Published in final edited form as:
Socius. 2021 ; 7: . doi:10.1177/23780231211029499.

Digital Traces of Sexualities:

Understanding the Salience of Sexual Identity through Disclosure on Social Media

Connor Gilroy[*], **Ridhi Kashyap**[†]

[*]Department of Sociology, University of Washington

[†]Department of Sociology, Leverhulme Centre for Demographic Science and Nuffield College, University of Oxford

Abstract

We analyze the expression of sexualities in the contemporary United States using data about disclosure on social media. Through the Facebook advertising platform, we collect aggregate counts encompassing 200 million Facebook users, 28% of whom disclose sexuality-related information. Stratifying by age, gender, and relationship status, we show how these attributes structure the propensity to disclose different sexual identities. We find a large generational difference; younger social media users share their sexualities at high rates, while for older cohorts marital status substitutes for sexual identity. Consistent with gendered expectations, women more often express a bisexual interest in men and women; men are more explicit about their heterosexuality. We interpret these variations in sexuality disclosure on social media to reflect the salience of sexual identity, intersected at times with availability. Our study contributes to the sociology of sexuality with a quantitative analysis, using novel digital data, of how sexuality is signaled socially.

Keywords

digital trace data; disclosure; sexual identity; sexual orientation; social media

Introduction

Digital social media hold the potential to expand quantitative knowledge about the disclosure of sexuality. Systematically understanding how people who are lesbian, gay, bisexual, or queer (LGBQ) express their orientations and identities poses a number of challenges, from the sheer smallness of the populations in question to the long history of stigma facing these groups. The benefits of new data about sexuality are not confined to sexual minorities, however. This same lens of identity and expression can also apply toward understanding heterosexuality. Indeed, heterosexual identity may take on heightened salience at a time when LGBQ people are more and more visible (Dean 2014). Here, we consider in tandem how heterosexual and LGBQ people express their sexualities online.

Sexuality is signaled socially. A sexual identity is not inherently visible or self-evident, and can be shared or withheld in different social settings. Whether a sexual identity is shared depends on its salience to a given social situation, or to a given individual’s self-identity (Doan and Mize 2020). Working against social disclosure, sexuality has long been treated as something stigmatized, shameful, secret, or simply private; for sexual minorities, concealing sexual identity may be a matter of safety (Goffman 1963). Yet sexual identity is also a core social and demographic trait. In a place like the contemporary United States, sexuality is a central part of people's identities that affects individual worldviews, life experiences, and trajectories (Schnabel 2018). While it is less often studied or measured than other socially salient attributes, like gender or race, survey data have shown marked generational shifts in the underlying distributions of different sexual identities, with young people increasingly identifying as sexual minorities (England, Mishel, and Caudillo 2016; Gates 2014, 2017; Jones 2021). These generational shifts have occurred against a backdrop of a growing digitalization of life and the life course, where digital social media are an active and salient space of social interaction and expression. In turn, the digitalization of our lives has also generated new data opportunities that offer a lens through which to understand this shifting terrain.

When people use social media platforms, they create "digital traces" of their activities and identities. These digital traces offer the unique opportunity to study how people manage and disclose information about themselves — including their sexualities — in a relevant social context, one that now forms an integral part of our lives. Our focus here is on the online disclosure of sexual identities as a social process, rather than an unmediated expression of underlying identities or "authentic selves" (Haimson and Hoffmann 2016). We do not consider these digital trace data as a straightforward measure of the demographic prevalence of different sexual identities, although we contextualize social disclosure using such measures from offline survey data sources.

We examine the disclosure of sexuality, and how it varies by other social attributes, using aggregate data from the population of Facebook users available from its advertising platform. Facebook is ideal for this purpose because it is the largest social media platform in the United States; 69% of all American adults have Facebook accounts (Auxier and Anderson 2021; Gramlich 2021). To obtain aggregate counts of users disclosing different sexualities, stratified by other characteristics we expect to be salient, we use the targeted advertising capabilities of the Facebook advertising platform. In this way, these data can be conceptualized as a type of "digital census" of the online population of Facebook users (Cesare et al. 2018). Existing work has used Facebook advertising data to model demographic and social indicators linked to migration (Alexander, Polimis, and Zagheni 2019; Zagheni, Weber, and Gummadi 2017), male fertility (Rampazzo et al. 2018), and internet access gender gaps (Fatehkia, Kashyap, and Weber 2018; Kashyap et al. 2020), and to understand the demographic biases of this online population by validating against "ground truth" measures (Alexander, Polimis, and Zagheni 2020; Ribeiro, Benevenuto, and Zagheni 2020). We contribute to this growing body of literature by using these data for understanding the expression of sexuality. In contrast to more demographic approaches however, we emphasize the interpretive opportunity offered by these data and consider the sociological implications of the patterns we find.

Our study contributes to the sociology and demography of sexuality by leveraging quantitative digital traces arising from a real social context as a unique lens for observing the social process of disclosure. These data can add to existing descriptive knowledge about sexual minorities, while also demonstrating the varying salience of heterosexual identity. We show at scale and in detail how the disclosure of these sexualities interacts with other social markers like age, gender, and marital status. Our results corroborate and extend findings from surveys and qualitative interviews, and suggest new directions for research on sexualities.

We proceed as follows. We begin by discussing how sexuality is disclosed and expressed in general and how people can express themselves on social media specifically. We then describe our approach to collecting data from Facebook's advertising platform, and to modeling that data in a way that conveys uncertainty. Next we present our results broken down by different characteristics, then reassemble them into a full picture of how sexuality is disclosed on Facebook in the United States. We close by considering both the implications of our findings and the potential futures of this sort of digital trace research.

Background

When people share information about their sexualities on social media, what might they mean to express? Social media platforms offer structured and unstructured ways for people to share their sexualities. On Facebook, this includes a profile field labeled “interested in,” for which the options are “men,” “women,” or “men and women.” But if a Facebook user's profile shows that they are, for instance, “interested in men and women,” it is not clear a priori which aspect of sexual orientation this statement might signal. Sexual orientation is multifaceted, encompassing behavior, attraction, and identity (Laumann et al. 1994). Based on prior research about social disclosure of sexuality (Doan and Mize 2020) and the systematic patterns we find, we argue that in this digital context “interested in” is primarily a marker of sexual identity. As we will show, for people in some social positions this profile field operates as an unambiguous signal of identity, and for others identity intersects with related considerations like sexual or romantic availability.

A fundamental distinction shaping social disclosure is that minority sexual identities are marked categories, while the majority identity of heterosexuality is taken for granted (Brekhus 1996; Zerubavel 2018). This means that people are generally presumed to be straight, and LGBQ people must disclose their non-heterosexuality in social situations by coming out. Because sexual minorities have long faced stigma, this repeated process of disclosing a non-normative sexuality in new contexts can be difficult and fraught (Goffman 1963; Orne 2011, 2013). Counterbalancing that stigma, factors like identity commitment positively mediate disclosure for sexual minorities (Doan and Mize 2020). Due to the emotionally charged, sometimes risky nature of outing oneself, and due to the potential strength of their identification with their sexual identities, we expect that LGBQ people would, on the whole, find this process of disclosure to be highly salient. In other words, they are accustomed to actively managing information and impressions about their sexualities.

Even though heterosexuality is unmarked, heterosexuals still also manage impressions about their sexualities. Heterosexuality is a privileged and normative social identity. Critical scholars of heterosexuality have analyzed this normative appeal of straight culture, showing how heterosexuals can lay claim to a heterosexual identity even when other aspects of their sexual orientation, like behavior, do not completely align (Budnick 2016; Carrillo and Hoffman 2018; Silva 2017a, 2017b; Ward 2015, 2020). The cultural meanings of heterosexuality also vary by gender, and this shapes the meaning of disclosure for heterosexual men and heterosexual women. For men, avoiding being perceived as gay forms a crucial part of heterosexual masculinity (Mishel, Bridges, and Caudillo 2021; Pascoe 2011; Ward 2015). Although homophobia is no longer central to all expressions of heterosexual masculinities (Dean 2014), men continue to experience pressure from peers to perform compulsory heterosexuality (Duckworth and Trautner 2019). Women, by contrast, are culturally afforded more flexibility in terms of both identity and behavior. Sociologists have argued that this is one reason for the rise in non-heterosexual identities and activities among young women (England, Mishel, and Caudillo 2016; Mishel et al. 2020). What this means is that explicit indications of sexuality have a different valence, depending on whether someone is part of a minority or the majority, and also depending on their age and gender. These varied potential meanings of online disclosure are reflected in the variations in disclosure by social categories (e.g. age, gender) that we will show in our results, which can be interpreted through an intersectional lens (Crenshaw 1989).

People express their sexual identities online because digital social media have become key venues for self-expression and impression management more generally. Where online spaces were once niche and separate social contexts, they increasingly overlap with offline social worlds (boyd 2014; Jurgenson 2011; Orne 2017). This context collapse means that LGBQ people must manage their sexual identities online as they do offline. As Duguay's (2016) interviews of queer British youth about their use of Facebook reveal, young LGBQ people are generally mindful of Facebook as a potential medium for disclosing sexualities and yet vary in how visibly they signal their sexualities, with about half of these queer young people using the "interested in" profile field to signal their identities. The remainder declined to use it, either out of privacy concerns or because they found it too rigidly binary to express queer identities. There is less prior research about the expression of heterosexual identities online, though Pascoe and Diefendorf (2018) show that men use homophobic language online to express a heterosexual style of masculinity, consistent with their offline behavior.

Our digital trace data provide a quantitative tool for studying the social disclosure and expression of sexualities, with unique strengths. We obtain a census of the complete population of US Facebook users, including those who do not disclose their sexualities. These comprehensive data equip us to examine the social process of identity disclosure in a novel way. Of course, online data are also constrained by their origin as found data (Salganik 2018). Technology companies decide how to structure the data that they collect, and researchers must work within that structure when using social media data. For instance, the way Facebook collects gender data and makes those data available constrains our analysis of sexualities. While Facebook allows individual users to select from a range of custom genders to appear on their profiles, Bivens (2017) shows that in internal databases userspecified genders are reduced down to three categories: women, men, and nonbinary people. Advertisers, subsequently, may only target advertisements toward women, toward men, or toward people of all genders (Bivens and Haimson 2016). Because advertisements cannot be targeted toward nonbinary people at all, nor toward more specific binary gender identities (for instance, only cisgender men, or only transgender men), data cannot be collected from

Facebook’s advertising platform for these groups. We therefore cannot analyze sexuality disclosure among these groups in this paper. This is one inherent limitation of found data, and we will return to the consequences of it in our conclusion.

In the following analysis, we investigate the disclosure of sexuality on Facebook, with the goal of assessing which characteristics shape the disclosure of sexual identities and how. We collect and model count data about US Facebook users, then aggregate predictions from the model to explore variations by age, gender, and relationship status. For some combinations of characteristics, rates of social disclosure are high across all sexual identities; in other cases identity disclosure is more conditional and varied.

Data

We collect data from Facebook using its advertising platform. Facebook links social, demographic, and behavioral information about individual Facebook users to categories that advertisers can use to select accounts that will be targeted with advertising. Potential advertisers on social media platforms such as Facebook can specify a desired audience for their ads based on targeting criteria, such as gender, age, geography, and other characteristics. For some characteristics (e.g. political preferences), these attributes are algorithmically-inferred categorizations, and the relation to concrete user-disclosed information is opaque. In other cases, including ours, specific user profile fields are directly linked to targeting options in the Facebook Ads Manager. Before an ad is actually launched, Facebook’s Ad Manager shows aggregate counts of audience sizes of the queried targeting attributes; this is the data we collect. Figure 1 shows both sides of this system, from user and advertiser perspectives. The figure highlights our main variable operationalizing sexuality, the “interested in” profile field and targeting option.

