability. Average over the set of treatments that satisfy the configuration rules, we can define the treatment effect of configuration c^1 v.s. c^2 on node i as:

$$\begin{aligned} m_i^{c_1,c_2} &= \mathbb{E}_{c_1}[Y_i(d^{c_1})] - \mathbb{E}_{c_2}[Y_i(d^{c_2})] \\ &:= \frac{1}{|\mathcal{D}^{c_1}|}\sum_{d^{c_1}\in\mathcal{D}^{c_1}} Y_i(d^{c_1}) - \frac{1}{|\mathcal{D}^{c_2}|}\sum_{d^{c_2}\in\mathcal{D}^{c_2}} Y_i(d^{c_2}) \end{aligned}$$

Finally, by averaging over the set of egos, we can easily define the average treatment effect of configuration c^1 v.s. c^2 as:

$$\begin{aligned} m^{c_1,c_2} &= \mathbb{E}_i\left[\mathbb{E}_{c_1}[Y_i(d^{c_1})] - \mathbb{E}_{c_2}[Y_i(d^{c_2})]\right] \\ &:= \frac{1}{N}\sum_{i=1,...,N}\big(\frac{1}{|\mathcal{D}^{c_1}|}\sum_{d^{c_1}\in\mathcal{D}^{c_1}} Y_i(d^{c_1}) - \frac{1}{|\mathcal{D}^{c_2}|}\sum_{d^{c_2}\in\mathcal{D}^{c_2}} Y_i(d^{c_2})\big) \end{aligned}$$

Lemma 1 (Unconfoundedness when treatment is defined over all links)**.**

$$Pr(D_i = d^c|Y_i^{pot}, \mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N) = Pr(D_i = d^c|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N)$$

and

$$Pr(D_i = d^c|Y_i^{pot}, e(\mathbf{U}_i, \mathbf{V}_1), ..., e(\mathbf{U}_i, \mathbf{V}_N)) = Pr(D_i = d^c|e(\mathbf{U}_i, \mathbf{V}_1), ..., e(\mathbf{U}_i, \mathbf{V}_N))$$

Proof. The first half of the proof is identical to that of Proposition 1. For the last part, instead we have

$$\begin{aligned} &Pr(D_i = d^c|\mathbf{U}_1, ..., \mathbf{U}_N, \mathbf{V}_1, ..., \mathbf{V}_N, \mathbf{Y}_i^{pot}) \\ =&Pr(D_i = d^c|\mathbf{U}_1, ..., \mathbf{U}_N, \mathbf{V}_1, ..., \mathbf{V}_N) \\ =&Pr(D_i = d^c|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N) \end{aligned}$$

The first equation holds because we have ruled out any confounders that affect any of the links, which means there are no confounders to affect all of i's links. The second equation comes from equation (2). □

Assumption 1 (Overlap for all links)**.** $0 < Pr(\mathbf{D}_i = d^{c_1}|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N) < 1$

Proposition 1. Under assumption 1,2,3 and 1, m^{c_1,c_2} is identified:

$$m^{c_1,c_2} = \mathbb{E}\left[\mathbb{E}_i\left[\mathbb{E}_{d^{c_1}}[Y_i(\mathbf{D}_i = d^{c_1})|e(\mathbf{U}_i,\mathbf{V}_1),...,e(\mathbf{U}_i,\mathbf{V}_N),\mathbf{D}_i = d^{c_1}]\right]\right]$$
$$- \mathbb{E}\left[\mathbb{E}_i\left[\mathbb{E}_{d^{c_2}}[Y_i(\mathbf{D}_i = d^{c_2})|e(\mathbf{U}_i,\mathbf{V}_1),...,e(\mathbf{U}_i,\mathbf{V}_N),\mathbf{D}_i = d^{c_2}]\right]\right]$$

Proposition (1) is proved in Section 2.4.5.

Notice here in order to estimate this estimand, we need to condition not just on the single pairwise propensity score $e(\mathbf{U}_i,\mathbf{V}_j)$, but rather on the vector of propensity scores $e(\mathbf{U}_i,\mathbf{V}_1),...,e(\mathbf{U}_i,\mathbf{V}_N)$. To gain some intuition, first recall that in the main analysis, the hypothetical intervention was on a single pair, and the estimand is the average of potential outcomes under repeated hypothetical interventions over different pairs each time. Here the hypothetical intervention, however, is on all the relationships of node i, thus the need to condition on the propensity scores of all relationships being formed.

Finally, note that as N goes to infinity, the overlap condition will fail to hold. To see why, write the generalised propensity score as the product of individual pairwise propensity score:

$$\begin{aligned}&Pr(\mathbf{D}_i = d^{c_1}|\mathbf{U}_i,\mathbf{V}_1,...,\mathbf{V}_N)\\ &= \prod_{j=1}^{N}\left(Pr(D_i^j = 1|\mathbf{U}_i,\mathbf{V}_j)\right)^{d_i^{c_1}}\left(1 - Pr(D_i^j = 1|\mathbf{U}_i,\mathbf{V}_j)\right)^{1-d_i^{c_1}}\end{aligned}$$

Since $0 < Pr(D_i^j = 1|\mathbf{U}_i,\mathbf{V}_j) < 1$, this product goes to 0 as N goes to infinity, causing the overlap condition to fail.

2.1.2 Alternative estimands

In the main analysis the treatment effect of sender-j relationship on receiver i's potential outcome is defined as the following contrast of potential outcomes:

$$\tau_i^j = Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}) - Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})$$

where all the non-sender-j relationships of receiver i are fixed at their observed level. This is only one of the many ways we can define the pair level estimand. In fact, for any i, j and $\mathbf{d}_i^{-j}$ we could define

$$\tilde{\tau}_i^j(\mathbf{d}_i^{-j}) = Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \mathbf{d}_i^{-j}) - Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \mathbf{d}_i^{-j}) \tag{2.1}$$

In this case, we could define an average linking effect for link receivers with characteristic $R_i = r$ and link senders with characteristic $A^j = a$ by averaging the pair level treatment effects over the probability distribution of the linking status of i's other (than j) relationships:

$$\tilde{\tau}_r^a = \mathbb{E}_{(i,j):R_i=r,A^j=a} \sum_{\mathbf{d}_i^{-j} \in \mathfrak{D}^j} \tilde{\tau}_i^j(\mathbf{d}_i^{-j}) Pr(\mathbf{D}_i^{-j} = \mathbf{d}_i^{-j}) \tag{2.2}$$

With a slight abuse of notation, I use $\mathbb{E}_{(i,j):R_i=r,A^j=a}[\cdot]$ to represent $\mathbb{E}_{(i,j)}[\cdot|R_i = r, A^j = a]$. $\mathfrak{D}^j = \cup_i \mathbf{d}_i^{-j}$.[1] Next I will prove that $\tilde{\tau}_r^a$ is identified.

[1] This estimand is similar to the kind of estimands usually defined in the literature of treatment interference, e.g. Forastiere et al. (2021). The difference is that in the treatment inference literature the "direct" or main estimand is defined by averaging over the treatments of interfering units, while here we average over the non-focal links of the same receiver.

Proof.

$$
\begin{aligned}
&\mathbb{E}_{(i,j):R_i=r,A^j=a}\big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j} Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})\big]\\
&=\mathbb{E}\big[\mathbb{E}_{(i,j):R_i=r,A^j=a}\big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j} Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N}\big]\big]\\
&=\mathbb{E}_{(i,j):R_i=r,A^j=a}\Big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j}\mathbb{E}\big[Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N}\big]\Big]\\
&=\mathbb{E}_{(i,j):R_i=r,A^j=a}\Big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j}\mathbb{E}\big[Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N}\big]\\
&\qquad\times Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j}|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N})\Big]\\
&=\mathbb{E}_{(i,j):R_i=r,A^j=a}\Big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j}\mathbb{E}\big[Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N},D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j}\big]\\
&\qquad\times Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j}|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N})\Big]\\
&=\mathbb{E}_{(i,j):R_i=r,A^j=a}\Big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j}\mathbb{E}\big[Y_i^{obs}|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N},D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j}\big]\\
&\qquad\times Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j}|\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N})\Big]
\end{aligned}
$$

The first equation comes from the law of iterated expectations, the second equation is due to linearity of expectations, the third equation is due to the independence between potential outcome and linking probability conditional on $(\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N})$ (same d-separation argument as before), the fourth equation comes from the unconfoundedness of $Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})$ conditional on $(\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N})$ (1), and the fifth equation holds because when $D_i^j=1$ and $\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j}$, $Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})=Y_i^{obs}$. This means if $(\mathbf{U}_i,\mathbf{V_1},...,\mathbf{V_N})$ were observed, or equivalently if $\{e(\mathbf{U}_i,\mathbf{V}_1),...,e(\mathbf{U}_i,\mathbf{V}_N)\}$ were observed,

$$
\mathbb{E}_{(i,j):R_i=r,A^j=a}\big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j} Y_i(D_i^j=1,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})\big]
$$

is identified, and can be estimated with observed data. The same proof holds for

$$\mathbb{E}_{(i,j):R_i=r,A^j=a}\big[\sum_{\mathbf{d}_i^{-j}\in\mathfrak{D}^j} Y_i(D_i^j=0,\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})Pr(\mathbf{D}_i^{-j}=\mathbf{d}_i^{-j})\big].$$

This means estimand $\tilde{\tau}_r^a$ is identified. □

2.1.3 Other types of linking effect to explore in the future

Indirect linking effect

As shown in Figure 2.1, we can define an indirect effect that contrasts i's potential outcome when some link sender j is linked to one of i's existing direct peer and its potential outcome when j is not linked to one of i's existing direct peer, while keeping i's existing peers fixed at the realised value. This requires the relaxation of L-SUTVA and is similar to the study of spillover effects in traditional setting (Forastiere et al., 2021).

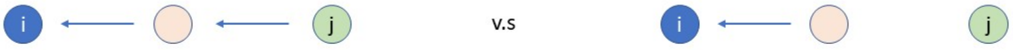


Figure 2.1: Indirect linking effect

Triangle reinforced linking effect

The triangle reinforced linking effect contrasts i's potential outcome when its direct peer j also sends a link to one of i's other existing direct peer and its potential outcome when j is not linked to one of i's existing direct peer, while keeping i's existing peers fixed at the realised value. This could be used to study whether direct linking effect could be reinforced by an additional indirect link. If the underlying mechanism for the peer effect is information flow, then triangle reinforced effect shouldn't exist. It also requires the relaxation of L-SUTVA to allow for interference.

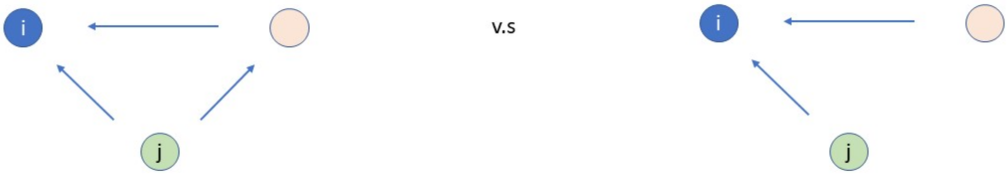


Figure 2.2: Triangle reinforced linking effect

2.1.4 Small networks

When networks are small, the estimation of propensity scores might be difficult, even if we have a large number of such small networks. This is because the estimation of propensity score is based on each single network. If the individual network is small, there is very little information for the inference of sufficient confounders and their propensity scores.

In this case, we could still make causal discovery based on additional assumptions. The idea is to assume that the effective treatment is some characteristic of the node, instead of the identity of the node. Let $Y_i^g(\cdot)$ denote the potential outcome of link receiver i in network g, this assumption is formalised as Assumption 2.

Assumption 2. For some function $l : \{0,1\}^N \to \mathbb{R}^M$

$$\begin{aligned} Y_i^g(D_{i1}, D_{i2}, ..., D_{in}) &= Y_i^g(l(D_{i1}, D_{i2}, ..., D_{iN})) \\ &= Y_i^g(l_1, ..., l_M) \end{aligned}$$

Function $l(\cdot)$ defines the effective treatment. For example if $l(\cdot) = \sum_{j=1,...,n} D_{ij}X_j$ where X is a dummy variable, Assumption 2 means i's links affect i's potential outcome only through the total number of links with characteristics X. Similarly, if $l(\cdot) = \frac{\sum_{j=1,...,n} D_{ij}X_j}{\sum_{j=1,...,n} D_{ij}}$, Assumption 2 means i's links affect i's potential outcome only through the share of i's links with characteristics X. Note that here we do not assume that the potential outcome is a linear function of $l(\cdot)$ as in the linear-in-means and linear-in-sum models. In both examples, we have $M = 1$, but this is not necessary. For example, $l(D_{i1}, D_{i2}, ..., D_{iN}) = (\sum_{j=1,...,n} D_{ij}X_j^1, \sum_{j=1,...,n} D_{ij}X_j^2)$ means the effective treatment is the total number of links with characteristics X^1 and the total number of links with characteristics X^2.

Next I show that under Assumption 2, causal identification and estimation of linking effect could be achieved by inferring sufficient confounders that render the distribution of effective treatment conditionally independent, as long as $M \geq 2$.

Definition 2.1.1. Let N^g be the number of nodes in network g, and $N = \sum_{g=1} N^g$. $o_1, ..., o_N$ and $q_1, ..., q_M$ are two vectors of random variables that satisfy the following condition:

$$Pr(l_{i1}, ..., l_{iM} | o_i, q_1, ..., q_M) = \prod_{m=1}^{M} Pr(l_{im} | o_i, q_m) \quad i = 1, ..., N$$

Effectively $l_{i1}, ..., l_{iM}$ is the multiple treatment vector of link receiver i, and since M is a fixed number, we are in the standard case studied in Wang and Blei (2019). Therefore $o_1, ..., o_N$ and $q_1, ..., q_M$ are sufficient confounders in the sense that after conditioning on them, treatment $(l_{i1}, ..., l_{iM})$ is independent of the potential outcome $Y_i(l_{i1}, ..., l_{iM})$.

Assumption 2 makes it possible to identify and estimate linking effects when networks are small. The intuition is that since nodes from different networks all share the same set of possible treatment $l_1, ..., l_M$. we could pool the link receivers across networks together to infer the sufficient confounders and their propensity scores. Note that in this case the estimators from the statistical network analysis literature, such as the neighbourhood smoothing estimator, won't work. But the factor models can still be used to estimate the propensity scores.

Finally, note that if this assumption doesn't hold, we will get biased causal estimates. This is because the sufficient confounders are defined as variables that make the supposedly effective treatments conditionally independent. If treatments are in fact at a more disaggregated level, these sufficient confounders are no long 'sufficient'.

2.2 Alternative 2nd-step treatment effect estimatiors

As mentioned earlier, the inverse probability weighting estimator described in Section 1.4.2 is not the only 2nd-step estimator we could use to estimate the linking effect. Two of the popular ones in the causal inference literature are propensity score matching and propensity score subclassification. Here I will explain in detail how subclassification works and omit the details for matching. The case of propensity score matching is similar to subclassification. The only difference is that instead of dividing pairs into blocks based on similarity of propensity scores, we will find for each pair its M-nearest neighbour(s) in terms of their propensity scores. As in traditional propensity score matching, we could do both matching with replacement or without replacement. Next I will start with a simple example to illustrate the steps of subclassfication. Then I will provide formal justification of the subclassfication estimator.

