Three Essays on The Economics of Sexually Transmitted Infections

by

Yifan Kang

Thesis submitted to the University of Ottawa
In partial fulfillment of the requirements for the degree of:
Doctorate in Philosophy Economics

Department of Economics
Faculty of Social Sciences
University of Ottawa

Declaration

All three chapters of this thesis are self-containing research articles, which are co-authored with Professor Roland Pongou. I acknowledge his contribution, and his contribution is equal to my own.

I understand that my thesis may be made electronically available to the public.

Abstract

Sexually transmitted infections (STIs) have important consequences for individuals and society. Extensive literature has shown that various individual factors impact STIs. However, much less is known about their structural causes and how they affect sexual behavior and sexual network formation. In the first two chapters of this dissertation, I investigate how sex ratios and ethnic divisions affect sexual activity and the spread of STIs. In the third chapter, I analyze the effect of ethnic-based romantic homophily on STIs. I provide a brief description of each chapter below.

Chapter 1. We extend a theory of fidelity in a two-sided economy, and empirically discriminate between different rationales of sexual network formation by testing their implications for how sex ratios affect sexual activity, relationship stability, and the spread of sexually transmitted diseases in men versus women. We use a unique individual-level dataset in combination with census data from England and Wales, a setting where adult women outnumber adult men. Exploiting variation in cohort/ethnicity/region-specific sex ratios as a quasi-natural experiment, we find that a decrease in sex ratio imbalance decreases sexual infidelity and the number of serial partners, and increases the likelihood of safe sex. This in turn reduces the likelihood of acquiring a range of sexually transmitted infections and diseases, including chlamydia, gonorrhoea, genital warts, and herpes. Consistent with the rationale underlying the formation of egalitarian (in)fidelity networks, the effects of the sex ratio on sexual activity are larger for men compared to women, while its effects on sexual diseases are larger for women compared to men. The causality of these effects is established using classical and recent instrumental variables approaches and various robustness checks. For falsification, we show that sex ratios have no impact on several

"atheoretical" health conditions, such as Parkinson's disease, chronic lung disease, heart attack, stroke, and diabetes, which do not arise from sexual interactions.

Chapter 2. In societies organized around distinct racial and ethnic groups, limited communication between these groups might increase the search cost of sexual partners outside of own group, leading to racially segregated sexual networks and low risks of sexually transmitted diseases. At the same time, because sexual infidelity is more likely to be discovered when the cheated-upon individuals are co-ethnics, individuals in multiracial societies might find it cheaper to select sexual partners from diverse ethnic groups to hide their infidelity, which would lead to large interethnic sexual networks and high risks of STIs. We test these conflicting hypotheses by analyzing the causal effect of neighborhood-level racial diversity on sexual activity and STIs, using unique individual-level data from England, Wales, and Scotland. We find that individuals residing in multiracial neighborhoods have a greater number of sexual partners and are more likely to be infected with a wide range of STIs than their counterparts residing in more racially homogeneous neighborhoods. We use traditional and new instrumental variables approaches and various robustness checks to establish causation. Analyzing mechanisms, we find that within racially diverse neighborhoods, individuals who select sexual partners from diverse racial groups are more likely to be infected with STIs, holding the number of partners and other individual characteristics fixed. For falsification, we conduct a reverse-placebo test showing that racial diversity has no effect on a wide range of health conditions that do not arise from sexual interactions. From a policy perspective, our analysis implies that policies that promote racial and ethnic integration are likely to reduce unhealthy sexual activity and the spread of STIs in racially heterogeneous societies.

Chapter 3. A classical hypothesis in social network theory holds that central individuals are more likely to receive and spread information than are their peripheral counterparts. We test this hypothesis in the context of sexual networks and sexually transmitted diseases, using data from the United Kingdom. Romantic homophily - the tendency to select sexual partners with similar ethnic background - is used as a measure of the extent to which an individual is peripheral in a sexual network. We find that more sexually homophilous

individuals have a lower risk of sexual infections. This effect is causal, and larger for women, Whites, and heterosexuals.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Dr. Roland Pongou for his continuous support of my Ph.D. study and for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Without his precious support, it would not have been possible to complete this thesis.

Besides my adviser, I would like to thank the other members of my thesis committee: Dr. Ana C. Dammert and Dr. Myra A. Yazbeck, for their time, interest, and insightful comments and encouragement. I would also like to acknowledge their hard questions which incented me to widen my research from various perspectives. I would also like to thank Dr. Gilles Grenier and Dr. Daniel Rosenblum, who were respectively the internal examiner and the external examiner of my thesis.

Last but not least, I would like to thank my parents for all their love and encouragement. My parents allowed me to pursue my potential and have cherished every important moment and supported me whenever I needed it.

Table of Contents

List of Tables

List of Figures

General Introduction

Sexually transmitted infections and diseases spread essentially through sexual networks. The configuration of these networks plays an important role in disease prevalence. However, very little is known about the forces that influence the formation of these networks and shape their configuration, and how these networks in turn may affect their members very differently. This dissertation sheds light on these important questions.

In the first chapter, we develop a simple extension of the dynamic theory of fidelity networks and empirically test the different implications for how sex ratios affect sexual activity, relationship stability, and the spread of sexually transmitted infections (STIs), and how these effects vary by gender. The empirical findings allow discriminating between three rationales or processes of sexual network formation. Our basic extension of these processes sheds light on the different mechanisms through which the relative numbers of men and women in a sexual economy shape the incentive to engage in various forms of sexual relationships, which in turn affects the spread of STIs in the population.

Using data from the United Kingdom, a country where female adults outnumber male adults. We make use of variations in age cohort, ethnicity, and region-specific sex ratios as a quasi-natural experiment, and we find that an increase in the sex ratio decreases the number of concurrent, overlapping, and serial sexual partners and increases the likelihood of safe sex. This, in turn, decreases the likelihood of acquiring a range of STIs. Consistent with the rationale underlying the formation of egalitarian infidelity networks, we find that the sex ratio's effect on sexual behavior is larger for men than women. In comparison, its effect on STIs is larger for women compared to men. It is worth noting that even though a non-negligible fraction of individuals choose partners outside of their ethnic groups, our

estimates show that the cross-ethnic sex ratio has no effect on sexual behavior and sexual infections. This analysis allows us seeing that an individual is more responsive to his/her ethnic group's sexual economy than to the outside economy.

The study explored in the second chapter aims to widen the focus of structural factors on racial diversity and STI status to shed light on how the ethnically residential circumstances could impact an individual's sexual behavior and the likelihood of being infected with a sexually transmitted disease. Our paper studies the effect of racial and ethnic heterogeneity on sexual activity and the spread of STIs in the United Kingdom based on a model of sexual infidelity. Incomplete or partial integration among ethnic groups due to cultural barriers drives incentives for agents to optimize their sexual partners' number by choosing them from diverse ethnic groups. Ethnic heterogeneity, therefore, affects the spread of STIs by increasing multiple partnerships and by inducing large sexual networks that interconnect different ethnic groups. Using a rich dataset from the United Kingdom, we validate these predictions by showing that ethnic heterogeneity increases sexual activity and the likelihood of acquiring various STIs. Analyzing the mechanism, we find that individuals residing in more racially diverse neighborhoods are more likely to diversify sexual partners. Moreover, within racially diverse areas, individuals who diversify sexual partners are more likely to be infected with STIs. The findings suggest that these individuals are more central to the sexual network and are more likely to be infected and infect others.

The research question examined in the third chapter is associated with racial homophily in sexual networks and how it affects STIs. The structure of romantic and sexual networks reflects various sexual behaviors and is likely to affect the spread of STIs. Our analysis relies on a classical hypothesis in social network theory that central individuals in networks are more relevant and influential in learning, contagion, and diffusion processes than their peripheral counterparts. Therefore, individuals who are central to sexual networks may have an increased probability of getting infected with sexually transmitted diseases. In societies organized around distinct ethnic groups, agents who sexually connect several ethnic groups may be more central in sexual networks. Therefore, homophilous behavior may be more likely to characterize peripheral individuals. We test the causal effect of romantic

homophily on STIs using individual-level data from England, Scotland, and Wales. We find that romantic homophily decreases the probability of acquiring a wide range of STIs. For all infections, it is obvious that individuals who ethnically homogenize sexual partners are less likely to get infected with STIs compared to their counterparts who racially diversify sexual partners. These effects are larger for women, majority individuals, and heterosexuals.

Our empirical findings of three chapters make it possible to understand the dynamics of link formation, which allows us to gain some knowledge of the properties and the configuration of sexual networks that are likely to form in our environment, in spite of the fact that these networks are hard to observe in reality. But how are these three different factors related? Does each factor have an independent effect, or do they have a joint effect on the outcomes that we analyze? To find answers to these questions, we include all three determinants in one regression with a full set of individual and regional controls and fixed effects. Our findings are consistent with those documented in each of the chapters (See Table AA1).

Our findings offer another perspective on how balancing sex ratios, and increasing ethnic integration in an ethnically diverse society discourage risky sexual behavior and reduce STIs. From a policy perspective, our analysis implies that more research should be done to understand factors that affect male-female differences in mortality and hence determine sex ratios. Such research is likely to lead to scientific discoveries that can be used to reduce the excess mortality of males in societies where women outnumber men, and of females in societies where their men outnumber women. This, in turn, will reduce risky sexual behavior and gender differences in STIs prevalence. Our analysis also implies policies of ethnic integration should be implemented in ethnically heterogeneous societies. If implemented successfully, these policies will increase information circulation between different ethnic groups, which in turn will increase trust among different groups.

Chapter 1

Fidelity Networks, Sex Ratios, and Sexual Diseases: Theory and Evidence From the United Kingdom

1.1 Introduction

Sexual infidelity has important consequences on individuals and society. For example, it leads to relationship instability and divorce, and is a main factor in the spread of sexually transmitted diseases. In the economic literature, infidelity has been analyzed in terms of time allocation between work and leisure (Fair 1978), and in terms of strategic network formation among individuals who enjoy having multiple partners but do not want their partners to do likewise (Pongou and Serrano 2013).[1] The latter approach proposes a unified framework for analyzing link formation in a competitive sexual economy, and identifies social mechanisms that give rise to such sexual patterns as (serial) monogamy, polygamy, and relationship concurrency. It shows that these different patterns arise under different *rationales* or processes of sexual network formation, where each rationale corresponds to a different culture of gender relations. By showing theoretically how different gender norms determine sexual network configuration and affect the spread of sexually transmitted dis-

[1] See Pongou and Serrano (2016) for a generalization to other two-sided economies.

eases, this approach implies that prevention policies that do not account for *structural* factors will yield low payoffs.

In this paper, we develop a simple extension of the dynamic theory of fidelity networks proposed by Pongou and Serrano (2013, 2016), and empirically discriminate between the different rationales of sexual network formation proposed in this theory by testing their distinct implications for how sex ratios affect sexual activity, relationship stability, and the spread of sexually transmitted infections (STIs), and how these effects vary by gender. The empirical analysis is based on micro-level data from the United Kingdom, a country where adult women outnumber adult men. A main contribution of our paper is to analyze a rich and comprehensive set of outcomes describing various kinds of sexual behavior and diseases in a multicultural context. Owing to their distinct properties, these diseases are likely to have different effects on individual and aggregate economic productivity. Our study can therefore be viewed as documenting novel mechanisms through sex ratios affect the economy. Moreover, our study is the first to show how the effects of male scarcity on sexual behavior and STIs vary by gender. In so doing, it sheds light on an important structural cause of the gender gap in these outcomes. Furthermore, our empirical findings make it possible to understand the dynamics of link formation, which allows us to gain some knowledge of the properties and the configuration of sexual networks that are likely to form in our environment, in spite of the fact that these networks are hard to observe in reality.

The model of fidelity networks assumes a two-sided economy with men and women. Each individual desires sexual relationships with individuals of the other sex. However, having more than one partner at a time is viewed as an act of infidelity, which is punished if detected. Punishment of infidelity is assumed to be more severe for women than for men, which results in the optimal number of partners being greater for the latter. In this environment, they propose three dynamic *rationales* or *processes* of relationship formation, with each corresponding to a different assumption about individual rationality and gender norms. In each of these models, men and women meet in discrete time and decide to form new matches and/or sever existing ones, which is a type of dynamics which goes

back to Roth and Vande Vate (1990). Link severance is unilateral and link formation is bilateral. The first model, which they call the "unperturbed process", assumes that individuals are perfectly rational when selecting their sexual partners. They show that the *long-run predictions* -steady or recurrent states- of this process coincide with the set of pairwise stable networks, in which each woman obtains her optimal number of partners, whereas each man obtains at most his optimal number of partners. The assumption of perfect rationality underlying this model is relaxed by assuming that individuals may make mistakes in partner selection, resulting in the two other dynamic processes, qualified as "perturbed". Under the first perturbed process, agents tend to match with individuals who have fewer other partners. In the long run, this behavioral pattern leads to *egalitarian* pairwise stable networks, in which men and women have the same number of partners, which is the optimum for women. Serial monogamy arises as a special case of this process. The second perturbed process, on the contrary, describes a pattern of herd behavior, in which agents tend to match with individuals with more partners. Here, the number of partners that an individual has can be taken as a correct or an incorrect signal of personal quality. This leads to long-run equilibrium networks that are *anti-egalitarian* in the sense that a few men share all the women. Polygamy is a special case of this process. Each of the perturbed processes significantly refines the set of steady-state networks.

An assumption made in the original model of fidelity networks is that there is an equal number of men and women in the economy. For our purpose, we relax this assumption in order to study the effects of sex ratios on sexual behavior under each of the three behavioral patterns described above. We find that under the unperturbed process, increasing the sex ratio does not affect the configuration of long-run stable networks. Under the first perturbed process, increasing the sex ratio reduces the number of concurrent and serial partners, so that women now have the ability to switch from more connected men to match with men who have fewer other partners. It follows that increasing the sex ratio reduces infidelity more in men than in women, but reduces STI risk more in women than in men. Under the second perturbed process, increasing the sex ratio has no effect on relationship patterns.

We empirically test these predictions using data from the United Kingdom. Women outnumber men in this country, and sexually transmitted diseases represent an important public health concern. Indeed, in the 2010-12 British National Survey of Sexual Attitudes and Lifestyles, 166 out of every 1,000 women and 123 out of every 1,000 men reported having been diagnosed with a sexually transmitted infection (STI) in their lifetime. Moreover, according to data from Public Health England, the number of new STI cases in the UK exceeds 415,000 every year. It is shown that if the current rates of infection continue, total treatment costs will reach £363 million between 2015 and 2020 (Lucas 2015). In addition to these costs, STIs impose a significant burden on the health of the infected individuals and on their economic productivity.

We use the third British National Survey of Sexual Attitudes and Lifestyles conducted in 2010-2012 (Natsal3), in conjunction with aggregate data from censuses conducted in England and Wales in 2001 and 2011. The Natsal3 data contain socioeconomic and demographic information on males and females aged 16-74 years. Most importantly, interviewees also provided detailed information on their sexual behavior and on whether they were diagnosed with certain sexually transmitted infections at some point in their lifetime, within the past five years before the interview, or within the year preceding the interview. Infections included diseases like chlamydia, gonorrhoea, genital warts, and herpes. We generate a variable for cohort/ethnicity/region-specific sex ratios (see also Charles and Luoh (2010)). Because the average age gap between male and female partners in our dataset is about 5 years, following a similar approach as in Angrist (2002), our sex ratio variable is computed as the ratio of the number of men of a given age group to the number of women who are up to 5 years younger than men, within each region and ethnicity.[2] This variable is constructed using data from the 2001 and 2011 censuses.

We analyze the effects of sex ratios on a number of variables that measure sexual behavior. These variables comprise: (1) the number of sexual partners in the past five years; (2) the number of serial partners in the past five years; (3) a binary indicator for

[2]Our approach is consistent with Becker (1973), who argues that "a group of women must be matched with the men they are most likely to marry, e.g., college-educated women with college-educated men, or women aged 20-24 with men aged 25-29." (p. 838).

whether an individual had overlapping partners; (4) two binary indicators for whether an individual had any concurrent relationships in the past five years and in the last year, respectively; (5) the number of partners with whom an individual had unprotected sex in the past year; and (6) a binary indicator for whether an individual had unprotected sex in the past year. We also analyze the effects of sex ratios on several sexually transmitted diseases: (1) chlamydia; (2) gonorrhoea; (3) warts; (4) herpes; and (5) any STI.[3]

Estimating ordinary least squares (OLS) regressions, we find that an increase in sex ratios (measured using the 2011 census data) reduces the number of concurrent sexual partners, the number of serial partners, and the probability of overlapping between sexual partners. These effects are larger for men compared to women. We also find that increasing sex ratios decreases unprotected sex, which suggests that a reduced scarcity of males improves females' bargaining power, allowing them to negotiate safer sex. Consistent with these findings, an increase in sex ratios reduces the likelihood of acquiring any of the aforementioned sexual infections or diseases. For instance, a one-unit increase in the sex ratio decreases the likelihood of being infected with chlamydia by 7 percentage points for women and 3 percentage points for men. These findings imply that equilibrium sexual networks tend to be more *egalitarian*, which validates the first perturbed process in described above.

Our OLS regressions may suffer from potential endogeneity issues due to a possible measurement error associated with the sex ratio. For instance, members of some population subgroups may be undercounted in the census. Also, skewed sex ratios due to sex-selective abortion have been observed in some Asian groups in some Western countries (Dubuc and Coleman 2007, Abrevaya 2009). It is possible that the social norms that lead to sex-selective abortion in these groups in fact reflect an imbalance of power between genders that limits the ability of women to negotiate safe sex, or that allows men to be as unfaithful as they wish. We address these issues in our OLS regressions in several ways. First, we control for ethnic group fixed effect, which therefore accounts for ethnic-related sexual and gender norms. Second, we restrict our analyses to whites, in which it is believed that sex ratios (in the absence of a major war, as in recent years in the United Kingdom) are determined by

[3]Information was collected on other STIs such as syphilis, trichomonas, pubic lice or crabs, and hepatitis, but we do not analyze them because their prevalence was very low.

natural causes. This implies that variation in sex ratios among white individuals is likely to be random.[4] Also, because the overwhelming majority of the white individuals in our data are British, it is unlikely that they were undercounted in the census. It follows that the sex ratio is more likely to be accurate in this population and suffers less from endogeneity. These additional analyses do not change our conclusions regarding the effects of sex ratios on sexual behavior and on the likelihood of getting infected with any of the aforementioned STIs. Our conclusions are also robust to restricting the analysis to white individuals born more than twenty years after World War II. Interestingly, the effects of sex ratios among whites are larger than in the overall sample, which implies that the aforementioned first perturbed process leading to egalitarian sexual networks is more valid in this subgroup.

Third, in order to address any remaining endogeneity concerns, we use cohort/ ethnicity/ region-specific sex ratios computed from the 2001 census as an instrumental variable (IV) for sex ratios computed from the 2011 census.[5] Our assumption is that the 2001 sex ratio is correlated with the 2011 sex ratio but does not have a direct effect on sexual behavior in 2011. For most outcomes, our IV estimates are close to the OLS estimates.

Our findings are qualitatively robust to two other recent identification strategies that relax the exclusion restriction assumption made in the traditional instrumental variables approach. One strategy, developed by Nevo and Rosen (2012), is used to identify bounds around the true effect of the endogenous variable. This identification approach assumes that the instrumental variable could be correlated with the unobserved structural error term. Since it is difficult to prove empirically that an instrumental variable satisfies the exclusion restriction assumption, relaxing this assumption can be viewed as a conservative approach to identification.[6] The other strategy, developed by Lewbel (2012), relies on a

[4]In fact, it is well known that in the absence of gender bias in the allocation of resources likely to affect survival, males are more likely to die than females due to sex differences in preconception and biological factors. Therefore, the sex ratio is primarily determined by these factors, especially in the absence of a major war, mass migration, or gender-biased mass incarceration.

[5]This identification approach follows one similar to Alesina et al. (1999), who instruments current ethnic diversity index using past ethnic diversity index.

[6]Perhaps because of this reason, this methodology is being increasingly used in empirical research (see, e.g., Okoye and Pongou (2017) and Bhalotra and Clarke (2019)).

heteroscedastic covariance assumption. In our analysis, the estimates based on the Lewbel IV approach are bit larger and more precise than those of the standard IV approach. These findings strongly support our main conclusion that increased sex ratios decrease sexual infidelity, promote relationship stability, and decrease the spread of sexually transmitted infections, and these effects differ by gender.

As an additional check, we compute sex ratios across ethnic groups and estimate their effects on sexual behavior and STIs. We find that their effects are much smaller than the effects of within-ethnicity sex ratios described above, and are in fact close to zero. This implies that individuals respond more strongly to structural changes in the internal or within-group sexual economy. More importantly, these findings show that the effects of within-ethnicity sex ratios are not mechanical.

We conduct some falsification tests to rule out alternative interpretations of our findings. More precisely, we investigate the effect of sex ratios on several "atheoretical" health conditions, which are conditions unlikely to be driven by sexual interactions. Those conditions include Parkinson's disease, chronic lung disease, heart attack, stroke, and diabetes. Unlike sexually transmitted infections, these diseases are noncontagious. An economically significant effect of the sex ratio on any of these "atheoretical" outcomes would call into question our interpretation of the negative effect of the sex ratio on STIs. We find no such effect, which lends credibility to our interpretations and mechanisms.

The remainder of this paper is organized as follows. Section 2 situates our study within the extant literature. Section 3 presents our theoretical framework. Section 4 describes the data. Section 5 describes our identification strategies and presents the results as well as the sensitivity and robustness analyses we conduct. Section 6 concludes.

1.2 Contributions to the Related Literature

Our paper contributes to the growing literature that studies the endogenous formation of social networks and documents their effects on individual and social outcomes (see Vega-Redondo (2007) and Jackson (2010) for authoritative monographs). Within this

literature, Pongou and Serrano (2013) propose a model of (in)fidelity networks and use it to explain the gender gap in HIV/AIDS prevalence in a two-sided economy. Their theory is generalized in Pongou and Serrano (2016) to several other two-sided economies found in real life, showing how the perception and interpretation of the degree of connectivity by agents can lead either to market concentration or to market fragmentation in the long run.[7] To our knowledge, the literature on social networks has paid little attention to how exogenous changes in structural factors affect the formation and configuration of networks and their outcomes. Moreover, model predictions that rely on network configuration may be hard to test, especially if networks cannot be observed easily. Because of the sensitive nature of information on intimate behavior, data on sexual networks are very difficult to collect. However, in our paper, sexual networks have distinct equilibrium properties and predictions for how the sex ratio affects sexual behavior and sexually transmitted diseases, and testing these predictions allows us to gain some knowledge of the nature of these networks (e.g., *egalitarian* versus *anti-egalitarian*) in the context of the United Kingdom.

By documenting the effect of sex ratios on sexual behavior and sexual diseases, our paper also contributes to the theoretical and empirical literature that studies the determinants of family behavior using economic models. The pioneering works of Becker (1973, 1974) have spawned a fascinating theoretical and empirical literature on this topic (see Dupuy and Galichon (2014) and the references therein). Inspired by the original framework of Becker, Fair (1978) models infidelity in terms of time allocation between work and leisure, where there are two types of leisure activities, which are time spent with spouse and time spent with "paramour". Our paper differs from this literature in its scope and analyses. First, from a theoretical viewpoint, we follow a network approach to modeling the formation of links in a competitive two-sided economy. This approach makes it possible to accommodate a wide range of sexual behaviors (infidelity, serial relationships, overlapping relationships, and concurrency) in a unified framework. Second, our network approach enables the analysis of the effect of network configuration on the spread of sexually transmitted diseases by gender (see, e.g., Pongou and Serrano (2013)). We view this as a contribution in the sense that the aforementioned literature does not analyze sexually

[7]However, the implications of these studies have not been empirically tested.

transmitted diseases.[8]

Our paper is also related to the general literature on the economic and social impacts of sex ratios. Over the last three decades, numerous theoretical and empirical studies have explored the effects of sex ratios on a number of interesting outcomes including marriage outcomes (Rao 1993, Angrist 2002, Brainerd 2007, Charles and Luoh 2010, Abramitzky et al. 2011, Francis 2011, Lafortune 2013); pre-marital investment (Rao 1993, Angrist 2002, Nick and Walsh 2007, Abramitzky et al. 2011, Lafortune 2013); labor market outcomes (Edlund 2001, Chiappori et al. 2002, Amuedo-Dorantes and Grossbard 2007, Mechoulan 2011); the racial gap in HIV (Johnson and Raphael 2009); savings (Du and Wei 2010, Wei and Zhang 2011, Bhaskar and Hopkins 2016); and crime (Edlund et al. 2013, Cameron et al. 2017). In particular, Angrist (2002) finds that an increase in the sex ratio increases women's bargaining power, both within the household and in the marriage market. Also, exploiting the large shock WWI inflicted on French men, Abramitzky et al. (2011) find that men improved their marriage market outcomes in regions where they were scarcer. Johnson and Raphael (2009) find that higher imprisonment of black males explains the black-white gap in HIV among black women in the US. Similarly, Charles and Luoh (2010) find that higher male imprisonment decreases the likelihood of marriage for women, reduces the quality of their spouses, and increases their schooling and labor supply.

It follows that our paper clearly differs from the aforementioned studies in its scope and analysis. We analyze the effects of sex ratios on a different set of outcomes, namely sexual infidelity, relationship stability, safe sex, and STIs. We are also the first to show how these effects vary by gender. Our theoretical framework, which relies on endogenous network configuration, is also new. This network approach is necessary because it makes it possible to understand why sex ratios may affect men and women differently depending on the nature of the network formation process at work. Indeed, the analysis has allowed us to discriminate between different rationales of sexual network formation (Pongou and Serrano 2013, 2016).

[8]There is also a large empirical literature on marriage inspired by the pioneering work of Gale and Shapley (1962). The issues studied in this literature are different from ours. We focus on sexual behavior and STIs.

In addition, to our knowledge, thanks to the richness of our data, our paper is the first to consider a comprehensive set of sexual activities (*concurrent* versus *overlapping* versus *serial* relationships; *safe* versus *unsafe* sex) and STIs. In particular, STIs have different biological and transmission properties and characteristics, and may therefore have different consequences on economic productivity. Some STIs are caused by bacteria, others by viruses, and others by parasites. Also, bacterial STIs such as chlamydia and gonorrhoea are curable, whereas viral STIs such as genital warts and herpes, like HIV/AIDS, are still not curable at present. By making this distinction, we show that the effect of sex ratios might vary across diseases, therefore documenting novel mechanisms through which male scarcity might affect health and human capital accumulation in an economy.

Finally, we view our various identification strategies as complementing those employed so far in the literature documenting the impacts of sex ratios. Some of these studies have relied on abnormally high sex ratios caused by large aggregate shocks such as wars, mass migration, male incarceration, or the transatlantic slave trade. We instead focus on a population in which sex ratios have been relatively stable over the past five decades, and where small variations in sex ratios could be viewed as random. In order to lend more credence to the causal interpretation of the documented effects, we consider white individuals in a robustness analysis, among which sex ratios are highly likely to be naturally determined in the absence of major wars. We also employ traditional and new instrumental variables techniques for identification. Moreover, we show that our documented effects are not mechanical by finding that cross-ethnicity sex ratios computed similarly as within-ethnicity sex ratios have no effects on outcomes. Furthermore, in a falsification analysis, we show that the sex ratio has no effect on several health conditions that are not driven by sexual interactions, which lends credence to our proposed mechanisms and interpretation of results.

1.3 Theoretical Framework

In this section, we briefly present the dynamic theory of fidelity (or infidelity) networks developed by Pongou and Serrano (2013, 2016), and modify it for the purpose of studying the effect of the sex ratio on sexual behavior and STIs. This theory proposes three different rationales (or processes) of sexual network formation in an economy where it is assumed that there are an equal number of men and women. We relax this assumption and derive testable implications for how sex ratios might affect sexual infidelity as well as the spread of sexually transmitted infections. We do this using a simple example, as our goal is simply to apply the same rationales underlying link formation as developed in the original studies to a context where the number of men and women is not necessarily the same.

The theory of fidelity networks assumes a two-sided economy with men and women. Each agent desires sexual relationships with individuals of the other sex. However, having multiple partners is viewed as infidelity and is punished if detected. This tradeoff leads to each agent having a single-peaked utility function. Punishment of infidelity is more severe for women than for men, resulting in the optimal number of partners being greater for the latter. This theory develops three dynamic rationales or processes of network formation. We describe these processes below and derive implications for the effect of sex ratios on sexual infidelity under each of them.

Process $\mathbf{P}_0$ (Unperturbed Process). The unperturbed Markov process is modeled as a dynamic Markov matching process in which time is discrete. In each period, a man and a woman are randomly selected with positive probability and given the opportunity to terminate their relationship if they indeed have one in the current network, or to form a new relationship if they do not have any. Terminating a relationship is a unilateral decision whereas forming a relationship is a bilateral decision, as this requires two parties to agree. While entering a new relationship, each agent is allowed to sever any of the relationships in which he/she is involved within the current network. Under natural assumptions on the size of the economy, it is shown show that the l*ong-run predictions* -steady or recurrent states- of this process coincide with the set of networks in which each woman obtains her optimal number of partners, whereas each man obtains at most his optimal number of

partners.

As mentioned previously, this study assumes that there are an equal number of men and women in the economy. In order to study the effect of sex ratios on partnership formation, we will relax this assumption, which will allow us to determine how sex ratios affect sexual behavior and link formation. Consider the following example of an economy involving 2 men, m_1 and m_2, and 3 women, w_1, w_2 and w_3. Each man desires 2 female partners and each woman desires only 1 male partner.[9] The network g_0 in which m_1 is matched with w_1 and m_2 is matched with w_2 and w_3 is a steady state (or an equilibrium) network under the process $\mathbf{P}_0$. Now assume that a new m_3 enters the economy, increasing the male-to-female ratio of the economy. Under the process $\mathbf{P}_0$, the links in g_0 will not be affected by this change (see Figure 1.1).

The testable implication is that, under the process $\mathbf{P}_0$, increasing sex ratios does not affect infidelity among sexually active individuals, and as a result also does not have any effect on the likelihood of transmission of a sexually transmitted infection.

In order to improve the predictive power of the process $\mathbf{P}_0$, two perturbations of this process are analyzed. Each perturbation can be viewed as corresponding to a different culture of gender relations. While under $\mathbf{P}_0$, each agent only cares about the number of partners he/she has, under each of the perturbed processes, agents also care about the quality of their relationships.

Process P_1^ϵ (The First Perturbed Process). Time is discrete as in the unperturbed process. In each period, a man and a woman are selected at random with positive probability and given the opportunity to sever their link if they have a relationship or form a new relationship if they are not matched in the current network. It is assumed that if a relationship is mutually beneficial or if severing a link is beneficial to the severer, it occurs with probability 1. However, any action (severance or agreement to form a relationship) that makes its initiator worse off, which is a mistake, is taken with a small probability $\epsilon > 0$. Any action that leaves its initiator indifferent (such an action consists of breaking

[9]Note that in the general model in Pongou and Serrano (2013, 2016), a woman can have multiple partners. Assuming that a woman desires only 1 partner in this example simplifies the exposition.

an existing link with an agent k and forming a new link with another agent) is taken with probability $\epsilon^{f(1/s_k)}$, where s_k is the number of partners that agent k has in the current network, and the exponent is an increasing function mapping into (0,1). The interpretation is that $1/s_k$ measures the amount of time invested by k in each of his/her relationships, and $f(1/s_k)$measures the strength of each of these relationships. Since f is increasing, a relationship linking agent i to agent k is harder to break by i if k invests more time in it or equivalently, if k has fewer other partners.

It is shown that under the perturbed process P_1^ϵ, only egalitarian networks in which men and women have the same number of partners-the optimal number of partners for each woman-form in the long run, where the equilibrium concept used to capture rationality is that of stochastic stability (Kandori et al. 1993, Young 2001). This prediction is valid under the assumption that there are an equal number of men and women in the economy. But extending the finding to an economy with an unequal number of men and women would not be difficult. One can show that if the sex ratio deviates from 1, then the long-run networks in this economy are still the most egalitarian networks in the sense that (1) each agent will have at most his/her optimal number of partners; and (2) and the number of female (resp. male) partners obtained by each man (resp. woman) will not exceed that obtained by another man (resp. woman) by more than 1. In other words, individuals on the same side of the economy will tend to have the same number of partners of the opposite sex.

Now, in order to analyze how sex ratios affect sexual infidelity under the process P_1^ϵ, consider the same example analyzed previously. The network g_0 is still the only stable network under P_1^ϵ. However, once m_3 enters the economy, g_0 is no longer stable; it will transition to g_2 in which each individual has one partner and infidelity is absent (see Figure 1.2).

The testable implication is that under the process P_1^ϵ, increasing sex ratios decreases sexual infidelity among the sexually active individuals, and as a consequence reduces the likelihood of acquiring a sexually transmitted infection.

Process P_2^ϵ (The Second Perturbed Process). This process is identical to the

process P_1^ϵ with one exception; any action that leaves its initiator indifferent (such an action consists of breaking an existing link with an agent k and forming a new link with another agent) is taken with probability $\epsilon^{f(s_k)}$ in the second process, where $f(\cdot)$ is a strictly increasing function mapping into (0,1). The interpretation is that the number of partners that agent k has signals his/her quality, and the higher the quality, the lower the probability of breaking a relationship with such an agent. This describes a pattern of herd behavior in which people tend to match with other people who have more partners.

It is shown that under the perturbed process P_2^ϵ, only anti-egalitarian networks in which a few men are matched to all the women are stable in the long run. In these stochastically stable networks, each woman obtains her optimal number of male partners whereas each man who is matched has his optimal number of partners except at most one matched man, and the remaining men have no partners. This result is proven under the assumption that there are an equal number of men and women, but one can show that it will not change qualitatively even if the sex ratio deviates from 1.

We now analyze the effect of sex ratios on sexual infidelity under the process P_2^ϵ, still considering our previous example. The network g_0 is still the only stable network under P_2^ϵ. If m_3 enters the economy, the links in g_0 will remain stable (see Figure 1.3).

The testable implication here is that under the process P_2^ϵ, increasing the sex ratio has no effect on sexual infidelity among the sexually active individuals, and as a consequence has no effect on the likelihood of acquiring a sexually transmitted infection.

Summary of Testable Implications

1) Under the unperturbed process $\mathbf{P}_0$ and the second perturbed process P_2^ϵ, sex ratios have no effect on sexual infidelity among sexually active individuals, and so do not affect the likelihood of acquiring an STI.[10]

2) Under the first perturbed process P_1^ϵ, increasing sex ratios decreases sexual infidelity among sexually active individuals, and reduces the likelihood of acquiring a sexually transmitted infection. Moreover, increasing sex ratios reduces infidelity more in men than in

[10]The comparative statics analysis is conducted on equilibrium networks.

women, and should therefore reduce STIs more in women than in men.[11]

We will next test these predictions empirically using data from the United Kingdom and Wales. As we will see, this empirical analysis shows that increasing sex ratios reduces multiple partnerships and STI transmission; the effect on multiple partnership is in general larger for men, and the effect on STIs is in general larger for women. The analysis therefore demonstrates that the first perturbed process P_1^ϵ is the process at work in this context of the United Kingdom.

1.4 Data

This section describes the data we use to analyze the causal effect of sex ratios on sexual behavior and the likelihood of acquiring a sexually transmitted infection. We use the third British National Survey of Sexual Attitudes and Lifestyles 2010-2012 (Natsal3), which we matched to data on sex ratios, calculated using data from two censuses conducted in England and Wales in 2001 and 2011.[12] The survey contains 6293 males and 8869 females. We exclude *LGBT* (lesbian, gay, bisexual, and transgender) individuals from our sample, which is in line with the model of a two-sided economy we use. They also represent a very small sample. Table 1.1 presents descriptive statistics for all key variables used in the analysis. The main variables are described below.

[11] In fact, in a highly concentrated sexual network in which a few men share all the women, an infected man who is not single will infect several women, whereas an infected woman will fewer men, which creates an infection bias against women. As infidelity decreases more for men than for women following an increase in the sex ratio, infected men who are not single infect fewer women, reducing the infection bias against women.

[12] The Natsal3 survey we use is a multistage, clustered, and nationally stratified probability sample survey conducted between September 2010 and August 2012. It collected data on over 15000 adults aged 16-74 using face-to-face interviews. The survey selected postcode sectors as the primary sampling units (PSUs); addresses within each postcode were chosen at the second stage; then one eligible adult in each household was randomly selected at the final stage (Erens et al. 2013). Two previous Natsal surveys were conducted in 1990 and 2000, respectively, but Natsal3 is the first survey that considers individuals up to the age of 74, because of the recognition that many people remain sexually active into their later life, and that sexual health issues affect both the younger and older age groups.

1.4.1 Sex Ratios

The sex ratio is defined as the ratio of the number of men to the number of women in a population. The sex ratio at any age is a function of the sex ratio at birth. In societies where sex-selective abortion is absent and females are not discriminated against in the allocation of food and other resources likely to affect survival, the male-to-female ratio at birth is 1.05 on average, but it varies between 1.03 and 1.07.[13] This surplus of males at birth generally turns into a deficit at later ages due to the fact that mortality rates are higher in males than in females (see, e.g., Waldron (1985), Pongou (2013)). In these societies, the sex ratio is thought to be determined by natural causes, which is why it can be viewed as a natural experiment for the purpose of our analysis, especially given the fact that the determinants of the sex ratio at birth (see Pongou (2013)) are not likely to affect sexual behavior or the transmission of sexually transmitted infections.

In the context of a sexual economy, the computation of the sex ratio usually takes into account the age gap between potential sexual partners (see, e.g., Becker (1973), Angrist (2002)). In our data, the average age gap between heterosexual partners is slightly greater than 4 years, which is consistent with the observation made by Angrist (2002) that men tend to marry younger women. We therefore consider a five-year age gap in computing sex ratios. Using aggregate data from censuses conducted in England and Wales in 2001 and 2011, we calculate sex ratios by first partitioning each gender category into five-year age groups (except for the first age group since individuals aged 15 at the time of the survey were not interviewed): 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, and 70-74 years old. Following a similar approach as in Angrist (2002), the sex ratio is computed as the number of men in a given cohort (e.g., 25-29) to the number of women in the preceding cohort (e.g., 20-24), given the average age gap between partners.[14] This calculation is done within each ethnicity and each region of residence. The

[13] Pongou (2013) finds that the sex ratio of at birth in sub-Saharan countries is 1.032, which is close the sex ratio among African Americans; in white populations, the sex ratio at birth is around 1.06.

[14] This is in line with Becker (1973), who argues that "a group of women must be matched with the men they are most likely to marry, e.g., college-educated women with college-educated men, or women aged 20-24 with men aged 25-29." (p. 838). However, we do not stratify on education as this variable can be

five ethnicity categories we consider are: White, Mixed, Asian (excluding Chinese)/Asian British, Black/Black British, and Chinese and Others. Nine regions are considered: North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, South West, Eastern, London, South East, and Wales.[15]

Each individual is matched to the sex ratio corresponding to his/her age group in 2011 and to the sex ratio corresponding to the age group to which he/she belonged in 2001. For example, a man aged 27 years at the time of the survey is matched to the sex ratio calculated by dividing the number of men aged 25-29 by the number of women aged 20-24 in the census conducted in 2011. Similarly, a woman aged 27 years at the time of the survey is matched to the sex ratio calculated by dividing the number of men aged 30-34 by the number of women aged 25-29 in the 2011 census. Moreover, a man aged 27 years at the time of the survey is matched to the sex ratio calculated by dividing the number of men aged 15-19 by the number of women aged 10-14 in the 2001 census, and a woman aged 27 years at the time of the survey is matched to the sex ratio calculated by dividing the number of men aged 20-24 by the number of women aged 15-19 in the 2001 census.

Table A2 in the appendix shows descriptive statistics on sex ratios. Sex ratios tend to vary across regions as well as across years. We also find that sex ratios vary across ethnic groups, being higher among the Asian and Chinese/Others categories, perhaps due to sex-selective abortion (Dubuc and Coleman 2007). Figure 1.4 further shows changes in sex ratios by region between the two censuses.

1.4.2 Sexual Behaviors

Sexual behavior is measured using several variables. These variables are: (1) the number of sexual partners in the past five years; (2) the number of serial partners in the past five years; (3) any binary indicator for whether an individual had two overlapping partners in

viewed as endogenous. We control for it in some of the regressions, and our estimates are unaffected.

