Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June 2018: 1–14

DOI: 10.1177/2053951718761141
journals.sagepub.com/home/bds
SAGE

Predicting ethnicity with first names in online social media networks

Bas Hofstra[1] and Niek C de Schipper[2]

Abstract

Social scientists increasingly use (big) social media data to illuminate long-standing substantive questions in social science research. However, a key challenge of analyzing such data is their lower level of individual detail compared to highly detailed survey data. This limits the scope of substantive questions that can be addressed with these data. In this study, we provide a method to upgrade individual detail in terms of ethnicity in data gathered from social media via the use of register data. Our research aim is twofold: first, we predict the most likely value of ethnicity, given one's first name, and second, we show how one can test hypotheses with the predicted values for ethnicity as an independent variable while simultaneously accounting for the uncertainty in these predictions. We apply our method to social network data collected from Facebook. We illustrate our approach and provide an example of hypothesis testing using our procedure, i.e., estimating the relation between predicted network ethnic homogeneity on Facebook and trust in institutions. In a comparison of our method with two other methods, we find that our method provides the most conservative tests of hypotheses. We discuss the promise of our approach and pinpoint future research directions.

Keywords

Social networks, Facebook, ethnicity, first names, prediction, Big Data

Introduction

Research on social media is rapidly expanding in the social sciences. Studies on social media (e.g., Bond et al., 2012; boyd and Ellison, 2008; Ellison et al., 2007; Lewis et al., 2008) are among the most highly cited articles in the social sciences. An increasing number of scientific journals cover social media, Big Data, and their relationship to society (e.g., *Big Data & Society, Social Media + Society, Journal of Computer-Mediated Communication, New Media & Society*). Even in science's most prestigious outlets—e.g., *Science, Nature*, and *PNAS*—there are a number of studies using Big Data from social media (e.g., Bakshy et al., 2015; Bond et al., 2012; Hobbs et al., 2016; Kramer et al., 2014).

This increased scholarly interest in social media is no surprise, as social media data provides scholars with novel ways to analyze human social interactions. On social media, for instance, individuals have new ways to communicate, to spread information, and to coordinate collective action (cf. Corten, 2012; see Gonzàlez-Bailòn and Wang, 2016). Furthermore, individuals increasingly use these platforms to maintain their interpersonal social relationships (Ellison et al., 2011). In addition, computational approaches to social science make it relatively easy to collect data on online interactions, such as those documented on Facebook or Twitter, because these interactions generate digital time-stamped footprints of large social networks (Golder and Macy, 2014).

It has been argued that the networked footprints these platforms automatically archive as part of their daily operations have revolutionized social science (Lazer et al., 2009; Spiro, 2016; Watts, 2011). These features of social media data make it relatively

[1]Stanford University, Graduate School of Education, Stanford, CA, USA
[2]Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

Corresponding author:
Bas Hofstra, Stanford University, Graduate School of Education, 520 Galvez Mall, Stanford, CA 94305, USA.
Email: bhofstra@stanford.edu

straightforward to study larger networks that reach beyond the small number of core ties (e.g., fewer than five network contacts) and the closed social contexts (e.g., classrooms or departments) usually under consideration in the study of social relationships, i.e., in social network analyses (cf. Hofstra et al., 2017; see, e.g., Hofstra et al., 2015a; Kalmijn and Van Tubergen, 2006; Marsden, 1987; Smith et al., 2014). These types of data are occasionally labeled "big," as they often contain information about the online behavior of millions of users (McFarland and McFarland, 2016). However, Big Data obtained from social media has yet to reach its full potential regarding its use in social science research.

Analyzing such online social media data comes with major challenges. One of the core challenges is that the level of individual detail in data gathered by researchers from social media is often considerably lower when compared to information gathered in survey research (Golder and Macy, 2014; Spiro, 2016; Stopczynski et al., 2014). For instance, key individual characteristics such as gender, age, ethnicity, education, or occupation are often either missing or even misreported by respondents in Big Data gathered by researchers online (Golder and Macy, 2014; Spiro, 2016). Individual privacy considerations may be one reason why social media data are often *broad but shallow*, as people might close their social media profiles to safeguard their personal information (e.g., see Hofstra et al., 2016). Therefore, these data may be *big*, but the level of detail is thin. This low level of detail limits the scope of the substantive questions that can be addressed when studying data obtained from social media platforms. Ethnicity and gender, for instance, are key individual characteristics by which social network analysts often study the patterns in their data (e.g., Hofstra et al., 2017; Mayer and Puller, 2008; McPherson et al., 2006; Smith et al., 2014; Van Tubergen, 2015; Wimmer and Lewis, 2010).

Hence, a key question in the growing field of analyzing online social network data is how one can use the plethora of opportunities of these data while at the same time maintaining (at least some of) the "richness" that is usually found in survey data. Effectively dealing with this issue increases the number of substantive questions one is able to answer. These considerations motivate the two aims of this study:

1. *We propose a procedure to upgrade the level of individual detail in online social network data by predicting the most likely value of ethnicity, given one's first name using register data.*
2. *We propose a procedure on how to test hypotheses using these "upgraded" social media data.*

It is well established that names are a clear signal of ethnicity. There are profound differences in how people from different ethnic backgrounds name their children (Lieberson, 2000; see, e.g., Bloothooft and Onland, 2011; Chang et al., 2010; Coldman et al., 1998; Fiscella and Fremont, 2006; Lauderdale and Kerstenbaum, 2000; Mateos et al., 2011). Scholars seem to increasingly use this empirical regularity to enrich social media data (see Cesare et al., 2017 for a recent overview). Logically, because (first) names are often among the only indicators researchers have about individuals in social media data. We follow this burgeoning line of research and thus make use of names as a signal of ethnicity.

We make two key contributions to this growing field. First, we extend prior work because we consider the possibility that those who carry the same first name can each have a different ethnicity. There are two studies that are most related to our procedure, that of Chang et al. (2010), who use a probabilistic Bayesian approach, and that of Hofstra et al. (2017), who use a supervised learning approach. Chang et al. (2010) and Hofstra et al. (2017) assign the most likely value of ethnicity to people on Facebook, given their surnames (Chang et al., Hofstra et al.) and first names (Hofstra et al.). While both studies aim to validate their ethnicity predictions using a source of ground truth (Chang et al.: MySpace data; Hofstra et al.: survey data), they do not model the possibility of different ethnicities among people carrying the same names. Or, as Chang et al. (2010: 25) put it: "we [. . .] have not yet theoretically modeled error throughout our calculations." We statistically take this uncertainty into account for a more realistic representation of the relationship between ethnicity and (first) names. To show the promise of our approach, we directly compare the Hofstra et al. (2017) method with the method described in this study and show which method is the least prone to false-positive results. Reducing such false-positives is crucial in order to obtain more meaningful hypotheses tests and correct statistical inference. By testing whether our method is less prone to false-positive results than other methods, we may show the importance of future implementation of our work in substantive empirical studies that aim to make knowledge claims using online social media data.