2.2.1 An example of subclassification estimator

In this example there are 8 link receivers with characteristic $R = r$ (labelled 1 to 8) and 7 link senders with characteristic $A = a$ (labelled a to g). The treatment assignment for the link receivers is given in Table 2.1. Here I omit the link receivers with characteristic $R \neq r$ and the link senders with characteristic $A \neq a$ because they are not needed for the estimand τ_r^a . Note that the matrix in Table 2.1 is not an adjacency matrix itself, but the intersection of a selection of rows and columns from the underlying adjacency matrix.

Table 2.1: Example treatment assignment

	a	b	c	d	e	f	g
1	0	0	0	1	0	0	0
2	0	0	1	0	0	1	0
3	1	0	0	1	0	0	0
4	0	0	0	0	0	0	1
5	0	0	1	0	0	0	1
6	1	0	0	0	1	1	0
7	0	0	0	0	0	0	0
8	1	0	1	0	0	0	0

Table 2.2: Example propensity scores

	a	b	c	d	e	f	g
1	0	0.1	0	0.11	0.33	0	0
2	0	0	0.5	0	0	0.33	0.16
3	0.25	0	0	0.67	0	0.25	0
4	0.15	0.33	0	0.33	0.1	0	0.27
5	0	0	0.2	0.2	0	0	0.3
6	0.33	0	0	0	0.6	0.56	0
7	0	0.2	0.3	0	0	0	0.1
8	0.5	0	0.1	0	0.3	0	0

The matrix of propensity scores are shown in Table 2.2. These propensity scores are fictional and are only meant for illustration purpose, meaning they are not estimated. The observed outcomes of the link receivers are: $Y_1 = Y_2 = Y_4 = Y_7 = 1$, and $Y_3 = Y_5 = Y_6 = Y_8 = 0$.

The main idea of subclassification is that if we divide the estimated propensity scores into small intervals, or subclasses, units within the same subclass will have similar estimated propensity scores and therefore can be viewed as having the same potential outcome distributions due to unconfoundedness. Here a unit is a pairwise link. Then, within the same subclass, the average of the missing potential outcomes for the treated units can be unbiasedly estimated by the observed outcomes of the control (untreated) units. Going back to the data above, I divide the propensity scores into three subclasses: $b_1 = (0, 0.3)$, $b_2 = [0.3, 0.5)$, $b_3 = [0.5, 1)$, with the assumption that uncounfoundedness holds within each subclass. Note that some pairs have an estimated propensity score of 0, which violates the positivity condition, so I leave them out in the data analysis. This means the estimator is now unbiased for the average effect only for those pairs within positive treatment probability.[2]

This leads to the classification of link receiver link sender pairs as shown in Table 2.3. The estimator is then:

$$\frac{13}{13+8+5}(\frac{Y_3+Y_5+Y_8+Y_1+Y_4}{5}-\frac{Y_4+Y_1+Y_7+Y_5+Y_4+Y_3+Y_2+Y_7}{8})$$
$$+\frac{8}{13+8+5}(\frac{Y_6+Y_2+Y_5}{3}-\frac{Y_4+Y_7+Y_4+Y_1+Y_8}{5})+\frac{5}{13+8+5}(\frac{Y_8+Y_2+Y_3+Y_6}{4}-Y_7)$$

Notice that the outcome of the same link receiver could be used multiple times, such as Y_4. They can appear both in the treated group and the control group, across multiple subclasses of propensity scores. This is because the propensity score is based on the pair, while the outcome is based on the link receiver only, and the same link receiver could appear in multiple pairs.

Note that unconfoundedness given propensity scores doesn't imply pairs with the same propensity scores have the same u_i, u_j. Instead, it means that the treated units and control unis have the same distribution of u_i, u_j, and that treated units and and control units have the same distribution of potential outcomes.

2.2.2 Subclassification formally

For exposition purpose, let's focus on the estimand

$$\tau_a^r = \mathbb{E}\left[\mathbb{E}_{(i,j)}[Y_i^{obs}|e(\mathbf{U}_i,\mathbf{V}_j), D_i^j=1]\right] - \mathbb{E}\left[\mathbb{E}_{(i,j)}[Y_i^{obs}|e(\mathbf{U}_i,\mathbf{V}_j), D_i^j=0]\right]$$

[2] In fact, in subclassfication analysis, researchers often leave out units with too low or too high propensity scores, even if they are not exactly 0 or 1. This is because with finite sample, there are often too few treated units within the subclass of very low propensity scores and two few control units within the subclass of very high propensity scores.

Table 2.3: Subclassification of pairs

	(0,0.3)	[0.3,0.5)	[0.5,1)
D_i^j=1	(3,a)	(6,a)	(8,a)
	(5,c)	(2,f)	(2,c)
	(8,c)	(5,g)	(3,d)
	(1,d)		(6,e)
	(4,g)		
D_i^j=0	(4,a)	(4,b)	(7,e)
	(1,b)	(7,c)	
	(7,b)	(4,d)	
	(5,d)	(1,e)	
	(4,e)	(8,e)	
	(3,f)		
	(2,g)		
	(7,g)		
number of pairs	13	8	5

Suppose we decide to divide the propensity scores into B subclasses and assume the propensity scores within the same subclass are roughly constant, then τ_a^r can also be written as

$$\begin{aligned}\tau_a^r &= \frac{1}{B}\sum_{b=1}^{B}\frac{N_b}{N}\,\mathbb{E}_{(i,j)}[Y_i^{obs}|(i,j)\in b, D_i^j=1]\\ &-\frac{1}{B}\sum_{b=1}^{B}\frac{N_b}{N}\,\mathbb{E}_{(i,j)}[Y_i^{obs}|(i,j)\in b, D_i^j=0]\\ &=\frac{1}{B}\sum_{b=1}^{B}\tau_{a,b}^r\end{aligned}$$

where N_b is the number of (i,j) pairs in subclass $b\in B$, and $\tau_{a,b}^r=\frac{N_b}{N}(\mathbb{E}_{(i,j)}[Y_i^{obs}|(i,j)\in b, D_i^j=1]-\mathbb{E}_{(i,j)}[Y_i^{obs}|(i,j)\in b, D_i^j=0])$. To estimate $\tau_{a,b}^r$, we can simply compare the sample mean of the outcomes of the link receiver in treated pairs ($D_i^j=1$) and the sample mean of the outcomes of the link receiver in control pairs ($D_i^j=0$) belonging to the subclass b. Alternatively, we could use linear regressions to estimate $\tau_{a,b}^r$ for all $b\in B$, thanks to the equivalence between $\tau_{a,b}^r$ and β_b of the following regression function:

$$Y_i=\alpha_b+\beta_b D_i^j+\epsilon_i^j$$

where observation is at the pair level. Within each subclass b, D_i^j is as good as random and independent of potential outcome. This means $\mathbb{E}[\epsilon_i^j|D_i^j] = 0$, and that $\tau_{a,b}^r = \beta_b$:

$$\begin{aligned}\tau_{a,b}^r &= \mathbb{E}_{(i,j)}[Y_i^{obs}|(i,j) \in b, D_i^j = 1] - \mathbb{E}_{(i,j)}[Y_i^{obs}|(i,j) \in, D_i^j = 0] \\ &= \mathbb{E}_{(i,j)}[\alpha_b + \beta_b + \epsilon_i^j|(i,j) \in b, D_i^j = 1] - \mathbb{E}_{(i,j)}[\alpha_b + \epsilon_i^j|(i,j) \in b, D_i^j = 0] \\ &= \beta_b\end{aligned}$$

Expressing $\tau_{a,b}^r$ as a regression coefficient allows the easy incorporation of additional covariates into the analysis. Including pre-treatment predictors of the outcome in the regression could help reduce the bias coming from the variation of propensity scores within the same subclass, as well as increasing estimation precision, the same as in the conventional subclassification method Imbens and Rubin (2015).

2.3 Discussion of Assumption 2

2.3.1 Super Population

We are interested in the super population if the estimands of interest are functions of the infinite population, for example the contrast in the mean potential outcomes for all units in the infinite population, including the ones not sampled. Assumption 2 is automatically satisfied if the sample network **D** is viewed as constructed by uniform random sampling of nodes from an infinite super population network with infinite number of nodes, where a link is recorded in the sample if it exists in the super population network. Under this construction, the randomness in link formation, or in other words, the assignment mechanism, solely comes from random sampling.

To see why random node sampling from super population implies Assumption 2, we proceed in 3 steps. First, based on the definition in Crane (2018), Assumption 2 is equivalent to **D** being vertex exchangeable. Second, under the Aldous-Hoover theorem, the equivalence of the De-Finetti theorem for network data, the distribution of vertex exchangeable network links can *always* be represented by some graphon process:

Definition 2.3.1 (Graphon (Crane, 2018))**.** Function $\phi \in \Phi : [0,1] \times [0,1] \to [0,1]$ has 0 diagonal. Fix any $\phi \in \varPhi$ and draw $w_1, w_2, ...$ i.i.d. Uniform[0,1]. Given $w_1, w_2, ...,$ assign D_i^j conditionally independently with probabilities

$$Pr(D_i^j = 1|w_1, w_2, ...; \phi) = \phi(w_i, w_j) \tag{2.3}$$

This way of constructioning $\mathbf{D}$ is called a *graphon process.*

Therefore random node sampling guarantees that there exists i.i.d. $\{w_i\}_{1 \leq i \leq N}$ such that

$$Pr(\mathbf{D} = \mathbf{d}|w_1, w_2, ..., w_N) = \prod_{i=1}^{N} \prod_{j \neq i}^{N} Pr(D_i^j = d_i^j|w_i, w_j) \tag{2.4}$$

Finally, as the third step let us compare equation (2.4) to equation (1.1). We can see the difference is that here $\mathbf{U}_i = \mathbf{V}_i = w_i$. This is not restrictive because we could always transform a vector of random variables to a random variable with standard uniform distribution. Moreover, it is always possible to find another function ϕ' such that $\phi(w_i, w_j)$=$\phi'(\mathbf{U}_i, \mathbf{V}_j)$

In conclusion, when the sample network is constructed by random node sampling from an infinite super population network, the assumption of doubly individualistic assignment mechanism must be true. This is similar to the case of conventional causal inference where random sampling from super population guarantees that the assignment mechanism is individualistic Imbens and Rubin (2015). Note that only random node sampling guarantees Assumption 2. Other sampling schemes, such as random link sampling, do not enjoy this property. An example of link sampling is in the study of co-authorship network where article is the sampling unit instead of the authors being the sampling unit.

2.3.2 Finite Population

In Leung (2015)'s network formation model, i's linking decision could depend on the anticipated network structure. Network nodes simultaneously form directed links to maximise expected utility given their beliefs about the state of the network. Because the objective is the expected utility, i's linking probability will be a function of equilibrium beliefs about others' linking decisions, conditioning on the observed attributes of all agents in the network. For this reason, equilibrium linking decisions are functions of the exogenous attributes only.

As such, the pairwise linking decision can be expressed as

$$D_i^j = h(Z_i, Z_j, \theta_{ij})$$

where Z_i includes both i's equilibrium beliefs about the the state of the network and i's observed exogenous attributes. Observed exogenous attributes are assumed to be common knowledge. Leung (2015) assumes that θ_{ij} are unobserved node or pairwise attributes that are private information and satisfy $\theta_{ij} \perp\!\!\!\perp \theta_{kl}$ for $i \neq k$. This allows θ_{ij} to be correlated with θ_{il}, which means by just conditioning on Z_i and Z_j we couldn't yet write the probability distribution of the entire network links as a conditionally independent process in the form of equation (1.1). But if we partition θ_{ij} into $(v_{1,i}, v_{2,ij})$ where $v_{1,i}$ are unobserved shocks to link formation common to more than one sender j, and $v_{2,ij}$ are mutually independent pairwise shocks. The idea is that we could always separate out variables that cause correlations among θ_{ij} and V_{il} for $j \neq l$, and put them in $v_{1,i}$. Then

$$D_i^j = h(Z_i, Z_j, \theta_{ij})$$

becomes

$$D_i^j = h(Z_i, Z_j, v_{1,i}, v_{2,ij}) = \tilde{h}(\mathbf{U}_i, \mathbf{V}_j, v_{2,ij})$$

where $\mathbf{U}_i = (Z_i, v_{1,i})$. Conditioning on $\mathbf{U}_i, \mathbf{V}_j$, the probability distribution of network links then becomes exactly as in equation (1.1). Therefore, network formation games with network externalities as specified in Leung (2015) satisfy the individualistic assignment mechanism assumption.

2.4 Proofs

As before, subscript d in all probabilities and expectations indicate the distribution is over random link assignment, subscript (i, j) and i indicate the distribution is over sampling from the super-population of pairs of nodes and nodes, respectively. I use $\mathbb{E}$ to indicate expectations over the distributions of both random link assignment and sampling from the super-population.

2.4.1 Proof of Lemma 1

Proof. LHS:

$$Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j, e(\mathbf{U}_i, \mathbf{V}_j)) = Pr_d(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j) = e(\mathbf{U}_i, \mathbf{V}_j)$$

The first equality holds because $e(\mathbf{U}_i, \mathbf{V}_j)$ is a function of $\mathbf{U}_i, \mathbf{V}_j$, the second equality holds from the definition of $e(\mathbf{U}_i, \mathbf{V}_j)$.

RHS:

$$\begin{aligned} Pr_d(D_i^j = 1|e(\mathbf{U}_i, \mathbf{V}_j)) &= \mathbb{E}_d[D_i^j|e(\mathbf{U}_i, \mathbf{V}_j)] = \mathbb{E}[\mathbb{E}_d[D_i^j|\mathbf{U}_i, \mathbf{V}_j, e(\mathbf{U}_i, \mathbf{V}_j)]|e(\mathbf{U}_i, \mathbf{V}_j)] \\ &= \mathbb{E}[\mathbb{E}_d[D_i^j|\mathbf{U}_i, \mathbf{V}_j]|e(\mathbf{U}_i, \mathbf{V}_j)] = \mathbb{E}[e(\mathbf{U}_i, \mathbf{V}_j)|e(\mathbf{U}_i, \mathbf{V}_j)] \\ &= e(\mathbf{U}_i, \mathbf{V}_j) \end{aligned}$$

□

2.4.2 Proof of Lemma 2

Proof.

$$\begin{aligned} &Pr_d(D_i^j = 1|Y_i^{pot}, e(\mathbf{U}_i, \mathbf{V}_j)) \\ &= \mathbb{E}_d[D_i^j = 1|Y_i^{pot}, e(\mathbf{U}_i, \mathbf{V}_j)] \\ &= \mathbb{E}\left[\mathbb{E}_d[D_i^j = 1|Y_i^{pot}, \mathbf{U}_i, \mathbf{V}_j, e(\mathbf{U}_i, \mathbf{V}_j)]\big|Y_i^{pot}, e(\mathbf{U}_i, \mathbf{V}_j)\right] \end{aligned}$$

The inner expectation is equal to $\mathbb{E}_d[D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j, e(\mathbf{U}_i, \mathbf{V}_j)]$ by unconfoundedness given $\mathbf{U}_i, \mathbf{V}_j$. And by the balancing property of the propensity score, this is $\mathbb{E}_d[D_i^j = 1|e(\mathbf{U}_i, \mathbf{V}_j)]$.

