PRELIMINARY DRAFT - PLEASE DO NOT CITE

Estimating internet adoption around the world using a sample of Facebook users

Dennis M. Feehan[*]
Curtiss Cobb[†]

April 2, 2018

Abstract

The internet has the potential to improve economic and social wellbeing but billions of people around the world have never been online. Reliable estimates of internet adoption traditionally require expensive and time-consuming household surveys. We describe an alternative approach that is dramatically faster and cheaper. Our approach is based on the insight that internet users are connected to other people through in-person social networks such as kin, friendship, and contact networks. By interviewing a sample of Facebook users and anonymously asking about the members of these offline social networks, we can learn about both people who are online and people who are not online. We describe how we derived our estimator and show initial results that suggest that our approach is promising. Our design could potentially be adapted to many other settings, offering one way to overcome some of the major challenges facing survey research in the information age.

[*]University of California, Berkeley, feehan@berkeley.edu
[†]Facebook, ccobb@fb.com

1 Overview

The internet has the potential to improve economic and social wellbeing through a wide range of different mechanisms, but billions of people around the world have never been online (World Bank 2016; Hjort and Poulsen 2017). This digital divide in access to a critical modern technology is an important dimension of inequality in the modern world. People in poor countries use the internet much less than people in wealthy countries (World Bank 2016). Even within countries that enjoy high levels of internet adoption, research suggests that access to the internet can be very unequal by age, gender, income, and race (Van Deursen and Van Dijk 2014; Friemel 2016; Haight, Quan-Haase, and Corbett 2014; Vigdor, Ladd, and Martinez 2014).

Researchers need to be able to measure the digital divide in order to understand its implications for inequality and opportunity. Entrepreneurs need to understand what sort of people currently face barriers in using the internet, and what kind of new products or technologies might help overcome those barriers. And policymakers who want to implement and evaluate strategies for making internet access more widely available rely on being able to measure the level and rate of change in the number of people who have access to the internet.

Reliable estimates of internet adoption are typically based on methodologically rigorous household surveys or censuses. However, this rigor comes at a price: these surveys can be very costly and typically take months to design and implement. These limitations are especially problematic because internet adoption appears to be changing on a much faster time-scale than many conventional indicators of social and economic wellbeing.

To help address this measurement challenge, we develop an alternative approach to estimating internet adoption that is dramatically faster and cheaper than conventional surveys: we interviewed a sample of Facebook users and asked them whether or not members of their offline personal networks use the internet. Our approach is based on the insight that internet users are connected to many other people through in-person social networks such as kin, friendship, and contact networks. By interviewing a probability sample of Facebook users and asking about the members of these offline social networks, we can learn about both people who are online and people who are not online.

Asking survey respondents to report about others is an idea that has independently arisen in many different substantive areas (see, for example, Sirken 1970; Lavallee 2007; Hill and Trussell 1977; Bernard et al. 1991; Marsden 2005). Our approach can be seen as an extension of this previous work to the situation where the goal is to learn about everyone in a population, but respondents are only sampled and interviewed online. Thus, our study is an illustration of one way to overcome many of the challenges that face the sampling and survey research community

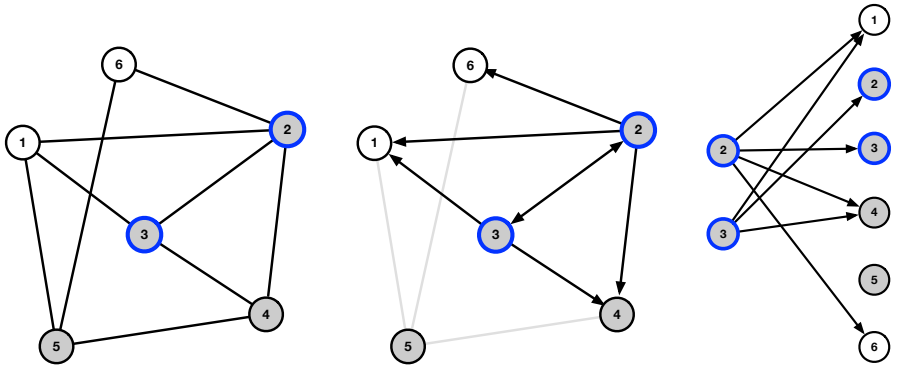


Figure 1: Network reporting setup: asking people on Facebook to report about their offline personal networks. (a) An example of a social network based on a symmetric (undirected) social relationship. (b) A reporting network generated by interviewing nodes 2 and 3. (c) The bipartite reporting network associated with the reporting network in Figure 1b.

in the information age.

The remainder of the paper begins with Section 2, which provides an overview of the formal setup that we use to determine which quantities need to be estimated, and how to estimate them; technical details are provided in Appendix A. Section 3 describes important design decisions, data collection procedures, and the actual estimators that we developed. Section 4 turns to the empirical results, which suggest that our strategy can be effective as a fast and inexpensive approach to estimating internet adoption; the results also point to several areas that the methodology could potentially be improved. Finally, Section 5 concludes by discussing the implications of our approach for the changing landscape of sampling and survey research in the information age.

2 Setup

Our strategy for obtaining fast and inexpensive estimates of internet adoption is based on asking Facebook users to report about internet adoption among other people they are connected to in their everyday, offline personal networks. The challenge is to determine how to turn people's anonymous reports about their personal network members into estimates of internet adoption. We now explain how we used a formal framework called network reporting to understand which quantities we need to estimate in order to accomplish our goal (Feehan 2015; Feehan and Salganik 2016a).

Figure 1 illustrates the general setup with an example. Figure 1a shows six people who are

connected together in a social network. The network relation is symmetric, meaning that whenever person A is connected to person B, then B is also connected to A. We make a distinction between nodes that can potentially be sampled and interviewed—the *frame population*—and other nodes. For example, a frame population might be cell phone users; the users of a specific app such as Facebook; or people who live at addresses that can be reached by postal mail. In Figure 1, nodes 2 and 3 are in the frame population.

Figure 1b shows the *reporting network* that is generated when both nodes 2 and 3 are interviewed about the people they are connected to in the social network. The reporting network is different from the social network: the social network has an undirected edge $A - B$ when A and B are socially connected; the reporting network, on the other hand, has a directed edge $A \to B$ whenever A reports about B. When reporting is accurate, there will be structural similarities between the social network and the reporting network, but this need not be true in general. The reporting network is a useful formalism that can help researchers develop estimators, understand possible sources of reporting errors, and derive self-consistency checks.

Figure 1c shows a rearrangement of Figure 1b that is helpful for deriving estimators from a reporting network. On the left-hand side of Figure 1c is the set of nodes that makes reports (the frame population), and on the right hand side is the set of nodes that can be reported about (the universe)[1]. Drawn this way, every report must connect a node on the left-hand side to a node on the right-hand side. Thus, the total number of reports that leaves the left-hand side must equal the total number of reports that arrives at the right-hand side. Mathematically, this means that when everyone in the frame population is interviewed, we have the following identity:

$$\#\text{ internet users} = N_H = \underbrace{\frac{\overbrace{y_{F,H}}^{\substack{\#\text{ reported connections from} \\ \text{people on FB to internet users}}}}{\bar{v}_{H,F}}}_{\substack{\text{average number of times each} \\ \text{internet user gets reported}}} \tag{1}$$

The denominator of Equation 1 is a quantity called the *visibility* of internet users. Intuitively, if we simply added up the number of reported internet users, we would get a number that is larger than the total number of internet users because each internet user can be reported more than once. Dividing by the visibility accounts for this fact.

[1]Note that a particular node can appear in both sides if it is in the frame population and in the universe.

3 Data collection and estimators

Each survey interview took place in two phases: in the first phase, survey respondents were asked to report the size of their personal networks (e.g., "How many people did you share food or drink with yesterday?"). Next, we wanted to obtain information about internet use among the members of each respondent's personal network. Ideally, we would ask for information about every single person in the respondent's network one by one. However, this approach seemed likely to produce unacceptable levels of respondent fatigue. Therefore, in the second phase of the interview respondents were asked for information about the three members of their personal networks who 'came to mind' first (Figure 2). We call these people that we obtain additional information about *detailed alters*[2].

The identity in Equation 1 would hold if we obtained a census of monthly active Facebook users. In practice, we have a sample and not a census; thus, we construct an estimator for the number of internet users by developing sample-based estimators for the numerator and the denominator of Equation 1. We now describe these two components in more detail.

