

The why (not) and how (not) of survey to digital footprint linkages: a use-case of ethnic background and social relationships

Bas Hofstra

To cite this article: Bas Hofstra (2025) The why (not) and how (not) of survey to digital footprint linkages: a use-case of ethnic background and social relationships, Journal of Ethnic and Migration Studies, 51:12, 3117-3134, DOI: [10.1080/1369183X.2025.2487745](https://doi.org/10.1080/1369183X.2025.2487745)

To link to this article: <https://doi.org/10.1080/1369183X.2025.2487745>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 17 May 2025.



[Submit your article to this journal](#)



Article views: 454



[View related articles](#)



[View Crossmark data](#)



Citing articles: 3 [View citing articles](#)

The why (not) and how (not) of survey to digital footprint linkages: a use-case of ethnic background and social relationships

Bas Hofstra 

Department of Sociology/ICS, Radboud University, Nijmegen, the Netherlands

ABSTRACT


In the last decade, there has been a rise of mixed-method-big data approaches in the social sciences. One such approach is linking survey data with digital footprint data. Yet, some of the fundamental challenges of linking survey data with digital footprint data are often unaddressed in prior work. Yet especially because of the importance of ethnic background dynamics in social relationships and the opportunities that survey/digital footprint data linkages afford, addressing such challenges is key. This is because statistical and methodological errors compound and are reinforced when one links (selective) data sources; or when social categories in footprint data are not as detailed as they could be. This contribution lays out *why* linking survey data with digital footprint data provides insights into ethnic inequalities and intergroup social cohesion, yet simultaneously addresses some of the challenges of doing so with a particular use-case as an example. Consequently, this contribution provides case-examples of methodologies for linking survey and footprint data – i.e. how to be aware of and account for sampling bias, and how to extract meaning in the context of ethnic background from unstructured data.

KEYWORDS

Social networks;
computational sociology;
ethnic background; digital
traces; survey research

Introduction

A defining feature of social life is how individuals are embedded in a web of social relationships. Some individuals are able to forge and maintain many social relationships and are deeply centred between them, whereas others are less able to find or maintain even a few good friendships. These social relationships come about through a set of micro-, meso-, and macro-level mechanisms (Blau 1977; McPherson, Smith-Lovin, and Cook 2001), and they influence our everyday lives, behaviours, and opinions (e.g. Allport 1954; Hofstra 2022). These social relationships affect many dimensions of everyday life and they are often inextricably linked to issues surrounding in- and out-group dynamics. So much so, that one of the stylised facts on social relationships is that

CONTACT Bas Hofstra  bas.hofstra@ru.nl

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

‘birds of a feather flock together’; social relationships heavily segregate, and overwhelmingly so by ethnic background (and race) (McPherson, Smith-Lovin, and Cook 2001). This has far-reaching societal consequences. For instance, for how prejudice (e.g. by ethnic background) diminishes in specific relationship configurations. Or for how political talk between ethnically similar friends affects voting intentions for ethnic minority interest parties. Decades of quantitative forays into these topics predominantly relied on data originating from questionnaires. This yielded invaluable insights into dynamics that entrench or ameliorate inequality and that deepen or weaken social cohesion along lines of ethnic background.

Yet over the last (two) decade(s) or so, some questions on social relationships have been studied and partly answered with a data lens that sometimes observes with *greater radius* and *greater resolution*. Specifically, social relationship dynamics have been analysed with new sources of so-called ‘digital footprint’ or ‘digital trace’ data capturing or measuring social relationships and other behavioural dynamics. These new sources of data comprise, among others, information from social media websites such as Twitter/X or Facebook. Analysing such data, is sometimes coined a computational approach to the social sciences, and such approaches had a rise to prominence in the last decade(s) (Edelmann et al. 2020; Golder and Macy 2014).

How does this ‘computational social science’ research trend (or fad?) fit in a mixed-method approach centred in this special issue? A main argument of this essay is that the *mixing* of survey data and digital footprint data – and the associated *mixing* of ‘traditional’ and computational methods – benefits the study of social relationships in relation to ethnic background, though there are challenges that undermine those benefits. I specifically reflect on the empirical innovations, practical considerations, and opportunities and pitfalls of *mixing and analysing survey data with quantitative digital footprint data sources, for the study of ethnic background vis-à-vis causes and consequences of social relationships*. I do so by explicitly focusing on experiences, chances, and challenges by means of a particular use-case. This use-case studied causes and consequences of social relationships of Dutch ethnic majority members with both parents born inside the Netherlands and those ethnic minority members whose parents were born outside of the Netherlands. Ethnic background conceptualised by parents’ birth countries is often standard practice in studies on relationships and ethnic background (e.g. see Bubritzki et al. 2018; Vermeij, Van Duijn, and Baerveldt 2009). Yet, some of the arguments later in the essay are not wholly unique to such a conceptualisation, and others (e.g. self-identification, dual identifying, etc.) may equally apply. As such, this essay both contributes to ‘providing insights into *doing* mixed methods’ and ‘understanding and lessening barriers’ vis-à-vis mixed methods concerning this research topic, neatly aligning with both the goals of the Special Issue (Geurts, Davids, and Spierings 2025).

With this essay, I build on prior overview essays and reviews that discuss computational approaches to the social sciences (e.g. Drouhot et al. 2023; Golder and Macy 2014; Lazer et al. 2009, 2020; Lewis 2015; McFarland and McFarland 2015; Salganik 2019). These all provide insightful, and often in-depth discussions on those approaches, or provide general outlooks, or reviews and introductions. Yet, this essay contributes a unique angle that directly ties some of the promises and pitfalls often discussed to the use-case that directly empirically identifies and pragmatically addresses them. The essay will describe in detail how the use-case can identify and address them because of a unique

interface of survey data and digital trace data on social relationships. To my knowledge, such a focus has hardly been employed in-depth elsewhere.

Given the aims of this essay, I envision at least three readerships. First, (survey) scholars of in- and out-group dynamics, perhaps with a specific interest in ethnic background and social relationships such as defined earlier, might be keen to learn about mixing different types of quantitative data and computational methods and its chances and pitfalls to answer innovative questions in their study realm. Second, computational social scientists that look for a deep dive into a specific use-case that directly matches, empirically identifies, and solves some of the often discussed, yet fundamental challenges and opportunities in their field in one domain (ethnic background and social relationships). Note that ‘computational social scientist’ is a very broad label. This is intentional, and I pragmatically define ‘computational social scientists’ later on in the essay (in short: the definition is inconsequential). Third, a general readership in ethnic and migration studies might be interested in some of the more general questions surrounding the large field studying ethnic background and social relationships, or in mixing various quantitative data sources and methods. For the readers interested in computational social science and ethnic and migration studies, I refer to the special issue solely dedicated to that topic in this journal (Drouhot et al. 2023).

