

# Voting Together: Network Effects in the Turnout Decisions of Naturalized Citizens in U.S. Elections

ECON 21130: FINAL PROJECT

Raymond Han

3/19/18

## Abstract

A large body of evidence suggests that social pressure and social learning play an important role in the turnout decision for citizens of democracies. These network effects may be especially crucial for naturalized citizens, for whom baseline political participation is often lower than for the rest of the population. This paper considers the econometric problem of using recent U.S. Census data on the demographics and turnout decisions of naturalized citizens to lend empirical support to this hypothesis. In particular, I employ an empirical strategy using the size of individuals' local nativity group to infer the extent of individuals' social networks and exploit variation in voting rates across nativity groups to tackle sources of endogeneity in network formation.

## 1 Introduction

An array of experimental studies have documented the fact that the political attitudes and demographic characteristics of the neighbors, friends and families who comprise an individual's social network are an important determinant of individual political participation choices. At the same time, economists have paid increasing attention to the social and economic effects of ethnic networks.

Do ethnic networks affect individual political participation? This paper attempts to shed light on the role of ethnic network effects in immigrant political participation by considering the econometric problem of using recent data from the Current Population Survey of the U.S. Census to study turnout decisions of naturalized citizens in national elections.

Immigrants face difficult challenges of acculturation and assimilation. As a result, immigrants may rely heavily on other more established immigrants to navigate the labor market and secure the basic necessities of life. Indeed, a large proportion of immigrants tend to settle in ethnic enclaves,

areas with dense ethnic networks. The presence of ethnic networks may have an important effect on the rate at which immigrants assimilate. However, while economists have long been concerned with the determinants of immigrant assimilation as measured in wage convergence, less attention has been paid to the determinants of social and cultural outcomes.

The directional effects of ethnic networks on economic and social outcomes are also not immediately clear. In the case of wages for example, on one hand, a dense ethnic network may provide labor market connections and easier access to crucial information necessary to navigate life in the host country. On the other hand, dense networks may preclude their members from broader cultural assimilation and acquisition of host country skills. The net effect of living in an ethnic enclave may therefore be positive or negative. A growing literature has documented the importance of the characteristics of the reference group in determining the net effect of living in enclaves. Edin et al. (2003) use a natural experiment resulting from a Swedish government policy assigning placement of refugees to study the effect that living in ethnic “enclaves” and find that dense ethnic networks improve labor market outcomes. The effect however, is much larger for members of ethnic groups with high-incomes.

This project investigates a similar type of reference group heterogeneity in the context of political participation: Do ethnic networks have a differential effect on individual political participation depending on the political participation of the reference group? In particular, do politically active groups provide a larger network effect than groups which are less politically active?

I hope to contribute to a large literature related to social network effects in political participation. Gerber, Green, and Larimer 2008 devised a seminal get-out-the-vote experiment relying on the importance of social pressure to incentive people to vote by using a series of mailings which informed households about the turnout records of their neighbors. Since then, many experimental studies have shed light on the mechanisms by which social network effects may act. Various studies have provided evidence for partisan signaling (Perez-Truglia and Cruces 2017), avoidance of shame (DellaVigna et al. 2016) and informational network effects (Fafchamps, Vaz, and Vicente 2017) as potential channels. While experimental designs have the advantage of being cleaner to interpret than typical observational studies, the social forces exerted by door-to-door surveyors or randomly distributed flyers are not necessarily indicative of the forces present in daily social life.

In addition to the considerations above, focusing on the voting behavior of naturalized citizens is also driven by two empirical motivations which allow me to consider econometric specifications of the type used by Bertrand, Luttmer, and Mullainathan (2000) to study network effects in welfare usage. First, the study of naturalized citizens allows me to infer the size of an individual’s social network using the co-ethnic group concentration of an individual’s neighborhood. This allows for the inclusion of both group and neighborhood fixed effects, avoiding several major identification challenges associated with any observational study of network effects. Second, there is significant

variation in the rates at which different ethnicities vote. This heterogeneity allows me to control for potential unobserved differences between individuals who select into areas with dense co-ethnic concentration and those who choose to live further from their ethnic group.

To conduct this analysis, I plan to use recent data on the turnout decisions of naturalized citizens from the Current Population Survey of the U.S. Census in addition to population estimates from the American Community Survey. In this project paper, I intend to focus primarily on the econometric challenges imposed by an observational study of network effects. To do so, I will largely rely on simulated data. However, I will nonetheless try to keep the qualities of available data in mind.