Given information about respondents' network sizes and the detailed alters' internet use, the numerator of Equation 1 ($y_{F,H}$) can be estimated from our sample with:

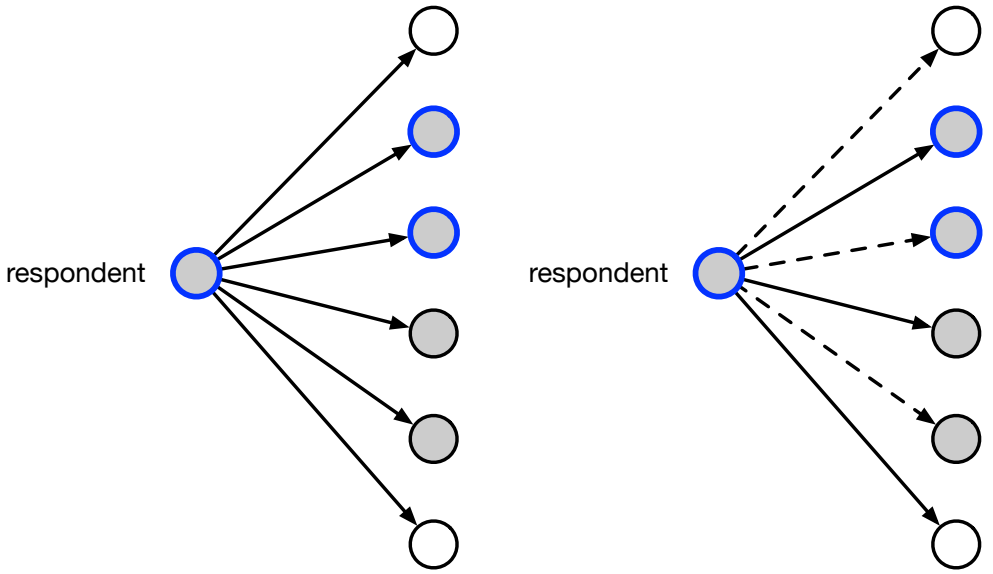
$$\widehat{y}_{F,H} = \sum_{i \in s} w_i \frac{d_i}{r_i} o_i, \tag{2}$$

where

- s is the sample of Facebook users
- w_i is the expansion weight for $i \in s$
- d_i is the network size (degree) of $i \in s$
- r_i is the number of detailed alters from $i \in s$ ($r_i \in \{0, 1, 2, 3\}$)
- o_i is the number of detailed alters reported to be online

In order to use information about the r_i detailed alters to make inferences about the d_i people in the respondent's network, the estimator in Equation 2 makes the additional assumption that the detailed alters are a simple random sample of respondents' personal networks. Thus, $\frac{d_i}{r_i}$ can be seen as a weight that accounts for sampling r_i out of the d_i personal network members. Previous work on egocentric survey research suggests that, instead of being a simple random sample, network members who come to mind first may be more likely to come from the same social context, and may be more likely to be strongly connected to the respondent (Marsden

[2]We did not ask for any sensitive or personally identifying information about these three detailed alters.



(a) (b)

Figure 2: (a) A survey respondent who is sampled online can be asked to report about members of one of her offline personal networks (e.g. her kin, friendship, or contact networks). Her responses contain information about both people who are online and people who are offline. (b) In order to reduce respondent burden, we asked for more detailed information about internet use, gender, and age for three *detailed alters* in each respondent's personal network.

2005). Therefore, we develop two different ways to assess this assumption: first, we introduce internal consistency checks that can detect systematic biases that would emerge if detailed alters are very different from other personal network members (Section 4.1); and, second, we introduce a sensitivity framework which enables us to formally assess the impact that different magnitudes of selection bias among the detailed alters would have on our estimates (Appendix B).

The denominator of Equation 1 ($\bar{v}_{H,F}$) is a quantity called the *visibility* of internet users, which is defined as the number of times that the average internet user would be reported in a census of active Facebook users. Many different strategies could be used to estimate or approximate the visibility of internet users; here, we adopt a simple approach: we use the average number of times that a Facebook user shares a meal with another Facebook user to approximate the visibility of internet users. This means that we make the assumption that people who are on Facebook share meals with each other at the same rate that they share meals with people who are on the internet, but who are not on Facebook. Mathematically, this assumption can be written

$$\bar{d}_{H,F} = \bar{d}_{F,F}. \tag{3}$$

The condition in Equation 3 requires that two quantities be equal: (1) the rate at which someone who is on the internet shares a meal with someone who is on Facebook ($\bar{d}_{H,F}$); and, (2) the rate at which someone who is on Facebook shares a meal with someone who is also on Facebook ($\bar{d}_{F,F}$). This assumption could be violated if, for example, people frequently organize sharing a meal together using Facebook (without inviting other people). We explore how violating this condition affects estimates as part of a sensitivity analysis in Appendix B and, in Section 5, we discuss how additional data collection could remove the need for this condition altogether.

Given the condition in Equation 3, we can estimate $\bar{v}_{H,F}$ with an estimator for $\bar{d}_{F,F}$, the average number of meals that someone on Facebook reports sharing with someone else on Facebook. In order to estimate $\bar{d}_{F,F}$, we use

$$\widehat{\bar{d}}_{F,F} = \frac{\sum_{i \in s} w_i \frac{d_i}{r_i} f_i}{\sum_{i \in s} w_i}, \tag{4}$$

where the new quantity, f_i, is the number of Facebook users that respondent i reports among her detailed alters.

Putting Equation 2 and Equation 4 together, we have

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{d}}_{F,F}} \tag{5}$$

$$= \frac{\sum_{i \in s} w_i \frac{d_i}{r_i} o_i}{\sum_{i \in s} w_i \frac{d_i}{r_i} f_i} \times \sum_{i \in s} w_i. \tag{6}$$

Appendix A has a detailed derivation of the estimator and a precise description of all of the conditions it relies upon, and Appendix B has a framework for sensitivity analysis which can be used to understand how estimates are affected by violations of the assumptions that the estimator relies upon.

Instrument design

We had to make several important design decisions when developing our survey. First, we had to determine who respondents should be asked to report about. In principle, people can be asked to report about any type of personal network relationship that is symmetric. Thus, the specific type of personal network that respondents are asked to report about—the *tie definition*—is a study design parameter that researchers are free to vary (Feehan et al. 2016). To explore the impact of this study design parameter, we randomized survey respondents to report about one of

two different tie definitions: the meal tie definition and the conversational contact tie definition (Table 1). We chose these two tie definitions because (1) previous research led us to believe that respondents can plausibly report the number of people that they interacted with in the previous day, avoiding the need to indirectly estimate personal network sizes; (2) researchers have had success using versions of these tie definitions in previous studies (Feehan et al. 2016; Mossong et al. 2008).

Table 1: The two different networks survey respondents were asked about.

Meal network	Conversational contact network
How many people did you share food or drink with yesterday? These people could be family members, friends, co-workers, neighbors, or other people. Please include all food or drink taken at any location, including at home, at work, at a cafe, or in a restaurant.	How many people did you have conversational contact with yesterday? By conversational contact, we mean anyone you spoke with face to face for at least three words.

4 Results

We used Facebook's survey infrastructure to obtain a simple random sample of people who actively use Facebook in five countries around the world: Brazil, Colombia, Great Britain, Indonesia, and the United States[3]. We chose these countries because they span a breadth of expected levels of internet adoption and economic development. Figure 3 shows the age and gender distribution of survey respondents for each tie definition[4]. Respondents were slightly more female than male in all countries except for Indonesia, and age distributions are typical of monthly active Facebook users in these countries. All estimates below are weighted to account for the sample design and to be representative of the universe of monthly active Facebook users in each country. Estimates of sampling uncertainty are based on the rescaled bootstrap method (Rao, Wu, and Yue 1992; Rao and Wu 1988; Feehan and Salganik 2016b).

Figure 4 shows the distribution of personal network sizes reported by respondents from each country, and for each tie definition. The average size of meal networks was smaller than conversational contact networks in all countries (Table 2): the average reported size of the meal

[3]We consider users to be active if they have logged onto Facebook in the 30 days before the survey; we also restrict responses to people over 15 years old.

[4]In order to ensure that the survey instrument and methods worked well, we started with a smaller sample in Great Britain (which is why there are fewer respondents in that country).

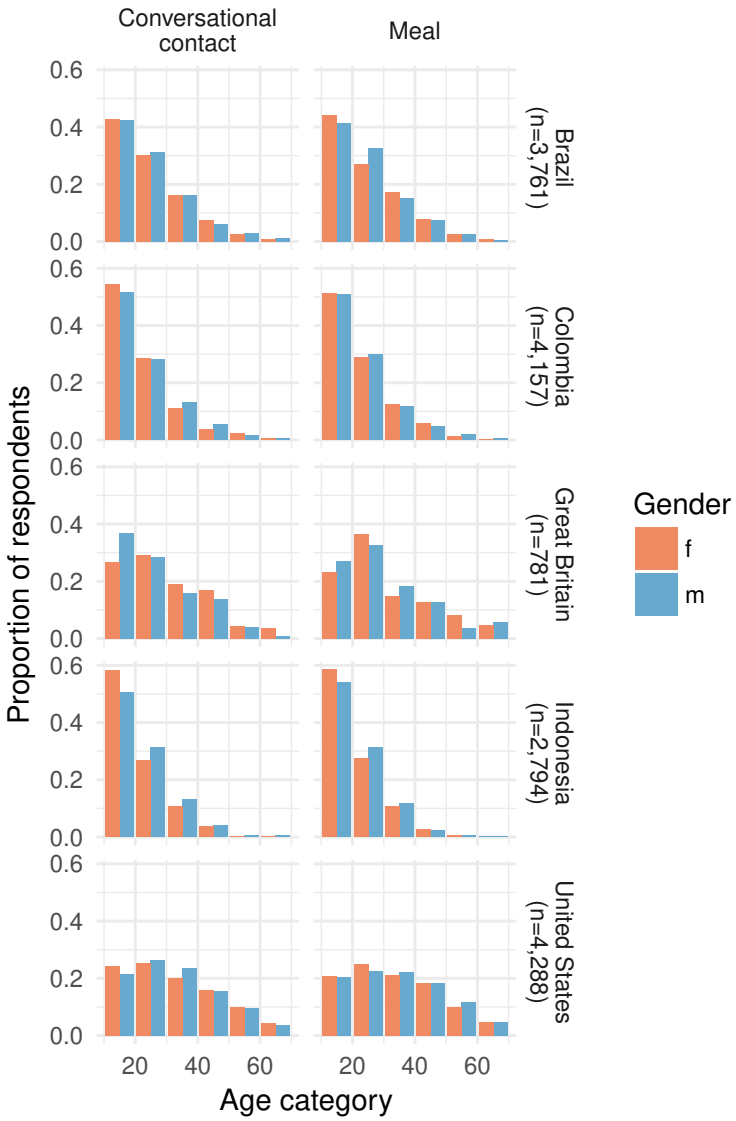


Figure 3: Age and gender of survey respondents in each country. Estimates throughout this article use sampling weights to account for sample design and nonresponse.

network varied from about 4 (Great Britain) to about 8 (Indonesia), while the average reported size of the conversational contact network varied from about 11 (Colombia and Indonesia) to about 13 (Brazil, Great Britain, and the United States). For both networks, Figure 4 suggests that there may be heaping in reported network sizes that are multiples of five and ten; this heaping is more evident in the reported number of conversational contacts than for meals, suggesting that reports about the meal network may more accurate than reports about the conversational contact network.