Second, we show how to use the predicted values of ethnicity among Facebook networks in standard regression models to tests hypotheses. More specifically, we show how to test hypotheses with the predicted variable as an independent variable while simultaneously accounting for the uncertainty in the predicted values of this new variable.[1] To show the promise of our approach, we provide a toy example. In recent years, there has been a sharp increase in studies

investigating the claim that ethnic diversity has detrimental effects on trust and social cohesion (Abascal and Baldassarri, 2015; Putnam, 2000; Van der Meer and Tolsma, 2014). As societies increasingly grow ethnically diverse, it is crucial to understand whether and how ethnic diversity across different contexts—e.g., in cities, neighborhoods, workplaces, and among social contacts—affects trust. As an example, we explore and engage some of the claims regarding the relationship between ethnic diversity and social cohesion. Specifically, we consider ethnic homogeneity in Facebook networks and investigate its relationship to trust in institutions. Note that we do not aim to test theoretically derived hypotheses. In tying our method to this substantive example, we merely push future work to theoretically consider the consequences of the (ethnic) composition of online (Facebook) networks. Apart from ethnic diversity online and trust, ethnic diversity relates to a myriad of socially relevant issues. One classic argument is that (ethnic) diversity in social networks facilitates the free flow of ideas through networks (e.g., Granovetter, 1973, 1983). Another argument is that a higher level of ethnic diversity in networks may reduce interethnic prejudice (Allport, 1954; Pettigrew and Tropp, 2006). Furthermore, the coevolution of ethnic homophily online and social influence may cause "echo chambers" (Halberstam and Knight, 2016) and intergroup polarization of attitudes and opinions (Mäs and Flache, 2013). Here, we provide a statistically plausible method to test hypotheses with predicted individual characteristics (i.e., ethnicity) in online Big Data as independent variables. As an engaging starting point in doing so, we consider trust in institutions, while the other abovementioned examples are similarly relevant.

Therefore, we contribute to the growing field of analyzing online Big Data by showing how to enrich and make innovative use of such data to test hypotheses in a novel way. We aspire to make the description of the procedure as accessible as possible so that the applied empirical scientist can adopt the method with relative ease using free and open source software.

To illustrate our approach, we use three data sources. However, the approach that we describe is general in nature and is not limited to the data sources that we use. At the end of this article, we propose some alternative, but similar data sources one may use to implement our approach. Here, we use (1) survey data from a large, diverse sample of adolescents; (2) online social network data containing more than a million network members downloaded from the Facebook pages of these same adolescents; and (3) register data that capture the frequency of first names and the proportion of the name carriers and their parents who have been born in specific countries. The specific data sets we use are: (1) survey data from the "Children of Immigrants Longitudinal Survey in Four European Countries" (CILS4EU; Kalter et al., 2016) and the "Children of Immigrants Longitudinal Survey in the Netherlands" (CILSNL; Jaspers and Van Tubergen, 2017), (2) Facebook data from the "Dutch Facebook Survey" (Hofstra et al., 2015b), and (3) register data of the Dutch Civil Registration (DCR) of 2010 (Bloothooft and Schraagen, 2011).

The concept of ethnicity in the Dutch context

Some define ethnicity based on an individual's self-identification (e.g., Verkuyten and Kwa, 1994), such that is up to individuals to decide whether they "identify as a member of group X." Others use objective measures of ethnicity, occasionally based on parents' birth countries (e.g., Stark and Flache, 2012; Vermeij et al., 2009). Here, for simplicity, we use the regularly applied definition by Statistics Netherland, which is standard practice in research on ethnicity and social networks in the Netherlands (e.g., Statistics Netherlands, 2012; Vermeij et al., 2009). This means that we classify individuals' ethnicity by their biological parents' birth country (cf. Stark and Flache, 2012; Vermeij et al., 2009). When individuals have one parent born in the Netherlands, we classify them as belonging to the ethnicity of the parent not born in the Netherlands, and when they have parents born in two different non-Dutch countries, we classify them in the birth country of the mother.

Because the data we use is from a sample of adolescent respondents in the Netherlands, it is informative to define the ethnic groups most salient in Dutch society. Dutch adolescents can be classified among six large ethnic origin groups (Castles et al., 2013). The first group comprises adolescents whose parents were born in the Netherlands and who are members of the Dutch majority. The second and third groups consist of children of immigrants from Turkey and Morocco. These children's parents originated from the low-educated labor force that the Netherlands recruited in the 1950s and 1960s from Turkey and Morocco or from parents who arrived more recently (e.g., because of family reunions). These two groups constitute the largest minority group in the Netherlands. Another group originates from postcolonial countries in the Dutch Caribbean (e.g., Aruba and Suriname). A fifth group originates from other Western countries (e.g., neighboring countries such as Germany), and a sixth group originates from other non-Western countries (e.g., conflict areas such as Afghanistan). These ethnic groups are rather similar across Western European countries in

the type of immigrants that settled there, despite varying specific countries of origin of the ethnic groups that are present (Smith et al., 2014).

Ethnic background is thus occasionally defined by classifying individuals into one of these six large ethnic background groups in the Netherlands: Dutch, Turkish, Moroccan, Dutch Caribbean, other Western backgrounds, and other non-Western backgrounds. We also use this categorization throughout this study.

Data sources

CILS4EU and CILSNL

We use the third, fourth, and fifth waves of the CILS4EU and CILSNL on adolescents in the Netherlands (Jaspers and Van Tubergen, 2017; Kalter et al., 2016).[2,3] In the CILS4EU, adolescents were followed for three consecutive years (2010–2013). Data were collected in the Netherlands, Sweden, Germany, and England. The CILSNL followed the Dutch panel of the CILS4EU for an additional four years (2014–2017). We analyze the Dutch part of the data because our variables of interest are only included in the Dutch section. We analyze waves 3, 4, and 5 because these waves are the anchor of and the closest in time to the Facebook data. Essentially, the Facebook data is collected via respondents' survey answers in waves 3 and 4 (elaborated below). All surveys include detailed information on respondents' background, attitudes, and leisure time activities.

The wave 1 sample was stratified by the proportion of immigrants of non-Western origin within a school. Within these strata, schools were chosen with a probability proportional to their size (using the number of pupils in the relevant educational level), and two classes were randomly sampled within the schools. In wave 1 (2010–2011), 4963 Dutch pupils participated.[4] There were 4272 respondents in wave 3 (2012–2013), 4072 in wave 4 (2013–2014), and 3836 in wave 5 (2014–2015). In waves 3, 4, and 5, respondents could self-complete the questionnaire online, on paper, or via telephone interview.[5]

The Dutch Facebook survey

The Dutch Facebook Survey (DFS) (Hofstra et al., 2015b) was collected to enrich the Dutch part of the CILS4EU survey. It consists of behavioral data obtained from Facebook.[6] The data were collected between June 2014 and September 2014. In waves 3 and 4 of the surveys, participants were asked about their membership on Facebook. In waves 3 and 4 combined, 4864 respondents indicated that they had a membership on Facebook in at least one of these waves (wave 3 $N = 3423$, wave 4 $N = 3595$). Seven coding assistants manually tracked down Facebook profiles. This manual tracking process was based on the respondents' names and their cities of residence which were obtained from the CILS4EU survey. The coding assistants tracked down Facebook profiles on the basis of these two indicators—names and cities—and were able to match respondents to Facebook profiles with substantial certainty. The cases in which coding assistant were unsure whether the Facebook profile matched to a respondent ($N = 47$; .96%, Hofstra et al., 2015b: 11) were not taken into account in this study. A total of $N = 4463$ (91.8%) Facebook profiles were tracked. From these tracked profiles, the coding assistants downloaded the complete friend lists of the respondents in HTML code. These Facebook friend lists are the focus of this study, i.e., the first names of the individuals found among these friend lists.[7]

Approximately 73% of the respondents kept a public friend list, and we collected the first names out of these friend lists (i.e., we parsed their names out of the HTML code). See Hofstra et al. (2016) for a discussion of these respondents' privacy settings. We have information on complete Facebook networks of 3352 respondents, and they had a combined total of 1,156,285 Facebook friends.[8] Together, these individuals have 52,651 unique first names. We refer to the technical report of these data for information about how the random sample of the CILS4EU respondents compares to the sample of Facebook users (Hofstra et al., 2015b).