## 2 Empirical Strategy

The identification of network effects poses several challenges. If we suppose that one’s network has an effect on an individual’s own outcome, we must be able to measure the presence of network influence. Since data on the type and strength of social pressure one’s network exerts, or the information an individual learns from their contacts is usually unavailable, a proxy must be found.

If we consider that network effects should affect the outcomes of an entire network base, an intuitive choice for network strength proxy is to consider mean characteristic at the neighborhood level. This leads to the simple linear-in-means model of network effects studied by Manski 1993:

$$Pr(Vote_{ij}) = \alpha \overline{Vote_j} + \beta X_i + \epsilon_{ij}$$

where  $i$  indexes individuals and  $j$  indexes areas,  $Vote_{ij}$  is an indicator for whether an individual votes and  $X_i$  is a vector of personal characteristics. Here, the mean voting rate of neighborhood  $j$  serves as a proxy for the “network effect strength” of individual  $i$ ’s network.

The issue with the above is that we cannot be sure whether one’s network increases the probability that individual  $i$  votes or whether omitted characteristics correlated with both  $Pr(Vote_{ij})$  and  $\overline{Vote_j}$  are driving the correlation. These omitted characteristics can be related to either personal characteristics of the group (e.g. individuals who live in wealthy areas tend to be more civically engaged) or to characteristics of the area itself (e.g. some areas have more polling places than others). Manski termed this simultaneity issue the “reflection problem” (1993).

Ideally, we would like to control for both omitted group characteristics and omitted neighborhood characteristics. The empirical framework I employ is adapted from Bertrand, Luttmer and Mullainathan (2000), who study the role the network effects of language groups in welfare usage. As a first step, the authors use language spoken at home as a measure of social links and infer the size of an individual’s social network using the number of people from the same language group in the same area. This strategy exploits variation in the size of social networks across neighborhoods

within a given language group as well as variation in the size of social networks within neighborhoods across the language groups living there. This allows for the inclusion of both neighborhood and ethnic group fixed effects.

As in Borjas 1992, I use ancestry as a measure of ethnicity instead of language since my prior is that voting behavior depends on the political context in which one grows up more than language spoken. Although individuals sharing a given ancestry may be loosely connected general, this concern is alleviated since I only consider first-generation immigrants in the present analysis. I follow Bertrand et al., and refer to the density of immigrants from the same country of origin  $j$  in area  $k$  as individual  $i$  as individual  $i$ 's *contact availability*.

If we suspect that simply having a large social network increases the probability of voting, we arrive at the following specification:

$$Pr(Vote_{ijk}) = \alpha CA_{jk} + \beta X_i + \gamma_j + \delta_k + \epsilon_{ijk}$$

where  $CA_{jk}$  denotes the contact availability in area  $j$  of group  $k$ ,  $\gamma_j$  are neighborhood fixed effects and  $\delta_k$  are group fixed effects. However, a concerning endogeneity remains even in the presence of both sets of fixed effects if individuals who choose to live in ethnic enclaves are differ from those who select away from their co-ethnic group on unobservables. For example, suppose “civic engagement” is an unobserved characteristic which affects the voting decision. If individuals who choose to live in ethnic enclaves are more civically engaged then  $\hat{\alpha}$  will be biased upwards.

To deal with this concern, Bertrand, Luttmer, and Mullainathan 2000 construct a measure of “network strength” as the interaction between contact availability and “contact quality”, defined as the welfare usage of the members of the reference group. Intuitively, network effects should be stronger for members of groups who are more knowledgeable about or have stronger attitudes towards a particular outcome. Crucially, this allows for the inclusion of the lower-order term  $CA_{jk}$  in the regression as a control for unobserved differences correlated with  $CA_{jk}$ .

In the context of voting, I proxy for “contact quality” using the average group voting rate. As Bertrand et al. note, using the average group mean at the neighborhood level may bias the coefficient if individuals share unobserved characteristics with their reference group so I use the average national voting rate instead. Explicitly, I construct my measure of “network strength” as

$$Netw_{jk} = \underbrace{\left( \text{Density of ethnic} \right)_{jk}}_{CA_{jk}} \times \underbrace{\left( \begin{array}{c} \text{Political} \\ \text{participation of} \\ \text{nativity group } k \\ \text{in area } j \end{array} \right)_{jk}}_{\overline{Vote}_k}$$

The basic specification that I consider is then given by

$$Pr(Vote_{ijk}) = \alpha(CA_{jk} * \overline{Vote}_k) + \theta CA_{jk} + \beta X_i + \gamma_j + \delta_k + \epsilon_{ijk}$$

where  $\alpha$  is the coefficient of interest.