Therefore the last expression is

$$
\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}_d[D_i^j = 1|e(\mathbf{U}_i, \mathbf{V}_j)]\Big|Y_i^{pot}, e(\mathbf{U}_i, \mathbf{V}_j)\Big] \\
&= \mathbb{E}_d[D_i^j = 1|e(\mathbf{U}_i, \mathbf{V}_j)] \\
&= Pr_d(D_i^j = 1|e(\mathbf{U}_i, \mathbf{V}_j))
\end{aligned}
$$

□

2.4.3 Proof of proposition 2

Proof.

$$
\begin{aligned}
\tau_a^r &:= \mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}) - Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})|R_i = r, A^j = a] \\
&= \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&- \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})|\mathbf{U}_i, \mathbf{V}_j, D_i^j = 1, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&- \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})|\mathbf{U}_i, \mathbf{V}_j, D_i^j = 0, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})|e(\mathbf{U}_i, \mathbf{V}_j), D_i^j = 1, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&- \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})|e(\mathbf{U}_i, \mathbf{V}_j), D_i^j = 0, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i^{obs}|\mathbf{U}_i, \mathbf{V}_j, D_i^j = 1, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&- \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i^{obs}|\mathbf{U}_i, \mathbf{V}_j, D_i^j = 0, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i^{obs}|e(\mathbf{U}_i, \mathbf{V}_j), D_i^j = 1, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big] \\
&- \mathbb{E}\Big[\mathbb{E}_{(i,j)}[Y_i^{obs}|e(\mathbf{U}_i, \mathbf{V}_j), D_i^j = 0, R_i = r, A^j = a]\big|R_i = r, A^j = a\Big]
\end{aligned}
$$

The second equation is from law of iterated expectations. The third and fourth are from unconfoundedness given both $\mathbf{U}_i, \mathbf{V}_j$ and $e(\mathbf{U}_i, \mathbf{V}_j)$. The fifth and the last equalities are from no multiple versions of treatment assumption.

□

2.4.4 Proof of unbiasedness of IPW estimator

Proof. Here I will only prove that

$$\mathbb{E}_{(i,j)}\Big[\frac{1}{\sum_{i=1}^{N} R_i = r}\cdot\frac{1}{\sum_{j=1}^{N} A^j = a}\sum_{i:R_i=r}\sum_{j:A^j=a}\frac{D_i^j\cdot Y_i^{obs}}{e(\mathbf{U}_i,\mathbf{V}_j)}\Big] = \mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}|R_i = r, A^j = a].$$

The case for

$$\mathbb{E}_{(i,j)}\Big[\frac{1}{\sum_{i=1}^{N} R_i = r}\cdot\frac{1}{\sum_{j=1}^{N} A^j = a}\sum_{i:R_i=r}\sum_{j:A^j=a}\frac{(1-D_i^j)\cdot Y_i^{obs}}{1-e(\mathbf{U}_i,\mathbf{V}_j)}\Big] = \mathbb{E}_{(i,j)}[Y_i(D_i^j = 0, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}|R_i = r, A^j = a$$

can be similarly proved.

$$\begin{aligned}
&\mathbb{E}_{(i,j)}\Big[\frac{1}{\sum_{i=1}^{N} R_i = r}\cdot\frac{1}{\sum_{j=1}^{N} A^j = a}\sum_{i:R_i=r}\sum_{j:A^j=a}\frac{D_i^j\cdot Y_i^{obs}}{e(\mathbf{U}_i,\mathbf{V}_j)}|R_i = r, A^j = a\Big]\\
=&\mathbb{E}_{(i,j)}\big[\frac{Y_i^{obs}\cdot D_i^j}{e(\mathbf{U}_i,\mathbf{V}_j)}|R_i = r, A^j = a\big]\\
=&\mathbb{E}_{(i,j)}\big[\frac{Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})\cdot D_i^j}{e(\mathbf{U}_i,\mathbf{V}_j)}|R_i = r, A^j = a\big]\\
=&\mathbb{E}_{(i,j)}\Big[\mathbb{E}_{(i,j)}\big[\frac{Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})\cdot D_i^j}{e(\mathbf{U}_i,\mathbf{V}_j)}\big|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a\big]|R_i = r, A^j = a\Big]
\end{aligned}$$

The second equation holds because $Y_i^{obs} = Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})$ when $D_i^j = 1$, the third equation is from iterated expectations.

Then the inner expectation can be re-written as

$$\begin{aligned}
&\mathbb{E}_{(i,j)}\big[\frac{Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})\cdot D_i^j}{e(\mathbf{U}_i,\mathbf{V}_j)}\big|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a\big]\\
=&\frac{\mathbb{E}_{(i,j)}[D_i^j|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a]\cdot\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a]}{e(\mathbf{U}_i,\mathbf{V}_j)}\\
=&\frac{e(\mathbf{U}_i,\mathbf{V}_j)\cdot\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a]}{e(\mathbf{U}_i,\mathbf{V}_j)}\\
=&\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a]
\end{aligned}$$

where the first equation holds because D_i^j and $Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j})$ are independent conditional on $\mathbf{U}_i, \mathbf{V}_j$, by unconfoundedness 1. Therefore

$$\begin{aligned}
&\mathbb{E}_{(i,j)}\Big[\mathbb{E}_{(i,j)}\big[\frac{Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}) \cdot D_i^j}{e(\mathbf{U}_i, \mathbf{V}_j)}, \mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a\big]|R_i = r, A^j = a\Big] \\
=&\mathbb{E}_{(i,j)}\big[\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}|\mathbf{U}_i, \mathbf{V}_j, R_i = r, A^j = a\big]|R_i = r, A^j = a\big] \\
=&\mathbb{E}_{(i,j)}[Y_i(D_i^j = 1, \mathbf{D}_i^{-j} = \bar{\mathbf{d}}_i^{-j}|R_i = r, A^j = a]
\end{aligned}$$

□

2.4.5 Proof of Proposition 1

Proof. $m^{c_1,c_2} = \mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(d^{c_1})]\big] - \mathbb{E}_i\big[\mathbb{E}_{d^{c_2}}[Y_i(d^{c_2})]\big]$. Here I will only prove that $\mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(d^{c_1})]\big]$ is identified. The identification of $\mathbb{E}_i\big[\mathbb{E}_{d^{c_2}}[Y_i(d^{c_2})]\big]$ follows similarly.

$$\begin{aligned}
&\mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(d^{c_1})]\big] = \mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(\mathbf{D}_i = d^{c_1})]\big] \\
&= \mathbb{E}\Big[\mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(\mathbf{D}_i = d^{c_1})|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N]\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(\mathbf{D}_i = d^{c_1})|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N, \mathbf{D}_i = d^{c_1}]\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(\mathbf{D}_i = d^{c_1})|Pr(\mathbf{D}_i = d^{c_1}|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N), \mathbf{D}_i = d^{c_1}]\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_i\big[\mathbb{E}_{d^{c_1}}[Y_i(\mathbf{D}_i = d^{c_1})|e(\mathbf{U}_i, \mathbf{V}_1), ..., e(\mathbf{U}_i, \mathbf{V}_N), \mathbf{D}_i = d^{c_1}]\big]\Big]
\end{aligned}$$

The first equation comes from the law of iterated expectations. The second equation follows the unconfoundedness condition in Lemma 1. The third equation comes from the balancing property of generalised propensity scores. The last equation holds because

$$\begin{aligned}
&Pr(\mathbf{D}_i = d^{c_1}|\mathbf{U}_i, \mathbf{V}_1, ..., \mathbf{V}_N) \\
=&\prod_{j=1}\big(Pr(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j)\big)^{d_i^{c_1}}\big(1 - Pr(D_i^j = 1|\mathbf{U}_i, \mathbf{V}_j)\big)^{1-d_i^{c_1}} \\
=&\prod_{j=1}\big(e(\mathbf{U}_i, \mathbf{V}_j)\big)^{d_i^{c_1}}\big(1 - e(\mathbf{U}_i, \mathbf{V}_j)\big)^{1-d_i^{c_1}}
\end{aligned}$$

□

2.5 Additional simulation results

2.5.1 Details of network formation model g_2

The second network link generation process $Pr(D_i^j = 1) = g_2(C_i, C_j)$ is a slgihtly more complicated version of a statistical block model. The linking probabilities are asymmetric, that is $g_2(C_i, C_j) \neq g_2(C_j, C_i)$. For any node i and j, the probability of i receiving a link from j is in general higher if i) C_i is larger and ii) C_j is slightly higher than C_j. If we think of C as the ability of the node, this is a model where higher ability nodes receive more friendships, but only from nodes who are slightly more able than themselves. This might be because they don't like people who are less able than them, and admire people who are more able, but become jealous of people who are too much more able than themselves.

$$g_2 : P_i^j = \begin{cases} 0.05 & \text{if } C_i \in [0.1, 0.2) \;\&\; C_j \in (0.2, 0.21] \\ 0.1 & \text{if } C_i \in [0.2, 0.3) \;\&\; C_j \in (0.3, 0.31] \\ 0.15 & \text{if } C_i \in [0.3, 0.4) \;\&\; C_j \in (0.4, 0.41] \\ 0.2 & \text{if } C_i \in [0.4, 0.5) \;\&\; C_j \in (0.5, 0.51], \text{ or if, } C_i \in [0.5, 0.6) \;\&\; C_j \in (0.6, 0.61] \\ 0.25 & \text{if } C_i \in [0.6, 0.7) \;\&\; C_j \in (0.7, 0.71], \text{ or if, } C_i \in [0.7, 0.8) \;\&\; C_j \in (0.8, 0.81] \\ 0.3 & \text{if } C_i \in [0.8, 0.9) \;\&\; C_j \in (0.9, 0.91], \text{ or if, } C_i \in [0.9, 1] \;\&\; C_j \in (0.99, 1] \\ 0.01 & \text{if } C_i \in [a, a+0.1) \;\&\; C_j \in [a, a+0.1) \text{ for } a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 \\ & \text{or if, } C_i \in [0, 0.1) \;\&\; C_j \in [0, 0.05) \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

2.6 Empirical application supplementary material

Table 2.4: Mean degree distribution for simulated g2 networks

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
N=100	0	0	0	0	0	0.2	0.8	1	1.1	1.9	4.3
N=300	0	0	0.2	1	1	1.5	2	2.6	3.3	4.7	10.2
N=500	0	0.3	1	1.5	2	2.7	3.2	4.1	5.5	7.4	15.7

Note: This table reports the mean degree distribution of the simulated networks. For each size N=100,300,500, and for each simulated network of that size, I caculate the deciles of the number of links each link receiver receives, and average over all the 500 simulated networks of that size.

Table 2.5: Simulation results for g_2

				IPW	Matching	Sub	Naive ols
Y_b	Bias	X_0					
			N=100	0.083264	0.096999	0.099578	0.135588
			N=300	0.048002	0.084757	0.087946	0.167218
			N=500	0.036638	0.088748	0.091242	0.186257
		X_1					
			N=100	0.074813	0.087677	0.089541	0.126791
			N=300	0.04956	0.086555	0.08927	0.167975
			N=500	0.035027	0.085468	0.089596	0.184303
	MAE	X_0					
			N=100	0.103605	0.134378	0.112983	0.143077
			N=300	0.050245	0.085861	0.087946	0.167218
			N=500	0.037016	0.088748	0.091242	0.186257
		X_1					
			N=100	0.094632	0.114537	0.10228	0.13395
			N=300	0.052468	0.087344	0.089483	0.167975
			N=500	0.03631	0.085468	0.089596	0.184303
Y_c	Bias	X_0					
			N=100	0.465683	0.529408	0.561574	0.779459
			N=300	0.2608	0.470754	0.494971	0.956848
			N=500	0.186274	0.526728	0.546989	1.148476
		X_1					
			N=100	0.456105	0.536314	0.54596	0.76797
			N=300	0.263195	0.482857	0.495973	0.954869
			N=500	0.177633	0.512275	0.537143	1.136513
	MAE	X_0					
			N=100	0.489148	0.601876	0.575155	0.784664
			N=300	0.262791	0.470754	0.494971	0.956848
			N=500	0.187562	0.526728	0.546989	1.148476
		X_1					
			N=100	0.465316	0.555527	0.548736	0.768234
			N=300	0.263612	0.482857	0.495973	0.954869
			N=500	0.179018	0.512275	0.537143	1.136513

Note: This table reports for the g_2 model the bias and the mean absolute error (MAE) of the inverse probability weighting estimator, the nearest neighbour matching estimator with replacement, the subclassification estimator and the narive ols estimator, compared to the true linking effects, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1, the number of subclasses for the subclassification estimator is 8. All the estimates are for the average treatment effect for the treated.

Table 2.6: True Propensity Score vs True Effects for g_1

				IPW	Matching	Sub
Y_b	Bias	X_0				
			N=100	-0.00618	0.00017	-0.00134
			N=300	0.002273	0.002384	0.006239
			N=500	-0.00044	-0.00046	0.004332
		X_1				
			N=100	-0.00422	-0.00806	-0.00598
			N=300	-0.00232	-0.00379	0.00157
			N=500	0.000664	0.00104	0.005368
	MAE	X_0				
			N=100	0.080231	0.097323	0.073752
			N=300	0.026164	0.029931	0.023283
			N=500	0.013928	0.018599	0.013309
		X_1				
			N=100	0.065956	0.085406	0.061449
			N=300	0.025165	0.029701	0.023626
			N=500	0.016689	0.018747	0.016217
Y_c	Bias	X_0				
			N=100	0.005795	0.011732	0.04541
			N=300	-0.00335	-0.00898	0.026149
			N=500	0.000463	-0.00044	0.033232
		X_1				
			N=100	-0.02251	-0.05712	-0.01709
			N=300	-0.01018	-0.01714	0.018337
			N=500	-0.0004	-0.00359	0.031436
	MAE	X_0				
			N=100	0.259116	0.281393	0.208147
			N=300	0.101655	0.104463	0.079084
			N=500	0.069394	0.070489	0.056754
		X_1				
			N=100	0.219053	0.215143	0.159517
			N=300	0.086806	0.088471	0.066411
			N=500	0.058299	0.058505	0.049941

Note: This table reports for the g_1 model the bias and the mean absolute error (MAE) of the inverse probability weighting estimator, the nearest neighbour matching estimator with replacement, and the subclassification estimator using true peopensity scores, compared to the true linking effects, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1, the number of subclasses for the subclassification estimator is 8. All the estimates are for the average treatment effect for the treated.

Table 2.7: True Propensity Score vs True Effects for g_2

				IPW	Matching	Sub
Y_b	Bias	X_0				
			N=100	0.000634	0.000271	0.003285
			N=300	-0.00094	0.00024	0.004268
			N=500	0.000682	0.000218	0.004587
		X_1				
			N=100	-0.00704	-0.01052	-0.00802
			N=300	0.000635	-0.00019	0.004349
			N=500	-0.00077	-0.00149	0.004058
	MAE	X_0				
			N=100	0.077348	0.094386	0.067707
			N=300	0.02476	0.030613	0.02275
			N=500	0.015357	0.019648	0.014188
		X_1				
			N=100	0.068187	0.088944	0.063137
			N=300	0.024323	0.029334	0.022893
			N=500	0.016496	0.017966	0.014982
Y_c	Bias	X_0				
			N=100	0.002911	-0.00577	0.02452
			N=300	-0.00701	-0.00254	0.028533
			N=500	0.003169	-0.00064	0.031012
		X_1				
			N=100	-0.01186	-0.01944	-0.00279
			N=300	-0.00522	-0.01368	0.020637
			N=500	-0.00502	-0.00856	0.027283
	MAE	X_0				
			N=100	0.270586	0.274076	0.199427
			N=300	0.100294	0.103225	0.080769
			N=500	0.068604	0.071309	0.055641
		X_1				
			N=100	0.211424	0.224245	0.162451
			N=300	0.084764	0.091543	0.068273
			N=500	0.062315	0.059699	0.049605

Note: This table reports for the g_2 model the bias and the mean absolute error (MAE) of the inverse probability weighting estimator, the nearest neighbour matching estimator with replacement, and the subclassification estimator using true peopensity scores, compared to the true linking effects, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1, the number of subclasses for the subclassification estimator is 8. All the estimates are for the average treatment effect for the treated.