“Interested in” is an optional field that Facebook users may fill out or leave empty, stating whether they are interested in men, women, or both men and women. In a user profile, this field is found alongside gender and pronouns, under “basic information.” From an individual user perspective, the field has no explicit definition or description beyond the options presented. For advertisers, the field is listed in the “Demographics” section, under “Relationships” alongside “Relationships Status.” The advertiser-facing descriptions are quite open-ended. They read, for instance, “People who are interested in *Men and Women* for friendship, dating, relationships or networking.” Despite the vagueness of that definition, each of these contexts supports the assertion that “interested in” relates to sexual identity.

We use the Facebook advertising platform to collect aggregate data by intersecting the “interested in” field with other attributes. The example in Figure 1, Panel B shows how all advertisements must be stratified by age, gender, and geography in some fashion, and optionally by other characteristics as well. In the example, the audience for a hypothetical advertisement is men of age 20 in the United States who are interested in men. The estimated audience size for this group, the “potential reach” highlighted in the figure, is 25,000 people.

By querying this system, we systematically collect estimates for each possible combination of our variables, to build up a complete multiple-way contingency table. We do not obtain these estimates by manually querying the Ads Manager user interface. Instead, we automate data collection, retrieving estimates programmatically using the Facebook Marketing application programming interface (API). We access the API using the *facebookads* software development kit (SDK) for Python, which Facebook officially develops and releases for use by registered third-party developers. At the time of data collection in September 2017, this registration process was relatively open to anyone with a verified Facebook account.

Our primary data set consists of estimates for the number of adult Facebook users in the United States for every combination of sexuality, gender, relationship status, and age. These data represent counts of monthly active users of Facebook. From the perspective of social media users, aggregate data such as these present fewer risks and ethical concerns than individual-level data (Fiesler and Proferes 2018). Nevertheless, a few aspects of these data have implications for our analyses. First, all of the estimates returned from the Marketing API are rounded to two significant figures, regardless of the magnitude of the estimate. At the time of data collection, the minimum value that could be returned was 20 individuals. This minimum helps to preserve k-anonymity (Sweeney 2002), preventing re-identification and protecting individual privacy. Combined, the minimum value and rounding to two digits mean that the precise true number of users is more uncertain for larger categories. We minimze the impact of these features by collecting more stratified estimates and adopting a modeling strategy that accounts for some uncertainty.

Second, Facebook does not guarantee that these estimates correspond to external population values, nor that they are useful for anything beyond advertising. However, more than most social media platforms, Facebook attempts to enforce a principle of "authenticity," whereby each person has only a single user account (Haimson and Hoffmann 2016). Still, human users with multiple accounts, as well as non-human or "bot" users, remain potential sources for systematic error in our estimates.

Finally, we note that Facebook's advertising platform is in considerable flux. Categories or fields that can be accessed, how these categories are defined in the Ads Manager, as well as minimum counts of audience sizes change continually and with little notice. These changes sometimes occur in response to public critiques, especially with regard to political advertisements or discriminatory advertising practices (Goldman and Himel 2018), but in other cases have occurred with minimal explanation. Relevant to this study, in 2018, Facebook removed the "interested in" profile field as a targeting option for advertisements, without any indication of the specific factors or circumstances motivating this change. To a large extent this unpredictability is not unique to Facebook's advertising platform, but applies more broadly to digital trace data sources. These developments shape the opportunities for and limitations of research with these types of data sources, and we will return to the implications of these changes in the conclusion.

Methods

Our aim is to investigate how the disclosure of sexuality, and of specific sexual identities, on Facebook intersects with three other variables: age, gender, and relationship status. To examine these associations, we treat the stratified aggregate estimates we obtained as the outcome of a statistical regression model. All four variables, including the focal variable of sexuality disclosure, are covariates. Because the estimates are count data, we use a log-linear model with the associations between variables expressed as interaction terms (Agresti 2012). Such interaction terms are one quantitative and intersectional approach for examining intercategorical complexity (McCall 2005).

The primary motivation for constructing a statistical model from the data is to investigate which covariates can be viewed as independent from each other, conditional on the other variables. However, our best-fitting model ultimately includes all two-way and three-way interactions. If a simpler model had fit the data equally well, then it would have been possible to conceptually simplify the relationships between sexuality, gender, relationship status, and age. Instead, as we show in the Results section below, none of these factors can be disentangled from the others.

Two additional considerations motivate us to build and present a statistical model, rather than simply describe the underlying data. First, a model regularizes the data by smoothing out noisy variation. This helps address potential issues of data quality. For instance, Facebook users at certain ages are more numerous than expected, so we have more confidence in our regularized estimates in those cases than in the original numbers. In this sense, our model is an alternative to nonparametric methods such as locally estimated scatterplot smoothing (LOESS). Second, a model generates a range of potential outcomes, providing a measure of uncertainty. This allows us to assess the strength of our evidence. We can see how likely the differences we observe are to be substantively meaningful.

We include the four variables in our model as follows:

> *Sexuality*: Our measure of sexual identity, the “interested in” profile field, is structured by binary gender. Combining responses to this field with the user’s gender as reported to advertisers, we recode this measure from “interested in men” and “interested in women” to “interested in the same gender” and “interested in a different gender.” We leave the remaining two categories, “interested in men and women” and “not specified,” unaltered. “Not specified” is an informative category, representing users who do not disclose any sexuality.
>
> *Gender*: The gender categories available to advertisers are “women,” “men,” and “all,” but we present results only for women and men separately. Due to rounding, it is not possible to recover the numbers of nonbinary people or people with custom genders from the combined estimates of all Facebook users.
>
> *Relationship status*: We restrict our data to the most common relationship statuses: “single” (20%), “in a relationship” (11%), “married” (28%), and “not specified” (41%). Together, these statuses account for 95% of Facebook users. We exclude individuals who specify other relationship statuses, for instance, “engaged,”

“divorced,” or “it’s complicated.” This simplifies our model and avoids issues of data quality; for sexual minorities, especially at older ages, the numbers in those remaining categories are too small to produce stable or meaningful estimates.

Age: While our other variables are categorical, we treat age as a continuous variable. Age has a nonlinear association with our outcome estimates, and a fifth-order polynomial produces the best fit. Only Facebook users between the ages of 18 and 64 are included in our model. Facebook’s advertising platform groups all people aged 65 and above into a single “65+” category, which which is not directly compatible with a continuous operationalization of age; the simplest approach for modeling and interpretation is to exclude this 65+ category. While Facebook is open to anyone above the age of 13, we did not gather data for any users below age 18.

With these variables, we fit a Bayesian negative binomial model using the *rstanarm* package in R (Goodrich et al. 2020). A negative binomial model is more appropriate than a Poisson model because the counts are over-dispersed. We take a Bayesian approach to estimation, with weakly informative priors, for several reasons. Adopting a Bayesian framework facilitates the simulation of potential data, allowing us to aggregate and transform our results when we present different aspects of our findings, while preserving quantile-based interval measures of uncertainty. Bayesian models have the further advantage of straightforward extensibility for future work that might incorporate measurement error or other data sources. Related Bayesian modeling strategies have been fruitfully applied to demographic research with Facebook advertising data (Alexander, Polimis, and Zagheni 2020).

Because our model contains a large number of estimated coefficients, we do not present and interpret the parameters of the model individually. Instead, we present posterior medians and 95% posterior predictive intervals of estimated counts and proportions graphically. Importantly, posterior predictive intervals are wider than intervals based on the predicted means or expected values alone, making them a conservative way to examine evidence of differences (Goodrich et al. 2020). From this single underlying model, we aggregate posterior predictions to explore how disclosure rates for different sexual orientations vary by age, gender, relationship status, and finally by all of these characteristics together. Full information about our modeling approach, including comparisons to alternative models, is presented in the Appendix.

Results

We first describe the distribution of the outcome of interest, sexuality as measured by the “interested in” profile field on Facebook. Next, we show how the disclosure of sexuality is associated with the demographic attributes of age and gender. We then explore associations with the conceptually-connected variable relationship status. Our model reveals that these three factors all matter and intersect in complex ways, so we close by considering them all together. While the majority of the data are “missing,” the distribution of disclosure and non-disclosure is itself informatively patterned. Based on the “interested in” field, there are three possible sexual identities Facebook users might disclose. Whether individuals choose to disclose or not is informed by which sexual identity they hold, but we present results in two stages where relevant — first collapsed into the overall tendency to disclose any sexual identity, and then separated into specific identity categories.

More than a quarter of US Facebook users share information about their sexualities. In 2017, of the approximately 200 million women and men aged 18–64 who used Facebook in the United States, 56.3 million (28%) specified the genders in which they were interested, while 143 million (72%) did not disclose information about their sexuality in this way. Among Facebook users who disclose their sexualities, we identify 4 million as sexual minorities: 1.68 million (0.8%) interested in their own gender, and 2.21 million (1.1%) interested in both men and women. 52.4 million people (26.4%) indicate an exclusive interest in a different gender, and we construe these Facebook users as heterosexual. Figure 2 shows the rate of overall disclosure and what sexual identities are disclosed, aggregated from the model; Table 1 in the Appendix provides these and other descriptive estimates from the underlying data.

We now turn to how the distribution of disclosure and non-disclosure varies by other basic demographic characteristics. Figure 3 shows the disclosure of sexualities by age, collapsing all identities together; here, it is clear that age has a strong and nonlinear association with the propensity to disclose a sexuality. Only 20% of 18-year-olds disclose any sexuality, but this rises sharply to 50% of those in their mid-twenties. It falls again to 20% by age 40 and declines thereafter, to 10% by age 60.

Compared to age, gender alone has less of an association with the propensity to disclose any sexuality. Overall, women and men disclose their sexualities at similar rates, and this similarity in disclosure rates largely hold across the life course: While men may disclose at slightly higher rates in their late 20s and 30s, the predictive intervals for women and men overlap substantially, as shown in Figure 4.a. However, the specific sexualities that women and men disclose are not the same. Women are much more likely than men to be interested in both men and women, and they are slightly more likely to be interested in their own gender. Many more men, by contrast, explicitly express interest only in a different gender. Reframed in the language of sexual identity, women are more often openly bisexual or lesbian; men are less often bisexual or gay. Compared to women, men are more often explicitly heterosexual. Figure 4.b shows these differences, revealing that the overall magnitude of the gap between heterosexual women and men is much larger than the gender gaps among sexual minorities. Figure 4.c shows these gender differences by age. The differences are most pronounced among Facebook users at younger ages, most prominently the difference between bi women and bi men. The proportion of bi men is relatively constant across all age groups, while younger women identify as interested in men and women at more than twice the rate of older women, as well as at more than twice the rate of men.

Beyond these demographic traits, Facebook users have the option to share another piece of information that may be related to sexuality: their relationship status. As we show in Figure 5, people with different relationship statuses disclose sexualities at markedly different rates. On the one hand, the majority of Facebook users who declare themselves single also disclose a sexuality. On the other hand, a substantial number of users leave both fields unspecified. Those in relationships and those who are married disclose their sexualities at more intermediate levels.

However, the association between relationship status and sexuality is conditioned by age. Relationship statuses shape the disclosure of sexuality more for the old than the young. Whether single, in a relationship, or married, young Facebook users who specify any of these relationship statuses disclose their sexualities at nearly identical and high rates. For instance, the sexuality disclosure rate peaks among users who are age 25. Among these users, an estimated 58% of married individuals, 64% of individuals in relationships, and 68% of single individuals disclose their sexualities. The exact values differ somewhat, but the 95% posterior predictive intervals for the estimates all overlap. Figure 6 shows this overlap.

Disclosure rates diverge among older Facebook users. Single people at older ages continue to disclose at high rates, while disclosure rates among people in relationships fall slightly among older ages. Married users present a strong contrast: from age 40 and above, they disclose their sexualities at rates of 20% or below. Those who do not specify a relationship status are the least likely to specify a sexuality at all ages, though they too disclose at higher rates if they are young. Further examination reveals that these trends are largely driven by explicitly straight individuals, simply because the number of users at older ages who specify minority sexualities and relationship statuses such as “single” are quite small.