The Dutch civil registration

The DCR data are register data of those who have Dutch nationality and were alive and living in the Netherlands in 2010 ($N = 15{,}785{,}208$; Bloothooft and Schraagen 2011). The Facebook networks constitute 52,651 unique first names. These first names (up to the first space or hyphen) were matched to the first names in the DCR data of those having a Dutch nationality and were living in the Netherlands in 2010. We were able to match 36,151 (68.66%) of the first names in the DFS to the DCR. The names comprise ~92% of the total Dutch population ($N = 14{,}447{,}100$) and ~95% of the respondents' total Facebook friends ($N = 1{,}106{,}675$). The goal of matching these two data sets is to know the ethnic distribution of names from the register data (DCR) in the Facebook data (DFS). Thus, we have register data on 36,151 unique first names, and these comprise the major part of the Facebook friend lists. This is an indication that the vast majority of first names in the networks on Facebook are of sufficient quality to match them to the register data, i.e., Facebook friends seem to

provide realistic first names on their Facebook profiles instead of fictive pseudonyms (e.g., "Captain Fantastic"). However, it may be that individuals provide fictive first names that match the register data (e.g., "Jane," whereas the real or legal name is "Alice"). Unfortunately, with the current data, we are unable to filter such cases among individuals' Facebook friends. Because we matched Facebook profiles to survey respondents on the basis of their names (and cities), this is not an issue for the names of the respondents themselves. The register data contains information on the birth country of parents of each of these first names. For instance, for all persons named "Patrick" in the Facebook networks, it shows percentage of the parents of the Patrick's born in the Netherlands or Morocco. Table 1 provides a schematic overview of the three data sources used in this study.

First names and ethnicity

The misclassification ratio

We calculate a weighted misclassification ratio for assigning ethnicity based on first names using the register data. This is done to provide an indication of how feasible it is to predict ethnicity using first names. As an example, we use a simple *majority rule* to assign ethnicity. Per first name, we assign ethnicity (i.e., a categorization into one of the six ethnicity groups mentioned before) based on the largest proportion of the name carrier's mother's country of birth (which we obtain from the register data). Next, we subtract the maximum proportion of mother's birth countries per name. For instance, if 90% of mothers of individuals named John are born in the Netherlands, we assign a Dutch ethnicity to John. However, this would mean that 10% of all John's are misclassified as Dutch. We calculate a weighted average of this number based on all of the first names. We weigh the average by how many times a name occurs in the register data. Hence, names that occur few times have less impact on this ratio than names that occur many times. The weighted misclassification rate is 1.3% (unweighted = 5.1%). This figure implies that predicting ethnicity via first names is (in our view) sufficiently accurate and can be used as such, even though it may be somewhat driven by the Dutch majority members. To make our estimates as precise as possible, we try to further adjust for misclassifications by using a procedure that accounts for further uncertainties arising from assigning ethnicity based on first names.

The diversity of first names in the Netherlands

Especially in the Netherlands, the diversity of first names is large. There are many different spellings of nearly similar names across ethnic backgrounds and social status groups (Meertens Institute, 2016). Given that there is much variation within the set of first names in the Netherlands makes them a suitable signal of ethnicity in our study context. This is also evidenced by the low misclassification ratio discussed above. However, we acknowledge that in different contexts, last names may be more suitable signal of ethnicity (our method is not restricted to using either first or last names). Scholars, in their decision to use either first or last names, should contemplate on the diversity of the set of first and last names and how they signal ethnicity in the specific country under consideration. For instance, see Mislove et al. (2011) who use last names are used to predict ethnicity in the USA and Rao et al. (2011) who used both names to predict ethnicity in Nigeria.

Outline of the procedure

General outline

Next, we outline the general procedure we use to predict ethnicity. The idea is to predict ethnicity on the basis of people's first names. This new variable (i.e., predicted ethnicity of a Facebook friend) can then be used as an independent variable for hypothesis

Table 1. An overview of the three data sources that are used for this study.

Name	Type	Description	Abbreviation
Children of Immigrants Longitudinal Survey in four European Countries[a]	Survey	Questionnaires to adolescents (wave 3)	CILS4EU
Children of Immigrants Longitudinal Survey in the Netherlands	Survey	Questionnaires to adolescents (waves 4, 5)	CILSNL
Dutch Facebook Survey	Behavioral	Facebook networks of respondents in the surveys	DFS
Dutch Civil Registration data	Register	Dutch register data	DCR

[a]The CILS4EU and CILSNL are the same data source only under a different name, as the CILSNL continued to survey the Dutch panel of the CILS4EU four additional waves.

testing. However, when one wants to test hypotheses with predicted independent variables, one should seek to take into account the uncertainty in the predicted values of this independent variable. The uncertainty about the predictions should be adjusted for in the model coefficients of interest as well.

In our specific case, we need to consider two types of uncertainty. First, the analyses should consider the possibility that different individuals who carry the same first names can each have a different ethnicity. Hence, the most likely prediction of ethnicity, given one's first name is not always correct. For instance, a specific first name can be popular across more than one ethnic subgroup. Second, the analyses should consider that the parameters (i.e., the model coefficients) of the prediction model may also carry uncertainty, as the prediction model is estimated from the register data. The fact that we use *register* data that cover nearly the entire Dutch population may imply that the parameters are less uncertain than if they had been estimated from sampled data. Nevertheless, to elaborate our example and to be as precise as possible, we do adjust for uncertainty in the model parameters. In our example, the parameters of the prediction model are the conditional probabilities of one's ethnicity, given one's first name.

A convenient way to consider the first type of uncertainty is to use bootstrap standard errors for hypothesis testing. A bootstrap sample is a sample from the original sample of the same size drawn *with* replacement. In our method, in each bootstrap sample, a newly predicted data set is used to obtain the parameter estimates of the hypothesis testing model of interest. The bootstrap standard error is the standard deviation (*SD*) of all parameter estimates obtained by estimating the model using data from the different bootstrap samples.

If we predict the ethnicity for each Facebook friend repeatedly for each bootstrap sample, we consider the possibility that the most likely ethnicity for this Facebook friend, given his or her name may not be correct. Hence, each time a new bootstrap sample is used to estimate the parameters of our model of interest, we consider the possibility of different ethnicities among similar names.

We also take into account the second type of uncertainty, i.e., the fact that the estimates in our prediction model of interest carry uncertainty. For each bootstrap sample, we use different probabilities to predict ethnicity conditional upon one's first name. These conditional probabilities are obtained from a Dirichlet posterior distribution (see Tu, 2017 for an explanation of the Dirichlet distribution) estimated from the register data (i.e., the DCR). If we use different conditional probabilities in our prediction model to predict ethnicity for each of the bootstrap samples, we account for the uncertainty of the parameters used in the prediction model.