[15] Fossett and Kiecolt (1991) suggest constructing sex ratios by race, age and county. However, county-level sex ratios constructed from Census data could contain serious measurement error caused by few observations in some cells. For this reason, sex ratios are computed at the regional level in this paper, as in Angrist (2002).

the past five years; (4) a binary indicator for whether an individual had any concurrent relationships in the past five years; (5) a binary indicator for whether an individual had any concurrent relationships in the last year; (6) the number of partners with whom an individual had unprotected sex in the past year; and (7) a binary indicator for whether an individual had unprotected sex in the past year. Interestingly, these variables not only reflect the traditional concept of relationships such as concurrency (or infidelity), but they also capture relationship stability, measured by the rate of partner change in a pattern of serial monogamy. We are also interested in overlapping relationships because the stochastic nature of our link-formation processes implies that, in the transition period leading to the new network configuration following an increase in the sex ratio, an individual might need to experiment with a new partner before switching definitively from the current partner. The variable measuring protection during sex could be viewed as reflecting the bargaining power of women to negotiate safe sex. All of these variables capture behaviors that play an important role in the transmission of sexually transmitted infections.

The summary statistics of these variables are reported in Table 1.2. The average number of heterosexual partners in the past five years is 2.187 for females and 3.239 for males, if we do not consider whether the relationship is concurrent or part of a pattern of serial monogamy. The average number of serial partners is 2.043 for women and 2.815 for men. Consistent with our theoretical framework in which infidelity is punished more severely for women, the latter report fewer concurrent and overlapping partners than men. Indeed, 17% of men versus 10% of women had overlapping sexual partners in the past five years before the survey. Similarly, 16.2% of men versus 11.9% of women have been involved in concurrent relationships in the past five years, whereas 7.3% of men versus 4.8% of women have been engaged in concurrent relationships in the year before the interview. The number of partners with whom an average individual had unprotected sex in the past year is 0.920 for men and 0.828 for women. Finally, the likelihood of involvement in an unprotected relationship in the year preceding the survey was 5.8% for women and 6.3% for men.

1.4.3 Sexually Transmitted Infections

In the third British National Survey of Sexual Attitudes and Lifestyles 2010-2012, interviewees were asked to provide information on whether they have been infected with any of the following STIs in their life time, within the past 5 years, and within the year preceding the survey: chlamydia, gonorrhoea, warts, syphilis, trichomonas, herpes, and/or any STI.

Table 1.3 shows some descriptive statistics on these diseases. Chlamydia is relatively more common and tends to be more prevalent among women compared to men. It is followed by gonorrhoea, which is more prevalent in men than in women. Given that the prevalence of syphilis and trichomonas is very low, we do not analyze these diseases. Therefore, the STIs we analyze are chlamydia, gonorrhoea, genital warts, herpes, and the variables capturing contact with any STI in the past five years and in the past year prior to the interview. For each infection type, we generate a binary variable indicating whether an individual is infected or not.

1.5 Identification Strategies

We use four main identification strategies to analyze the causal effect of sex ratios on sexual behavior and STI transmission. First, we estimate an OLS regression as our baseline strategy. Second, we resort to the instrumental variables approach. These analyses are complemented by a series of sensitivity analyses and robustness checks. In particular, we also employ two recent identification methodologies that relax the exclusion restriction assumption underlying the classical instrumental variables approach. In another robustness check, we restrict the analysis to white individuals born twenty years after World War II. Given that these individuals have not been affected by any major shock likely to affect the sex ratio, any variation in sex ratio in these populations is likely to be random. Finally, we analyze the effect of sex ratio on "atheoretical" health conditions, which are health conditions that are not driven by sexual interactions (Section 1.5.5). Our assumption is that sex ratios should not have any effect on these conditions. The current section only deals with the first two identification strategies.

1.5.1 OLS Estimation

We estimate the following equation using OLS:

$$y_{iael} = \alpha + \beta * R^{2011}_{iael} + \gamma_a + \alpha_e + \delta_l + X_i^{'}\gamma_0 + \epsilon_{iael} \tag{1}$$

In equation (1), y_{iael} is any of the outcome variables described in Sections 1.4.2 and 1.4.3 for individual i, in age group a, in ethnic group e, in a region l; R^{2011}_{iael} is the sex ratio corresponding to the age group of i (if i is a man, then $R^{2011}_{iael} = \frac{M^{2011}_{iael}}{W^{2011}_{i(a-5)el}}$ where the numerator is the number of men in the same age group, ethnic group, and region as i and the denominator is the number of women in the age group preceding that of i (e.g., if $a = 25-29$, then $a-5 = 20-24$) and belong to the same ethnic group and region as i; if i is a woman, $R^{2011}_{iael} = \frac{M^{2011}_{i(a+5)el}}{W^{2011}_{iael}}$ where the numerator represents the number of men in the age group following that of i (e.g., if $a = 25-29$, then $a+5 = 30-34$) and belonging to the same ethnic group and region as i, and the denominator is the number of women in the same age group, ethnic group, and region as i.); γ_a is age group fixed effect; α_e is ethnic group fixed effect; δ_l is region fixed effect; ϵ_{iael} is a disturbance term; and $X_i^{'}$ indicates individual i's education level. We estimate heteroscedasticity-robust standard errors when estimating the model to deal with the potential clustering of observations at the region level. Due to the small number of clusters in our data, we compute two p-values for all estimates; one based on the critical t-distribution value, and the other based on a wild cluster-bootstrap (Cameron et al. 2008). All regressions are estimated for men and women separately.

1.5.2 IV-2SLS Estimation

Our OLS regressions may have potential endogeneity issues due to a measurement error associated with sex ratios. Also, skewed sex ratios due to sex-selective abortions have been observed in some Asian groups in Western countries (Dubuc and Coleman 2007, Abrevaya 2009). It is possible that the rationale or norm that leads to sex-selective abortions in these groups in fact reflects an imbalance of power between genders that limit the ability

of women to negotiate safe sex or that allows men to be as unfaithful as they wish. We allay this issue in our OLS regressions by controlling for ethnic group fixed effect. We also restrict our analyses to whites in which it is believed that sex ratios are determined by natural reasons and variation in sex ratios is likely to be random.

In order to address any potential endogeneity concern, a variable for sex ratio calculated from the 2001 census data is used as an instrument for the sex ratio variable calculated from the 2011 census data. This method involves estimating a two-stage model as follows:

The first stage regresses the endogenous variable on the instrumental variable as in the following equation:

$$R_{iael}^{2011} = \alpha_1 + \beta_1 * R_{iael}^{2001} + \gamma_a + \alpha_e + \delta_l + X_i^{'}\gamma_1 + \epsilon_{1iael} \tag{2}$$

The second stage regresses our outcome variable on the predicted value of the endogenous variable obtained in the first stage as in the following equation:

$$y_{iael} = \alpha_2 + \beta_2 * \widehat{R_{iael}^{2011}} + \gamma_a + \alpha_e + \delta_l + X_i^{'}\gamma_2 + \epsilon_{2iael} \tag{3}$$

Our identification assumption is that the sex ratio of 2001 is correlated with the 2011 sex ratio, and does not directly affect the outcome of interest. In the absence of events such as wars, plagues, and natural disasters, the 2001 sex ratio should indeed be highly correlated with the 2011 sex ratio. And given that the ethnicity and region-specific cohorts on which the calculation of the sex ratio for 2001 is based do not represent the same groups on which the calculation of the 2011 sex ratio is based, it is highly unlikely that the 2001 sex ratio would directly affect sexual behavior and STIs in 2011.

The first-stage regressions for men and women are presented in Table 1.4. Columns 1 and 3 only include the 2001 sex ratio, and show that this variable strongly predicts the 2011 sex ratio. These effects are robust to controlling for age group fixed effect, ethnic group fixed effect, region fixed effect and education (columns 2 and 4). The first-stage F-statistics show that the instrument is strong.

1.5.3 Main Empirical Results

We report the results on the analysis of the effects of sex ratio on sexual behavior and acquisition of STIs, respectively. Figure 1.5 illustrates the relationship between the sex ratio and each of the aforementioned sexual behaviors, where individuals are grouped by the sex ratio quintile. This graph controls for age as both sexual behavior and the sex ratio vary by age. Overall, we find that increasing the sex ratio decreases risky sexual behavior, and that the effect is larger for men compared to women for most outcomes. Figure 1.6 describes the relationship between the sex ratio and the probability of getting infected with each of the aforementioned STIs. We find that increasing the sex ratio decreases the probability of getting infected with an STI, with the effect being larger for women than for men for most outcomes.[16] These descriptive findings are confirmed in our multavariate analyses below.

Sexual Behavior

We examine the effect of sex ratios on several variables capturing sexual behavior, as described in Section 1.4.2. Both the OLS and the IV estimates are presented in Table 1.5. All regressions control for age group fixed effect, ethnic group fixed effect, region of residence, and individual's education level. We find that increasing sex ratios significantly reduces the number of sexual partners in the past five years, the number of serial partners, the likelihood of concurrent relationships, the probability of involvement in overlapping relationships, the number of partners with whom an individual has had unprotected sex in the past five years, and the likelihood of unprotected sex in the past year preceding the survey. For instance, the OLS estimates indicate that increasing the sex ratio by one standard deviation causes the likelihood of having a concurrent partnership in the past five years to decrease by 0.050 standard deviations for women and by 0.078 standard deviations for men. For the majority of outcomes, the effect is larger for men compared to women.

All these results are consistent with the prediction of the first perturbed process P_1^ϵ.

[16] Bertocchi and Dimico (2019) find that the slave trade has positively affected polygamy, and that this in turn has increased the likelihood of HIV, with a larger effect on women.

Under this process, increasing the relative number of men increases competition for female partners, which results in women switching from highly connected male partners to match with men who have fewer other partners. In equilibrium, this results in an overall decrease in the average number of concurrent and overlapping partners, and therefore in the prevalence of infidelity.

The fact that an increased sex ratio significantly reduces the number of partners with whom an individual does not use a condom or even the likelihood of an unprotected sexual relationship shows that an increased sex ratio does not only increase competition among men, but it also improves the bargaining power of women, allowing them to negotiate safer sex. In all, the findings show that an increased sex ratio decreases risky sexual behavior, which, as we will see in the next section, results in a decrease in the likelihood of acquiring a sexually transmitted infection.

Sexually Transmitted Infections

This section reports the findings regarding the effects of sex ratio on the likelihood of acquisition of a certain number of sexually transmitted diseases. These diseases are: (1) chlamydia; (2) gonorrhoea; (3) warts; (4) herpes; and (5) any STI.

In our report, we mainly focus on infections that were diagnosed in the past five years preceding the survey, however results for infections diagnosed in the year preceding the survey are also presented. In fact, as some of these infections are asymptomatic, it may take time for an infected person to discover his/her infection status. It follows that a measure of an infection that only focuses on the year preceding the survey is likely to underestimate its prevalence, and thus analyses of the effects of sex ratios on sexually transmitted infections that occurred within the past five years preceding the survey may be more reliable.

The OLS and IV estimates are reported in Table 1.6. Increasing sex ratios clearly decreases the likelihood of acquiring any of the STIs analyzed in this paper, with the magnitude and statistical significance of the effect varying across outcomes for men and women. For example, the findings imply that that an increase in the sex ratio from 0.995

(the sample mean of census) to 1 is estimated to decrease the probability of being infected with chlamydia in the past five years by 0.035 percentage points for women. The OLS effect of sex ratio is more significant for chlamydia and any STI in the past five years. As expected, the results are less significant for infections contracted in the year preceding the survey.

The IV estimates are generally somewhat smaller and less statistically significant compared to OLS estimates. The IV effects are large for males' likelihood of getting infected with chlamydia in the past five years. They are also in general statistically significant for all other infections diagnosed in the past five years, with the exception of warts and genital herpes.

Table 1.6 (Columns (3) and (4)) also shows the negative effects of sex ratios on STI acquisition, which are smaller for men than for women. The OLS estimates are also more statistically significant than the IV estimates, and the estimates are stronger for infections diagnosed within the past five years than for those diagnosed in the year preceding the survey.

1.5.4 Alternative Identification Strategies and Robustness Checks

In this section, we investigate the robustness of our findings to alternative identification strategies and a number of specification choices. First, due to the fact that individuals might have sexual partners outside of their ethnic groups, we test the sensitivity of our results to the inclusion of cross-ethnicity sex ratios.[17] This analysis allows to see if an individual is more responsive to his/her ethnic group's sexual economy than to the outside economy. Second, we analyze the subsample of white individuals younger than 50 years, given that this is arguably the subgroup that experiences the most sexual interactions.

[17]The cross-ethnicity sex ratio for a man is the ratio of the number of men in his age group and ethnic group to the number of women from other ethnic groups in the age group that is up to 5 years younger than his age group. The cross-ethnicity sex ratio for a woman is the ratio of the number of men from other ethnic groups in the age group that is up to 5 years older than her age group to the number of women in her age group and ethnic group.

Moreover, individuals in this subgroup have not been affected by any major shock likely to affect the sex ratio, such as a war.[18] It could therefore be argued that variation in the sex ratio in this subgroup is likely to be random. Third, we use the methodology developed by Nevo and Rosen (2012) to set-identify the effect of an endogenous variable using an imperfect instrument. An instrument is imperfect when it is correlated with the unobserved error term of the equation to be identified. Given the difficulty of proving formally that an instrument is perfect in the sense of satisfying the exclusion restriction, assuming that an instrument is imperfect is realistic and can be viewed as a conservative assumption. Fourth, we use the approach developed by Lewbel (2012), who provides an identification method when the valid external instrumental variables are challenged or are not available.

Sex Ratio across Ethnic Groups

In the third British National Survey of Sexual Attitudes and Lifestyles 2010-2012, interviewees were also asked to provide information on their recent partners' perceived ethnicity. We found that a nonnegligible fraction of individuals choose partners outside of their ethnic groups. We therefore measure the impact of cross-ethnicity sex ratio on variables measuring sexual behavior and STIs using the following regression:

$$y_{iael} = \alpha + \beta_{within} * R^{2011}_{iael\_withinethnic} + \beta_{across} * R^{2011}_{iael\_acrossethnic} + \gamma_a + \alpha_e + \delta_l + X_i^{'}\gamma_0 + \epsilon_{iael} \quad (5)$$

In equation (5), $R^{2011}_{iael\_withinethnic}$ is the sex ratio that we studied so far as described in Section 1.5.1, and $R^{2011}_{iael\_acrossethnic}$ is the cross-ethnicity sex ratio corresponding to the age group of i.

We generate an instrumental variable for the cross-ethnic sex ratio following the same approach as for the within-ethnicity sex ratio. Table 1.7 and 1.8 present the results. Both the OLS and IV estimates show that the cross-ethnic sex ratio has no effect on sexual behavior and sexual infections. Its effect is much smaller than the effect of within-ethnicity sex ratio, and is not statistically significant. In addition, controlling for the cross-ethnic

[18] They were born at least twenty years after World War II.

sex ratio does not affect the effect of within-ethnicity sex ratio in general. This shows that individuals are more responsive to structural changes in their ethnic group's sexual economic than changes in the outside economy.

White Individuals under 50

We analyze the effect of the sex ratio on sexual behavior and infections among white individuals under 50 years old. As already mentioned, this subsample is interesting because sex-selective abortion has not been found among white Europeans and there has not been any major shock likely to affect the sex ratio since the 1960s. One can therefore view variation in the sex ratio in this population as close to random.

We show the descriptive analysis of the effects of the sex ratio on sexual behavior and STIs in Figures A1 and A2. These effects are larger compared to those obtained over the whole sample (see Figures 1.5 and 1.6), implying that the first perturbed process in Pongou and Serrano (2013, 2016) is more valid in the white population. We report the results of the regression in Tables 1.9 and 1.10. The OLS and IV effects of sex ratio on the outcome variables are in general larger compared to the estimates obtained over the whole sample. For example, in Table 1.9, the IV estimate of the effect of sex ratios on the number of sexual partners in the past five years has increased from -1.503 (for the whole sample) to -2.317 for women and from -1.230 to -3.049 for men, while the effect on the number of serial partners in the past five years has increased from -0.796 to -1.643 for women and from -0.541 to -3.001 for men. Similarly, the IV estimate of the effect of sex ratio on chlamydia infection has increased from -0.053 to -0.195 for women and from -0.064 to -0.117 for men, whereas the effect on gonorrhea has increased from -0.012 to -0.049 for women and from -0.010 to -0.023 for men. The findings in Table 1.10 also indicate that the effect of sex ratio is larger for women than for men, which shows that reducing the relative scarcity of male individuals carries greater benefits for women in terms of the reduction of sexually transmitted diseases.

The Imperfect Instrumental Variable Approach

We also employ a novel identification strategy developed by Nevo and Rosen (2012) to set-identify the effect of sex ratio on sexual behavior and infections. This technique relaxes the assumption of instrument exogeneity that underlies the traditional instrumental variable approach. Its underlying assumption can therefore be viewed as realistic given that it is almost impossible to prove formally that an instrument is fully exogenous. Its main assumption is that the correlation between the instrument and the unobserved error term of the equation to be identified is smaller than the correlation between the error term and the endogenous regressor; that is: $|\rho_{XU}| \geqslant |\rho_{ZU}|$, where X stands for the endogenous variable, Z is an instrumental variable, and U is a mean zero unobserved error term. It is also assumed that the two correlations are of the same sign.

We summarize the signs of the key correlations in Table 1.11. Because the correlations between our endogenous variable (the sex ratio in 2011) and the instrument (the sex ratio in 2001) are positive for both females and males, the true effect of the sex ratio lies in a one-sided instead of a two-sided bounded interval, with the sign of $\sigma_{XU}\sigma_{ZU}$ determining whether the finite bound is an upper bound or a lower bound. If the estimated interval does not contain zero, then we will be able to say something about the sign of the true effect of the sex ratio. Table 1.12 presents the results. We control for age group fixed effect, ethnic group fixed effect, region fixed effect, and education. The 95% confidence intervals are calculated using bootstrap sampling with 1000 replications. These imperfect IV estimates of the upper (resp. lower) bound of the effect of sex ratio is well within or around the confidence intervals of the 2SLS estimates for all outcomes (see columns (1) and (3) of Table 1.12). Though the imperfect instrument results do not give two-sided bounds, it does tighten the upper bound or lower bound of the true effect. It also informs us about the true sign of the effect of sex ratio for most outcomes.

Robustness to Lewbel IV

In this section, we employ the instrumental variable approach developed by Lewbel (2012) to identify the effect of an endogenous regressor when valid external instrumental vari-

ables can be challenged or are not available. Here, an instrument is found by construction. According to this methodology, β in Equation 1 can be identified based only on the heteroscedastic distribution of the error term ϵ_{1iael} in the first-stage (Equation 2). The endogenous regressor can be estimated as in a linear 2SLS regression by using X and $\left(Z - \overline{Z}\right) \widehat{\epsilon_{1iael}}$ as the instruments, where X is a vector of observed exogenous variables such as age cohort, ethnicity, residential area and education level,[19] and Z can either include the entire set of variables X or a non-empty subset of X. The identification assumption requires that $Cov(Z, \epsilon_j^2) \neq 0$ and $Cov(Z, \epsilon_1 \epsilon_2) = 0$ (for $j = 2$ in the presence of a triangular model, and for both $j = 1$ and $j = 2$ in a reverse causality system), and there are no further restrictions imposed on Z.

We find that the Pagan and Hall (1983) test rejects the null hypothesis of homoskedasticity with a P-value equals to 0.000 in our model. It follows that the Lewbel IV approach can be applied to our analysis. Table 1.13 reports the findings. Columns (2) and (4) present results from the Lewbel IV approach with both the external instrument (the 2001 sex ratio) and the internally generated instruments for females and males, respectively. The results show that estimates obtained from the Lewbel's approach are a bit larger than those obtained from the classical IV approach. However, as demonstrated in Lewbel (2012), the Lewbel IV estimates have smaller standard errors and are more efficient. Moreover, the p-value of the Hansen's J statistic of all estimations cannot reject the null hypothesis of exogeneity of the Lewbel IVs. Taken together, we can conclude that our findings are robust across different identification strategies.

1.5.5 Non-Communicable Conditions: A Falsification Test

In this section, we conduct a falsification test by analyzing the effect of the sex ratio on "atheoretical" health conditions, which are health conditions unlikely to be caused by sexual interactions. The conditions we analyze are non-communicable conditions, including Parkinson's disease, chronic lung disease, heart attack, stroke, and diabetes. Our test

[19] We acknowledge that education might not be entirely exogenous in our context. We find that removing this variable from the analysis does not change our results.

attempts to shed additional light on the mechanism through which the sex ratio affects sexually transmitted infections. Although an increase in the sex ratio could also increase the stress associated with competition for female partners, which may in turn increase the likelihood of stroke in males, we expect the sex ratio to have little effect on most of these health conditions. We estimate Equation 1, where the outcome variable is each of the aforementioned health conditions. The OLS and IV estimates are presented in Table 1.14. We find that the sex ratio has no effect on any of these conditions, with the exception of diabetes for females, but the IV estimate are not statistically significant. In general, the estimates are very close to zero. The fact that we do not find any consistent effect of the sex ratio on any of these conditions lends more credence to the proposed mechanism and interpretation of our main findings.

1.6 Conclusion

In this paper, we extend a theory of infidelity networks and test its implications for how the sex ratios affect sexual infidelity, relationship stability, and the likelihood of being infected with a sexually transmitted disease, and how these effects vary by gender. The empirical findings allow to discriminate between three *rationales* or processes of sexual network formation proposed by Pongou and Serrano (2013, 2016). Our basic extension of these processes sheds light on the different mechanisms through which the relative numbers of men and women in a sexual economy shape the incentive to engage in various forms of sexual relationships, which in turn affects the spread of STIs in the population. Using data from United Kingdom, a country where female adults are outnumber male adults, we find that an increase in the sex ratio decreases the number of concurrent, overlapping, and serial sexual partners and increases the likelihood of safe sex. This in turn decreases the likelihood of acquiring a range of STIs. Consistent with the rationale underlying the formation of egalitarian infidelity networks, we find that the effect of the sex ratio on sexual behavior is larger for men compared to women, while its effect on STIs is larger for women compared to men. These findings are robust to different identification strategies and sensitivity checks. In a context in which there are fewer men than women in the sexual economy, our findings

suggest that medical and institutional progress likely to decrease excess male mortality will have a positive effect on relationship stability and will carry substantial benefits for both men and women in terms of the reduction of sexually transmitted infections.

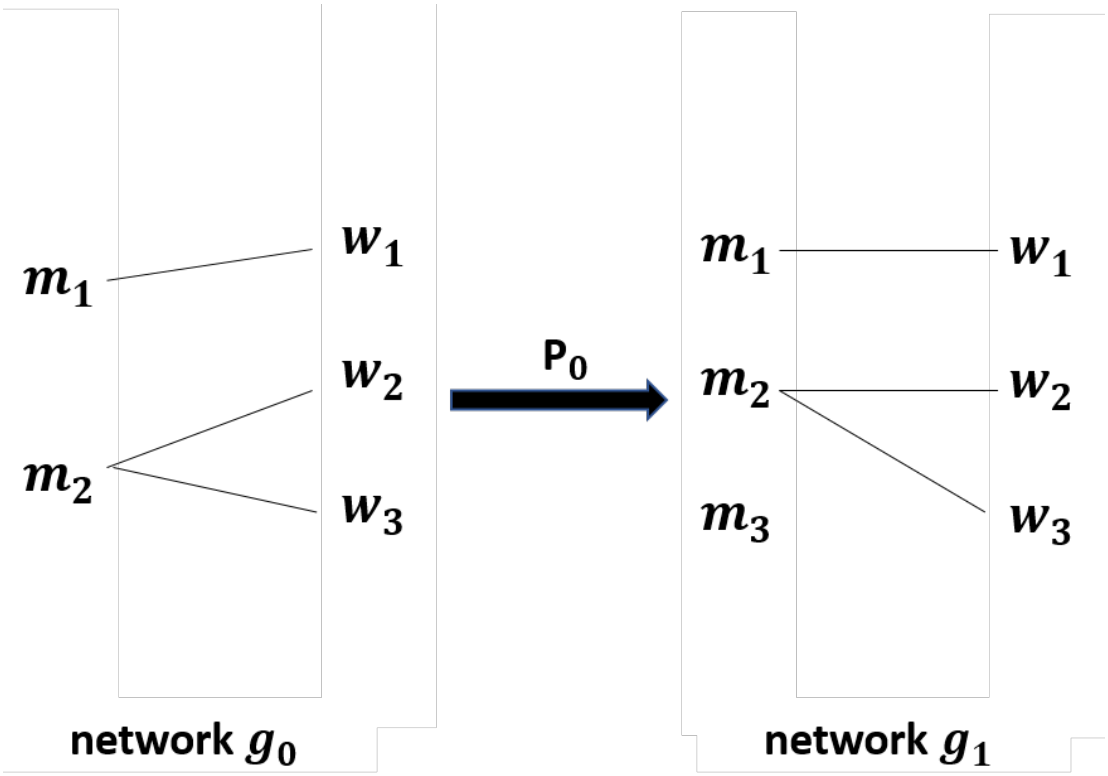


Figure 1.1: An illustration of the unperturbed process $\mathbf{P}_0$

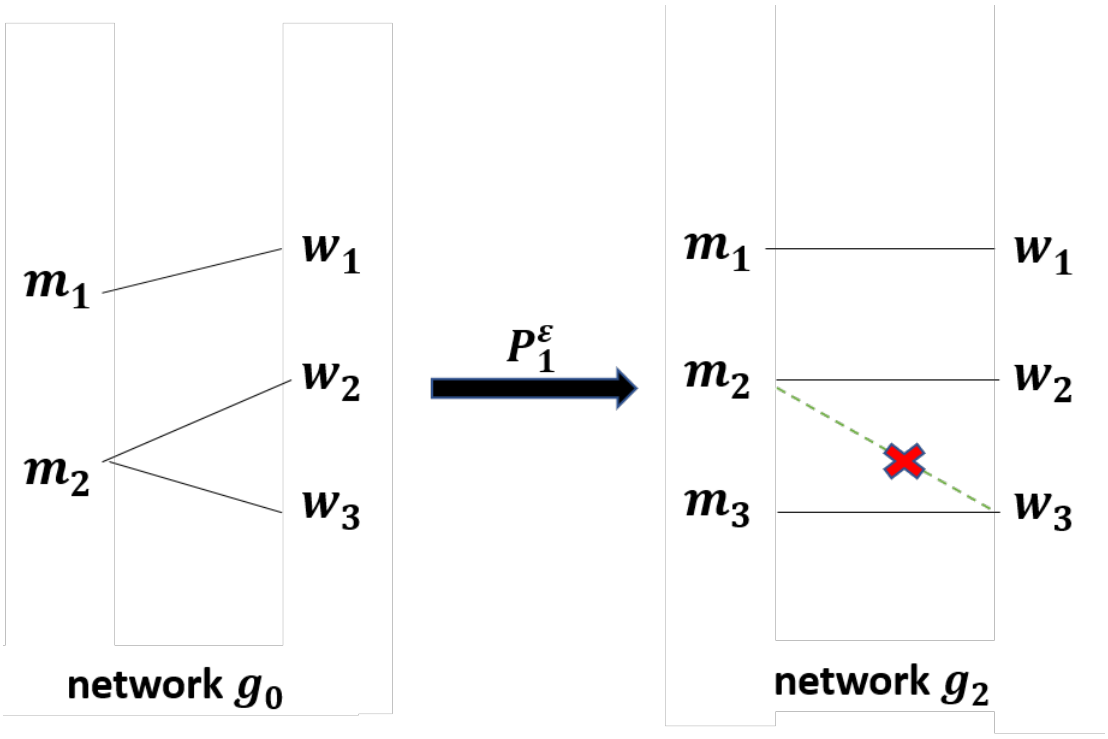


Figure 1.2: An illustration of the first perturbed process P_1^{ϵ}

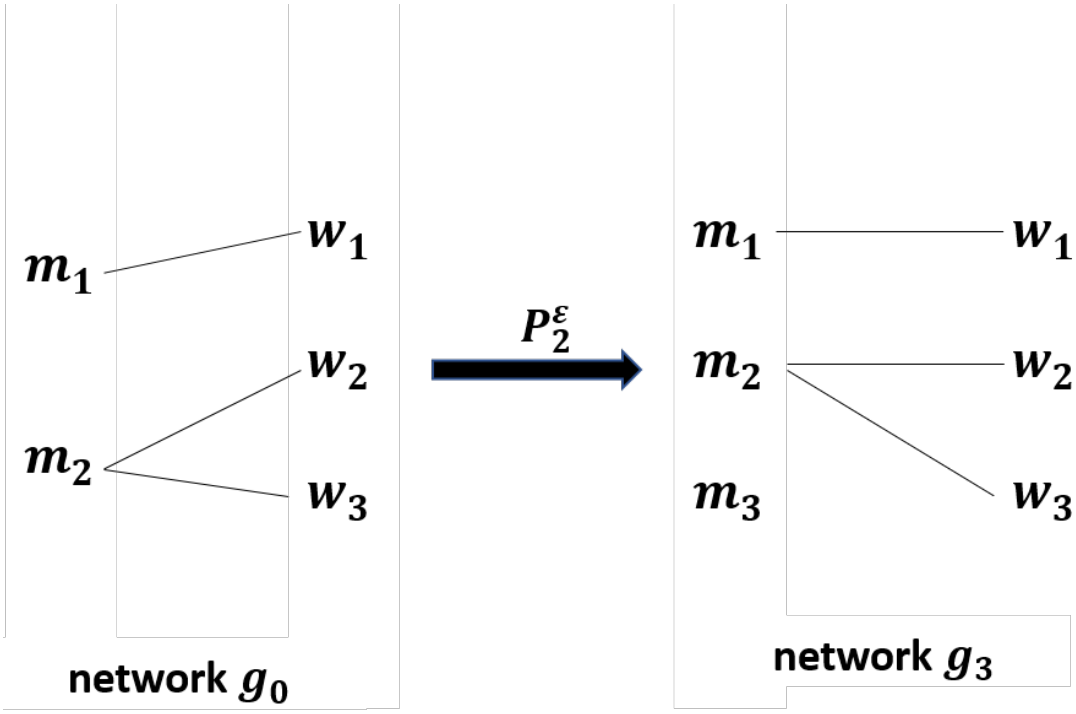


Figure 1.3: An illustration of the second perturbed process P_2^{ϵ}

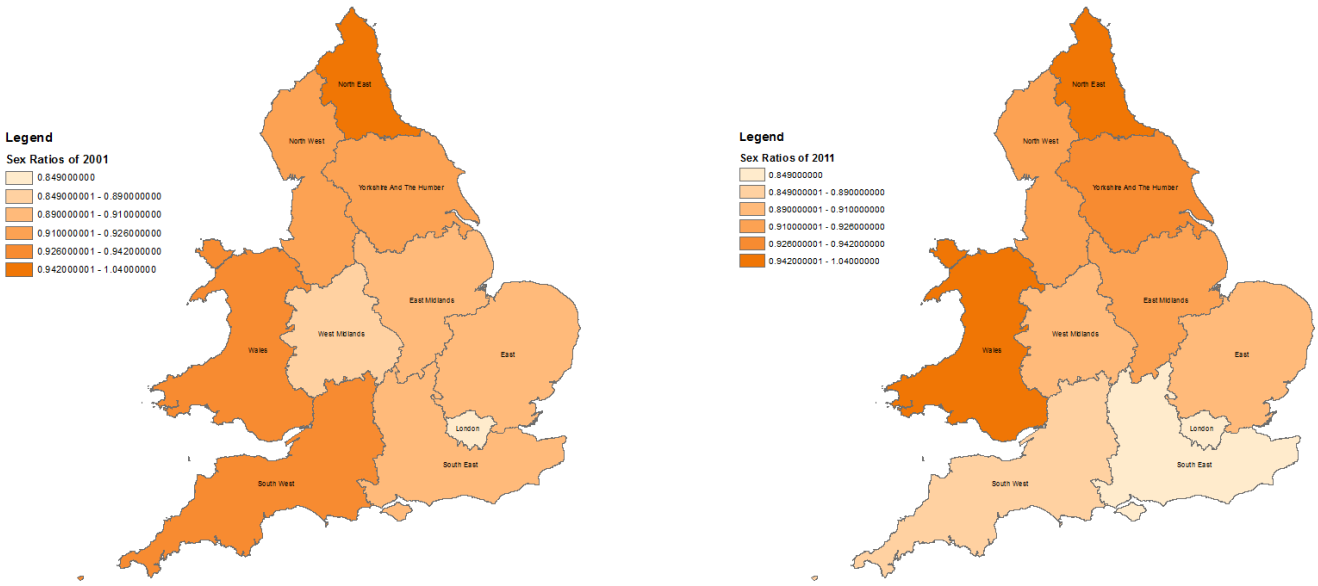


Figure 1.4: Sex Ratios by region in the 2001 and 2011 censuses

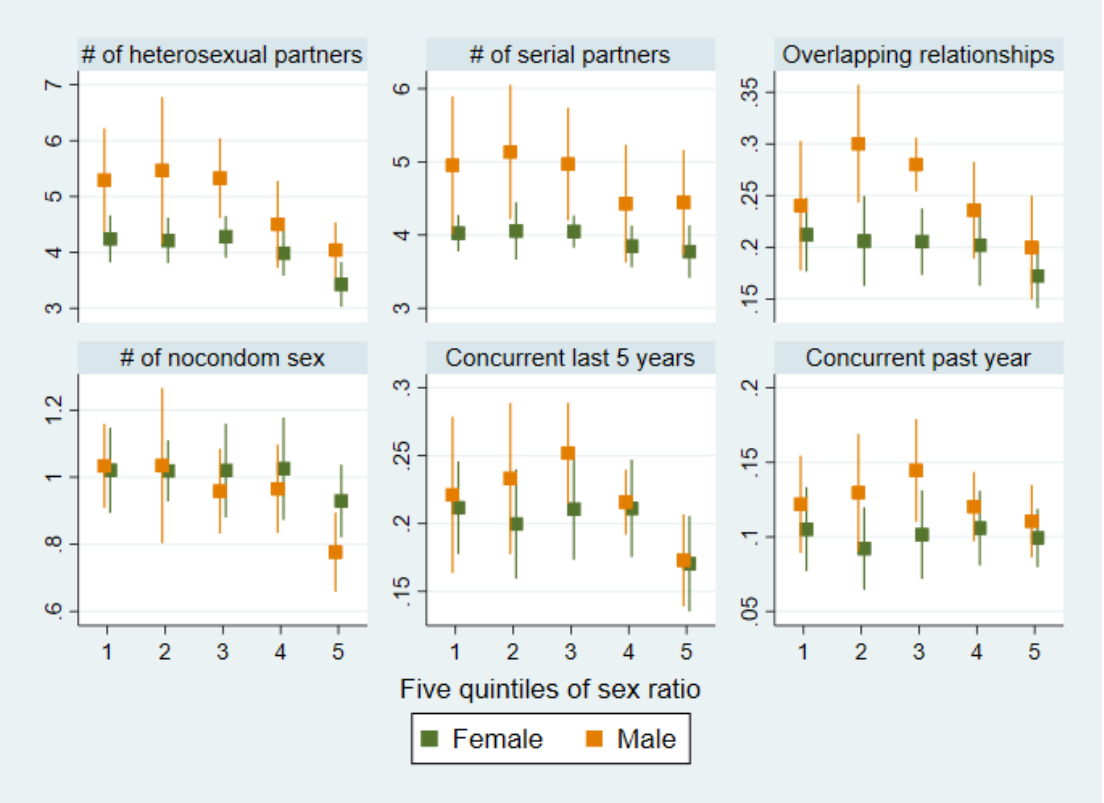


Figure 1.5: The effect of the sex ratio on sexual behavior

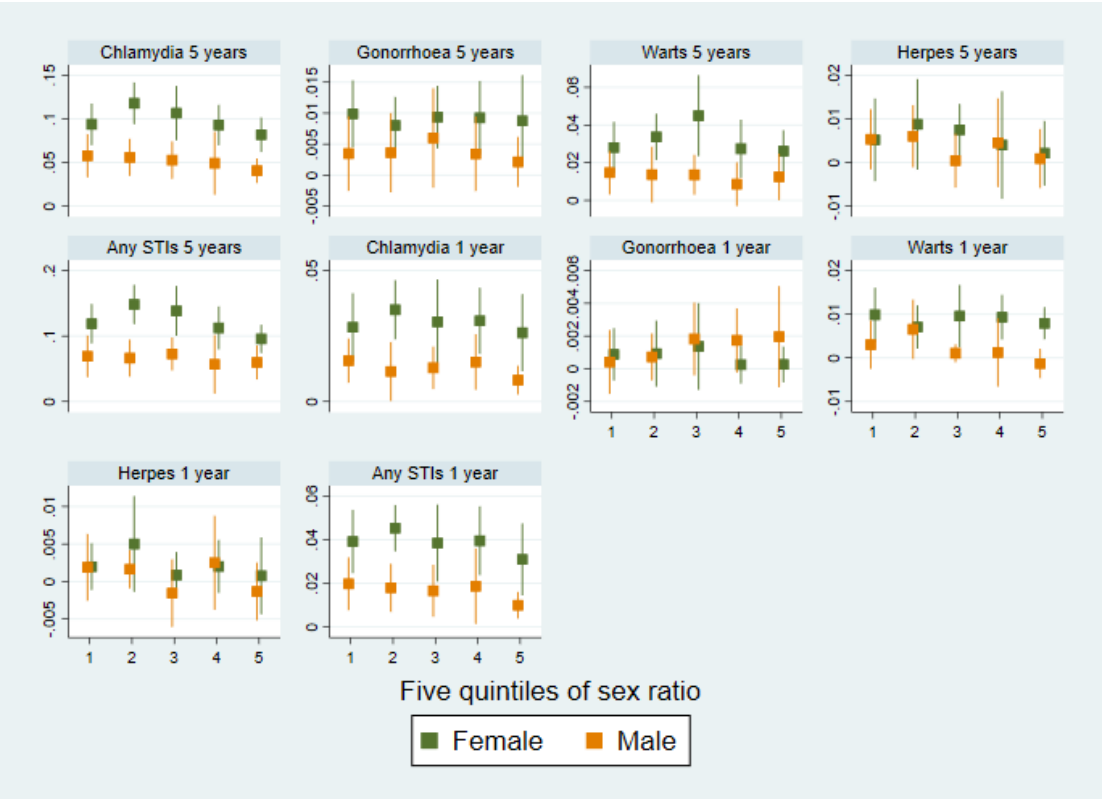


Figure 1.6: The effect of the sex ratio on STIs

Table 1.1: Summary Statistics

Variables	Description	Mean (s.d.)	N
Sex Ratios 2011	Sex Ratios of 2011	0.995	12,806
		(0.202)	
Sex Ratios 2001	Sex Ratios of 2001	0.990	12,806
		(0.142)	
Age	Respondent's age at interview	37.034	12,806
		(15.638)	
Gender	Male	0.415	12,806
		(0.493)	
Region	North East	0.055	688
		(0.229)	
	North West	0.140	1741
		(0.347)	
	Yorkshire And The Humber	0.093	1160
		(0.291)	
	East Midlands	0.090	1123
		(0.287)	
	West Midlands	0.097	1202
		(0.295)	
	South West	0.087	1085
		(0.282)	
	East	0.111	1376
		(0.314)	
	London	0.124	1545
		(0.330)	
	South East	0.144	1795
		(0.351)	
	Wales	0.058	725
		(0.234)	
Ethnic Group	White	0.866	10,773
		(0.341)	
	Mixed	0.024	300
		(0.153)	
	Asian/Asian British	0.058	720
		(0.234)	
	Black/Black British	0.038	467
		(0.190)	
	Chinese/Others	0.014	180
		(0.119)	

Notes: The table shows the mean and the standard deviation for some variables used in the analysis. The standard deviations are in parentheses. The variables Sex Ratios 2011 and Sex Ratios 2001 are calculated from censuses conducted in England and Wales in 2011 and 2001, respectively.