Age, gender, and relationship status interact to shape which Facebook users disclose a sexuality and which sexual identities they disclose. Young people and single people are more likely to disclose any sexuality, and women are more likely to be sexual minorities. At the same time, all these factors are conditionally dependent, which means they should be considered simultaneously. (For instance, women are more likely to be in relationships, where men are more likely to be single, which in turn shapes the distribution of relationship statuses across sexualities.) Figure 7 shows the full set of variables, with results presented in terms of counts rather than rates. Just like the proportions shown in previous figures, the counts are derived from a model that smooths the value of the estimates, rather than showing the underlying raw data (which are shown in the Appendix). This model is the best fit to the data; simplifying the model by dropping any interaction term, even three-way interactions, worsens model fit. Beyond the results we have already presented, Figure 7 shows various other demographic differences. For instance, while there are as many single gay men as lesbian women, there are more partnered lesbians than gay men. Across all disclosed sexual identities, relationship statuses also follow a life course pattern; the counts of single Facebook users have the youngest peak age, with older peaks for those in a relationship or married. Finally, because of the relatively high model-based uncertainty inherent in posterior predictive intervals (especially when estimated for 1-year age groups), even distributions that appear to overlap somewhat have a high probability of actually being distinct. This means that, for example, young men are more likely than young women to be explicitly heterosexual and to leave their relationship status unstated; conversely, young women are more likely than young men to state that they are in a relationship or married while not disclosing their sexual identity. Numerous comparisons of this sort are possible.

Discussion

In this section, we interpret and contextualize our findings about the patterns of disclosure in this social media population using prior survey data, qualitative research, and social theory. Because we found strong interactions between different social categories, we pay particular attention to how explanations might vary according to the intersections among those categories (Collins 2015; Crenshaw 1989; Nelson 2021). As we discuss disclosure among different groups in turn — LGB and straight, younger and older, partnered and single, men and women — we expand and complicate our preceding interpretations.

Deriving estimates from responses to the “interested in” field, we found 4 million lesbian, gay, or bisexual (LGB) Facebook users. For comparison, Gallup and the Williams Institute estimate that there are 10 million LGBT Americans over the age of 18 (Gates 2017). Of course, the categories used in the two data sources do not exactly correspond (sexual or romantic partner preference versus a combined measure of sexual orientation and gender identity), and not all American adults are Facebook users (Auxier and Anderson 2021; Gramlich 2021). Still, an underlying population of 10 million LGB(T) people would imply a higher disclosure rate (approximately 40%) for LGB people on Facebook than for Facebook users overall (28%). This number, in fact, corresponds to the result of Pew’s nationally representative survey of LGBT Americans, which found that four of ten LGBT adults overall — and 54% of those who use social networking sites — have disclosed their sexual identity on social media (Pew Research Center 2013). These numbers, along with the qualitative evidence from Duguay (2016), validate our belief that the “interested in” field is a strong indicator for sexual identity among LGB people. Compared to heterosexuals, LGB people disclose their sexual identities on Facebook at relatively high rates. From this comparatively high disclosure rate, we conclude that many LGB people do find sexuality a salient facet of identity to manage and disclose in the online social context of their Facebook profiles; for them, the salience of sexual identity outweighs potential stigma or risk of sharing their status as a sexual minority. Of course, a large fraction of LGB people on Facebook still do not disclose their identities through the “interested in” field.

Accounting for age adds important nuance to this interpretation. Several recent surveys show that young people identify as LGB at high and increasing rates (England, Mishel, and Caudillo 2016; Gates 2014, 2017; Jones 2021), and we find that young people are also disproportionately likely to disclose sexuality-related information on Facebook — which could drive part of the high LGB disclosure rate we previously discussed. What this generational divide in online disclosure might mean more broadly is not clear-cut, and the possibilities are not mutually exclusive. Young people could be willing to share their sexualities online because sexual identity is generally more salient to them, or because sexuality overall is less subject to stigma for younger cohorts. Sharing sexuality information online may also be a practical matter for those in a life stage where seeking sexual and romantic partners is common. By contrast, older cohorts on Facebook either may not find sexuality a salient axis of identity to express, may perceive expressing information related to sexuality to be taboo, or else may not find sharing their sexual identity to be practically relevant.

The youngest adult Facebook users, especially those under age 20, present a puzzle. Rather than disclose at the high rate of the cohort just ahead of them, they appear more similar to those in their mid-30s and beyond who share their sexual orientations on their profiles less often. However, we do not think 18- and 19-year-olds find sexuality irrelevant or inappropriate to disclose in social contexts. Some young people may find sexual identities expressed strictly in terms of binary gender to be unnecessarily limiting, preferring to identify as queer or pansexual rather than LGB or straight (Duguay 2016; Hammack et al. 2021) — though bisexual identity is even more common among the youngest generations (Jones 2021). More significantly, we suspect that, either due to privacy concerns or due to disengagement with Facebook as a social media platform, these youngest Facebook users are more generally reluctant to share personal information in their Facebook profiles. Where previous cohorts of youth may have struggled with “context collapse” online between their peers and their parents (boyd 2014), teenagers may have become increasingly adept, on average, at online impression management, presenting more curated facets of their identities to parents and other adults (Rafalow 2020).

Examining the interaction of age and relationship status helps disambiguate the practical and expressive aspects of disclosure: on Facebook, young people do not condition whether they disclose their sexuality based on whether they are presently single or partnered. Among both LGB and straight users, young people use the “interested in” field to express sexual identity, not to signal availability. Of course, the relationship statuses of young people may also be in greater flux, making them less likely to micromanage their “interested in” profile field accordingly.

Unlike people in their 20s, older cohorts’ willingness to signal their sexual identity is highly dependent on other characteristics. In particular, among many older users being married appears to serve as a substitute signal of sexual identity. Because the majority of marriages are heterosexual, marital status itself works to convey the unmarked status of heterosexuality. The fact that older single Facebook users disclose like their younger counterparts suggests that, for them, the “interested in” field signals not identity but availability.

Gender differences add yet another layer of complexity to our explanation above. For sexual minorities, the gender differences we observe in online disclosure are consistent with estimates of prevalence from nationally representative surveys. Specifically, a large number of those who express LGB sexual identities on Facebook are young women interested in both men and women, far outnumbering young bisexual men. This result echoes a key finding from the 2002–2013 National Survey of Family Growth (NSFG), where England, Mishel, and Caudillo (2016) show that the increasing number of young women identifying as bisexual is large enough to drive an observed overall increase in LGB identity over time. Previously, we observed a close correspondence between national surveys and our data about sexual minorities, due to the relatively high salience of sexual identity disclosure for LGB people online. In this case, we also conclude that our finding about the large number of young bisexual women reflects not only a sociological fact about disclosure, but also a demographic fact about the distribution of sexual identities in the population.

While young women are more likely to disclose a sexual minority identity than men, we find that at all ages men are more likely than women to explicitly signal that they are heterosexual. We argue that demographic prevalence is not the primary driver of this heterosexual gender disparity — the number of additional bi and lesbian women does not equal the magnitude of the gap between explicitly heterosexual men and explicitly heterosexual women. Instead, we attribute this disparity to the powerful interaction between heterosexuality and masculinity. As Pascoe (2011) has shown, a central goal of the public social performance of heterosexual masculinity is to enable straight boys and men to avoid being perceived or labeled as gay. Ward (2015) and Carrillo and Hoffman (2018) have gone further, showing that even when heterosexual men engage in homosexual or homoerotic behavior, they aim to hold onto the privileges associated with straight identity. Our finding that young men who explicitly indicate their interest in women also disproportionately disclose that they are single, or omit their relationship status, fits squarely into this framework of masculinities studies and critical heterosexualities studies. These men are using the "interested in" field as an opportunity to signal their heterosexual identity, and often their heterosexual availability. The audience for this social signaling is not only women who might be potential sexual or romantic partners, but also male members of their social worlds who might otherwise call their masculinity into question. (In fact, it could be primarily the latter.) "Interested in women" lays claim to straight privilege (Dean 2014) and meets potential peer pressure for publicly performing heterosexuality (Duckworth and Trautner 2019). This social media profile field affords a low effort way to signal heterosexuality, while potentially avoiding costs that may now come with being perceived as homophobic (Dean 2014). Altogether, the demands of heterosexual masculinity reasonably explain why straight men might be unusually motivated to disclose their sexualities.

We see the relative absence of openly heterosexual women through the flip side of the same lens. Both online and offline, women may face violent, sexist, or sexual harassment from men in public and private social spaces (Amundsen 2020; Nakamura 2019; Rubin, Blackwell, and Conley 2020). This risk of harassment shapes how and whether women, and heterosexual women in particular, disclose their sexualities. We find that a disproportionate number of women who leave their "interested in" information unspecified also report themselves to be in a relationship or married. While (explicitly) straight men choose the opposite responses (single or relationship status unspecified) to reinforce their heterosexuality and signal their availability, women are more likely to *avoid* signaling availability. This gap in sexuality disclosure between women and men is a uniquely heterosexual phenomenon, rather than purely a gendered one. As we discussed previously, LGB disclosure rates are generally high, there are slightly more lesbian women than gay men, and there are many more bi women than bi men. In a heterosexual context, the aggregate behavior of potentially-straight women strikes us not as an anomaly, but as a rational response to the behaviors of straight men.

To sum up, our findings should be understood in both demographic and sociological terms. Relatively speaking, for millions of LGB people and young people Facebook's "interested in" field represents sexual identity, and the salience of this identity appears to outweigh the potential stigma of disclosing a sexual minority identity. Consequently, for these subpopulations Facebook profile information provides a reasonable demographic proxy about sexual identity. Given the wide uptake of Facebook in the US, our findings correspond to population trends beyond the platform's users (Ribeiro, Benevenuto, and Zagheni 2020). By contrast, among heterosexual or older Facebook users, the "interested in" question is sociologically entangled with relationship statuses, revealing as much about gender dynamics and sexual or romantic availability as it does about sexuality or identity.

Conclusion

We have analyzed the disclosure of social information about sexuality in a digitally-mediated social context. Tens of millions of people in the United States have used Facebook's "interested in" profile field to disclose their sexualities. Digital trace data allow us to observe the disclosure behavior not only of LGB youth (Duguay 2016), for instance, but also of heterosexual people whose sexual identities may be taken for granted. And, while the design of the "interested in" field is outside of our control as researchers — as digital trace data generally are (Salganik 2018) — it is a vehicle for people to express their sexualities online. As the digitalization of our lives has generated new spaces for social expression and interaction, we have shown that careful measurement and analysis of digital traces can uncover some of the complexity of social lives and identities.

There are several important limitations to this work. First, the data are cross-sectional, descriptive, and aggregate. This means that we cannot discern the meanings of social disclosures at the individual level. Nor can we analytically disentangle age, period, and cohort effects from cross-sectional data alone. For instance, we might expect the mid-20s cohort to maintain their high sexual identity disclosure rate as they age and form partnerships, while 18-year-olds may continue to find Facebook profiles less salient as a vehicle for identity disclosure as they age. Accordingly, a second limitation is that our findings are specific to Facebook and the time period of our data collection (2017); they do not necessarily generalize to other platforms. General social trends like attitudinal changes toward privacy and data collection, or migration to new social media platforms with different architectures, might affect future disclosure patterns.

Finally, fundamental data limitations constrained which characteristics we were able to measure, limiting our ability to account for the full complexity of sexual identity. For instance, we could not distinguish asexual identities from non-disclosure. Asexual identities are increasingly visible and salient (Carroll 2020), but our study reinforces their invisibility. Nor could we account for other characteristics known to intersect with sexual orientation in the United States, like race (Silva and Evans 2020). Gender identity also interacts with sexual identity (James et al. 2016), but we could not separate out cisgender and transgender women and men; nonbinary people were excluded from our count data entirely (Bivens 2017; Bivens and Haimson 2016). This last limitation means that the present study reinforces a binary understanding of gender, furthering what Bivens (2017) labels *symbolic violence* against nonbinary people. This is a serious shortcoming, albeit one shared by many quantitative studies of sexual orientation and gender identity. Others have argued that this issue of "data violence" (Hoffmann 2018) may be intrinsic and insurmountable (Keyes 2019).