This essay employs an emphasis on research about social relationships and ethnic background. This is for three reasons. First, I count survey and digital footprint social relationships linkages along my own expertise. This allows me to reflect on my own experiences of making those linkages, which is among the essay’s goals. Second, social relationship causes and consequences are often intersected with ethnic background. This renders the research topics that this essay engages with particularly relevant – e.g. who are our social ties; are they ethnically similar or dissimilar and why and on what scale; what consequences does this have? Who is connected to whom relates to levels of ethnic prejudice and to polarised subsections of society where people of different ethnic backgrounds hardly meet, let alone share positive experiences and interactions. Note that the ‘social relationship’ focus is a sociologically meaningful use-case, but that some of the mixing quantitative data/methods experiences here apply broader too. Third and most importantly, promises and pitfalls of digital footprints to study social relationships and ethnic background are particularly salient. I zoom in on those particularly salient promises and pitfalls in the context of social relationships and ethnic background. This is why the use-case is essential, as it uniquely serves to highlight how prone it is to face such issues – and particularly so in the context of (biases by) ethnic background – and how the interface between survey and digital trace data might help to find solutions to some of those issues.

I structure this essay as follows. I start out with a short, general discussion on new types of digital footprint data, computational social science, and the promises and pitfalls of a solely ‘computational’ approach in relation to social relationships. This serves as a baseline to understand digital footprint data and its opportunities and challenges to subsequently engage in a deeper reflection on how this fits into research on ethnic background and social relationships and the case-specific opportunities and challenges. I highlight that this serves a starting-point introduction and that I merely report and summarise earlier work. Other work covers these promises and pitfalls more in-depth (e.g. Drouhot et al. 2023; Golder and Macy 2014; Salganik 2019). Yet, the baseline serves the somewhat

unacquainted reader (e.g. the first and third readership discussed before) and is discussed with a focus on social relationships to preamble the use-case. In latter sections, I discuss the use-case, how they tie to the promises and pitfalls discussed before, and discuss how the interface of survey/digital traces addresses those pitfalls. I include tools, recommendations, and ideas to employ similar approaches, while I simultaneously reflect on my experiences and tools in this realm. I end this essay with three take-ways.

Digital footprints: computational trend or fad?

The introduction mentions a new ‘data lens’ because a metaphorical ‘telescope’ has been used to describe observing new, big data that capture social dynamics. Watts (2011), for instance, mentions that with the advent and access to new types of digitally accessible, online network information – social media data or other types of *online, digital footprint data* – ‘we [social scientists] have finally found our telescope’ (266). The claim is that as new data sources become available, accessible, and far-reaching, we ‘finally’ have a measurement tool to study human (relationship) behaviour at scale. In this essay, I define new data sources pragmatically: quantitative data sources that are *not* survey data, and, in the context of mixing methods in essay, which ought to *match with high recall* to survey data. One can then analyse these new data in conjunction with the available survey data. Many survey/new-data linkages consist of linking survey data with so-called *digital footprint data*. These are behavioural signals that respondents ‘leave’ in the digital world (mostly the internet) and that are logged automatically. Digital footprints are ‘likes’ on Twitter/X, relationships on Facebook, replies on Instagram, posts and messages on specific fora, WhatsApp conversations, and so on. Other ‘new’ quantitative data types are not always from the internet. For instance, register data may include tax records identifying individuals employed by the same organisation. These data might be fitting to study social relationships as well depending on one’s research question. Yet, this essay focuses on the interface between *survey data* and *online, digital footprints*, such as found in the use-case.

Computational social scientists are commonly interested in digital footprints capturing some social dynamic (e.g. ‘liking’ or ‘following’ on Instagram). Here, I propose a pragmatic definition of computational social science (as we do in Tolsma and Hofstra, n.d.): ‘Problem-driven, [quantitative social science], but with the empirical part specifically containing some form of digital footprint data and/or some new [i.e. not standard practice in social science] methodological technique’. This inclusive definition prevents a discussion of which specific methods or data are or are not computational social science. New computational social science approaches can increasingly integrate in ‘regular’ quantitative research methods, and this provides many opportunities for the study of social relationships and ethnic background. To analyse digital footprint data, one can first ‘webscrape’ such information through manual data collection, by automated collection through code, or through application programming interfaces of specific platforms.

Promises and pitfalls of digital footprints

Before I discuss the use-case, we need a short, general introduction where I summarise prior work that mentions some of the promises and pitfalls of digital footprint data.

Digital footprint data have been discussed in-depth elsewhere and with greater detail than here (e.g. Drouhot et al. 2023; Edelman et al. 2020; Golder and Macy 2014; Lazer et al. 2009, 2020; Lewis 2015; Stier et al. 2020). The summary of prior work helps contextualise them in the use-case-specific reflection later.

Common opportunities

Features of digital footprint data that provide research opportunities relate (among many others) to (a) social relationships, (b) behavioural signals, and (c) size. First, many digital footprints include a form of social interaction. For instance, ‘liking’ on Twitter/X is a directed social relationship between the liker (the one who ‘likes’) and likee (the one who receives a ‘like’) in a dyad.¹ Such information is a strength of digital footprint data; it is relatively easy to webscrape data of many social relationships and their interactions.² Respondents likely find it challenging to remember their thousandth, hundredth, or even fiftieth social relationship in surveys, whereas digital footprints often store those automatically (and longitudinally). This feature may be a key reason computational social scientists often study of social relationships with digital footprint data.

Second, webscraped digital footprints can capture behavioural signals sometimes difficult to gauge in questionnaires. Rare events such as illicit activities may be hard to measure through surveys, whereas webscraping from the (dark) web (e.g. Norbutas 2018) provides opportunities to study illicit activity. Some attitudes or opinions are prone to social desirability bias, whereas through Facebook signals, for instance on topics pertaining to migration and politics can be unobtrusively observed. Behavioural signals in digital footprints like emoji’s or music sharing likewise capture emotions or cultural tastes otherwise difficult to measure. Though not all behavioural or attitudinal signals directly proxy offline attitudes, nor are they all without social desirability challenges; still, digital footprint data can illuminate behavioural signals otherwise difficult to measure.