Table 1.2: Descriptive statistics of sexual behavior by gender

Variables	All Sample	Observations	Female	Observations	Male	Observations
	(1)	(2)	(3)	(4)	(5)	(6)
het5yrs	2.623	12,390	2.187	7,253	3.239	5,137
	(4.585)		(3.495)		(5.728)	
serial5yrs	2.355	9,052	2.043	5,392	2.815	3,660
	(3.787)		(2.950)		(4.722)	
overlap5y	0.133	12,400	0.105	7,256	0.173	5,144
	(0.340)		(0.307)		(0.378)	
concu5y	0.137	12,437	0.119	7,279	0.162	5,158
	(0.343)		(0.323)		(0.368)	
concu1y	0.058	12,428	0.048	7,274	0.073	5,154
	(0.234)		(0.213)		(0.259)	
nocondom1y	0.866	12,336	0.828	7,237	0.920	5,099
	(1.087)		(0.833)		(1.368)	
totusex1y	0.060	12,459	0.058	7,283	0.063	5,176
	(0.238)		(0.234)		(0.244)	

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table.

Table 1.3: Descriptive statistics of STIs by gender

Variables	**Ever Been Diagnosed with STI**		**Diagnosed with STI in Past Five Years Before Interview**		**Diagnosed with STI in Past One Year Before Interview**	
	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)
Chlamydia	0.076	0.048	0.037	0.030	0.010	0.007
	(0.265)	(0.214)	(0.188)	(0.170)	(0.099)	(0.081)
Gonorrhoea	0.011	0.013	0.003	0.003	0.001	0.001
	(0.104)	(0.113)	(0.054)	(0.057)	(0.029)	(0.028)
Warts	0.037	0.027	0.012	0.008	0.002	0.002
	(0.189)	(0.161)	(0.108)	(0.092)	(0.044)	(0.045)
Herpes	0.016	0.009	0.008	0.004	0.002	0.002
	(0.126)	(0.093)	(0.088)	(0.066)	(0.047)	(0.040)
Any STI	0.166	0.123	0.051	0.045	0.014	0.011
	(0.372)	(0.328)	(0.221)	(0.208)	(0.118)	(0.106)
Observations	7103	5015	7103	5015	7103	5015

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table.

Table 1.4: First-stage regressions: The effect of the 2001 sex ratio on the 2011 sex ratio

Dependent Variable: 2011 Sex Ratio

Variables	Female		Male	
	(1)	(2)	(3)	(4)
2001 Sex Ratio	0.838***	0.631***	0.867***	0.682***
	(0.140)	(0.153)	(0.160)	(0.181)
Age Group FE	No	Yes	No	Yes
Ethnicity Group FE	No	Yes	No	Yes
Region FE	No	Yes	No	Yes
Education Control	No	Yes	No	Yes
First-stage F statistic	35.83	16.93	29.19	14.16
Observations	7,231	7,226	5,099	5,096

Note: Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 1.5: The effect of the sex ratio on sexual behavior (OLS and 2SLS)

		Female		Male	
Dependent Variables	Regressor	OLS	2SLS	OLS	2SLS
		(1)	(2)	(3)	(4)
concu5y	sexratio	-0.075***	-0.039	-0.147***	-0.177***
	2011	(0.015)	(0.024)	(0.036)	(0.065)
		P^A=0.001;P^W=0.006	P^A=0.105;P^W=0.136	P^A=0.003;P^W=0.010	P^A=0.007;P^W=0.042
het5yrs	sexratio	-2.083***	-1.503***	-3.358***	-1.230
	2011	(0.265)	(0.229)	(0.663)	(0.953)
		P^A=0.000;P^W=0.000	P^A=0.000;P^W=0.000	P^A=0.001;P^W=0.000	P^A=0.197;P^W=0.296
serial5y	sexratio	-0.937***	-0.796**	-1.906***	-0.541
	2011	(0.269)	(0.316)	(0.511)	(0.652)
		P^A=0.007;P^W=0.006	P^A=0.012;P^W=0.060	P^A=0.005;P^W=0.014	P^A=0.407;P^W=0.436
overlap5y	sexratio	-0.092***	-0.064***	-0.129**	-0.057
	2011	(0.024)	(0.020)	(0.046)	(0.067)
		P^A=0.004;P^W=0.000	P^A=0.002;P^W=0.014	P^A=0.019;P^W=0.004	P^A=0.392;P^W=0.523
concu1y	sexratio	-0.008	-0.020	-0.043*	-0.029
	2011	(0.019)	(0.019)	(0.021)	(0.031)
		P^A=0.681;P^W=0.677	P^A=0.287;P^W=0.372	P^A=0.077;P^W=0.088	P^A=0.359;P^W=0.489
nocondom1y	sexratio	-0.228**	-0.063	-0.564***	-0.525***
	2011	(0.0980)	(0.0936)	(0.112)	(0.196)
		P^A=0.045;P^W=0.128	P^A=0.502;P^W=0.555	P^A=0.001;P^W=0.000	P^A=0.007;P^W=0.076
totusexly	sexratio	-0.0342**	-0.0921***	-0.0127	0.0523
	2011	(0.012)	(0.021)	(0.033)	(0.059)
		P^A=0.017;P^W=0.054	P^A=0.000;P^W=0.014	P^A=0.709;P^W=0.693	P^A=0.377;P^W=0.440
Age FE		Yes	Yes	Yes	Yes
Ethnic FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. The instrumental variable for the variable Sex Ratios 2011 is Sex Ratios 2001. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. 1000 wild bootstrap replications were used. P^A refers to asymptotic p-value, and P^W refers to wild-cluster bootstrap p-value.

Table 1.6: The effect of the sex ratio on STIs (OLS and 2SLS)

		Female		Male	
Dependent Variables	Regressor	OLS	2SLS	OLS	2SLS
		(1)	(2)	(3)	(4)
anysti5y	sexratio	-0.097***	-0.069***	-0.043**	-0.058
	2011	(0.016)	(0.018)	(0.014)	(0.035)
		P^A=0.000;P^W=0.000	P^A=0.000;P^W=0.022	P^A=0.015;P^W=0.014	P^A=0.100;P^W=0.232
chlamydia5y	sexratio	-0.070***	-0.053**	-0.033**	-0.064***
	2011	(0.013)	(0.021)	(0.010)	(0.022)
		P^A=0.000;P^W=0.000	P^A=0.013;P^W=0.074	P^A=0.011;P^W=0.000	P^A=0.004;P^W=0.034
gonorrh5y	sexratio	-0.000	-0.012*	-0.004	-0.010*
	2011	(0.006)	(0.007)	(0.004)	(0.006)
		P^A=0.933;P^W=0.983	P^A=0.070;P^W=0.082	P^A=0.362;P^W=0.477	P^A=0.098;P^W=0.138
warts5y	sexratio	-0.029**	-0.009	-0.013*	-0.008
	2011	(0.009)	(0.006)	(0.007)	(0.016)
		P^A=0.011;P^W=0.004	P^A=0.146;P^W=0.214	P^A=0.099;P^W=0.040	P^A=0.621;P^W=0.687
herpes5y	sexratio	-0.004	0.007	-0.006	-0.001
	2011	(0.003)	(0.011)	(0.003)	(0.015)
		P^A=0.272;P^W=0.210	P^A=0.535;P^W=0.653	P^A=0.113;P^W=0.084	P^A=0.967;P^W=0.945
chlamydialy	sexratio	-0.015	-0.001	-0.008	-0.012
	2011	(0.012)	(0.012)	(0.005)	(0.011)
		P^A=0.234;P^W=0.326	P^A=0.915;P^W=0.967	P^A=0.145;P^W=0.144	P^A=0.256;P^W=0.356
gonorrhly	sexratio	-0.001	-0.004*	0.002	0.002
	2011	(0.003)	(0.002)	(0.003)	(0.003)
		P^A=0.646;P^W=0.790	P^A=0.079;P^W=0.192	P^A=0.515;P^W=0.513	P^A=0.447;P^W=0.471
wartsly	sexratio	-0.004	-0.004	-0.010**	-0.004
	2011	(0.004)	(0.004)	(0.004)	(0.006)
		P^A=0.380;P^W=0.477	P^A=0.313;P^W=0.412	P^A=0.046;P^W=0.072	P^A=0.513;P^W=0.581
herpesly	sexratio	-0.002	0.006	-0.002	-0.003
	2011	(0.004)	(0.007)	(0.001)	(0.003)
		P^A=0.564;P^W=0.653	P^A=0.328;P^W=0.448	P^A=0.128;P^W=0.166	P^A=0.379;P^W=0.463
anystily	sexratio	-0.024**	-0.009	-0.015**	-0.016**
	2011	(0.010)	(0.011)	(0.005)	(0.008)
		P^A=0.046;P^W=0.086	P^A=0.367;P^W=0.412	P^A=0.025;P^W=0.050	P^A=0.043;P^W=0.074
Age FE		Yes	Yes	Yes	Yes
Ethnic FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. The instrumental variable for the variable Sex Ratios 2011 is Sex Ratios 2001. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. 1000 wild bootstrap replications were used. P^A refers to asymptotic p-value, and P^W refers to wild-cluster bootstrap p-value.

Table 1.7: The effects of within-ethnicity sex ratio and cross-ethnicity sex ratio on sexual behavior (OLS and IV)

		Female		Male	
Dependent Variable	Regressor	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)
concu5y	SR_withinethnic	-0.0712***	-0.0332	-0.154***	-0.180**
	2011	(0.0156)	(0.0243)	(0.0357)	(0.0733)
		P^A=0.001;P^W=0.002	P^A=0.164;P^W=0.2823	P^A=0.002;P^W=0.016	P^A=0.013;P^W=0.044
	SR_acrossethnic	-0.000901***	-0.00105***	0.000311	-0.000296
	2011	(0.000178)	(0.000148)	(0.000616)	(0.000307)
het5yrs	SR_withinethnic	-2.048***	-1.422***	-3.430***	-1.310
	2011	(0.272)	(0.270)	(0.638)	(0.917)
		P^A=0.000;P^W=0.000	P^A=0.000;P^W=0.000	P^A=0.000;P^W=0.000	P^A=0.142;P^W=0.2242
	SR_acrossethnic	-0.00957	-0.0127***	0.00651	0.00382
	2011	(0.00552)	(0.00479)	(0.00962)	(0.00782)
serial5y	SR_withinethnic	-0.924***	-0.758**	-1.911***	-0.637
	2011	(0.268)	(0.320)	(0.514)	(0.648)
		P^A=0.007;P^W=0.006	P^A=0.02;P^W=0.0681	P^A=0.005;P^W=0.016	P^A=0.339;P^W=0.3744
	SR_acrossethnic	-0.00375	-0.00516	0.00239	0.0104
	2011	(0.00391)	(0.00370)	(0.0109)	(0.00954)
overlap5y	SR_withinethnic	-0.0892***	-0.0599***	-0.139**	-0.0593
	2011	(0.0242)	(0.0227)	(0.0458)	(0.0647)
		P^A=0.005;P^W=0	P^A=0.005;P^W=0.024	P^A=0.014;P^W=0.002	P^A=0.338;P^W=0.4925
	SR_acrossethnic	-0.000885*	-0.00101***	0.00117*	0.000353
	2011	(0.000455)	(0.000349)	(0.000597)	(0.000564)
concu1y	SR_withinethnic	-0.00767	-0.0178	-0.0482**	-0.0332
	2011	(0.0186)	(0.0190)	(0.0213)	(0.0306)
		P^A=0.689;P^W=0.6847	P^A=0.363;P^W=0.3964	P^A=0.05;P^W=0.0561	P^A=0.245;P^W=0.2843
	SR_acrossethnic	0.000132	-0.000120	0.000202	-0.000237
	2011	(0.000206)	(0.000159)	(0.000567)	(0.000352)
nocondom1y	SR_withinethnic	-0.225*	-0.0525	-0.571***	-0.528***
	2011	(0.100)	(0.0991)	(0.119)	(0.193)
		P^A=0.052;P^W=0.1261	P^A=0.551;P^W=0.5886	P^A=0.001;P^W=0	P^A=0.006;P^W=0.0641
	SR_acrossethnic	-0.00249	-0.00361*	-0.000416	-0.00258**
	2011	(0.00180)	(0.00188)	(0.00151)	(0.00129)
totusex1y	SR_withinethnic	-0.0337**	-0.0905***	-0.0175	0.0481
	2011	(0.0113)	(0.0212)	(0.0308)	(0.0548)
		P^A=0.015;P^W=0.036	P^A=0;P^W=0.018	P^A=0.585;P^W=0.6486	P^A=0.374;P^W=0.4525
	SR_acrossethnic	-0.0000299	-0.000134	0.000396	0.0000852
	2011	(0.000312)	(0.000284)	(0.000582)	(0.000453)
Age FE		Yes	Yes	Yes	Yes
Ethnic FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. The instrumental variables for the 2011 within-ethnicity sex ratio (SR_withinethnic 2011) and the 2011 cross-ethnicity sex ratio (SR_acrossethnic 2011) are the 2001 within-ethnicity sex ratio and the 2001 cross-ethnicity sex ratio, respectively. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. 1000 wild bootstrap replications were used. P^A refers to asymptotic p-value, and P^W refers to wild-cluster bootstrap p-value, respectively.

Table 1.8: The effects of within-ethnicity sex ratio and cross-ethnicity sex ratio on STIs (OLS and IV)

		Female		Male	
Dependent Variable	Regressor	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)
anysti5y	SR_withinethnic	-0.0936***	-0.0648***	-0.0446**	-0.0600*
	2011	(0.0154)	(0.0184)	(0.0139)	(0.0342)
		P^A=0;P^W=0	P^A=0;P^W=0.012	P^A=0.011;P^W=0.002	P^A=0.068;P^W=0.1381
	SR_acrossethnic	-0.000786**	-0.000769***	0.000229	0.000220
	2011	(0.000301)	(0.000215)	(0.000154)	(0.000156)
chlamydia5y	SR_withinethnic	-0.0675***	-0.0491**	-0.0324**	-0.0620***
	2011	(0.0128)	(0.0198)	(0.0109)	(0.0203)
		P^A=0.001;P^W=0.002	P^A=0.016;P^W=0.0741	P^A=0.016;P^W=0.000	P^A=0.002;P^W=0.028
	SR_acrossethnic	-0.000653**	-0.000567***	-0.0000800	-0.0000533
	2011	(0.000232)	(0.000150)	(0.0000819)	(0.000145)
gonorrh5y	SR_withinethnic	-0.000431	-0.0122*	-0.00444	-0.0103
	2011	(0.00550)	(0.00679)	(0.00443)	(0.00632)
		P^A=0.939;P^W=0.967	P^A=0.077;P^W=0.0581	P^A=0.343;P^W=0.4324	P^A=0.104;P^W=0.1742
	SR_acrossethnic	-0.0000259	-0.0000481	0.0000633	0.0000566
	2011	(0.0000545)	(0.0000465)	(0.0000666)	(0.0000846)
warts5y	SR_withinethnic	-0.0285**	-0.00778	-0.0138*	-0.00983
	2011	(0.00897)	(0.00668)	(0.00685)	(0.0155)
		P^A=0.011;P^W=0.01	P^A=0.227;P^W=0.2843	P^A=0.074;P^W=0.006	P^A=0.536;P^W=0.6507
	SR_acrossethnic	-0.0000791	-0.000153	0.0000983	0.0000749
	2011	(0.000129)	(0.000139)	(0.000141)	(0.000163)
herpes5y	SR_withinethnic	-0.00428	0.00788	-0.00514	-0.000766
	2011	(0.00366)	(0.0117)	(0.00330)	(0.0157)
		P^A=0.273;P^W=0.2302	P^A=0.505;P^W=0.6126	P^A=0.154;P^W=0.1181	P^A=0.976;P^W=0.955
	SR_acrossethnic	-0.0000499	-0.0000641	-0.0000556*	-0.0000346
	2011	(0.000140)	(0.000119)	(0.0000303)	(0.0000269)
chlamydialy	SR_withinethnic	-0.0149	-0.000835	-0.00704	-0.0117
	2011	(0.0119)	(0.0129)	(0.00519)	(0.00928)
		P^A=0.242;P^W=0.2983	P^A=0.945;P^W=0.9389	P^A=0.208;P^W=0.1822	P^A=0.203;P^W=0.2462
	SR_acrossethnic	0.0000177	0.0000371	-0.000104	-0.0000513
	2011	(0.000120)	(0.0000888)	(0.0000765)	(0.000150)
gonorrhly	SR_withinethnic	-0.00132	-0.00384*	0.00160	0.00186
	2011	(0.00261)	(0.00217)	(0.00241)	(0.00240)
		P^A=0.624;P^W=0.7047	P^A=0.079;P^W=0.1662	P^A=0.524;P^W=0.4925	P^A=0.434;P^W=0.4464
	SR_acrossethnic	0.0000138	0.0000793	0.0000405	0.0000613e
	2011	(0.0000149)	(0.0000946)	(0.0000556)	(0.0000836)
wartsly	SR_withinethnic	-0.00363	-0.00386	-0.0101**	-0.00498
	2011	(0.00384)	(0.00406)	(0.00419)	(0.00626)
		P^A=0.369;P^W=0.4805	P^A=0.338;P^W=0.4484	P^A=0.039;P^W=0.0901	P^A=0.433;P^W=0.5225
	SR_acrossethnic	0.0000661*	0.0000538	0.0000404**	0.0000284**
	2011	(0.0000343)	(0.0000397)	(0.0000125)	(0.0000119)
herpesly	SR_withinethnic	-0.00233	0.00632	-0.00201	-0.00277
	2011	(0.00372)	(0.00638)	(0.00151)	(0.00382)
		P^A=0.547;P^W=0.6126	P^A=0.311;P^W=0.311	P^A=0.216;P^W=0.2002	P^A=0.479;P^W=0.5606
	SR_acrossethnic	0.0000793	0.0000699	-0.0000855**	-0.0000499***
	2011	(0.0000897)	(0.0000872)	(0.0000343)	(0.0000135)
anystily	SR_withinethnic	-0.0240**	-0.00976	-0.0138**	-0.0166**
	2011	(0.0106)	(0.0112)	(0.00560)	(0.00717)
		P^A=0.05;P^W=0.1141	P^A=0.391;P^W=0.4344	P^A=0.035;P^W=0.028	P^A=0.018;P^W=0.0501
	SR_acrossethnic	0.000141	0.000153*	-0.000110	-0.0000211
	2011	(0.000109)	(0.0000908)	(0.000115)	(0.000163)
Age FE		Yes	Yes	Yes	Yes
Ethnic FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. The instrumental variables for the 2011 within-ethnicity sex ratio (SR_withinethnic 2011) and the 2011 cross-ethnicity sex ratio (SR_acrossethnic 2011) are the 2001 within-ethnicity sex ratio and the 2001 cross-ethnicity sex ratio, respectively. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. 1000 wild bootstrap replications were used. P^A refers to asymptotic p-value, and P^W refers to wild-cluster bootstrap p-value, respectively.

Table 1.9: The effects of the sex ratio on sexual behavior (OLS and IV): **White Individuals under 50**

		Female		Male	
Dependent Variables	Regressor	OLS	2SLS	OLS	2SLS
		(1)	(2)	(3)	(4)
concu5y	sexratio	-0.130***	-0.108	-0.249***	-0.441***
	2011	(0.032)	(0.091)	(0.064)	(0.159)
		P^A=0.003;P^W=0	P^A=0.233;P^W=0.459	P^A=0.004;P^W=0.010	P^A=0.005;P^W=0.034
het5yrs	sexratio	-3.091***	-2.317**	-5.249***	-3.049
	2011	(0.359)	(0.972)	(1.309)	(3.023)
		P^A=0.000;P^W=0.000	P^A=0.017;P^W=0.126	P^A=0.003;P^W=0.000	P^A=0.313;P^W=0.523
serial5y	sexratio	-1.397***	-1.643**	-3.045***	-3.001*
	2011	(0.323)	(0.810)	(0.915)	(1.689)
		P^A=0.002;P^W=0.002	P^A=0.043;P^W=0.002	P^A=0.009;P^W=0.008	P^A=0.076;P^W=0.038
overlap5y	sexratio	-0.154***	-0.071	-0.283**	-0.282*
	2011	(0.037)	(0.099)	(0.097)	(0.165)
		P^A=0.002;P^W=0.010	P^A=0.473;P^W=0.613	P^A=0.017;P^W=0.002	P^A=0.088;P^W=0.270
concu1y	sexratio	0.007	0.045	-0.075**	-0.079
	2011	(0.029)	(0.056)	(0.030)	(0.106)
		P^A=0.818;P^W=0.835	P^A=0.421;P^W=0.184	P^A=0.032;P^W=0.056	P^A=0.457;P^W=0.671
nocondom1y	sexratio	-0.277*	-0.070	-0.761**	-0.925**
	2011	(0.138)	(0.331)	(0.275)	(0.431)
		P^A=0.077;P^W=0.070	P^A=0.833;P^W=0.969	P^A=0.022;P^W=0.004	P^A=0.032;P^W=0.034
totusexly	sexratio	-0.026	-0.093*	-0.071	-0.035
	2011	(0.024)	(0.055)	(0.044)	(0.098)
		P^A=0.304;P^W=0.280	P^A=0.090;P^W=0.044	P^A=0.139;P^W=0.106	P^A=0.721;P^W=0.907
Age FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. The instrumental variable for the variable Sex Ratios 2011 is Sex Ratios 2001. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. 1000 wild bootstrap replications were used. P^A refers to asymptotic p-value, and P^W refers to wild-cluster bootstrap p-value, respectively.

Table 1.10: The effects of the sex ratio on STIs (OLS and IV): **White Individuals under 50**

		Female		Male	
Dependent Variables	Regressor	OLS	2SLS	OLS	2SLS
		(1)	(2)	(3)	(4)
anysti5y	sexratio	-0.192***	-0.251***	-0.067**	-0.029
	2011	(0.047)	(0.078)	(0.028)	(0.128)
		P^A=0.003;P^W=0.000	P^A=0.001;P^W=0.018	P^A=0.039;P^W=0.008	P^A=0.821;P^W=0.981
chlamydia5y	sexratio	-0.149***	-0.195***	-0.053**	-0.117*
	2011	(0.030)	(0.065)	(0.021)	(0.061)
		P^A=0.001;P^W=0.000	P^A=0.003;P^W=0.016	P^A=0.032;P^W=0.010	P^A=0.055;P^W=0.038
gonorrh5y	sexratio	0.004	-0.049	-0.010	-0.023
	2011	(0.013)	(0.038)	(0.009)	(0.019)
		P^A=0.744;P^W=0.899	P^A=0.194;P^W=0.012	P^A=0.282;P^W=0.310	P^A=0.233;P^W=0.148
warts5y	sexratio	-0.062**	-0.046***	-0.018	0.012
	2011	(0.020)	(0.018)	(0.013)	(0.043)
		P^A=0.014;P^W=0.010	P^A=0.009;P^W=0.044	P^A=0.187;P^W=0.152	P^A=0.774;P^W=0.619
herpes5y	sexratio	-0.007	-0.007	-0.008	0.005
	2011	(0.009)	(0.025)	(0.007)	(0.049)
		P^A=0.514;P^W=0.529	P^A=0.780;P^W=0.805	P^A=0.332;P^W=0.300	P^A=0.923;P^W=0.737
chlamydialy	sexratio	-0.029*	-0.011	-0.019***	-0.018
	2011	(0.015)	(0.026)	(0.006)	(0.011)
		P^A=0.088;P^W=0.110	P^A=0.659;P^W=0.761	P^A=0.009;P^W=0.016	P^A=0.119;P^W=0.156
gonorrhly	sexratio	-0.002	-0.012	-0.000	0.001
	2011	(0.006)	(0.008)	(0.003)	(0.004)
		P^A=0.764;P^W=0.945	P^A=0.137;P^W=0.056	P^A=0.868;P^W=0.857	P^A=0.885;P^W=0.769
wartsly	sexratio	-0.003	-0.007	-0.020**	-0.017
	2011	(0.007)	(0.005)	(0.008)	(0.018)
		P^A=0.665;P^W=0.861	P^A=0.129;P^W=0.070	P^A=0.036;P^W=0.104	P^A=0.339;P^W=0.611
herpesly	sexratio	-0.001	0.017	-0.003	-0.008
	2011	(0.006)	(0.021)	(0.004)	(0.009)
		P^A=0.917;P^W=0.943	P^A=0.414;P^W=0.172	P^A=0.405;P^W=0.394	P^A=0.401;P^W=0.513
anystily	sexratio	-0.035**	-0.024	-0.032**	-0.032*
	2011	(0.012)	(0.017)	(0.011)	(0.017)
		P^A=0.018;P^W=0.032	P^A=0.153;P^W=0.314	P^A=0.014;P^W=0.022	P^A=0.059;P^W=0.182
Age FE		Yes	Yes	Yes	Yes
Ethnic FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. The instrumental variable for the variable Sex Ratios 2011 is Sex Ratios 2001. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. 1000 wild bootstrap replications were used. P^A refers to asymptotic p-value, and P^W refers to wild-cluster bootstrap p-value, respectively.

Table 1.11: Identifying Assumptions

	Female		Male	
$\sigma_{X,Z}$	0.0159		0.0166	
	σ_{XU}	σ_{ZU}	σ_{XU}	σ_{ZU}
	(1)	(2)	(3)	(4)
Panel A	**Sexual behavior**			
het5yr	-	-	+	+
serial5y	-	-	-	-
overlap5y	-	-	+	+
concu5y	-	-	+	+
concu1y	-	-	+	+
nocondom1y	-	-	+	+
totusex1y	-	-	+	+
Panel B	**Sexually transmitted infections**			
chlamydia5y	-	-	-	-
gonorrh5y	-	-	+	+
warts5y	-	-	+	+
herpes5y	-	-	+	+
anysti5y	-	-	-	-
chlamydialy	-	-	+	+
gonorrhly	+	+	-	-
wartsly	-	-	+	-
herpesly	-	-	+	+
anystily	-	-	+	+

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. σ_{XU}, σ_{ZU} and $\sigma_{X,Z}$ are estimated from data.

Table 1.12: Estimated Bounds employing Nevo and Rosen (2012)

Dependent Variables	Regressor	Female			Male		
		IV	Imperfect IV(bootstrap at 95% CI)		IV	Imperfect IV(bootstrap at 95% CI)	
		(bootstrap at 95% CI)	Lower Bound	Upper Bound	(bootstrap at 95% CI)	Lower Bound	Upper Bound
		(1)	(2)	(3)	(4)	(5)	(6)
Panel A				**Sexual behavior**			
het5yr	SexRatios 2011	[-2.145, -.8792]	-2.01592	$+\infty$	[-4.345, .8803]	$-\infty$	-1.885543
serial5y	SexRatios 2011	[-1.568, .02859]	-1.501219	$+\infty$	[-2.671, .8077]	-1.9893767	$+\infty$
overlap5y	SexRatios 2011	[-.1148, -.0157]	-0.10855935	$+\infty$	[-.2222, .1105]	$-\infty$	-0.09825052
concu5y	SexRatios 2011	[-.114, .01166]	-0.09365651	$+\infty$	[-.3271, -.006236]	$-\infty$	-0.02982497
concu1y	SexRatios 2011	[-.06565, .03649]	-0.06266317	$+\infty$	[-.105, .05197]	$-\infty$	-0.01695551
nocondom1y	SexRatios 2011	[-.2203, .216]	-0.27286232	$+\infty$	[-.9849, .09577]	$-\infty$	-0.0866363
totusexly	SexRatios 2011	[-.1391, -.04079]	-0.03696791	$+\infty$	[-.1487, .1769]	$-\infty$	0.05935141
Panel B				**Sexually transmitted infections**			
chlamydia5y	SexRatios 2011	[-.1031, .007738]	-0.09962763	$+\infty$	[-.121, -.009019]	$-\infty$	-0.01466502
gonorrh5y	SexRatios 2011	[-.03033, .001449]	-0.01194421	$+\infty$	[-.02362, .01049]	$-\infty$	0.00347845
warts5y	SexRatios 2011	[-.0252, .004844]	-0.02307949	$+\infty$	[-.05689, .03115]	$-\infty$	0.02599635
herpes5y	SexRatios 2011	[-.02746, .03332]	-0.01849024	$+\infty$	[-.03681, .0354]	$-\infty$	0.03069916
anysti5y	SexRatios 2011	[-.1123, -.01976]	-0.10949405	$+\infty$	[-.1553, .02771]	$-\infty$	-0.01119657
chlamydialy	SexRatios 2011	[-.04159, .02576]	-0.02873134	$+\infty$	[-.03639, .01356]	$-\infty$	0.00293689
gonorrhly	SexRatios 2011	[-.009413, .001162]	n.a.	n.a.	[-.003944, .0091]	n.a.	n.a.
wartsly	SexRatios 2011	[-.01325, .004223]	-0.0118952	$+\infty$	[-.02192, .009325]	n.a.	n.a.
herpesly	SexRatios 2011	[-.009246, .02467]	-0.00822213	$+\infty$	[-.01216, .00393]	n.a.	n.a.
anystily	SexRatios 2011	[-.0376, .01529]	-0.03275397	$+\infty$	[-.03318, .003895]	$-\infty$	0.00176821
Age FE		Yes	Yes		Yes	Yes	
Ethnic FE		Yes	Yes		Yes	Yes	
Region FE		Yes	Yes		Yes	Yes	
Education Control		Yes	Yes		Yes	Yes	

Table 1.12 calculates bounds for the estimated impact of sex ratios on sexual behavior and the likelihood of getting infected with STIs under the assumption that the instrument is correlated with unobservable determinants of our dependent variables, and the bounds estimates use Sex Ratios 2001 as an instrument. Because $\sigma_{X,Z} > 0$, we have a one-sided bound, and a sign identification for $\sigma_{X,U}$ and $\sigma_{Z,U}$ listed in Table 1.11 determines whether this is an upper or a lower bound.1000 wild-cluster bootstrap replications are used for both IV and IIV estimations.

Table 1.13: The effect of the sex ratio on sexual behavior and STIs: **Standard IV** vs. **Lewbel IV**

Dependent Variables	Regressor	Female		Male	
		Standard IV	Lewbel IV	Standard IV	Lewbel IV
		(1)	(2)	(3)	(4)
Panel A		**Sexual behavior**			
concu5y	sexratio	-0.039	-0.083***	-0.177***	-0.152***
	2011	(0.024)	(0.019)	(0.065)	(0.037)
het5yrs	sexratio	-1.503***	-2.240***	-1.230	-3.307***
	2011	(0.229)	(0.233)	(0.953)	(0.697)
serial5y	sexratio	-0.796**	-1.010***	-0.541	-1.898***
	2011	(0.316)	(0.341)	(0.652)	(0.539)
overlap5y	sexratio	-0.064***	-0.102***	-0.057	-0.131***
	2011	(0.020)	(0.022)	(0.067)	(0.049)
concu1y	sexratio	-0.020	-0.019	-0.029	-0.057***
	2011	(0.019)	(0.021)	(0.031)	(0.022)
nocondom1y	sexratio	-0.063	-0.262***	-0.525***	-0.562***
	2011	(0.094)	(0.101)	(0.196)	(0.155)
totusexly	sexratio	-0.092***	-0.044***	0.052	-0.012
	2011	(0.021)	(0.013)	(0.059)	(0.029)
Panel B		**Sexually transmitted infections**			
anysti5y	sexratio	-0.069***	-0.108***	-0.058	-0.022
	2011	(0.018)	(0.015)	(0.035)	(0.029)
chlamydia5y	sexratio	-0.053**	-0.076***	-0.064***	-0.035**
	2011	(0.021)	(0.014)	(0.022)	(0.016)
gonorrh5y	sexratio	-0.012*	-0.003	-0.010*	-0.007
	2011	(0.007)	(0.004)	(0.006)	(0.005)
warts5y	sexratio	-0.009	-0.027***	-0.008	-0.005
	2011	(0.006)	(0.008)	(0.016)	(0.009)
herpes5y	sexratio	0.007	-0.006*	-0.001	0.001
	2011	(0.011)	(0.003)	(0.015)	(0.004)
chlamydialy	sexratio	-0.001	-0.013	-0.012	-0.002
	2011	(0.012)	(0.014)	(0.011)	(0.011)
gonorrhly	sexratio	-0.004*	-0.001	0.002	0.001
	2011	(0.002)	(0.002)	(0.003)	(0.003)
wartsly	sexratio	-0.004	-0.004	-0.004	-0.006
	2011	(0.004)	(0.004)	(0.006)	(0.004)
herpesly	sexratio	0.006	-0.004	-0.003	-0.002
	2011	(0.007)	(0.005)	(0.003)	(0.001)
anystily	sexratio	-0.009	-0.025*	-0.016**	-0.007
	2011	(0.011)	(0.015)	(0.008)	(0.009)
Age FE		Yes	Yes	Yes	Yes
Ethnic FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Note: Refer to Table A1 for the detailed definitions of the variables listed in the table. The instrumental variable for the variable Sex Ratios 2011 is Sex Ratios 2001. Weighted robust standard errors are in parentheses.

Table 1.14: Falsification Tests: The effects of the sex ratio on non-communicable health conditions (OLS and IV)

		Female		Male	
Dependent Variables	Regressor	OLS	2SLS	OLS	2SLS
		(1)	(2)	(3)	(4)
Parkinson's disease	sexratio	0.005	0.013	-0.000	0.003
	2011	(0.003)	(0.011)	(0.001)	(0.004)
		P^A=0.107;P^W=0.547	P^A=0.223;P^W=0.503	P^A=0.931;P^W=0.959	P^A=0.464;P^W=0.557
chronic lung disease	sexratio	-0.010	-0.024	-0.007	-0.016
	2011	(0.012)	(0.033)	(0.005)	(0.010)
		P^A=0.419;P^W=0.591	P^A=0.460;P^W=0.517	P^A=0.161;P^W=0.130	P^A=0.135;P^W=0.226
heart disease	sexratio	-0.004	-0.008	0.012	0.065
	2011	(0.007)	(0.021)	(0.010)	(0.040)
		P^A=0.597;P^W=0.613	P^A=0.700;P^W=0.781	P^A=0.284;P^W=0.194	P^A=0.105;P^W=0.376
stroke	sexratio	0.011	0.066	0.008*	0.032
	2011	(0.012)	(0.047)	(0.004)	(0.025)
		P^A=0.390;P^W=0.553	P^A=0.162;P^W=0.489	P^A=0.080;P^W=0.192	P^A=0.208;P^W=0.394
diabetes	sexratio	-0.058*	-0.140	-0.016	0.020
	2011	(0.027)	(0.087)	(0.024)	(0.064)
		P^A=0.058;P^W=0.328	P^A=0.108;P^W=0.507	P^A=0.537;P^W=0.641	P^A=0.750;P^W=0.731
Age FE		Yes	Yes	Yes	Yes
Ethnic FE		Yes	Yes	Yes	Yes
Region FE		Yes	Yes	Yes	Yes
Education Control		Yes	Yes	Yes	Yes

Notes: The instrumental variable for the variable Sex Ratios 2011 is Sex Ratios 2001. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. 1000 wild bootstrap replications were used. P^A refers to asymptotic p-value, and P^W refers to wild-cluster bootstrap p-value.

Chapter 2

The Mingling of Nations: Racial Diversity, Sexual Activity, and STIs in the UK

2.1 Introduction

Sexually transmitted infections (STIs) continue to be a major public health concern in low and high income countries.[1] In addition to their negative effects on individual economic productivity, these infections engender significant treatment costs.[2] Moreover, it is argued that STIs reinforce each other, as STIs such as herpes simplex virus type 2 and syphilis might be particularly important in the early stages of a localized HIV epidemic (Galvin

[1] In 2016, more than 1 million sexually transmitted infections (STIs) were acquired every day worldwide. According to data gathered from the World Health Organization (WHO), there were an estimated 376 million new infections of the four curable STIs, including chlamydia (127million), gonorrhea (87 million), syphilis (6 million), and trichomoniasis (156 million) (WHO 2018). Moreover, women were more vulnerable to STIs than men.

[2] According to data supplied by the Public Health England, every year since 2008, the number of new STI cases exceeded 415,000 in the United Kingdom. If the current rates of infections continue, treatment costs will reach £3.631 billion between 2015 and 2020, comparable to the health costs of providing services to children, which amounted to £3.225 billion over the same period (Lucas 2015).

and Cohen 2004).

The destructive impacts of STIs have made it important to study their behavioral and social causes. Multiple sexual partnership and sexual network configuration are acknowledged as the two most important proximate determinants of STIs in all societies.[3] It is evident that these two factors are themselves endogenous to structural factors that might be viewed as more exogenous. However, very little research has been conducted on the structural causes of sexual behavior and STIs. In this paper, we study the effect of racial diversity on sexual behavior and STIs risks in the United Kingdom.

Our analysis draws on the theoretical framework proposed by Pongou (2009) to study the effect of ethnic heterogeneity on sexual infidelity and HIV in sub-Saharan African countries. This study argues that, in societies organized around distinct ethnic groups, limited communication between these groups lowers the cost of sexual infidelity and hence makes it more prevalent. The rationale is that if an individual has two sexual partners who are not co-ethnics, then it will be more difficult for any of these cheated individuals to find out about the infidelity of their common partner than if the cheated individuals are co-ethnics. This implies that ethnic heterogeneity provides incentives for sexually active individuals to choose sexual partners from diverse ethnic groups, which leads to large inter-ethnic sexual networks, therefore increasing STIs risks. This theoretical framework also acknowledges that limited communication between ethnic groups might increase the search cost of sexual partners outside of their group, leading to ethnically segregated sexual networks and low risks of sexually transmitted diseases.

We test the conflicting hypotheses generated by this theory in the context of the United Kingdom. This context is different from the African context in several respects. One main difference is that our focus is on racial diversity. Race is more distinguishing individual characteristic than tribe in a society that is racially homogeneous. It might therefore be more or less likely to discover cross-racial infidelity than cross-tribe infidelity. It follows that the empirical implications of this theory are likely to be different in a multiracial

[3]See Morris and Kretzschmar (1997, 2000), Halperin and Epstein (2004), and Mah and Halperin (2010) on multiple sexual partnership, and Adimora and Schoenbach (2005), Helleringer and Kohler (2007), and Pongou and Serrano (2009, 2013) on sexual network configuration.

context.

We empirically test these hypotheses by using individual-level data from England, Scotland, and Wales. In particular, we use the third British National Survey of Sexual Attitudes and Lifestyles conducted in 2010-2012 (Natsal3). The Natsal3 data contain socioeconomic and demographic information on males and females aged 16-74 years. Most importantly, interviewees not only provided detailed information on their sexual behavior but also on whether they were diagnosed with certain STIs at some point in their lifetime, within the past five years before the interview, or within the year preceding the interview. We are, therefore, able to analyze the effects of racial diversity on a number of sexual behaviors and activities including: (1) the number of sexual partners in the past five years; (2) the number of heterosexual partners in the past five years; and (3) the number of most recent sexual partners. We also analyze the effect of racial diversity on several sexually transmitted diseases: (1) chlamydia; (2) gonorrhoea; (3) warts; (4) herpes; and (5) any STI.

Our empirical analysis comprises two main estimation strategies. Firstly, estimating ordinary least squares (OLS) regressions, our baseline strategy, we find that an increase in racial diversity (measured by the index of neighborhood-level ethnic heterogeneity) increases the number of sexual partners, the number of heterosexual partners, and the number of recent sexual partners, which suggests that individuals who live in more racially diverse areas are embedded in large sexual networks. Consistent with these findings, an increase in racial diversity raises the likelihood of acquiring any of the aforementioned sexual infections and diseases.

OLS estimates might be biased in that the choice of neighborhood might not be random. To address this concern, we control for a large number of covariates, including region fixed effects, ethnicity fixed effects, demographic characteristics (gender and age), and individual covariates (namely, educational attainment, religion, sexual identity, employment status, marital status, and whether the respondent is a native). The estimates of interest change very little and remain statistically significant. Specifically, we find that a one-point increase in the racial diversity index increases the total number of sexual partners by 0.644, the

number of heterosexual partners in the past five years by 0.530, and the number of most recent sexual partners by 0.147. As to the prevalence of STIs, we find that increasing neighborhood-level racial diversity clearly increases the likelihood of acquiring infections such as chlamydia, genital warts, and any STIs that have been diagnosed in one's lifetime, in past five years, and one year before the interview. For instance, increasing the racial diversity index by one standard deviation causes the likelihood of getting infected with chlamydia by 0.024 standard deviation.