Nevertheless, our work gives us a foundation for assessing which directions for future research are promising, and which might be less viable. For instance, platform changes undermine the possibility of continuous monitoring for temporal comparison. The removal of the "interested in" field as an ad-targeting option as of early 2018, together with decreasing researcher access to APIs more generally in the "post-API age" (Freelon 2018), forecloses the possibility of conducting an ongoing digital census of sexuality data. As a potential solution, we believe platform-based surveys, whether microtargeted or aimed at characterizing a platform's entire user base, hold greater potential moving forward as a way to understand their user populations. These surveys could combine the best of digital and traditional methods (Salganik 2018); to survey sexual identity and its expression on social media, they would also need to incorporate best practices in survey design for sexual orientation and gender identity (Lagos and Compton 2021; Westbrook, Budnick, and Saperstein 2021).

Algorithmically-inferred categorizations are on the rise on social media platforms (Simpson and Semaan 2021), but we do not believe these will be as useful for studying sexual identity in digital contexts as explicitly user-disclosed information. For example, a social media user might be classified as liking "LGBTQ community." This could be a basis for targeted advertising, but it would tell us little about deliberate disclosure or conscious identification. Moreover, the attendant harms of algorithm ascription for LGBTQ+ people have already been noted in rising algorithm-centric platforms such as TikTok (Simpson and Semaan 2021). To study social disclosure, researchers might focus instead on how marked (i.e. LGBQA+) sexual identities are made visible in online social contexts, such as through explicit identification in symbols, emojis, or unstructured text. Initial promising work in this vein already exists (Andalibi 2019; Haimson and Veinot 2020).

Of equal importance are the experiences of the unmarked heterosexual majority. We interpret gendered disparities in heterosexual disclosure behavior in terms of theories of masculinity and critical heterosexuality studies. We believe that further work should consider interviews with heterosexual social media users to confirm or complicate this interpretation, with implications for understanding when heterosexuality becomes salient to disclose, and for promoting the wellbeing of straight women online.

We began this project optimistic about the potentials of digital trace data for extending the bounds of knowledge about sexuality. Even in the relatively short time since we collected the data presented here, the world has changed. Detailed data collection that fuels targeted advertising still undergirds much of the contemporary social internet, but public attitudes have shifted and it is unclear if this corporate model is viable in the long term. For researchers, this may mean disruptive changes to digital data sources and how publicly accessible they are, even as online and offline lives come to increasingly overlap for individuals. We believe that these changes make it all the more urgent and valuable, at any given moment, to document and understand online social life through digital trace data. And, though digital data ethics is an unsettled subject that will continue to evolve beyond this writing, we believe we have undertaken this work with minimal symbolic or material harm, especially toward sexual minorities and other marginalized social groups. We hope to have shown that even when data are partial and flawed, the effort to make meaning out of them remains worthwhile.

Acknowledgments

We would like to thank Emilio Zagheni for the initial inspiration to work with this type of digital trace data. We are indebted to the intellectual guidance of Kate Stovel, through extensive conversations and detailed comments which led to a stronger sociological understanding and framing for this paper. Mark Igra, Emma Mishel, Tristan Bridges, Hannah Curtis, Daiki Hiramori, as well as panelists and audiences at PAA and ASA, also provided valuable feedback on various iterations of this project. We thank two anonymous reviewers at Socius for their thoughtful comments. We would also like to gratefully acknowledge the 2017 Russell Sage Foundation Summer Institute in Computational Social Science for providing a vibrant space for the exchange of knowledge and ideas, including the genesis of this work.

Funding

Connor Gilroy's participation in this project was funded by a National Institutes of Health Big Data to Knowledge (BD2K) traineeship, grant n. 3T32HD007543-15S1, through the Center for Studies in Demography & Ecology at the University of Washington (CSDE), and also partially supported by a Shanahan Endowment Fellowship, a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant (P2C HD042828) to the Center for Studies in Demography & Ecology, a Training Grant (T32 HD101442-01) and the Schmidt Endowment. Ridhi Kashyap would like to acknowledge support from the Leverhulme Trust through the Leverhulme Centre for Demographic Science.

Appendix:: Descriptive statistics and model comparisons

Table 1 summarizes univariate descriptive statistics for the data presented in the main text of the paper. In aggregate, 199 million US Facebook users are included in the analysis, and 32.9 million more are included in the collected data but excluded from the analysis.

Table 1:

Descriptive statistics

	Count, in millions	Percent
Sexuality		
Interested in same gender	1.7	0.8%
Interested in men and women	2.2	1.1%
Interested in different gender	52.4	26.4%
Not specified	142.6	71.7%
Gender		
Women	103.4	52.0%
Men	95.5	48.0%
Relationship status		
Single	39.7	19.9%
In a relationship	21.6	10.8%
Married	55.5	27.9%
Not specified	82.3	41.3%
Age		
18–24	41.9	21.1%
25–34	56.3	28.3%
35–44	40.9	20.6%

	Count, in millions	Percent
45–54	33.8	17.0%
55–64	26.1	13.1%
Total included in analysis	198.9	100.0%
Total excluded from analysis	32.9	

The final data set comprises 1,504 cells of count values, stratified by sexuality (four possible values, including “not specified”), gender (two values), relationship status (four values, including “not specified”), and age (47 values). The cell values are counts rounded to two significant figures. The largest observed estimate in our data is 1.9 million, representing a true count value between 1,850,001 and 1,949,999 inclusive. The smallest observed estimate is 280, representing a value between 275 and 285 inclusive. This implies that stratifying Marketing API queries into smaller categories results in more precise estimates. Accordingly, all broader aggregations in the main text are derived from the original stratified estimates, rather than queried anew from the API.

As discussed in the main text, nonbinary Facebook users and users with other custom gender values cannot be targeted through the Facebook advertising platform, and so are excluded from our estimated counts. Due to rounding, it is not possible to recover the number of nonbinary Facebook users by subtracting the counts for men and women separately from the counts of all Facebook users. From the data we collected, we also exclude less common relationship statuses (“engaged,” “divorced,” “widowed,” and so on, comprising nine additional categories) and users age 65 and older from our model and presentation in the main text.

We make claims about an online population of user accounts in terms of human behavior. The main plausible systematic source of error in the data would be inauthentic user accounts, that is, accounts not corresponding to individual humans. We in fact observe unusually large numbers of users with limited account information—both sexuality and relationship status unspecified—at ages ending in seven (27, 37, and so on), corresponding to birth years in 1990, 1980, etc. For these observations we believe our model provides a more reliable estimate than the underlying data. We considered removing and interpolating those observations, but judged the smoothing induced by the polynomial age coefficients to be adequate.

To the four-way contingency table of the data, we fit a log-linear model (Agresti 2012). A log-linear model is a generalized linear model for count data, with an analytic focus on the interactions between variables. The goal is to determine which variables are associated with each other in the data, and which variables might be mutually, jointly, or conditionally independent. What we found, however, is that none of the relationships between sexuality (S), gender (G), relationship status (R), and age (A) can be simplified or removed. (See Table 2 for Bayesian Information Criterion (BIC) comparisons of different models.) Following Agresti (2012), we represent our best-fitting model with the shorthand notation *SGR, SGA, SRA, GRA*, which nests all lower-order interactions and main terms. This is Model 13 in Table 2, with the lowest BIC value, written in expanded notation in Eq. (2). A four-way interaction, $SGRA$, is unnecessary according to this measure of model fit. (Note that in a four-variable model, this interaction term would saturate the model — if age is encoded categorically — and cannot be fit with maximum likelihood estimation. Compare Model 15 to Model 16 in Table 2.) We compare candidate models fit with maximum likelihood estimation for speed and efficiency, and re-fit our final selected model in a Bayesian estimation framework.

We extend the basic log-linear modeling framework in two ways. First, we use a negative binomial distribution rather than a Poisson distribution to model the response variable y_{hijk}, the count in each cell, because the counts are overdispersed; the variance is larger than the mean. In the final Bayesian model, the reciprocal dispersion parameter $\frac{1}{\phi}$ is estimated to be 67.4 (95% credible interval 62.2–72.8); an estimate near 0 would have indicated no overdispersion. Second, while log-linear models typically include only categorical covariates, we model age as a continuous variable with nonlinear terms. This results in a more parsimonious model, and one that is preferred by measures of model fit such as the BIC (see Table 2 for further comparisons).

Table 2:

Model comparisons

Model	Notation	Description	Age encoding	Parameters	Deviance	BIC
M1	(Intercept only)	Null model	—	1	2001.51	36966.50
M2	S, G, R, A	No interactions	5th-order polynomial	13	1601.13	33281.04
M3	SG, SR, GR, SA, GA, RA	All two-way interactions	5th-order polynomial	63	1519.78	30322.38
M4	SGR, SA, GA, RA	One three-way interaction	5th-order polynomial	72	1520.09	30228.41
M5	GRA, SG, SR, SA	One three-way interaction	5th-order polynomial	78	1519.26	30323.39
M6	SGA, SR, GR, RA	One three-way interaction	5th-order polynomial	78	1518.58	30130.17
M7	SRA, SG, GR, GA	One three-way interaction	5th-order polynomial	108	1520.25	29851.11
M8	SGR, SGA, GRA	Three three-way interactions	5th-order polynomial	102	1519.16	29980.28
M9	SGR, SRA, GRA	Three three-way interactions	5th-order polynomial	132	1527.59	29538.77
M10	SGR, SGA, SRA	Three three-way interactions	5th-order polynomial	132	1535.08	29129.07
M11	SGA, SRA, GRA	Three three-way interactions	5th-order polynomial	138	1534.65	29287.21
M12	SGR, SGA, SRA, GRA	All three-way interactions	3rd-order polynomial	101	1516.17	30309.65
M13	SGR, SGA, SRA, GRA	All three-way interactions	5th-order polynomial	147	1551.24	28872.35
M14	SGR, SGA, SRA, GRA	All three-way interactions	Categorical	1090	1708.46	33763.13
M15	SGRA	Four-way interaction	5th-order polynomial	192	1559.93	28971.68

Model	Notation	Description	Age encoding	Parameters	Deviance	BIC
M16	SGRA	Saturated model	Categorical	1504	0.00	—

We then fit the following Bayesian model, using the weakly informative priors recommended by the Stan developers (Goodrich et al. 2020):

$$y_{hijk} \sim \text{ NegativeBinomial } (\mu, \phi) \tag{1}$$

$$\begin{aligned} \log(\mu_{hijk}) = {} & \lambda_0 + \lambda_h^S + \lambda_i^G + \lambda_j^R + \lambda_{hi}^{SG} + \lambda_{hj}^{SR} + \lambda_{ij}^{GR} + \lambda_{hij}^{SGR} + \\ & (\lambda_1^A + \lambda_{h1}^{SA} + \lambda_{hi1}^{SGA} + \lambda_{hj1}^{SRA} + \lambda_{ij1}^{GRA})x_k + \\ & (\lambda_2^A + \lambda_{h2}^{SA} + \lambda_{hi2}^{SGA} + \lambda_{hj2}^{SRA} + \lambda_{ij2}^{GRA})x_k^2 + \\ & (\lambda_3^A + \lambda_{h3}^{SA} + \lambda_{hi3}^{SGA} + \lambda_{hj3}^{SRA} + \lambda_{ij3}^{GRA})x_k^3 + \\ & (\lambda_4^A + \lambda_{h4}^{SA} + \lambda_{hi4}^{SGA} + \lambda_{hj4}^{SRA} + \lambda_{ij4}^{GRA})x_k^4 + \\ & (\lambda_5^A + \lambda_{h5}^{SA} + \lambda_{hi5}^{SGA} + \lambda_{hj5}^{SRA} + \lambda_{ij5}^{GRA})x_k^5 \end{aligned} \tag{2}$$

$$\lambda_0 \sim \text{Normal}(0, 10) \tag{3}$$

$$\lambda_{(h)(i)(j)(n)} \sim \text{Normal}(0, 2.5/s_x) \tag{4}$$

$$\frac{1}{\phi} \sim \text{ Exponential}(1) \tag{5}$$

λ_0 is the intercept, and the other λ coefficients represent indicators for categories and their interactions. Parameters for reference categories are constrained to equal 0:

$$\lambda_H^S = \lambda_I^G = \ldots = \lambda_{HI}^{SG} = \ldots = \lambda_{HIJ}^{SGR} = \ldots = 0$$

The scale parameter for each coefficient $\lambda_{(h)(i)(j)(n)}$ is rescaled by dividing by the standard deviation of the the corresponding centered covariate s_x. The polynomial terms are orthogonal polynomials.