Third, digital footprint data can contain a lot of observations (into the millions). This can be an advantage and disadvantage – i.e. more data are not always better data. Yet, it might make it easier to observe relationships that have small true effect sizes that may otherwise be eclipsed by random variability (cf. Golder and Macy 2014, 132) or might make it less challenging to capture either hard to reach or very small populations or groups, which is particularly relevant to studying minority populations, including ethnic minorities and minorities among them. For instance, recently arrived asylum seekers in the Netherlands, are hard to reach, but perhaps by sampling Facebook groups in addition to on the ground observation (see Roblain, Mazzola, and Politi 2025), one can study their social interactions, and migration experiences.

Common pitfalls

Digital footprint data/methods have challenges (among others) pertaining to (a) data size and structure, (b) sampling, and (c) unobserved variables. These may apply to survey data too, yet they might be amplified in digital footprints. First, data size and structure are advantages (see above) but also pitfalls. As observations grow into the millions, data become harder to process and analyse. Digital footprint data are also often structured differently than standard data frames where rows are usually respondents and columns variables. Webscraped data is often stored in nested structures or contains

raw textual data.³ Additional manipulation is needed before ‘standard’ statistical packages can read and analyse such data, whereas standard methods curricula often do not include training to properly execute such tasks.

Second, scientists collecting and using digital footprints should carefully delineate target populations vis-à-vis their sampling frame and realised sample. It is easy to be impressed by (small but significant correlations in) the sheer data sizes (see Lewis 2015; McFarland and McFarland 2015). Imagine that a study considers three million careers in the Netherlands and how they relate to ethnic background by using web-scraped LinkedIn data. But what if only certain ethnic background groups and specific occupations use LinkedIn? And what if ethnic backgrounds and occupations correlate? This implies that sampling methods affects both the independent and dependent variable side, possibly biasing inference. Online platforms are often disproportionately adopted by specific groups and generalisability is therefore more often than not limited to the platform under consideration.

Third, digital footprints often do not contain (detailed) information on demographics. Yet, such information often contains variables of interest, almost by definition so in studies on ethnic inequalities, discrimination, and migration experiences. Imagine you webscraped reviews from AirBnB to study how Dutch majority members review their hosts in Marrakech, Morocco, and Sevilla, Spain. This may show how destination country versus gender dynamics in reviewing may (re)produce inequality in perceived trustworthiness of apartment rentals. Yet, how do we know which Dutch reviewers are men and women? How do we know whether they are Dutch majority members? These labels often do not come with digital footprints.

Surveys with digital footprints: best of both worlds?

Digital footprint data thus contain unique features, and enable, sometimes, innovative tests of hypotheses, and perhaps tests of entirely new ones. Yet some pitfalls negatively affect the analytical strength of those tests. Ideally, you would have a digital footprint data ‘telescope’ with great observational radius and resolution. You want to sample many groups without bias and a large number of observations to find relevant signals (radius), but proper sampling is one of the more common pitfalls in analysing digital footprint data. Additionally, you want a high level of detail (resolution) within observations, though that might be a challenge in digital footprints as we often do not know much about these individual observations (e.g. selectivity by privacy settings). Table 1 depicts a summary of these features. Digital footprint data often fall into the quadrants in the right-hand panel; often high *resolution* (signal of interest captured)

Table 1. Digital footprint observational ‘telescope’.

		Resolution	
		Low	High
Radius	Low	High N, but biased observations/few variables, few signals	High N, but biased observations/many variables, signals of interest
	High	High N, target population of interest/few variables, few signals	High N, target population of interest/many variables, signals of interest

with some sliding scale on the *radius* as to how much we know about who these individuals are and how they generalise to some general target population of interest. Radius can be referred to as a large number of cases ('big data', 'high N', etc.). Yet, given that digital footprints often contain many cases, I would commend a shift towards referring to a large radius as being High N, but with unbiased sampling from a target population – i.e. one needs to know more about these observations with respect to demographic detail.

How can we leverage digital footprints such that both resolution and radius are high and ameliorate some of the common pitfalls? This is where the main mixed-method approach of this essay arrives; by linking survey data with digital footprint data. The relevant use-case is the study of ethnic background and social relationships, and in that use-case some of these common pitfalls are directly (and pragmatically) confronted. I also use this case to sometimes make a larger argument for such a mixing of data and methods, yet, as I mentioned before: the study of social relationships remains inherently linked to issues surrounding ethnic background.

By mixing these two data types (surveys/digital footprints), studies have a larger chance to be categorised in the 'high/high' cell of Table 1 so as to leverage the 'best of both worlds'. Why so? A careful tradition of survey research often utilises decades of knowledge on sampling so as to enable generalising findings to well-described target populations. In short, issues of biased inference when you link carefully sampled individual survey records with digital footprint data may become less salient if you match with high recall or if it enables to explicitly account for sample selection bias. The discussion of the use-case will provide an example of this.

Additionally, survey data often contains a lot of detailed, demographic information on ethnic backgrounds, national origins, genders, socioeconomic statuses, parental information, self-identification, and so forth, whereas digital footprint data might not. Hence, digital footprint data that contains information on social relationships of focal survey respondents becomes a powerful mixed-methods approach. This implies a *substantial integration* of both data types (Woolley 2009), insofar that such a mixed-data/mixed-methods approach of quantitative/quantitative data allows to ask and answer new questions on, among other topics, ethnic background and social relationships.

So what does this look like in practice? With this mixing there are pitfalls that mirror those discussed before, but by providing in-depth information on the use-case I can address those, discuss what this mixing of data and methods looks like in practice and what they add substantively to existing knowledge on ethnic background and social relationships. Therefore, I reflect and expand on a (selective) set of tools, recommendations, and experiences that I distilled from and designed in my own prior research in the next section.