To address remaining endogeneity concerns, we use two main classes of methodologies. The first class of methodologies consists of assessing bias due to unobserved variables. We use three different methodologies in this class to assess the degree of selection on unobservables in order to gauge the severity of endogeneity bias due to unobserved characteristics from the OLS estimates. These methodologies are proposed by Altonji et al. (2005), Oster (2017) and Gonzalez and Miguel (2015), respectively. They are based on the stability of the coefficient of the endogenous variable and/or the change in the R-squared as more controls are added to the regression. We find that, for each of our outcome variables, selection on unobservables has to significantly greater compared to selection on observables to explain away the estimated effect of racial diversity. This strongly suggests that the OLS estimates are not biased and the estimated effects are causal.

The second class of methodologies we employ are based on an instrumental variable. Inspired by Miguel and Gugerty (2005), we compute an index of racial diversity based on the racial composition of "natives" only, which we use as the instrument for our main predictor. Estimating 2SLS regressions, we confirm our findings. The classical IV estimates are moderately larger and more statistically significant than the OLS estimates.

It is empirically impossible to "prove" that an instrumental variable is perfect in the sense of satisfying the exclusion restriction. For this reason, several recent techniques have been proposed for the estimation of causal effect when the instrumental variable cannot be proved to be perfect. We employ three such techniques for robustness checks. First, we use a methodology developed by Nevo and Rosen (2012) to set-identify the effect of racial diversity on outcomes. This identification strategy weakens the assumption of zero correla-

tion between the instrumental variable and the unobserved structural error that underlies the classical IV methodology and only relies on the directions of the correlation between the IV and the regressor and the correlation between the IV and the error term. Using this methodology, we bound the true effect of racial diversity, and find that our results are in line with those obtained from OLS and IV regressions. Second, we use an alternative identification developed by Lewbel (2012) which relies on a heteroskedastic covariance restriction for causal identification even the instrument does not satisfy the exclusion restriction assumption. This approach yields estimates that are in general larger than our OLS estimates and that are more efficient than the standard IV estimates. Third, we also implement the identification technique introduced by Conley et al. (2012) and that relaxes the exclusion restriction assumption of the classical IV approach. This methodology assumes that the instrumental variable and the dependent variable can be correlated and applies the so-called "union of confidence intervals (UCI)" and "local to zero (LTZ)" approaches to estimate informative bounds around the true effect of the endogenous variable. Our findings from this approach are consistent with the other methods. Altogether, the findings show that racial diversity causally increases risky sexual activity and the spread of STIs.

Investigating, causal mechanisms, we find that individuals residing in more racially diverse neighborhoods tend to select their sexual partners from diverse racial groups. Moreover, within racially diverse neighborhoods, individuals who choose sexual partners from more than one racial group are more likely to be infected with any of the aforementioned STIs than their counterparts who choose partners only from one group. Importantly, the analysis controls for the number of partners and a wide range of individual characteristics.

Furthermore, we conduct some falsification (or reverse-placebo) tests by analyzing the effects of racial diversity on health conditions that are not driven by sexual interactions. These conditions include heart attack, heart disease, stroke, chronic lung disease, Parkinson's disease and epilepsy. Unlike sexually transmitted infections, these diseases are non-contagious. An economically significant effect of racial diversity on any of these falsification outcomes would question our interpretation of the positive impact of racial diversity on

STIs. We find no such effect, which lends more credibility to our mechanisms.

The remainder of this paper is organized as the follows. Section 2 situates our study within the extant literature on the economic and social impacts of ethnic heterogeneity. Section 3 presents our theoretical framework. Section 4 describes the data and identification strategies. Section 5 presents our main findings and findings from robustness and falsification checks. Section 6 concludes.

2.2 Closely Related Literature

The theoretical model used in this paper is an extension of the model of a two-sided fidelity economy in Pongou and Serrano (2009, 2013) to a multi-ethnic environment in which agents derive utility not only from the number of sexual partners but also from their ethnic composition. Pongou (2009) was the first to show a causal link between ethnic heterogeneity and HIV in Africa. Our paper can be viewed as extending this finding to a developed country.

Our paper is also related to the literature on matching. The pathbreaking contributions date back to Koopmans and Beckmann (1957), and Gale and Shapley (1962), and the works of Becker (1973, 1974). Inspired by the original framework of assortative matching by Becker (1973, 1974), a body of recent models on marriage in diverse areas has been produced (Dagsvik 2000, Hatfield and Milgrom 2005, Fox 2010). These studies have either explored models of frictionless matching with transferable utility or non-transferable utility, or investigated models of matching with frictions (Shimer and Smith 2000, Eeckhout and Kircher 2010, 2011, Hagedorn et al. 2017). The insights derived from these theoretical models have been applied to a host of empirical studies, including the marriage market with unbalanced gender distribution (Dahl and Moretti 2008, Abramitzky et al. 2011, Banerjee, Duflo, Ghatak and Lafortune 2013), online dating market (Fisman et al. 2006, 2008, Hitsch et al. 2010), the effect of assortative mating on household income inequality (Cancian and Reed 1998, Fernández and Rogerson 2001, Schwartz 2010, Greenwood et al. 2014), and many others.

Our paper is also related to an extensive literature on homophily. This literature shows that similarity along several dimensions breeds connections in a wide variety of areas including marriage, and friendship (Kalmijn 1998, Fong and Isajiw 2000, Baerveldt et al. 2004, Currarini et al. 2009, 2010, Boucher 2015); information transfer (Schneider et al. 1997, Golub and Jackson 2010, 2012, Acemoglu et al. 2017); organizations (Ruef et al. 2003, Ozcan and Eisenhardt 2009), and employment (Bertrand and Mullainathan 2004, Patacchini and Zenou 2012), among others. These papers show that homophily takes place in relationships that are socially accepted. In contrast, our paper focuses on sexual promiscuity and infidelity, which is a different type of relationship often morally and socially prohibited. We show that individuals tend to form sexual links with other individuals that are culturally distant, in an environment where anonymity is highly prevalent.

Our paper also contributes to a growing literature that studies the impact of ethnic diversity on social, economic, and political outcomes such as social conflict (Collier and Hoeffler 1998, 2004, Fearon and Laitin 2003, Miguel et al. 2004, Blattman and Miguel 2010, Esteban et al. 2012); economic development (Easterly and Levine 1997, Alesina and Ferrara 2005, Montalvo and Reynal-Querol 2005); social trust (Zak and Knack 2001, Alesina and La Ferrara 2002, Putnam 2007); public goods provision (Poterba 1997, Alesina et al. 1999, Miguel 2004, Miguel and Gugerty 2005, Habyarimana et al. 2007, Fernandez and Levy 2008); and the quality of governance (Easterly et al. 2006, Alesina and Zhuravskaya 2011, Lieberman and McClendon 2013). In general, most studies view ethnic diversity as an obstacle rather than a contributor to socio-economic development. These studies argue that, for a variety of reasons, individuals from different cultural and social groups find it difficult to cooperate and make optimal decisions regarding the allocation of public goods. Our paper postulates a different mechanism due to the fact that sexual activity is not a public good. In fact, if individuals prefer to form sexual links only with others from the same racial group, sexual networks would be racially segregated and the spread of STIs would be low. We do not find that this is the case in the UK (See Pongou (2009) for sub-Saharan Africa). Our findings support a mechanism according to which individuals strategically choose their sexual partners from diverse racial groups to hide their infidelity. This is the case despite the fact that it might be more costly to find a sexual partner in

own ethnic group than outside of their group.

2.3 Theory and Testable Implications

In this section, we briefly describe the infidelity model developed by Pongou (2010) to show how ethnic heterogeneity affects sexual network formation, multiple partnerships, and the spread of STIs.

Let e be the number of ethnic groups in a society, and i is an agent who derives his/her utility from having intercourse with sexual partners irrespective of their ethnic background. Let $s = \sum_{h=1}^{e} s_h$ be the total number of partners that i has, s_h is the number of partners from ethnic group h.

Having multiple partners is associated with two types of cost; the first is the search cost, and the second is the infidelity cost. We will assume that the searching partners in group h costs c_h per partner. The infidelity cost arises from the fact that having multiple partners is viewed as infidelity and is costly if found out. Assume agent i has two sexual partners, and let the probability of the two partners finding out about each other's relationship with i be cheated p if they are co-ethnics, but q if they are inter-ethnic, where $p > q > 0$. That is, an individual who has two partners is more likely to be detected if the two partners are co-ethnics than if they are not. We assume that the cost incurred by i per infidelity detection is c. Agent i's utility function is therefore

$$\underset{(s,s_1,...,s_e)}{max}\, u(s, s_1, ..., s_e) = v(s) - \sum_{h=1}^{e} c_h s_h - \sum_{h=1}^{e} s_h[p(s_h - 1) + q(s - s_h)]c$$

subject to:

$$s = \sum_{h=1}^{e} s_h \ (s_h \geq 0, h = 1, ..., e)$$

where v is an increasing and concave function such that $v(0) = 0$ and u is assumed to be twice continuously differentiable. Without loss of generality, assume that agent $i's$ search cost in group G_1 is the least: $c_1 \leq c_h \ \forall h \in \{2, ..., e\}$.

His goal consists of choosing the optimal number and and ethnic composition of partners by solving the following problem:

Let s^* be the optimal number of partners of i, and let $\overrightarrow{s^*} = (s_1^*, ..., s_e^*)$ be the optimal composition of these partners. For simplicity, assume that $c_h = c_2$ for any $h \in \{2, ..., e\}$. Pongou (2010) shows the following result.

Proposition 1

$$\frac{\partial s^*}{\partial e} \begin{cases} = 0 & if\ c_1 \geq c^* & (1) \\ = 0 & if\ c_1 < c^*\ and\ c_2 \geq c^* & (2) \\ > 0 & if\ c_1 < c^*\ and\ c_2 < c^* & (3) \end{cases}$$

Assertions (1) and (2) describe the situations in which either the searching cost for a partner in any groups is very high ($c_2 \geq c_1 \geq c^*$), which precludes having a partner (assertion 1) or forces agent i to select partners only in one group (assertion 2). In both cases, area-level ethnic heterogeneity has no effect on multiple partnerships. In contrast, assertion 3 implies that if the search costs are sufficiently low, then an increase in ethnic heterogeneity increases multiple partnerships.

The main testable implication is that ethnic heterogeneity increases multiple sexual partnerships and STIs risks.

2.4 Data and Identification Strategy

In this section, we test the proposition that area-level racial diversity encourages multiple sexual partners and directly affects the probability of becoming infected with a sexually transmitted infection using data from the National Survey of Sexual Attitudes and Lifestyles 2010 (also named Natsal-3). We describe both our data and the identification strategy below.

2.4.1 National Survey of Sexual Attitudes and Lifestyles

The National Survey of Sexual Attitudes and Lifestyles are ten-yearly cross-sectional national surveys conducted in Britain conducted since 1990. We use the third round of

this survey (Natsal-3) conducted in 2010-2012 as it has a larger sample size and covers a larger age range than the two preceding surveys. These surveys collect information on respondents' socioeconomic and demographic characteristics and sexual behavior. They also collection information on exposure to a range of STIs.

The Natsal-3 selected a representative sample of over 15000 men and women aged 16-74 years dwelling in the 10 regions of England, Wales and Scotland to collect a range of sexual behavior, attitudinal and biological data. The survey selected postcode sectors as the primary sampling units (PSUs), and addresses within these PSUs were chosen at the second stage; then one eligible adult was randomly selected at the final step per household (Erens et al. 2014). Most information was collected using a computer-assisted self-completion format. The methodology and main results of the survey are detailed in Johnson and Mercer (2017).

Data were weighted by age, gender, and geographic region so as to reflect the total population of Britain and to reduce potential bias. Using responses to questions on sexual behavior and the characteristics of sexual partners, we were able to construct variables on number of serial and concurrent partners, and the extent to which an individual selects sexual partners from diverse racial groups. We also have information on whether an individual has been diagnosed with a range of common STIs at some point in their lifetime. Data were also collected on the gender, age, and sexual orientation of respondents and each of their sexual partners.

For our study, the most crucial information provided by the interviewees is that concerning their ethnic group, the number of most recent sexual partners, the number of sexual partners in the year and in the past five years before the survey, and information on their health conditions.[4] The three countries altogether have over 16 ethnic groups, but

[4]The most recent sexual partners represent the total number of opposite-sex and same-sex partners of the interviewees but not limited to partners in the last five years before the interview. Interviewees were asked detailed questions about their most recent partners, such as their gender, age, ethnic belonging, and when and where to meet them, etc.; questions about partners were asked only for a maximum of four partners.

Natsal-3 has classified interviewees into six categories[5] according to their own perceived ethnic group and cultural background, with the white-majority representing over 88% of the pooled sample.

According to Natsal-3, over 90% of the interviewees have opposite-sex partners, and over 6% percent have same-sex partners one year before the survey, but less than 2% have bisexual partners. The prevalence of STIs depends on the type of the infection. Natsal-3 asked questions on whether an individual was diagnosed with several common STIs at some point in their lifetime, within five calendar years before the interview, and within one calendar year preceding the interview. These infections are bacterial infections (chlamydia and gonorrhea), viral infections (genital warts and herpes), and any other infection excluding thrush. The critical distinction between bacterial and viral-induced diseases is that the former are curable, but the latter one can only be treated but are not curable.

It is well-known that in surveys, respondents consistently underreport socially undesirable behaviors and overreport others that socially desirable (Tourangeau and Yan 2007, Krumpal 2013). Empirical studies show that women usually underreport sexual partners, unlike men (Smith 1992). However, self-administered surveys such as Natsal-3 could decrease social desirability bias in answering questions about sexual activity, same-gender sex, and sexually transmitted infections Krumpal (2013). Furthermore, Tourangeau and Smith (1996) show that the discrepancy of the self-reported number of sexual partners between men and women diminished when the questions are self-administered.

In any case, our estimated effect of racial diversity on sexual behavior is likely to be biased downward, implying that our estimates are a lower bound of the true effect.

[5]These six ethnic and racial groups that the respondent considers themselves belonging to are: white(white British, white Irish, and white others), mixed(mixed white and Black Caribbean, mixed white and Black African, mixed white and Asian, and mixed others), Asian/Asian British(Indian, Pakistani, Bangladeshi, and other Asian), Black/Black British(Caribbean, African and other Black), Chinese, and other ethnic group.

Independent variable

Our main independent variable is the Hirschman-Herfindahl index (HHI) of ethnolinguistic fractionalization computed at the neighborhood level. The index measures the probability that two randomly selected individuals from the population belong to different ethnic groups. It is computed as follows:

$$HHI = 1 - \sum_{i=1}^{e} \frac{N_i^2}{N^2} \tag{1}$$

where e is the number of ethnic groups in the area, N_i is the size of ethnic group i, and N is the total population size in the neighborhood. The summary statistics of the variables used in this analysis are presented in Table 2.1. It can be seen from Table 2.1 that the average individual dwells in a neighborhood where two random selected persons belong to different ethnic groups with a probability of 0.19. Figure 2.1 presents the distribution of the ethnic heterogeneity index (HHI). The index ranges from 0 to 1, and has a large support, which shows that it varies substantially in the population, an essential feature for econometric identification.

Dependent variables: Various sexual activities and STI status

The dependent variables are: (1) the reported total number of sexual partners (hetero and same-sex); (2) the number of heterosexual partners; (3) the number of most recent partners; (4) an indicator variable for whether an individual has been diagnosed with chlamydia at some point in their lifetime; (5) an indicator variable for whether an individual has been diagnosed with gonorrhea at some point in their lifetime; (6) an indicator variable for whether an individual has been diagnosed with genital warts at some point in their lifetime; (7) an indicator variable for whether an individual has been diagnosed with herpes at some point in their lifetime; (8) an indicator variable for whether an individual has been diagnosed with any STI at some point in their lifetime; (9) an indicator variable for whether an individual has been diagnosed with any STI in the past five years preceding the survey; and (10) an indicator variable for whether an individual has been diagnosed with any STI in one year preceding the survey.

Table 2.2 shows summary statistics for the outcome variables. In Panel A, we see that an average individual has 2.752 partners (including opposite-sex and same-sex partners) and 2.573 heterosexual partners. But the average number of recent partners is lower (1.647). Panel B shows the probability of infection with each of the STIs listed above. Chlamydia (6.3%) is the most commonly diagnosed bacterial STI in the UK.[6] Gonorrhea (1.3%) is the second most commonly reported bacterial STI. The prevalence of other STIs is comparatively low. The probability of being infected with any STI was 17% in one's life, 4.8% in the past five years, and 1.3% just one year preceding the interview.

2.4.2 Ethnic Heterogeneity and Number of Sexual Partners and STIs: A Graphical Description

Figure 2.3 describes the relationship between neighborhood-level ethnic heterogeneity and each of our dependent variables. Individuals are grouped by HHI quintile. The average number of sexual partners and STI prevalence are calculated for each quintile. In general, ethnic heterogeneity has a positive effect both the total number of opposite-sex and same-sex partners, the number of heterosexual partners in the past five years, and the number of most recent sexual partners (Figure 2.3a, Figure 2.3b, and Figure 2.3c). Similarly, individual living in a more ethnically heterogeneous neighborhood is associated with a higher probability of being infected with any of the STIs we study (Figure 2.3d, Figure 2.3e, and Figure 2.3f).

These effects cannot be interpreted as causal because individuals who have a high demand for sex might self-select into more ethnically heterogeneous neighborhoods. We directly address this potential endogeneity issue in several ways: (1) we control for pertinent variables; (2) we use classical and recent instrumental variables approaches; and (3) we assess and control for the selection of unobservable variables.

[6] 3.3% of the interviewees were diagnosed with chlamydia in the past five year preceding the survey and only 0.8% were diagnosed with this disease in the past year. Because of these low prevalences, we do not analyse STI in the past five year or in the past year. Instead, we analyse being diagnosed with any STI in the past five years and in the past year.

2.4.3 Regression-based Results: Baseline Model

We use the following baseline regression model to estimate the causal effect of racial diversity on number of sexual partners and STI status:

$$y_{ien} = \alpha + \beta * HHI_n + X'_{ien}\theta + \sigma_e + \phi_r + \epsilon_{ien} \tag{2}$$

where y_{ien} is an outcome variable of interest (number of sexual partners or STI) for individual i who belongs to an ethnic group e and dwells in the neighborhood n; HHI_n is the index of ethnic heterogeneity for neighborhood n; and X'_{ien} is a vector of individual-level and neighborhood-level controls. These control variables are thought to be correlated both with our main predictor (HHI) and outcomes, and therefore address potential endogeneity problems. Our main parameter of interest is β, which measures the marginal effect of HHI on the outcome. The term σ_e is an ethnicity fixed effect; ϕ_r is a region fixed effect; and ϵ_{ien} is an unobserved error term assumed to be uncorrelated with regressor given the controls. We use ordinary least square (OLS) regressions to estimate the effect of HHI on number of sexual partners and STI status.[7]

Controls are included in two steps; we first include a restricted set of controls, and then include all the controls to evaluate the sensitivity of the coefficient of interest. Table 2.3 and Table 2.4 present the OLS estimates for number of sexual partners and STI status, respectively. Columns (1) and (2) of Table 2.3 report the estimated effect of HHI on the total number of partners in the past five years. The controls included in Column (1) are ethnicity and region fixed effects. Column (2) additionally controls for gender, age, educational attainment, religion, sexual identity, employment status, marital status,

[7]Remember that the dependent variables of sexual behavior are the total number of sexual partners in the last five years, the number of heterosexual partners in the last five years, and the number of most recent partners preceding the interview. We estimate the model using an ordinary least squares (OLS) regression for the ease of interpretation. However, since all three sexual behavior variables are not genuinely continuous, we also estimated a Poisson regression model with the count dependent variables. These results, which are reported in Table B2, reveal that our inferences are robust to the choice of estimator. Finally, all OLS analyses employ a wild cluster-bootstrap (Cameron et al. 2008) due to the small number of clusters in our data, and robust standard errors clustered at the regional level.

and an indicator whether the respondent is a native. The estimates indicate that an individual who resides in a fully heterogeneous neighborhood has 2 more partners than his/her counterpart residing in a fully homogeneous neighborhood. But the coefficient of HHI drops to 0.6 when all the controls are included. These results show that racial diversity significantly increases sexual activity.

In Columns (3) and (4), the outcome variable is the number of opposite-sex partners in the past five years. Here too, racial diversity has a positive and significant effect. Columns (5) and (6) analyze the effect of HHI on the number of most recent sexual partners, also finding a positive and statistically significant effect. Altogether, the findings suggest that living in a more racially diverse neighborhood significantly increases sexual activity as measured by the number of sexual partners.

Table 2.4 reports the estimated effect of racial diversity of STI status. We analyze binary indicators for being diagnosed with the following infections: chlamydia, gonorrhea, warts, herpes, any STI in one's lifetime, any STI in the past five years, and any STI in the year preceding the survey. To compare coefficients across different outcomes, we also report the beta coefficient in brackets, which measures how much the dependent variable changes as the result of a one-percent increase in the standard deviation of HHI. We find that racial diversity significantly increases the likelihood of any of these outcomes. For instance, moving from a fully homogeneous neighborhood to a fully heterogeneous neighborhood increases the likelihood of being infected with chlamydia by 5.5 percentage points with a restricted set of controls and by 2.6 percentage points with a full set of controls. Similarly, a marginal increase in racial diversity increases the likelihood of being infected with any STI in one's life time by 11 (resp. 9.9) percentage points with a restricted (resp. full) set of controls. We also note that the coefficients are relatively stable for most outcomes, which indicates endogeneity problems are not severe.

2.4.4 Addressing Endogeneity Issues

The OLS analysis in Section 2.4.3 could suffer from an endogeneity problem owing to the fact that individuals are not randomly assigned to their neighborhoods. It is possible that

individuals who reside in more racially diverse neighborhoods are different from those who reside in more homogeneous neighborhoods, which would prevent a causal interpretation of the estimates presented in Tables 2.3 and 2.4. We attempt to address this endogeneity concern by controlling for a wide range of variables likely to affect both our predictor and the outcome variables. We have seen that including these controls do not affect our conclusion qualitatively. However, there might be remaining endogeneity problems due to unobservables. We address them using classical and new econometric approaches to dealing with such issues. First, we assess potential bias from unobservables (Section 2.4.4.1). Second, we use the classical instrumental variable approach (Section 2.4.4.2). Third, we use three additional recent approaches that deal with imperfect instruments.

Assessing Potential Bias from Unobservables

In this section, we assess the degree of selection on unobservables to gauge the severity of endogeneity bias due to unobserved characteristics from the OLS estimates. We use three different approaches proposed by Altonji et al. (2005), Oster (2017) and Gonzalez and Miguel (2015), respectively. The approach was proposed by Altonji et al. (2005) who measures the degree of selection on the unobservables by using the degree of selection on observed variables, in order to determine how much stronger selection on unobservables would have to be compared to selection on observables in order to fully explain away our result. For this purpose, we compute the ratio $\frac{\hat{\beta_F}}{(\hat{\beta_R} - \hat{\beta_F})}$, where $\hat{\beta_F}$ is the estimated coefficient of HHI obtained from a regression that includes a full set of controls, and $\hat{\beta_R}$ is the estimated coefficient of HHI obtained from a regression that includes only a restricted set of controls.

In column (1) of Table 2.5, we report the Altonji et al. (2005) ratios for each of the 10 outcome variables listed in Table 2.2. The full set of controls includes gender, age, educational attainment, religion, sexual identity, employment status, marital status, region fixed effect, ethnicity fixed effect, and an indicator for whether the respondent is a native. The restricted set of controls only includes age, region fixed effect, and ethnicity fixed effect. The reported ratios in Table 2.5 range from -29.583 to 10.588. which are obviously

very large. All these ratios imply that the influence of unobservable factors would have to be far greater (at least 10 times greater) than that of observable characteristics to fully account for our results. It is therefore unlikely that the estimated effect of HHI on any of the outcomes is driven by unobservables.

The second approach we employ was developed by Oster (2017), who introduced the following equation of the bias-adjusted estimator: $\beta^* = \hat{\beta_F} - (\hat{\beta_R} - \hat{\beta_F}) \times \frac{R_{max} - R^*}{R^* - R}$. This estimator calculates the lower bound for the coefficient of interest, where $\hat{\beta_F}$ is the coefficient estimates with a full set of controls, $\hat{\beta_R}$ is the coefficient estimate with a restricted set of controls, R^* is the R^2 (coefficient of variation) from the regression with a full set of controls, and R is the R^2 from a regression with a restricted set of controls. For the most conservative case, $R_{max} = 1$ when the regression includes all the observable and unobservable variables. In the real world, however, measurement errors are inevitable, and the value of R_{max} should therefore be less than one suggested by Gonzalez and Miguel (2015). We assume $R_{max} = 1$ and three other values of $R_{max} = min\{\Pi R^*, 1\}$ suggested by Oster (2017), with the three suggested values of Π being 1.2, 1.4 and 1.6.

Columns (2)-(5) in Table 2.5 produce the lower bounds of the coefficients of HHI corresponding $R_{max} = 1$ and three suggested values of R_{max}. The most conservative case shown in the column (2), where no measurement error is assumed, is too unrealistic to produce informative bounds; therefore, most of the coefficient intervals do not exclude zero. Now we consider the most realistic cases suggested by Oster (2017), where $R_{max} = min\{\Pi R^*, 1\}$ with varying values of Π. For the total number of sexual partners as the dependent variable, for instance, the coefficient of HHI lies in the following intervals [0.476, 0.644], [0.309, 0.644], and [0.141, 0.644] when Π equals to 1.2, 1.4 and 1.6, respectively. We see that the lower bound is greater than zero in each of these cases, which implies that racial diversity has a significantly positive effect on sexual activity.

The third approach proposed by Gonzalez and Miguel (2015) proposes that the formula for the computation of the lower bound of the parameter of interest should assumes $R_{max} = R^* + (R^* - R)$. The results are reported in Column (6) in Table 2.5. Here too, we find that the lower bound is greater than zero for most outcomes. All these results imply that it is

unlikely that our OLS estimates are biased by some unobserved characteristics; therefore, the relationship between ethnic heterogeneity and each of our outcomes can be interpreted as causal.

Instrumental Variable

In this section, we use the instrumental variable approach to address possible endogeneity issues in our OLS estimates. Inspired by Miguel and Gugerty (2005) who use the ethnic heterogeneity of natives as the instrument for school ethnic diversity to estimate the impact of the latter on public goods provisions in sub-Saharan Africa, we use the ethnic diversity of natives as an instrument for HHI. This instrument is sensible because it is not driven by migration of respondents. Figure 2.2 presents the distribution of native-induced ethnic heterogeneity. The identification assumption is that the instrument only affects the outcome through HHI.[8] We compute an index for native-induced ethnic diversity using Equation (1), where N_i and N stand for the size of a group i natively and the total native population of natives, respectively.

In the first stage, we regress our endogenous HHI on the instrumental variable following the equation below:

$$HHI_n = \alpha_1 + \beta_1 * HHI_n^{native} + X_{ien}^{'}\theta_1 + \sigma_e + \phi_r + \epsilon_{1ien} \tag{3}$$

In the second stage, we regress our outcome variable on the predicted value of the endogenous variable obtained from the first stage regression as in the following equation:

$$y_{ien} = \alpha_2 + \beta_2 * \widehat{HHI_n} + X_{ien}^{'}\theta_2 + \sigma_e + \phi_r + \epsilon_{2ien} \tag{4}$$

The first-stage regressions are presented in Table 2.6. Column (1) Column (1) controls only for region fixed effect and ethnicity fixed effect, and column (2) includes the full set of controls as discussed in Section 2.4.3. The results imply that our instrument is a strong and significant predictor of neighborhood-level racial diversity. The F-test for the excluded

[8] A native is someone who has not lived elsewhere since birth.

instrument is above 10 for both specifications, while the p-values based on the Anderson-Rubin test statistical significance are small. The instrumental variable (IV) estimates are reported in Column (2) of Table 2.7, with Column (1) reporting the OLS estimates for comparison. The IV estimates are generally somewhat larger and more statistically significant compared to the OLS estimates.

Robustness to Alternative IV Approaches

A critical assumption in the classical IV approach is the exclusion restriction, which means that an instrumental variable should affect the outcome variable only through the endogenous variable. In practice, this assumption is very difficult to test. And given the difficulty of proving formally that an instrument is perfect in the sense of satisfying the exclusion restriction, it has become common to assume that instrumental variables are not so perfect, even when there is no evidence showing that this is indeed the case. This can be viewed as a conservative assumption. Several recent econometric techniques have been introduced to deal with imperfect instruments by relaxing the exclusion restriction assumption. In this section, we use three such techniques to assess the robustness of our findings. First, we use the methodology developed by Nevo and Rosen (2012) to set-identify the effect of an endogenous variable using an imperfect instrument. Here, an instrument is imperfect when it is correlated with the unobserved error term of the equation to be identified. Second, we use the approach developed by Lewbel (2012) who provides an identification method when the validity of the external instrumental variables can be challenged or when external instrumental variables are not available. In this approach, a "perfect" instrument is found by construction, although it is not always possible to tell a "real-life" story about it. The third methodology that we use is the "plausibly exogenous" IV approach developed by Conley et al. (2012). In this approach, the definition of plausibly exogenous instruments is based on priors over the failure of the exclusion restriction.

Robustness to the Nevo and Rosen's (2012) Imperfect IV Approach We use Nevo and Rosen (2012)'s IIV approach to set-identify the effect of the ethnic heterogeneity

on sexual behavior and STIs. This methodology relaxes the assumption of instrument exogeneity that underlies the traditional instrumental variable approach. Its underlying assumption can there be viewed as realistic given that it is almost impossible to prove formally that an instrument is fully exogenous. Its main assumption is that the correlation between the instrument and the unobserved error term of the equation to be identified is smaller than the correlation between the structural error and the endogenous regressor; that is: $|\rho_{XU}| \geqslant |\rho_{ZU}|$, where X stands for the endogenous variable, Z is an instrumental variable, and U is a mean zero unobserved error term. It is also assumed the two correlations are of the same sign.

Because the correlations between HHI and native-induced HHI (our instrument) is positive ($\sigma_{XZ} > 0$), we may actually only have a one-sided bounded interval instead of a two-sided bounded interval, and the sign of $\sigma_{XU} * \sigma_{ZU}$ will determine whether the return result is an upper bound or a lower bound.[9] In order to obtain a two-sided bounded interval following Nevo and Rosen (2012), we use $1 - HHI_{native}$ as the instrument instead of the native-induced HHI. Note that this does not have any impacts on the estimated coefficients, but only allows to derive a two-sided bound.

Column (3) and (4) in Table 2.7 presents the estimates of the upper and lower bound of the true effect of HHI on each of the three outcomes measuring the number of sexual partners and on each STI outcome analyzed in this paper. Remark that the upper bound is given by the standard IV estimate. Then, under Nevo and Rosen (2012)'s propositions listed in Table 2.8, what we need is to determine the sign of σ_{XU} and check whether the Assumption 4 is satisfied or not. Based on the lower bound estimator and upper bound estimates presented in columns (3) and (4), When $\sigma_{XU} < 0$ and Assumption 4 is satisfied, the effect (β) of HHI on the total number of sexual partners in the past five years (tot5yrs), the number of heterosexual partners in the past five years (het5yrs), chlamydia, herpes, the binary indicator for ever been diagnosed with an STI (eversti), and the binary indicator for being diagnosed with an STI in the past five years (anysti5yrs) is such that $\beta^{iv}_{v(1)} \leqslant \beta \leqslant \beta^{iv}_{native}$, where $\beta^{iv}_{v(1)}$ is the 2SLS estimator of a transformed valid instrument,

[9]The positive sign of $\sigma_{XU} * \sigma_{ZU}$ indicates that the imperfect IV estimates of the upper bound of the effect of HHI are well within or around the confidence intervals of the 2SLS estimates.

denoted $V(\lambda^*) = (\sigma_X Z - \lambda^* \sigma_Z X)$ with $V(1) = (\sigma_X Z - 1\sigma_Z X)$. We find that bounds are wider when the independent variable of interest is more exogenous, vice versa. However, in a few cases where Assumption 4 is not satisfied and $\sigma_{XU} > 0$, such when the dependent variable is the number of most recent sexual partners (mostrecent), gonorrhea status, or warts status, the effect (β) of HHI is such that $\beta^{iv}_{native} \leqslant \beta \leqslant \beta^{ols}$; and when $\sigma_{XU} < 0$, for instance, the binary indicator for being diagnosed with an STI in the year preceding the survey (anysti1yr), the effect (β) of HHI is such that $\beta^{ols} \leqslant \beta \leqslant \beta^{iv}_{native}$. In either case, we note that, for most outcomes, the effect of HHI lies in an interval that does not contain zero, which means that ethnic diversity causally increases sexual activity and the likelihood of getting infected with an STI.

Robustness to Lewbel IV In this section, we employ the instrumental variable approach developed by Lewbel (2012) to identify the effect of an endogenous regressor when the validity of external instrumental variables can be challenged or when such instruments are not available. Here, an instrument is found by construction. According to this methodology, β in Equation 2 can be identified based only on the heteroskedastic distribution of the error term ϵ_{1ien} in the first-stage (Equation 3). The endogenous regressor can be estimated as in a linear 2SLS regressions by using X and $\left(Z - \overline{Z}\right)\widehat{\epsilon_{1ien}}$ as instruments, where X is a vector of observed exogenous variables such as age, gender, ethnicity, residential area and education level, and Z can be a subset of the regressors. The identification assumption requires that $Cov(Z, \epsilon_j^2) \neq 0$ and $Cov(Z, \epsilon_1 \epsilon_2) = 0$ (for $j = 2$ in the presence of a triangular model, and for both $j = 1$ and $j = 2$ in a reverse causality system), and there are no further restrictions imposed on Z.

We find that the Pagan and Hall (1983) test rejects the null hypothesis of homoskedasticity with a P value equals to 0.000 in our model. It follows that the Lewbel IV approach can be applied to our analysis. The findings are reported in Column (5) of Table 2.7. Column (5) presents the result of Lewbel IV with both the external instrument (native-induced HHI) and the internally generated instruments. The results generally show that our estimates obtained from the Lewbel's approach are a bit larger than those obtained from the OLS estimation, but are a bit smaller compared to the classical IV estimates.

However, as demonstrated in Lewbel (2012), the Lewbel IV estimates are more efficient, with smaller standard errors for the coefficients. Moreover, the p-value of the Hansen's J statistic of all estimations cannot reject the null hypothesis of exogeneity of the Lewbel IVs.

Robustness to Conley et al.'s (2012) IV bounds Conley et al. (2012) provide practical methods for performing causal inference when the exclusion restriction that underlies the validity of the classical IV inference is debatable. Their approaches assume that the instrument is only plausibly (or approximately) exogenous. That is, the instrument may be correlated with the error term of the equation to be identified, but the correlation is not high. To illustrate, consider the following equation that follows from Conley et al. (2012) and that is applied to our context:

$$y_{ien} = \alpha + \beta * HHI_n + \gamma * HHI_n^{native} + X_{ien}^{'}\theta + \sigma_e + \phi_r + \epsilon_{ien} \tag{5}$$

In Equation (5), when HHI_n is endogenous, the parameters β and γ are not jointly identified. The exclusion restriction assumption underlying the classical IV approach implies that $\gamma = 0$. According to Conley et al. (2012), rather than assuming that $\gamma = 0$, one can define plausible exogeneity as a situation where γ is close to 0, but is not exactly equal to zero. They propose four complementary inference strategies that use prior information about γ. The first strategy specifies the set of possible values for γ. The second and third strategies use prior information about the distribution of potential values of γ but do not fully specify the error distributions. The fourth strategy corresponds to a Bayesian analysis that requires prior information about all model parameters and assumptions about the error distributions. Two of these four approaches are commonly used in the literature (e.g., Bhalotra and Clarke (2018)), which are the ones we use in this paper.

First, we assume a range of values that γ may take: $\gamma \in [\gamma_{min}, \gamma_{max}]$, where the values could be entirely positive or negative, or could contain zero. We can then obtain interval estimates for β, the treatment parameter of interest, for each value of γ. A union of these interval estimates across different γ values provides a conservative interval estimate for β. This approach is also known as the "union of confidence intervals (UCI)" approach.

Second, we assume that γ follows some arbitrary distribution, and estimate the bounds on the parameter β using the entire assumed distribution for γ. This second approach is known as the "local to zero (LTZ) " approach. We follow a similar procedure as Bhalotra and Clarke (2018) did to implement both strategies. We first simply assume a broad range of values for γ in the asymmetric case where we assume a positive bias formalized as $\gamma \in [0, 2\hat{\gamma}]$, where the true $\hat{\gamma}$ in each case will lie precisely in the middle of the confidence interval . Conley et al. (2012) show that, in our case, as long as $\gamma \in [0, 2\hat{\gamma}]$, the confidence intevals will asymptotically contain the true estimates of β at the 95% level. The second method used is based on the LTZ approach assumes a normal distribution of γ with mean 0 and standard deviation δ: $\gamma \sim N(0, \delta^2)$.

Figures (2.4) and (2.5) present both the midpoints in the case of the LTZ approach and the asymptotic coverage of the 95% confidence intervals in the case of the UCI approach. Figure (2.4) provides the estimates for the dependent variables measuring sexual behavior, and Figure (2.5) provides estimates for the dependent variables measuring STI status. In these figures, the x-axis indicates the possible values of δ, the standard deviation of γ.

The estimates derived from the two approaches are presented in Table 2.9. Columns (1) and (2) show the lower bound and the upper bound from the UCI approach and columns (3) and (4) show similar results from the LTZ approach. The mid-point of the bounds is always positive, and the bounds are most informative for the binary indicators for ever been diagnosed with an STI (eversti) and being diagnosed with an STI in the past five years (anysti5yrs). In general, we find that the bounds are informative even assuming that the exclusion restriction is violated. Taken together, the consistently statistically significant estimates from this alternative identification method provide us with robust evidence on the positive effect of the HHI on sexual activity and the likelihood of being infected with various STIs.

2.5 Mechanism

2.5.1 Strategic Diversification of Sexual Partners

Our theoretical framework suggests one mechanism that may explain the greater number of sexual partners for individuals residing in multiracial neighborhoods than for their counterparts living in the more racially homogeneous neighborhoods. It proves that "partial" ethnic integration leads to the strategic diversification of sexual partners across ethnic groups so as to hide infidelity and decrease its cost. This optimizing behavior implies a mechanism in which ethnic heterogeneity causes individuals to diversify their sexual partners across ethnic groups, leading to a large interethnic sexual network, which increases the risk of getting infected with sexually transmitted diseases (see the mechanism framework in Figure 2.6).

To test for this mechanism, we compute a partner ethnic diversification index, and use it to test two hypotheses. The first hypothesis is that individuals who live in a multiracial neighborhood are more likely to diversify their sexual partners than their counterparts residing in more racially homogeneous areas. The second hypothesis is that, fixing the number of sexual partners, an individual is more likely to be infected with an STI if his/her partners are more ethnically diverse. We first compute an index that measures the fraction of inter-ethnic links of an individual of a specific ethnicity, and we use this index to determine whether the individual diversifies his/her sexual partners or not. This measure is the ethnic diversification of i's partners and is defined as follows:

$$EHI(i's\ partner) = 1 - \sum_{j=1}^{e} \frac{s_j^2}{s^2}$$

where s_i denotes the number of links that an individual i has in a racial group j, and s indicates the total number of links that i has across the e ethnic groups. The index ranges from zero (no diversity) to one (full diversity). From this partner ethnic heterogeneity index, we construct our partner diversification index which is a dummy variable that takes the value of one if $EHI(i's\ partner)$ is greater than zero, zero otherwise.

Figure 2.7 describes the relationship between neighborhood-level ethnic heterogeneity and the extent of sexual partner diversification. Similar to what we have done in Section (2.4.2), individuals are grouped by the HHI quintile. The average degree of partner diversification is calculated for each quintile. In general, individuals living in a more ethnically heterogeneous neighborhood are associated with a higher probability of diversifying sexual partner(s), no matter whether we consider the region and ethnicity fixed effects (Figure 2.7a and Figure 2.7b). In addition, Figure 2.8 demonstrates the relationship between neighborhood-level ethnic heterogeneity and concurrency.[10] Not surprisingly, individuals living in more ethnically heterogeneous neighborhoods are more likely to be involved in overlapping relationships.