To fit this model, we draw samples from the posterior distribution with a Markov chain Monte Carlo (MCMC) method, specifically the No U-Turn Sampler (NUTS), using the rstanarm R package (Goodrich et al. 2020). The resulting fit consists of 4,000 draws from four Markov chains, each run for 2000 iterations; the first half of each chain is discarded as a warmup. Model diagnostics indicate acceptable convergence and model fit. $\widehat{R}$ values for every parameter are below 1.01 (the largest is 1.003); parameter effective sample sizes are generally above 0.5 (i.e. 2000 effective samples). According to the pareto-k diagnostic, the model fits 11 observations poorly, with k > 0.7 (out of 1,504 observations). A further 36 observations have k values between 0.5 and 0.7. These observations are at the tails of the age distribution, especially the upper tail, suggesting that the polynomial specification for age in the model may be misspecified for ages very close to 18 or 64. For computational efficiency, we do not replicate every comparison in Table 2, in part because Gelman et al. (2020) suggest it is not necessary to estimate models already known to fit poorly with MCMC. However, select Bayesian model comparisons with PSIS-LOO and 10-fold cross-validation confirm that this model is the best fit among the models under consideration.

To present model results, we combine the samples from the posterior distribution. We summarize the predicted value for each combination of variables with the posterior median using the *tidybayes* R package (Kay 2020). We use the posterior predictive distribution to convey uncertainty, through 95% quantile-based posterior predictive intervals. Compared to credible intervals, posterior predictive intervals incorporate the additional uncertainty from the likelihood of the generative model; Lynch and Bartlett (2019) among others discuss the merits of this approach. To present higher-level summaries, we simply aggregate the samples before calculating the point estimates and intervals. Where appropriate, we transform these point and interval values from counts to proportions by dividing by group sums. To convey the fit of our model graphically, Figure 8 replicates Figure 7 from the main text, and additionally shows the actual estimates in the data alongside the model predictions.

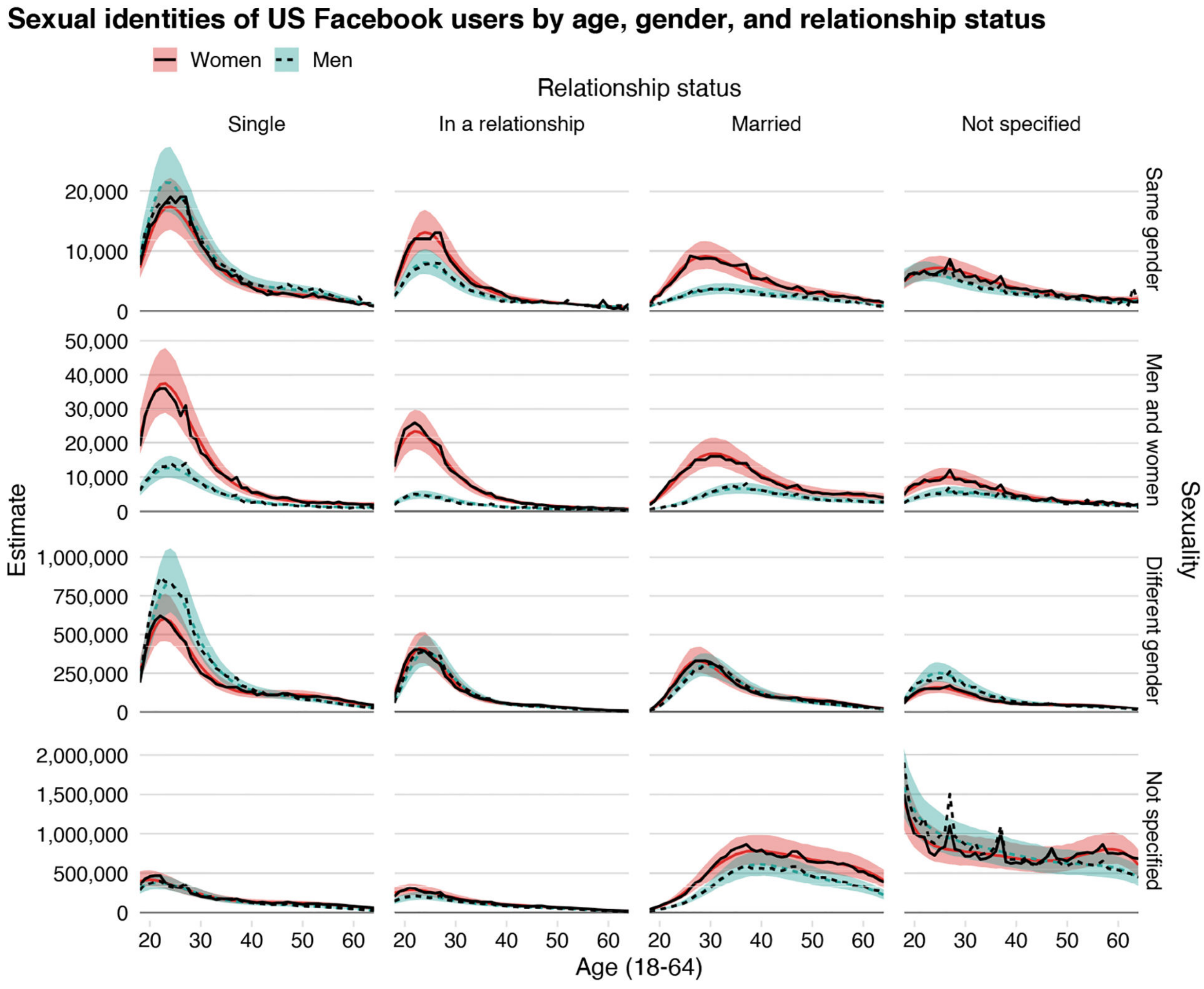


Figure 8:
Underlying data estimates of US Facebook users of each sexual identity, relationship status, age, and gender (black lines), overlaid on full model estimates from Figure 7. *Note*: Data were collected in September 2017 from the Facebook Marketing API. For model estimates, posterior medians and 95% posterior predictive intervals are shown.

All research code for data collection, data processing, statistical modeling, and visualization is available at https://github.com/ccgilroy/digital-traces-sexualities.

References

Agresti Alan. 2012. Categorical Data Analysis. Somerset: Wiley.

Alexander Monica, Polimis Kivan, and Zagheni Emilio. 2019. "The Impact of Hurricane Maria on Out-Migration from Puerto Rico: Evidence from Facebook Data." Population and Development Review 45(3):617–30. doi: 10.1111/padr.12289.

Alexander Monica, Polimis Kivan, and Zagheni Emilio. 2020. "Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States." Population Research and Policy Review. doi: 10.1007/s11113-020-09599-3.

Amundsen Rikke. 2020. "'A Male Dominance Kind of Vibe': Approaching Unsolicited Dick Pics as Sexism." New Media & Society. doi: 10.1177/1461444820907025.

Andalibi Nazanin. 2019. "What Happens After Disclosing Stigmatized Experiences on Identified Social Media: Individual, Dyadic, and Social/Network Outcomes." Pp. 1–15 in Proceedings of the 2019 CHI conference on human factors in computing systems, CHI '19. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3290605.3300367.

Auxier Brooke, and Anderson Monica. 2021. "Social Media Use in 2021." Retrieved June 3, 2021 (https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/).

Bivens Rena. 2017. "The Gender Binary Will Not Be Deprogrammed: Ten Years of Coding Gender on Facebook." New Media & Society 19(6):880–98. doi: 10.1177/1461444815621527.

Bivens Rena, and Haimson Oliver L.. 2016. "Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers." Social Media + Society 2(4):1–12. doi: 10.1177/2056305116672486.

boyd danah. 2014. It's Complicated: The Social Lives of Networked Teens. New Haven: Yale University Press.

Brekhus Wayne. 1996. "Social Marking and the Mental Coloring of Identity: Sexual Identity Construction and Maintenance in the United States." Sociological Forum 11(3):497–522. doi: 10.1007/bf02408390.

Budnick Jamie. 2016. "'Straight Girls Kissing?' Understanding Same-Gender Sexuality Beyond the Elite College Campus." Gender & Society 30(5):745–68. doi: 10.1177/0891243216657511.

Carrillo Héctor, and Hoffman Amanda. 2018. "'Straight with a Pinch of Bi': The Construction of Heterosexuality as an Elastic Category Among Adult US Men." Sexualities 21(1–2):90–108. doi: 10.1177/1363460716678561.

Carroll Megan. 2020. "What Can Asexuality Offer Sociology? Insights from the 2017 Asexual Community Census." doi: 10.31235/osf.io/bh7t3.

Cesare Nina, Lee Hedwig, McCormick Tyler, Spiro Emma, and Zagheni Emilio. 2018. "Promises and Pitfalls of Using Digital Traces for Demographic Research." Demography 55(5):1979–99. doi: 10.1007/s13524-018-0715-2. [PubMed: 30276667]

Collins Patricia Hill. 2015. "Intersectionality's Definitional Dilemmas." Annual Review of Sociology 41(1):1–20. doi: 10.1146/annurev-soc-073014-112142.

Crenshaw Kimberlé. 1989. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." University of Chicago Legal Forum 1989(1):139–67.

Dean James Joseph. 2014. Straights: Heterosexuality in Post-Closeted Culture. New York: NYU Press.

Doan Long, and Mize Trenton D.. 2020. “Sexual Identity Disclosure Among Lesbian, Gay, and Bisexual Individuals.” Sociological Science 7:504–27. doi: 10.15195/v7.a21.

Duckworth Kiera D., and Trautner Mary Nell. 2019. “Gender Goals: Defining Masculinity and Navigating Peer Pressure to Engage in Sexual Activity.” Gender & Society 33(5):795–817. doi: 10.1177/0891243219863031.

Duguay Stefanie. 2016. “‘He Has a Way Gayer Facebook Than I Do’: Investigating Sexual Identity Disclosure and Context Collapse on a Social Networking Site.” New Media & Society 18(6):891–907. doi: 10.1177/1461444814549930.

England Paula, Mishel Emma, and Caudillo Monica L.. 2016. “Increases in Sex with Same-Sex Partners and Bisexual Identity Across Cohorts of Women (but Not Men).” Sociological Science 3:951–70. doi: 10.15195/v3.a42.

Fatehkia Masoomali, Kashyap Ridhi, and Weber Ingmar. 2018. “Using Facebook Ad Data to Track the Global Digital Gender Gap.” World Development 107:189–209. doi: 10.1016/j.worlddev.2018.03.007.

Fiesler Casey, and Proferes Nicholas. 2018. “‘Participant’ Perceptions of Twitter Research Ethics.” Social Media + Society 4(1):2056305118763366. doi: 10.1177/2056305118763366.

Freelon Deen. 2018. “Computational Research in the Post-API Age.” Political Communication 35(4):665–68. doi: 10.1080/10584609.2018.1477506.

Gates Gary J. 2014. LGBT Demographics: Comparisons Among Population-Based Surveys. Williams Institute, UCLA School of Law.

Gates Gary J. 2017. “In US, More Adults Identifying as LGBT.” Retrieved April 29, 2021 (https://news.gallup.com/poll/201731/lgbt-identification-rises.aspx).