Table 2.8: Using estimated propensity scores vs true propensity scores for g_1

				IPW	Matching	Sub
Y_b	Bias	X_0				
			N=100	0.083621	0.096681	0.095233
			N=300	0.049644	0.084352	0.085735
			N=500	0.033614	0.084658	0.083421
		X_1				
			N=100	0.082936	0.102901	0.098823
			N=300	0.049853	0.087267	0.086033
			N=500	0.033867	0.08433	0.083669
	MAE	X_0				
			N=100	0.08604	0.13704	0.096486
			N=300	0.0501	0.085676	0.085735
			N=500	0.033836	0.084672	0.083421
		X_1				
			N=100	0.084721	0.124468	0.09916
			N=300	0.050546	0.087735	0.086033
			N=500	0.034166	0.08433	0.083669
Yc	Bias	X_0				
			N=100	0.488645	0.579477	0.538106
			N=300	0.26446	0.492194	0.47262
			N=500	0.172692	0.512657	0.499917
		X_1				
			N=100	0.476862	0.591576	0.556474
			N=300	0.26786	0.4877	0.475234
			N=500	0.17529	0.510615	0.501656
	MAE	X_0				
			N=100	0.488645	0.635377	0.541816
			N=300	0.264799	0.492258	0.47262
			N=500	0.172846	0.512657	0.499917
		X_1				
			N=100	0.476916	0.619324	0.556625
			N=300	0.267956	0.4877	0.475234
			N=500	0.175635	0.510615	0.501656

Note: This table reports for the g_1 model the bias and the mean absolute error (MAE) of the inverse probability weighting estimator, the nearest neighbour matching estimator with replacement, and the subclassification estimator using factor model estimated propensity scores, compared to the linking effects estimated using the true propensity socres, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1, the number of subclasses for the subclassification estimator is 8. All the estimates are for the average treatment effect for the treated.

Table 2.9: Using estimated propensity scores vs true propensity scores for g_2

				IPW	Matching	Sub
Yb	Bias	X_0				
			N=100	0.08263	0.096729	0.096293
			N=300	0.048945	0.084517	0.083678
			N=500	0.035957	0.088529	0.086655
		X_1				
			N=100	0.081855	0.098198	0.097559
			N=300	0.048925	0.086743	0.084921
			N=500	0.035799	0.086957	0.085537
	MAE	X_0				
			N=100	0.086134	0.138018	0.097608
			N=300	0.049154	0.086261	0.083678
			N=500	0.036224	0.088529	0.086655
		X_1				
			N=100	0.084542	0.120627	0.098128
			N=300	0.049115	0.086914	0.084921
			N=500	0.036091	0.086957	0.085537
Yc	Bias	X_0				
			N=100	0.462772	0.535179	0.537054
			N=300	0.267808	0.473296	0.466438
			N=500	0.183105	0.527369	0.515976
		X_1				
			N=100	0.467964	0.555751	0.548748
			N=300	0.268415	0.496538	0.475336
			N=500	0.182649	0.520831	0.50986
	MAE	X_0				
			N=100	0.462914	0.601572	0.537354
			N=300	0.267808	0.473863	0.466438
			N=500	0.183854	0.527369	0.515976
		X_1				
			N=100	0.46845	0.574089	0.549009
			N=300	0.268415	0.496538	0.475336
			N=500	0.182914	0.520831	0.50986

Note: This table reports for the g_2 model the bias and the mean absolute error (MAE) of the inverse probability weighting estimator, the nearest neighbour matching estimator with replacement, and the subclassification estimator using factor model estimated propensity scores, compared to the linking effects estimated using the true propensity socres, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1, the number of subclasses for the subclassification estimator is 8. All the estimates are for the average treatment effect for the treated.

Table 2.10: Matching and Subclassification with increasing matches and subclasses vs True Effects for g_1

				Matching	Sub
Y_b	Bias	X_0			
			N=100	0.096851	0.093895
			N=300	0.088153	0.091161
			N=500	0.083725	0.085986
		X_1			
			N=100	0.094838	0.092844
			N=300	0.082749	0.08666
			N=500	0.084929	0.087267
	MAE	X_0			
			N=100	0.137418	0.111374
			N=300	0.088258	0.091161
			N=500	0.083725	0.085986
		X_1			
			N=100	0.11907	0.103271
			N=300	0.082933	0.086674
			N=500	0.084929	0.087267
Y_c	Bias	X_0			
			N=100	0.591209	0.583515
			N=300	0.483253	0.493814
			N=500	0.508767	0.522097
		X_1			
			N=100	0.534451	0.539381
			N=300	0.467582	0.488238
			N=500	0.507616	0.521799
	MAE	X_0			
			N=100	0.62927	0.595459
			N=300	0.483253	0.493814
			N=500	0.508767	0.522097
		X_1			
			N=100	0.558012	0.542142
			N=300	0.467582	0.488238
			N=500	0.507616	0.521799

Note: This table reports for the g_1 model the bias and the mean absolute error (MAE) of the nearest neighbour matching estimator with replacement and the subclassification estimator with factor model estimated propensity scores, compared to the true linking effects, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1 for networks with $N = 100$, 3 for networks with $N = 300$, 5 for networks with $N = 500$. The number of subclasses for the subclassification estimator is 8 for networks with $N = 100$, 10 for networks with $N = 300$, 12 for networks with $N = 500$. All the estimates are for the average treatment effect for the treated.

Table 2.11: Matching and Subclassification with increasing matches and subclasses vs True Effects for g_2

				Matching	Sub
Y_b	Bias	X_0			
			N=100	0.096999	0.099578
			N=300	0.083618	0.086948
			N=500	0.088433	0.089523
		X_1			
			N=100	0.087677	0.089541
			N=300	0.086064	0.08834
			N=500	0.085855	0.087839
	MAE	X_0			
			N=100	0.134378	0.112983
			N=300	0.083826	0.086948
			N=500	0.088433	0.089523
		X_1			
			N=100	0.114537	0.10228
			N=300	0.086486	0.08858
			N=500	0.085855	0.087839
Y_c	Bias	X_0			
			N=100	0.529408	0.561574
			N=300	0.470578	0.489446
			N=500	0.525757	0.535877
		X_1			
			N=100	0.536314	0.54596
			N=300	0.47555	0.490552
			N=500	0.513496	0.526042
	MAE	X_0			
			N=100	0.601876	0.575155
			N=300	0.470578	0.489446
			N=500	0.525757	0.535877
		X_1			
			N=100	0.555527	0.548736
			N=300	0.47555	0.490552
			N=500	0.513496	0.526042

Note: This table reports for the g_2 model the bias and the mean absolute error (MAE) of the nearest neighbour matching estimator with replacement and the subclassification estimator with factor model estimated propensity scores, compared to the true linking effects, for link sender with $X_j = 0$ and $X_j = 1$ separately. The number of matches for the matching estimator is 1 for networks with $N = 100$, 3 for networks with $N = 300$, 5 for networks with $N = 500$. The number of subclasses for the subclassification estimator is 8 for networks with $N = 100$, 10 for networks with $N = 300$, 12 for networks with $N = 500$. All the estimates are for the average treatment effect for the treated.

Table 2.12: Variable definitions for CFP friendship re-analysis

Variable	Definition in the original papers	Definition in this paper
Post college education for parents	Dummy variable equal to 1 if the respondent reports that the highest level of education attained by their residential father and residential mother has a post-college education, and 0 otherwise. If a student either does not have a residential father/mother or the information is missing, that parent's level of education is imputed using the other parent's education [a].	Same definition. The difference is that in-home data is used instead. If the in-home data is missing, in-school data is used. This is because for saturated schools, data from in-home interviews have less missing values than data from the in-school survey.
log family income	log of total household income (thousands). If family income is missing, family income is set to the mean value for the school and a dummy is included for missing family income.	Same. In addition, for families with 0 annual family income, their income is replaced with 0.1, in order for the log income to take real values.
Grade	Grade point average is calculated based on self-reported student grades in math, science, english, and history from the Wave I in-home survey where A=4, B=3, C=2, and D or lower=1.	Same. Note: If the respondent didn't take the subject, I code the grade as missing.
MaleFrac (FemaleFrac) high	They are the fraction of male and female high flyers (those with at least one post-college parent) in the grade and school.	Same
Bachelor's degree	Dummy variable equal to 1 if the respondent has completed a bachelor's degree (four-year college) and 0 otherwise.	Same
LFP	Dummy variable equal to 1 if the respondent is currently working at least 10 hours per week, is on sick leave or temporarily disabled, is on maternity/paternity leave, or is unemployed and looking for work, and is equal to zero otherwise.	Same
Ever married	Dummy variable equal to one if the respondent resported they have ever been married	Same
Children	Total number of (non-deceased) biological children they have.	Same

[a] For example, if the residential father's education is missing, but the residential mother has a high-school education, they impute a value for father post-college by taking the average value of father post-college among students of the same gender within the school who also have a residential mother with a high-school education. If there are no students with equivalent mother's education and non-missing information on father's education, they impute father post-college using the value of father post-college among all students in the school who have a residential mother with a high-school education.

Table 2.13: Naive OLS estimates for the effect of friendship

	Bachelor's Degree (p.p)	Want (p.p)	Will (p.p)	Intelligence (p.p)
F_FL	0.638^{***}	0.195	-0.109	0.214
	(0.163)	(0.208)	(0.206)	(0.209)
F_ML	1.150^{***}	0.525	0.284	0.175
	(0.342)	(0.379)	(0.363)	(0.441)
F_FH	2.984^{***}	4.441^{***}	2.600^{**}	0.955
	(1.118)	(0.987)	(1.065)	(1.699)
F_MH	2.052	1.474	1.429	3.451^{**}
	(1.435)	(1.344)	(1.120)	(1.741)
M_FL	0.473^{*}	0.152	0.147	-0.697^{**}
	(0.282)	(0.272)	(0.283)	(0.314)
M_ML	0.499^{**}	-0.058	-0.253	-0.327
	(0.202)	(0.189)	(0.217)	(0.244)
M_FH	4.145^{**}	1.971	-3.514	0.021
	(1.777)	(2.013)	(2.480)	(2.833)
M_MH	3.262^{**}	1.765	2.540^{**}	4.102^{***}
	(1.561)	(1.378)	(1.123)	(1.145)

Note: This table reports the naive OLS estimated effects of high school friendship on students' bachelor's degree attainment (column 1), and their intermediate outcomes (column 2-4). The dependent variable in Column (2) is a dummy variable recording whether the student reported a scale 5 (1 is the lowest and 5 is the highest) on the the extent of how much they want to go to college (Wave II). The dependent variable in Column (3) is a dummy variable recording whether the student reported a scale 5 (1 is the lowest and 5 is the highest) on the likelihood that they will go to college (Wave II). The dependent variable in Column (4) is a dummy variable recording whether the student reported a scale 5 or 6 (1 is the lowest and 6 is the highest) on their intelligence compared to other people of their age (Wave II). Each row corresponds to a characterisation of the friendship, based on the character of the receiver and the sender. Receiver characteristics is shown before the underbar _, and sender characteristics is shown after. "F" and "M" are used to refer to the gender female and male respectively. "H" and "L" are used to refer to whether the individual is a high flyer or non-high flyer (low flyer) respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high flyer link senders. The regressions reported in all columns include cohort dummies, whether the student was born in the US, their PVT score, whether their PVT score is above the population median PVT score, whether their mother's and father's highest degree is high school, some college, college, or post college, whether their mother's and father's highest education level is missing, the student's log family income, whether family is missing, the age of the student during Wave I, whether the student's mother and father were in the household, dummies for whether the student is black, hispanic, white, asian and indian. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. *p<0.1; **p<0.05; ***p<0.01

Table 2.14: Effect of friendship on long-term outcomes

	LFP	Num Children	Married
	(1)	(2)	(3)
F_FL	0.002	0.003	0.007^{***}
	(0.002)	(0.005)	(0.002)
F_ML	0.001	-0.031^{***}	-0.001
	(0.004)	(0.008)	(0.004)
F_FH	-0.016	-0.081^{***}	0.046^{***}
	(0.012)	(0.031)	(0.012)
F_MH	0.024	-0.064^{***}	0.006
	(0.015)	(0.018)	(0.010)
M_FL	0.010^{***}	0.004	0.009^{***}
	(0.003)	(0.008)	(0.003)
M_ML	0.003	-0.003	-0.003
	(0.003)	(0.006)	(0.003)
M_FH	-0.055^{**}	0.024	-0.015
	(0.024)	(0.031)	(0.016)
M_MH	-0.034^{**}	-0.057^{***}	0.025^{*}
	(0.013)	(0.021)	(0.013)

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

Note: This table reports the estimated effects of high school friendship on students' long term outcomes measured in Wave IV. The dependent variable in Column (1) is a dummy variable recording whether the respondent was part of the labour force. The dependent variable in Column (2) is the number of children the respondent. The dependent variable in Column (3) is a dummy variable recording whether the respondent has ever been married. The estimands are all ATT. Each row corresponds to a characterisation of the friendship, based on the character of the receiver and the sender. Receiver characteristics is shown before the underbar _, and sender characteristics is shown after. "F" and "M" are used to refer to the gender female and male respectively. "H" and "L" are used to refer to whether the individual is a high flyer or non-high flyer (low flyer) respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high flyer link senders. The regressions reported in all columns include cohort dummies, whether the student was born in the US, their PVT score, whether their PVT score is above the population median PVT score, whether their mother's and father's highest degree is high school, some college, college, or post college, whether their mother's and father's highest education level is missing, the student's log family income, whether family is missing, the age of the student during Wave I, whether the student's mother and father were in the household, dummies for whether the student is black, hispanic, white, asian and indian. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. *p<0.1; **p<0.05; ***p<0.01

Table 2.15: Heterogeneous effects of friendship on desire and likelihood to go to college

	Dependent variable:			
	Want		Will	
	PVT Median -	PVT Median +	PVT Median -	PVT Median +
	(1)	(2)	(3)	(4)
F_FL	-2.100^{***}	0.603	-1.256^{***}	-0.328
	(0.508)	(0.385)	(0.350)	(0.407)
F_ML	0.216	-0.462	-1.396^{*}	0.091
	(0.798)	(0.876)	(0.774)	(0.697)
F_FH	5.494^{***}	-0.387	2.600	-0.410
	(2.089)	(1.708)	(3.457)	(1.584)
F_MH	4.506^{*}	-0.722	5.365^{***}	-0.441
	(2.393)	(1.548)	(1.816)	(1.260)
M_FL	0.153	-0.893^{*}	0.394	-0.890^{*}
	(0.767)	(0.496)	(0.746)	(0.522)
M_ML	1.010^{**}	-0.525	-0.352	-0.884^{**}
	(0.436)	(0.437)	(0.480)	(0.436)
M_FH	1.805	0.648	-5.984^{**}	0.746
	(2.537)	(1.928)	(2.426)	(2.587)
M_MH	-4.820^{**}	10.053^{***}	-3.905	9.147^{***}
	(2.295)	(2.411)	(2.912)	(2.593)