Model specification

Here, we describe the full estimation model in more detail. First, we outline the prediction model; subsequently, we describe the bootstrap procedure. For our prediction model, we assume that someone's ethnicity $y_1, \ldots, y_L$ is distributed according to a multinomial distribution $Mult(\theta_1,\ldots, \theta_L)$, where $y_1, \ldots, y_L$ denote indicator variables of the L ethnicity categories and $\theta_1,\ldots, \theta_L$ denote the probabilities of belonging to the Lth ethnicity category, given one's first name. We assume $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_L)$, the vector of conditional probabilities, to be distributed according to a Dirichlet distribution $Dir(a_1, \ldots, a_L)$, where $a_1, \ldots, a_L$ denote the parameters of the Dirichlet distribution. To obtain a distribution for $\boldsymbol{\theta}$, we consider the posterior distributions $f(\boldsymbol{\theta} \mid D)$, where D refers to the register data, for each occurring first name. We make use of the fact that $f(\boldsymbol{\theta} \mid D)$ is proportional to $f(D, \boldsymbol{\theta}) = f(D \mid \boldsymbol{\theta})f(\boldsymbol{\theta})$, where $f(D \mid \boldsymbol{\theta})$ is the density of the data (or likelihood of the data) and $f(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ (see, e.g., Tu, 2017). Essentially, this means that we obtain $f(\boldsymbol{\theta} \mid D)$ by multiplying the density of the data with our prior distribution. Because our prior distribution is Dirichlet and the density of the data is a multiproduct of multinomial distributions, our posterior distribution $f(\boldsymbol{\theta} \mid D)$ is also a Dirichlet distribution, which is convenient to sample from. More specifically, $f(\boldsymbol{\theta} \mid D)$ is given by $Dir(b_1, \ldots, b_L)$, where b_1 is provided by the sum of the value on the corresponding prior parameter and the number of people with ethnicity l.

As an example, consider a posterior distribution of people named John, where L is 3 (i.e., there are three ethnic categories). When the posterior distribution $f(\boldsymbol{\theta} \mid D)$ is equal to $Dir(100+1,\ 300+1,\ 30+1)$, we have observed 100 John's in the first ethnicity category, 300 in the second, and 30 in the third. The plus 1 appears from the fact that we have used an uninformative prior $f(\boldsymbol{\theta})$.

To obtain realistic bootstrap confidence intervals for our model parameters of interest, we predict ethnicity, given one's first name for each $k = 1, \ldots, K$ bootstrap sample using a freshly drawn $\boldsymbol{\theta}$ for each occurring name from $f(\boldsymbol{\theta} \mid D)$. For each of the K samples, we then obtain the regression parameters by estimating our model of interest. The bootstrap standard error of the regression parameters is then given by

$$\mathrm{SE}(\hat{\beta}) = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(\hat{\beta}_i - \bar{\beta})^2}. \quad (1)$$

Our estimation procedure is summarized in Figure 1. We aggregate the data because the Facebook networks are nested in individuals (i.e., Facebook friends of respondents in the surveys). Therefore, we obtain one value for the ethnic composition of Facebook networks per respondent (which we will elaborate upon later). We adjust the model parameters for confounding respondent-level variables (e.g., respondent's own gender and ethnicity).

Relation to imputation of missing data

We briefly note how our described procedure relates to the missing data literature. Procedures of data enrichment can be seen as a missing data problem, specifically, not having certain information about subjects while still aiming to use the missing information in an analysis. In our example, ethnicity is completely unobserved. We predict ethnicity based on a prediction model, and this prediction model resembles a missing data imputation model. For an imputation model to perform well, it should be specified according to the missing data assumption. Here, we adopt the missing at random (MAR) assumption (Rubin, 1976). Essentially, this implies that the missing data can be accounted for by the observed variables. We assume that, conditional upon one's first name, we can determine ethnicity and that there is no correlation between ethnicity and other variables. MAR is a rather strong assumption, but one that is made generally in the imputation of missing data. In our case, MAR is indeed a strong assumption, and we do not expect it to hold entirely. However, we think that the imputation model is reasonable because a first name is a strong signal of ethnicity (e.g., Bloothooft and Onland, 2011; Chang et al., 2010; Lieberson, 2000; Mateos et al., 2011) and because the weighted misclassification ratio is low.

Additionally, imputation models should account for the uncertainty in their parameters (Schafer and Graham, 2002). We do this by keeping the parameters of our imputation model random (instead of fixed). Moreover, each time a new data set is imputed, the parameters of our imputation model are sampled again from the posterior distribution.

Application of the procedure to our data

In this section, we describe the application of our procedure. As an example, we examine to what extent we can statistically relate *trust in institutions* to a predicted measure of *ethnic diversity* in Facebook networks. There is an ongoing discussion about whether ethnic

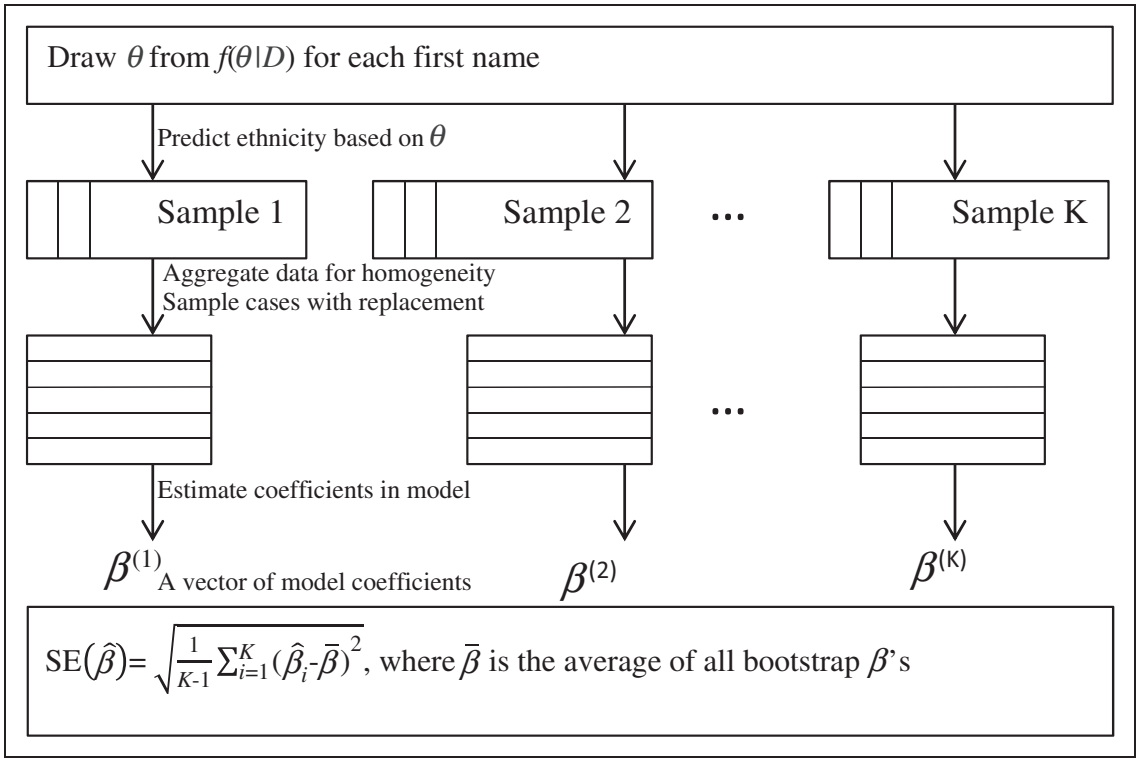

Figure 1. A graphical outline of the method to predict ethnicity.

diversity is detrimental or beneficial to social trust (see Abascal and Baldassarri, 2015; Van der Meer and Tolsma, 2014).[9] As such, we think it is engaging to explore relations between ethnic diversity on Facebook and trust using our method. However, the choice of this dependent variable is arbitrary and may as well be something different, i.e., it serves as an example. No specialized software is needed to replicate this approach, nor is the approach limited to the data we described. All computations of the procedure were performed using custom code written for the software package R (R Core Team, 2016).