Gelman Andrew, Vehtari Aki, Simpson Daniel, Margossian Charles C., Carpenter Bob, Yao Yuling, Kennedy Lauren, Gabry Jonah, Bürkner Paul-Christian, and Modrák Martin. 2020. “Bayesian Workflow.” Retrieved April 29, 2021 (http://arxiv.org/abs/2011.01808).

Goffman Erving. 1963. Stigma: Notes on the Management of Spoiled Identity. Englewood Cliffs, N.J.: Prentice-Hall.

Goldman Rob, and Himel Alex. 2018. “Making Ads and Pages More Transparent.” Retrieved April 29, 2021 (https://newsroom.fb.com/news/2018/04/transparent-ads-andpages/).

Goodrich Ben, Gabry Jonah, Ali Imad, and Brilleman Sam. 2020. “rstanarm: Bayesian Applied Regression Modeling via Stan.” Retrieved April 29, 2021 (https://mc-stan.org/rstanarm).

Gramlich John. 2021. “10 Facts about Americans and Facebook.” Retrieved June 3, 2021 (https://www.pewresearch.org/fact-tank/2021/06/01/facts-about-americans-andfacebook/).

Haimson Oliver L., and Hoffmann Anna Lauren. 2016. “Constructing and Enforcing ”authentic” Identity Online: Facebook, Real Names, and Non-Normative Identities.” First Monday 21(6). doi: 10.5210/fm.v21i6.6791.

Haimson Oliver L., and Veinot Tiffany C.. 2020. “Coming Out to Doctors, Coming Out to ‘Everyone’: Understanding the Average Sequence of Transgender Identity Disclosures Using Social Media Data.” Transgender Health 5(3):158–65. doi: 10.1089/trgh.2019.0045. [PubMed: 32923666]

Hammack Phillip L., Hughes Sam D., Atwood Julianne M., Cohen Elliot M., and Clark Richard C.. 2021. “Gender and Sexual Identity in Adolescence: A Mixed-Methods Study of Labeling in Diverse Community Settings.” Journal of Adolescent Research. doi: 10.1177/07435584211000315.

Hoffmann Anna Lauren. 2018. “Data Violence and How Bad Engineering Choices Can Damage Society.” Retrieved April 29, 2021 (https://medium.com/s/story/data-violenceand-how-bad-engineering-choices-can-damage-society-39e44150e1d4).

James Sandy, Herman Jody, Rankin Susan, Keisling Mara, Mottet Lisa, and Anafi Ma’ayan. 2016. The Report of the 2015 US Transgender Survey. National Center for Transgender Equality.

Jones Jeffrey M. 2021. “LGBT Identification Rises to 5.6% in Latest U.S. Estimate.” Retrieved April 29, 2021 (https://news.gallup.com/poll/329708/lgbt-identification-riseslatest-estimate.aspx).

Jurgenson Nathan. 2011. “Digital Dualism Versus Augmented Reality.” Retrieved April 29, 2021 (https://thesocietypages.org/cyborgology/2011/02/24/digital-dualism-versusaugmented-reality/).

Kashyap Ridhi, Fatehkia Masoomali, Al Tamime Reham, and Weber Ingmar. 2020. “Monitoring Global Digital Gender Inequality Using the Online Populations of Facebook and Google.” Demographic Research 43(27):779–816. doi: 10.4054/DemRes.2020.43.27.

Kay Matthew. 2020. “tidybayes: Tidy Data and Geoms for Bayesian Models.”

Keyes Os. 2019. “Counting the Countless.” Real Life, 4 8. Retrieved April 29, 2021 (https://reallifemag.com/counting-the-countless/).

Lagos Danya, and Compton D Lane. 2021. “Evaluating the Use of a Two-Step Gender Identity Measure in the 2018 General Social Survey.” Demography 58(2):763–72. doi: 10.1215/00703370-8976151. [PubMed: 33834217]

Laumann Edward O., Gagnon John H., Michael Robert T., and Michaels Stuart. 1994. The Social Organization of Sexuality: Sexual Practices in the United States. Chicago: University of Chicago Press.

Lynch Scott M., and Bartlett Bryce. 2019. “Bayesian Statistics in Sociology: Past, Present, and Future.” Annual Review of Sociology 45(1):47–68. doi: 10.1146/annurev-soc-073018022457.

McCall Leslie. 2005. “The Complexity of Intersectionality.” Signs: Journal of Women in Culture and Society 30(3):1771–1800. doi: 10.1086/426800.

Mishel Emma, Bridges Tristan, and Caudillo Mónica L.. 2021. “Google, Tell Me. Is He Gay? Masculinity, Heterosexuality, and Gendered Anxieties in Google Search Queries about Sexuality.” Sociological Perspectives. doi: 10.1177/07311214211001906.

Mishel Emma, England Paula, Ford Jessie, and Caudillo Mónica L.. 2020. “Cohort Increases In Sex With Same-Sex Partners: Do Trends Vary by Gender, Race, and Class?” Gender & Society 34(2):178–209. doi: 10.1177/0891243219897062.

Nakamura Lisa. 2019. “Gender and Race in the Gaming World.” in Society and the Internet: How Networks of Information and Communication Are Changing Our Lives, edited by Graham M and Dutton WH. Oxford: University Press.

Nelson Laura K. 2021. “Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South.” Poetics 101539. doi: 10.1016/j.poetic.2021.101539.

Orne Jason. 2011. “‘You Will Always Have to “Out” Yourself’: Reconsidering Coming Out Through Strategic Outness.” Sexualities 14(6):681–703. doi: 10.1177/1363460711420462.

Orne Jason. 2013. “Queers in the Line of Fire: Goffman’s Stigma Revisited.” The Sociological Quarterly 54(2):229–53. doi: 10.1111/tsq.12001.

Orne Jason. 2017. Boystown: Sex and Community in Chicago. Chicago ; London: University of Chicago Press.

Pascoe CJ 2011. Dude, You’re a Fag: Masculinity and Sexuality in High School. University of California Press.

Pascoe CJ, and Diefendorf Sarah. 2018. “No Homo: Gendered Dimensions of Homophobic Epithets Online.” Sex Roles 1–14. doi: 10.1007/s11199-018-0926-4.

Pew Research Center. 2013. “A Survey of LGBT Americans.” Retrieved April 29, 2021 (https://www.pewresearch.org/social-trends/2013/06/13/a-survey-of-lgbt-americans/).

Rafalow Matthew H. 2020. Digital Divisions. University of Chicago Press.

Rampazzo Francesco, Zagheni Emilio, Weber Ingmar, Testa Maria Rita, and Billari Francesco. 2018. “Mater Certa Est, Pater Numquam: What Can Facebook Advertising Data Tell Us about Male Fertility Rates?” Proceedings of the International AAAI Conference on Web and Social Media 12(1, 1).

Ribeiro Filipe N., Benevenuto Fabrício, and Zagheni Emilio. 2020. “How Biased Is the Population of Facebook Users? Comparing the Demographics of Facebook Users with Census Data to Generate Correction Factors.” Pp. 325–34 in 12th ACM Conference on Web Science. Southampton United Kingdom: ACM. doi: 10.1145/3394231.3397923.

Rubin Jennifer D., Blackwell Lindsay, and Conley Terri D.. 2020. “Fragile Masculinity: Men, Gender, and Online Harassment.” Pp. 1–14 in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI ‘20. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3313831.3376645.

Salganik Matthew J. 2018. Bit by Bit: Social Research in the Digital Age. Princeton University Press.

Schnabel Landon. 2018. “Sexual Orientation and Social Attitudes.” Socius 4:1–18. doi: 10.1177/2378023118769550.

Silva Tony. 2017a. “A Quantitative Test of Critical Heterosexuality Theory: Predicting Straight Identification in a Nationally Representative Sample.” Sexuality Research and Social Policy 1–14. doi: 10.1007/s13178-017-0307-8. [PubMed: 28824733]

Silva Tony. 2017b. “Bud-Sex: Constructing Normative Masculinity Among Rural Straight Men That Have Sex With Men.” Gender & Society 31(1):51–73. doi: 10.1177/0891243216679934.

Silva Tony, and Evans Clare R.. 2020. “Sexual Identification in the United States at the Intersections of Gender, Race/Ethnicity, Immigration, and Education.” Sex Roles 83(1112):722–38. doi: 10.1007/s11199-020-01145-x.

Simpson Ellen, and Semaan Bryan. 2021. “For You, or For ”You”?: Everyday LGBTQ+ Encounters with TikTok.” Proceedings of the ACM on Human-Computer Interaction 4:252:1–34. doi: 10.1145/3432951.

Sweeney Latanya. 2002. “K-Anonymity: A Model for Protecting Privacy.” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05):557–70. doi: 10.1142/S0218488502001648.

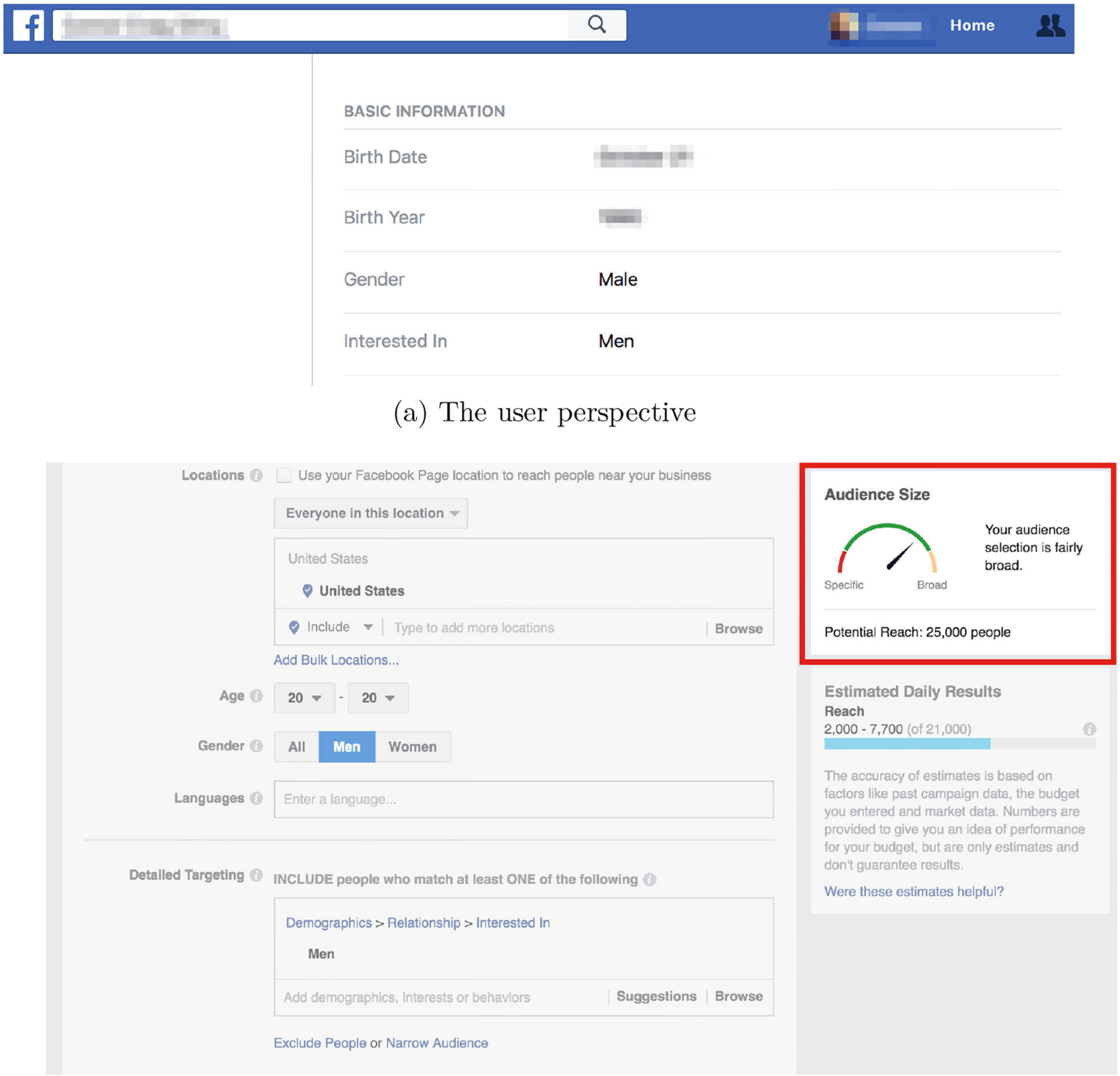
Ward Jane. 2015. Not Gay: Sex Between Straight White Men. New York: NYU Press.