Note: This table reports the estimated heterogeneous effects of high school friendship on students' desire and likelihood of going to college. The dependent variable in Column (1) and Column (2) is a dummy variable recording whether the student reported a scale 5 (1 is the lowest and 5 is the highest) on the the extent of how much they want to go to college (Wave II). The dependent variable in Column (3) and Column (4) is a dummy variable recording whether the student reported a scale 5 (1 is the lowest and 5 is the highest) on the likelihood that they will go to college (Wave II). Column (1) and (3) reports results for ego whose PVT score is below population median PVT score. Column (2) and (4) reports results for ego whose PVT score is above population median PVT score. The estimands are all ATT. Each row corresponds to a characterisation of the friendship, based on the character of the receiver and the sender. Receiver characteristics is shown before the underbar _, and sender characteristics is shown after. "F" and "M" are used to refer to the gender female and male respectively. "H" and "L" are used to refer to whether the individual is a high flyer or non-high flyer (low flyer) respectively. For example, "F_FL" means the linking effect is estimated for female link receivers and female non-high flyer link senders. The regressions reported in all columns include cohort dummies, whether the student was born in the US, their PVT score, whether their PVT score is above the population median PVT score, whether their mother's and father's highest degree is high school, some college, college, or post college, whether their mother's and father's highest education level is missing, the student's log family income, whether family is missing, the age of the student during Wave I, whether the student's mother and father were in the household, dummies for whether the student is black, hispanic, white, asian and indian. Standard errors are estimated with subsample bootstrapping with 900 subsamples drawn randomly. At each bootstrap, 90% of the individuals (nodes) within each school are sampled without replacement. *p<0.1; **p<0.05; ***p<0.01

Chapter 3

Gender difference in preference for competition

3.1 Introduction

Competition exists everywhere in our daily life. Students compete in exams, athletes compete in tournaments, and firms compete in the market. While some competitions are used to select the best performer, other competitions are designed solely to improve the performance of their participants. In fact, is it shown that competition improves performance even without tangible rewards. For example, Delfgaauw et al. (2013) finds that regardless of being rewarded or not for winning sales competition, workers in retail stores perform better if they are put in competition with each other. The assumption underlying the design of these competitions is that human beings are status-seeking, meaning they care about their relative position among their peers. As formally shown by Hopkins and Kornienko (2004), status-seeking people spend more on status-enhancing goods when there is competition for status than when there is no competition for status. Tincani (2018) translates this model to an education setting where students seek higher ranking in terms of academic performance within their classroom.

A common feature of the theoretical models in Hopkins and Kornienko (2004) and Tincani (2018) is that everyone has the same preference for status or ranking. However, from the empirical literature on competition, it is clear that there are gender differences in attitudes toward competition. The general consensus in the empirical literature, most of which use insights from lab experiments, is that women shy away from competition, largely due to their stronger risk aversion and more negative beliefs about their ability (Niederle and Vesterlund, 2011; Gneezy et al., 2003).

It has been difficult to measure the gender difference in attitudes towards competition that is purely due to preference because, in most settings, the competition is constructed in a way that involves risk and/or belief. In contrast, I am able to isolate the gender difference in preference for competition thanks to the special features of the Duolingo setting. During the weekly Duolingo leaderboard competition, learners could see the real-time performance of all competitors in their group. This eliminates the risk component in the competition. Moreover, all lessons and exercises on Duolingo are standardized. Each time a learner makes a mistake in the exercise, the correct answer is provided, and the same question will reappear at the end of the exercise. In order to complete an exercise, the learner only needs to correctly answer all the questions. Each exercise has about 10 questions and takes about 1-5 minutes to finish. Due to the known and fixed production technology that produces performance, it is also unlikely that there is a gender difference in beliefs about ability. Finally, the fact that learners are randomly assigned to competition groups eliminates any concern about self-selection.

In order to empirically examine the differences in preference for ranking between female and male Duolingo players, I introduce preference heterogeneity into the utility function specified in Hopkins and Kornienko (2004) and Tincani (2018), and I derive testable implications of the model. The empirical results show that female learners exhibit a stronger preference for ranking than male learners. This suggests that by designing competition in a way that does not involve risks, e.g. through real-time performance feedback, and makes it clear to participants how efforts are transformed to achievements, women could outperform men due to their stronger innate preference for ranking. This is in line with the findings in Alan and Ertac (2019), which suggest the gender difference in willingness to compete disappears when students are introduced a worldview that emphasizes the importance of effort.

Section 3.2 introduces the institutional setting of the Duolingo leaderboard competition and details the data collection procedure. Section 3.3 develops a theoretical model of group competition that incorporates gender differences in preference for ranking. Section 3.4 introduces the empirical strategy. Section 3.5 discusses the empirical results. Section 3.6 concludes and suggests future improvements.

3.2 Duolingo and data collection

Duolingo is one of the most popular online language learning platforms used by more than 300 million users for second language learning. All contents on Duolingo are freely available to its users.[1] It is accessible both via web (http://www.duolingo.com) and mobile apps. Duolingo divides learning targets into skills (e.g. Food, Plural, etc). Each skill is composed of several lessons that include tasks such as completing the sentence, translation, speaking, and listening. It usually takes between 1-5 minutes to complete one lesson. Users earn experience points (XP) for completing lessons.

Since the year 2019, Duolingo introduced a new feature called leaderboard, where users could compete with other people in the same league based on their XP. Every Monday at 12am UTC, leaderboards are reset, and 30 people who complete their first lesson at around the same time after the reset are paired to be in the same league.[2] Users could opt out of the leaderboard competition by setting their account as private.

Only XP earned during that week counts for the leaderboard competition. During the week, users could see the rank of everyone else in the same league, as well as how many XP they have. There are 10 leagues level, starting from Bronze to Diamond. At the end of the week, the top 10 will advance to the next league level, while the bottom 5 will be demoted to the previous league level.[3] The rest remains at the same league level.[4] Users who finish in the top 3 of their league will get some *lingots*, the Duolingo currency that can be used to

[1] Users could pay a monthly fee to become a premium member. Among other things, premium users do not see ads on the platform and can access the content offline. However, premium users and non-premium users have the same access to all Duolingo content.

[2] Note that users learning different languages could be paired into the same league as long as they completed their first lesson at around the same time.

[3] There are some exceptions. For the Bronze league, the top 20 will be promoted to the next league, Silver. For the Silver league, the top 15 will be promoted to Gold. For the Bronze league, no one will be demoted as there is no league lower.

[4] You will not be demoted if you don't enter the leaderboard competition during the week. That is if you don't do any lessons during that week.

purchase certain features, such as streak freeze.[5] In general, users have more lingots than they could spend as it is easy to earn them in many ways. In addition, if you finish as top 1 in any league level, you will earn the badge *Winner*, and if you finish as top 1 in the Diamond league, you will earn the badge *Legendary*. Other users are able to see these badges if they go to your profile page. In conclusion, even though no tangible reward or punishment results from the leaderboard competition, users have incentives to perform well to gain recognition from others as well as to fulfill their need to succeed.

For data collection, I created hundreds of Duolingo accounts, all controlled exclusively by me. I call these accounts the *seed players*. All seed players are learning the Italian language with English as their "mother tongue". The nicknames of the seed accounts are randomly selected from the five most common female names and the five most common male names.

In order to identify all the learners in the same group during a certain week, it is necessary to have at least one seed player enter that group. seed players enter the leaderboard competition sequentially at the beginning of the week. Every seed player will only do 2 lessons when they enter their group, earning about 25-30 XP, and do nothing else during the rest of the week.[6] The order of the seed accounts entering the competition is random. Because within a certain time interval, there could be more or fewer real learners entering the leaderboard, sometimes two or more seed accounts enter the same leaderboard competition. Once a seed player enters a group, the profiles of the learners in the same group will be scraped. After the competition ends, learner data is scraped again in order to obtain their learning activities during the competition week. The gender of the learners is inferred from their Duolingo username, which is often variation of their real name.

The current data is collected during two waves. The first wave contains the competition data for the week of 29 May -5 June 2022. These are 363 groups, all of which are in the league level "Gold". The second wave contains the competition data for the week of 17 July -24 July 2022. In the second wave, some groups are in the league level "Gold" (326 groups), and some groups are in the league level "Sapphire" (259 groups), which is one level above

[5] Users have an n-day streak if they have non-zero daily XP for n consecutive days. If a user is inactive for 24 hours, then his/her n-day streak is lost, a new streak will start once he/her starts earning XP again. The streak freeze allows users to not lose the n-day streak if they are inactive for a day.

[6] There are some exceptions where seed players do more than 2 lessons due to technical issues.

"Gold".

Table 3.1 shows the summary statistics of the seed players, and Table 3.2 shows the summary statistics of the real active learners. "XP past 7d" gives the total number of XP earned 7 days prior to the start of the corresponding competition. For learners with non-zero "XP past 7d", this statistic gives the total XP earned during their last competition week. "Total crowns" gives the total number of crowns the learner earned until the moment the learner data is scraped.[7] "Total XP" refers to the total XP earned until the current competition started.

None of the seed players had any activities during the last 7 days prior to entering the competition, as reflected in Table 3.1. The seed players in the second wave on average have more XP and more crowns than in the first wave. This is natural because the seed players who entered the competition during the first wave only did 2 lessons during the entire week and therefore are demoted to the lower "Silver" league. After the first wave of competition ended, these seed players are put into another competition in order to bring them back up to the "Gold" league.

Looking at Table 3.2, we see that the active learners in Wave II Gold league on average earned more than twice as many XPs during the days prior to the recorded competition than active learners in Wave I Gold league did. They also have more than 10 percent more total crowns and total XPs. This difference could be due to a seasonal effect: learners might have more time and be more active during July than during May due to holidays. The active learners in the Sapphire league on average have higher statistics than the active learners in the Gold league. This should not come as surprise either since Sapphire is a level higher than the Gold league, and in general, learners need to perform better to be in the higher league.

[7] Note that the time a learner's data is scraped is later than the time that learner started the competition. This means some of the crowns might have been earned after the competition started. Therefore, strictly speaking, the total number of crowns is not a pre-treatment variable. However, since the time interval between a learner's start of the competition and the time that her data is scraped is usually short, making it difficult to earn a lot of crowns within this period, I still use it as a pre-treatment variable.

Table 3.1: Descriptive statistics for seed players

	Mean	SD	Median
Wave I Gold (363 groups)			
XP past 7d	0.00	0.00	0.00
total crowns	14.15	2.23	15.00
total XP	758.15	110.46	771.00
Wave II Gold (326 groups)			
XP past 7d	0.00	0.00	0.00
total crowns	16.61	1.74	17.00
total XP	918.10	137.90	898.50
Wave II Sapphire (259 groups)			
XP past 7d	0.00	0.00	0.00
total crowns	16.78	1.51	17.00
total XP	919.97	126.28	909.00

Table 3.2: Descriptive statistics for active players

	Mean	SD	Median
Wave I Gold (363 groups)			
Female	0.54	-	-
XP past 7d	33.30	56.87	29.00
Total crowns	95.34	125.47	53.00
Total XP	7644.82	11128.89	3798.00
Wave II Gold (326 groups)			
Female	0.52	-	-
XP past 7d	85.43	55.23	92.00
Total crowns	105.74	152.12	51.00
Total XP	8994.71	14322.41	3740.00
Wave II Sapphire (259 groups)			
Female	0.52	-	-
XP past 7d	88.91	47.92	103.00
Total crowns	141.95	163.28	87.00
Total XP	13021.82	15334.62	7501.00

3.3 Conceptual framework

A general model of group competition in classrooms is considered in Tincani (2018), which is a derivation from the model of status-seeking by Hopkins and Kornienko (2004). First, a simple model of group competition directly adapted from Tincani (2018) is described in Section 3.3.1. Then in Section 3.3.2, I allow heterogeneous preferences for ranking as an extension of the baseline model, allowing the study of gender differences in preference for competition.

3.3.1 A model of group competition

The level of effort a learner exerts, denoted as e, can be chosen by the learner, and it has a positive effect on their learning outcome, denoted as y and measured by cumulative XP earned during some time period. The effort is costly, and learners differ on this cost, denoted as q. q is a function of e and learner type c, $q(e, c)$. In the case of Duolingo, c could represent time constraints, language learning talent, existing language knowledge, the language itself, how different the language is from the mother tongue, etc.[8] The model imposes that q decreases in c, so higher type c means lower marginal cost of effort. Learners compete within their group, which consists of 30 learners. Learners gain utility from improvement in their language skills, as well as their ranking within their leaderboard group. Additionally, utility decreases with the cost of effort. In the baseline model, type c is the only source of heterogeneity.

Formally denote utility as $U(y, q)$. It has two components, the utility that only depends on own learning outcome y and cost of effort q, denoted as $V(y, q)$, and the utility that depends on the relative ranking R of the learner, $S(R)$. As conventional in the study of status-seeking, ranking R_i is represented by the proportion of learners whose outcome y is lower than y_i. This is $F_Y(y_i)$, where $F_Y(\cdot)$ is the c.d.f of y. The overall utility is given by $U(y, q) = V(y, q)(S(F_Y(y)) + \phi)$, where $\phi > 0$ is to make sure that even the learner with the lowest rank still gets positive utility from their absolute level of learning outcome. Furthermore, the following assumptions on the utility function are made:

[8] For an individual with many other time-consuming commitments will face a trade-off between learning a language and doing other things. Therefore the cost related to time constraints is an opportunity cost.

Assumption 1.

1.1 $y(e) = \alpha e$. The learning outcome y measured by XP is a deterministic linear function of effort, and with zero effort, learners earn zero XP.[9] Because of the deterministic linear relationship between y and e, ranking $F_Y(y)$ can be equivalently expressed as $F_E(e)$ where $F_E(\cdot)$ is the c.d.f of effort.

1.2 $V(y, q) = k_1 y - k_2 q - k_3 yq$ where $k_1, k_2 > 0, k_3 \geq 0$. That is, the utility from own achievement is linearly increasing in the learning outcome, linearly decreasing in the effort cost, and the marginal utility from the learning outcome is weakly lower at a higher cost.

Learner type c is private information, but the distribution of c, $G(c)$, is common knowledge. Learners choose their own effort to maximize their overall utility by solving the following first order condition:

$$\alpha V_1 + \frac{V}{F_E(e) + \phi} f_E(e) = -V_2 \frac{\partial q}{\partial e} \tag{3.1}$$

where $V_1 = \frac{\partial V}{\partial y}$, $V_2 = \frac{\partial V}{\partial q}$, $f_E(e)$ is the p.d.f of effort e. One important result from Tincani (2018) and Hopkins and Kornienko (2004) is that the symmetric equilibrium strategy $e(c)$ is a strictly increasing function of c. This means the first order condition can also be expressed as

$$\alpha V_1 + \frac{V}{G_(c(e)) + \phi} g(c(e))c'(e) = -V_2 \frac{\partial q}{\partial e} \tag{3.2}$$

where $c(e)$ is an increasing function mapping the equilibrium effort to type. It is useful to note that in equilibrium, learners always choose a higher level of effort than they would in the situation of no rank concerns.