A use-case of adolescent online social relationships and ethnic background

Much of my prior and current work considers antecedents and consequences of social relationships in different domains (social media, work, academia, etc.). One dimension in which I studied those processes are social network dynamics along *ethnic background* lines among Dutch adolescents. Ethnic background here is broadly categorised as Dutch majority adolescents and minoritised groups such as Moroccan Dutch, Turkish Dutch, or Caribbean Dutch youth of whom at least one of their parents is born in Morocco, Turkey,

or the Dutch Caribbean (e.g. Hofstra 2017). When we set out to study social relationships and ethnic background, we found out that most work in this area considers so-called strong ties. For instance, adolescents' best friends or friendships in schools and school classes. These stronger social relationships are important, as adolescents spent a good amount of time in school and likely spend most time with good friends rather than with relationships beyond those core ties. This provided important insights into many meso- and micro-level mechanisms for the causes and consequences of (adolescents') closest social ties.

The collection of an individual's social relationships, however, include both strong and weaker ones. Essentially, the strength of a social relationship between two individuals is a function of time, emotional intensity, intimacy, and reciprocity (cf. Granovetter 1973, 1361). And as individuals have only a limited amount of each of these 'resources' to invest in their contacts, the majority of social relationships are weak rather than strong bonds. These weak ties are crucially important to study for numerous reasons. A classic argument is that weak ties are instrumental in diffusion dynamics and social capital extracted from social relationships. This is because weaker social relationships provide access to information, on job openings for instance, not received from stronger social relationships (Granovetter 1973). Weaker ties thus directly relate to labor market outcomes. They also affect health and well-being (Holt-Lunstad, Smith, and Layton 2010), and relate to out-group attitudes, as even weak ties reduce intergroup prejudice (Hofstra 2022).

Whereas ethnic background is often linked to stronger social relationships, weaker social relationships thus seem understudied in tandem with ethnic background. We developed the idea to make use of adolescents' contemporary social life which is increasingly happening *online*. Perhaps *online social relationships* of adolescents could measure weaker ties as these online networks capture hundreds of offline social ties (e.g. Mayer and Puller 2008)? In surveys it is often difficult to gain insight into social relationships other than the closest ones. Hence, we set up a study of adolescent online networks (i.e. digital footprints) and the role of ethnic background therein. We came up with a data collection plan to mix recently gathered survey data with the online social media networks of these youth. The reflection below moves from survey data to digital footprint data – i.e. the interface between surveys and digital footprints starts with the survey. I can imagine examples where this mixing is reversed, where a set of digital footprints is enriched with survey data. Much of the benefits of the linkage that I discuss below, however, will not apply in the latter case. I discuss three specific topics, and link those to a toolset, lessons learned, and recommendations. These mirror the common pitfalls described in the prior section posed by prior work and offers pragmatic solutions.

Data mixing and sampling

My involvement as a data collection coordinator for the Dutch part of the Children of Immigrant Longitudinal Survey in Four European Countries (CILS4EU; Kalter et al. 2022), provided opportunities to collect the online social media profiles and their social relationships of a representative set of Dutch adolescents. This rendered the interface between the survey and digital footprints easier to establish. We first achieved this interface with a manual approach: a set of student assistants tracking social media profiles and downloading those in HTML format (all in a protected environment, with

Careful data protection measures, and with the approval of an ethical committee) on the basis of names, educational institution, and place of residence that were reported in the survey by respondents when the respondent-provided URLs (nowadays that might be considered respondent 'data donation') in the survey did not prove sufficient.⁴ The URL was not sufficient in many cases, as respondents have difficulties remembering their URLs. Logically so, as these URLs were not always logically structured. (Purposively wrong URLs or ones not given were checked/validated through the survey-online data interface: by cross-validating names, places of residence, educational institution between both data sources.)

Herein lies one of the first difficulties; collecting digital footprint data does not always happen automatically, and in this case a number of student assistants helped. Specifically, the architecture of the platform – here, how URLs are generated – might dictate the data collection and sampling. This is the first step where the actual *mixing* of survey/digital footprints actually occurred: by identifying the online social media profiles of our respondents, we could extract their HTMLs and could link matches between survey respondents and Facebook profiles by their individual-level identifiers. We first also set out to collect information from Hyves, a former and popular Dutch social media website akin to Facebook. Yet we quickly turned to Facebook taking over as the dominant platform among youth right around that time (2014–2016). Herein lies the second pitfall and something that Salganik (2019) calls 'drifting': social media platforms are highly volatile in both their popularity (and design changes to their platforms). If I currently were to study social relationships, ethnic backgrounds, and adolescents in the Netherlands, Facebook would likely not be the prime empirical setting anymore.

That same pitfall that seemed a challenge, however, were of substantive interest. As the data collections from Hyves and Facebook were around the same time, we could study and contrast selection into late-stage membership in Hyves and early-adoption membership in Facebook. This use-case is a particular example of those selection dynamics described in prior work. Yet, what is new is how we addressed that selection. Those in minoritised groups – Moroccan Dutch and Turkish Dutch youth – adopt newer platforms later (Hofstra, Corten, and Van Tubergen 2016a, but see Jacobs and Spierings 2016, 119, on Twitter/X adoption). In other words, depending on a platform's age and stage, some ethnic background groups' social relationships may be over- or underrepresented in your data. In a direct follow-up study, we showed how the privacy settings of social media of Dutch majority members were less restrictive than those of minoritised groups; social relationships contacts were more often visible among majority than among minority members (Hofstra, Corten, and Van Tubergen 2016b). Because the use-case included multiple studies outlining these issues, we could use that knowledge to account for sample selection biases.

What does such sampling selection bias look like? Imagine survey data of 10,000 respondents. Yet, you can only track 6000 of those on some relevant platform that contains some sort of social network measure (linkage recall = 60%). And from those 6000, only 3000 of those observations allow you to see their social interaction patterns due to their privacy settings (measure-of-interest recall = 30%). In both of those steps, some groups are overrepresented (to make it more complex: not even exclusively by ethnic background). In sum, compounding sample selection bias creeps into online activity levels pertaining to both membership and privacy in social media networks (Hofstra, 2017, 23).