Ward Jane. 2020. The Tragedy of Heterosexuality. NYU Press.

Westbrook Laurel, Budnick Jamie, and Saperstein Aliya. 2021. “Dangerous Data: Seeing Social Surveys Through the Sexuality Prism.” Sexualities. doi: 10.1177/1363460720986927.

Zagheni Emilio, Weber Ingmar, and Gummadi Krishna. 2017. “Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants.” Population and Development Review 43(4):721–34. doi: 10.1111/padr.12102.

Zerubavel Eviatar. 2018. Taken for Granted: The Remarkable Power of the Unremarkable. Princeton ; Oxford: Princeton University Press.



[redacted] Home

BASIC INFORMATION

Birth Date	[redacted]
Birth Year	[redacted]
Gender	Male
Interested In	Men

(a) The user perspective

Locations ☐ Use your Facebook Page location to reach people near your business

Everyone in this location

United States

United States

Include | Type to add more locations | Browse

Add Bulk Locations...

Age 20 - 20

Gender All **Men** Women

Languages Enter a language...

Detailed Targeting INCLUDE people who match at least ONE of the following

Demographics > Relationship > Interested In

Men

Add demographics, interests or behaviors | Suggestions | Browse

Exclude People or Narrow Audience

Audience Size

Your audience selection is fairly broad.

Potential Reach: 25,000 people

Estimated Daily Results

Reach

2,000 - 7,700 (of 21,000)

The accuracy of estimates is based on factors like past campaign data, the budget you entered and market data. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.

Were these estimates helpful?

(b) The advertiser perspective, estimated count highlighted

Figure 1:

The “interested in” profile field from user and advertiser perspectives. Screenshots from 2017 by one of the authors.

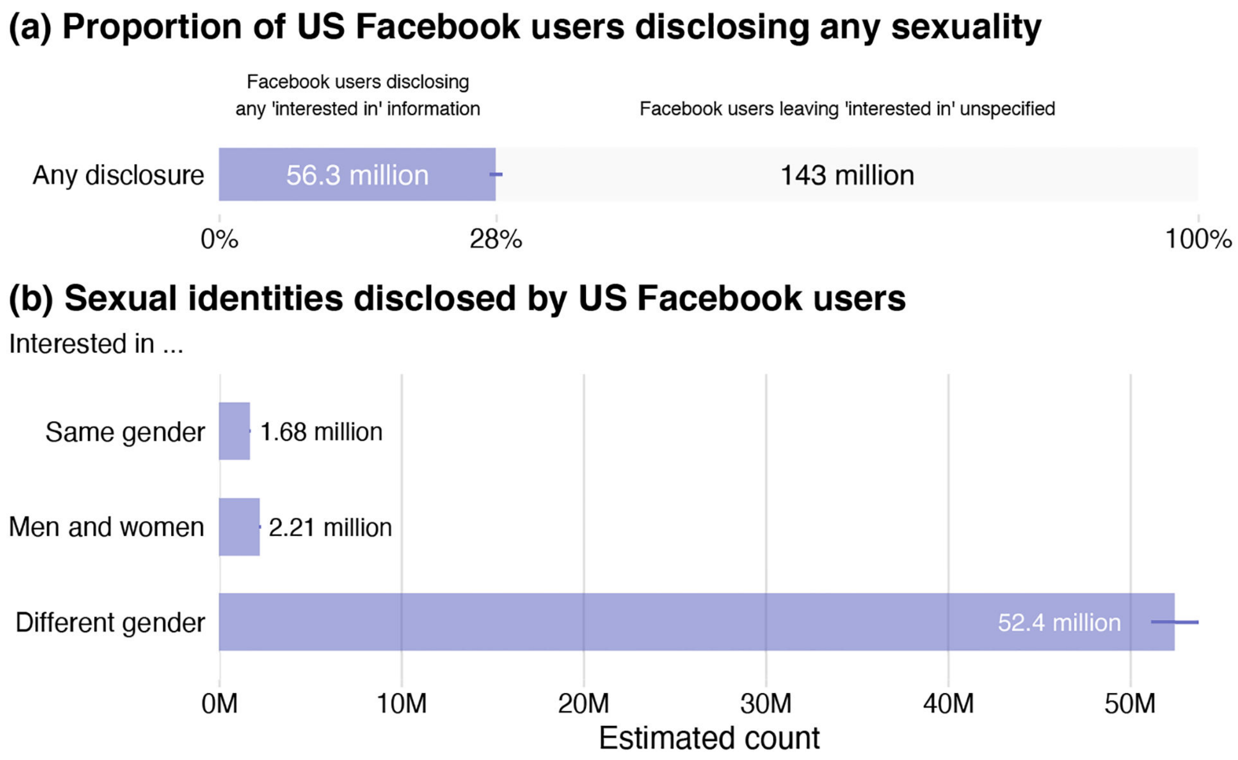


Figure 2:

Overall disclosure of sexualities by US Facebook users. *Note*: Data include women and men aged 18–64 with the four most common relationship statuses, and were collected in September 2017 from the Facebook Marketing API. Posterior medians and 95% posterior predictive intervals are shown, aggregated from the full model shown in Appendix Eq. (1) – Eq. (5).

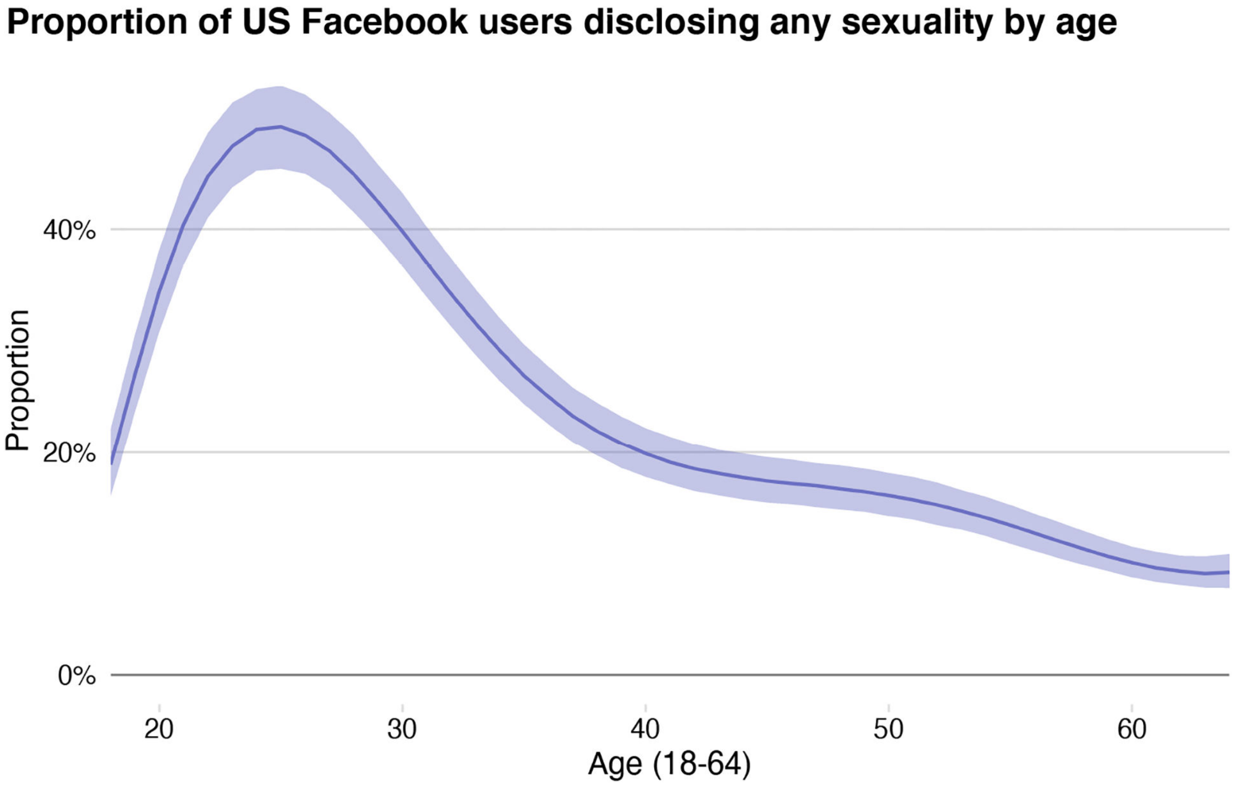


Figure 3:

Disclosure of any sexuality on Facebook, by age. *Note*: Data include US women and men with the four most common relationship statuses, and were collected in September 2017 from the Facebook Marketing API. Posterior medians and 95% posterior predictive intervals are shown.

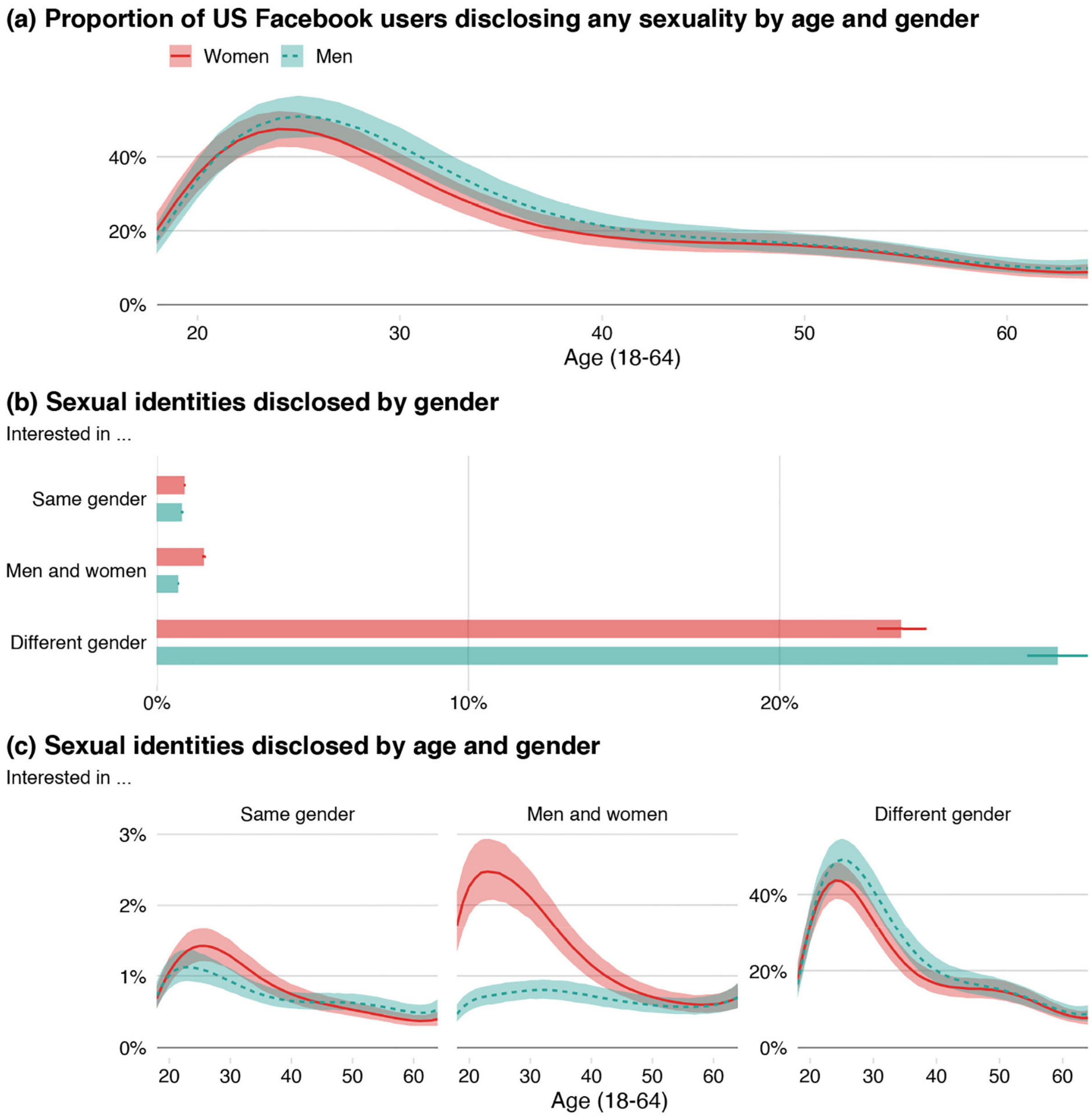


Figure 4:

Disclosure of any sexuality and specific sexual identities on Facebook, by gender and age. *Note*: Data include the four most common relationship statuses, and were collected in September 2017 from the Facebook Marketing API. Posterior medians and 95% posterior predictive intervals are shown.