We also test our second hypothesis that individuals who diversify sexual partners more are more likely to be infected with an STI. Figure 2.9 shows how the prevalence of several STIs varies according to whether individuals diversify sexual partners or not. Clearly, for each STI, we see that those who diversify sexual partners are more likely to be infected compared to their counterparts who do not diversify. We further explore this relationship in a regression-based framework in which we control for the number of sexual partners and a wide range of controls. We estimate the following model:

$$y_{ien} = \beta_0 + \beta_1 diverse_{ien} + \beta_2 partners_i + X'_{ien}\gamma + \sigma_e + \phi_r + \epsilon_{ien} \tag{6}$$

The dependent variable y_{ien} is a binary indicator for whether individual i has an STI or not. Our explanatory variable of interest, $diverse_{ien}$, indicates whether an individual i who belongs to an ethnic group e and resides in a neighborhood n diversifies sexual partners; $partners_i$ is a vector of categorical variable for individual i's number of sexual partners; X'_{ien} is a vector of individual-level and neighborhood-level controls. The term σ_e is an ethnicity fixed effect; ϕ_r is a region fixed effect; and ϵ_{ien} is an unobserved error term assumed to be uncorrelated with the main regressor given the controls. Finally, to correct for the possibility that the error terms ϵ_{ien} are correlated across regions and racial

[10]Here, the concurrent relationship means whether an individual had two overlapping partners in the past five years.

groups, we use the multi-way cluster strategy proposed by Cameron et al. (2011) and cluster standard errors on the basis of both the region and the ethnicity.

Results from Table 2.10 show positive and significant effects of the sexual partner diversification index on the risk of getting infected with STIs. With respect to chlamydia (column (1) in Table 2.10), for instance, the results suggest the risk of infection is 2 percentage points higher if an individual diversifies sexual partners than if he/she does not. This effect is statistically significant at the one percent level. Although the robust standard errors clustered within regions are larger than those adjusted for two-way clustering, most of the estimates are still statistically significant at least at the 10 percent level. Figure 2.10 presents the estimated effect of sexual partner diversification on each outcome of interest with the 95% of the confidence interval.

The findings show that the positive effects of neighborhood-level racial diversity on sexual activity and STIs are primarily driven by individuals who diversify sexual partners in more racially diverse areas. These individuals are more central to the sexual network and therefore are more likely to be infected and to infect others.

2.5.2 Non-Communicable Conditions: A Reverse-Placebo Test

In this section, we conduct a falsification or reverse-placebo test by analyzing the effect of HHI on health conditions unlikely to be caused by sexual interactions. The conditions we analyze are non-communicable conditions, including heart attack, heart disease, stroke, chronic lung disease, Parkinson's disease and epilepsy. Our test attempts to rule shed additional light on the mechanism through which HHI affects sexually transmitted infections. We estimate Equation (2), where the outcome variable is each of the aforementioned health conditions. The OLS and IV estimates are presented in Table 2.11. We find that neighborhood-level ethnic heterogeneity has no effect on any of these conditions. In general, the estimates are very close to zero. The fact that we do not find any consistent effect of ethnic heterogeneity on any of these conditions lends more credibility to our proposed mechanism that ethnic heterogeneity affects the spread of STIs by increasing sexual activity.

2.6 Conclusion

This paper studies the effect of racial and ethnic heterogeneity on sexual activity and the spread of STIs in the United Kingdom based on a model of sexual infidelity. Incomplete or partial integration among ethnic groups due to cultural barriers drives incentives for agents to optimize their number of sexual partners by choosing them from diverse ethnic groups. Ethnic heterogeneity therefore affects the spread of STIs by increasing multiple partnership and by inducing large sexual networks that interconnect different ethnic groups. Using a rich dataset from the United Kingdom, we validate these predictions by showing that ethnic heterogeneity increases sexual activity and the likelihood of acquiring various types of STIs. Different identification techniques are employed to show that these effects are causal. Analyzing the mechanism, we find that individuals residing in more racially diverse neighborhoods are more likely to diversify sexual partners. Moreover, within racially diverse neighborhoods, individuals who diversify sexual partners are more likely to be infected with STIs. The findings suggest that these individuals are more central to the sexual network and therefore are more likely to be infected and to infect others.

Our study offers another perspective on how ethnic segregation encourages risky sexual behavior and harms the prevention of STIs in ethnically heterogeneous societies. Therefore, public policies aimed at reducing the spread of sexually transmitted infections should promote social and ethnic integration. From a policy perspective, our analysis implies that policies of ethnic integration should be implemented in ethnically heterogeneous societies. If implemented successfully, these policies will increase information circulation between different ethnic groups, which in turn will discourage the tendency to choose sexual partners from diverse ethnic groups to hide infidelity. Policies that raise awareness about the danger of strategic cross-ethnic sexual mixing should can also be encouraged, and should be easy to implement.

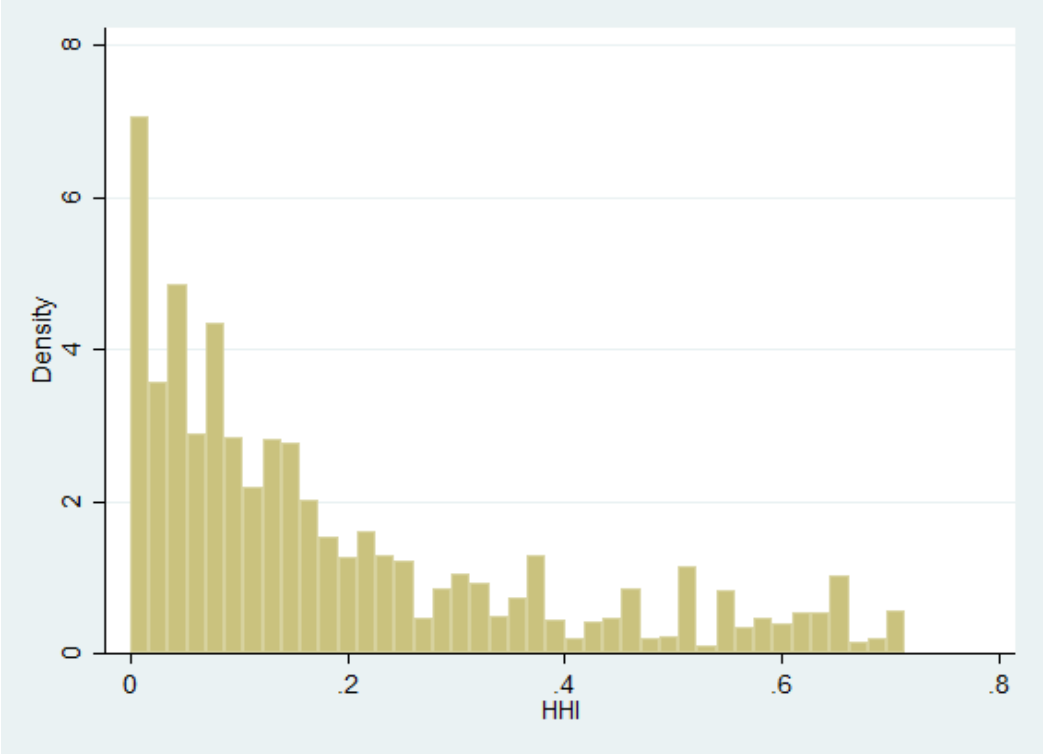


Figure 2.1: Distribution of neighborhood-level ethnic heterogeneity (HHI)

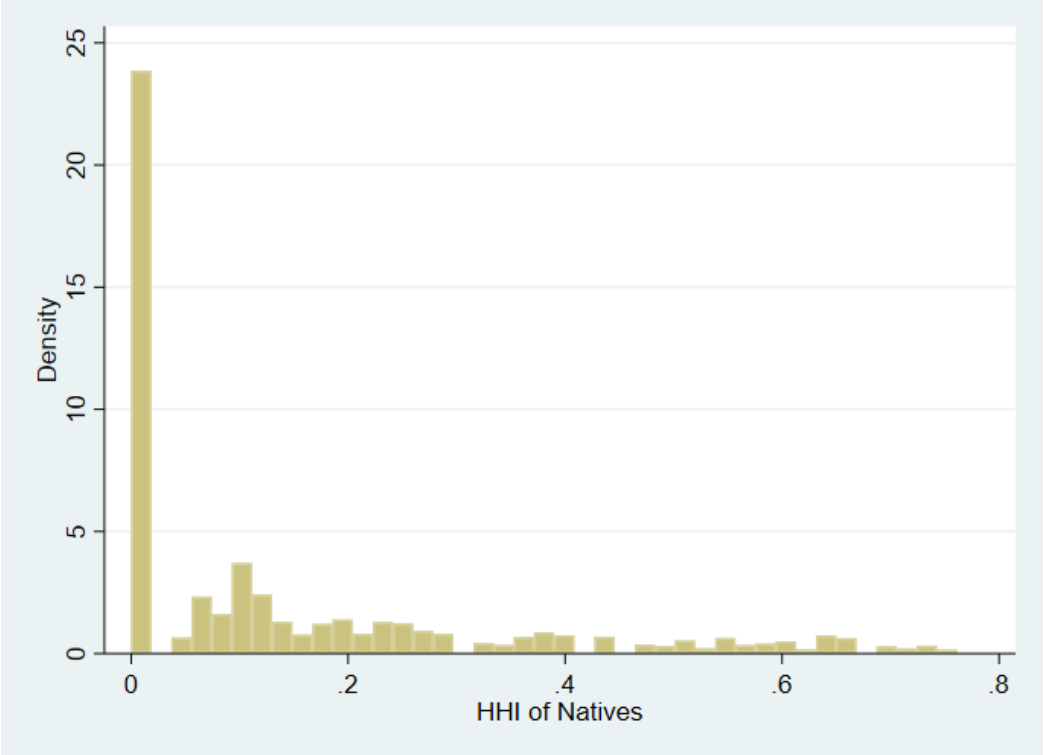
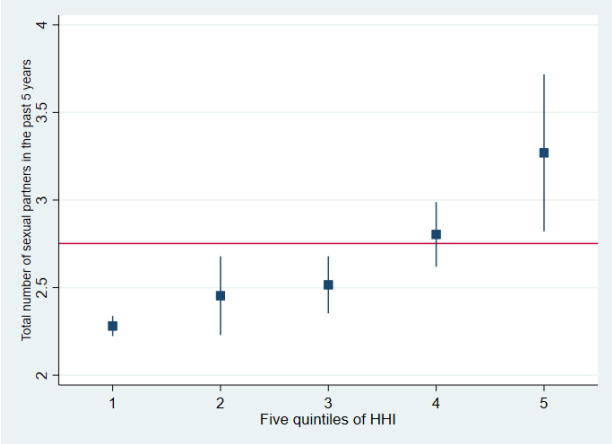
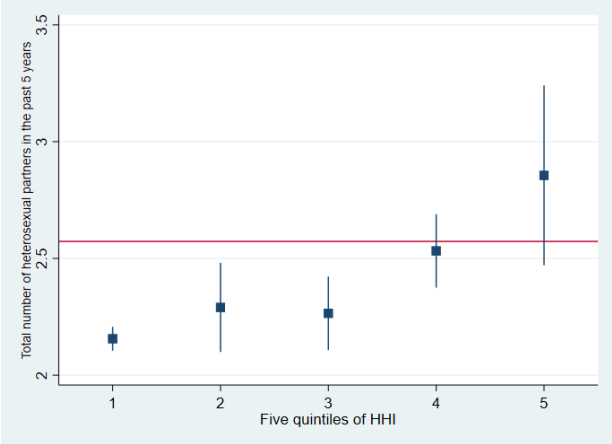


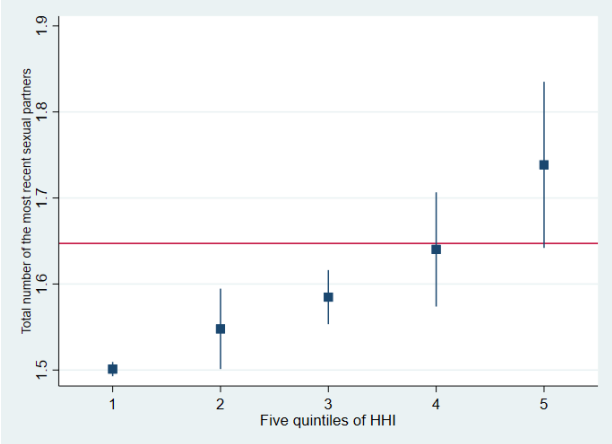
Figure 2.2: Distribution of native-induced neighborhood-level ethnic heterogeneity (native_HHI)



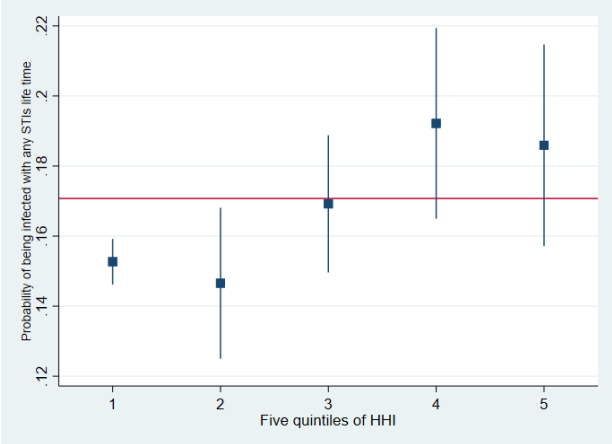
(a) Total number of sexual partners in past 5 years



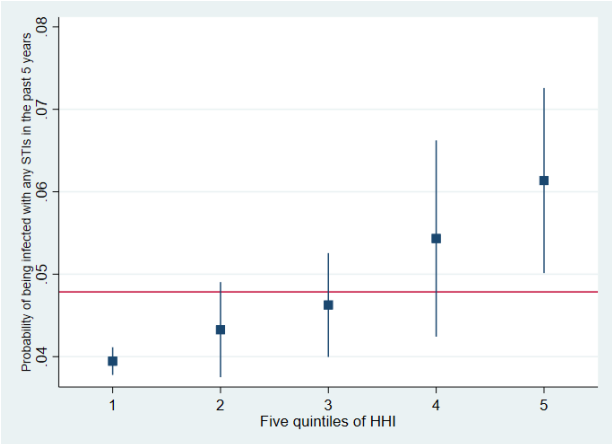
(b) Number of heterosexual partners in past 5 years



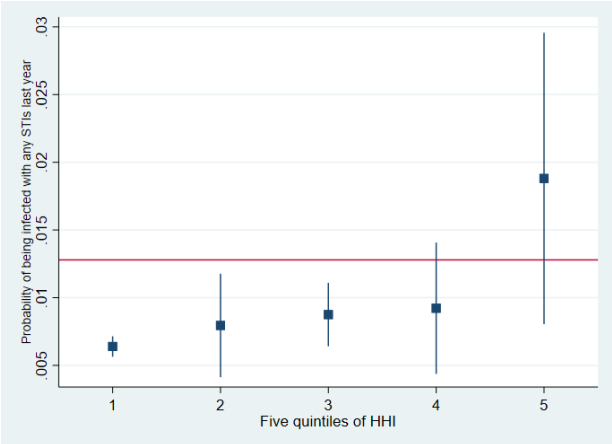
(c) Number of most recent sexual partners



(d) Ever been diagnosed with STIs

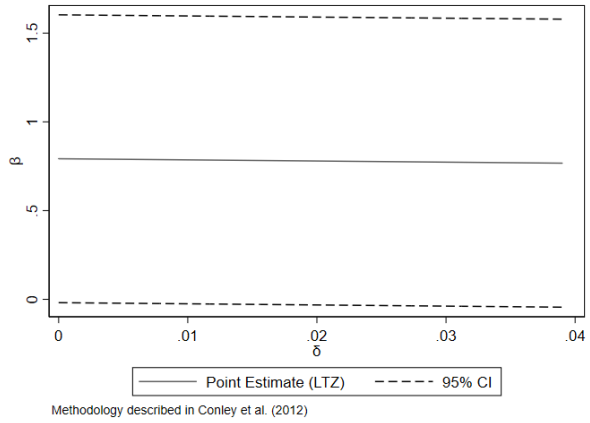


(e) Diagnosed with STIs five years

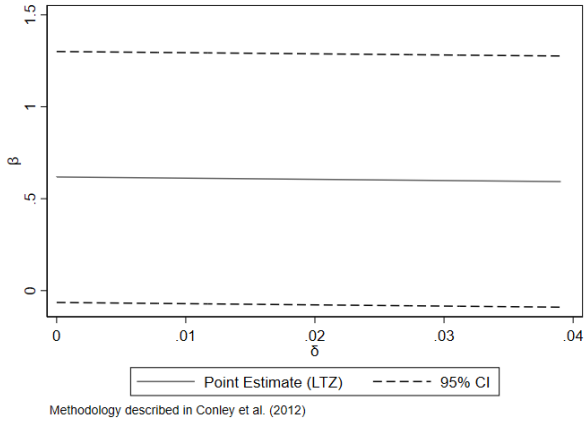


(f) Diagnosed with STIs one year

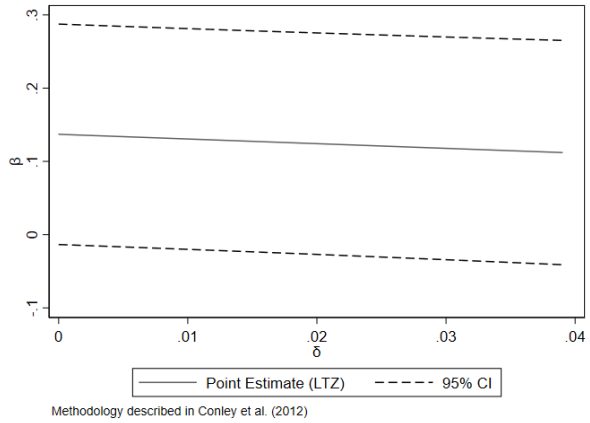
Figure 2.3: Effect of neighborhood-level ethnic heterogeneity (HHI) on total number of sexual partners, the number of heterosexual partners, the most recent sexual partners and STIs infection at different periods



(a) Total number of sexual partners in past 5 years



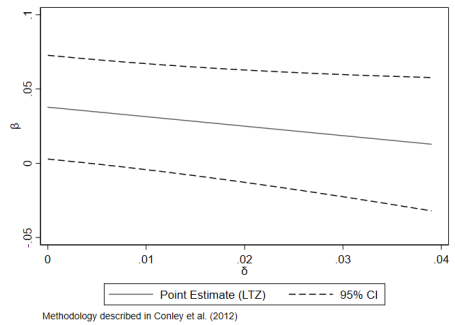
(b) Number of heterosexual partners in past 5 years



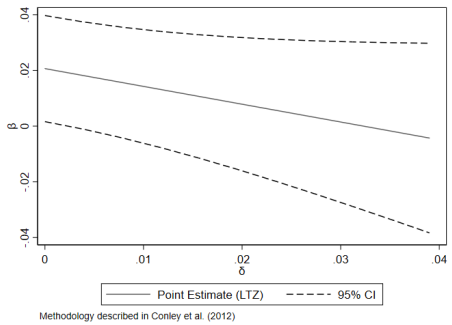
(c) Number of most recent sexual partners

Figure 2.4: Plausibly Exogenous Bounds: Sexual behavior

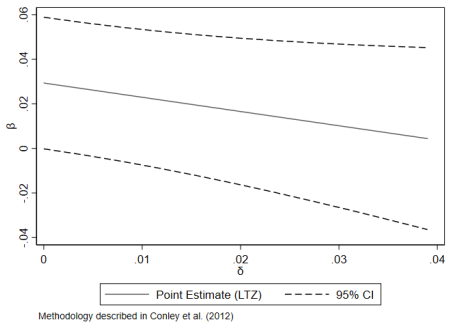
Notes: The figures present the upper and lower bound of the 95% confidence intervals of the second-stage coefficient on sexual behaviors. Point estimates are calculated according to Conley et al. (2012), using our baseline IV specification from column (2) in Table 2.7. The "local to zero (LTZ)" approach treats the uncertainty surrounding γ, the coefficient on the instrument as being normally distributed, with the mean and variance of a $N(0, \delta^2)$ variable.

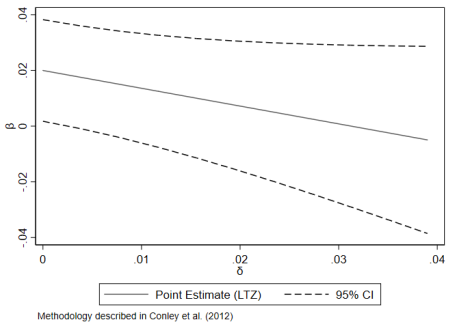


(a) Ever been diagnosed with chlamydia

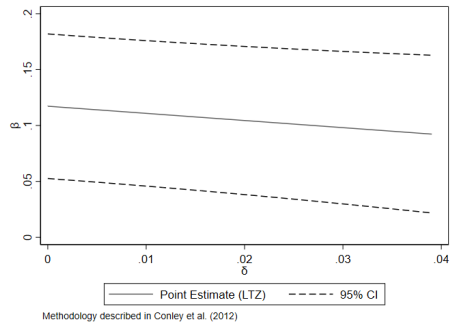


(b) Ever been diagnosed with gonorrhoea

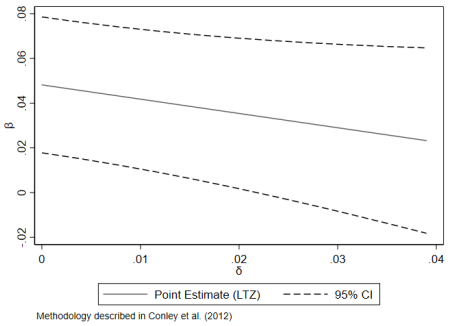




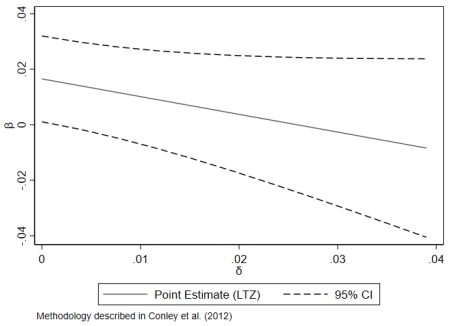
(c) Ever been diagnosed with warts (d) Ever been diagnosed with herpes



(e) Ever been diagnosed with STIs



(f) Diagnosed with STIs in past 5 years



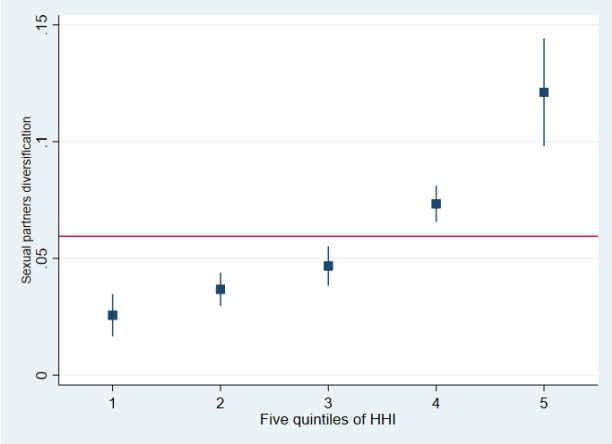
(g) Diagnosed with STIs last year

Figure 2.5: Plausibly Exogenous Bounds: Sexually transmitted infections

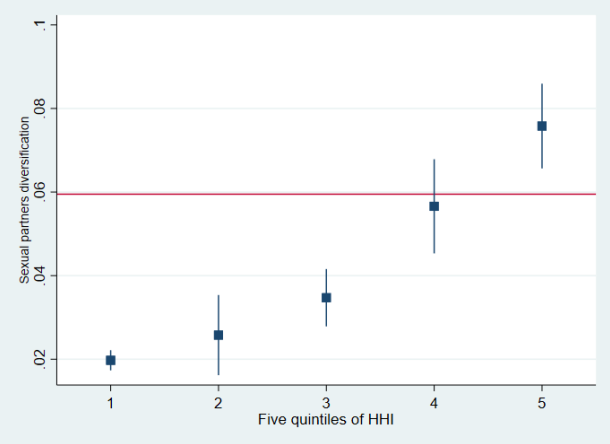
Notes: The figures present the upper and lower bound of the 95% confidence intervals of the second-stage coefficient on STIs. Point estimates are calculated according to Conley et al. (2012), using our baseline IV specification from column (2) in Table 2.7. The "local to zero (LTZ)" approach treats the uncertainty surrounding γ, the coefficient on the instrument as being normally distributed, with the mean and variance of a $N(0, \delta^2)$ variable.

Figure 2.6: Mechanism Framework

Notes: The mechanism framework is valid under the assumptions that the society is organized around different racial groups, and with ethnic integration being "partial". According to our theoretical framework, partial ethnic integration could lead to strategic diversification of sexual partners across ethnic groups so as to hide infidelity and decrease its cost. This optimizing behavior implies that individuals residing in multiracial neighborhoods are more likely to diversify their sexual partners than their counterparts living in the more racially homogeneous areas, which therefore increases the risk of getting infected with sexually transmitted diseases.

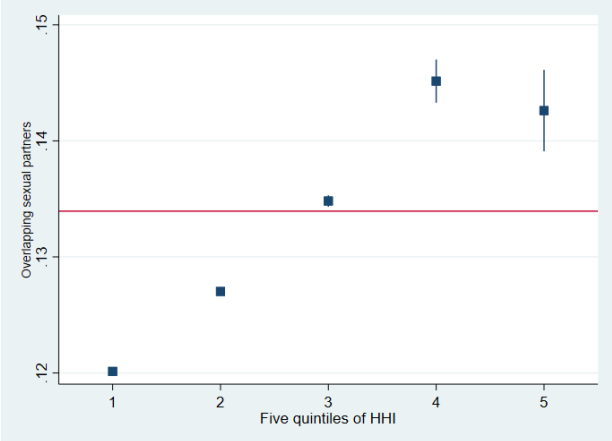


(a) Without fixed effects

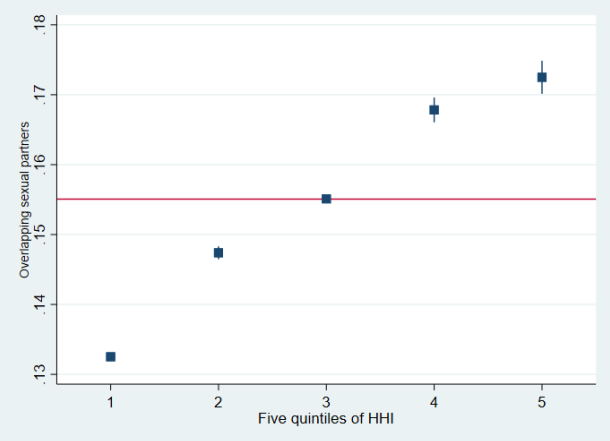


(b) With fixed effects

Figure 2.7: HHI and sexual partners diversification



(a) Without fixed effects



(b) With fixed effects

Figure 2.8: HHI and overlapping sexual partners

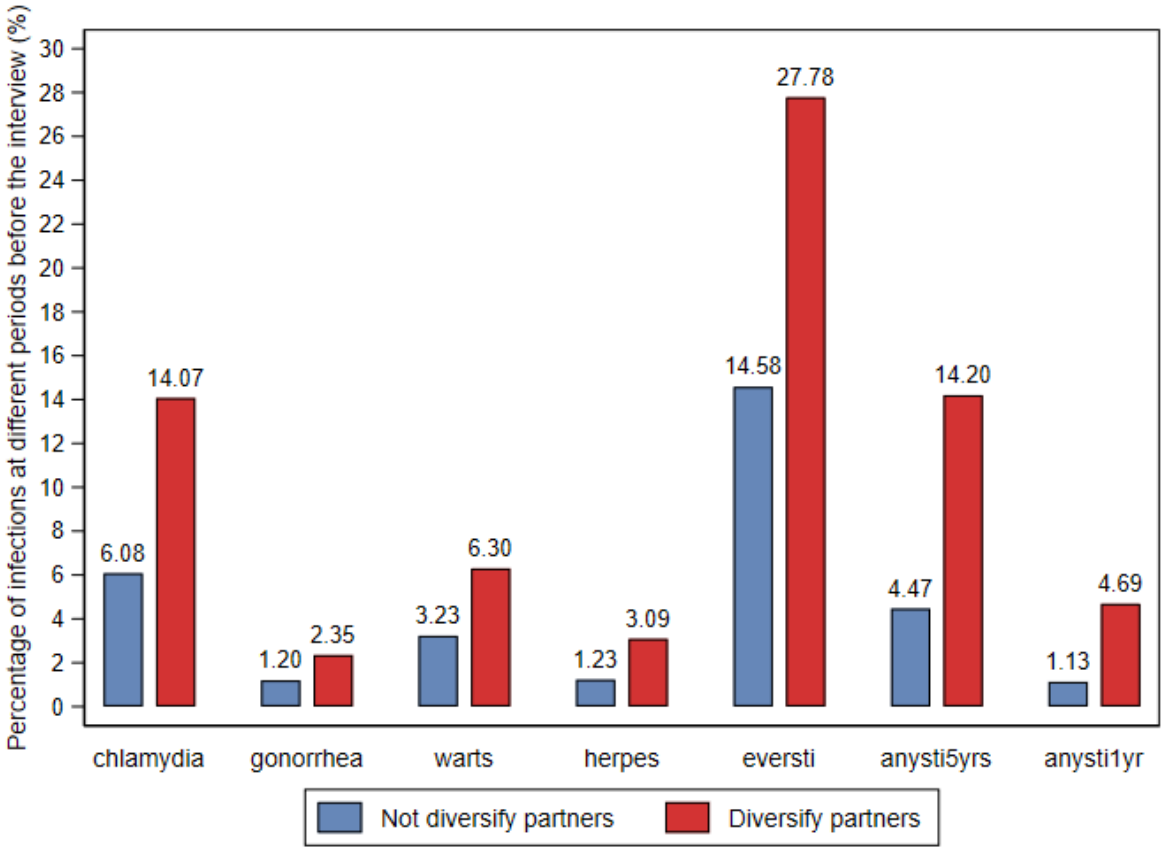


Figure 2.9: Prevalence of STIs among individuals diversify vs homogenize sexual partner

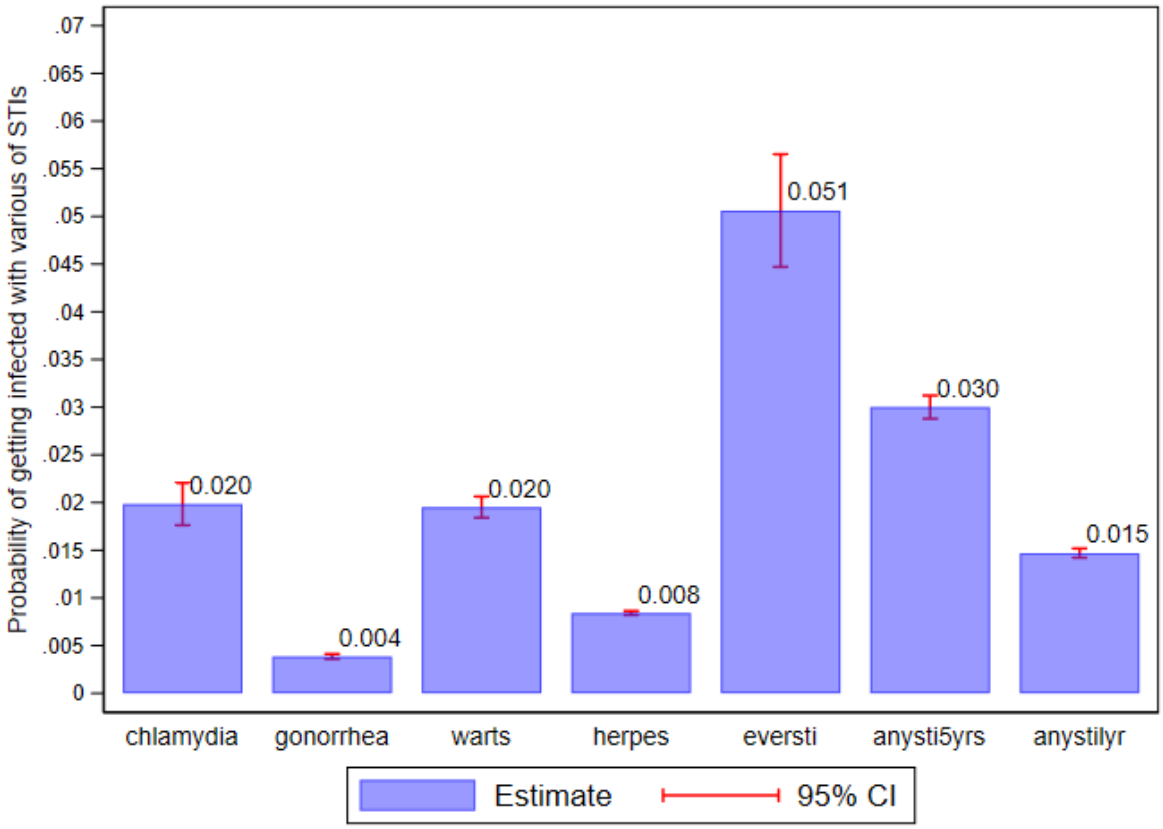


Figure 2.10: The estimate of sexual partner diversification on STIs infection

Table 2.1: Summary Statistics

Variables	Observations	Mean	Std. Dev.
HHI	15162	0.191	0.188
Controls			
Age	15162	38.390	16.802
Male	15162	0.415	0.493
Native	15132	0.410	0.492
Region			
North East	15,162	0.051	0.220
North West	15,162	0.129	0.336
Yorkshire and The Humber	15,162	0.085	0.278
East Midlands	15,162	0.083	0.276
West Midlands	15,162	0.088	0.284
South West	15,162	0.081	0.273
East	15,162	0.102	0.303
London	15,162	0.110	0.312
South East	15,162	0.133	0.339
Wales	15,162	0.054	0.226
Scotland	15,162	0.084	0.278
Marriage status			
Married	15,124	0.353	0.478
Cohabitation	15,124	0.118	0.323
Previously married	15,124	0.141	0.348
Single and never married	15,124	0.388	0.487
Education			
Degree	15,117	0.238	0.426
Higher education	15,117	0.274	0.446
GCSE	15,117	0.352	0.478
Foreign or others	15,117	0.012	0.108
None	15,117	0.124	0.330
Ethnic groups			
White	15,637	0.854	0.353
Mixed	15,637	0.028	0.165
Asian/Asian British	15,637	0.055	0.227
Black/Black British	15,637	0.040	0.197
Chinese/Others	15,637	0.023	0.150
Religion			
None	15,114	0.511	0.500
Christian - Church of England/Anglican	15,114	0.143	0.350
Christian - Roman Catholic	15,114	0.097	0.296
Christian - other	15,114	0.182	0.386
Non-Christian	15,114	0.066	0.249
Employment status			
Employed	15,116	0.542	0.498
Full-time education	15,116	0.129	0.336
Unemployed	15,116	0.205	0.404
Retired	15,116	0.124	0.329
Sex identity			
Heterosexual/straight	15,109	0.967	0.177
Gay/lesbian	15,109	0.014	0.118
Bisexual	15,109	0.015	0.121
Other	15,109	0.004	0.059

Table 2.2: Descriptive statistics of outcomes

Variable	Observations	Mean	Standard Deviation
Panel A	**Sexual behavior**		
tot5yrs	14,615	2.752	5.331
het5yrs	14,621	2.573	4.939
mostrecent	15,162	1.647	1.001
Panel B	**Sexually transmitted infection**		
chlamydia	14,255	0.063	0.243
gonorrhoea	14,251	0.013	0.112
warts	14,252	0.033	0.178
herpes	14,252	0.013	0.112
eversti	12,394	0.171	0.376
anysti5yrs	14,545	0.048	0.213
anysti1yr	14,544	0.013	0.112

Notes: Refer to Table B1 for the detailed definitions of the variables listed above.

Table 2.3: Neighborhood-level ethnic heterogeneity and sexual behavior: OLS estimates

	tot5yrs		het5yrs		mostrecent	
Variable	(1)	(2)	(3)	(4)	(5)	(6)
HHI	2.062***	0.644**	1.462***	0.530**	0.472***	0.147**
	(0.305)	(0.228)	(0.264)	(0.184)	(0.081)	(0.058)
	[0.082]	[0.026]	[0.078]	[0.025]	[0.100]	[0.031]
R-squared	0.008	0.127	0.007	0.127	0.014	0.186
Region & Ethnicity FE	✓		✓		✓	
Fully controlled		✓		✓		✓

Notes: Refer to Table B1 for the detailed definitions of the dependent variables. Beta estimates in the brackets demonstrate that increasing the ethnic heterogeneity by one standard deviation will result in how many standard deviations increase for sexual behaviors. Fully controlled regressions include a gender indicator, age, five education level fixed effects, seven religion fixed effects, four sexual identity fixed effects, four employment status fixed effects, four marital status fixed effects, 11 region fixed effect, 5 ethnicity fixed effect, and an indicator whether the respondent is a native. Robust standard errors are in parenthesis, *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 2.4: Neighborhood-level ethnic heterogeneity and sexually transmitted infections: OLS estimates

	chlamydia		gonorrhoea		warts		herpes		eversti		anysti5yrs		anysti1yr	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
HHI	0.055***	0.026*	0.018**	0.017*	0.037***	0.035***	0.010	0.008	0.114***	0.099***	0.053***	0.021*	0.015	0.006
	(0.014)	(0.013)	(0.008)	(0.009)	(0.011)	(0.009)	(0.007)	(0.006)	(0.026)	(0.021)	(0.012)	(0.011)	(0.010)	(0.009)
	[0.051]	[0.024]	[0.031]	[0.029]	[0.040]	[0.038]	[0.018]	[0.015]	[0.061]	[0.053]	[0.056]	[0.023]	[0.032]	[0.012]
R-squared	0.006	0.043	0.005	0.016	0.004	0.017	0.002	0.008	0.011	0.053	0.007	0.044	0.003	0.018
Region & Ethnicity FE	✓		✓		✓		✓		✓		✓		✓	
Fully controlled		✓		✓		✓		✓		✓		✓		✓

Notes: Dependent variables are dichotomous scored one if the respondent reported being infected within different periods, otherwise zero. Refer to Table B1 for the detailed definitions of the dependent variables. Beta estimates in the brackets indicate that increasing the ethnic heterogeneity by one standard deviation causes the likelihood of getting infected with STIs by how many standard deviations. Fully controlled regressions include a gender indicator, age, five education level fixed effects, seven religion fixed effects, four sexual identity fixed effects, four employment status fixed effects, four marital status fixed effects, 11 region fixed effect, 5 ethnicity fixed effect, and an indicator whether the respondent is a native. Robust standard errors are in parenthesis, *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 2.5: The importance of bias from unobservables: Altonji et al. (2005), Oster (2017), and Gonzalez and Miguel (2015)

	Coeff. Ratio after Altonji et al. (2005)	Mininum Coeff. Lower Bound after Oster (2017) ($R_{max} = \Pi R^*$)				Coefficient after Gonzalez and Miguel (2015) ($R_{max} = R^* + (R^* - R)$)
Outcomes	AET	R_{max}=1	$R_{max} = 1.2 * R^*$	$R_{max} = 1.4 * R^*$	$R_{max} = 1.6 * R^*$	β^*
	(1)	(2)	(3)	(4)	(5)	(6)
tot5yrs	1.451	-5.108	0.476	0.309	0.141	0.200
het5yrs	3.437	-2.309	0.448	0.366	0.283	0.376
mostrecent	1.282	-0.750	0.106	0.065	0.024	0.032
Panel B Sexually transmitted infection						
chlamydia	10.558	-0.114	0.025	0.023	0.022	0.023
gonorrhoea	4.543	-0.354	0.016	0.015	0.014	0.013
warts	-9.012	0.363	0.036	0.037	0.038	0.039
herpes	8.902	-0.162	0.008	0.008	0.007	0.007
eversti	-29.583	0.178	0.100	0.101	0.101	0.102
anysti5yrs	3.039	-0.484	0.017	0.012	0.007	0.014
anysti1yr	2.414	-0.320	0.005	0.003	0.002	0.003

Notes: This table shows the selection on unobservables. The independent variable is the ethnic heterogeneity and the dependent variables are outcomes of both sexual behaviors and the sexually transmitted infections which are listed on the left side of the table. The cells in conlumn (1) report the coefficient for the ethnic heterogeneity in two regressions, a restricted set of controls and a full set of controls. The regression with a restricted set of controls only includes the region and ethnicity fixed effects, and the gender and the age of each individual, and the latter one with a full set of controls. The ratio is calculated as $\frac{\hat{\beta_F}}{\hat{\beta_R} - \hat{\beta_F}}$, where $\hat{\beta_F}$ is the coefficient estimates in a full controls regression, and $\hat{\beta_R}$ is the restricted one. Columns (2)-(5) present coefficient lower bound by employing Oster (2017), with the bias-adjusted estimator equation as $\hat{\beta_F} - (\hat{\beta_R} - \hat{\beta_F}) * \frac{R_{max} - R^*}{R^* - R}$. R^* and R are the R^2 for the regression with the full set of controls and the restricted controls, respectively. The cells in conlumn (2) report the coefficient lower bound with the most conservative value equals to one, and columns (3)-(5) with $R_{max} = min\{\Pi R^*, 1\}$ suggested by Oster (2017), and the values of the three multipliers are 1.2, 1.4 and 1.6. Column (6) reports the coefficient using $R_{max} = R^* + (R^* - R)$ which suggested by Gonzalez and Miguel (2015).