So what if we still want to study survey respondents' social relationships drawn from these social media profiles? An advantage of linking representative survey data to online network data is that you can gauge and then model data selectivity. In contrast, such a baseline for representativeness to some target population is often lacking when exclusively focusing on digital footprints. Say we aim to estimate a number of relevant social relationship outcomes; the number of friends on social media profiles and the percentage of co-ethnic friends across social media contacts of our CILS4EU respondents by a number of selected covariates (e.g. ethnic background, gender, education, and so forth). Given that we are aware of ethnic background biases in our survey/social media sample linkage, we first needed to ascertain that selectivity does not affect substantive findings. One way to do this is through sample selection corrections; by estimating two equations simultaneously, a substantive regression equation and a 'selection' equation where error terms of both are allowed to correlate (Heckman 1979). We found out that for the latter outcome (% co-ethnic in social media contacts) there was hardly any sampling bias, whereas for the former (number of social relationships on social media networks) there was. Specifically, the correlations of the number of social contacts with ethnic background move from not significantly deviating from zero in models that do not correct for sample selection bias to negatively and significantly deviating from zero in models that do account for sample selection bias (see Hofstra, Corten, and van Tubergen 2021). Because of the linkage between the survey and digital footprints, we were among the first to account for such biases. And note that this issue directly relate to the first common pitfall on sampling described before.

What does this imply? That people from minoritised groups who publicly show their profiles are likely to have more online social relationships than those who have closed profiles. See Figure 1 that depicts the loss of observations in each step (Survey → Profiles → Networks), ending with a biased set of observations: researchers observe as many minoritised persons with a small and large number of relationships, even though on average there are more among this group with smaller than larger set of social relationships but that difference remains unobserved. By not explicitly modelling and considering such possibilities, the biased inference would have implied that there are no differences in the number of contacts online between majority and minoritised groups. This is a crucial observation, as biased sample selections in mixed-methods, survey/digital footprint linkages thus vary across social relationship outcomes of interest and change the substantive interpretation of covariates related to ethnic background. Here, and in the next two sections, I summarise a toolset for mixing data/methods, the lessons I learned, and the recommendations I distilled from these experiences for future research.

- **Toolset:** Linkages of representative survey data by means of manual or API collection of digital footprint data on online social relationships – i.e. the focal mixed-method approach discussed in this contribution.
- **Lessons learned:** A series of papers point to sample selection biases in digital footprint data pertaining to social media membership and privacy that, if unaccounted for, can cause biases in research findings *and particularly so if one contrast majority and minoritised ethnic groups*. This does depend on the metric that is studied, and yet such seemingly trivial differences do call for scrutiny on sample bias in relation to social relationship outcomes.

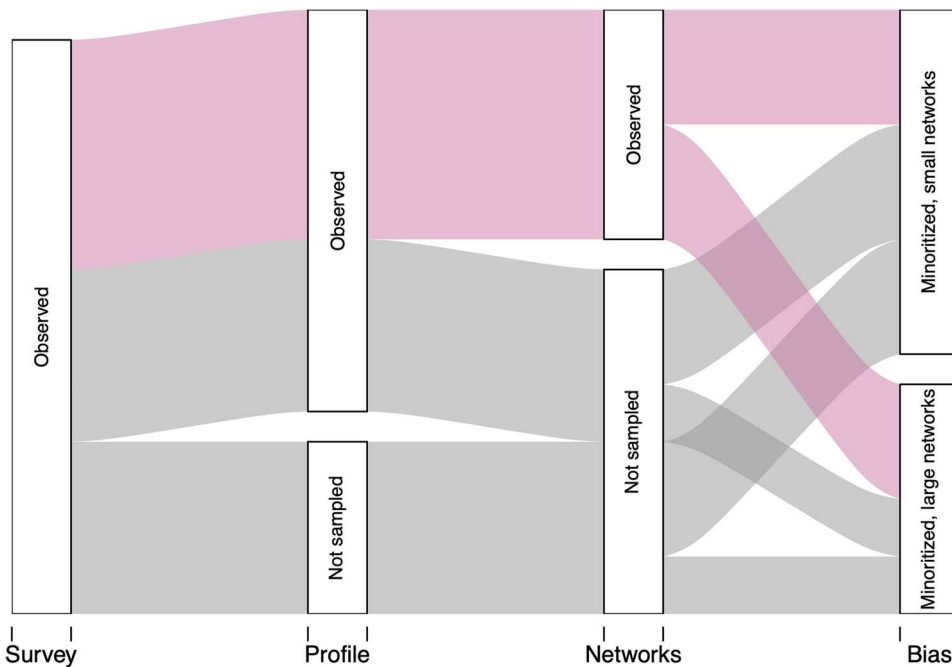


Figure 1. Introduction of bias in an observed sample of minoritised persons' networks. Notice the equal number of observations of minoritised persons' small and large networks, despite having smaller networks on average if one were to observe the entire set of minoritised persons' networks. This is due to the observed group that in each sampling is smaller, eventually observing in the final step an equal number of minoritised persons with small and large networks (the pink lines), whereas for the entire set minoritised persons have smaller networks (the grey lines).

- **Recommendations:** By carefully matching representative survey data to online social relationships data, one can explicitly account for group biases in the digital footprints through statistical models. This is a necessary precondition, and not doing so potentially biases our knowledge base on online social relationships, their ethnic composition, and how it relates to ethnic background and relationship consequences.

Data size and structure

The second example in the use-case directly relates to the second pitfall described earlier: the data sizes and structures are challenging to process and then analyse. Very crudely put, quantitative analyses often imply manipulation of a data frame (individuals as rows, variables as columns) containing survey answers so as to construct a set of variables and use those in some variation of regression analyses. When I was first confronted with thousands of text files in raw HTML format containing social media profiles, it took some trial and error to extract some substantive meaning from those data. How would you even process that number of files? Let alone extract meaning (e.g. ethnic backgrounds of Facebook friends, network size, etc.) from specific pieces of text hiding somewhere in between all that messy code? Such challenges can ultimately render scholars pretty

fluent in a number of programming languages. Yet, I cannot help but wonder how much faster prospective researchers and faculty will be able to analyse such new datatypes and extract meaning from them if we equip them with the necessary toolset and programming skills from the outset. This is not an argument for an overhaul of methodology courses, as the prior section points out how important those sampling and analysis competencies are, but it is a recommendation to extend curricula so as to equip prospective scholars with the correct toolset to be able to link and analyse survey with digital footprints (for the study of networks and ethnicity). Courses on basic-level programming, text analyses/natural language processing, and the analyses of social relationships may already come a long way.