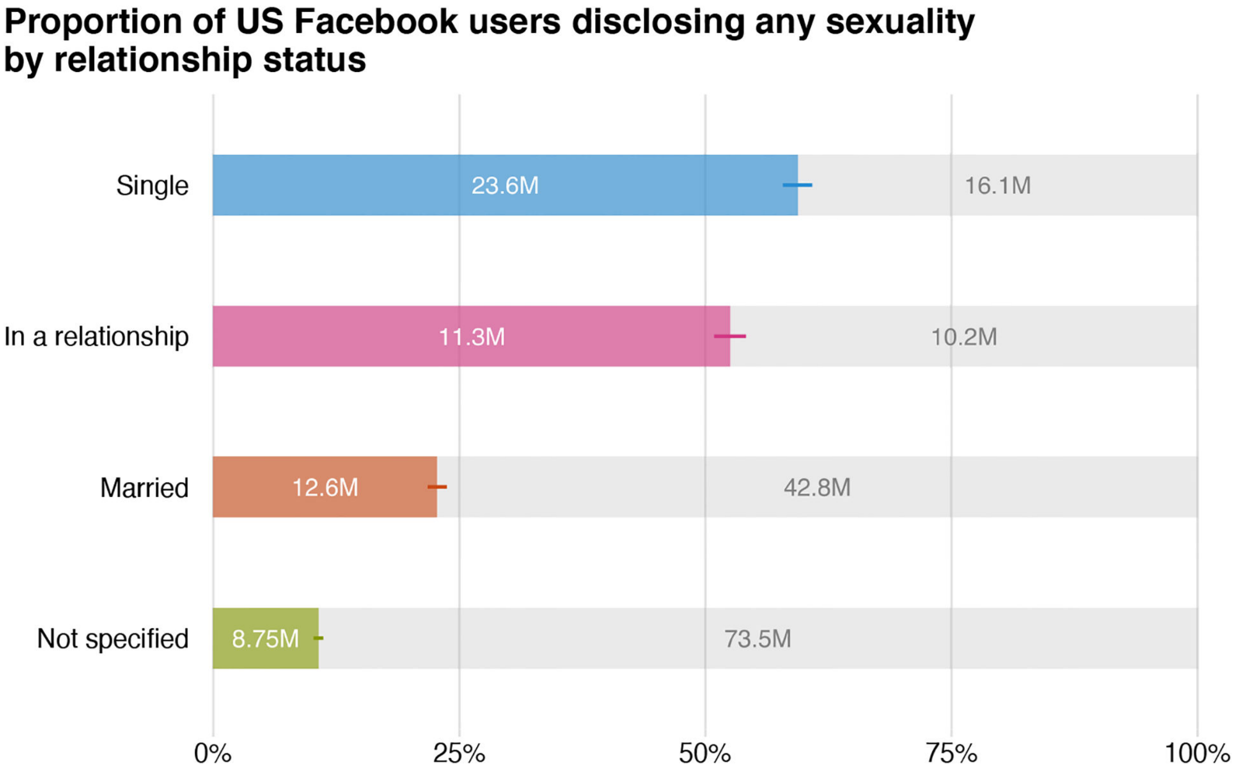


Figure 5:

Disclosure of any sexuality on Facebook, by relationship status. *Note*: Data include US women and men aged 18–64, and were collected in September 2017 from the Facebook Marketing API. Posterior medians and 95% posterior predictive intervals are shown.

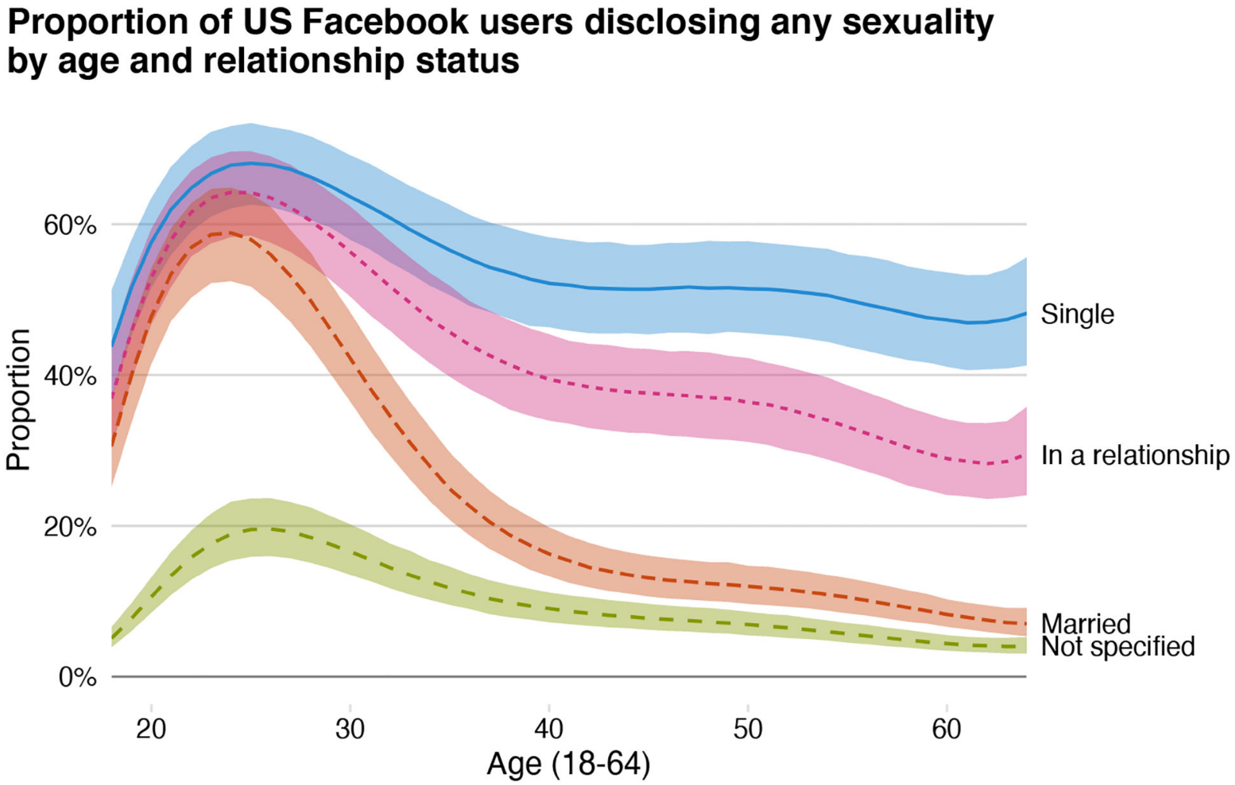


Figure 6:

Disclosure of any sexuality on Facebook, by relationship status and age. *Note*: Data include women and men in the US, and were collected in September 2017 from the Facebook Marketing API. Posterior medians and 95% posterior predictive intervals are shown.

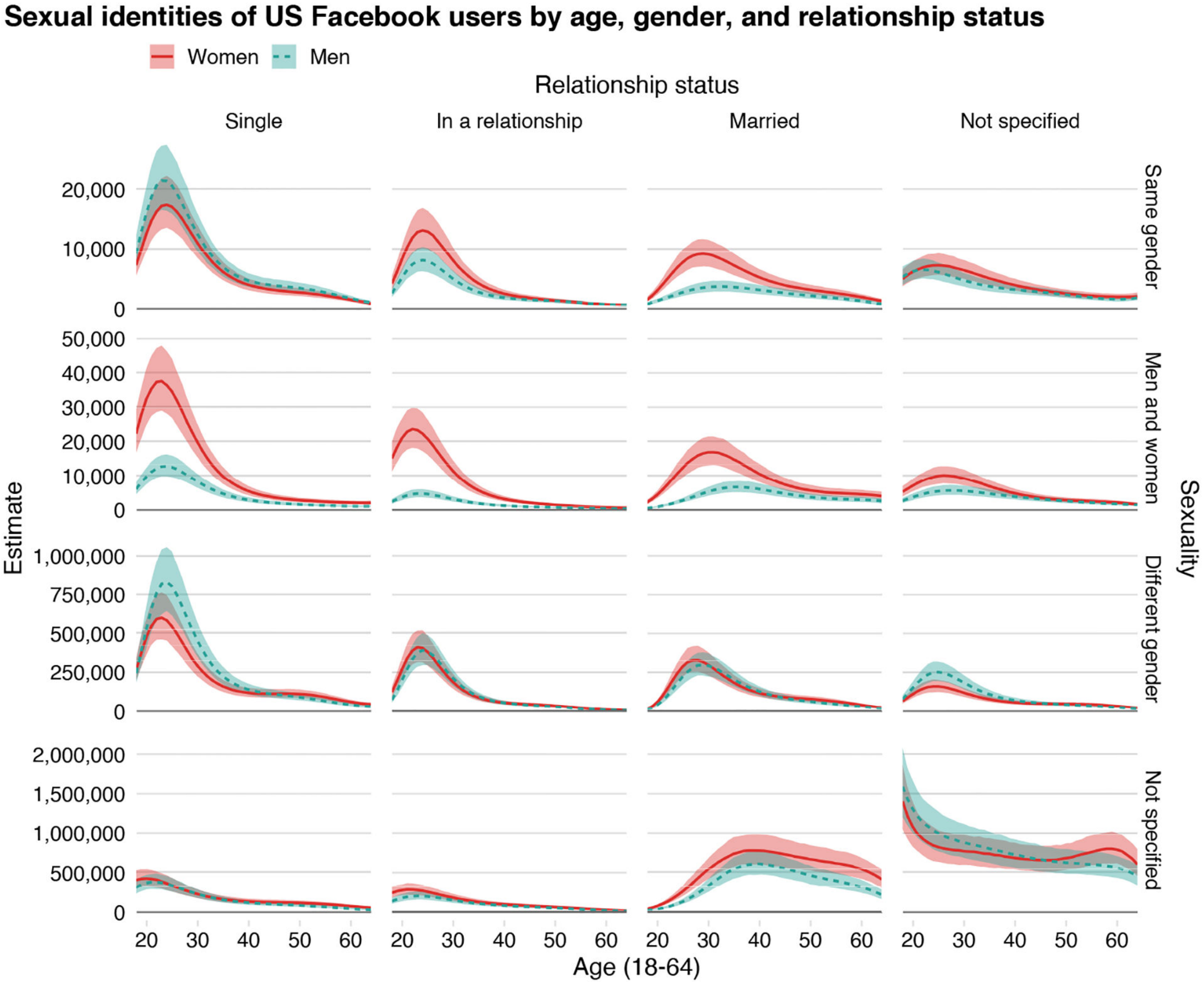


Figure 7:

Estimated counts of US Facebook users of each sexual identity, relationship status, age, and gender. *Note*: Data were collected in September 2017 from the Facebook Marketing API. Posterior medians and 95% posterior predictive intervals are shown.