Dependent and control variables

The variable that we thus want to relate to our predicted values of ethnicity is *trust in institutions*. We define trust in institutions as a specific form of *generalized trust*. Where generalized trust is occasionally specified as whether individuals trust "most others" (e.g., Paxton, 2007), we define trust in institutions as the extent to which individuals trust "institutions." This variable is constructed as follows. In wave 5 of the CILSNL, respondents indicated on a 10-point scale on four items how much confidence (1, no confidence whatsoever—10, a lot of confidence) they had in politicians, judges, scientists, and the police. We took the mean score out of these four to calculate trust in institutions ($M = 6.358$; $SD = 1.413$; Cronbach's $\alpha = .803$).

Next, we construct five control variables from the third and fourth waves of the CILS4EU and CILSNL to examine whether we can isolate the effect of our newly predicted variable from confounding factors. In constructing these variables, we took the survey answers from wave 4. If these survey answers were missing, we took answers from wave 3. First, we construct the *ethnicity* of the respondents themselves. We classify respondents into one of the six largest ethnic groups in the Netherlands: "Native Dutch" (76.9%), "Turkish" (3.7%), "Moroccan" (3.5%), "Dutch Caribbean" (3.9%), "Other Western" (5.4%), and "Other non-Western" (6.6%). As we mentioned before, this is based on their parents' country of birth (Statistics Netherlands, 2012; Vermeij et al., 2009) that they reported in the surveys. Second, we measured whether the respondents indicated whether they were a *girl* (59.6%) or a *boy* (40.4%). Third, we measured how satisfied respondents reported being with their "life in general" (1, very dissatisfied—10, very satisfied; $M = 7.585$; $SD = 1.613$). Fourth, we measured whether respondents were in a romantic relationship (38.7%) or not (62.3%). We adjust for these factors according to their availability, and the questions were posed in a similar way to respondents across the multiple waves of survey data. Finally, using dummy variables, we measured in which of the waves the respondents participated (waves, 3, 4, and 5 = 78.4%; waves 4 and 5 = 15.8%; or waves 3 and 5 = 57.5%). We listwise delete missing values across these six variables and realize a data set consisting 3445 cases.

Predicting ethnicity and estimation results

Next, we want to obtain the most likely value of ethnicity, given the first name of each of the Facebook friends of a respondent. We have register data for 36,151 of the first names on Facebook (95% of all friends). These register data contain (1) the number of occurrences of each of these first names in the Netherlands and (2) for each first name, the fraction of the name carriers' mothers who were born in the six major origin countries (the Netherlands, Turkey, etc.). We thus predict the ethnicity of friends on Facebook via the birth country of the mothers. This is a small deviation from the regularly applied definition in the cases where the mothers were born in the Netherlands, but the fathers were born elsewhere. To keep our method parsimonious, we consider only mothers' birth countries. For 2208 out of the 3445 of respondents (from which we have all values across the dependent and control variables), we can observe the first names in their Facebook friend lists, i.e., these respondents have public friend lists. This final set of 2208 respondents have a combined total of 776,135 Facebook friends for which we want to predict ethnicity.

For each bootstrap sample, we draw the conditional probabilities for ethnicity, given one's first name on Facebook using the posterior distribution obtained with the register data. Next, we count the number of friends who have the same predicted ethnicity as the ethnicity of the respondent himself or herself and divide this count by the total number of friends and multiply it by 100. As such, we calculate the percentage of co-ethnic Facebook friends per respondent, or co-ethnic$_{\text{FACEBOOK}}$, as a measure of ethnic homogeneity. This means that we aggregated the predicted values for ethnicity across the friends' first names from the respondent's Facebook network.

This aggregated predicted variable is then used as an independent variable in a linear regression model—also adding the control variables mentioned before—with trust in institutions as the dependent variable. We repeat this process 10,000 times, each time bootstrapping new conditional probabilities for ethnicity and each time bootstrapping the model coefficients from the linear regression models. We obtain a distribution of 10^4 bootstrap coefficients per variable in the linear regression model. Assessing these 10^4 bootstrap

coefficients, we can obtain the bootstrap confidence interval and observe whether the middle 95% of the bootstrapped coefficients are either above or below zero. If we plot these 10^4 coefficients, they visually resemble normal distributions. Addressing this bootstrap confidence interval is a nonparametric alternative to standard null-hypothesis significance testing.

Table 2 shows the bootstrap results of 10^4 linear regressions. We show upper and lower quantiles and the means of the coefficients. We briefly discuss the results of these regression models. First, those *with* a romantic partner seem to report *less trust* in institutions, and those who report higher *life satisfaction* report *more trust* in institutions. Those of *Turkish* and *Dutch Caribbean* ethnicity report *less trust* in institutions than members of the Dutch majority. The magnitude of these ethnicity effects seems rather high if we consider the mean of all coefficients. Those who participated in waves 4 and 5 report less trust in institutions than those who participated in waves 3, 4, and 5. Finally, we do not observe a relation between the percentage of co-ethnic Facebook friends and trust, given our bootstrapped coefficients.

Table 2. Regression results of co-ethnic$_{\text{FACEBOOK}}$ and control variables on trust in institutions (using 10^4 bootstraps and $N = 2208$).

	Bootstrap coefficients			
	M	Lower (2.5%)	Upper (97.5%)	Deviates from 0?[a]
Co-ethnic$_{\text{FACEBOOK}}$	0.113	−0.795	1.037	No
Intercept	6.058	5.198	6.909	Yes
Wave				
Wave 3, 4, and 5	Ref.	Ref.	Ref.	
Wave 4 and 5	−0.340	−0.484	−0.195	Yes
Wave 3 and 5	−0.017	−0.260	0.224	No
Girls (ref. boys)	−0.012	−0.113	0.091	No
Romantic partner	−0.110	−0.214	−0.006	Yes
Life satisfaction	0.088	0.054	0.123	Yes
Ethnicity				
Dutch	Ref.	Ref.	Ref.	
Turkish	−1.277	−1.897	−0.628	Yes
Moroccan	−0.593	−1.316	0.167	No
Dutch Caribbean	−0.875	−1.641	−0.083	Yes
Other Western	−0.165	−0.932	0.621	No
Other non-Western	−0.427	−1.235	0.398	No

[a]If the middle 95% of the coefficients do not contain zero, we can safely assume that the coefficient is nonzero.

Model performance

Next, we provide an indication of the performance of our method. We provide confidence intervals for three methods of estimating the effects of co-ethnic$_{\text{FACEBOOK}}$ on trust in institutions. First, we provide confidence intervals based on 10^4 bootstrapped coefficients as outlined in this study. Second, we assign ethnicity based on the majority rule explained before and calculate the fraction of co-ethnic Facebook friends. We then ran a linear regression of this measure—controlling for the variables mentioned before—on trust. Third, we replicated the co-ethnic$_{\text{FACEBOOK}}$ measure of Hofstra et al. (2017). They used a training data set, where they knew the self-reported ethnic background and first names of the respondents and sought which proportions of parents' birth country in the register data correlated best with self-reported ethnicity in the survey. We regress trust in institutions on the fraction of co-ethnic Facebook friends using this measure (while controlling for the set of variables mentioned before).