Table 2.6: First-stage regressions: The effect of native-induced ethnic heterogeneity on the neighborhood-level ethnic heterogeneity

Dependent Variable: HHI

Variable	(1)	(2)
hhinative	0.779***	0.774***
	(0.051)	(0.049)
Region & Ethnicity FE	✓	
Fully controlled		✓
First-stage F statistic	234	254.41
Anderson-Rubin F-test	0.007	0.004
Number of clusters	11	11
Observations	12,376	12,267

Notes: Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 2.7: Effects of neighborhood-level ethnic heterogeneity on sexual behavior and STI infection: Standard IV, Imperfect IV and Lewbel IV (Independent Variable: HHI of 2011)

Dependent Variables	OLS	Standard IV	Imperfect IV (Estimator)		Lewbel IV
			Lower Bound	Upper Bound	
	(1)	(2)	(3)	(4)	(5)
Panel A Sexual behavior					
tot5yrs	0.644**	0.791**	0.712	0.791	0.710**
	(0.228)	(0.375)			(0.341)
het5yrs	0.530**	0.596**	0.561	0.596	0.535**
	(0.184)	(0.260)			(0.235)
mostrecent	0.147**	0.136**	0.136	0.147	0.138***
	(0.058)	(0.055)			(0.053)
Panel B Sexually transmitted infection					
chlamydia	0.026*	0.036**	0.030	0.036	0.028*
	(0.013)	(0.016)			(0.016)
gonorrhoea	0.017*	0.021**	0.017	0.021	0.020**
	(0.009)	(0.009)			(0.009)
warts	0.035***	0.029***	0.029	0.035	0.025***
	(0.009)	(0.010)			(0.008)
herpes	0.008	0.021	0.014	0.021	0.016
	(0.006)	(0.013)			(0.011)
eversti	0.099***	0.119***	0.108	0.119	0.108***
	(0.021)	(0.034)			(0.030)
anysti5yrs	0.021*	0.048**	0.034	0.048	0.042**
	(0.011)	(0.020)			(0.019)
anysti1yr	0.006	0.016	0.006	0.016	0.014
	(0.009)	(0.011)			(0.010)
Region & Ethnicity FE	Yes	Yes	Yes		Yes
All Controls	Yes	Yes	Yes		Yes

Notes: Dependent variables in Panel B are dichotomous scored one if the respondent reported being infected within different periods, otherwise zero. Please refer to Table B1 for the detailed definitions of the dependent variables. The instrumental variable for HHI of 2011 in regressions is native-induced neighborhood level HHI of the same year. All regressions include a constant, and a full set of controls. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 2.8: Interpretation the Imperfect IV estimators from the propositions of Nevo and Rosen (2012)

	$\sigma_{XZ} < 0$	
	Assumption4	No Assumption4
One Instrument	$\lvert\rho_{XU}\rvert \geqslant \lvert\rho_{ZU}\rvert$	$\lvert\rho_{XU}\rvert < \lvert\rho_{ZU}\rvert$
$\rho_{XU} > 0$	$\beta^{iv}_{native} \leqslant \beta \leqslant \beta^{iv}_{v(1)}$	$\beta^{iv}_{native} \leqslant \beta \leqslant \beta^{ols}$
$\rho_{XU} < 0$	$\beta^{iv}_{v(1)} \leqslant \beta \leqslant \beta^{iv}_{native}$	$\beta^{ols} \leqslant \beta \leqslant \beta^{iv}_{native}$

Notes: The table was adapted from Clarke and Matta (2018). In our paper, the instrument Z is denoted as $1 - HHI_{native}$ for two-side bounds or $\sigma_{XZ} < 0$ purpose. $\beta^{iv}_{v(1)}$ is the 2SLS estimator of a transformed valid instrument, denoted $V(\lambda^*) = (\sigma_X Z - \lambda^* \sigma_Z X)$ with $V(1) = (\sigma_X Z - 1\sigma_Z X)$.

Table 2.9: UCI and LTZ Bounds

	UCI: $\gamma \in [0, 2\hat{\gamma}]$		LTZ: $\gamma \sim N(0, \delta^2)$	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
	(1)	(2)	(3)	(4)
Panel A: Sexual Behavior				
tot5yrs	-0.2563	1.8316	-0.0191	1.6025
het5yrs	-0.2470	1.4740	-0.0650	1.2999
mostrecent	-0.0737	0.3385	-0.0140	0.2869
Panel B: Sexually transmitted infections				
chlamydia	-0.0100	0.0764	0.0023	0.0722
gonorrhoea	-0.0203	0.0524	0.0011	0.0392
warts	-0.0107	0.0600	-0.0007	0.0584
herpes	-0.0093	0.0403	0.0012	0.0377
eversti	0.0233	0.2021	0.0521	0.1814
anysti5yrs	0.0040	0.0832	0.0173	0.0781
anysti1yr	-0.0098	0.0338	0.0005	0.0315

Notes: This table presents upper and lower bounds of a 95% confidence interval for the effect of neighborhood-level ethnic heterogeneity on individuals' sexual behavior and the prevalence of STIs. The methodology is developed by Conley et al. (2012). Columns (1) and (2) show the lower bound and upper bound from the UCI approach assumes the true $\gamma \in [0, 2\hat{\gamma}]$, and columns (3) and (4) from the LTZ approach assuming γ follows a normal distribution with the mean and variance of $\gamma \sim N(0, \delta^2)$.

Table 2.10: Effects of sexual partner diversification on STI infection

Variable	(1) chlamydia	(2) gonorrhea	(3) warts	(4) herpes	(5) eversti	(6) anysti5yrs	(7) anysti1yr
diverse	0.020***	0.004***	0.020***	0.008***	0.051***	0.030***	0.015***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.003)	(0.001)	(0.000)
	[0.011]	[0.007]	[0.008]	[0.006]	[0.013]	[0.007]	[0.008]
Observations	13,375	13,371	13,372	13,372	11,552	13,373	13,372
R-squared	0.068	0.017	0.017	0.010	0.083	0.078	0.033
Region FE	Y	Y	Y	Y	Y	Y	Y
Ethnicity FE	Y	Y	Y	Y	Y	Y	Y
Individual Controls	Y	Y	Y	Y	Y	Y	Y

Notes: Independent variable is a binary variable scored one if the respondent diverse his/her sexual partner composition, otherwise zero. Dependent variables are dichotomous scored one if the respondent reported being infected within different periods, otherwise zero. Refer to Table B1 for the detailed definitions of the dependent variables. The table reports OLS estimates. The unit of observation is an individual. Below each coefficient two standard errors are reported. The first, reported in parentheses, is standard errors adjusted for two-way standard errors clustering within ethnic groups and within regions. The second, reported in square brackets, is standard errors adjusted for clustering within regions. Fully controlled regressions include number of most recent sexual partners, a gender indicator, age, five education level fixed effects, seven religion fixed effects, four sexual identity fixed effects, four employment status fixed effects, four marital status fixed effects, 11 region fixed effect, 5 ethnicity fixed effect, the majority group size of the neighborhood where the respondent lives, and an indicator whether the respondent is a native. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Table 2.11: Falsification test: Effects of neighborhood-level of ethnic heterogeneity on various common non-transmitted diseases (OLS v.s. 2SLS)

Dependent Variable: Falsification Outcomes

Dependent	(1)	Independent	(2)	(3)
Variables	Mean	Variable	OLS	IV
heart attack	0.014	HHI(2011)	0.019	0.009
	(0.116)		(0.012)	(0.012)
heart disease	0.025	HHI(2011)	0.009	0.009
	(0.156)		(0.013)	(0.019)
hypertension	0.101	HHI(2011)	0.033	0.045
	(0.301)		(0.026)	(0.033)
stroke	0.012	HHI(2011)	0.004	0.003
	(0.108)		(0.007)	(0.008)
chronic lung disease	0.011	HHI(2011)	0.005	0.011
	(0.103)		(0.009)	(0.010)
Parkinson's disease	0.001	HHI(2011)	-0.001	-0.001
	(0.027)		(0.001)	(0.002)
epilepsy	0.013	HHI(2011)	0.008	0.002
	(0.114)		(0.007)	(0.009)
All controls			Yes	Yes
Region FE			Yes	Yes
Ethnicity FE			Yes	Yes

Notes: Dependent variables are dichotomous scored one if the respondent reported having the marked disease, otherwise zero. Column (1) states the mean values of these non-transmitted diseases, with the standard deviation in the parentheses. The instrumental variable for HHI of 2011 in regressions is native-induced neighborhood level HHI of the same year. All regressions include a constant, and a full set of controls. Weighted robust standard errors are in parentheses. Regression disturbance terms are clustered at the region level. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Chapter 3

Romantic Homophily and Sexually Transmitted Infections

3.1 Introduction

The structure of romantic and sexual networks reflects a variety of sexual behaviors and is likely to affect the spread of sexually transmitted diseases (STDs). In social and economic contexts, it is frequently the case that links occur between similar individuals with relevant characteristics, such as racial/ ethnicity, age, religious affiliation, and educational attainment, etc., and these attributes are often related to systematic connectivity patterns. This ecological phenomenon is known as homophily, which encourages similar connections of multiple varieties ranging from friendship, marriage, to sexual relationship. Taking high-school romantic networks, for instance, Currarini et al. (2009) find that larger racial groups have a larger fraction of intra-race friendships. Likewise, Bearman et al. (2004) demonstrates adolescents at Jefferson High School tend to select partners that are similar in terms of socioeconomic status, suspension from school, sexual experience, and smoking. What is little known is the effect of homophily patterns in sexual network formation on the spread of STDs. We address this question in this paper using data from the United Kingdom.

Our analysis relies on a classical hypothesis in social network theory that central indi-

viduals in networks are more relevant and influential in learning, contagion, and diffusion processes than their peripheral counterparts. Individuals who are central in sexual networks may therefore have an increased probability of getting infected with sexually transmitted diseases. In societies organized around distinct ethnic groups, agents who sexually connect several ethnic groups may be more central in sexual networks. Therefore, homophilous behavior may be more likely to characterize peripheral individuals.[1] It follows that building sexual links only with members of own ethnic group is likely to decrease the risk of STDs. We test this hypothesis in this paper.

We test the causal effect of romantic homophily on STIs using rich individual-level data from England, Scotland, and Wales. In particular, we use the third British National Survey of Sexual Attitudes and Lifestyles conducted in 2010-2012 (Natsal3). The Natsal3 data contain socioeconomic and demographic information on males and females aged 16-74 years. Most importantly, interviewees not only provided detailed information on their most recent sexual partners but also on whether they were diagnosed with certain STIs at some point in their lifetime, within the past five years before the interview, or within the year preceding the interview. We were, therefore, able to analyze the effects of homophilous sexual behavior on several sexually transmitted diseases: (1) chlamydia; (2) gonorrhoea; (3) warts; (4) herpes; and (5) any STI within the past five years before the interview.

We find that romantic homophily decreases the probability of acquiring a wide range of STDs. Figure 3.1 reports the average infection rate in the past five years for various diseases, including chlamydia, gonorrhea, genital warts, herpes, and any STI diagnosed in the past five years before the interview of the United Kingdom. For all infections, it is obvious that individuals who ethnically homogenize sexual partners are less likely to get infected with STIs compared to their counterparts who racially diversify sexual partners.

[1]The method that we employ in this paper of measuring ego centrality in a network is similar but not the same as the ways described in Jackson et al. (2017). These methods are known as connectivity, closeness, intermediation, and having well-connected neighbors. In our paper, individuals who totally diversify sexual partners across racial groups are in the central position of the sexual network; otherwise, non-centrality. See more studies work on measures of network centrality for various applications by Bonacich (1987), Valente (2012), Jackson et al. (2017), Banerjee et al. (2019), and Bloch et al. (2019).

We find that these results are robust to controlling for a large number of covariates, including region fixed effects, ethnicity fixed effects, the number of most recent sexual partners, demographic characteristics (gender and age), and individual covariates (namely, educational attainment, religion, sexual identity, employment status, marital status, whether the intercourse is unprotected, and whether the respondent is a native) using OLS regressions. For many of these diseases, the estimates of interest vary moderately but are still statistically significant after we condition on these covariates. Specifically, the OLS regressions predict that, romantic homophily significantly decreases the likelihood of acquiring infections such as chlamydia, genital warts, and any STIs that have been diagnosed in the past five years. Take chlamydia, for instance. Moving from an intra-ethnic relationship to an inter-ethnic relationship increases the probability of acquiring chlamydia by 1.49 percentage points with a restricted set of controls and by 1.65 percentage points with a full set of controls.

OLS estimates are likely to be confounded by unobserved factors. We address this potential endogeneity problem using three recently developed methods. The first approach consists of assessing bias attributable to unobserved characteristics (Altonji et al. 2005, Oster 2019, Gonzalez and Miguel 2015). The second method is an instrumental variable technique developed by Lewbel (2012). It constructs internal instruments when an external instrument is lacking, and it does not rely on the classical exclusion restriction assumption for identification. Third, we use a propensity score matching (PSM) technique as a complementary and alternative method for causal inference with individual survey data. All of these methods corroborate our main conclusion that romantic homophily causally decreases the probability of acquiring an STD.

We also conduct some falsification tests that consist of a reverse-placebo analysis. These tests consist of analyzing the effect of the racial homogeneity of sexual partners on health conditions that are not driven by sexual interactions. These conditions include arthritis, heart disease, stroke, diabetes, and chronic lung disease. Unlike sexually transmitted infections, these diseases are non-contagious. An economically significant effect of homophilous partner composition on any of these falsification outcomes would question our interpreta-

tion of the positive impact of romantic homophily on STIs. We find no such effect, which lends more credibility to our mechanisms.

We also analyze the heterogeneous effect of romantic homophily on STDs by gender, minority status, and sexual orientation. We find that this effect is substantially greater for women, members of the majority group, and heterosexual individuals. While previous studies show that homophily can put minority groups at a disadvantage (Ananat and Washington 2009, Centola 2011, Halberstam and Knight 2016, Karimi et al. 2018), our findings seem to suggest that minority individuals are less affected by romantic homophily.

The paper proceeds as follows: Section 2 situates our study within the extant literature. Section 3 describes the data. Section 4 presents the main finding about the causal impact of romantic homophily on STIs. Section 5 presents sensitivity analyses and robustness checks. Section 6 concludes.

3.2 Related Literature

A body of theoretical and empirical research examines the impact of homophilous effects in social networks. This literature shows that similarity along several dimensions breeds connections in a wide variety of areas including marriage, and friendship (Kalmijn 1998, Fong and Isajiw 2000, Baerveldt et al. 2004, Bearman et al. 2004, Currarini et al. 2009, 2010, Skopek et al. 2011, Boucher 2015); information diffusion (Schneider et al. 1997, Golub and Jackson 2010, 2012, Halberstam and Knight 2016, Acemoglu et al. 2017); organizations (Ruef et al. 2003, Ozcan and Eisenhardt 2009), and employment (Bertrand and Mullainathan 2004, Calvo-Armengol and Jackson 2004, Patacchini and Zenou 2012, Zenou 2013, 2015), among others. As the paper by Currarini et al. (2009) concludes that the majority groups tend to form more same-type friendships and fewer other-type connections than small groups. Likewise, Bearman et al. (2004) demonstrates that adolescents at Jefferson High School tend to select partners that are similar in terms of socioeconomic status, suspension from school, sexual experience, and smoking. These papers indicate that homophily takes place in relationships that are socially accepted.

Our paper is also related to the literature on the racial homogeneity of networks, from the friendships of adolescence (Joyner and Kao 2000, Kao and Joyner 2004, 2006, Wang et al. 2006, Wimmer and Lewis 2010) to adult marriage (Shimer and Smith 2000, Eeckhout and Kircher 2010, 2011, Hagedorn et al. 2017), which has been well documented by sociologists and economists. Studies on the sexual network structures also show strong racial homophily (Kenyon and Colebunders 2013, Logan 2013, Birkett et al. 2015). In particular, for young men who have sex with men, Birkett et al. (2015) show that most of the respondents reported sexual relationships with those of the same race, and a significant preference for older sexual partners. Their results indicate that only egos whose sexual partners were African American and having sexual partners who were all male that significantly correlates with HIV infection. Logan (2013) shows a similar homophilous pattern for lesbians.

Our paper contributes to these literatures, but we differ in our research question. To our knowledge, our paper is the first to document the effect of romantic homophily on the probability of acquiring an STI. Our findings suggest that individuals who display a homophilous behavior in sexual partner selection occupy more peripherical positions in sexual networks, as these individuals are less likely to be infected with STDs. The results have implications for the design of STD prevention policies, especially in a context in which it is almost impossible to design policies that target the configuration of sexual networks these networks are hard to observe in reality.

Our study is also related to a large body of literature on the role of centrality in a network on the diffusion of information, products, and contagion (Ballester et al. 2006, Calvó-Armengol et al. 2009, Acemoglu et al. 2012, 2017, Banerjee, Chandrasekhar, Duflo and Jackson 2013, Leduc et al. 2017, Banerjee et al. 2019). The fundamental question for these studies is to find the most effective node for diffusion. The optimal nodes are those that are central, according to appropriate measures of centrality. Our empirical results validate the notion that agents who occupy the central positions within and across area play more critical roles in STD spread. The spread of infectious diseases is analogous to information diffusion within social networks but not identical. The information diffusion

depends not only on the characteristics of the information, the depth of the relationships between two nodes, and the willingness of the information holder to share it with the specific connection. In contrary, the spread of a disease within networks is more rapid and somewhat random.

3.3 Data

This section describes the data we use to analyze the causal effect of sexual homophily on the likelihood of acquiring a sexually transmitted infection. We use the third British National Survey of Sexual Attitudes and Lifestyles 2010-2012 (Natsal3). This survey includes individuals between the age of 16 and 74 years, and it contains detailed information on respondents' socioeconomic and demographic characteristics and sexual behavior.[2] Data were weighted by age, gender, and geographic region so as to reflect the total population of Britain and to reduce potential bias. Questions relating to patterns of partnership formation (such as serial monogamy and concurrency), sexual mixing (including ethnicity, sexual identity, location, and age of each of the recent sexual partners), and on whether one was diagnosed with several common sexually transmitted infections at some point were asked using computer-assisted self-completion format. The methodology and main results of the survey are detailed in Johnson and Mercer (2017).The survey contains 6293 males and 8869 females.

[2]The Natsal3 survey we use is a multistage, clustered, and nationally stratified probability sample survey conducted between September 2010 and August 2012. It collected data on over 15000 adults aged 16-74 using face-to-face interviews. The survey selected postcode sectors as the primary sampling units (PSUs); addresses within each postcode were chosen at the second stage; then one eligible adult was randomly selected at the final stage household (Erens et al. 2014). Two previous Natsal surveys were conducted in 1990 and 2000, respectively, but Natsal3 is the first survey that considers individuals up to the age of 74, because of the recognition that many people remain sexually active into their later life, and that sexual health issues affect both the younger and older age groups.

3.3.1 Sexually Transmitted Infections

In the third British National Survey of Sexual Attitudes and Lifestyles 2010-2012, interviewees were asked to provide information on whether they have been infected with any of the following STIs in their lifetime, within the past 5 years, and within the year preceding the survey: chlamydia, gonorrhoea, warts, syphilis, trichomonas, herpes, and/or any STI.

Given that the very low prevalence of syphilis and trichomonas, we do not analyze these diseases. Therefore, the STIs we analyze are chlamydia, gonorrhoea, genital warts, herpes, and the variables capturing contact with any STI in the past five years prior to the interview. For each infection type, we generate a binary variable indicating whether an individual is infected or not.

Figure 3.2 shows some descriptive statistics on these diseases (3.2a), and in each of the sub-samples by gender (3.2b), ethnic status (3.2c), and sexual identity (3.2d). Chlamydia is relatively more common among each of the four samples of interest. It is followed by genital warts, herpes, and then gonorrhoea. Figures (3.2b-3.2d) demonstrate that the proportions of infections that varied between different groups. In general, all the sexually transmitted infections listed above tend to be more prevalent among women compared to men, more prevalent among the majority than the minority, and widely different between heterosexual and other sexual identities.

3.3.2 Sexual Homophily Index

For our study, the most crucial information provided by the interviewees is that concerning their ethnic group, the number of their most recent sexual partners, the gender, and the ethnicity of their partners belong. In Natsal3, all of the interviewees and their partners are classified into six categories according to their own perceived ethnic groups and cultural background, however, only five ethnicity categories we consider are: White, Mixed, Asian (excluding Chinese)/Asian British, Black/Black British, and Chinese and Others.

It is helpful to first provide a general picture of the patterns of mixed-ethnic sexual links in the UK. This analysis is based on the interviewees' most recent partners' ethnic

identity. Figure 3.3 shows the distribution of individuals by the ethnicity of their most recent partners. Except for the Mixed group, the majority of interviewees find a partner from their own ethnic group. Take the most recent partner, for instance, over 95% percent of White/White British have a White opposite-sex partner; 65% percent of mixed individuals have a White partner, but only 16% of mixed individuals choose an opposite-sex partner from their group. Both Asians and Blacks have very high rates of homogamous partnerships with their groups (77% and 68%, respectively). Chinese have a comparable partnering rate between Whites (37%) and Chinese (34%). These patterns are consistent with what Currarini et al. (2009) found in the United States as the inbreeding homophily for most groups, with some exceptions for the smallest groups.

We compute an index that measures the fraction of intra-ethnic links of an individual of a specific ethnicity, and we use this index to indicate whether the individual diversifies his/her sexual partners or not. This measure is the homophily index proposed by Currarini et al. (2009) and defined as follows,[3]

$$HI_i = \frac{s_i}{s_i + d_i} \tag{1}$$

where s_i denotes the average number of intra-ethnic links that an individual of ethnic group i has, and d_i indicates the number of links that this individual has outside of his/her ethnic group. The index ranges from zero (full diversity) to one (complete uniformity).

Figure 3.4 reports the full distribution of the sexual homophily index (HI) in each of the four subsamples of interest (Figures (3.4a)-(3.4d)). Around 90% of the interviewees have intraethnic links, and only 6% of the interviewees select their sexual partners outside of theirethnic groups (see Figure (3.4a)), with the distribution variying between different sample groups. In general, in terms of finding sexual partners within own ethnic groups,

[3]We also calculate the inbreeding homophily index (IH_i) which is defined as $IH_i = \frac{H_i - \omega_i}{1 - \omega_i}$ to measure the amount of bias concerning our baseline homophily (HI_i). The bias relates to the denominator term $1 - \omega_i$, where ω_i is the relative fraction of type i in the population, defines as $\omega_i = \frac{N_i}{N}$, $N = \sum_K N_K$, K denotes different types of agents in the population. As to the details on the inbreeding homophily index, see Currarini et al. (2009). Here, IH_i varies little from HI_i; therefore, we use the baseline homophily index HI_i as the variable we are interested in the following descriptive analysis and regression estimations.

women do so more than men (See Figure(3.4b)), majority individuals are more homophilous than the minorities (See Figure (3.4c)), and heterosexuals are more homophilous than their counterparts of other sexual identities (See Figure (3.4d)). According to Figures (3.4c) and (3.4d), we see the different patterns of interaction between the majority groups (White and heterosexual) and minority (non-White and non-straight). In particular, for the minority groups, links from their own group are smaller compared to those from other groups due to the size effect. Due to the group size effect therefore, the ego in the minority group has the highest proportion of the reachable alters from the majority group (Bramoullé et al. 2012).

Apart from our explanatory variable of interest, which is the sexual homophily index, Table 3.1 presents the summary statistics of several potential confounding factors that we control for in regression-based analyses. An average individual had more than 1.5 sexual partners recently, and over 50% of the interviewees had one sexual partner, 12% had two partners, and around 22% individuals had three sexual partners. About 40% of the interviewees are male, and the average age of the individual is 38 at the time of the survey. Whites constitute 87% of the sample. About 35% of interviewees are married, compared with around 39% of whom are single or have never been married. In addition, variables included in the regressions, such as the level of education, and employment status may be correlated with the ability to choose partners from other ethnic groups. Religion could capture moral norms and principles. We also control for sexual orientation.

3.4 The Determinants of Romantic Homophily

This section analyses the determinants of sexual homophily We use a combination of descriptive analysis and modeling methods. Initially, we examine the extent to which individuals display homophilous behavior and how this behavior depends on the number of sexual partners, gender, and majority status. We then extend this to estimate our results in a multivariate model in which we control for other variables. For the ease of interpretation, we construct a partner uniformity index from the homophily index based on Equation 1.

This uniformity index is a dummy variable that takes the value of one if the homophily index equals to one, zero otherwise.

Descriptive analysis. Table 3.2 shows how the ethnic homogeneity of sexual partners depends on the number of sexual partners, gender, and majority status. It can be seen from column (1) that the ethnic homogeneity of sexual partners is a non-increasing function of the number of partners. Specifically, the sexual partners' uniformity index of an average person is 0.935 if he/she has only one sexual partner, 0.853 if he/she has two partners, and 0.713 if he/she has three sexual partners. It is noticeable, however, that having four partners is solely associated with having both heterosexual and same-sexual partnerships, which shows a little bit higher homophily rate compared to individuals who have three partners.

Noticeably, the descriptive results only predict a non-increasing trend in the ethnic homogeneity of partners, which is represented by our uniformity index. However, it is easy to note from Table 3.2 that the uniformity index is less than 0.5 for an average minority individual who has two partners. Therefore, if the number of partners that an individual has increased from one to two, the marginal decrease in ethnic homogeneity of partners should be more significant for an individual who belongs to an ethnic minority than his/her counterpart who belongs to a majority group, which is precisely what columns (4) and (5) show. Also, when the number of partners increases from two to three, the ethnicity homogenous of partners decreases from 0.897 to 0.783 for majorities, and from 0.449 to 0.220 for minorities. These results continue to hold if we compare the outcomes of majority versus minority men (column (6) and (8)) and the majority versus minority women (column (7) and (9)).

Regarding ethnically homogenizing partners, the last row of Table 3.2 shows that women do so more than men (the index of sexual homophily is 0.764 and 0.749 for females and males, respectively), but the difference is insignificant; and minorities are less homophilous than majorities (sexual homogeneity is 0.159 versus 0.819). Even more strikingly, although the variation between majority men and women is not significant (sexual homogeneity of 0.816 versus 0.822), the results show that both are more homophilous than minority

men (0.255) and minority women (0.455). These patterns continue to hold after keeping constant the number of sexual partners, which shows that the difference in the ethnic homogeneity of partners do not result from the variations in the number of partners.

Multivariate analysis. Table 3.3 represents the same analysis as Table 3.2 does in a multivariate regression framework. The regression model estimated is as follows:

$$homogenized_{ien} = \alpha_0 + \alpha_1 gender_i + \alpha_2 gpsize_{ie} + \alpha_3 nbpartners_i + X_{ien}^{'}\gamma + \delta_r + \epsilon_{ien} \quad (2)$$

Where $homogenized_{ien}$ represents the uniformity index, or the ethnic homogeneity of the sexual partners of individual i who belongs to the ethnic group e and resides in a neighborhood n. This binary variable takes the value of one if individual fully choose his/her sexual partners from own ethnic group, and zero otherwise; $gender_i$ is a binary indicator for men with the reference group being women; $gpsize_{ie}$ indicates $i's$ ethnic group size or its majority status; $nbpartners_i$ is a vector of categorical variables for individual i' number of partners; X_{ien} is a vector that includes a set of individual-level controls; δ_r is a region fixed effect, and ϵ_{ien} is a disturbance term assumed to be uncorrelated with the predictors. The coefficients on the number of partners, gender(male) and own group size are the primary predictors of the ethnic homogeneity of partners. We use ordinary least squares (OLS) regression to estimate Equation (2), with and without a set of demographic and socioeconomic variables (including age, education, religion, migration status, occupation, employment status, sexual identity, marriage status, and unprotected sex status).

The results in Table 3.3 confirm the descriptive results of Table 3.2. The results reveal that men are significantly less homophilous than women, and the difference is statistically significant (columns (1)-(4)). Majorities significantly are homophilous than their minority counterparts (columns (1)-(2)), and an individual who belongs to a larger group is more homophilous than an individual who belongs to a smaller group (columns (3)-(4)). The results also confirm the negative relationship between the number of partners and their ethnic homogeneity. The last four columns show that majority men homogenize partners less than majority women (columns (5)-(6)); similarly, minority men homogenize less than minority women (columns (7)-(8)). However, only the difference between the males and

females of the minority groups is significant after controlling for other variables as shown in Table 3.2.

3.5 Effects of Romantic Homophily on STIs

We use three main identification strategies to analyze the causal effect of romantic homophily on STIs. First, we estimate an OLS regression as our baseline strategy. Second, we resort to a recent IV method that relaxes the exclusion restriction assumption underlying the classical instrumental variables approach. Third, we employ propensity score matching. The idea is to reproduce a randomized experiment in a nonexperimental setting with a treatment group and a control group. These analyses are complemented by a reverse-placebo test. In particular, we analyze the effect of the uniformity index on health conditions that are not driven by sexual interactions (Section 3.5.4). Our assumption is that these conditions should not be affected by romantic homophily.

3.5.1 Identification Strategies

Baseline Strategy: OLS with Individual-level Controls

Our baseline strategy estimates the following OLS regression:

$$y_{ien} = \beta_0 + \beta_1 homogenized_{ien} + \beta_2 nbpartners_i + X'_{ien}\gamma_1 + \sigma_e + \phi_r + \epsilon_{ien} \quad (3)$$

In equation (3), y_{ien} is any of the outcome variables described in Section 3.3.1 for individual i, in ethnic group e, living in neighborhood n. $homogenized_{ien}$ and $nbpartners_i$ are defined as in Equation 2; and X_{ien} is a vector that includes a set of individual-level controls (including a gender indicator, age, education level fixed effects, religion fixed effects, sexual identity fixed effects, employment status fixed effects, marriage status fixed effects, an indicator for whether the respondent is a native, and an indicator for whether the respondent had unprotected sex in last 12 months). The term σ_e is an ethnicity fixed

effect; ϕ_r is a region fixed effect; and ϵ_{ien} is an unobserved error term assumed to be uncorrelated with the regressor given the controls.

The fact that ϵ_{ien} is uncorrelated with $homogenized_{ien}$ given controls implies that if two individuals belong to the same gender and the ethnic group, having the same number of sexual partners, then any difference in the extent to their partners' ethnic diversification is by accident.

Accessing Potential Bias from Unobservables

Despite attempts to include a set of observed individual controls, our main identification strategy could yet be biased by some unobservable factors which are correlated with intra-ethnic sexual link formation and our outcomes. Therefore, we assess the degree of selection on unobservables to gauge the severity of endogeneity bias due to unobserved characteristics from the OLS estimates. We use three different approaches proposed by Altonji et al. (2005), Oster (2019) and Gonzalez and Miguel (2015), respectively. The approach was proposed by Altonji et al. (2005) who measure the degree of selection on unobservables by using the degree of selection on observed variables, in order to determine how much stronger selection on unobservables would have to be compared to selection on observables in order to fully explain away our result. For this purpose, we compute the ratio $\frac{\hat{\beta_F}}{(\hat{\beta_R} - \hat{\beta_F})}$, where $\hat{\beta_F}$ is the estimated coefficient of the uniformity index obtained from a regression that includes a full set of controls, and $\hat{\beta_R}$ is the estimated coefficient of the uniformity index obtained from a regression that includes only a restricted set of controls.

The second approach we employ was developed by Oster (2019), who introduced the following equation of the bias-adjusted estimator: $\beta^* = \hat{\beta_F} - (\hat{\beta_R} - \hat{\beta_F}) \times \frac{R_{max} - R^*}{R^* - R}$. This estimator calculates the lower bound for the coefficient of interest, where $\hat{\beta_F}$ is the coefficient estimate with a full set of controls, $\hat{\beta_R}$ is the coefficient estimate with a restricted set of controls, R^* is the R^2 (coefficient of variation) from the regression with a full set of controls, and R is the R^2 from a regression with a restricted set of controls. For the most conservative case, $R_{max} = 1$ when the regression includes all the observable and unobservable variables. In the real world, however, measurement errors are inevitable, and

the value of R_{max} should therefore be normally less than one as suggested by Gonzalez and Miguel (2015). We assume $R_{max} = 1$ and three other values of $R_{max} = min\{\Pi R^*, 1\}$ suggested by Oster (2019), with the three suggested values of Π being 1.2, 1.4 and 1.6.

The third approach proposed by Gonzalez and Miguel (2015) proposes that the formula for the computation of the lower bound of the parameter of interest should assume $R_{max} = R^* + (R^* - R)$.

3.5.2 Alternative Identification Strategies

Lewbel IV Estimation

We also use a recent instrumental variables approach developed by Lewbel (2012) to address remaining endogeneity issues.

The conventional IV approach to addressing endogeneity issues assumes that the instrument should affect t the outcome of interest only through the endogenous regressor. This assumption is however very hard to satisfy, and even to test. The Lewbel IV method relaxes this assumption. Moreover, it does not necessarily redly on an external instrument for identification. Indeed, there is a sufficient number of exogenous variables, it constructs an instrument based on these variables. To summarize, when valid external instrumental variables can be challenged or are not available, an instrument is found by construction. This is important in our context as we do not have an external IV.

We estimate the coefficient β_1 in equation (3). However, due to the fact that the regressor $homogenized_{ien}$ could be mismeasured endogenous, we actually observe $homogenized^*_{ien}$, where

$$homogenized^* = homogenized + U, \quad E(U) = 0, \quad U \perp homogenized, y, X' \tag{4}$$

If we define V as the residual from the linear projection of $homogenized$ on X, then we have

$$homogenized = X'\gamma_2 + V, \quad E(XV) = 0 \tag{5}$$

After substituting out in equation (4), we obtain the triangular models with measurement errors as follows:

$$y_{ien} = \beta_0 + \beta_1 homogenized^*_{ien} + X'_{ien}\gamma_1 + \epsilon_1, \quad \epsilon_1 = \epsilon_{ien} - \beta_1 U \tag{6}$$

$$homogenized^*_{ien} = \alpha_0 + X'_{ien}\gamma_2 + \epsilon_2, \quad \epsilon_2 = V + U \tag{7}$$

In the presence of heteroskedastic disturbances in equation (6) or the "first-stage" regression, Lewbel (2012) demonstrates that the endogenous regressor can be estimated as in a linear 2SLS regression by using X and $\left(Z - \overline{Z}\right)\widehat{\epsilon_2}$ as the instruments, where X is a vector of observed exogenous variables such as individual's gender, age cohort, ethnicity, residential area, occupation, and education level etc., and Z can either include the entire set of variables X or a non-empty subset of X, and ϵ_2 is the vector of residuals from the "first stage regression" with all exogenous regressors. The identification assumption requires that $Cov(Z, \epsilon_j^2) \neq 0$ and $Cov(Z, \epsilon_1\epsilon_2) = 0$ (for $j = 2$ in the presence of a triangular model, and for both $j = 1$ and $j = 2$ in a reverse causality system), and there are no further restrictions imposed on Z.

Propensity Score Matching

A complementary and alternative method for causal inference with individual survey data is the propensity score matching (PSM). The method attempts to reproduce a randomized experiment in a nonexperimental setting with a treatment group and a control group. Intuitively, the idea is to obtain the effect of the uniformity index by estimating the mean difference between the STI status of individuals who homogenized sexual partners and matched sets of individuals who ethnically diversify sexual partners. The matching is based on an estimated probability known as the propensity score.

In our analysis, we define the treatment as same as the variable "homogenized" in Equation 2. Therefore, the treatment group consists of all individuals whose homophily index equals to one, whereas the control group consists of individuals whose homophily index is smaller than one.

To match the treatment group to the control group, we employ methods that exploit the implementation of propensity score matching (PSM) in two ways: the first is the estimation of the PSM model and the second is the implementation of the matching for clustered data. We first calculate a propensity score for each individual by running a logit model using all control variables X'_{ien} in our benchmark model (that of Table 3.4, i.e., columns (2) with full controls). Then, we conduct a balancing test for misspecification of the propensity score function to make sure that the observables are independent of the uniformity index conditional on the estimated propensity score. Inspired by Rothstein (2013), the third step that we use to match each treated individual to nontreated individuals with similar propensities is by using an Epanechnikov kernel estimator with a bandwidth of .05 to obtain a treatment effect on the treated.

However, when using propensity score matching strategies, we might face an issue related to the average treatment effect on the treated (ATT) being biased, especially in the presence of unobserved cluster-level covariates. This problem is common in multilevel structured large-scale surveys where data are collected at the individual level, while no sufficient information is available at higher levels.[4] Consequently, the contextual heterogeneity may not be fully captured by the observed variables. We therefore employ a "preferential within-cluster matching (PW)" proposed by Arpino and Cannas (2016), which is a way of implementing PSM when the dataset has a two-level structure. PW is based on a single-level logit propensity score model, but considers the hierarchical structure in the implementation of the matching. This approach starts by searching control units within clusters; if none is found, control units are searched in other clusters. This approach not only improves the balancing of cluster-level variables but also reduces the number of unmatched treated units that are only searched within the same cluster (Arpino and Cannas 2016).[5]

[4] In our analysis, we consider the survey as a two-level structural data where individual-level units are nested in 11 regions (or clusters). This problem arises when we omit a cluster-level covariate, which is correlated with the individual one. As a consequence, the error term at the cluster-level will be correlated with the individual-level covariate, leading to inconsistent parameter estimates (Arpino and Mealli 2011).

[5] See Thoemmes and West (2011), Arpino and Mealli (2011), Li et al. (2013), and Arpino and Cannas (2016) for more details on PSM with clustered data.

3.5.3 Main Empirical Results

The Uniformity Index and STIs: OLS Estimates

We start by reporting the findings regarding the effects of the uniformity index on the likelihood of acquiring the following sexually transmitted diseases: (1) chlamydia; (2) gonorrhoea; (3) warts; (4) herpes; and (5) any STI that being diagnosed in the past five years prior to the interview.

Estimates of equation (3) are reported in Table 3.4. Controls are included in two steps; we first include a restricted set of controls (in Columns (1), (3), (5), (7), and (9)), and then include all the controls to evaluate the sensitivity of the coefficient of interest (in Columns (2), (4), (6), (8), and (10)). The controls included in Column (1) are the number of most recent sexual partners. Column (2) additionally controls for gender, age, educational attainment, religion, sexual identity, employment status, marital status, an indicator of unprotected sex in the past year before the interview, and an indicator of whether the respondent is a native. We find that ethnically homogenizing sexual partner(s) decreases the likelihood of being infected with chlamydia in the past five years. Specifically, moving from an inter-ethnic relationship to an intra-ethnic relationship reduces the possibility of being infected with chlamydia by 1.49 percentage points with a restricted set of controls and by 1.65 percentage points with a full set of controls. In Columns (3) and (4), the outcome variable is gonorrhea status in the past five years, and here the uniformity index has a negative and significant effect as well. Columns (5)-(8) analyze the effect of the ethnically homogenizing sexual partner(s) on genital warts and herpes in the past five years before the interview; we find a negative impact that is not statistically significant for both outcomes after including the full set of controls. Similarly, Columns (9) and (10) report the effect of the uniformity index on whether one was diagnosed with any STI in the past five years, and we find a negative and statistically significant effect. A one standard deviation increase in the uniformity index leads to a 0.053 (resp. 0.051) standard deviation decrease in the predicted likelihood of being infected with any STI with a restricted (resp. full) set of controls.