## 2.1 A stylized framework

The empirical methodology we have chosen is by no means obvious. In particular, it is not immediately evident that the network effect for individuals from high voting groups should be different for those from low voting groups. In other words, the interaction term requires some theoretical motivation. I outline a simple voting model to show that the coefficient of interest indeed captures the action of network effects and that this coefficient can be identified using a regression specification of the form that has been described. It will hopefully also become clear through this exercise which assumptions are necessary for identification and which can be relaxed.

When do agents decide to vote? Let  $\tilde{R}_i$  be a latent variable capturing individual  $i$ 's return to voting. In particular, we assume

$$\tilde{R}_i = B_i - C_i(e_i) + S_i(e_i) + \epsilon_i \quad , \quad R_i = \begin{cases} 1 & \text{if } \tilde{R}_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where  $B_i$  are benefits to voting unrelated to network effects,  $C_i$  are the costs of voting and  $S_i$  is the benefit derived from social image considerations and  $e_i \in [0, \infty]$  is a measure of ethnic concentration. We leave ambiguous, for the moment, how  $e$  is explicitly defined. An individual votes if the expected return is greater than 0. That is, We parameterize  $B_i$  using a vector of individual characteristics  $X_i$ , a group fixed effect,  $\pi_k^B$  (e.g. internalized group civic culture) and a neighborhood fixed effect  $\gamma_j^B$ . That is,

$$B_i = \beta^B X_i + \delta_k^B + \gamma_j^B.$$

For now, we do not include the instrumental benefit to voting since the probability of being pivotal is so small in national elections that this term is negligible.<sup>1</sup> We also assume that election closeness does not enter into the voting decision in any other way. We explore the effect of election closeness in Section X.

Voters incur significant temporal and logistical costs corresponding to  $C_i$ .  $C_i$  is a function of  $e_i$  since it is easy to imagine that the costs of voting depend on the size and quality of one's social network. For instance, having a politically active network may lower the costs of information acquisition necessary to register and vote. Like in the case of  $B_i$ ,  $C_i$  may also depend on the characteristics of one's group (e.g. political transferability of group civic culture) as well as neighborhood characteristics (e.g. number of polling places) so we again include group and neighborhood fixed effects.

---

<sup>1</sup>The instrumental benefit to voting is given by the expected benefit from an election that is won relative to an election that is lost multiplied by the probability of being pivotal.

Explicitly,

$$-C_i = \beta^C X_i + \nu(e_i * \pi_k) + \nu_0 e_i + \delta_k^C + \gamma_j^C.$$

The interaction term captures the fact that the size of one's network should have a differential effect on voting costs depending on the quality of one's contacts. I will refer to the effect of ethnic concentration on changes in voting cost (captured by  $\nu$  and  $\nu_0$ ) as the *information channel* of network effects on voting.

$S_i$  captures the fact that voting is socially desirable. Agents recognize two types of individuals differing in their level of civic engagement indexed by  $\sigma_i \in \{l, h\}$ . The social component arises since agents use voting as a way to signal their type. Formally, we model  $S_i$  using the basic framework for social image presented in Bénabou and Tirole 2006 and further developed by Bursztyn and Jensen 2017 to write

$$S_i(e_i) = \lambda E_i(\omega_k) \Pr_{-i}[\sigma_i = h | R_i](e_i).$$

Here,  $\Pr_{-i}[\sigma_i = h | R_i](e_i)$  denotes the probability that an individual is perceived by their group as being of type  $h$  conditional on their voting decision  $R_i$ . We assume that  $\Pr_{-i}[\sigma_i = h | R_i](e_i)$  is a linearly increasing function of  $e_i$  so that  $\Pr_{-i}[\sigma_i = h | R_i](e_i) = m e_i$ .<sup>2</sup>  $E_i(\omega_k)$  is the individual's expectation about the social desirability  $\omega_k$  of being perceived by group  $k$  as being of type  $h$ . We specify that  $E_i(\omega_k) = \pi_0 + \pi_k$ , where  $\pi_k$  is a measure of the civic culture in group  $k$  and  $\pi_0$  is a constant.  $\lambda$  is a parameter to be estimated and is a measure of the overall importance of social image. In this construction, variation in social image is introduced through individuals' choice of  $e$  and the strength of civic culture within the group. Note that we impose the restriction that  $\lambda$  does not also vary systematically between individuals. With these assumptions, we have that