4.1 Internal consistency checks

In order to more formally assess the accuracy of reports about each network, we develop *internal consistency checks* (Feehan et al. 2016; Bernard et al. 2010). These internal consistency checks use the information about the age group and gender of each detailed alter that respondents reported about. The idea is to find reported quantities that can be estimated from the data in two different ways. To the extent that these independent estimates of the same quantity agree, the reported network connections are internally consistent. For example, using survey responses from only men, we can estimate the number of connections between men and women; similarly, using survey responses from only women, we can estimate the number of connections between women and men. By definition, these two quantities are equal; thus, under perfect conditions where our survey does not suffer from any reporting errors or selection biases, we would expect these two independent estimates to agree (up to sampling noise):

$$\begin{matrix}\#\text{ connections from} \\ \text{men to women}\end{matrix} = \begin{matrix}\#\text{ connections from} \\ \text{women to men}\end{matrix}.$$

We devised internal consistency checks based on reported connections to and from each of twelve different age-sex groups, by country and by tie definition. For each age-sex group α, we estimate the average number of connections from Facebook users in age-sex group α to Facebook users not in α ($d_{F_\alpha, F_{-}\alpha}$). We also estimate the average number of connections from Facebook users not in age-sex group α to Facebook users who are in age-sex group α ($d_{F_{-}\alpha, F_\alpha}$). We then define the average normalized difference Δ_α to be

$$\Delta_\alpha = \frac{1}{N_F}(\widehat{d}_{F_\alpha, F_{-}\alpha} - \widehat{d}_{F_{-}\alpha, F_\alpha}),$$

where N_F is the number of Facebook users in the country, a scaling factor that is intended to make it easier to compare across countries. In the absence of any reporting error, selection biases, or sampling variation, we would expect $\Delta_\alpha = 0$. On the other hand, it is possible that

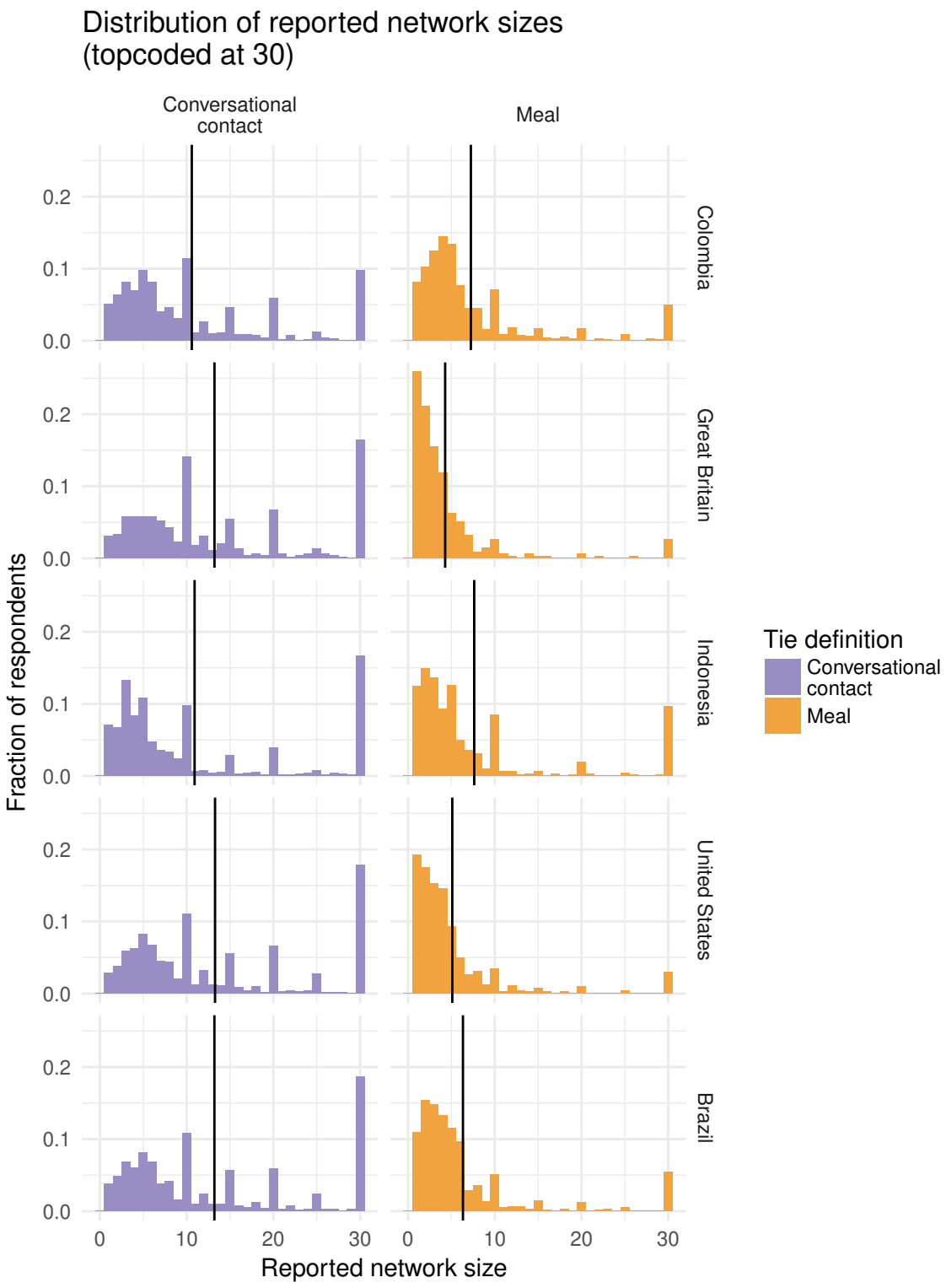


Figure 4: Estimated degree distributions for the conversational contact network (left) and the meal network (right). The vertical line on each panel shows the average. Average personal network size is smaller for the meal network than for the contact network; further, the contact network shows greater evidence of heaping on multiples of 5 and 10 than the meal network. These findings are consistent with a hypothesized tradeoff between the quality and the quantity of information reported in personal networks. Responses higher than 30 are coded as 30 in these plots.

$\Delta_\alpha > 0$; that can happen if either there is homophilic selection bias in the respondents' choice of detailed alters or if members of group α are especially inconspicuous; similarly, $\Delta_\alpha < 0$ can result from heterophilic selection bias in respondents' choice of detailed alters or if members of group α are especially conspicuous.

Figure 5 shows the average normalized difference (Δ_α) for internal consistency checks based on reported connections to and from each of twelve different age-sex groups, by country and by tie definition. Several notable features emerge from Figure 5. First, for many of the internal consistency checks, the averaged normalized differences are very close to zero, or have confidence intervals that contain zero. Second, Figure 5 suggests that reports based on the meal network are more internally consistent than reports based on conversational contact (confirmed below). Third, there appears to be no universal pattern that describes deviations in internal consistency checks that are not close to zero. In Indonesia the average normalized differences for younger age groups suggest that females may be relatively inconspicuous while males are relatively conspicuous[5]. On the other hand, in Brazil and Colombia, younger women appear to be particularly inconspicuous. And in Great Britain and the United States, most of the IC checks suggest that reports are internally consistent.

Figure 6 directly compares the difference in internal consistency results for the conversational contact and meal networks. The figure shows the estimated sampling distrubution of TSE, the total squared difference between the internal consistancy checks for the conversational contact network and the internal consistency checks for the meal network:

$$\text{TSE} = \sum_\alpha \left(\Delta_{\text{IC}}^{\text{cc}} - \Delta_{\text{IC}}^{\text{meal}}\right)^2$$

For all countries except for Indonesia, most of the mass of the estimated distribution is greater than 0, suggesting that the meal network reports are more internally consistent than conversational contact network reports (Appendix C).

Estimates of internet adoption

Figure 7 shows estimated internet adoption for each country in our sample, using each tie definition[6]. Two findings emerge from Figure 7. First, estimated internet adoption rates are very similar for the conversational contact and for the meal networks; in all countries, the

[5]Conspicuousness and homophilic reporting are not distinguishable from the data. In this discussion, we focus on conspicuousness; however, instead of Indonesian women being inconspicuous, it could also be the case that Indonesian women have homophilic selection biases in choosing their detailed alters (i.e., they tend to report other women at a higher rate than would be expected from simple random sampling of their network members).

[6]For the purposes of this study, we say that a person has adopted the internet if she has used the internet on a computer or a phone in the last 30 days.