Table 3 presents confidence intervals for effects on trust in institutions of the three methods of calculating the percentage of co-ethnic Facebook friends. It presents conservative, regular, and nonconservative confidence interval boundaries of the coefficients researchers usually consider for standard statistical significance across three panels of results.[10]

We observe that the method outlined in this study provides the most-conservative tests of the effects of the predictor on the dependent variable. Using a simple majority rule for only one predicted data set may lead more often to false-positive results. One may conclude that the more homogeneous Facebook networks are the more trust individuals have in institutions (in the consideration of the third panel of results). However, when we take into account the two types of uncertainty incorporated in our method, i.e., the possibility of different ethnicities among similar first names and the uncertainty in the model coefficients in the prediction

Table 3. A comparison of three methods for predicting ethnicity, presented are confidence intervals for effects of co-ethnic$_{\text{FACEBOOK}}$ on trust.

	Conservative		Regular		Nonconservative	
	2.5%[a]	97.5%	5%	95%	10%	90%
This study	−0.795	1.037	−0.658	0.891	−0.482	0.707
Majority rule	−0.198	1.421	−0.068	1.291	0.082	1.141
Hofstra et al. (2017)	−0.618	0.238	−0.549	0.169	−0.470	0.090

[a]These are the confidence intervals for the different methods.

model, we no longer observe such a relationship. The method provided in Hofstra et al. (2017) does not seem to lead to false inferences using this example. However, the confidence intervals are more conservative using the method provided in this article. These findings suggest that using our method in the process of hypothesis testing may provide more-conservative tests of hypotheses that are less prone to false-positive results.[11]

Conclusions and discussion

The research aim of this study was twofold. First, we outlined a procedure to predict ethnicity in social media data using register data. Second, we showed how to use these predicted values in standard regression models to tests hypotheses. As such, we contributed to an expanding amount of prior work aiming to enrich social media data (e.g., Chang et al., 2010; Hofstra et al., 2017) based on the idea that names are a signal of individual characteristics such as ethnicity (Bloothooft and Onland, 2011; Fiscella and Fremont, 2006; Lieberson, 2000). We provided a contribution that accounted for (1) the possibility of people with similar names having different ethnicities and (2) uncertainty in model estimates when using predicted values as independent variables in linear regression models. We did so by bootstrapping conditional probabilities, given one's first name and bootstrapping standard errors from a set of 10^4 linear regression models. The percentage of misclassifications of ethnic background using a simple majority rule was approximately 1.3%, which is relatively low and may be a further illustration of the promise of our approach.

We provided a toy example showcasing how we could predict the ethnicity of respondents' friends on Facebook (usually not readily available) and related respondents' percentage of co-ethnic friends on Facebook to trust in institutions. As such, we provided a way to illuminate long-standing substantive discussions in future research. In this example, we explored the relation between ethnic diversity and trust (e.g., Abascal and Baldassarri, 2015; Van der Meer and Tolsma, 2014). We found no significant relation between trust in institutions and the predicted values of the percentage of co-ethnic Facebook friends.

We compared the method outlined in this study with two other, more straightforward ways to predict ethnicity on the basis of first names. First, we compared it with a simple majority rule within the register data and, second, with a supervised learning method. The results suggest that the method outlined in this study is the least prone to false-positive results. We showed that using more-straightforward ways to assign ethnicity based on first names may lead to the conclusion that the percentage of co-ethnic Facebook friends positively predicts trust in institutions. However, we showed that this result may be a false-positive finding as a consequence of not accounting for the two types of uncertainty in the analyses. This comparison highlights the crucial importance of a nuanced data-analytic approach to enrich social media data; if one relies on simpler methods to do so, statistical inference may be incorrect.

Limitations of this study

Five limitations of this study merit acknowledgement. First, there may be selection biases in these data. Previous work, however, showed that ethnic homogeneity estimates on Facebook do not differ when taking into account sample selections or not (Hofstra et al., 2017). Furthermore, the observed results from the Facebook data may be driven by the fact that those of non-Dutch ethnic background more often have opted for private Facebook profiles (see Hofstra et al., 2016). According to the homophily principle (Lazarsfeld and Merton, 1954; McPherson et al., 2001), individuals prefer to befriend ethnically similar others above those of ethnically dissimilar backgrounds. When Dutch majority members more often have open profiles and they more often befriend Dutch friends on Facebook (which they do, see Hofstra et al., 2017), this will be reflected in an oversampling of Dutch majority members and Dutch first names among the Facebook friend lists. In itself, this is an interesting intuition, because then the question arises of which individuals we as researchers *by design* can study in online social media data. Future researchers in the field should be aware of such selection patterns. The goal of this paper was to outline our method, but an oversampling of typical Dutch first names may have led to insufficient variation in the ethnic homogeneity measure. However, the two alternative methods we compared our method with experienced this same issue. The finding that our method was the least prone to false-positives strongly underlines the importance of the method, assuming that one aims for conservative hypotheses tests.

Second, we lumped "Other Western" and "Other non-Western" roots together as ethnic background categories, whereas it is only reasonable to assume that there is substantial variation in naming habits within these categories. For instance, those of Afghan origin are labeled under other non-Western ethnicities, as are those with Chinese roots. However, the naming practices between these countries vary substantially. In future research, scholars should consider how to strike a compromise between sufficient observations within ethnic-racial background categories and the precision of the ethnicity predictions. This is also highly

dependent on the substantive question the researcher aspires to address. One can, for instance, decide to only study groups for which the data cells are sufficiently filled, without examining possible residual categories. Another example would be to use more-precise "Other Western" categories if one is interested in interethnic ties on Facebook between majority and minority members from neighboring countries to the Netherlands. In this study, we used these categories because studies on ethnicity in the Netherlands usually apply this categorization (Statistics Netherlands, 2015).

Third, there were approximately 16.6 million inhabitants in the Netherlands in 2010. This means that the number of cases in the register data cover ~95% of all inhabitants in the Netherlands. Moreover, there were approximately 800,000 inhabitants in the Netherlands in 2010 not carrying a Dutch nationality. There may be situations where first names of ethnic minorities who do not have a Dutch nationality vary from those of ethnic minorities who do. This may result in predictions for these first names that are less precise because we do not have register data on their mothers' country of birth. Unfortunately, we cannot adjust for these situations with the current data.

Fourth, our proposed method relies on the reliability of names in social media platforms. This seems particularly suited to study social network sites where the ratio of non-fictive versus fictive names is high—e.g., Facebook or LinkedIn. However, there may be platforms where oftentimes aliases are used (e.g., Twitter, a microblogging website, see also Pennacchiotti and Popescu, 2011) and where our method may be less suited for. Furthermore, a related issue is the noise that fictive accounts and *bots* may add to the data. In this study, we used a controlled data set of Facebook profiles which likely reduces such noise and if the distribution of fictive accounts and bots is similar across ethnicities this will further reduce noise.

Finally, register data on the prevalence of names and their associated demographics might not be readily available in every study context. However, register data that contains demographic information on names is not unique to the register data from the Netherlands and does exist in other (country-specific) contexts. For instance, the US Census Bureau provides a list of first names *and* last names in the US (occurring >100 times) and the percentage of these first and last name carriers' racial backgrounds (United Status Census Bureau, 2014). One could implement our method using these data and compare how the predictions across first and last names are similar. Another context would be Germany, where Statistics Germany provides lists of first names by region and gender (Statistics Germany, 2016). Furthermore, the US Social Security Administration (2017) provides a list of nearly all given names and their genders from 1900 to 2016. These alternative data sources might be suitable points of departure for replication of our study and implementation of our method.