$$S_i(e_i) = m\lambda(e_i * \pi_k) + m\lambda\pi_0 e_i.$$

I will refer to  $S_i(e_i)$  as the *social pressure* channel. We let  $N_i(e_i) = S_i(e_i) - C_i(e_i)$  denote the portion of the voting decision affected by network effects, since we do not distinguish between the information channel and the social pressure channel for the moment. Putting these pieces together,

$$\begin{aligned} \tilde{R}_i &= N_i(e_i) + B_i + \epsilon_i \\ &= (m\lambda + \nu)(e_i * \pi_k) + (m\lambda\pi_0 + \nu_0)e_i + \beta X_i + \delta_k + \gamma_j + \epsilon_i \\ &= \phi(e_i * \pi_k) + \theta e_i + \beta X_i + \delta_k + \gamma_j + \epsilon_i \end{aligned}$$

where the last line follows from the specification that  $\phi := m\lambda + \nu$  and  $\theta := m\lambda\pi_0 + \nu_0$  and where unsuperscripted coefficients  $\beta, \delta$  and  $\gamma$  are the sum of the coefficients superscripted by  $B$  and  $C$ .

---

<sup>2</sup>The specified relationship between  $\Pr_{-i}[\sigma_i = h | R_i](e_i)$  and the *absolute* size of an individual's network will of course depend on how our measure  $e$  is defined.

We see that this simple model recommends the regression model presented in Section X in a binary choice framework. The coefficients  $\phi$  and  $\theta$  capture the effect of networks on the voting decision. However, I cannot distinguish between the social pressure and the information channels in this baseline specification since both  $\phi$  and  $\theta$  are determined by the social pressure as well as the information channels. Nonetheless, we might look to both coefficients for evidence of network effects given only the model we have outlined. Concretely,  $\theta$  captures the effect of ethnic concentration on voting in the absence of civic culture (i.e. when  $\pi_k = 0$ ) while  $\phi$  provides a measure of how the strength of network effects change as civic culture  $\pi_k$  and network size increase.

The lower-order  $e_i$  term from the regression might be removed if it were suspected that there are no informational or social pressure effects when  $\pi_k = 0$ , i.e. when other members of one's group have no interest in civic participation. However, considering ethnic concentration as a *choice* variable suggests at least one other good reason for including the  $e_i$  term and also a rationale for focusing on  $\phi$  as the coefficient of interest.<sup>3</sup> In particular, a serious endogeneity concern arises if we imagine that ethnic concentration is not randomly assigned to individuals and immigrants consider ethnic concentration as an important factor when choosing an emigration destination. Our worry is that unobserved characteristics underlying an individual's choice of  $e_i$  may affect the voting decision through causal channels unrelated to one's network. For example, suppose individuals with a greater sense of civic duty are more likely to live in areas with higher ethnic concentration. A relationship between ethnic concentration and voting may simply reflect the greater civic commitment of individuals living in ethnic enclaves.<sup>4</sup> Including  $e_i$  in our regression serves as a way to control for the unobserved differences between those who choose to live in ethnic enclaves and those who don't. But if we interpret  $e_i$  as a control for the non-network effects of unobserved differences,  $\theta$  cannot be interpreted as a causal network effect.  $\phi$  is therefore treated as the coefficient of interest.

## 2.2 Measuring ethnic concentration

It is not immediately clear how to define a measure of ethnic concentration. Considering the raw number of ethnic group members in one's neighborhood will tend to assign systematically low values of  $e$  to those living in more sparsely populated areas. Defining ethnic concentration as the share of the local population which is co-ethnic alleviates the previous issue but will still assign systematically low values of  $e$  to those who belong to smaller ethnic groups. An appropriate choice of  $e$  must also correctly specify a linear interaction term with  $\pi$ .

---

<sup>3</sup>In addition to the fact that dropping one of the constituent terms of an interaction we would like to interpret causally is itself hard to justify.