Another important result from Tincani (2018) and Hopkins and Kornienko (2004) concerns comparative statistics for any two groups A and B with distributions of type c given

[9] In Tincani (2018), α and u are a function of group mean type to incorporate technological spillover (Blume et al., 2015). α is allowed to differ by gender. This technological spillover should not exist in the case of Duolingo, because learners can not discuss or chat with each other, which means their own learning outcome is only a function of their own effort.

by $G_A(c)$ and $G_B(c)$. In particular, they show that if $G_B(c)$ is a mean-preserving spread of $G_A(c)$, for example, if $G_A(c)$ and $G_B(c)$ have the same mean but $G_B(c)$ has a larger variance than $G_A(c)$, the following results hold:

1. A learner with middle c exerts less effort and achieves lower performance in group B than in group A.

2. A learner with low $c$exerts more effort and achieves better performance in group B than in group A.

3. A learner with high c may perform better or worse in group B than in group A, depending on the relative strength of the preference for achievement versus rank.

The intuition is that when the type distribution of group B is a mean-preserving spread of that of group A, there are fewer learners with middle type, but more learners with high and low types in group B. This means for middle type learners, improving their rank is more difficult in group B, while for low type and high type learners, improving their rank is easier in group B. To see why, notice that if there are more people of similar type, then for the same increase in effort, it is easier to surpass more people and therefore improve one's rank by a high amount. For this reason, the low type learners will respond to an increase in the density of their type by exerting more effort, and the middle type learners will respond to a decrease in the density of their type by exerting more effort while being able to maintain their rank. For the high type learner, there are two forces that push toward the opposite direction. On the one hand, just like low type learners, the high type learners have an incentive to exert more effort because for the same amount of increase in effort, they are now able to surpass more people. On the other hand, since the middle type learners are exerting less effort, the high type learners also have the incentive to exert less effort while maintaining their rank. The final result depends on the relative strength of preference for higher ranks and the preference for the absolute level of achievement.

3.3.2 Incorporating heterogeneous preferences

The baseline model assumes that learners only differ in their type c, which is the determinant of how costly a unit effort is. Next, I will allow learners to have heterogeneous preferences for both ranking and absolute achievement based on their gender. I do this by assuming the utility function is instead given by (3.3)

Assumption 2 (Utility function with heterogeneous preferences)**.**

$$U_i(y, q; \gamma_i, \phi_i) = V(y, q)(\gamma_i S(F_Y(y)) + \phi_i) \quad \gamma_i = \{\gamma_F, \gamma_M\}, \phi_i = \{\phi_F, \phi_M\} \tag{3.3}$$

Wlog, suppose $\gamma_F > \gamma_M$, the utility function given by (3.3) means for any given level of achievement y and cost of effort q, therefore fixing the absolute utility V, an equal increase in the rank $F_Y(y)$ will give female learners more utility than the male learners. In other words, γ_i measures the preference for ranking. Similarly, wlog, suppose $\phi_F > \phi_M$, the utility function given by (3.3) means that for any given rank, an increase in the absolute utility gives more overall utility to female learners than male learners. That is, ϕ_i measures the strength of preference for the absolute achievement net of the cost of effort.

In order to analyze how heterogeneous preferences affect the equilibrium strategy, let us first rewrite the overall utility function for female and male learners separately.

Female learners:

$$U_F(y, q) = \phi_F V(y, q)(\frac{\gamma_F}{\phi_F} S(F_Y(y)) + 1) \tag{3.4}$$

Male learners:

$$U_M(y, q) = \phi_M V(y, q)(\frac{\gamma_M}{\phi_M} S(F_Y(y)) + 1) \tag{3.5}$$

An immediate observation from equations (3.4) and (3.5) is that preferences heterogeneity only affects the equilibrium strategy through $\frac{\gamma_M}{\phi_M}$. This is because when taking the FOCs, ϕ_F and ϕ_M in front of V will disappear. In other words, it is the relative strength of preference for ranking and absolute achievement that determines the equilibrium strategy. Therefore, it is only possible to identify $\frac{\gamma_F}{\phi_F}(\frac{\gamma_M}{\phi_M})$ as a whole, and not separately for γ_F and ϕ_F, nor γ_M and ϕ_M.

Denote $\theta_F = \frac{\gamma_F}{\phi_F}$ and $\theta_M = \frac{\gamma_M}{\phi_M}$. Because it is still true that for all learners, the equilibrium effort with ranking concern is still more than the optimal effort level without ranking concern, the equilibrium effort continues to be a compromise between decreasing utility from the absolute achievement and increasing utility from a higher ranking. A higher θ therefore represents a higher willingness to increase ranking at the cost of a lower absolute utility. Delaying rigorous proofs to the future and only resorting to the intuitions developed by Tincani (2018) and Hopkins and Kornienko (2004), next, I will give some conjectures on how the equilibrium strategies and the comparative statistics should look like.

Proposition 1. Under Assumptions 1 and 2,

1.1 The equilibrium effort is a function of both c and θ, denoted by $e(c, \theta)$.

1.2 Conditional on θ, the equilibrium effort is a strictly increasing function of c. That is, among female learners, the ones with higher c will exert more effort, and among male learners, the ones with higher c will exert more effort.

1.3 Conditional on c, the equilibrium effort is higher for the learners with higher θ.

The intuition for 1.3 is that the equilibrium effort with ranking concern must be higher than the equilibrium effort without the ranking concern, meaning that the equilibrium effort with ranking concern is sub-optimal (too high) when we only look at the direct utility from absolute achievement. Therefore, learners will only have the incentive to increase their effort if they can get extra utility from ranking. This extra utility is higher for learners with a stronger preference for ranking, hence their higher equilibrium effort.

In terms of comparative statistics, when we change the distribution of types within a group, the results from Tincani (2018) and Hopkins and Kornienko (2004) need to be modified to accommodate the heterogeneity of preferences. This is because the equilibrium effort is no longer only a function of type c, but is also dependent on gender, meaning the relative competitiveness between pairs of groups r and r' can no longer be characterized by $G_r(c)$ and $G_{r'}(c)$.[10]

[10] For example, let us assume $\theta_F > \theta_M$. Then in the case where $G_B(c)$ is a mean-preserving spread of $G_A(c)$, if group B also has more middle c female learners who have a stronger preference for ranking, a given middle c learner, if they were put in group B compared to if they were put in group A, could either face an either higher degree of competition when the effect of more female middle c learners dominates or face a lower degree of competition when the effect of less middle c learners dominates.

Assumption 3. Distribution of type in the population is independent of gender:

$$G(c_i|Female_i = 1) = G(c_i|Female_i = 0) = G(c_i) \tag{3.6}$$

Because type completely determined the effort cost for any given level of effort, Assumption 3 essentially assumes female and male learners have the same production technology in turning effort into achievement. This assumption could be violated in situations where female learners are better at completing the Duolingo exercises, for example, if female learners need less time to complete a standardized lesson. In the next stage of this project, I will test this assumption by inferring the task completion time from the data. I will also explore ways to relax this assumption.

Proposition 2 (Comparative statistics)**.** Under Assumptions 1, 2 and 3, for any pair of groups A and B where $G_B(c)$ is a mean preserving spread of $G_A(c)$, and the share of female learner are the same in both groups, the following is true:

2.1 Given learner gender, in expectation (over all such pairs of groups), a learner with middle c exerts less effort and achieves lower performance in group B than in group A; a learner with low c exerts more effort and achieves higher performance in group B than in group A; a learner with high c may perform better or worse in group B than in group A, depending on the value of θ.

2.2 Given learner type c, in expectation (over all such pairs of groups), the difference in a low or high (middle) type learner's performance is positive (negative) when moving from group B to group A if the learner has higher θ.

The intuition for Proposition 2.2 is that for any pair of groups A and B, learners with a higher relative preference for ranking θ will exert more effort and obtain more XP when they are placed in the group with higher density of similar type learners. If B is a mean-preserving spread of A, high type and low type learners will have a higher density of similar type learners if they were put in group B than group A, and middle type learners will have a lower density of similar type learners if they were put in group B than group A. This means high and low type learners with a higher relative preference for ranking will exert more effort than learners with the same type c but lower relative preference for ranking, while middle type learners with a higher relative preference for ranking will exert less effort than learners with the same type c but lower relative preference for ranking.

3.4 Empirical strategy

Proposition 1 and 2 suggest if we were to observe every learner's type c, the relationship between θ_F and θ_M could be empirically examined with two tests. The first one is a direct result of Proposition 1, and tests whether $y_{rF}(c) > y_{rM}(c)$. The second one comes from Proposition 2.2 and looks within pairs of groups where one is a mean-preserving spread of the other in terms of the distribution of c, whether the effect of higher density of similar types causes female learners to change their effort more than male learners. However, both of the two tests require conditioning on the unobserved type c. For this reason, I will first discuss how to measure the unobserved type (or substitute type, as defined later) in Section 3.4.1. Then I will propose two empirical tests in Section 3.4.2.

3.4.1 Conditioning on unobserved type c

In order to test the relationship between θ_F and θ_M, we need to condition on type c. However, type c is not observed by the econometrician. One way to deal with this problem is by assuming c is a function of some individual characteristics and estimate the parameters of this linear function by minimizing the distance between the predicted achievement y and the observed y (Tincani, 2018).

In this paper, I propose another solution by observing that instead of conditioning on c itself, it is sufficient to condition on any variable that is a monotone transformation of c. Recall that the equilibrium achievement of learner i is a function of their own type c, their preference parameter θ, and the distribution of types and gender in their group g in week t.

$$y_{it} = f(c_i, \theta_i, s_{it})$$

for some function f.

Assumption 4.

1. The distribution of types and gender in i's group affects learner i only through a demeaned summary statistic s_{it}, with $E(s_{it}) = 0$.

2. learner type c and relative preference for ranking θ do not change over time. This means the observed achievement of learner i during week t can be expressed as

$$y_{it} = f(c_i, \theta_i, s_{it})$$

   for some function f.

3. $f(c_i, \theta_i, s_{it})$ can be parameterized as the following linear function

$$y_{it} = m(c_i) + s_{it} + m(c_i) \times s_{it} + m(c_i) \times l(\theta_F) \times Female_i + m(c_i) \times l(\theta_M) \times (1 - Female_i) \tag{3.7}$$

$m(\cdot)$ is a strictly increasing function of c because the equilibrium effort, hence achievement, is proved to be a strictly increasing function of c, for any given θ and distribution of types and gender. $l(\cdot)$ is also an increasing function because everything else equal, the higher the relative preference for ranking, the higher the equilibrium effort and achievement. Parameterization as in equation (3.7) assumes that i) the effect of relative preference θ, which is determined by gender, is separable from the effect of group type and gender distribution s_{it}; ii) the effect of relative preference for ranking depends linearly on own type c_i; and iii) the effect of within-group distribution of types and gender depends on own type c_i linearly.

Now suppose after collecting leaderboard competition data during period $t = 0$, we follow all learners and collect data on their weekly achievements y_{it} for an additional T periods (weeks), and defined the average achievement for learner i across periods T as

$$\begin{aligned} \bar{y_i}^T &:= \frac{1}{T}\sum_{t=1}^{T} y_{it} \\ &= m(c_i) + \bar{s}_i + m(c_i)\bar{s}_i + m(c_i) \times l(\theta_F) \times Female_i + m(c_i) \times l(\theta_M) \times (1 - Female_i) \end{aligned} \tag{3.8}$$

where $\bar{s}_i := \frac{1}{T}\sum_{t=1}^{T} s_{it}$. As $T \to \infty$, $\bar{s}_i \to 0$, and because of randomization of leaderboard competition groups, s_{it} is independent of learner type c_i, therefore $m(c_i)\bar{s}_i \to 0$.[11] This means

$$\bar{y}_i := \lim_{T\to\infty} \bar{y}_i^T = m(c_i) + m(c_i) \times l(\theta_F) \times Female_i + m(c_i) \times l(\theta_M) \times (1 - Female_i) \quad (3.9)$$

Next define $\bar{y}_F := \mathbb{E}[\bar{y}_i | Female_i = 1]$ and $\bar{y}_M := \mathbb{E}[\bar{y}_i | Female_i = 0]$. Plugging in equation (3.9), we get

$$\begin{aligned} \bar{y}_F =& \mathbb{E}[m(c_i) | Female_i = 1] \\ &+ \mathbb{E}[m(c_i) \times l(\theta_F) \times Female_i | Female_i = 1] \\ &+ \mathbb{E}[m(c_i) \times l(\theta_M) \times (1 - Female_i) | Female_i = 1] \end{aligned}$$

Next notice that Assumption 3 means $\mathbb{E}[m(c_i)|Female_i = 1] = \mathbb{E}[m(c_i)|Female_i = 0] = \mu_c$, $\mathbb{E}[m(c_i) \times l(\theta_F) \times Female_i|Female_i = 1] = l(\theta_F)\mu_c$, and $\mathbb{E}[m(c_i) \times l(\theta_M) \times (1 - Female_i)|Female_i = 1] = 0$. That is,

$$\bar{y}_F = \mu_c + l(\theta_F)\mu_c$$

And similarly

$$\bar{y}_M = \mu_c + l(\theta_M)\mu_c$$

Therefore, $\frac{\bar{y}_F}{\bar{y}_M} = \frac{l(\theta_F)}{l(\theta_M)}$, and $\bar{y}_F - \bar{y}_M = (l(\theta_F) - l(\theta_M))\mu_c$. Plug this back in to equation (3.9),

[11] The fact that learners could be promoted/demoted to a different or remain at the same league level after every week's competition might complicate things. My conjecture is that learners will eventually have a stable range of league levels they compete at, so approximately, in the steady state, the type and gender distribution of their leaderboard group is orthogonal to their type. A proper way to account for this is left for future work.

we get

$$\begin{aligned}\bar{y}_i &= m(c_i) + m(c_i) \times (l(\theta_F) - l(\theta_M)) \times Female_i + m(c_i) \times l(\theta_M) \\ &= m(c_i)(1 + l(\theta_M) + \frac{\bar{y}_F - \bar{y}_M}{\mu_c} \times Female_i) \\ &= \frac{m(c_i)}{\mu_c}(\mu_c + \mu_c l(\theta_M) + (\bar{y}_F - \bar{y}_M) \times Female_i) \\ &= \frac{m(c_i)}{\mu_c}(\bar{y}_M + (\bar{y}_F - \bar{y}_M) \times Female_i)\end{aligned}$$

Rearrange we get

$$\tilde{c}_i := \frac{\bar{y}_i}{\bar{y}_M + (\bar{y}_F - \bar{y}_M) \times Female_i} = \frac{m(c_i)}{\mu_c} \tag{3.10}$$

This means $\tilde{c}_i$ is an increasing function of c_i, and we could use

$$\tilde{c}_i^T := \frac{\bar{y}_i^T}{\bar{y}_M^T + (\bar{y}_F^T - \bar{y}_M^T) \times Female_i} \tag{3.11}$$

as an estimator of $\tilde{c}_i$.

3.4.2 Testing gender difference in relative preference for ranking

A further assumption is needed to translate Proposition 1 and Proposition 2, which are based on the unobserved type c, to empirical tests based on the substitute type $\tilde{c}$.

Assumption 5. $m(\cdot)$ is linear.

With Assumption 5, it is easy to show that if $\mu_{\tilde{c}}^B = \mu_{\tilde{c}}^A$ and $\sigma_{\tilde{c}}^B > \sigma_{\tilde{c}}^A$, it is also true that $\mu_c^B = \mu_c^A$ and $\sigma_c^B > \sigma_c^A$. Let $h_{F(M)}^r(\tilde{c})$ denote the achievement of a female (male) learner with substitute type $\tilde{c}$ when she (he) is put in group r. This means if $\theta_F > \theta_M$, $h_F^B(\tilde{c}) - h_F^A(\tilde{c}) > h_M^B(\tilde{c}) - h_M^A(\tilde{c})$ for high and low levels of c, and $h_F^B(\tilde{c}) - h_F^A(\tilde{c}) < h_M^B(\tilde{c}) - h_M^A(\tilde{c})$ for middle level of $\tilde{c}$.