We note that the coefficients are relatively stable for most outcomes, which indicates endogeneity problems are not so severe.[6]

Robustness Checks

In this sub-section, we undertake a number of robustness checks in order to validate our findings regarding the link between the uniformity index and STI prevalence.

Accessing Potential Bias from Unobservables The OLS analysis in Section 3.5.3.1 could suffer from unobservable factors correlated with sexual partner selection and subsequent STI prevalence. We assess the degree of selection on unobservables to gauge the severity of endogeneity bias due to unobserved characteristics from the OLS estimates. We use three different approaches proposed by Altonji et al. (2005), Oster (2019) and Gonzalez and Miguel (2015), respectively.

In column (1) of Table 3.5, we report the Altonji et al. (2005) ratio for each of the 5 outcome variables listed in Table 3.4. The full set of controls includes the number of most recent sexual partners, gender, age, educational attainment, religion, sexual identity, employment status, marital status, region fixed effect, ethnicity fixed effect, an indicator of unprotected sex, and an indicator for whether the respondent is a native. The restricted set of controls only includes the number of most recent sexual partners, region fixed effect, and ethnicity fixed effect. The reported ratios in Table 3.5 range from -137 to 72, which are obviously very large. All these ratios imply that the influence of unobservable factors would have to be far greater than that of observable characteristics to fully account for our results. It is therefore unlikely that the estimated effect of romantic homophily on any of the outcomes is driven by unobservables.

The second approach we employ was developed by Oster (2019). Columns (2)-(6) in Table 3.5 produce the lower bounds of the coefficients of the uniformity index correspond-

[6]Finally, all OLS analyses employ a wild cluster-bootstrap (Cameron et al. 2008) due to the small number of clusters in our data, and robust standard errors clustered at the regional level and multiple levels. These results are similarly robust.

ing to $R_{max} = 1$ and to three suggested values of R_{max}. The most conservative case shown in column (2), where no measurement error is assumed, is too unrealistic to produce informative bounds; therefore, most of the coefficient intervals do not exclude zero (warts, herpes, and any STI in the past five years). Now we consider the most realistic cases suggested by Oster (2019), where $R_{max} = min\{\Pi R^*, 1\}$ with varying values of Π. When Π equals to 1.2, 1.4, 1.6 and 1.8, respectively, all lower bound estimates remain negative.

The third approach proposed by Gonzalez and Miguel (2015) proposes that the formula for the computation of the lower bound of the parameter of interest should assumes $R_{max} = R^* + (R^* - R)$. The results are reported in Column (7) in Table 3.5. Here too, we find that the lower bound is smaller than zero for all outcomes. All these results imply that it is unlikely that our OLS estimates are biased by some unobserved characteristics.

Lewbel IV we also employ the instrumental variable approach developed by Lewbel (2012) to identify the effect of an endogenous regressor when external instrumental variables are not available or when their validity can be challenged. Here, an instrument is found by construction. The findings are reported in Table 3.6 with the robust standard errors clustered at the region level and multiple levels, respectively. The results generally show that our estimates obtained from the Lewbel's approach are a bit larger and significant than those obtained from the OLS estimation, but the standard errors are also large. In Column (3) of Table 3.6, for instance, we detect a negative and significant effect of the uniformity index on genital warts in the past five years from the regression with a full set of controls.

It's worth noting that the endogenous regressor here is a binary variable, and the estimator is also applied to estimate a treatment effect (i.e., ethnically homogenizing sexual partners or not) when the treatment is not randomly assigned. Therefore, some studies have raised questions about whether Lewbel (2012) can be applied to such a binary regressor case, especially when the external instrumental variable is not available (Emran et al. 2014, Lewbel 2018). Lewbel (2012) does not explicitly assume that the endogenous regressor is continuous. The larger standard errors of the Lewbel IV estimates may be reflecting the inefficiency of the heteroskedasticity-based estimator noted in Lewbel (2018) due to the

endogenous regressor being binary.

Propensity Score Matching We also tackle endogeneity using propensity score matching. As described in Section 3.5.2.2, we use two different model specifications to estimate propensity scores in this paper, and each of the matching procedures is executed in three steps. After estimating a propensity score for each individual using a logit model with all covariates that represent individual-specific features, we then conduct a balancing test for misspecification of the propensity score function. Here, we use a similar method as Rothstein (2013) with a higher-order polynomial.[7] The third step of each PSM model involves two different matching algorithms, using an Epanechnikov kernel function with a bandwidth of 0.05 and 3% trimming; and a one-to-one caliper matching with replacement, and common support, respectively.[8] The latter model is also based on the ATT estimation but deals with the potential bias by matching units within region cluster.

Figures 3.5 illustrates the comparative distribution graphic for the balance check using an Epanechnikov kernel estimator. As Figure 3.5 shows, after matching, a good overlap between treated and control groups in terms of their adoption propensities.[9] For the preferential within-cluster matching (PW), the average percentage of Absolute Standardized Bias (ASB) after matching is 4.2%, which is considered acceptable for values of ASB after the matching which is smaller than 5% (Caliendo and Kopeinig 2008).

Entries in Table 3.7 reports propensity score matching estimates using the two ways described above. Column (1) presents unweighted propensity score matching estimates with a 50-replication bootstrapped standard errors. The propensity score is estimated using a logit model with all individual-specific characteristics as explanatory variables; the matching method uses an Epanechnikov kernel with a bandwidth of .05 to obtain

[7]The repeated testing procedure that is adding higher-order and interaction terms to the logistic model for balancing the treatment and comparison groups, which will cause lower quality of the match (Dehejia and Wahba 2002, Smith and Todd 2005).

[8]Here, we choose a caliper as 0.25 of standard deviations of the estimated propensity score, which in our case could effectively avoid bad matches.

[9]However, if we use Mahalanobis-distance kernel matching, we will have two distinguishable groups in terms of their adoption propensities, the graphic is not shown here.

ATT. Common support is imposed by dropping 3% of the treated observations in which the propensity score density of the control observations is the lowest (Rothstein 2013). Column (2) shows the relative bias obtained with bootstrap, which almost has no difference from zero for all outcomes. Column (5) reports the PW estimation of implementing the matching at the regional level. We see the PW matching is robust to the Epanechnikov kernel matching presented in Column (1). Therefore, our PSM results do not suffer from the so-called second-level endogeneity problem in the multilevel dataset.

Interestingly, if we compare the results from the propensity score matching with those from the OLS estimation (Table 3.4 and 3.7), we find a higher effect of homogenizing sexual partners in both ways of the propensity score matching estimates. This is also in line with the results that we obtain by running Lewbel IV estimations, although the Lewbel IV estimates might be inefficient due to the binary endogenous regressor. Hence, both the two ways of propensity score matching and the IV technique suggest that OLS estimates suffer from downwards bias due to endogeneity issues. Taken together, the consistently statistically significant estimates from this alternative identification method provide us with robust evidence on the negative effect of the ethnically homogenizing sexual partner(s) and the likelihood of being infected with various STIs.

3.5.4 Heterogeneous Effects of Romantic Homophily by Gender, Minority Status, and Sexual Identity

Having established a robust effect of romantic homophily on various sexually transmitted infections and diseases, we would now like to investigate whether this effect varies by gender, minority status, and sexual identity. We split the sample by gender, minority status, and sexual identity, and estimate the effect of homophily in each sub-sample. The results are presented in Table 3.8. Entries in Columns (1) and (2) show the effect for women and men, separately. For most of the sexually transmitted diseases, the magnitude of the impact is larger and statistically significant for women compared to men. In fact, the effect of homophily is not statistically significant for men.

Entries in Columns (3)-(6) show the effects of romantic homophily on STIs for majority

and minority individuals, and for heterosexuals and non-heterosexuals. The effects are much larger for majorities compared to minorities. They are also larger for heterosexuals compared to non-heterosexuals. The analysis complements studies of how the impact of homophily varies by minority status and population subgroups (Ananat and Washington 2009, Centola 2011, Halberstam and Knight 2016, Karimi et al. 2018).

3.5.5 Romantic Homophily and Non-Communicable Health Conditions: A Falsification Test

In this section, we conduct a reverse-placebo test by analyzing the effect of romantic homophily on health conditions not caused by sexual interactions. The conditions we analyze are non-communicable conditions, including arthritis, heart disease, stroke, diabetes, and chronic lung disease. Our test attempts to rule out alternative interpretations of the documented effect of romantic homophily on STDs. We estimate Equation (3), where the outcome variable is each of the aforementioned health conditions. The OLS, Lewbel IV and PSM estimates are presented in Table 3.9. We find that the uniformity index does not affect most of these conditions (except for chronic lung disease when using, although the magnitude of the effect is close to zero). In general, the estimates are close to zero. The fact that we do not find any consistent effect of ethnically homogenizing sexual partners on any of these conditions lends more credibility to the interpretation of our main findings.

3.6 Conclusion

This paper studies the effect of romantic homophily on the spread of STIs using unique data from the United Kingdom. Individuals who are central in social networks are more likely to acquire and spread information of all sorts. If a sexual network, information can be a sexually transmitted infection, and in this case, central individuals are at a higher risk of contracting and spreading STIs. But information can also be a new technology to prevent STIs, and in this case, central individuals are less likely to be infected with STIs. We test these conflicting hypotheses by analyzing the causal effect of sexual homophily on

a wide range of STIs. The homophily index measures the extent to which one is peripheral (as opposed to central) in a sexual network. We find that romantic homophily significantly decreases the probability of being infected with chlamydia, gonorrhoea, warts, herpes, and any STI diagnosed in the past five years prior to the interview. These effects are robust to controlling for a wide range of factors, including the number of sexual partners and the use of protective contraception. The findings suggest that the first hypothesis dominates. These effects are larger for women, majority individuals, and heterosexuals.

Last but not least, in Appendix Table AA1, we estimate a regression that controls for homophily, sex ratio, and ethnic heterogeneity along with the range of individual and contextual factors included in most of the regressions, and we find that these variables still affect sexual behavior and STI status.

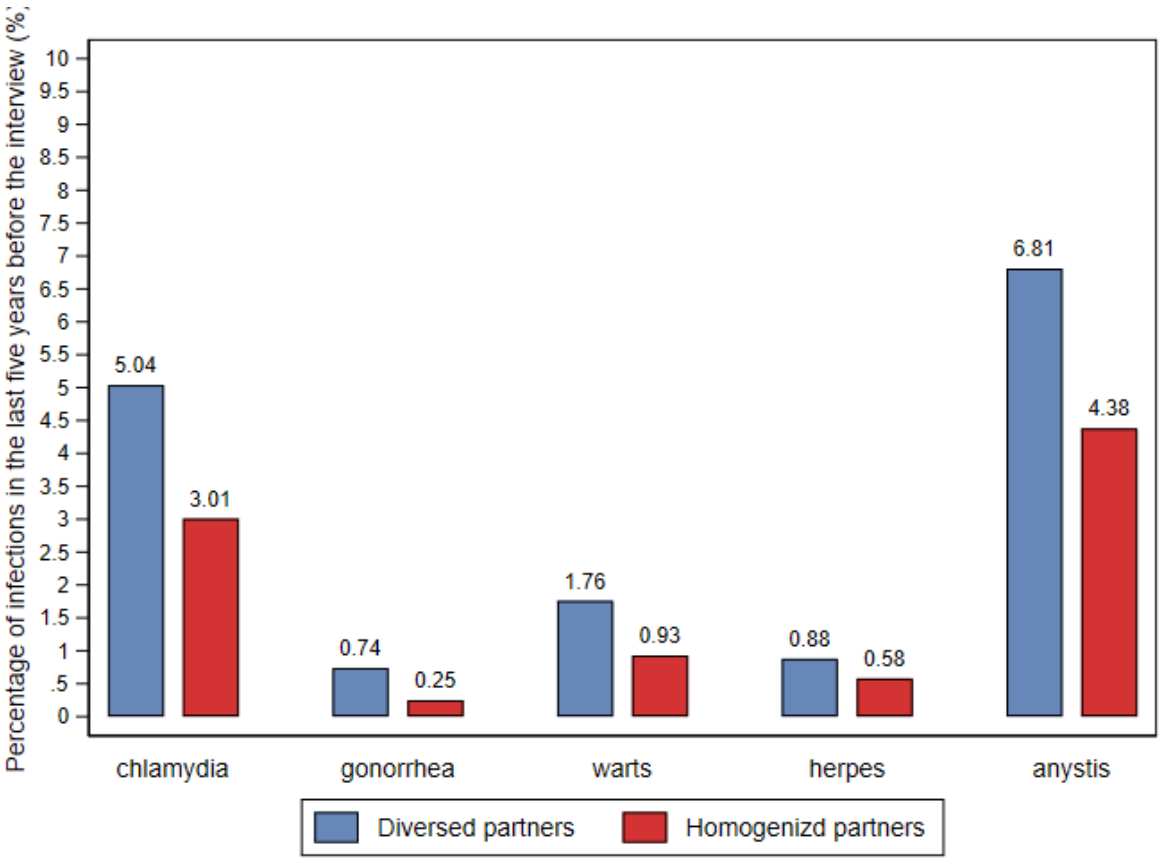
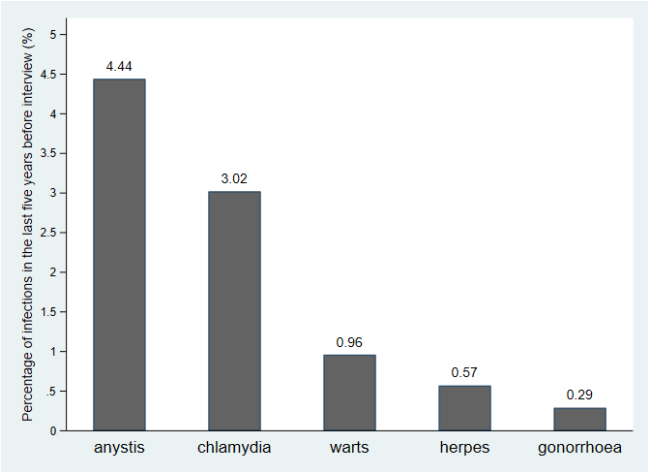
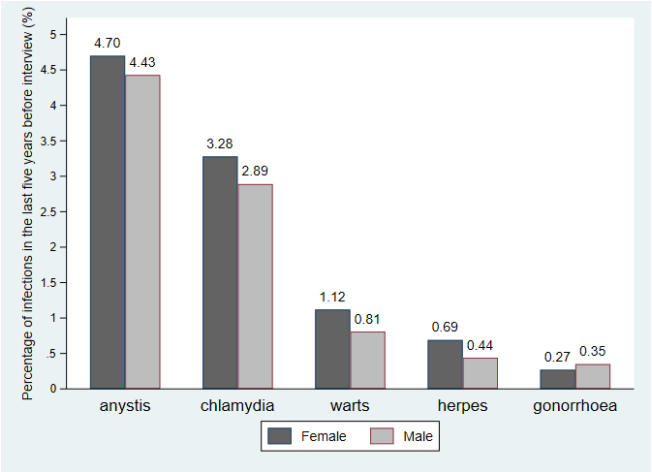


Figure 3.1: Prevalence of STIs between individuals who homogenize vs diversify sexual partners

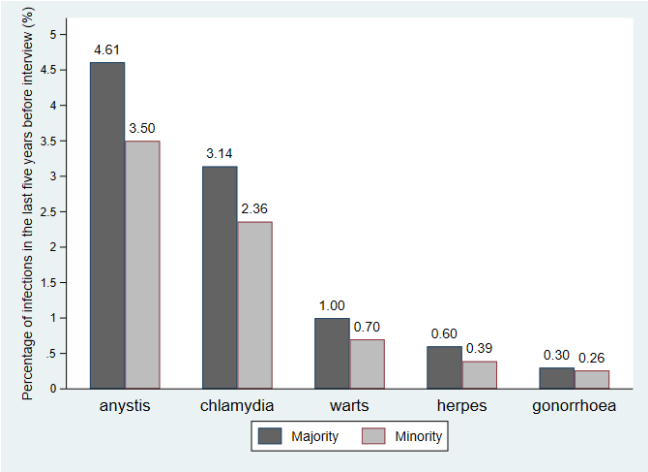
Data source: The third British National Survey of Sexual Attitudes and Lifestyles 2010-2012 (Natsal3).



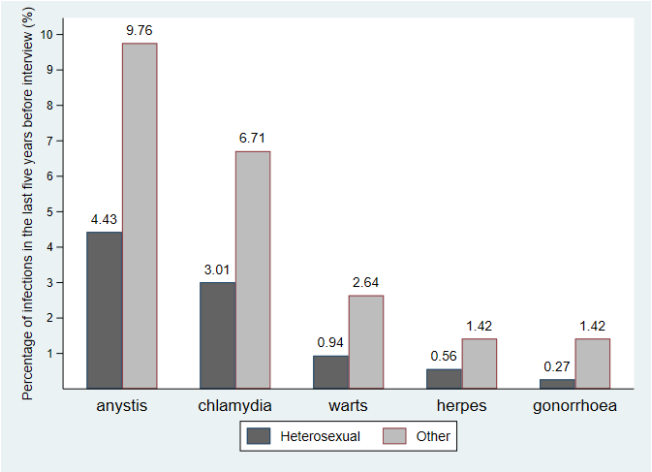
(a) Percentage of infections



(b) By gender



(c) By ethnic group



(d) By sexual identity

Figure 3.2: Descriptive statistics of infection rate

Data source: The third British National Survey of Sexual Attitudes and Lifestyles 2010-2012 (Natsal3).

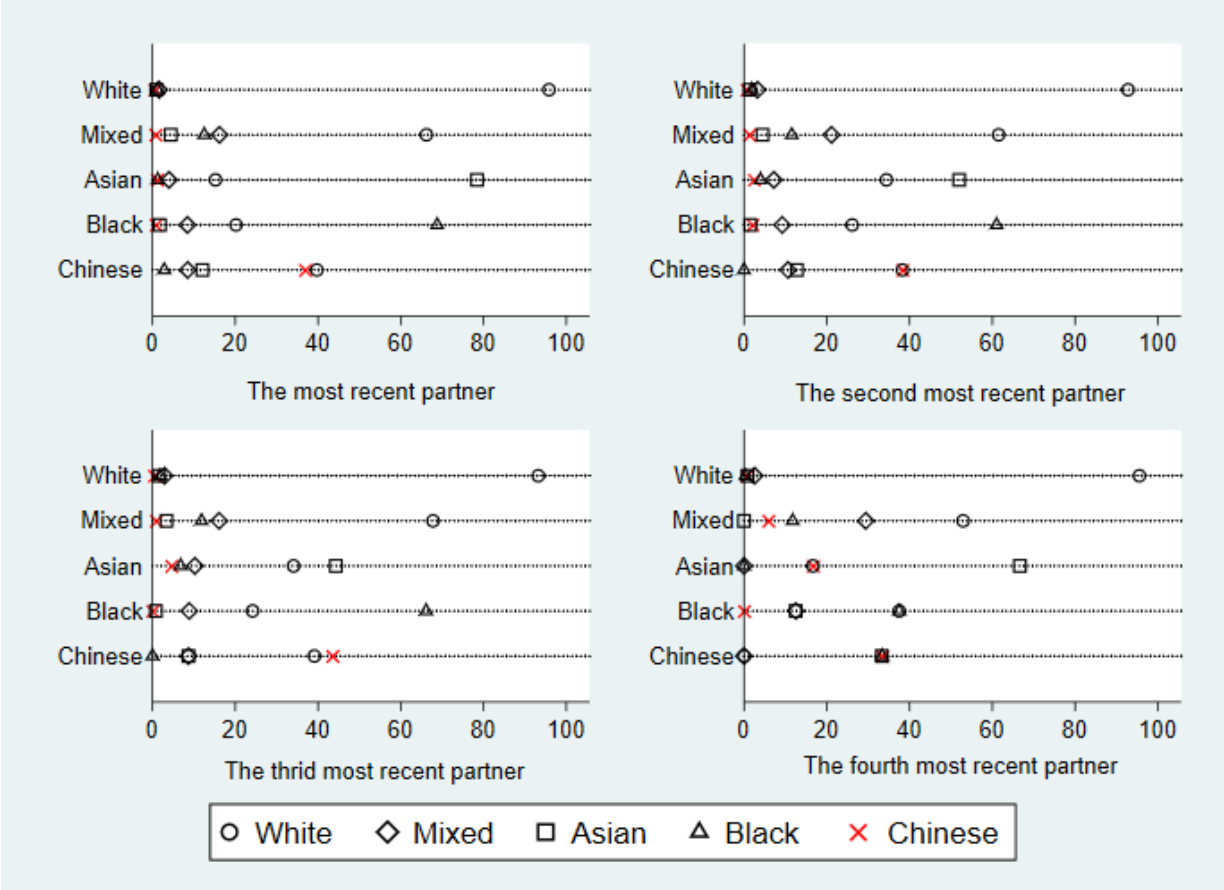
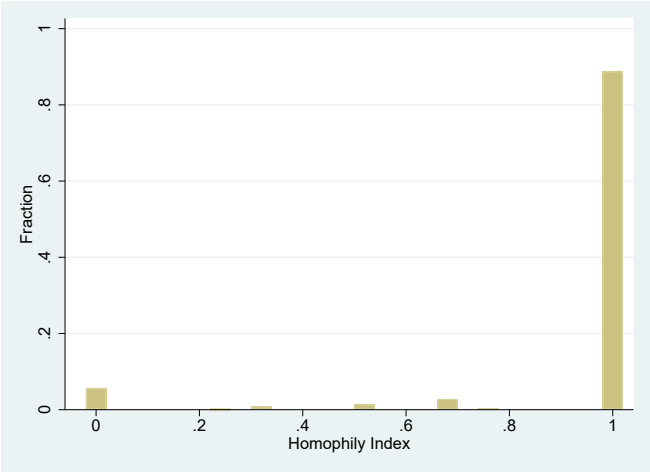
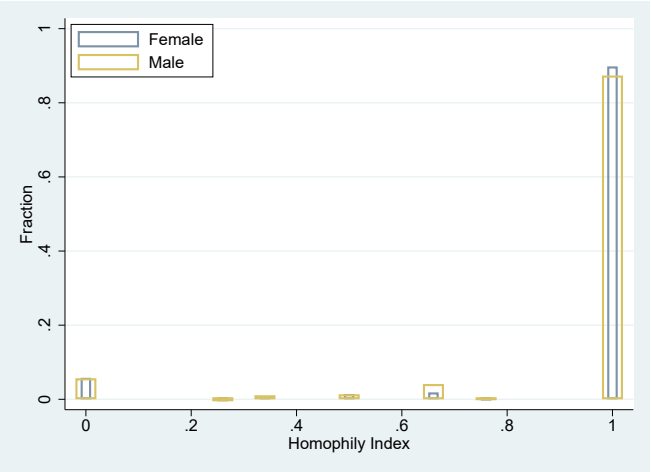


Figure 3.3: Percentage of partnerships within and across ethnic groups

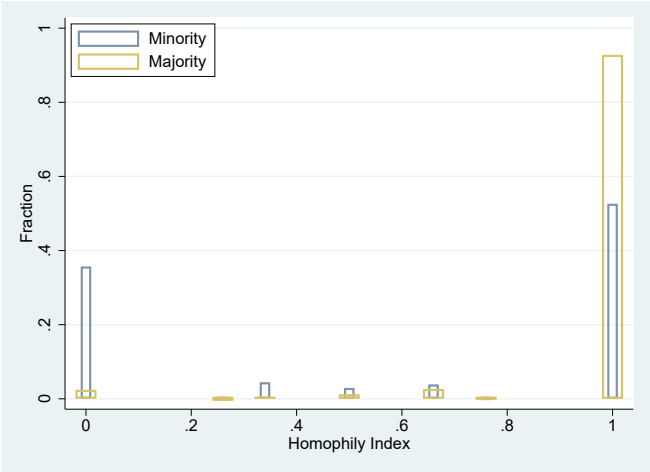
Data source: The third British National Survey of Sexual Attitudes and Lifestyles 2010-2012 (Natsal3).



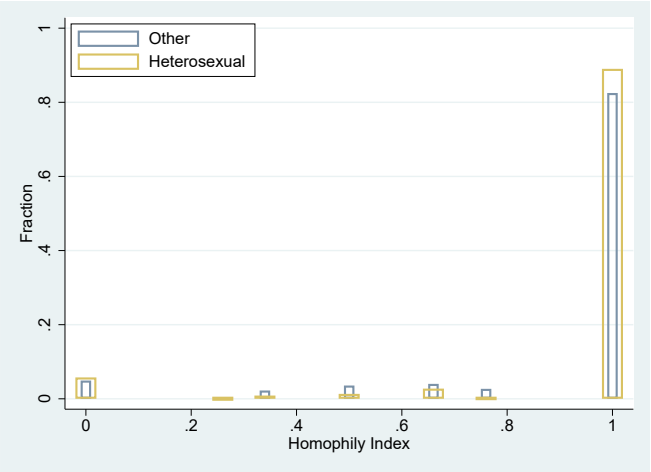
(a) All sample



(b) By gender



(c) By ethnic group



(d) By sexual identity

Figure 3.4: Distribution of the romantic homophily Index in different groups

Data source: The third British National Survey of Sexual Attitudes and Lifestyles 2010-2012 (Natsal3).

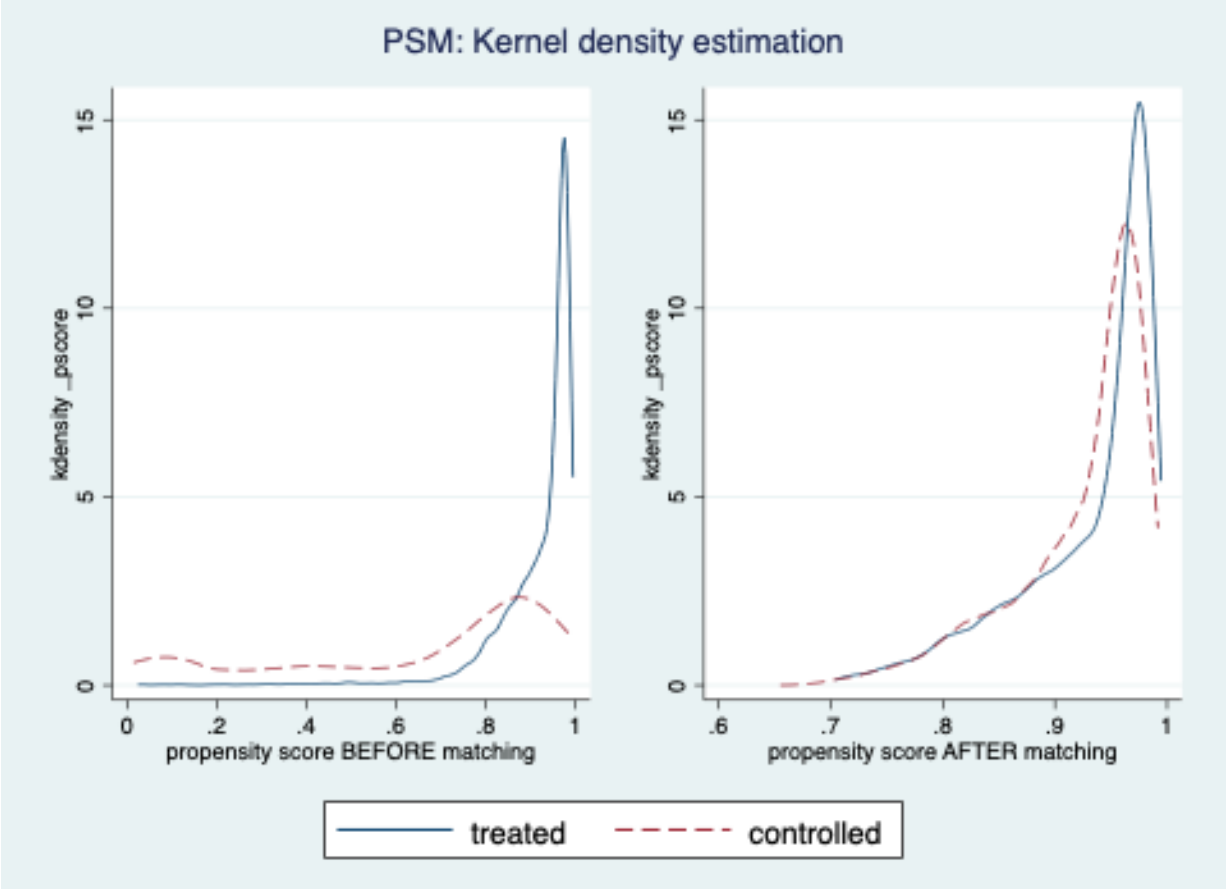


Figure 3.5: Distribution of the propensity score before and after matching by using Epanechnikov kernel density estimation

Notes: In each case, the solid line is a kernel density estimate of the conditional density of the propensity score among treated units (i.e., individuals who ethnically homogenize sexual partner(s) for a representative data set). The dashed line is for the conditional density among control units (i.e., individuals who racially diversifying sexual partner(s)).

Table 3.1: Summary Statistics

Variables	Observations	Mean	Std. Dev.
Sexual Homophily Index	13,712	0.919	0.249
homogenized	13,712	0.888	0.315
Controls			
No. of most recent partners	15,019	1.518	0.990
Has zero sex partner	15,019	0.093	0.291
Has one sex partner	15,019	0.548	0.498
Has two sex partners	15,019	0.124	0.330
Has three sex partners	15,019	0.215	0.411
Has four sex partners	15,019	0.019	0.136
Age	15,162	38.390	16.802
Male	15,162	0.415	0.493
Majority	15,162	0.876	0.329
Native	15,132	0.410	0.492
Total unprotected sex	13,438	0.066	0.248
Marriage status			
Married	15,124	0.353	0.478
Cohabitation	15,124	0.118	0.323
Previously married	15,124	0.141	0.348
Single and never married	15,124	0.388	0.487
Education			
Degree	15,117	0.238	0.426
Higher education	15,117	0.274	0.446
GCSE	15,117	0.352	0.478
Foreign or others	15,117	0.012	0.108
None	15,117	0.124	0.330
Religion			
None	15,114	0.511	0.500
Christian-Church of England/Anglican	15,114	0.143	0.350
Christian - Roman Catholic	15,114	0.097	0.296
Christian - Other	15,114	0.182	0.386
Non-Christian	15,114	0.066	0.249
Employment Status			
Employed	15,116	0.542	0.498
Full-time education	15,116	0.129	0.336
Unemployment	15,116	0.205	0.404
Retired	15,116	0.124	0.329
Sex identity			
Heterosexual/straight	15,109	0.967	0.177
Gay/lesbian	15,109	0.014	0.118
Bisexual	15,109	0.015	0.121
Other	15,109	0.004	0.059

Notes: The table shows means and standard deviations for key variables used in the analysis. The standard deviations are in parentheses. The variables measuring sexual homophily are constructed using the third British National Survey of Sexual Attitudes and Lifestyles 2010-2012 (Natsal3)

Table 3.2: Ethnic homogeneity of most recent partners by number of partners, gender, and majority status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Number of most recent partners	All	Male	Female	Majority	Minorities	Majority Male	Majority Female	Minority Male	Minority Female
1	0.935	0.931	0.937	0.966	0.656	0.962	0.968	0.624	0.672
2	0.853	0.855	0.851	0.897	0.449	0.905	0.892	0.394	0.482
3	0.713	0.711	0.715	0.783	0.220	0.785	0.782	0.216	0.224
4	0.758	0.727	0.773	0.804	0.150	0.753	0.831	0.200	0.133
2+	**0.698**	**0.749**	**0.764**	**0.819**	**0.159**	**0.816**	**0.822**	**0.255**	**0.455**

Table 3.3: OLS estimates of the effects of gender, majority status, and the number of partners on ethnic homogeneity of partners

Dependent variable: Uniformity Index

	All Samples				Majority Group		Minority Groups	
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male	-0.0113**	-0.0119*	-0.0126**	-0.0126*	-0.0049	-0.0046	-0.0503	-0.0887*
	(0.0050)	(0.0058)	(0.0050)	(0.0060)	(0.0043)	(0.0044)	(0.0377)	(0.0436)
Majority	0.377***	0.444***						
	(0.0493)	(0.0589)						
Own Group Size			0.0066***	0.0068***				
			(0.0005)	(0.0006)				
Had 2 partners	0.869***	0.873***	0.869***	0.873***	0.912***	0.914***	0.476***	0.497***
	(0.0022)	(0.0017)	(0.0021)	(0.0017)	(0.0016)	(0.0013)	(0.0062)	(0.0030)
Had 3 partners	0.798***	0.816***	0.800***	0.817***	0.856***	0.865***	0.284***	0.388***
	(0.0100)	(0.0079)	(0.0096)	(0.0076)	(0.0072)	(0.0060)	(0.0296)	(0.0140)
Had 4 partners	0.728***	0.760***	0.731***	0.761***	0.799***	0.816***	0.092*	0.278***
	(0.0179)	(0.0140)	(0.0171)	(0.0136)	(0.0128)	(0.0106)	(0.0529)	(0.0251)
Observations	13,685	13,560	13,685	13,560	12,321	12,225	1,364	1,335
Number of regions	11	11	11	11	11	11	11	11
Adjusted R-squared	0.164	0.180	0.144	0.151	0.042	0.056	0.134	0.256
Reginal FE	Y	Y	Y	Y	Y	Y	Y	Y
Controls	N	Y	N	Y	N	Y	N	Y

Table 3.4: Romantic homophily and STI prevalence: OLS estimates

	chlamydia5years		gonorrhoea5years		warts5years		herpes5years		anystis5years	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Homogenized	-0.0149***	-0.0165***	-0.0053*	-0.0074*	-0.0079*	-0.0071	-0.0035	-0.0027	-0.0302***	-0.0294***
	(0.0041)	(0.0038)	(0.0028)	(0.0035)	(0.0038)	(0.0044)	(0.0035)	(0.0039)	(0.0057)	(0.0076)
	[0.0002]	[0.0002]	[0.0000]	[0.0000]	[0.0001]	[0.0001]	[0.0000]	[0.0000]	[0.0007]	[0.0009]
Observations	14,231	14,105	14,228	14,102	14,229	14,103	14,229	14,103	14,522	14,390
R-squared	0.052	0.067	0.009	0.014	0.014	0.019	0.005	0.010	0.068	0.080
Full Individual Controls	N	Y	N	Y	N	Y	N	Y	N	Y
FE	N	Y	N	Y	N	Y	N	Y	N	Y

Notes: Each dependent variable is a dichotomous variable equal to one if the respondent reported being infected with the indicated infection in the past five years prior to the interview, and zero otherwise. The independent variable is a binary variable equal to one if the respondent's index of romantic homophily is 1, and zero otherwise. Below each coefficient, two standard errors are reported. The first, reported in parentheses, is the standard error adjusted for clustering within regions. The second, reported in square brackets, is the standard error adjusted for two-way standard errors clustering within ethnic groups and within regions. Fully controlled regressions include the number of most recent sexual partners, a gender indicator, age, education level fixed effects, religion fixed effects, sexual identity fixed effects, employment status fixed effects, marital status fixed effects, region fixed effects, ethnicity fixed effects, the majority group size of the neighborhood where the respondent lives, an indicator for the frequency of unprotected sex in the past year before the interview, and an indicator for whether the respondent is a native. *, ** and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 3.5: The importance of bias from unobservables: Altonji et al. (2005), Oster (2019), and Gonzalez and Miguel (2015)

	Coeff. Ratio after Altonji et al. (2005)	Mininum Coeff. ($R_{max} = \Pi R^*$)	Lower Bound after Oster (2019)			Coefficient after Gonzalez and Miguel (2015) ($R_{max} = R^* + (R^* - R)$)
Outcomes	AET	R_{max}=1	$\Pi = 1.2$	$\Pi = 1.4$	$\Pi = 1.6$	$\Pi = 1.8$	β^*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
chlamydia5years	-17.321	-0.091	-0.018	-0.019	-0.020	-0.021	-0.018
gonorrhoea5years	-137.321	-0.023	-0.007	-0.007	-0.008	-0.008	-0.007
warts5years	21.735	0.075	-0.007	-0.006	-0.006	-0.006	-0.007
herpes5years	8.213	0.106	-0.003	-0.002	-0.002	-0.002	-0.002
anystis5years	72.360	0.011	-0.029	-0.028	-0.027	-0.027	-0.029

Notes: This table shows the selection on unobservables. The independent variable is the uniformity index, and the dependent variables are the sexually transmitted infections which are listed on the left side of the table. The cells in conlumn (1) report the coefficient for the uniformity index in two regressions, a restricted set of controls and a full set of controls. The regression with a restricted set of controls only includes the number of most recent sexual partners, region fixed effect, and ethnicity fixed effect; and the regression with a full set of controls includes all the independent variables The ratio is calculated as $\frac{\hat{\beta_F}}{\hat{\beta_R} - \hat{\beta_F}}$, where $\hat{\beta_F}$ is the coefficient estimates in a full-controls regression, and $\hat{\beta_R}$ is the coefficient estimate in the restricted-control regression. Columns (2)-(5) present coefficient lower bounds by employing Oster (2019), with the bias-adjusted estimator equation calculated as $\hat{\beta_F} - (\hat{\beta_R} - \hat{\beta_F}) * \frac{R_{max} - R^*}{R^* - R}$, where R^* and R are the R^2 for the regressions with the full set of controls and the restricted set of controls, respectively. The cells in conlumn (2) report the coefficient lower bounds with the most conservative value equals to one, and columns (3)-(5) with $R_{max} = min\{\Pi R^*, 1\}$ suggested by Oster (2019), and the values of the three multipliers are 1.2, 1.4 and 1.6. Column (6) reports the coefficient using $R_{max} = R^* + (R^* - R)$ which was suggested by Gonzalez and Miguel (2015).