<sup>4</sup>A simple extension of the voting model which endogenizes  $e_i$  by following Edin, Fredriksson, and Aslund 2003 is presented in Section 3.2.

As a first step, I follow Bertrand et al. and construct  $e_i$  as the share of the population within area  $j$  that is co-ethnic, divided by the share of the U.S. population that is co-ethnic. I then take the log of this measure. Explicitly,

$$e_i = CA_{jk} = \log \left( \frac{C_{jk}/A_j}{L_k/T} \right) \quad (2)$$

where  $C_{jk}$  is the number of people in area  $j$  who belong to ethnic group  $k$ ,  $A_j$  denotes the total population in area  $j$ ,  $L_k$  is the total number of people in the U.S. in ethnic group  $k$  and  $T$  is the total population of the U.S. This choice of  $CA_{jk}$  ensures that small ethnic groups are not systematically associated with smaller  $e_i$ . Further, if the distribution of co-ethnic members is uniform within group  $k$ , the quantity  $CA_{jk} = 0$  for all areas  $j$ .

### 3 Identification Concerns

I now consider several additional threats to identification.

#### 3.1 Proxying for political activism

A further consideration arises when we consider the related question of what should be used as a proxy for  $\pi_k$ , the level of “political activity” of group  $k$ . Following Bertrand et al. in considering the group mean at the national level, we might choose to use the national voting rate of group  $k$ . The frequency with which members of a given ethnic group turnout in elections would serve as a fine proxy were it not for the fact that the national voting rate already reflects the action of network effects. This complicates the interpretation of an OLS model which uses  $\overline{Vote}_k$  in the interaction term. To see this, note that

$$\overline{Vote}_k = \phi(\bar{e}_k * \pi_k) + \theta \bar{e}_k + \underbrace{\beta \bar{X}_k + \delta_k}_{:=\pi_k}$$

where we designate  $\pi_k$  as the component of  $\overline{Vote}_k$  unaffected by network effects. This is in some sense what the national voting rate would be in the absence of network effects. Then, if

$$\begin{aligned} Vote_i &= \hat{\phi}(e_i * \overline{Vote}_k) + \hat{\theta}e_i + \hat{\beta}X_i + \hat{\delta}_k + \hat{\gamma}_j + \hat{\epsilon}_i \\ &= \hat{\phi}(e_i * (\phi(\bar{e}_k * \pi_k) + \theta \bar{e}_k + \pi_k)) + \hat{\theta}e_i + \hat{\beta}X_i + \hat{\delta}_k + \hat{\gamma}_j + \hat{\epsilon}_i \\ &= (\hat{\phi} + \hat{\phi}\phi\bar{e}_k)(e_i * \pi_k) + \hat{\phi}\theta(e_i * \bar{e}_k) + \hat{\theta}e_i + \hat{\beta}X_i + \hat{\delta}_k + \hat{\gamma}_j + \hat{\epsilon}_i. \end{aligned} \quad (3)$$

the coefficient on  $(e_i * \pi_k)$  does not equal  $\hat{\phi}$  as desired. However, the above suggests that if  $\bar{e}_k = \bar{e}$  for all groups  $k$ , that is, all ethnic groups exhibit the same level of average ethnic concentration  $e$ ,



we ought to be able to make a simple correction to recover the true parameter  $\phi$ :

$$\phi = \frac{\hat{\phi}}{1 - \hat{\phi}\bar{e}}. \quad (4)$$

In this case we can run the regression above, making sure to correct the coefficient  $\hat{\phi}$  before interpreting the magnitude of network effects. If this is not the case, however, the OLS estimate  $\hat{\phi}$  may be biased.

Will this really work? To examine the issue, I simulate some data.<sup>5</sup> In particular, I assign 10,000 individuals to 50 ethnic groups (with the size of the groups drawn from an exponential distribution with rate  $\lambda = 0.5$ .) Individuals within ethnic groups are assigned to 50 different neighborhoods with likelihoods drawn again from an exponential distribution with rate  $\lambda = 0.5$ ). However, each individual within an ethnic group (from the second individual onwards) also has a chance  $\rho$  of being directly assigned to the same neighborhood as the previous co-ethnic individual in order to develop heterogeneity in  $e$  within ethnic groups. Ethnic concentration is defined as in Eq. (2). Parameters are shown in Table 1.