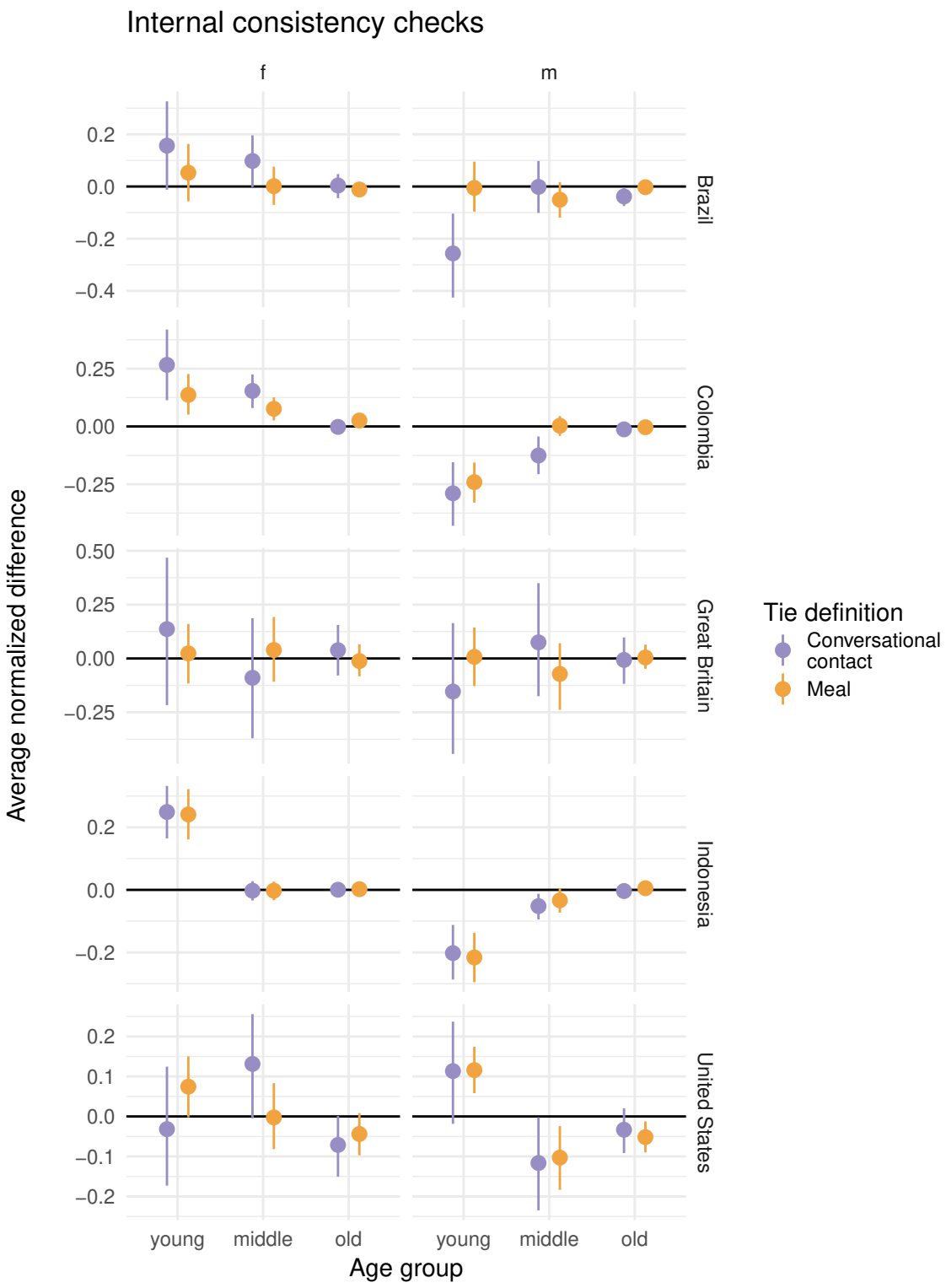


Figure 5: Internal consistency checks. By estimating the same quantity using independent parts of our sample, we can assess the internal consistency of respondents' network reports. Estimated difference between two independent estimates of the same quantity and 95% confidence intervals are shown for each age-gender group and each type of network; an estimate of 0 means that the two independent estimates are exactly the same. Across most age-sex groups, results are internally consistent with one another, within sampling error. Further, the results suggest that reports about the meal definition may be more internally consistent, even though meal networks are smaller than conversational contact networks.

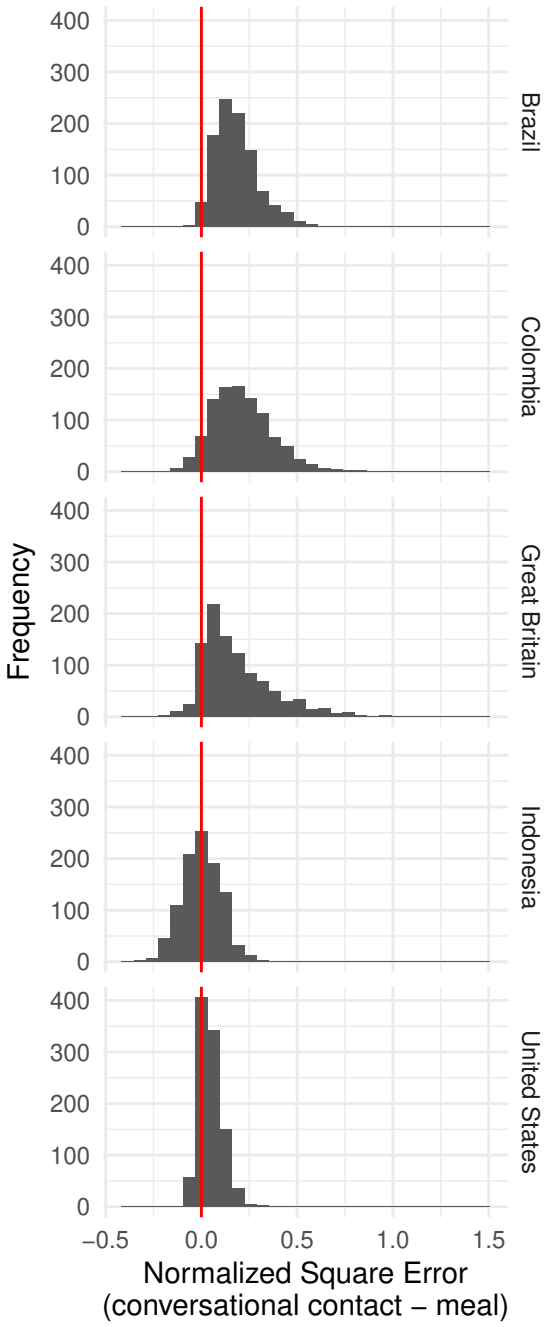


Figure 6: Estimated sampling distribution of the difference between the normalized square error for internal consistency checks from the conversational contact network and the normalized square error for internal consistency checks from the meal network. A difference greater than zero means that the meal definition was more internally consistent, while a difference less than zero means that the conversational contact definition was more accurate. Sampling uncertainty is estimated using the rescaled bootstrap. For all countries except for Indonesia, the meal network was more internally consistent than the conversational contact network ($P < 0.05$); in Indonesia, the two networks performed similarly (the meal network error was greater than the conversational contact network in about 52% of the bootstrap replications).

confidence intervals for estimates from the two tie definitions overlap. Second, the countries can be divided into three groups according to estimated adoption rates: the United States and Great Britain have the highest rates of internet adoption (above 75%); Brazil and Colombia have estimated internet adoption rates between 50% and 75%; and Indonesia has estimated adoption rates below 50%. This ordering is consistent with what would be predicted if economic factors such as GDP per capita were the main driver of internet adoption.

Ideally, we would evaluate our estimator by comparing it to gold standard measurements of internet adoption in each of the five countries. Unfortunately, no such gold standard currently exists. Therefore, in order to further assess the plausibility of the estimates presented in Figure 7, we compared our results to existing internet adoption estimates for Great Britain (Figure 8a) and for the United States (Figure 8b), two countries where several alternative estimates were available. The results show that the fast and inexpensive network reporting estimates are within the range of other estimates (in the United States) and similar to or slightly lower than other estimates (in Great Britain).

Summary and discussion

Several empirical findings emerged from the results of our study. We found that (1) reports from the stronger network tie produced information about fewer people than the weaker network tie in all five countries (Figure 4); but (2) reports from the stronger network tie appeared to produce more accurate information than reports from the weaker tie in all countries except for Indonesia (Figure 6). This finding is consistent with a hypothesized trade-off between the quantity and quality of information produced by network reports (Feehan et al. 2016); previous work found support for this theory in network reports about interactions in the 12 months before the interview. We find that this tie strength trade-off may operate even when reports are about interactions that took place the day before the interview. Future research could compare different time-windows to see if the hypothesized tradeoff between the quantity and quality of information operates across time within a fixed type of network tie. Over time, we hope that a deeper understanding of the relationship between reporting accuracy and the different dimensions of network tie definitions will accumulate, leading to useful guidance about how to design studies like ours.