Table 3.6: Romantic homophily and STIs: Lewbel IV estimates

	chlamydia5years	gonorrhoea5years	warts5years	herpes5years	anystis5years
Variable	(1)	(2)	(3)	(4)	(5)
Homogenized	-0.0164**	-0.0107*	-0.0112**	-0.0068	-0.0335***
	(0.0066)	(0.0065)	(0.0049)	(0.0048)	(0.0118)
	[0.0031]	[0.0057]	[0.0024]	[0.0022]	[0.0172]
Constant	0.0269	-0.0036	0.0092	0.0155***	0.0442**
	(0.0172)	(0.0038)	(0.0080)	(0.0054)	(0.0222)
Observations	14,105	14,102	14,103	14,103	14,390
R-squared	0.067	0.014	0.019	0.010	0.080
Individual Controls	Y	Y	Y	Y	Y
FE	Y	Y	Y	Y	Y

Notes: Each dependent variable is a dichotomous variable equal to one if the respondent reported being infected with the indicated infection in the past five years prior to the interview, and zero otherwise. The independent variable is a binary variable equal to one if the respondent's index of romantic homophily is 1, and zero otherwise. Below each coefficient, two standard errors are reported. The first, reported in parentheses, is the standard error adjusted for clustering within regions. The second, reported in square brackets, is the standard error adjusted for two-way standard errors clustering within ethnic groups and within regions. Fully controlled regressions include the number of most recent sexual partners, a gender indicator, age, education level fixed effects, religion fixed effects, sexual identity fixed effects, employment status fixed effects, marital status fixed effects, region fixed effects, ethnicity fixed effects, the majority group size of the neighborhood where the respondent lives, an indicator for the frequency of unprotected sex in the past year before the interview, and an indicator for whether the respondent is a native. *, ** and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 3.7: Romantic homophily and and STIs: PSM estimates

Variable	PSM				PW
	Estimator	Bias	[95% Conf. Interval]		
	(1)	(2)	(3)	(4)	(5)
chlamydia5years	-0.0247**	-0.0011	-0.0448	-0.0039	-0.0264**
	(0.0093)				(0.0086)
gonorrhoea5years	-0.0043*	-0.0001	-0.0086	-0.0005	-0.0040*
	(0.0021)				(0.0018)
warts5years	-0.0053	-0.0008	-0.0129	0.0014	-0.0025
	(0.0039)				(0.0041)
herpes5years	0.0003	-0.0004	-0.0022	0.0028	0.0012
	(0.0015)				(0.0018)
anystis5years	-0.0324***	0.0016	-0.0459	-0.0219	-0.0302**
	(0.0070)				(0.0099)
All Controls & FE			Y		Y

Notes: Column (1) reports unweighted propensity score matching estimates (the average treatment effect on the treated) with bootstrapped standard errors in parentheses (50 replications). The propensity score is estimated using a logit model with all individual-specific characteristics as explanatory variables; the matching method uses an Epanechnikov kernel with a bandwidth of .05 and 3% trimming. Column (2) shows the relative bias obtained with bootstrap, and Columns (3)-(4) report 95% bootstrap confidence intervals. Column (5) reports the PW estimation of implementing of the matching at the regional level. *, ** and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 3.8: Ethnically homogenizing sexual partner and STI prevalence: By sub-samples

Variables	By gender		By ethnicity		By sexual identity	
	Female	Male	Majority	Minority	Heterosexual	Others
	(1)	(2)	(3)	(4)	(5)	(6)
chlamydia5years	-0.0260***	-0.0051	-0.0195**	-0.0062	-0.0165***	-0.0104
	(0.0052)	(0.0066)	(0.0065)	(0.0156)	(0.0036)	(0.0298)
gonorrhoea5years	-0.0046*	-0.0086	-0.0083**	-0.0019	-0.0081**	0.0078
	(0.0024)	(0.0062)	(0.0030)	(0.0050)	(0.0034)	(0.0166)
warts5years	-0.0127	-0.0008	-0.0070	-0.0039	-0.0059	-0.0215
	(0.0071)	(0.0036)	(0.0044)	(0.0064)	(0.0048)	(0.0339)
herpes5years	-0.0019	-0.0040	-0.0045	0.0008	-0.0014	-0.0228
	(0.0046)	(0.0045)	(0.0063)	(0.0042)	(0.0030)	(0.0307)
anystis5years	-0.0403***	-0.0160	-0.0386***	0.0017	-0.0277***	-0.0467
	(0.0081)	(0.0092)	(0.0080)	(0.0203)	(0.0068)	(0.0309)
Individual Controls	Y	Y	Y	Y	Y	Y
FEs	Y	Y	Y	Y	Y	Y

Notes: Each dependent variable is a dichotomous variable equal to one if the respondent reported being infected with the indicated infection in the past five years prior to the interview, and zero otherwise. The independent variable is a binary variable equal to one if the respondent's index of romantic homophily is 1, and zero otherwise. Below each coefficient, two standard errors are reported. The first, reported in parentheses, is the standard error adjusted for clustering within regions. The second, reported in square brackets, is the standard error adjusted for two-way standard errors clustering within ethnic groups and within regions. Fully controlled regressions include the number of most recent sexual partners, a gender indicator, age, education level fixed effects, religion fixed effects, sexual identity fixed effects, employment status fixed effects, marital status fixed effects, region fixed effects, ethnicity fixed effects, the majority group size of the neighborhood where the respondent lives, an indicator for the frequency of unprotected sex in the past year before the interview, and an indicator for whether the respondent is a native. *, ** and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 3.9: Falsification test: Effects of romantic homophily on non-transmitted diseases

Variable	OLS	Lewbel IV	PSM
	(1)	(2)	(3)
arthritis	-0.0103	-0.0007	-0.0033
	(0.0058)	(0.0072)	(0.0141)
heart disease	-0.0043	-0.0051	-0.0022
	(0.0053)	(0.0071)	(0.0083)
stroke	-0.0010	-0.0034	-0.0012
	(0.0034)	(0.0034)	(0.0063)
diabetes	-0.0088	-0.0010	-0.0077
	(0.0052)	(0.0075)	(0.0085)
chronic lung disease	-0.0026	0.0037	-0.0097*
	(0.0027)	(0.0024)	(0.0046)
Individual Controls	Y	Y	Y
FE	Y	Y	Y

Notes: All regressions include a constant and a full set of controls. Robust standard errors are in parentheses. Regression disturbance terms are clustered at the regional level. *, ** and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

APPENDICES

Table AA1: How sex ratio, racial homophily, and ethnic diversity are related?

Variables	No. of hetero partners in last 5 years	No. of hetero partners are serial in last 5 years	Had concurrent relationship last 5 years	Been diagnosed Chlamydia in last 5 years	Been diagnosed any STI in last 5 years
	OLS		Marginal Effect		
	(1)	(2)	(3)	(4)	(5)
Sex Ratio	-2.384***	-1.564***	-0.088***	-0.018***	-0.031***
	(0.267)	(0.313)	(0.018)	(0.006)	(0.006)
Racial Homophily	-0.099	-0.999***	-0.044***	-0.007**	-0.009**
	(0.225)	(0.139)	(0.011)	(0.003)	(0.004)
Ethnic Diversity	0.426*	0.088	0.002	0.008**	0.011*
	(0.203)	(0.224)	(0.020)	(0.004)	(0.006)
Observations	12,350	9,023	12,397	10,860	12,291
Adjusted(Pseudo) R-squared	0.126	0.143	0.088	0.156	0.151
FE & Controls	Y	Y	Y	Y	Y

Notes: Dependent variables are sexual behaviors and infection status. Refer to Table A1 for the detailed definitions of the dependent variables. The table reports OLS (in columns (1)-(2)) and marginal probit (in columns (3)-(5)) estimates. The unit of observation is an individual. Standard errors are adjusted for clustering within regions. Fully controlled regressions include a gender indicator, age, education fixed effect, religion fixed effect, employment fixed effect, marital status fixed effect, region fixed effect, ethnic group fixed effect, the majority group's size in the neighborhood where the respondent lives, and an indicator for whether the respondent is a native. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

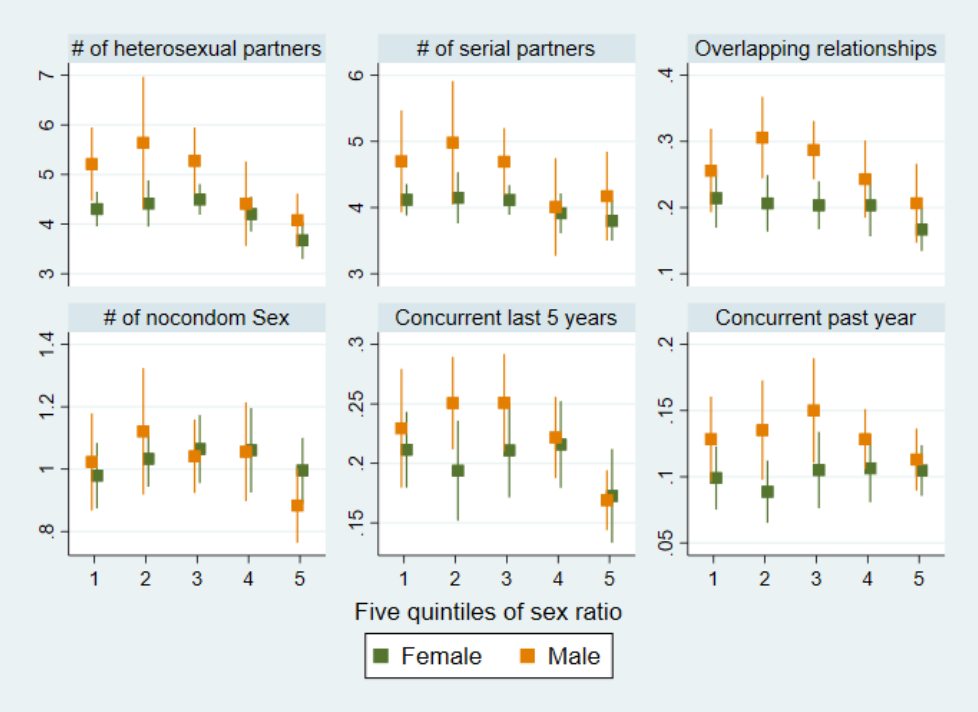


Figure A1: The effect of the sex ratio on sexual behavior (Whites under 50)

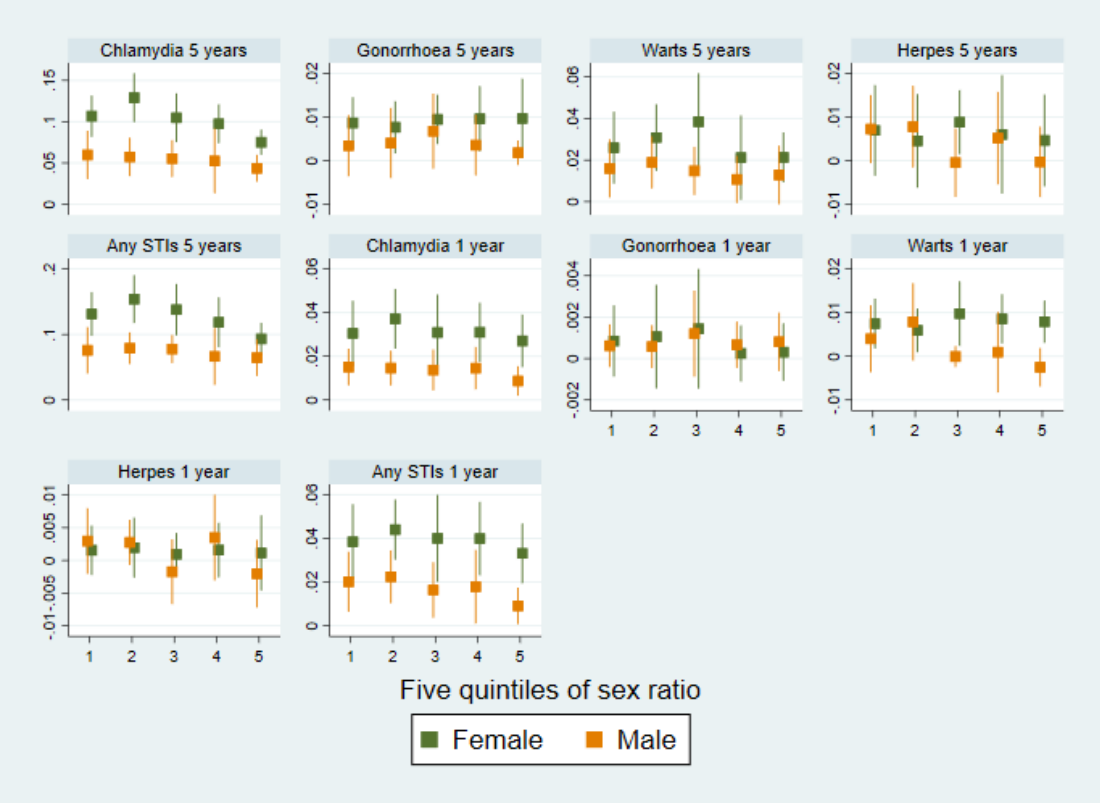


Figure A2: The effect of the sex ratio on STIs (Whites under 50)

Table A1: Definition of Sexual Behavior Variables and Sexually Transmitted Infections

Variables	**Description**
Sexual Behaviors	
het5yrs	Number of heterosexual partners in the last 5 years before the interview
serial5yrs	Number of heterosexual partners are serial monogamy in the last 5 years before the interview
overlap5y	Had any overlap between partners in the last 5 years before the interview
concu5y	Had a concurrent partnership in the last 5 years before the interview
concu1y	Had a concurrent partnership in the year before the interview
nocondom1y	Number of heterosexual and / or same-sex partners without a condom in the year before the interview
totusex1y	Had unprotected heterosexual and / or same-sex relationship in the year before the interview
Sexually Transmitted Infections	
chlamydia5y	Diagnosed with chlamydia in the last 5 years before interview
gonorrhoea5y	Diagnosed with gonorrhoea in the last 5 years before interview
warts5y	Diagnosed with genital warts in the last 5 years before interview
herpes5y	Diagnosed with herpes in the last 5 years before interview
anysti5y	Diagnosed with any STI excluding thrush in the last 5 years before interview
chlamydia1y	Diagnosed with chlamydia in the year before interview
gonorrhoea1y	Diagnosed with gonorrhoea in the year before interview
warts1y	Diagnosed with genital warts in the year before interview
herpes1y	Diagnosed with herpes in the year before interview
anysti1y	Diagnosed with any STI excluding thrush in the year before interview

Table A2: Sex ratio by ethnic group and region

(a) Panel A: Ethnic groups

Variables	**Statistics**	**Sex Ratios 2001**	**Sex Ratios 2011**
		(1)	(2)
White	mean	1.000	0.999
	sd	(0.125)	(0.177)
Mixed	mean	0.788	0.853
	sd	(0.119)	(0.170)
Asian/Asian British	mean	0.983	1.049
	sd	(0.127)	(0.337)
Black/Black British	mean	0.895	0.881
	sd	(0.239)	(0.233)
Chinese/Others	mean	1.044	1.104
	sd	(0.393)	(0.492)

(b) Panel B: Regions

Variables	Statistics	Sex Ratios 2001	Sex Ratios 2011
		(1)	(2)
North East	mean	1.040	1.041
	sd	(0.295)	(0.565)
North West	mean	0.926	0.953
	sd	(0.218)	(0.327)
Yorkshire & The Humber	mean	0.918	0.972
	sd	(0.272)	(0.362)
East Midlands	mean	0.910	0.937
	sd	(0.289)	(0.367)
West Midlands	mean	0.890	0.906
	sd	(0.280)	(0.302)
South West	mean	0.941	0.905
	sd	(0.303)	(0.265)
East	mean	0.907	0.907
	sd	(0.258)	(0.277)
London	mean	0.849	0.885
	sd	(0.231)	(0.328)
South East	mean	0.909	0.882
	sd	(0.274)	(0.235)
Wales	mean	0.942	1.025
	sd	(0.221)	(0.428)

Table B1: Definition description of Sexual Behavior and Sexually Transmitted Infections

Variables	**Description**
Sexual Behaviors	
tot5yrs	No. of hetero and same sex partners in the last 5 years
het5yrs	No. of opposite-sex partners in the last 5 years
mostrecent	No. of most recent sexual partners
Sexually Transmitted Infections	
chlamydia	Ever been diagnosed with chlamydia before interview
gonorrhoea	Ever been diagnosed with gonorrhoea before interview
warts	Ever been diagnosed with genital warts before interview
herpes	Ever been diagnosed with herpes before interview
eversti	Ever been diagnosed with any STI excluding thrush before interview
anysti5ys	Diagnosed with any STI excluding thrush last 5 years before interview
anysti1y	Diagnosed with any STI excluding thrush in 1 year before interview

Table B2: Effects of sexual partner diversification on number. of most recent partners and STI infection

Variable	mostrecent	chlamydia	gonorrhoea	warts	herpes	eversti	anysti5yrs	anysti1yr
	Poisson Regression	Marginal Probit						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
HHI	0.099**	0.024***	0.017***	0.032***	0.006	0.103***	0.015**	0.004
	(0.044)	(0.009)	(0.005)	(0.008)	(0.005)	(0.019)	(0.007)	(0.004)
Observations	13,207	12,461	12,016	12,302	12,458	10,798	12,723	11,274
Pseudo R-squared	-	0.117	0.077	0.052	0.058	0.061	0.145	0.148
FE & Controls	Y	Y	Y	Y	Y	Y	Y	Y

Notes: Dependent variables are sexual behaviors and infection status. Refer to Table B1 for the detailed definitions of the dependent variables. The table reports Poisson regression (in column (1)) and marginal probit (in columns (2)-(8)) estimates. The unit of observation is an individual. Standard errors are adjusted for clustering within regions. Fully controlled regressions include a gender indicator, age, education fixed effect, religion fixed effect, employment fixed effect, marital status fixed effect, region fixed effect, ethnic group fixed effect, the majority groups size in the neighborhood where the respondent lives, and an indicator for whether the respondent is a native. *, ** and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Bibliography

Abramitzky, R., Delavande, A. and Vasconcelos, L.: 2011, Marrying up: the role of sex ratio in assortative matching, *American Economic Journal: Applied Economics* **3**(3), 124–57.

Abrevaya, J.: 2009, Are there missing girls in the United States? Evidence from birth data, *American Economic Journal: Applied Economics* **1**(2), 1–34.

Acemoglu, D., Carvalho, V. M., Ozdaglar, A. and Tahbaz-Salehi, A.: 2012, The network origins of aggregate fluctuations, *Econometrica* **80**(5), 1977–2016.

Acemoglu, D., Makhdoumi, A., Malekian, A. and Ozdaglar, A.: 2017, Privacy-constrained network formation, *Games and Economic Behavior* **105**, 255–275.

Adimora, A. A. and Schoenbach, V. J.: 2005, Social context, sexual networks, and racial disparities in rates of sexually transmitted infections, *Journal of Infectious Diseases* **191**(Supplement 1), S115–S122.

Alesina, A., Baqir, R. and Easterly, W.: 1999, Public goods and ethnic divisions, *The Quarterly Journal of Economics* **114**(4), 1243–1284.

Alesina, A. and Ferrara, E. L.: 2005, Ethnic diversity and economic performance, *Journal of economic literature* **43**(3), 762–800.

Alesina, A. and La Ferrara, E.: 2002, Who trusts others?, *Journal of public economics* **85**(2), 207–234.

Alesina, A. and Zhuravskaya, E.: 2011, Segregation and the quality of government in a cross section of countries, *American Economic Review* **101**(5), 1872–1911.

Altonji, J. G., Elder, T. E. and Taber, C. R.: 2005, Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools, *Journal of political economy* **113**(1), 151–184.

Amuedo-Dorantes, C. and Grossbard, S.: 2007, Cohort-level sex ratio effects on women's labor force participation, *Review of Economics of the Household* **5**(3), 249–278.

Ananat, E. O. and Washington, E.: 2009, Segregation and black political efficacy, *Journal of Public Economics* **93**(5-6), 807–822.

Angrist, J.: 2002, How do sex ratios affect marriage and labor markets? Evidence from America's second generation, *The Quarterly Journal of Economics* **117**(3), 997–1038.

Arpino, B. and Cannas, M.: 2016, Propensity score matching with clustered data. an application to the estimation of the impact of caesarean section on the apgar score, *Statistics in medicine* **35**(12), 2074–2091.

Arpino, B. and Mealli, F.: 2011, The specification of the propensity score in multilevel observational studies, *Computational Statistics & Data Analysis* **55**(4), 1770–1780.

Baerveldt, C., Van Duijn, M. A., Vermeij, L. and Van Hemert, D. A.: 2004, Ethnic boundaries and personal choice. assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils networks, *Social Networks* **26**(1), 55–74.

Ballester, C., Calvó-Armengol, A. and Zenou, Y.: 2006, Who's who in networks. wanted: The key player, *Econometrica* **74**(5), 1403–1417.

Banerjee, A., Chandrasekhar, A. G., Duflo, E. and Jackson, M. O.: 2013, The diffusion of microfinance, *Science* **341**(6144), 1236498.

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

Banerjee, A., Duflo, E., Ghatak, M. and Lafortune, J.: 2013, Marry for what? caste and mate selection in modern india, *American Economic Journal: Microeconomics* **5**(2), 33–72.

Bearman, P. S., Moody, J. and Stovel, K.: 2004, Chains of affection: The structure of adolescent romantic and sexual networks, *American journal of sociology* **110**(1), 44–91.

Becker, G. S.: 1973, A theory of marriage: Part i, *Journal of Political economy* **81**(4), 813–846.

Becker, G. S.: 1974, A theory of marriage: Part ii, *Journal of political Economy* **82**(2, Part 2), S11–S26.

Bertocchi, G. and Dimico, A.: 2019, The long-term determinants of female hiv infection in africa: The slave trade, polygyny, and sexual behavior, *Journal of Development Economics* **140**, 90–105.

Bertrand, M. and Mullainathan, S.: 2004, Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination, *American economic review* **94**(4), 991–1013.

Bhalotra, S. and Clarke, D.: 2018, The twin instrument: fertility and human capital investment, *Journal of the European Economic Association* .

Bhalotra, S. and Clarke, D.: 2019, The twin instrument: Fertility and human capital investment, *Journal of the European Economic Association* .

Bhaskar, V. and Hopkins, E.: 2016, Marriage as a rat race: Noisy premarital investments with assortative matching, *Journal of Political Economy* **124**(4), 992–1045.

Birkett, M., Kuhns, L. M., Latkin, C., Muth, S. and Mustanski, B.: 2015, The sexual networks of racially diverse young men who have sex with men, *Archives of sexual behavior* **44**(7), 1787–1797.

Blattman, C. and Miguel, E.: 2010, Civil war, *Journal of Economic literature* **48**(1), 3–57.

Bloch, F., Jackson, M. O. and Tebaldi, P.: 2019, Centrality measures in networks, *Available at SSRN 2749124* .

Bonacich, P.: 1987, Power and centrality: A family of measures, *American journal of sociology* **92**(5), 1170–1182.

Boucher, V.: 2015, Structural homophily, *International Economic Review* **56**(1), 235–264.

Brainerd, E.: 2007, *Uncounted Costs of World War II: The Effects of Changing Sex Ratios on Marriage and Fertility of Russian Women*, Citeseer.

Bramoullé, Y., Currarini, S., Jackson, M. O., Pin, P. and Rogers, B. W.: 2012, Homophily and long-run integration in social networks, *Journal of Economic Theory* **147**(5), 1754–1786.

Caliendo, M. and Kopeinig, S.: 2008, Some practical guidance for the implementation of propensity score matching, *Journal of economic surveys* **22**(1), 31–72.

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

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

Cameron, A. C., Gelbach, J. B. and Miller, D. L.: 2008, Bootstrap-based improvements for inference with clustered errors, *The Review of Economics and Statistics* **90**(3), 414–427.

Cameron, A. C., Gelbach, J. B. and Miller, D. L.: 2011, Robust inference with multiway clustering, *Journal of Business & Economic Statistics* **29**(2), 238–249.

Cameron, L., Meng, X. and Zhang, D.: 2017, China's sex ratio and crime: Behavioural change or financial necessity?, *The Economic Journal* **129**(618), 790–820.

Cancian, M. and Reed, D.: 1998, Assessing the effects of wives' earnings on family income inequality, *Review of Economics and Statistics* **80**(1), 73–79.

Centola, D.: 2011, An experimental study of homophily in the adoption of health behavior, *Science* **334**(6060), 1269–1272.

Charles, K. K. and Luoh, M. C.: 2010, Male incarceration, the marriage market, and female outcomes, *The Review of Economics and Statistics* **92**(3), 614–627.

Chiappori, P.-A., Fortin, B. and Lacroix, G.: 2002, Marriage market, divorce legislation, and household labor supply, *Journal of political Economy* **110**(1), 37–72.

Clarke, D. and Matta, B.: 2018, Practical considerations for questionable ivs, *The Stata Journal* **18**(3), 663–691.

Collier, P. and Hoeffler, A.: 1998, On economic causes of civil war, *Oxford economic papers* **50**(4), 563–573.

Collier, P. and Hoeffler, A.: 2004, Greed and grievance in civil war, *Oxford economic papers* **56**(4), 563–595.

Conley, T. G., Hansen, C. B. and Rossi, P. E.: 2012, Plausibly exogenous, *Review of Economics and Statistics* **94**(1), 260–272.

Currarini, S., Jackson, M. O. and Pin, P.: 2009, An economic model of friendship: Homophily, minorities, and segregation, *Econometrica* **77**(4), 1003–1045.

Currarini, S., Jackson, M. O. and Pin, P.: 2010, Identifying the roles of choice and chance in network formation: Racial biases in high school friendships, *Proceedings of the National Academy of Sciences* **107**, 4857–4861.

Dagsvik, J. K.: 2000, Aggregation in matching markets, *International Economic Review* **41**(1), 27–58.

Dahl, G. B. and Moretti, E.: 2008, The demand for sons, *The Review of Economic Studies* **75**(4), 1085–1120.

Dehejia, R. H. and Wahba, S.: 2002, Propensity score-matching methods for nonexperimental causal studies, *Review of Economics and statistics* **84**(1), 151–161.

Du, Q. and Wei, S.-J.: 2010, A sexually unbalanced model of current account imbalances, *Technical report*, National Bureau of Economic Research.

Dubuc, S. and Coleman, D.: 2007, An increase in the sex ratio of Births to india-born mothers in England and wales: Evidence for sex-selective abortion, *Population and Development Review* **33**(2), 383–400.

Dupuy, A. and Galichon, A.: 2014, Personality traits and the marriage market, *Journal of Political Economy* **122**(6), 1271–1319.

Easterly, W. and Levine, R.: 1997, Africa's growth tragedy: policies and ethnic divisions, *The quarterly journal of economics* **112**(4), 1203–1250.

Easterly, W., Ritzen, J. and Woolcock, M.: 2006, Social cohesion, institutions, and growth, *Economics & Politics* **18**(2), 103–120.

Edlund, L.: 2001, Dear son-expensive daughter: Do scarce women pay to marry?, *Manuscript, Columbia Univ* .

Edlund, L., Li, H., Yi, J. and Zhang, J.: 2013, Sex ratios and crime: Evidence from China, *Review of Economics and Statistics* **95**(5), 1520–1534.

Eeckhout, J. and Kircher, P.: 2010, Sorting and decentralized price competition, *Econometrica* **78**(2), 539–574.

Eeckhout, J. and Kircher, P.: 2011, Identifying sorting—in theory, *The Review of Economic Studies* **78**(3), 872–906.

Emran, M. S., Robano, V. and Smith, S. C.: 2014, Assessing the frontiers of ultrapoverty reduction: evidence from challenging the frontiers of poverty reduction/targeting the ultra-poor, an innovative program in bangladesh, *Economic Development and Cultural Change* **62**(2), 339–380.

Erens, B., Phelps, A., Clifton, S., Mercer, C. H., Tanton, C., Hussey, D., Sonnenberg, P., Macdowall, W., Field, N., Datta, J. et al.: 2013, Methodology of the third British

National Survey of Sexual Attitudes and Lifestyles (Natsal-3), *Sexually transmitted infections* pp. 84–89.

Erens, B., Phelps, A., Clifton, S., Mercer, C. H., Tanton, C., Hussey, D., Sonnenberg, P., Macdowall, W., Field, N., Datta, J. et al.: 2014, Methodology of the third british national survey of sexual attitudes and lifestyles (natsal-3), *Sex Transm Infect* **90**(2), 84–89.

Esteban, J., Mayoral, L. and Ray, D.: 2012, Ethnicity and conflict: An empirical study, *American Economic Review* **102**(4), 1310–42.

Fair, R. C.: 1978, A theory of extramarital affairs, *Journal of Political Economy* **86**(1), 45–61.

Fearon, J. D. and Laitin, D. D.: 2003, Ethnicity, insurgency, and civil war, *American political science review* **97**(1), 75–90.

Fernandez, R. and Levy, G.: 2008, Diversity and redistribution, *Journal of public economics* **92**(5-6), 925–943.

Fernández, R. and Rogerson, R.: 2001, Sorting and long-run inequality, *The Quarterly Journal of Economics* **116**(4), 1305–1341.

Fisman, R., Iyengar, S. S., Kamenica, E. and Simonson, I.: 2006, Gender differences in mate selection: Evidence from a speed dating experiment, *The Quarterly Journal of Economics* **121**(2), 673–697.

Fisman, R., Iyengar, S. S., Kamenica, E. and Simonson, I.: 2008, Racial preferences in dating, *The Review of Economic Studies* **75**(1), 117–132.

Fong, E. and Isajiw, W. W.: 2000, Determinants of friendship choices in multiethnic society, *Sociological Forum*, Vol. 15, Springer, pp. 249–271.

Fossett, M. A. and Kiecolt, K. J.: 1991, A methodological review of the sex ratio: Alternatives for comparative research, *Journal of Marriage and the Family* pp. 941–957.

Fox, J. T.: 2010, Identification in matching games, *Quantitative Economics* **1**(2), 203–254.

Francis, A. M.: 2011, Sex ratios and the red dragon: using the Chinese Communist Revolution to explore the effect of the sex ratio on women and children in Taiwan, *Journal of Population Economics* **24**(3), 813–837.

Gale, D. and Shapley, L. S.: 1962, College admissions and the stability of marriage, *The American Mathematical Monthly* **69**(1), 9–15.

Galvin, S. R. and Cohen, M. S.: 2004, The role of sexually transmitted diseases in hiv transmission, *Nature Reviews Microbiology* **2**(1), 33.

Golub, B. and Jackson, M. O.: 2010, Using selection bias to explain the observed structure of internet diffusions, *Proceedings of the National Academy of Sciences* **107**(24), 10833–10836.

Golub, B. and Jackson, M. O.: 2012, How homophily affects the speed of learning and best-response dynamics, *The Quarterly Journal of Economics* **127**(3), 1287–1338.

Gonzalez, F. and Miguel, E.: 2015, War and local collective action in sierra leone: A comment on the use of coefficient stability approaches, *Journal of Public Economics* **128**, 30–33.

Greenwood, J., Guner, N., Kocharkov, G. and Santos, C.: 2014, Marry your like: Assortative mating and income inequality, *American Economic Review* **104**(5), 348–53.

Habyarimana, J., Humphreys, M., Posner, D. N. and Weinstein, J. M.: 2007, Why does ethnic diversity undermine public goods provision?, *American Political Science Review* **101**(4), 709–725.

Hagedorn, M., Law, T. H. and Manovskii, I.: 2017, Identifying equilibrium models of labor market sorting, *Econometrica* **85**(1), 29–65.

Halberstam, Y. and Knight, B.: 2016, Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter, *Journal of public economics* **143**, 73–88.

Halperin, D. T. and Epstein, H.: 2004, Concurrent sexual partnerships help to explain africa's high hiv prevalence: implications for prevention, *The Lancet* **364**(9428), 4–6.

Hatfield, J. W. and Milgrom, P. R.: 2005, Matching with contracts, *American Economic Review* **95**(4), 913–935.

Helleringer, S. and Kohler, H.-P.: 2007, Sexual network structure and the spread of hiv in africa: evidence from likoma island, malawi, *Aids* **21**(17), 2323–2332.

Hitsch, G. J., Hortaçsu, A. and Ariely, D.: 2010, Matching and sorting in online dating, *American Economic Review* **100**(1), 130–63.

Jackson, M. O.: 2010, *Social and economic networks*, Princeton university press.

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

Johnson, A. and Mercer, C.: 2017, National survey of sexual attitudes and lifestyles, 2010-2012: Biological data: Secure access.

Johnson, R. C. and Raphael, S.: 2009, The effects of male incarceration dynamics on acquired immune deficiency syndrome infection rates among African American women and men, *The Journal of Law and Economics* **52**(2), 251–293.

Joyner, K. and Kao, G.: 2000, School racial composition and adolescent racial homophily, *Social science quarterly* pp. 810–825.

Kalmijn, M.: 1998, Intermarriage and homogamy: Causes, patterns, trends, *Annual review of sociology* **24**(1), 395–421.

Kandori, M., Mailath, G. J. and Rob, R.: 1993, Learning, mutation, and long run equilibria in games, *Econometrica: Journal of the Econometric Society* pp. 29–56.

Kao, G. and Joyner, K.: 2004, Do race and ethnicity matter among friends? activities among interracial, interethnic, and intraethnic adolescent friends, *Sociological quarterly* **45**(3), 557–573.

Kao, G. and Joyner, K.: 2006, Do hispanic and asian adolescents practice panethnicity in friendship choices?, *Social Science Quarterly* **87**(5), 972–992.

Karimi, F., Génois, M., Wagner, C., Singer, P. and Strohmaier, M.: 2018, Homophily influences ranking of minorities in social networks, *Scientific reports* **8**(1), 1–12.

Kenyon, C. and Colebunders, R.: 2013, Birds of a feather: homophily and sexual network structure in sub-saharan africa, *International journal of STD & AIDS* **24**(3), 211–215.

Koopmans, T. C. and Beckmann, M.: 1957, Assignment problems and the location of economic activities, *Econometrica: journal of the Econometric Society* pp. 53–76.

Krumpal, I.: 2013, Determinants of social desirability bias in sensitive surveys: a literature review, *Quality & Quantity* **47**(4), 2025–2047.

Lafortune, J.: 2013, Making yourself attractive: Pre-marital investments and the returns to education in the marriage market, *American Economic Journal: Applied Economics* **5**(2), 151–78.

Leduc, M. V., Jackson, M. O. and Johari, R.: 2017, Pricing and referrals in diffusion on networks, *Games and Economic Behavior* **104**, 568–594.

Lewbel, A.: 2012, Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models, *Journal of Business & Economic Statistics* **30**(1), 67–80.

Lewbel, A.: 2018, Identification and estimation using heteroscedasticity without instruments: The binary endogenous regressor case, *Economics Letters* **165**, 10–12.

Li, F., Zaslavsky, A. M. and Landrum, M. B.: 2013, Propensity score weighting with multilevel data, *Statistics in medicine* **32**(19), 3373–3387.

Lieberman, E. S. and McClendon, G. H.: 2013, The ethnicity–policy preference link in sub-saharan africa, *Comparative Political Studies* **46**(5), 574–602.

Logan, L. S.: 2013, Status homophily, sexual identity, and lesbian social ties, *Journal of homosexuality* **60**(10), 1494–1519.

Lucas, S.: 2015, Unprotected nation: An update on the financial and economic impacts of restricted contraceptive and sexual health services.

Mah, T. L. and Halperin, D. T.: 2010, Concurrent sexual partnerships and the hiv epidemics in africa: evidence to move forward, *AIDS and Behavior* **14**(1), 11–16.

Mechoulan, S.: 2011, The external effects of black male incarceration on black females, *Journal of Labor Economics* **29**(1), 1–35.

Miguel, E.: 2004, Tribe or nation? nation building and public goods in kenya versus tanzania, *World politics* **56**(3), 327–362.

Miguel, E. and Gugerty, M. K.: 2005, Ethnic diversity, social sanctions, and public goods in kenya, *Journal of public Economics* **89**(11-12), 2325–2368.

Miguel, E., Satyanath, S. and Sergenti, E.: 2004, Economic shocks and civil conflict: An instrumental variables approach, *Journal of political Economy* **112**(4), 725–753.

Montalvo, J. G. and Reynal-Querol, M.: 2005, Ethnic diversity and economic development, *Journal of Development economics* **76**(2), 293–323.

Morris, M. and Kretzschmar, M.: 1997, Concurrent partnerships and the spread of HIV, *Aids* **11**(5), 641–648.

Morris, M. and Kretzschmar, M.: 2000, A microsimulation study of the effect of concurrent partnerships on the spread of hiv in uganda, *Mathematical Population Studies* **8**(2), 109–133.

Nevo, A. and Rosen, A. M.: 2012, Identification with imperfect instruments, *Review of Economics and Statistics* **94**(3), 659–671.

Nick, M. and Walsh, P. R.: 2007, Building the family nest: Premarital investments, marriage markets, and spousal allocations, *The Review of Economic Studies* **74**(2), 507–535.

Okoye, D. and Pongou, R.: 2017, Sea changes: the transatlantic slave trade and missionary activity in africa, *Unpublished manuscript. Retrieved December* **6**, 2017.

Oster, E.: 2017, Unobservable selection and coefficient stability: Theory and evidence, *Journal of Business & Economic Statistics* pp. 1–18.

Oster, E.: 2019, Unobservable selection and coefficient stability: Theory and evidence, *Journal of Business & Economic Statistics* **37**(2), 187–204.

Ozcan, P. and Eisenhardt, K. M.: 2009, Origin of alliance portfolios: Entrepreneurs, network strategies, and firm performance, *Academy of Management Journal* **52**(2), 246–279.

Pagan, A. R. and Hall, A. D.: 1983, Diagnostic tests as residual analysis, *Econometric Reviews* **2**(2), 159–218.

Patacchini, E. and Zenou, Y.: 2012, Ethnic networks and employment outcomes, *Regional Science and Urban Economics* **42**(6), 938–949.

Pongou, R.: 2009, Anonymity and infidelity: Ethnic identity, strategic cross-ethnic sexual network formation, and hiv/aids in africa, *Unpublished paper, Department of Economics, Brown University* .

Pongou, R.: 2010, *The economics of fidelity in network formation*, PhD thesis, Brown University.

Pongou, R.: 2013, Why is infant mortality higher in boys than in girls? A new hypothesis based on preconception environment and evidence from a large sample of twins, *Demography* **50**(2), 421–444.

Pongou, R. and Serrano, R.: 2009, A dynamic theory of fidelity networks with an application to the spread of HIV, *AIDS, Working paper* **2**.

Pongou, R. and Serrano, R.: 2013, Fidelity networks and long-run trends in HIV/AIDS gender gaps, *The American Economic Review* **103**(3), 298–302.

Pongou, R. and Serrano, R.: 2016, Volume of trade and dynamic network formation in two-sided economies, *Journal of Mathematical Economics* **63**, 147–163.

Poterba, J. M.: 1997, Demographic structure and the political economy of public education, *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* **16**(1), 48–66.

Putnam, R. D.: 2007, E pluribus unum: Diversity and community in the twenty-first century the 2006 johan skytte prize lecture, *Scandinavian political studies* **30**(2), 137–174.

Rao, V.: 1993, The rising price of husbands: A hedonic analysis of dowry increases in rural India, *Journal of political Economy* **101**(4), 666–677.

Roth and Vande Vate: 1990, Random paths to stability in two-sided matching, *Econometrica: Journal of the Econometric Society* pp. 1475–1480.

Rothstein, D. S.: 2013, Breastfeeding and children's early cognitive outcomes, *Review of Economics and Statistics* **95**(3), 919–931.

Ruef, M., Aldrich, H. E. and Carter, N. M.: 2003, The structure of founding teams: Homophily, strong ties, and isolation among us entrepreneurs, *American sociological review* pp. 195–222.

Schneider, M., Teske, P., Roch, C. and Marschall, M.: 1997, Networks to nowhere: Segregation and stratification in networks of information about schools, *American Journal of Political Science* pp. 1201–1223.

Schwartz, C. R.: 2010, Earnings inequality and the changing association between spouses' earnings, *American journal of sociology* **115**(5), 1524–1557.

Shimer, R. and Smith, L.: 2000, Assortative matching and search, *Econometrica* **68**(2), 343–369.

Skopek, J., Schulz, F. and Blossfeld, H.-P.: 2011, Who contacts whom? educational homophily in online mate selection, *European Sociological Review* **27**(2), 180–195.

Smith, J. A. and Todd, P. E.: 2005, Does matching overcome lalonde's critique of nonexperimental estimators?, *Journal of econometrics* **125**(1-2), 305–353.

Smith, T. W.: 1992, Discrepancies between men and women in reporting number of sexual partners: A summary from four countries, *Social biology* **39**(3-4), 203–211.

Thoemmes, F. J. and West, S. G.: 2011, The use of propensity scores for nonrandomized designs with clustered data, *Multivariate Behavioral Research* **46**(3), 514–543.

Tourangeau, R. and Smith, T. W.: 1996, Asking sensitive questions: The impact of data collection mode, question format, and question context, *Public opinion quarterly* **60**(2), 275–304.

Tourangeau, R. and Yan, T.: 2007, Sensitive questions in surveys., *Psychological bulletin* **133**(5), 859–883.

Valente, T. W.: 2012, Network interventions, *Science* **337**(6090), 49–53.

Vega-Redondo, F.: 2007, *Complex social networks*, number 44, Cambridge University Press.

Waldron, I.: 1985, What do we know about causes of sex differences in mortality? A review of the literature., *Population Bulletin of the United Nations* (18), 59–76.

Wang, H., Kao, G. and Joyner, K.: 2006, Stability of interracial and intraracial romantic relationships among adolescents, *Social Science Research* **35**(2), 435–453.

Wei, S.-J. and Zhang, X.: 2011, The competitive saving motive: Evidence from rising sex ratios and savings rates in china, *Journal of political Economy* **119**(3), 511–564.

WHO: 2018, Report on global sexually transmitted infection surveillance 2018.

Wimmer, A. and Lewis, K.: 2010, Beyond and below racial homophily: Erg models of a friendship network documented on facebook, *American journal of sociology* **116**(2), 583–642.

Young, H. P.: 2001, *Individual strategy and social structure: An evolutionary theory of institutions*, Princeton University Press.

Zak, P. J. and Knack, S.: 2001, Trust and growth, *The economic journal* **111**(470), 295–321.

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

Zenou, Y.: 2015, A dynamic model of weak and strong ties in the labor market, *Journal of Labor Economics* **33**(4), 891–932.