Table 1: Simulation Parameters

$N$	$K$	$J$	Iterations	$\phi/\beta$	$\theta/\beta$	$\rho$	$X$	$\delta_k$	$\gamma_j$	$\epsilon$
10000	50	50	20	10	1	0.95	$\sim N(0, 1)$	$\sim N(0, 0.2)$	$\sim N(0, 0.2)$	$\sim N(0, 1)$

For now, I estimate the model using OLS instead of a binary choice model due to the number of fixed effects. I therefore consider the ratio of the estimated coefficient  $\hat{\phi}$  and the estimated coefficient  $\hat{\beta}$  (since the binary choice framework grants us a choice of normalization). Estimating the model given by Eq. (3) then yields  $\hat{\phi} = 1.166$  (s.d.) = 0.125 while  $\hat{\beta} = 0.203$  (s.d.) = 0.011. The ratio of the coefficients is then  $\left(\frac{\hat{\phi}}{\hat{\beta}}\right) = 5.74$ . Note that this is significantly less than the underlying value of 10. To alleviate this problem, we can perform the correction given by Eq. (4). Doing so yields  $\hat{\phi}^{adj} = 1.977$  which recovers the ratio  $\left(\frac{\hat{\phi}}{\hat{\beta}}\right)^{adj} = 9.74$ .

However, suppose that some ethnic groups are more clustered than others, i.e. they have higher  $\rho$ . It is especially problematic if  $\rho$  is a function of  $\pi_k$ , that is, if groups which are more politically active are also more (or less) clustered. We simulate this dependency by specifying that  $\rho_k = \max(0.95 + 0.25 * \pi_k, 0.99)$ . Since  $E[\pi_k] = E[\beta\bar{X}_k + \delta_k] = 0$ , we maintain  $\overline{rho_k} \approx 0.95$  as before. Leaving the other parameters untouched,  $\hat{\phi} = 0.518$  (s.d. = 0.109) and  $\hat{\beta} = 0.163$  (s.d. = 0.018). The ratio of the coefficients is then given by  $\left(\frac{\hat{\phi}}{\hat{\beta}}\right) = 3.18$ . The same correction as before yields

<sup>5</sup>The associated code is presented in the accompanying R markdown file.

$\hat{\phi}^{adj} = 0.740$  so that  $\left(\frac{\hat{\phi}}{\beta}\right)^{adj} = 4.54$ , a significant underestimate, suggesting that the OLS estimate is biased.

### 3.2 Differential Selection

A second set of identification concerns arises if we consider the effect of endogenizing immigrants' choice of ethnic concentration,  $e_i$ . Following Edin, Fredriksson, and Aslund 2003, suppose immigrants derive utility from living with members of their own ethnic group and consumption. They maximize utility by making a location choice. We assume locations are characterized by a measure of ethnic concentration ( $e$ ). For simplicity, we assume there is a continuum of location choice such that  $e \in [0, \infty]$ . Utility is given by

$$U_i = u_i(e_i) + \ln y_i.$$

Individuals maximize utility subject to a labor market opportunity locus which is a function of ethnic concentration. That is, individuals solve

$$\begin{aligned} \max_{e_i} & u_i(e_i) + \ln y_i \\ \text{s.t.} & \ln y_i = z_i + b_i e_i. \end{aligned}$$

The FOC is given by  $u_i(e_i)' + b_i = 0$ . Specifying that  $u_i = a_i e - \frac{k}{2} e^2$ , the optimal (interior)  $e$  is then given by

$$e_i^* = (a_i + b_i)/k.$$

Here  $a_i$  denotes the importance the individual attaches to  $e$  (inherently) and  $b_i$  is individual  $i$ 's marginal return in consumption to  $e$ . Let  $Z_i = a_i + b_i$ . The straightforward case is that  $Z_i$  has a direct effect on the voting decision.. Since  $Z_i$  is unobserved so we cannot control for it directly. On the other hand,  $e_i^*$  is observed and since  $Z_i = k e_i^*$ , we can control for the direct effect of  $Z_i$  on the voting decision by including  $e_i^*$  as a "control". This is what is done in the main specification.