The internal consistency checks suggest that people's reports about their network members can suffer from reporting errors, and that these reporting errors vary by who is being reported about (Figure 5). One possible mechanism for this result would be differential salience of interactions; another possible mechanism could be homophilic selection of the detailed alters. Future research could study different designs to try and distinguish between the salience of different demographic

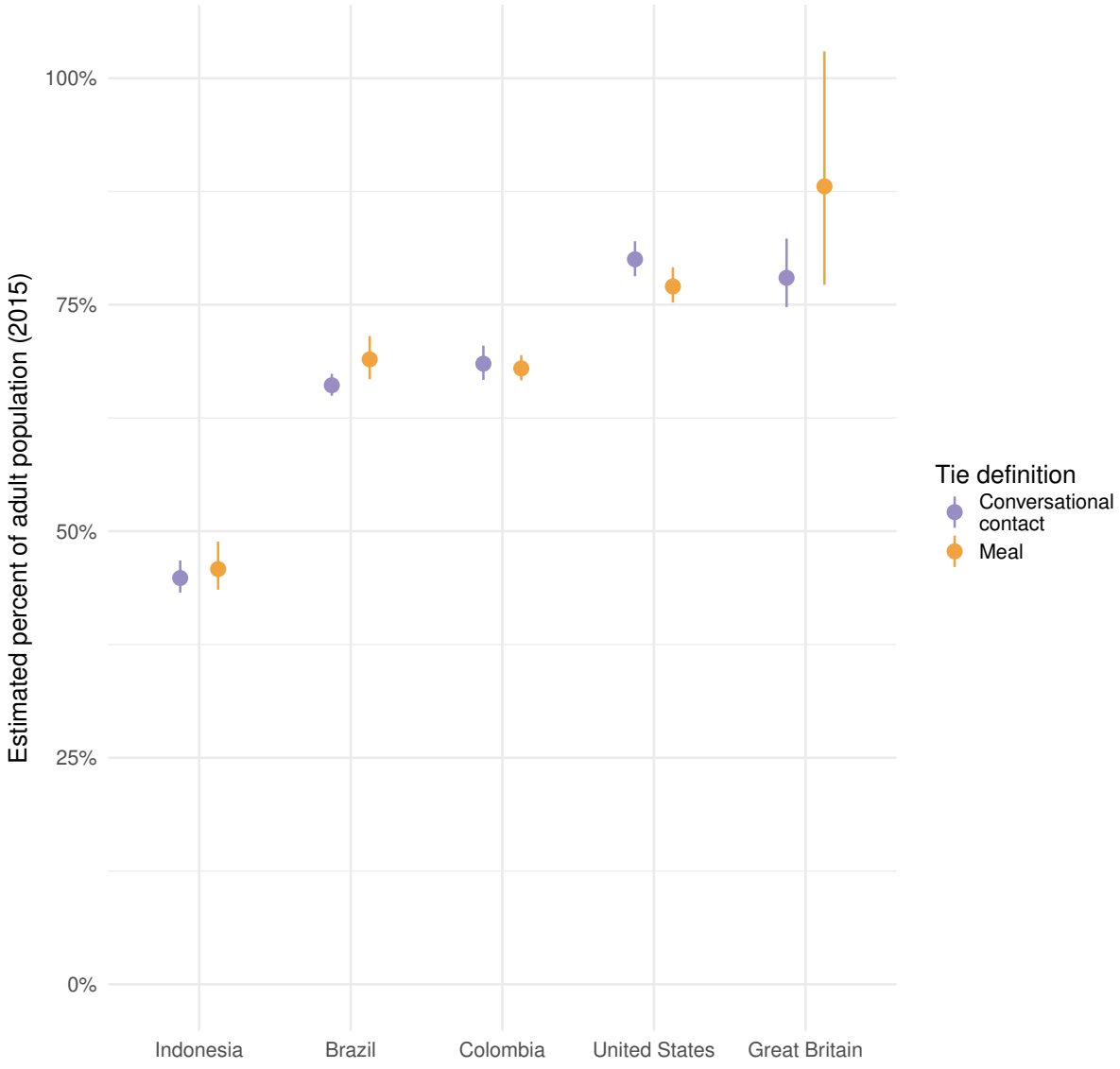
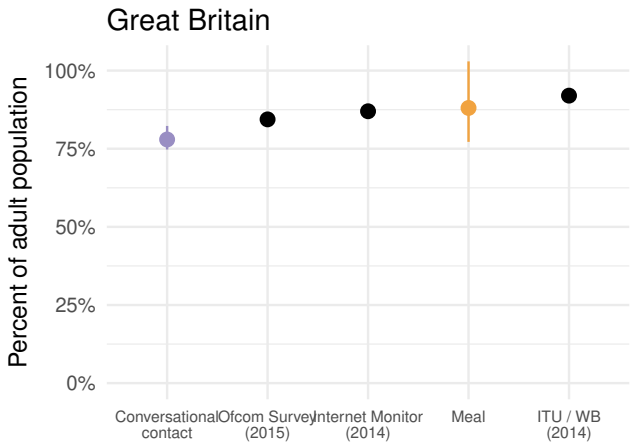
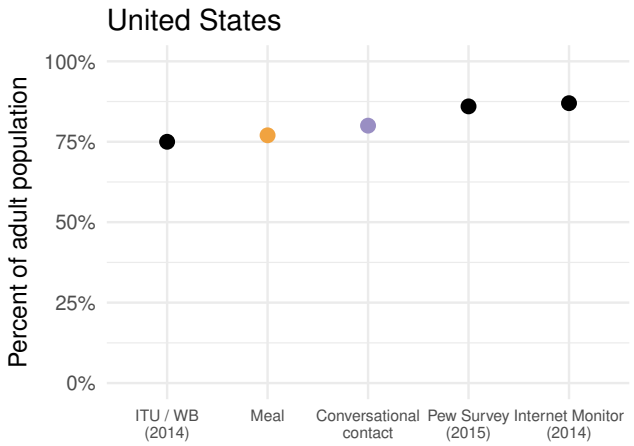


Figure 7: Estimated percentage of 2015 adult population that uses the internet, by country and for each of the two networks. 95% confidence intervals are based on the estimated sampling distribution from the rescaled bootstrap.



(a)



(b)

Figure 8: Comparison between network reporting estimates and other estimates for the United States and for Great Britain. (a) Internet access estimates for the UK. Comparison estimates are available for the UK from the ITU / World Bank; from the Internet Monitor; and from an Ofcom survey. Estimates from online network reports are generally slightly lower than other sources. (b) Internet access estimates for the US. Comparison estimates are available for the US from the ITU / World Bank; from the Internet Monitor; and from a Pew survey. Estimates from online network reports are generally slightly lower than estimates from Pew and Internet Monitor, but higher than estimates from the ITU / World Bank.

groups on the one hand and selection bias among the detailed alters on the other.

Finally, we also found that estimates of internet adoption from the two different networks were very similar (Figure 7). We could not validate our estimates by comparing them to gold-standard measurements of internet adoption rates because such a gold standard was not available. However, a comparison to other sources of estimates from the United States and Great Britain showed that the network reporting estimates are consistent with other sources of estimates in the United States and consistent or slightly lower than other estimates from Great Britain (Figure 8). Thus, we conclude that our fast and inexpensive strategy for obtaining approximate estimates of internet adoption is very promising.

5 Conclusion

Our results suggest several possible avenues for future work. In this study, we focused on simple, design-based estimators. A natural next step would be to start to build statistical models using these data. These models could exploit the relationships that are embedded in the internal consistency checks as a kind of constraint, estimating adjustments to ensure that reports are internally consistent. Such a model could potentially improve the accuracy of the resulting estimates. A second next step would be to use these data to produce estimates of internet adoption by age and gender; in principle, this should be possible with the data we collected.

More generally, we see this project as an example of how survey research can adapt and thrive in the information age. Decades of innovation, experimentation, and accumulated experience has led to the modern sample survey. This research program has been a huge success: today, surveys are the basis for a large share of quantitative research in demography, economics, political science, sociology, and other disciplines. Looking to the future, survey research in the information age faces challenges and opportunities (Groves 2011; Salganik 2017; Goel, Obeng, and Rothschild 2015). There are opportunities in the information age because it is increasingly fast and easy to sample some groups of people online, and it is increasingly possible to conduct interviews through computers, cell phones, tablets, or yet-uninvented means that can be much more flexible than traditional surveys. For example, surveys administered over the internet can incorporate an extremeley rich range of question styles, including items that feature video, audio, games, chatbots, and more (Tourangeau, Conrad, and Couper 2013; Salganik and Levy 2015; Nosek, Greenwald, and Banaji 2005; Couper, Conrad, and Tourangeau 2007; Fuchs 2009)

But survey research in the information age also faces big challenges. Response rates have been rapidly declining and traditional sampling frames like landline telephone numbers are increasingly inadequate (Meyer, Mok, and Sullivan 2015; Czajka and Beyler 2016; Kohut et

al. 2012). Conventional surveys are also time-consuming and expensive, meaning that they are only suitable for measuring quantities that only change over moderately long time-scales. Finally, many important problems concern people who cannot readily be sampled and interviewed using standard probability sampling (Sudman, Sirken, and Cowan 1988).

Researchers will have to adapt and innovate to overcome these challenges. In this paper, we describe one way forward. We showed that a sample of people who are online can be used to estimate internet adoption in five different countries around the world. Our approach is based on the idea that people know things about other people to whom they are connected through different kinds of personal networks.

We see our approach as a complement, rather than a replacement for conventional surveys. The ideal situation would combine frequent, inexpensive estimates, such as the ones described here, with less frequent conventional surveys. For example, a conventional probability sample of the general population in a country could be used to empirically estimate the average number of meals shared between an internet user and a Facebook user; with direct estimates of that quantity, the need for a key assumption in our estimator could be completely removed. More generally, a conventional probability sample survey can both be used to assess the accuracy of the fast and cheap estimates, and they can also be used to try to measure and relax some of the assumptions required by the faster, cheaper strategy.

6 References

Bernard, H. R., T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, and O. Scutelniciuc. 2010. "Counting Hard-to-Count Populations: The Network Scale-up Method for Public Health." *Sexually Transmitted Infections* 86 (Suppl 2): ii11–ii15.

Bernard, H. Russell, Eugene C Johnsen, Peter D Killworth, and Scott Robinson. 1991. "Estimating the Size of an Average Personal Network and of an Event Subpopulation: Some Empirical Results." *Social Science Research* 20 (2): 109–21.