Implications and future research directions

Next, we discuss implications and future research directions, and we pinpoint alternative data sources. We urge scholars to use (variants of) this method for future scientific endeavors, especially because of the growing use of online social network data and the challenges that come with it. A first key future research path we would commend is replicating the method using register data on social class. Social class is related to names as well, as parents from different societal strata name their children differently (Bloothooft and Schraagen, 2011; Lieberson, 2000). Defining social class disparities in online behavior relates to issues of digital inequalities (e.g., DiMaggio et al., 2001; Hargittai, 2002). Moreover, investigating online social network segregation or integration by social class is an understudied area that directly merits further investigation. One way to do this is to obtain register data on names and information on parents' educational background about these names. Another feature of first names is that they vary in their popularity *over time*. This may provide possibilities to study age cleavages in online social networks, i.e., one may aim to predict age or age-cohort based on first names.

A second future research endeavor would be to directly test hypotheses using "upgraded" social media data. Here, we provided an example using trust in institutions as the dependent variable. However, another example that may directly relate to ethnic homogeneity on Facebook (predicted via first names on Facebook) is *ethnic prejudice*. Literature suggests that even superficial contacts between members of different ethnic groups potentially reduce intergroup prejudice (Allport, 1954; Pettigrew and Tropp, 2006). Implementing our method using ethnic prejudice instead of trust in institutions as the dependent variable would be a direct test of the hypothesis of whether ethnic diversity among Facebook friends hampers ethnic prejudice.

Finally, we pinpoint data sources other than the ones we used. Using these data sources, scholars can replicate our method across different contexts and using different sources of data. Gathering network data form social media is relatively easy (Golder and Macy, 2014), and our predicted values of ethnicity are not limited to Facebook (nor are they to limited to ethnicity). For instance, the Application Programming Interface of Twitter is straightforward to access via custom-tailored packages in the statistical package R. Although,

a consideration for future scholars is the extent to which Twitter users use aliases instead of real (first) names. Another data source would be the networks found on LinkedIn, which tend to be related to the professional, work-related networks people have. An interesting question would be to what extent (ethnic) diversity in these networks by, for instance, social status is related to labor market outcomes (e.g., Granovetter, 1973, 1983). Additionally, predicting measures of gender diversity may be beneficial to examine relationships between gender diversity and gender attitudes. We described a method to predict the ethnic background of members in Facebook networks—which is unavailable in many cases (Cesare et al., 2017; Spiro, 2016)—on the basis of people's first names, but our procedure is not necessarily limited to specific individual characteristics. We recommend future replication efforts of our procedure on different individual characteristics and in different national contexts.

Acknowledgments

We thank Lukas Norbutas, Rense Corten, and Frank van Tubergen for invaluable feedback on an earlier draft of this study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research benefited from the support of the NORFACE research program on Migration in Europe—Social, Economic, Cultural, and Policy Dynamics and from the support of several grants from the Dutch Scientific Organization (NWO): NWO onderzoekstalent (grant number: 406-12-004), NWO middelgroot (grant number: 480-11-013), and NWO veranderingsstudies (grant number: 481-11-004).

ORCID iD

Bas Hofstra http://orcid.org/0000-0002-9052-956X.

Notes

1. Data enrichment resembles issues of missing data. We specify later on how these two bodies of work relate to one another.
2. One can apply for data access to waves 1, 2, and 3 of the CILS4EU via the following link: www.cils4.eu.
3. One can apply for data access to waves 4, 5, 6, and 7 of the CILSNL via the following link: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:65866.
4. Six hundred respondents in wave 1 were sampled who were not a part of the random sampling frame because some schools wanted to participate in the survey with more than two classrooms. Therefore, a *random* sample of 4363 pupils was drawn in wave 1. Because of attrition rates between waves 1 and 2, our sample cannot be guaranteed to be representative. We include as many respondents as possible in the sample for analyses, including newcomers (nonrandom) and the nonrandom sample of wave 1, to ensure a large sample size across waves.
5. A minority of the pupils in the higher educational track were still in high school in wave 3. These pupils were still surveyed at their respective schools while a researcher was present.
6. An anonymized version of the DFS is available from October 2017 onward via the following link: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:62379.
7. The collected information was publicly visible on Facebook. Coding assistants were instructed personally, and all followed strict coding procedures with password-protected files. All personal identifiers were removed from the data. The data collection, the coding procedure, and the use of these data for scientific purposes were reviewed and approved by an internal review board of the Faculty of Social and Behavioural Sciences at Utrecht University (project number: FETC14-019).
8. This total of 1,158,227 Facebook friends is a raw count of all of the friendships respondents have. A likely situation is that respondents have the same friends in their Facebook networks. Counting this unique set of Facebook friends would likely result in a lower number.
9. We use trust in institutions as one dimension of *social trust*. We find that *generalized trust* (often used as a measure for social trust), i.e., whether individuals think that "most people can be trusted," correlates with trust in institutions ($r = .342$; $p < .001$). We do not use generalized trust, as it is a dichotomous-dependent variable, and this makes our estimation procedure more complex and relatively slow. We acknowledge, however, that trust in institutions is only a dimension of social trust.
10. The control variables are omitted from Table 3. However, the control variables had qualitatively similar results over the different analyses, comparable with the coefficients found in Table 2.
11. One may also argue that our procedure is more prone to false-negative results in the case of a false null-hypothesis. In the situation of null-hypothesis testing, however, one should aim for conservative hypothesis tests.

References

Abascal M and Baldassarri D (2015) Love thy neighbor? Ethnoracial diversity and trust reexamined. *American Journal of Sociology* 121(3): 722–782.

Allport GW (1954) *The Nature of Prejudice*. Reading: Addison-Wesley.

Bakshy E, Messing S and Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239): 1130–1132.

Bloothooft G and Schraagen M (2011) Name fashion dynamics and social class. Working paper. Available at: www.let.uu.nl/~Gerrit.Bloothooft (accessed 8 August 2016).

Bond RM, Fariss CJ, Jones JJ, et al. (2012) A 61 million-person experiment in social influence and political mobilization. *Nature* 489: 295–298.

boyd dm and Ellison NB (2008) Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* 13(1): 210–223.

Castles S, De Haas H and Miller MJ (2013) *The Age of Migration: International Population Movements in the Modern World*, 5th ed. New York: The Guilford Press.

Cesare N, Grant C and Nsoesie EO (2017) Detection of user demographics on social media: A review of methods and recommendations for best practices. Available at: arxiv.org/abs/1702.01807 (accessed 13 February 2017).

Chang J, Rosenn I, Backstrom L, et al. (2010) ePluribus: Ethnicity on social networks. In: *Proceedings of the fourth international AAAI conference on weblogs and social media*, pp. 18–25. Available at: https://pdfs.semanticscholar.org/91c0/9bed0c0caa8e3df87bf33d50edd242a1b997.pdf.

Coldman AJ, Braun T and Gallagher RP (1998) The classification of ethnic status using name information. *Journal of Epidemiology and Community Health* 42(4): 390–395.

Corten R (2012) Composition and structure of a large online social network in the Netherlands. *PLoS ONE* 7(4): e3476.

DiMaggio P, Hargittai E, Neuman WR, et al. (2001) Social implications of the Internet. *Annual Review of Sociology* 27: 307–336.

Ellison NB, Steinfield C and Lampe C (2007) The benefits of Facebook ''friends'': Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12(4): 1143–1168.

Ellison NB, Steinfield C and Lampe C (2011) Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media & Society* 13(6): 873–982.

Fiscella K and Fremont AM (2006) Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research* 41(4): 1482–1500.

Golder SA and Macy MW (2014) Digital footprints: Challenges and opportunities for online social research. *Annual Review of Sociology* 40: 129–152.