However, a potential endogeneity issue arises if we suppose it is not precisely the sum  $a_i + b_i$  which has a direct effect on the voting decision. Suppose the underlying model is given by (dropping fixed effects and individual covariates for simplicity):

$$\tilde{R}_i = \phi(e_i * \pi_k) + \theta_1 A_i + \theta_2 B_i + \epsilon_i$$

which is equivalent to

$$\tilde{R}_i = \phi(e_i * \pi_k) + \alpha A_i + \theta e_i + \epsilon_i$$

where  $\alpha = \theta_1 - \theta_2$ . If we then consider the specification

$$\tilde{R}_i = \phi(e_i * \pi_k) + \theta e_i + \epsilon_i,$$

it is clear that our estimation will suffer from omitted variable bias. In particular,

$$\begin{aligned} E[\hat{\beta}] &= \beta + (X'X)^{-1}E[X'Z|X]\alpha \\ &= \beta + \begin{pmatrix} E[e_i^2] & E[e_i(e_i * \pi_k)] \\ E[e_i(e_i * \pi_k)] & E[(e_i * \pi_k)^2] \end{pmatrix}^{-1} E\left[\begin{pmatrix} \sum_{i=1}^N e_i A_i \\ \sum_{i=1}^N (e_i * \pi_k) A_i \end{pmatrix}\right] \alpha \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \end{aligned}$$

Letting  $\gamma = E[e_i^2]E[(e_i * \pi_k)^2] - E[e_i(e_i * \pi_k)]^2$ ,

$$= \beta + \frac{1}{\gamma} \begin{pmatrix} E[(e_i * \pi_k)^2] & -E[e_i(e_i * \pi_k)] \\ -E[e_i(e_i * \pi_k)] & E[e_i^2] \end{pmatrix} E\left[\begin{pmatrix} \sum_{i=1}^N e_i A_i \\ \sum_{i=1}^N (e_i * \pi_k) A_i \end{pmatrix}\right] \alpha$$

If we further assume  $E[e_i] = 0$  and  $E[e_i * \pi_k] = 0$ ,

$$\begin{aligned} &= \beta + \frac{1}{\gamma} \begin{pmatrix} \text{Var}(e * \pi) & -\text{Cov}(e, e * \pi) \\ -\text{Cov}(e, e * \pi) & \text{Var}(e) \end{pmatrix} \begin{pmatrix} \text{Cov}(e, A) \\ \text{Cov}(e * \pi, A) \end{pmatrix} \alpha \\ &\Rightarrow \hat{\phi} = \phi + \underbrace{\frac{\alpha}{\gamma} (\text{Var}(e)\text{Cov}(e * \pi, A) - \text{Cov}(e, e * \pi)\text{Cov}(e, A))}_{\text{bias}} \end{aligned}$$

We cannot in general guarantee that this bias term is 0, i.e. that there does not exist a determinant of immigrants' choice of ethnic concentration that is differentially predictive of  $e$  for groups with high  $\pi$ . If such a determinant also has an independent effect on the voting decision, the misspecification in our model may introduce a source of bias. This is a problem of *differential selection*-individuals who choose to live in politically active enclaves may differ on unobservables from individuals who choose to live in politically inactive enclaves.

The problem is easy enough to simulate. Concretely, suppose a sense of “civic duty” has an independent effect on the voting decision and individuals with high “civic duty” seek out other individuals of the same type. Then high “civic duty” individuals in groups with high civic culture may be more likely to live in enclaves than high “civic duty” individuals in groups with low civic culture.<sup>6</sup> I consider the following underlying model:

$$\tilde{R}_i = \phi(e_i * \pi_k) + \theta_1 A_i + \theta_2 B_i + \delta_k + \gamma_j + \epsilon_i$$

where  $e_i = \alpha \pi_k * A_i + B_i$ . The coefficient  $\alpha$  captures the fact that high civic duty individuals from groups with high civic culture may be more likely to live in enclaves. Then, suppose we estimate:

$$\tilde{R}_i = \phi(e_i * \pi_k) + \alpha A_i + \theta e_i + \epsilon_i$$

---

<sup>6</sup>While such a model is possible to envision, evidence suggests that in actuality, immigrants' choice of location decision has less to do with concerns about political participation and more to do with the urgent issues of finding a job, access to language, etc. However, this does not preclude the presence of other kinds of differential selection.

Then  $\phi$  will be biased. To show this, I simulate using the latent variable rather than the voting dummy to avoid noise and highlight the presence of bias. General simulation parameters are the same as in Table 1—the new parameters shown in Table 2.