Empirically, similar to Tincani (2018), we could estimate the functiona $h_F^r(\tilde{c})$ and $h_F^r(\tilde{c})$

for any group r non-parametrically:

$$\hat{h}^r_F(\tilde{c}) = \frac{\sum_{i\in r, Female_i=1} w_i K(\frac{c_i - c}{b}) y_i}{\sum_{i\in r, Female_i=1} w_i K(\frac{c_i - c}{b})} \tag{3.12}$$

$$\hat{h}^r_M(\tilde{c}) = \frac{\sum_{i\in r, Female_i=0} w_i K(\frac{c_i - c}{b}) y_i}{\sum_{i\in r, Female_i=0} w_i K(\frac{c_i - c}{b})} \tag{3.13}$$

where a standard normal Kernel $K(\cdot)$, the optimal bandwidth $b = 1.06\hat{\sigma}_c n^{-1/5}$ that minimizes the Approximated Mean Integrated Squared Error are used. The weights w_i are such that only observations i where the p.d.f. of c at c_i exceeds a small positive number are kept.

Next, we could again non-parametrically estimate the effect of an increase in group (substitute) type dispersion on the performance of a learner of any (substitute) type for female and male learners separately. This is done by finding all pairs of groups where one has higher variance than the other and calculating the difference in their expected performance for every $\tilde{c}$ (c) and gender, where the difference is weighted to make sure higher weights are given to pairs of groups with more similar mean (substitute) type.

$$\hat{\mathit{\Delta}}_F(\tilde{c}) = \frac{\sum_{r=1}\sum_{r'=r+1} \omega_{rr'}(\hat{h}^r_F(\tilde{c}) - \hat{h}^{r'}_F(\tilde{c}))}{\sum_{r=1}\sum_{r'=r+1} \omega_{rr'}} \tag{3.14}$$

$$\hat{\mathit{\Delta}}_M(\tilde{c}) = \frac{\sum_{r=1}\sum_{r'=r+1} \omega_{rr'}(\hat{h}^r_M(\tilde{c}) - \hat{h}^{r'}_M(\tilde{c}))}{\sum_{r=1}\sum_{r'=r+1} \omega_{rr'}} \tag{3.15}$$

where $\omega_{rr'} = \mathbb{1}\{\sigma_r > \sigma_{r'}\}\frac{1}{b_\mu}K(\frac{\mu_r - \mu_{r'}}{b_\mu})$.

Finally, the estimated difference between female and male learners in their response to increased type dispersion is given by

$$\hat{\delta}_{F,M}(\tilde{c}) = \hat{\mathit{\Delta}}_F(\tilde{c}) - \hat{\mathit{\Delta}}_M(\tilde{c}) \tag{3.16}$$

Again, according to the prediction of the model, if female learners have a higher relative preference for ranking, $\delta_{F,M}(\tilde{c})$ should be positive for high and low levels of $\tilde{c}$, and negative for middle level of $\tilde{c}$.

Additionally, as an implication of Proposition 1 is that if female learners have a stronger relative preference for ranking, $h^r_F(\tilde{c})]$ should be higher than $h^r_M(\tilde{c})]$ for all r and c. $\hat{\Psi}$ defined in (3.17) could be used to test this.

$$\hat{\Psi}_{F,M}(\tilde{c}) = \frac{1}{N_g} \sum_{r=1} (\hat{h}^r_F(\tilde{c}) - \hat{h}^r_M(\tilde{c})) \tag{3.17}$$

3.5 Empirical results

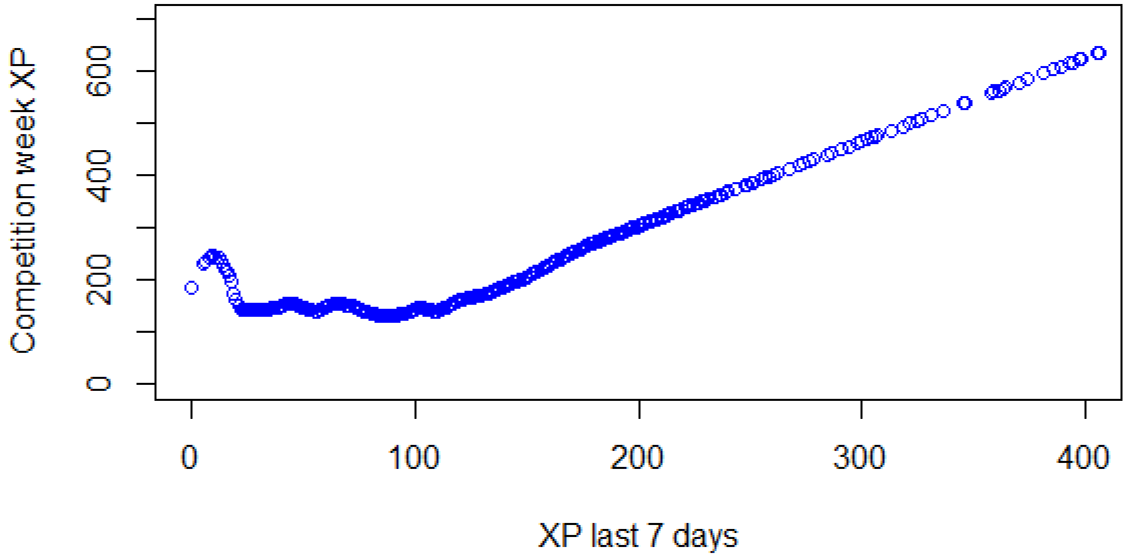
Analysis in Section 3.4 suggests the use of long-term average weekly XP to construct a substitute type. Unfortunately, by the time of this draft, the long-term performance data has not been collected. However, data on past performance is available. Particularly of interest is the XP earned during the 7 days (XP7d) before group competition data is collected.[12] Because this is a very short period, it is a noisy measure for the substitute type $\tilde{c}$. One main source of noise comes from the fact that the effect of group type and gender distribution cannot be averaged out by multiple periods since only one period is available, as can be seen from equations (3.8) and (3.9). Another source of noise comes from the fact that many learners did not participate in the leaderboard competition during the past week because they didn't do any lessons. For these learners, we have no direct information to infer their substitute type.

One way to examine the suitability of using XP7d as a substitute type is by looking at how well it predicts the competition week performance, as the model predicted higher type learners will in general exert more effort and accumulate more XP. Figure 3.1 shows the locally smoothed relationship between the average competition week XP and XP7d. As can

[12] Other measures are also available, for example, the total and average accumulated XP since the learner's Duolingo account registration and the total number of crowns the learner has earned. However, as Table 3.3 shows, these measures have almost 0 correlation with the competition week XP, hence bad measures for the substitute type. To see why this is the case, notice that a low type learner could have higher total XP and crowns if they have been active on Duolingo for a much longer period than a high type learner. In terms of average XP, notice that a high type learner could have a lower average XP if, for the majority of the weeks since their registration, they were inactive hence not participating in the competition at all. None of these will be a problem when we have data on learners' XP during each active week following the competition week.

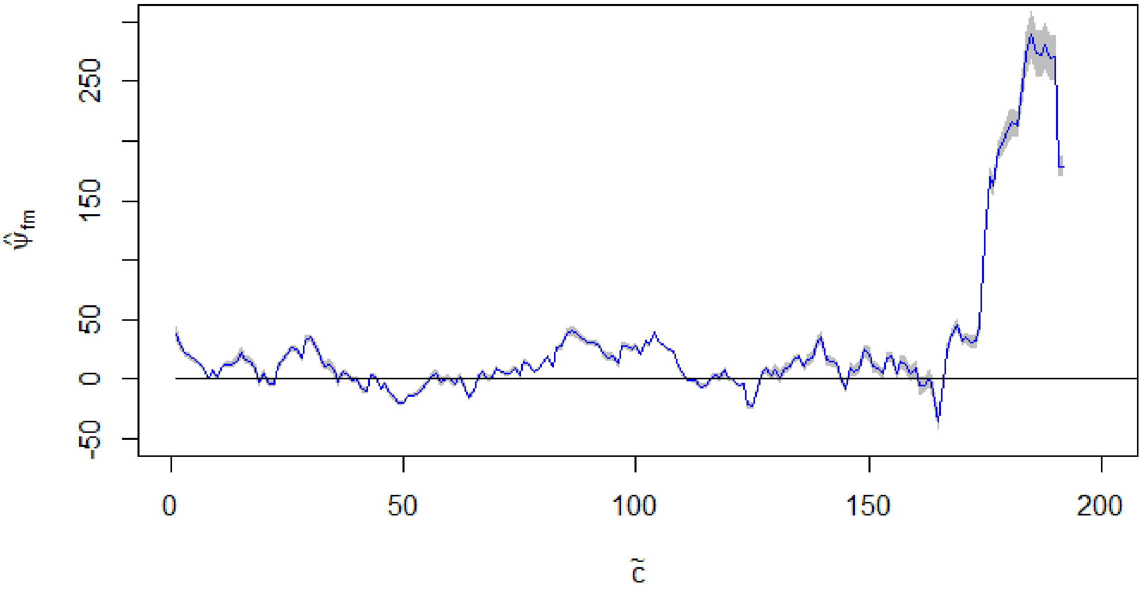
be seen in the figure, there is a positive relationship between past and current performance for learners with roughly more than 120 XP7d. However, no clear relationship is found for learners with less than 120 XP7d. This could be because when learners do not do a lot of lessons during a particular week, it is not because they have a lower type, but rather because of idiosyncratic negative productivity shocks. For example, a learner could learn less than their type predicted level because an unusually high workload at their job or a family emergency prevented them to spend time learning on Duolingo. Notice also that learners with 0 XP7d have an expected competition week XP of around 190 XP, similar to those with 150 XP7d. This is precisely because learners with 0 XP7d are, in fact, inactive during that week, and their substitute type cannot be directly inferred. To partly deal with this problem, I substitute the XP7d of the inactive learners with the average XP7d if the active learners with similar characteristics in terms of the number of latest streaks, the data collection wave, the total accumulated XP, and the number of crowns earned since Duolingo account registration, and their league level.

Figure 3.1: Smoothed relationship between average XP7d days and competition week XP



After constructing $\tilde{c}$ with XP7d, I conduct the two tests discussed in the previous section.

Figure 3.2: Expected Competition XP: difference between female and male learners

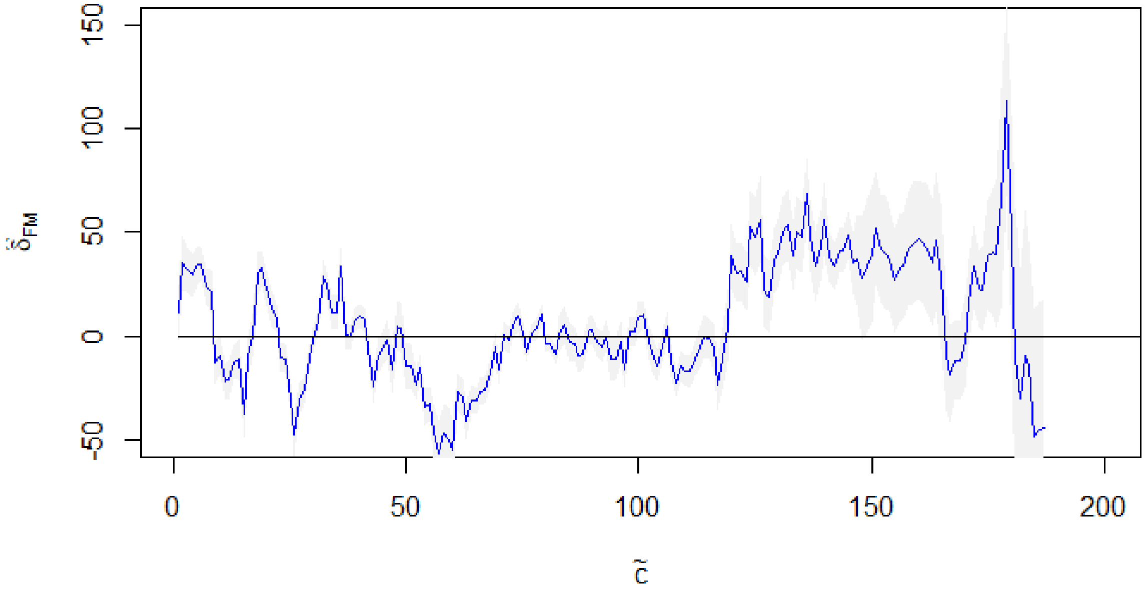


Note: Standard errors are calculated by bootstrapping groups at random for 1000 times. The confidence band is at the 1% level.

The first result is from the test based on Proposition 1. Figure 3.2 shows that in general female males achieve higher performance than male learners, conditional on their c. Note that at the right tail of the distribution of $\tilde{c}$, the difference first becomes much higher, then completely disappears due to missing data. This is likely to be caused by the fact that very few learners' substitute type $\tilde{c}$ are higher than 170 (less than 1%), as can be seen from Figure 3.5. Figure 3.2 provides the first piece of evidence that female learners have a higher relative preference for ranking.

The second result is from the test based on Proposition 2. Figure 3.3 shows that for learners with $\tilde{c}$ between around 120 to 170, female learners exert more effort and achieve higher performance when in groups with higher type variance but the same mean type. Referring to Figure 3.5, we can see that $\tilde{c}$ between around 120 to 170 represents the top 85-99% of the $\tilde{c}$ distribution, suggesting they are the high type learners. This means for high type learners, females have a higher preference for ranking than males. For the low and

Figure 3.3: Difference in female and male response to type dispersion



Note: Standard errors are calculated by bootstrapping groups at random for 1000 times. The confidence band is at the 10% level.

middle levels $\tilde{c}$, the result is less clear. Figure 3.1 suggests that this could be because XP7d is a bad measure to construct $\tilde{c}$ for learners with less than 120 XP7d.

3.6 Conclusion

This paper develops a model of group competition with heterogeneous preferences for ranking. Using web-scrapped data from Duolingo, I find evidence suggesting that women have a stronger preference for ranking than men. However, both the theoretical and the empirical analysis need to be dealt with more carefully. In terms of the theoretical model, rigorous proofs of the propositions, instead of just intuitions, will be developed. In terms of the empirical analysis, the main focus will be on collecting performance data for a much longer period in order to construct a better measure of substitute type $\tilde{c}$.