Couper, Mick P., Frederick G. Conrad, and Roger Tourangeau. 2007. "Visual Context Effects in Web Surveys." *Public Opinion Quarterly* 71 (4): 623–34.

Czajka, John L., and Amy Beyler. 2016. "Background Paper Declining Response Rates in Federal Surveys: Trends and Implications." *Mathematica Policy Research*.

Feehan, Dennis M. 2015. "Network Reporting Methods." PhD thesis, Princeton University.

Feehan, Dennis M., and Matthew J. Salganik. 2016a. "Generalizing the Network Scale-up Method: A New Estimator for the Size of Hidden Populations." *Sociological Methodology*.

———. 2016b. *Surveybootstrap: Tools for the Bootstrap with Survey Data*.

Feehan, Dennis M., Aline Umubyeyi, Mary Mahy, Wolfgang Hladik, and Matthew J. Salganik. 2016. "Quantity Versus Quality: A Survey Experiment to Improve the Network Scale-up Method." *American Journal of Epidemiology*, March, kwv287.

Friemel, Thomas N. 2016. "The Digital Divide Has Grown Old: Determinants of a Digital Divide Among Seniors." *New Media & Society* 18 (2): 313–31.

Fuchs, Marek. 2009. "Gender of Interviewer Effects in a Video-Enhanced Web Survey: Results from a Randomized Field Experiment." *Social Psychology* 40 (1): 37–42.

Goel, Sharad, Adam Obeng, and David Rothschild. 2015. "Non-Representative Surveys: Fast, Cheap, and Mostly Accurate." In *Working Paper*.

Groves, Robert M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75 (5): 861–71.

Haight, Michael, Anabel Quan-Haase, and Bradley A. Corbett. 2014. "Revisiting the Digital Divide in Canada: The Impact of Demographic Factors on Access to the Internet, Level of

Online Activity, and Social Networking Site Usage." *Information, Communication & Society* 17 (4): 503–19.

Hill, Kenneth, and James Trussell. 1977. "Further Developments in Indirect Mortality Estimation." *Population Studies* 31 (2): 313–34.

Hjort, Jonas, and Jonas Poulsen. 2017. "The Arrival of Fast Internet and Employment in Africa." National Bureau of Economic Research.

Kohut, Andrew, Scott Keeter, Carroll Doherty, Michael Dimock, and Leah Christian. 2012. "Assessing the Representativeness of Public Opinion Surveys." *Pew Research Center, Washington, DC*.

Lavallee, P. 2007. *Indirect Sampling*. New York: Springer-Verlag.

Maltiel, Rachael, Adrian E. Raftery, Tyler H. McCormick, and Aaron J. Baraff. 2015. "Estimating Population Size Using the Network Scale up Method." *Annals of Applied Statistics* 9 (3): 1247–77.

Marsden, Peter V. 2005. "Recent Developments in Network Measurement." *Models and Methods in Social Network Analysis* 8: 30.

Meyer, Bruce D., Wallace KC Mok, and James X. Sullivan. 2015. "Household Surveys in Crisis." *The Journal of Economic Perspectives* 29 (4): 199–226.

Mossong, Joël, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, et al. 2008. "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases." *PLoS Med* 5 (3): e74.

Nosek, Brian A., Anthony G. Greenwald, and Mahzarin R. Banaji. 2005. "Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity." *Personality and Social Psychology Bulletin* 31 (2): 166–80.

Rao, J. N. K., and Norma P. Pereira. 1968. "On Double Ratio Estimators." *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 30 (1): 83–90.

Rao, J. N. K., and C. F. J. Wu. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83 (401): 231–41.

Rao, JNK, CFJ Wu, and K Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18 (2): 209–17.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Salganik, Matthew J., and Karen EC Levy. 2015. "Wiki Surveys: Open and Quantifiable Social Data Collection." *PloS One* 10 (5): e0123483.

Sarndal, C. E., B. Swensson, and J. Wretman. 2003. *Model Assisted Survey Sampling*. Springer Verlag.

Sirken, Monroe G. 1970. "Household Surveys with Multiplicity." *Journal of the American Statistical Association* 65 (329): 257–66.

Sudman, Seymour, Monroe G. Sirken, and Charles D. Cowan. 1988. "Sampling Rare and Elusive Populations." *Science* 240 (4855): 991–97.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. *The Science of Web Surveys*. Oxford University Press.

Van Deursen, Alexander JAM, and Jan AGM Van Dijk. 2014. "The Digital Divide Shifts to Differences in Usage." *New Media & Society* 16 (3): 507–26.

Vigdor, Jacob L., Helen F. Ladd, and Erika Martinez. 2014. "Scaling the Digital Divide: Home Computer Technology and Student Achievement." *Economic Inquiry* 52 (3): 1103–19.

Wolter, Kirk. 2007. *Introduction to Variance Estimation*. 2nd ed. New York: Springer.

World Bank. 2016. "World Development Report 2016: Digital Dividends." Washington, D.C.

A Derivation of the estimators

Sampling setup

We assume a conventional probability sampling setup, following the theory of design-based sampling; see Sarndal, Swensson, and Wretman (2003) for an excellent overview. Our frame population F – the set of people who could potentially be sampled – is monthly active Facebook users in a given country[7]. The population whose size we are trying to estimate is H, the number of internet users in the country. The goal is to use information about people on Facebook's reported offline personal network connections in order to estimate the size of H.

We assume that we obtain a *probability sample* s from the frame population, where we use the same definition of a probability sample as Sarndal, Swensson, and Wretman (2003). To briefly review, we assume that the sample s is chosen from among the members of the frame population F using a known random sampling method. The probability that $i \in F$ is included in the sample s, called i's *inclusion probability*, is written π_i. We require that $\pi_i > 0$ for all $i \in F$. We call the $w_i = \frac{1}{\pi_i}$ the *expansion weight* for unit $i \in F$.

Finally, several of the estimators we study are ratio or compound ratio estimators. The literature on design-based sampling has established that if each component estimator is consistent and unbiased, then compound ratio estimators are design-consistent but, strictly speaking, compound ratio estimators are not unbiased. Fortunately, a large literature has studied this problem and such estimators are typically found to be very nearly unbiased, both in theory and in practice[8]. Thus, we refer to these compound ratio estimators as *essentially unbiased*. The following result formally establishes these important properties of compound ratio estimators; which we will use these properties below.

Result A.1. *Suppose that $\widehat{y}_1, \ldots, \widehat{y}_n$ are estimators that are consistent and unbiased for $Y_1, \ldots, Y_n$ respectively. Then the compound ratio estimator*

$$\widehat{R} = \frac{\widehat{y}_1 \ldots \widehat{y}_k}{\widehat{y}_{k+1} \ldots \widehat{y}_n}.$$

is consistent and essentially unbiased for $R = (Y_1 \ldots Y_k)/(Y_{k+1} \ldots Y_n)$.

[7] Throughout this paper, we use the term Facebook users to refer to monthly-active Facebook users.

[8] We do not expect the situations in which compound ratio estimators would be biased to be relevant to our study; the biggest concern is typically when the denominator of $\widehat{R}$ is very small, which is not likely in our applications.

Proof. See Rao and Pereira (1968), Wolter (2007) (pg. 233), and Feehan and Salganik (2016a) for more details.

□

Throughout this analysis, we work in the framework of design based sampling. Thus, when we refer to an estimator as 'consistent', we mean design-consistent (also called Fisher consistent; Sarndal, Swensson, and Wretman (2003)).

We adhere to the notation used in previous papers about network scale-up and network reporting (Feehan 2015; Feehan and Salganik 2016a; Feehan et al. 2016):

- $y_{i,B}$ is the number of reported connections from person i to members of group B
- $y_{A,B} = \sum_{i \in A} y_{i,B}$ is the number of reported connections from members of group A to group B
- $d_{i,B}$ is the number of undirected connections in the social network between i and members of group B
- $d_{A,B} = \sum_{i \in A} d_{i,B}$ is the total number of undirected connections in the social network between members of group A and members of group B
- $v_{i,A}$ is the *visibility* of i to group A – i.e., the number of times that i would be reported if everyone in A was interviewed
- $v_{B,A} = \sum_{i \in B} v_{i,A}$ is the total visibility of members of group B to group A
- $\widehat{y} \to Y$ is shorthand for '$\widehat{y}$ is a consistent and unbiased estimator for Y
- $\widehat{y} \rightsquigarrow Y$ is shorthand for '$\widehat{y}$ is a consistent and essentially unbiased estimator for Y
- $y^{+}_{F,H}$ is the number of reported connections from F to H that actually lead to H. If $y^{+}_{F,H} = y_{F,H}$ then we say that there are *no false positive reports*

Aggregate reporting framework

We develop an estimator using the network reporting framework, an approach that builds upon insights from several different streams of previous research on sampling (Feehan and Salganik 2016a; Feehan 2015; Sirken 1970; Lavallee 2007; Bernard et al. 1991). Feehan (2015) shows that researchers can develop estimators based on network reports using either an individual or an aggregate multiplicity approach. Since we do not collect information at the level of detail required by individual multiplicity estimation, we adopt an aggregate multiplicity approach in this study. This aggregate multiplicity approach is similar to the network scale-up method (Bernard et al. 1991; Bernard et al. 2010; Maltiel et al. 2015; Feehan and Salganik 2016a).

Result A.2. *Suppose that a census of the frame population F is interviewed and asked to report about their connections to a group Z. Call the total number of reported connections $y_{F,Z}$ and suppose $y_{F,Z} > 0$. Further, suppose that there are no false positive reports, so that $y_{F,Z} = y^{+}_{F,Z}$. Finally, suppose that $\bar{v}_{Z,F}$ is the average visibility of members of Z; that is, $\bar{v}_{Z,F}$ is the average number of times that a member of Z is reported by someone in F. Then*

$$N_H = \frac{y_{F,H}}{\bar{v}_{H,F}}. \tag{7}$$

Proof. See Feehan (2015) and Feehan and Salganik (2016a).