There may be two interrelated and arguably positive outcomes if we do so. First, it allows for the development of a certain level of standardisation and maturity in the field writ large. This way every new researcher in the field does not need to be self-taught and in some ways re-invent the wheel into these approaches, with its increased risks of errors. This allows for better interstudy comparisons, shared knowledge bases, and reviewer/researcher knowledge on how to address both strengths and weaknesses of these approaches. Second, fields other than the social sciences are increasingly studying social dynamics on issues pertaining to ethnic background with complex digital footprint data. However, some of the methodological pitfalls described before might not get as much recognition as they deserve as these may be less salient in other fields. More importantly, scholars of social relationships and ethnic background, or of in- and out-group dynamics, have an enormous shared knowledge base on theories and questions and may be excellent referees as to what are important or innovative questions and hypotheses and research directions.

- **Toolset:** A baseline level of some programming language, some introduction on text analysis, social network analyses.
- **Lessons learned:** It takes a while before prospective researchers with a ‘standard’ methodology toolset in sociology/the social sciences may be able to make use of survey/digital footprint linkages to answer research questions on social relationship dynamics and ethnic background. The knowledge base of scholars studying social network dynamics pertaining to ethnic background is invaluable for the study of survey/digital footprints.
- **Recommendation:** Start at the basis in undergraduate social science to offer basic programming courses and move forward with that, possibly with more in-depth course foci (of choice) later on in programs. In the long run, this will increase the number of scholars that provide much-needed theoretical and methodological contributions to the field.

Unobserved variables

The last reflection on how the mixing of surveys and digital footprints enrich one another is the issue of unobservables. This relates to the third common pitfall and the interface of surveys and digital footprints helps addressing this issue. In survey data, researchers often introduce tailor-suited (batteries of) question to measure some aspect of social life (e.g. social trust, some metric of core social ties) or to get information on respondents’ (ethnic) background. The CILS4EU was no exception to this. Yet, when matched to

the Facebook profiles and the associated friend lists in those profiles, we did not necessarily know much about the persons in those friend lists, nor how these sets of social relationships structurally varied by adolescent ethnic background. One of our core interests was studying the ethnic (and gender) compositions of those weaker social ties. And as we conjectured that and wanted to test if weaker ties likely include many more opportunities to included co-ethnically *dissimilar* others in contrast to best friends, we needed to develop a way to measure ethnic background signals of those respondents' social relationships.

What we did know, were the *names* of respondents' online contacts. Prior work already argues how names function as perceived signals for gender, ethnic/racial, and social class identities (Lieberson 2000). As such, we developed a prediction method to attach ethnic (and gender) labels to online network friends on the basis of first and last names (I reflect on the challenges of such methods below). To summarise the method, it needs individual-level baseline 'training data' where individuals self-label their ethnic (and gender, or other) identities and data covering many names in a target population on what percentage of name carriers in that population have which ethnic background – e.g. the % of those with the surname 'Spierings' with Dutch parents, or Moroccan parents, or Turkish parents, and so on. By combining both and through a set of permutations, one can then find out what best predicts which percentage of surname carriers (in the population data) best relates to 'true' self-labeled ethnic background (in the training data). The same distilled 'threshold' percentages (in steps of .1%) can subsequently be used to assign ethnic background signals for names in online network data. We wrote a paper later on whether different labelling techniques and accounting for uncertainty in labelling matters (Hofstra and De Schipper 2018). Hence, the mixing, here, is about the data (survey and digital footprints) *and* the methods associated with that mix. This is where the mixing of survey and digital footprints was essential. The survey data containing self-reports on ethnic background were used as the 'training data'. The online social relationships of these adolescents were likely to include many adolescents as well. Therefore, using the survey data names and ethnic background identification may have increased correct assignments of ethnic background to these online social relationships as naming habits within cohorts correlate.

There are substantive and methodological challenges to such methods that are not always well described. Assigned ethnic background signals on the basis of names are by no means a substitute to ethnic identities that respondents themselves report in surveys or interviews. It also does not capture the multifaceted ways in which people sometimes identify with several ethnic backgrounds. And for gender, in particular, it often instils a dichotomy for men and women, whereas identification that falls outside or somewhere inbetween that binary is more difficult to capture with such methods. Such metrics are thus simplified signals of identities that may capture how individuals are perceived by other individuals. Another difficulty is that the smaller the ethnic background group and the less distinct their first and last names are from other groups the greater the chance of misclassification – most often being misclassified as the ethnic majority group. As for this latter point, depending on the outcome under consideration this may push findings towards the conservative side, as misclassifying ethnic minorities as ethnic majorities pulls their estimate-of-interest closer together. I would commend

more studies of self-identification versus name-based signals (but see Lockhart, King, and Munsch 2023) to increase our knowledge base on the validity of model coefficients.

Despite these challenges, this does remain one of the key ways to attach some form of demographic information to observations in digital footprints data, third party services have been developed – e.g. off-the-shelf packages such as ‘genderizeR’, ‘ethnicolr’, ‘gender’, ‘predictrace’ – to help researchers assign some gender/ethnic background to names (sometimes as a paid service). Yet, it is not always entirely clear how these services exactly operate, nor what baseline register databases they include for their predictions – e.g. whether it contains a representative sample of names in countries and their ethnic origins. This is potentially problematic. For instance, if one were to make use of a majority rule to assign ethnic backgrounds (e.g. if >50% of name carriers has Dutch parents, we assign ‘Dutch’), it might not work as well possibly even leading to incorrect inference. As discussed in Hofstra and de Schipper (2018), we show how, with the help of the survey/digital footprint interface, that these majority rules lead to biased inference. Hence, ideally developing an own method that is tailor-made to the use-case and using training and population data matching to that specific use-case, or at the very least combine several off-the-shelf methods such that you can see the percentage of overlapping labels per method (and to increase recall) is strongly suggested. This is key to ascertain the validity of such metrics for the question at hand. This references back to my earlier point on integrating these new data types and methods into standard curricula so as to get social scientists up to necessary speed.

- **Toolset:** Text- and name-based approaches to attach ethnic (and gender or educational) signals to find relevant meaning/entities in online social networks.
- **Lessons learned:** Assigning ethnic background signals on the basis of first and last names needs a careful consideration of training and population databases, its associated techniques, and multi-method comparisons. The interfaces between survey data and digital footprint data was key in this particular use-case.
- **Recommendation:** A renewed and more-nuanced focus on name signals pertaining to ethnic background, gender, also in combination with other socioeconomic signals that more carefully describes and considers advantages and disadvantages.

A particular use-case?