Gonzàles-Bailòn S and Wang N (2016) Networked discontent: The anatomy of protest campaigns in social media. *Social Networks* 44: 95–104.

Granovetter MS (1973) The strength of weak ties. *American Journal of Sociology* 78(6): 1360–1380.

Granovetter MS (1983) The strength of weak ties: A network theory revisited. *Sociological Theory* 1: 201–233.

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.

Hargittai E (2002) Second level digital divide: Differences in people's online skills. *First Monday* 7(4). Available at: http://firstmonday.org/article/view/942/864 (accessed 13 March 2017).

Hobbs WR, Burke M, Christakis NA, et al. (2016) Online social integration is associated with reduced mortality risk. *Proceedings of the National Academy of Sciences* 113(46): 12980–12984.

Hofstra B, Corten R and Buskens V (2015a) Learning in social networks: Selecting profitable choices among alternatives of uncertain profitability in various networks. *Social Networks* 43: 100–112.

Hofstra B, Corten R and Van Tubergen F (2015b) Dutch Facebook survey: Wave 1 [dataset and codebook]. Available at: easy.dans.knaw.nl (accessed 1 November 2016).

Hofstra B, Corten R and Van Tubergen F (2016) Understanding the privacy behavior of adolescents on Facebook: The role of peers, popularity and trust. *Computers in Human Behavior* 60: 611–621.

Hofstra B, Corten R, Van Tubergen F, et al. (2017) Sources of segregation in social networks: A novel approach using Facebook. *American Sociological Review* 82(3): 625–656.

Jaspers E and Van Tubergen F (2017) Thematic collection: Children of immigrants longitudinal survey in the Netherlands (CILSNL). DANS. Available at: easy.dans.knaw.nl (accessed 1 November 2016).

Kalmijn M and Van Tubergen F (2006) Ethnic intermarriage in the Netherlands: Confirmations and refutations of accepted insights. *European Journal of Population* 22(4): 371–397.

Kalter F, Heath A, Hewstone M, et al. (2016) Children of immigrants longitudinal survey in four European countries (CILS4EU)—full version. ZA5353, GESIS Data Archive, Cologne, Data file Version 3.1.0.

Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790.

Lauderdale DS and Kestenbaum B (2000) Asian American ethnic identification by surname. *Population Research and Policy Review* 19: 283–300.

Lazarsfeld PF and Merton RK (1954) Friendships as a social process: A substantive and methodological analysis. In: Berg M (ed.) *Freedom and Control in Modern Society*. New York: Van Nostrand, pp. 18–66.

Lazer D, Pentland A, Adamic L, et al. (2009) Computational social science. *Science* 323(5915): 721–723.

Lewis K, Kaufman J, Gonzalez M, et al. (2008) Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks* 30(4): 330–342.

Lieberson S (2000) *A Matter of Taste: How Names, Fashions, and Culture Change*. New Haven & London: Yale University Press.

Marsden PV (1987) Core discussion networks of Americans. *American Sociological Review* 52(1): 122–131.

Mäs M and Flache A (2013) Differentiation without distancing: Explaining bi-polarization of opinions without negative influence. *PloS ONE* 8(11): e74516.

Mateos P, Longley PA and O'Sullivan D (2011) Ethnicity and population structure in personal naming networks. *PLoS One* 6(9): e22943.

Mayer A and Puller SE (2008) The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics* 92(1–2): 329–347.

McFarland DA and McFarland HR (2016) Big Data and the danger of being precisely inaccurate. *Big Data & Society* 2(2): 1–4.

McPherson M, Smith-Lovin L and Brashears ME (2006) Social isolation in America: Changes in core discussion networks over two decades. *American Sociological Review* 71(3): 353–375.

McPherson M, Smith-Lovin L and Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1): 415–444.

Meertens Institute (2016) De Nederlandse Voornamenbank. Available at: http://www.meertens.knaw.nl/nvb/ (accessed 1 November 2016).

Mislove A, Lehmann S, Ahn Y, et al. (2011) Understanding the demographics of Twitter users. In: *Proceedings of the fifth international AAAI conference on weblogs and social media*, pp. 554–557. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816

Paxton P (2007) Association memberships and generalized trust: a multilevel model across 31 countries. *Social Forces* 86(1): 47–76.

Pennacchiotti M and Popescu A (2011) A machine learning approach to Twitter user classification. In: *Proceedings of the fifth international AAAI conference on weblogs and social media*, pp. 281–288. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2886

Pettigrew TF and Tropp LR (2006) A meta-analytic test of integroup contact theory. *Journal of Personality Social Psychology* 90(5): 751–783.

Putnam RD (2000) *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Rao D, Paul M, Fink C, et al. (2011) Hiearchical Bayesian models for latent attribute detection in social media. In: *Proceedings of the fifth international AAAI conference on weblogs and social media*, pp. 598–601. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2881

R Core Team (2016) R: *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rubin DB (1976) Inference and Missing Data. *Biometrika* 63(3): 5810592.

Schafer JL and Graham JW (2002) Missing data: Our view of the state of the art. *Psychological Methods* 7(2): 147–177.

Smith S, Maas I and Van Tubergen F (2014) Ethnic ingroup friendships in schools: Testing the by-product hypothesis in England, Germany, the Netherlands and Sweden. *Social Networks* 39: 33–45.

Social Security Administration (2017) Popular baby names. Available at: https://www.ssa.gov/oact/babynames/limits.html (accessed 27 September 2017).

Spiro ES (2016) Research opportunities at the intersection of social media and survey data. *Current Opinion in Psychology* 9: 67–71.

Stark TH and Flache A (2012) The double edge of common interest. *Sociology of Education* 85(2): 179–199.

Statistics Germany (2016) Liste der Häufigen Vornamen 2015. Available at: www.govdata.de (accessed 16 August 2016).

Statistics Netherlands (2012) Migranten, Vreemdelingen en Vluchtelingen: Begrippen op het Terrein van Asiel en Buitenlandse Migratie. Available at: www.cbs.nl (accessed 16 June 2015).

Statistics Netherlands (2015) CBS Statline: Bevolking; kerncijfers. Available at: statline.cbs.nl (accessed 23 March 2015).

Stopczynski A, Sekara V, Sapiezunski P, et al. (2014) Measuring large-scale social networks with high resolution. *PLoS ONE* 9(4): e95978.

Tu S (2017) The Dirichlet-Multinomial and Dirichlet-categorical models for Bayesian inference. Available at: people.eecs.berkeley.edu/~stephentu (accessed 24 February 2017).

United States Census Bureau (2014) Frequently occurring surnames from the census 2000. Available at: www.census.gov (accessed 16 August 2016).

Van der Meer T and Tolsma J (2014) Ethnic diversity and its effects on social cohesion. *Annual Review of Sociology* 40: 459–478.

Van Tubergen F (2015) Ethnic boundaries in core discussion networks: A multilevel social network study of Turks and Moroccans in the Netherlands. *Journal of Ethnic and Migration Studies* 41(1): 101–116.

Verkuyten M and Kwa GA (1994) Ethnic self-identification and psychological well-being among minority youth in the Netherlands. *International Journal of Adolescence and Youth* 5(1–2): 19–34.

Vermeij L, Van Duijn MAJ and Baerveldt C (2009) Ethnic segregation in context: Social discrimination among native Dutch pupils and their ethnic minority classmates. *Social Networks* 31(4): 230–239.

Watts DJ (2011) *Everything is Obvious: How Common Sense Fails Us*. New York: Crown Business.

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.