□

To see the intuition behind the aggregate multiplicity approach from Result A.2, suppose we conducted a census of the frame population, asking every frame population member to tell us how many members of her personal network were online. Simply adding up the number of reported connections to internet users would produce a number that is larger than the number of internet users because each internet user can be reported more than once. Thus, in order to adjust for this over-counting, aggregate multiplicity estimators divide an estimate for the total number of reports by an estimate of hidden population members' *visibility*. The visibility is the number of times an average member of the hidden population would be reported if everyone on the frame population responded to the survey. In this study, the visibility is the number of times that the average internet user in a given country would be reported as an internet user, if everyone on Facebook in the country responded to the survey. Dividing the estimated total number of reported connections to people on the internet by the estimated visibility adjusts for the over-counting that would occur if the reports were used to directly estimate the number of internet users.

Given the aggregate multiplicity identity, our basic approach is to develop data collection strategies and statistical estimators that enable us to estimate the numerator and denominator of the identity in Equation 7. In the remainder of this Appendix, we develop necessary technical results to use the identity in Equation 7 to estimate the number of internet users in a given country.

Estimates about detailed alters

Result A.3 formalizes a situation where respondents are sampled and then asked about a sample of their network members. Result A.3 is stated in terms of an arbitrary dichotomous trait z that respondents report about their personal network members; for example, z could be Facebook usage, internet usage, gender, or membership in an age group.

Result A.3. *Suppose we have a sample s taken from the frame population using a probability sampling design. Call the expansion weights given by the sampling design w_i for each $i \in s$. Further, suppose that for each $i \in s$, we obtain information from a simple random subsample s_i of size r_i from the d_i people in i's personal network. Let z_{ij} be an indicator variable for whether or not i reports that j has trait Z, and let $z_i = \sum_{j \in s_i} z_{ij}$ be the total number of detailed alters respondent i reports having trait Z. Then the estimator*

$$\widehat{y}_{F,Z} = \sum_{i \in s} w_i \frac{d_i}{r_i} z_i$$

is consistent and unbiased for $y_{F,Z}$, the total number of reported connections to people with trait Z in a census of the frame population in which respondents report about everyone in their networks.

Proof. First, we note that we can consider this to be a multi-stage sample, where the first stage(s) lead to selection of the respondent and the final stage is the subsampling of detailed alters within each respondent's network. Since the final stage is a simple random sample of r_i out of d_i network members, the design weight for the final stage is $\frac{d_i}{r_i}$ for each detailed alter. In order to show that the estimator is unbiased, we take expectations with respect to the multi-stage sampling design:

$$\mathbb{E}[\widehat{y}_{F,Z}] = \mathbb{E}_I[\sum_{i \in s} w_i \mathbb{E}_i[\frac{d_i}{r_i} z_i | s]] = \sum_{i \in F} \pi_i w_i \mathbb{E}_i[\frac{d_i}{r_i} z_i | s]] = \sum_{i \in F} \pi_i w_i \left(\sum_{j \sim i} \pi_j^i \frac{d_i}{r_i} z_{ij} \right),$$

where the outer expectation $\mathbb{E}_I[\cdot]$ is taken with respect to the sampling of respondents and the inner expectation $\mathbb{E}_i[\cdot|s]$ is taken with respect to the sampling of detailed alters within each sampled respondent; $j \sim i$ indexes over all of the network members j that i could potentially report about; and we have written π_i for the inclusion probability of respondent i under the sampling design, and π_j^i for the inclusion probability of respondent i's jth network member under the subsampling design.

By definition, $w_i = \frac{1}{\pi_i}$ and $\pi_j^i = \frac{r_i}{d_i}$. Thus, continuing from above, we have

$$\mathbb{E}[\widehat{y}_{F,Z}] = \sum_{i \in F} \pi_i w_i \left(\sum_{j \sim i} \pi_j^i \frac{d_i}{r_i} z_{ij} \right) = \sum_{i \in F} \left(\sum_{j \sim i} \pi_j^i \frac{d_i}{r_i} z_{ij} \right) = \sum_{i \in F} y_{i,Z} = y_{F,Z}.$$

So we have shown that the estimator is unbiased for $y_{F,Z}$.

Finally, in a census of the frame population where every respondent reports about all of her

network members, $s = F$, $\pi_i = 1$, $\pi_j^i = 1$, $z_i = y_{i,Z}$, and $r_i = d_i$ for all i and j. Thus

$$\widehat{y}_{F,Z} = \sum_{i \in s} w_i \frac{d_i}{r_i} z_i = \sum_{i \in F} y_{i,Z} = y_{F,Z}$$

So the estimator is design-consistent.

□

Corollary A.1. *Under the conditions of Result A.1, the estimator*

$$\widehat{\bar{y}}_{F,Z} = \frac{\sum_{i \in s} w_i \frac{d_i}{r_i} z_i}{\sum_{i \in s} w_i}$$

is consistent and essentially unbiased for $\bar{y}_{F,Z}$.

Proof. By Result A.3, the numerator is consistent and unbiased for $y_{F,Z}$, and the denominator is a sample-based estimate for the size of the frame population, $\widehat{N}_F = \sum_{i \in s} w_i$. Thus, this is a Hajek-type estimator. See (Sarndal, Swensson, and Wretman 2003) for a proof that Hajek estimators are consistent and essentially unbiased.

□

Note that Result A.3 implies that Equation 2 is consistent and unbiased for $y_{F,H}$ and Corollary A.3 implies that Equation 4 s consistent and unbiased for $\bar{y}_{F,F}$.

Assembling the esitmator

The next estimator, Result A.4, shows that if we can estimate the total reported connections from frame population members to internet users, and if we can estimate the average visibility of internet users to frame population members, then we can estimate the number of internet users.

Result A.4. *Suppose that the* $\widehat{y}_{F,H}$ *is a consistent and unbiased estimator for* $y_{F,H}$ *and that* $\widehat{\bar{y}}_{F,F}$ *is a consistent and essentially unbiased estimator for* $\bar{y}_{F,F}$. *Further, suppose that reports are accurate in aggregate, so that* $y_{F,H} = d_{F,H}$ *and* $y_{F,F} = d_{F,F}$. *Finally, suppose that*

$$\bar{d}_{H,F} = \bar{d}_{F,F} \tag{8}$$

Then the estimator

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{y}}_{F,F}}$$

is consistent and essentially unbiased for N_H*.*

Proof. Since $\widehat{y}_{F,H} \to y_{F,H}$ and $\widehat{\bar{y}}_{F,F} \rightsquigarrow \bar{y}_{F,F}$, Result A.1 shows that $\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{y}}_{F,F}} \rightsquigarrow \frac{y_{F,H}}{\bar{y}_{F,F}}$. It remains to show that $\frac{y_{F,H}}{\bar{y}_{F,F}}$ is equal to N_H. By the condition that reports are accurate in aggregate, $y_{F,H} = d_{F,H}$ and $y_{F,F} = d_{F,F}$. Thus,

$$\frac{y_{F,H}}{\bar{y}_{F,F}} = \frac{d_{F,H}}{\bar{d}_{F,F}}.$$

Next, using the condition that $\bar{d}_{F,F} = \bar{d}_{H,F}$, we have

$$\frac{d_{F,H}}{\bar{d}_{F,F}} = \frac{d_{F,H}}{\bar{d}_{H,F}} = N_H \frac{d_{F,H}}{d_{H,F}} = N_H,$$

where the last step follows from the fact that we are assuming a symmetric type of network tie, meaning that the number of connections from F to H must be equal to the number of connections from H to F.

□

Result A.4 relies upon the condition that $\bar{d}_{H,F} = \bar{d}_{F,F}$ (Equation 8), which requires that two quantities be equal: (1) the rate at which someone who is on the internet shares a meal with someone who is on Facebook ($\bar{d}_{H,F}$); and, (2) the rate at which someone who is on Facebook shares a meal with someone who is also on Facebook ($\bar{d}_{F,F}$). This assumption could be violated if, for example, people frequently organize sharing a meal together using Facebook (without inviting other people).

To further understand the condition in Equation 8, note that since $F \subset H$ (i.e., everyone on Facebook is also on the Internet), it follows that

$$\bar{d}_{H,F} = p_{F|H}\bar{d}_{F,F} + (1 - p_{F|H})\bar{d}_{H-F,F} \tag{9}$$

where $p_{F|H} = \frac{N_F}{N_H}$ is the prevalence of F among H, i.e., the fraction of people on the internet that is also on Facebook. Therefore, when the condition in Equation 8 holds, then it is also the case that

$$\bar{d}_{F,F} = \bar{d}_{H-F,F}. \tag{10}$$

B Sensitivity framework

In this Appendix, we describe a framework that can be used to understand how to assess the sensitivity of the estimated number of people who use the internet to the various conditions that the results in Appendix A rely upon.