Throughout the use-case description and the set of used tools and analyses, the data requirements and access to specific data to achieve such a survey/digital footprint linkage on ethnic background and social relationships are high. One needs to know who respondents are and where their online identities reside to match individual survey records to online social networks. Ideally, one is practically involved in survey data collection so as to immediately integrate digital footprint data. And even in that ideal case, one might run into difficulties. Social media platforms change terms and policies and platform architecture constantly. In this use-case, this implies that doing follow-up studies might be difficult. In practice, the ideal case might even be difficult to achieve: hence it might be a ‘particular case’. One way to achieve the linkage, however, would be to add digital footprint data to existing secondary data – e.g. with initiatives to enrich and add information to already existing data and ongoing survey panels.⁵ That said, I have not observed many studies using the approach(es) I describe above to study ethnic

background differences in the number and composition of social relationships and its consequences. However, see the collection of papers described in Stier et al. (2020) who mention a range of related studies.

Conclusions

In this essay, I argued how the *mixing of* survey data and digital footprint data and the associated *mixing of* ‘traditional’ and computational methods benefits the study of ethnic background and social relationships. I reflected on empirical innovations, practical considerations, and opportunities and pitfalls of linking and analysing survey data with new quantitative online digital footprint data. As a use-case, I focused on the study of ethnic background vis-à-vis causes and consequences of social relationships. The essay therefore aligns with both goals of this Special Issue: providing insight into doing mixed methods as well as reflecting on its practices (Geurts, Davids, and Spierings 2025).

An inclusive definition of computational social science potentially draws in many researchers that may associate their research with that definition; *[p]roblem-driven, [quantitative social science], but with the empirical part specifically containing some form of digital footprint data and/or some new methodological technique*. Computational social scientists that study digital footprints meet opportunities and challenges that potentially benefit or undermine their work, as prior work summarises (e.g. Drouhot et al. 2023; Salganik 2019). In short, observing social interactions and their dynamics, in a large-scale way provides answers to novel questions, yet some data sizes and structure are difficult to manipulate, sampling bias might creep into the data, and the level of (demographic) detail in digital footprints might be thin.

By means of the use-case, I outlined how mixing survey data with digital footprints might ameliorate these issues. The use-case was a study of social relationships by ethnic background among Dutch adolescents (see Hofstra, 2017) that tested old hypotheses in novel ways and tested new hypotheses altogether. Three topics and common pitfalls need attention as they detail important tools and lessons learned. For instance, sample selection biases can be modelled when survey data are taken as a baseline to which digital footprint data are then matched. Additionally, data sizes and structures are difficult to manage with skillsets learned from traditional curricula. Finally, observational detail in digital footprints may be increased by using text-based metrics to find relevant signals. Yet, these methods come with challenges, particularly pertaining to ethnicity-related biases. It is crucial that researchers understand what happens ‘under the hood’ of such methods.

One important topic that I did not discuss much in this essay are issues of ethics and privacy. This does not directly touch upon the goal of the essay – i.e. a boots-on-the-ground reflection on linking survey data with digital footprints – and it would take another full-length essay to provide a nuanced perspective on those issues. Nuanced perspectives on these issues have been described carefully in, among others, Stier et al. (2020) or Boeschoten et al. (2022). In contrast to solely studying digital footprint data where informed consent is often practically impossible (given the data size and limited insights on individual cases), informed consent for linkages between surveys and digital footprints might be possible. For instance, through data donation practices by respondents.

Finally, I synthesise three main take-aways from the content of this essay. The first is a researcher's out-of-the-box capacity. The computational social science field in general, and linking survey with digital footprints specifically, calls for an out-of-the-box approach. In many cases, attempted linkages will not have been applied before. Perhaps new 'fuzzy' matching between the survey and digital footprint data is needed, or some probabilistic matching, or the textual data matched to the survey needs a non-standard social science method so as to extract meaning from it. In all cases, each of these challenges need some out-of-the-box approaches not very often seen in traditional survey approaches. This holds for reviewers and editors too (e.g. Spierings and Geurts 2025). These last two groups (justifiably) need some convincing from the authors' side as some of the tools are relatively unknown.

The second take-away is that decades of methodological rigour in quantitative survey analysis is invaluable to computational social science. Survey research can benefit from linked digital footprints to ask novel questions, whereas digital footprints may need sample selection corrections. 'Traditional' social scientists have a lot to bring to the table. Moreover, new methods are often seen as bringing something new and valuable, yet often cannot be leveraged without an existing knowledge base of the 'older' methods and/or theories.

The third and final take-away is that social scientists may need to increasingly engage with computational social science. The reason for this is not because it is in vogue, but because other fields (e.g. computer scientists) are employing computational social science methods and are answering social scientific questions. Yet they perhaps do so without the in-depth substantive expertise of social scientists or with only little reference to enormous social science literatures on specific topics. Social scientists have much to offer in particular in the research area of social relationships and ethnic background, for instance, with in-depth knowledge on nuanced categorisations of ethnic background groups.

Notes

1. A dyad in a social network is the smallest social structure: a pair of actors that are connected or not (in the case of binary social ties).
2. Webscraping means the process by which one collects data from the internet.
3. For instance, nested structures such as HTML (Hypertext Markup Language) – a text-based approach that dictates how web browsers should display text, images, or other elements on a webpage. Or it contains uncleaned textual phrases that need additional manipulation to analyze.
4. URL (Uniform Resource Locator) is an address on the Web.
5. The LISS panel in the Netherlands might be a good example of this: <https://www.centerdata.nl/en/liss-panel>.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