3.7 Appendix

Table 3.3: Correlation between pre-treatment variables and competition week XP

	XP past 7d	Total past XP	Average past XP	Total crowns
Pearson	0.09084***	-0.0571***	0.00874	-0.04115***
Kendall	0.15949***	-0.00204	0.03448***	-0.00436

Figure 3.4: Quantile of $\tilde{c}$ by gender: from 0 to 99%

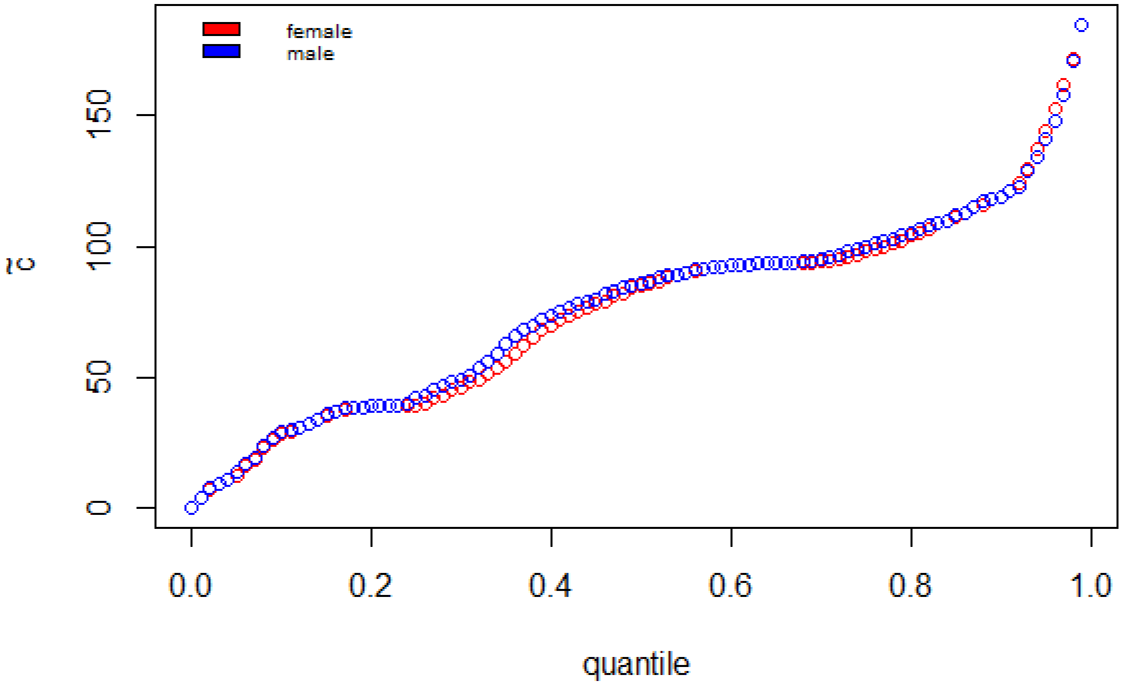


Figure 3.5: Quantile of $\tilde{c}$ by gender: from 99% to 100%

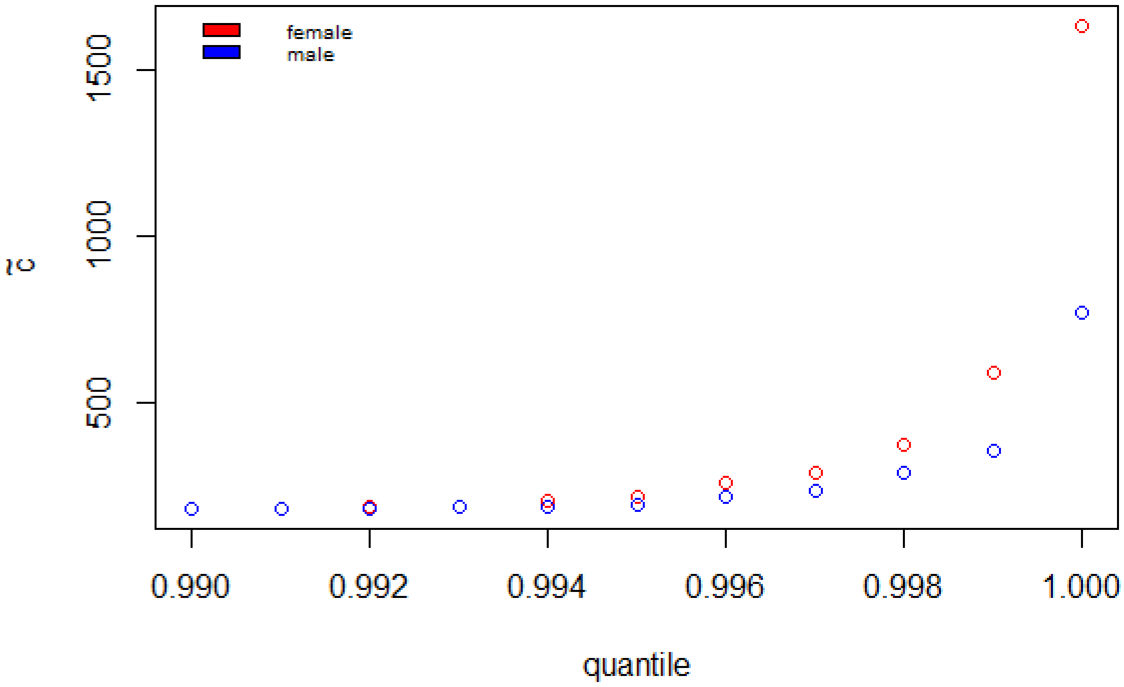
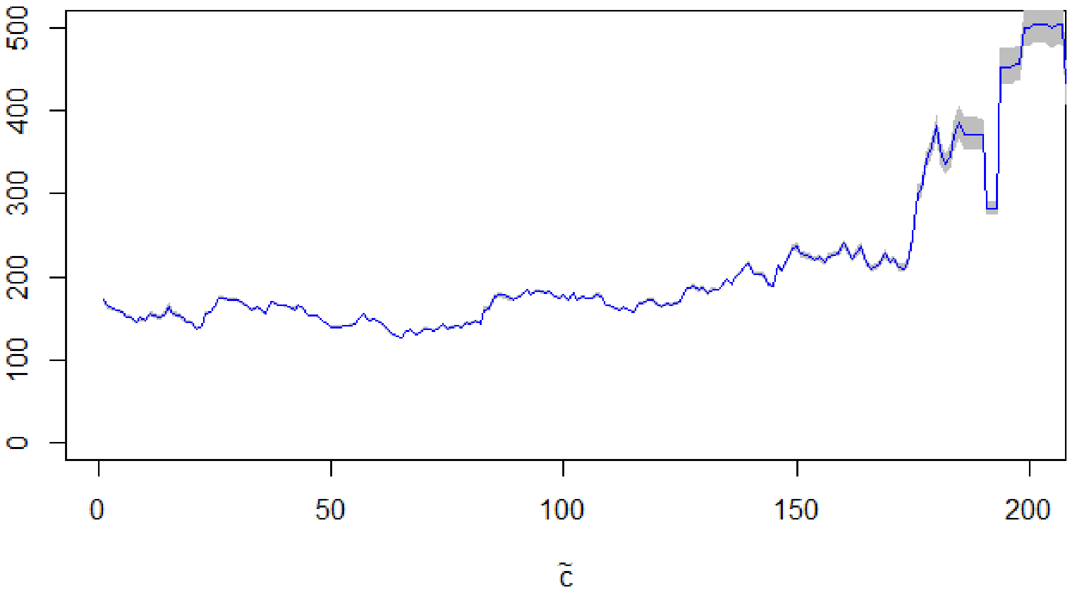
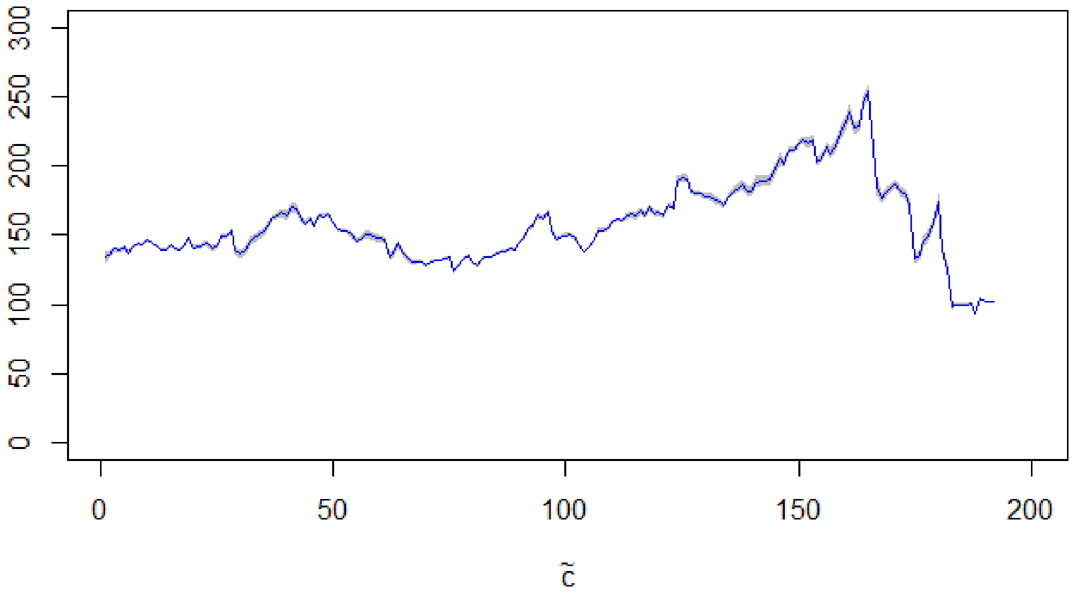


Figure 3.6: Expected Competition XP: female



Note: Standard errors are calculated by bootstrapping groups at random for 1000 times. The confidence band is at the 1% level.

Figure 3.7: Expected Competition XP: male



Note: Standard errors are calculated by bootstrapping groups at random for 1000 times. The confidence band is at the 1% level.

Chapter 4

References

Bibliography

Alan, Sule and Seda Ertac, "Mitigating the Gender Gap in the Willingness to Compete: Evidence from a Randomized Field Experiment," *Journal of the European Economic Association*, August 2019, *17* (4), 1147–1185.

Arduini, Tiziano, Eleonora Patacchini, and Edoardo Rainone, "Parametric and Semiparametric IV Estimation of Network Models with Selectivity," Technical Report 1509, Einaudi Institute for Economics and Finance (EIEF) October 2015. Publication Title: EIEF Working Papers Series.

Auerbach, Eric, "Identification and Estimation of a Partially Linear Regression Model Using Network Data," *Econometrica*, 2022, *90* (1), 347–365. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19794.

Badev, Anton, "Nash Equilibria on (Un)Stable Networks," *Econometrica*, 2021, *89* (3), 1179–1206. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12576.

Basse, Guillaume, Peng Ding, Avi Feller, and Panos Toulis, "Randomization tests for peer effects in group formation experiments," *arXiv:1904.02308 [stat]*, April 2019. arXiv: 1904.02308.

Bifulco, Robert, Jason M. Fletcher, Sun Jung Oh, and Stephen L. Ross, "Do high school peers have persistent effects on college attainment and other life outcomes?," *Labour Economics*, August 2014, *29*, 83–90.

Blume, Lawrence E, William A Brock, Steven N Durlauf, and Rajshri Jayaraman, "Linear Social Interactions Models," *Journal of Political Economy*, 2015, *123* (2), 444–496.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, "Peer Effects in Networks: a Survey," *Annual Review of Economics*, 2020.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West, "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, May 2013, *81* (3), 855–882.

Cools, Angela, Raquel Fernández, and Eleonora Patacchini, "Girls, Boys, and High Achievers," Technical Report w25763, National Bureau of Economic Research, Cambridge, MA April 2019.

_ , _ , and _ , "The asymmetric gender effects of high flyers," *Labour Economics*, December 2022, *79*, 102287.

Crane, Harry, *Probabilistic foundations of statistical network analysis*, CRC Press, 2018.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin, "An Economic Model of Friendship: Homophily, Minorities, and Segregation," *Econometrica*, 2009, *77* (4), 1003–1045.

Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke, "Tournament Incentives in the Field: Gender Differences in the Workplace," *Journal of Labor Economics*, April 2013, *31* (2), 305–326. Publisher: The University of Chicago Press.

Diaconis, Persi and Svante Janson, "Graph limits and exchangeable random graphs," *arXiv preprint arXiv:0712.2749*, 2007.

Forastiere, Laura, Edoardo M. Airoldi, and Fabrizia Mealli, "Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks," *Journal of the American Statistical Association*, April 2021, *116* (534), 901–918.

Gagete-Miranda, Jessica, "An aspiring friend is a friend indeed: school peers and college aspirations in Brazil," *Manuscript*, 2020, p. 46.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, "Performance in Competitive Environments: Gender Differences," *The Quarterly Journal of Economics*, 2003, *118* (3), 1049–1074. Publisher: Oxford University Press.

Goldsmith-Pinkham, Paul and Guido W. Imbens, "Social Networks and the Identification of Peer Effects," *Journal of Business & Economic Statistics*, July 2013, *31* (3), 253–264.

Graham, Bryan S., "Network data," in "Handbook of Econometrics," Vol. 7, Elsevier, 2020, pp. 111–218.

Hernán, MA and JM Robins, *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC, 2020.

Hopkins, Ed and Tatiana Kornienko, "Running to Keep in the Same Place: Consumer Choice as a Game of Status," *THE AMERICAN ECONOMIC REVIEW*, 2004, *94* (4).

Hoxby, Caroline, "Peer effects in the classroom: Learning from gender and race variation," Technical Report, National Bureau of Economic Research 2000.

Hsieh, Chih-Sheng and Lung Fei Lee, "A Social Interactions Model with Endogenous Friendship Formation and Selectivity," *Journal of Applied Econometrics*, March 2016, *31* (2), 301–319. 00118.

Imai, Kosuke and Zhichao Jiang, "Discussion of "The Blessings of Multiple Causes" by Wang and Blei," October 2019. arXiv:1910.06991 [stat].

Imbens, Guido W. and Donald B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, April 2015.

Jochmans, Koen, "Peer effects and endogenous social interactions," *arXiv preprint arXiv:2008.07886*, 2020.

Johnsson, Ida and Hyungsik Roger Moon, "Estimation of Peer Effects in Endogenous Social Networks: Control Function Approach," *The Review of Economics and Statistics*, May 2021, *103* (2), 328–345.

Leung, Michael P., "Two-step estimation of network-formation models with incomplete information," *Journal of Econometrics*, September 2015, *188* (1), 182–195.

Li, Xinran, Peng Ding, Qian Lin, Dawei Yang, and Jun S. Liu, "Randomization Inference for Peer Effects," *Journal of the American Statistical Association*, October 2019, *114* (528), 1651–1664.

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

Niederle, Muriel and Lise Vesterlund, "Gender and Competition," *Annual Review of Economics*, September 2011, *3* (1), 601–630.

Olhede, Sofia C. and Patrick J. Wolfe, "Network histograms and universality of block-model approximation," *Proceedings of the National Academy of Sciences*, October 2014, *111* (41), 14722–14727. Publisher: Proceedings of the National Academy of Sciences.

Olivetti, Claudia, Eleonora Patacchini, and Yves Zenou, "Mothers, Peers, and Gender-Role Identity," *Journal of the European Economic Association*, February 2020, *18* (1), 266–301.

Sacerdote, Bruce, "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *The Quarterly Journal of Economics*, 2001, *116* (2), 681–704.

Sävje, Fredrik, Peter M Aronow, and Michael G Hudgens, "Average treatment effects in the presence of unknown interference," *The Annals of Statistics*, 2021, *49* (2), 673–701. Publisher: Institute of Mathematical Statistics.

Tincani, Michela M, "Heterogeneous Peer Effects in the Classroom," *Technical report*, 2018, *Working paper.*

Wang, Yixin and David M. Blei, "The Blessings of Multiple Causes," *Journal of the American Statistical Association*, October 2019, *114* (528), 1574–1596.

Zhang, Yuan, Elizaveta Levina, and Ji Zhu, "Estimating network edge probabilities by neighbourhood smoothing," *Biometrika*, December 2017, *104* (4), 771–783.