Table 2: Simulation Parameters									
$N$	$K$	$J$	Iterations	$\phi$	$\theta$	$\beta$	$\theta_1$	$\theta_2$	$\alpha$
10000	50	50	20	10	1	1	2	1	1

The result of this exercise is that both  $\hat{\phi}$  and  $\hat{\theta}$  are biased. Estimation yields  $\hat{\phi} = 11.70$  while  $\hat{\theta} = 0.968$ .

Bertrand et al. suggest two methods of attempting to deal with differential selection. The first is to consider estimating the model using a larger geographical unit. In their paper, while most of the analysis is conducted at the PUMA level, they also consider specifications where neighborhoods are defined at the larger MSA level. Their estimates showed that the difference in effect size under the two different geographical schemes is small enough such that differential selection *between* MSAs would have to be greater than differential selection *within* MSAs. Since intuitively it should be the case that selection between MSAs should be much more costly, the authors interpret this as evidence that differential selection is not driving the results.

Second, making the underlying assumption that selection on unobservables might reflect patterns of selection on observables to some extent, the authors also look for evidence of differential selection on observables. In particular, the authors investigate the effect of dropping important controls, such as education dummies. They find that their point estimates are robust to dropping these controls, evidence that the results are largely unaffected by differential selection on observables.

## 4 Further Work

Moving forward, I plan to test the specification using a different proxy for group political participation, i.e. the quality of democratic institutions in the country of origin. This will serve as an important robustness check.

There are also many other specification and robustness checks to perform. These include

- Changing the functional form of  $CA_{jk}$
- Estimating binary choice models instead of linear probability (e.g. logit, probit)
- Testing for differential selection by dropping controls

- Testing that the interaction term is truly linear (e.g. by binning the specification)
- Considering inference (how to cluster standard errors)

I also hope to understand how the results change when we define groups differently. For example, do the network effects get stronger if we define groups at the ethnicity  $\times$  occupation level?

## References

- Bénabou, Roland and Jean Tirole (2006). “Incentives and Prosocial Behavior”. In: *American Economic Review* 96.5, pp. 1652–1678. ISSN: 0002-8282. DOI: 10.1257/aer.96.5.1652.
- Bertrand, Marianne, Erzo F. P. Luttmer, and Sendhil Mullainathan (2000). “Network Effects and Welfare Cultures\*”. In: *Quarterly Journal of Economics* 115.3, pp. 1019–1055. ISSN: 0033-5533. DOI: 10.1162/003355300554971.
- Borjas, G. J. (1992). “Ethnic Capital and Intergenerational Mobility”. In: *The Quarterly Journal of Economics* 107.1, pp. 123–150. ISSN: 0033-5533. DOI: 10.2307/2118325.
- Bursztyn, Leonardo and Robert Jensen (2017). “Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure”. In: *Annual Review of Economics* 9.1, pp. 131–153. ISSN: 1941-1383. DOI: 10.1146/annurev-economics-063016-103625.
- DellaVigna, Stefano et al. (2016). “Voting to Tell Others \*”. In: *NBER Working Paper No. w19832*.
- Edin, P.-A., P. Fredriksson, and O. Aslund (2003). “Ethnic Enclaves and the Economic Success of Immigrants—Evidence from a Natural Experiment”. In: *The Quarterly Journal of Economics* 118.1, pp. 329–357. ISSN: 0033-5533. DOI: 10.1162/00335530360535225.
- Fafchamps, Marcel, Ana Vaz, and Pedro C Vicente (2017). “Voting and Peer Effects: Experimental Evidence from Mozambique”. In: *Working Paper*.
- Gerber, Alan S, Donald P Green, and Christopher W Larimer (2008). “Social Pressure and Voter Turnout: Evidence from a Large-scale Field Experiment”. In: *American Political Science Review* 102.1. DOI: 10.1017/S000305540808009X.
- Manski, Charles F. (1993). “Identification of Endogenous Social Effects: The Reflection Problem”. In: *The Review of Economic Studies* 60.3, p. 531. ISSN: 00346527. DOI: 10.2307/2298123.
- Perez-Truglia, Ricardo and Guillermo Cruces (2017). “Partisan Interactions: Evidence from a Field Experiment in the United States”. In: *Journal of Political Economy* 125.4, pp. 1208–1243. ISSN: 0022-3808. DOI: 10.1086/692711.