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

References

- Allport, G. 1954. *The Nature of Prejudice*. Boston: Addison-Wesley.
- Blau, P. M. 1977. "A Macrosociological Theory of Social Structure." *American Journal of Sociology* 83 (1): 26–54.
- Boeschoten, L., J. Ausloos, J. E. Möller, T. Araujo, and D. L. Oberski. 2022. "A Framework for Privacy Preserving Digital Trace Data Collection through Data Donation." *Computational Communication Research* 4 (2): 388–423.
- Bubritzki, S., F. Van Tubergen, J. Weesie, and S. Smith. 2018. "Ethnic Composition of the School Class and Interethnic Attitudes: A Multi-Group Perspective." *Journal of Ethnic and Migration Studies* 44 (3): 482–502.
- Drouhot, L. G., E. Deutschmann, C. V. Zuccotti, and E. Zagheni. 2023. "Computational Approaches to Migration and Integration Research: Promises and Challenges." *Journal of Ethnic and Migration Studies* 49 (2): 389–407.
- Edelmann, A., T. Wolff, D. Montagne, and C. A. Bail. 2020. "Computational Social Science and Sociology." *Annual Review of Sociology* 46:61–81.
- Geurts, N., T. Davids, and N. Spierings. 2025. "Two Peas in a Pod? How to Mix Methods in Ethnic and Migration Studies." *Journal of Ethnic and Migration Studies* 51 (12): 3041–3059. <https://doi.org/10.1080/1369183X.2025.2487740>.
- Golder, S. A., and M. W. Macy. 2014. "Digital Footprints: Opportunities and Challenges for Online Social Research." *Annual Review of Sociology* 40: 129–152.
- Granovetter, M. S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78 (6): 1360–1380.
- Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica: Journal of the Econometric Society* 47:153–161.
- Hofstra, B. 2017. "Online Social Networks: Essays on Membership, Privacy, and Structure." Doctoral diss., Utrecht University.
- Hofstra, B. 2022. "Interethnic Weak Ties Online and Out-Group Attitudes among Dutch Ethnic Majority Adolescents." *European Societies* 24 (4): 463–492.
- Hofstra, B., R. Corten, and F. Van Tubergen. 2016a. "Who was First on Facebook? Determinants of Early Adoption among Adolescents." *New Media & Society* 18 (10): 2340–2358.
- Hofstra, B., R. Corten, and F. Van Tubergen. 2016b. "Understanding the Privacy Behavior of Adolescents on Facebook: The Role of Peers, Popularity and Trust." *Computers in Human Behavior* 60:611–621.
- Hofstra, B., R. Corten, and F. van Tubergen. 2021. "Beyond the Core: Who has Larger Social Networks?" *Social Forces* 99 (3): 1274–1305.
- Hofstra, B., and N. C. de Schipper. 2018. "Predicting Ethnicity with First Names in Online Social Media Networks." *Big Data & Society* 5 (1): 2053951718761141. <https://doi.org/10.1177/2053951718761141>.
- Holt-Lunstad, J., T. B. Smith, and B. J. Layton. 2010. "Social Relationships and Mortality Risk: A Meta-Analytic Review." *PLoS Medicine* 7 (7): e1000316. <https://doi.org/10.1371/journal.pmed.1000316>.
- Jacobs, K., and N. Spierings. 2016. *Social Media, Parties, and Political Inequalities*, 45–76. Basingstoke: Palgrave Macmillan.
- Kalter, F., A. F. Heath, M. Hewstone, J. O. Jonsson, M. Kalmijn, I. Kogan, F. van Tubergen, et al. 2022. "Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU)- Full Version. Data File for on-Site Use." GESIS Data Archive, Cologne. ZA5353 Data file Version 3.3.0. <https://doi.org/10.4232/cils4eu.5353.3.3.0>.
- Lazer, D., A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, et al. 2009. "Social Science. Computational Social Science." *Science* 323 (5915): 721–723. <https://doi.org/10.1126/science.1167742>.
- Lazer, D. M., A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, et al. 2020. "Computational Social Science: Obstacles and Opportunities." *Science* 369 (6507): 1060–1062. <https://doi.org/10.1126/science.aaz8170>.

- Lewis, K. 2015. "Three Fallacies of Digital Footprints." *Big Data & Society* 2 (2): 2053951715602496. <https://doi.org/10.1177/2053951715602496>.
- Liebersohn, S. 2000. *A Matter of Taste: How Names, Fashions, and Culture Change*. New Haven: Yale University Press.
- Lockhart, J. W., M. M. King, and C. Munsch. 2023. "Name-based Demographic Inference and the Unequal Distribution of Misrecognition." *Nature Human Behaviour* 7: 1–12.
- Mayer, A., and S. L. Puller. 2008. "The Old Boy (and Girl) Network: Social Network Formation on University Campuses." *Journal of Public Economics* 92 (1–2): 329–347. <https://doi.org/10.1016/j.jpubeco.2007.09.001>.
- McFarland, D. A., and H. R. McFarland. 2015. "Big Data and the Danger of Being Precisely Mnaccurate." *Big Data & Society* 2 (2): 2053951715602495. <https://doi.org/10.1177/2053951715602495>.
- McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27 (1): 415–444.
- Norbutas, L. 2018. "Offline Constraints in Online Drug Marketplaces: An Exploratory Analysis of a Cryptomarket Trade Network." *International Journal of Drug Policy* 56:92–100. <https://doi.org/10.1016/j.drugpo.2018.03.016>.
- Roblain, A., A. Mazzola, and E. Politi. 2025. "A Multi-Level Mixed-Methods Research Design in Studying Localized Experiences of Asylum Seekers: Challenges and Lessons Learned." *Journal of Ethnic and Migration Studies* 51 (12): 3096–3116. <https://doi.org/10.1080/1369183X.2025.2487744>.
- Salganik, M. J. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.
- Spierings, N., and N. Geurts. 2025. "Mixed Methods, Mixed Feelings: A Review of Hurdles Faced and Vaulting Poles to Apply When Wanting to Do and Publish Mixed-Methods Research." *Journal of Ethnic and Migration Studies* 51 (12): 3170–3191. <https://doi.org/10.1080/1369183X.2025.2487748>.
- Stier, S., J. Breuer, P. Siegers, and K. Thorson. 2020. "Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field." *Social Science Computer Review* 38 (5): 503–516. <https://doi.org/10.1177/0894439319843669>.
- Tolsma, J., and B. Hofstra. n.d. Social Network Analysis for Social Scientists. Accessed February 23, 2023. <https://jochemtolsma.github.io/SNA-4-Social-Scientists/>.
- Vermeij, L., M. A. Van Duijn, and C. Baerveldt. 2009. "Ethnic Segregation in Context: Social Discrimination Among Native Dutch Pupils and Their Ethnic Minority Classmates." *Social Networks* 31 (4): 230–239. <https://doi.org/10.1016/j.socnet.2009.06.002>.
- Watts, D. J. 2011. *Everything is Obvious: Once You Know the Answer*. New York: Currency.
- Woolley, C. M. 2009. "Meeting the Mixed Methods Challenge of Integration in a Sociological Study of Structure and Agency." *Journal of Mixed Methods Research* 3 (1): 7–25.