In order to develop the sensitivity framework, we adapt previous work on network scale-up and other network reporting methods (Feehan 2015; Feehan and Salganik 2016a). We start by introducing three quantities, called *adjustment factors*:

$$\eta_H = \frac{\text{avg \# reported connections from from F to H that actually lead to H}}{\text{avg \# reported connections F to H}} = \frac{y^+_{F,H}}{y_{F,H}}, \tag{11}$$

and

$$\eta_F = \frac{\text{avg \# reported connections from from F to F that actually lead to F}}{\text{avg \# reported connections F to F}} = \frac{y^+_{F,H}}{y_{F,H}}, \tag{12}$$

and

$$\nu = \frac{\text{avg \# in-reports to F from F}}{\text{avg \# in-reports to H from F}} = \frac{\bar{v}_{F,F}}{\bar{v}_{H,F}}. \tag{13}$$

Each of these new parameters is equal to 1 under ideal conditions, when the requirements of the results in Appendix A are satisfied. In general, ν can take on any value from 0 to ∞, while η_F and η_H can take on any value from 0 to 1.

The first sensitivity result reveals how estimated numbers of internet users will be affected if one or more of the three adjustment factors is not equal to 1.

Result B.1. *Suppose that the sampling conditions for Result A.3 hold, but that the reporting and network structure conditions do not. That is, suppose we have a sample s taken from the*

frame population using a probability sampling design. Call the expansion weights given by the sampling design w_i for each $i \in s$. Further, suppose that for each $i \in s$, we obtain information from a simple random subsample s_i of r_i out of the d_i people in i's personal network.

Now suppose that $\widehat{y}_{F,H}$ is consistent and unbiased for $y_{F,H}$ and that $\widehat{\bar{y}}_{F,F}$ is consistent and unbiased for $\bar{y}_{F,F}$, but that $\eta_{F,H} \neq 1$, $\eta_{F,F} \neq 1$, and $\nu \neq 1$; that is, assume that the remaining conditions in Result A.4 do not hold. Then the estimator

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{y}}_{F,F}}$$

is consistent and unbiased for $(\frac{\eta_F}{\eta_H}\nu)N_H$.

Proof. The proof follows along the lines of Feehan and Salganik (2016a). Briefly,

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{y}}_{F,F}} \rightsquigarrow \frac{y_{F,H}}{\bar{y}_{F,F}}$$

by the sampling conditions. Next, we wish to use the adjustment factors to relate the estimand to N_H:

$$\begin{aligned}
\frac{y_{F,H}}{\bar{y}_{F,F}} &= \frac{\eta_F}{\eta_H} \frac{y^+_{F,H}}{\bar{y}^+_{F,F}} \\
&= \frac{\eta_F}{\eta_H} \frac{v_{H,F}}{\bar{v}_{F,F}} \\
&= \frac{\eta_F}{\eta_H} \frac{\bar{v}_{H,F}}{\bar{v}_{F,F}} N_H \\
&= \frac{\eta_F}{\eta_H} \nu N_H.
\end{aligned}$$

Thus, we conclude that

$$\widehat{N}_H \rightsquigarrow \frac{\eta_F}{\eta_H}\nu N_H.$$

□

Corollary B.1. *Under the conditions listed in Result B.1,*

$$Bias[\widehat{N}_H] = \mathbb{E}[\widehat{N}_H] - N_H = N_H(\frac{\eta_F}{\eta_H}\nu - 1).$$

Now we show how problems with the sampling weights can affect estimates; this will be helpful in understanding what impact non simple random subsampling of detailed alters would have.

First, we must define *imperfect sampling weights*. We follow Feehan and Salganik (2016a) and repeat the definition here for convenience:

Imperfect sampling weights. Suppose a researcher obtains a probability sample s from the frame population F. Let I_i be the random variable that assumes the value 1 when unit $i \in F$ is included in the sample s, and 0 otherwise. Let $\pi_i = \mathbb{E}[I_i]$ be the true probability of inclusion for unit $i \in F$, and let $w_i = \frac{1}{\pi_i}$ be the corresponding design weight for unit i. We say that researchers have *imperfect sampling weights* when researchers use imperfect estimates of the inclusion probabilities π'_i and the corresponding design weights $w'_i = \frac{1}{\pi'_i}$. Note that we assume that both the true and the imperfect weights satisfy $\pi_i > 0$ and $\pi'_i > 0$ for all i.

Result B.2. *Suppose researchers have obtained a probability sample s, but that they have imperfect sampling weights. Call the imperfect sampling weights $w'_i = \frac{1}{\pi'_i}$, call the true weights $w_i = \frac{1}{\pi_i}$, and define $\epsilon_i = \frac{w'_i}{w_i} = \frac{\pi_i}{\pi'_i}$. Then*

$$Bias[\widehat{y}'_{F,Z}] = N_F \left[\bar{y}_{F,Z}(\bar{\epsilon} - 1) + cov_F(y_{i,Z}, \epsilon_i)\right],$$

where $\bar{\epsilon} = \frac{1}{N_F}\sum_{i \in F} \epsilon_i$ and $cov_F(\cdot, \cdot)$ is the finite population unit covariance in the frame population F.

Proof. See Result D.2 in Feehan and Salganik (2016a).

□

Result B.2 will be useful to us because we can use it to understand situations where respondents' reports about the detailed alters are different from simple random sampling.

Fact B.1.

$$\sum_{i \in A} a_i b_i = N_A[\bar{a}\bar{b} + \mathrm{cov}_A(a_i, b_i)]$$

Result B.3. *Suppose that respondents do not report about the detailed alters by picking r_i out of d_i of them uniformly at random, so that the estimator for $\widehat{y}_{F,Z}$ in Result A.3 uses imperfect weights $l'_{ij} = \frac{d_i}{r_i}$ for the final-stage subsampling of detailed alters, while the true weight for each of respondent i's detailed alters j is given by l_{ij}. Let $\epsilon_i = \frac{l'_i}{l_i}$. Then the bias of $\widehat{y}'_{F,Z}$ is given by*

$$Bias[\widehat{y}'_{F,Z}] = \sum_{i \in F} \sum_{j \sim i} z_{ij}(\epsilon_{ij} - 1).$$

Proof.

$$\begin{aligned}
\mathbb{E}[\widehat{y}'_{F,Z}] &= \mathbb{E}\left[\sum_{i\in s} w_i \times \mathbb{E}_i[\sum_{j\in s_i} l'_{ij} z_{ij}|s]\right] \\
&= \sum_{i\in F} w_i \mathbb{E}[I_i] \times \sum_{j\sim i} \mathbb{E}[I_{ij}|s] l'_{ij} z_{ij} \\
&= \sum_{i\in F}\sum_{j\sim i} \frac{l'_{ij}}{l_{ij}} z_{ij} \\
&= \sum_{i\in F}\sum_{j\sim i} \epsilon_{ij} z_{ij},
\end{aligned}$$

where $j \sim i$ indexes the people j that are reported in respondent i's network. Thus, the bias is

$$\begin{aligned}
\text{Bias}(\widehat{y}'_{F,Z}) &= \mathbb{E}[\widehat{y}'_{F,Z}] - y_{F,Z} \\
&= \sum_{i\in F}\sum_{j\sim i} \epsilon_{ij} z_{ij} - \sum_{i\in F}\sum_{j\sim i} z_{ij} \\
&= \sum_{i\in F}\sum_{j\sim i} z_{ij}(\epsilon_{ij} - 1).
\end{aligned}$$

□

To understand Result B.3 better, we manipulate the expression for $\text{Bias}[\widehat{y}'_{F,Z}]$ with the aim of producing a more interpretable expression:

$$\begin{aligned}
\text{Bias}(\widehat{y}'_{F,Z}) &= \sum_{i\in F}\sum_{j\sim i} z_{ij}(\epsilon_{ij} - 1) \\
&= \sum_{i\in F} y_i[\bar{z}_i(\bar{\epsilon}_i - 1) + \text{cov}_{j\sim i}(z_{ij}, \epsilon_{ij} - 1)] \\
&= \sum_{i\in F} y_i \bar{z}_i \bar{\epsilon}_i - \sum_{i\in F} y_i \bar{z}_i + \sum_{i\in F} y_i \sigma_i \\
&= \sum_{i\in F} z_i \bar{\epsilon}_i - \sum_{i\in F} y_i z_i + \sum_{i\in F} y_i \sigma_i,
\end{aligned}$$

where $j \sim i$ indexes the people j that are reported in respondent i's network; $y_i = y_{i,U} = \sum_{j\sim i} 1$ is the total number of people i would report about if there was no subsampling; $\bar{z}_i = y_i^{-1} \sum_{i\sim j} z_{ij}$ is the average z_{ij} among respondent i's reported network members; $\bar{z} = N_F^{-1} \sum_{i\in F} \bar{z}_i$ is the average $\bar{z}_i$ across people in the frame; $\bar{\epsilon}_i = y_i^{-1} \sum_{i\sim j} \epsilon_{ij}$ is the average ϵ_{ij} among respondent i's reported network members; $\bar{\epsilon} = N_F^{-1} \sum_{i\in F} \bar{\epsilon}_i$ is the average $\bar{\epsilon}_i$ across people in the frame; and $\sigma_i = \text{cov}_{i\sim j}(z_{ij}, \epsilon_{ij})$ is the covariance between the ϵ_{ij} and z_{ij} among respondent i's reported network members.

Finally, we use Fact B.1 twice–once within respondent and once between respondents:

$$\sum_{i \in F} z_i \bar{\epsilon}_i - \sum_{i \in F} y_i z_i + \sum_{i \in F} y_i \sigma_i = N_F[\bar{z}\bar{\epsilon} + \text{cov}_F(z_i, \bar{\epsilon}_i)] - y_{F,U} + N_F[\bar{y}_{F,U}\bar{\sigma} + \text{cov}_F(y_i, \sigma_i)]$$

$$= y_{F,U} \left[\underbrace{\frac{\bar{z}\bar{\epsilon} + \text{cov}_F(z_i, \bar{\epsilon}_i)}{\bar{y}_{F,U}}}_{\text{between respondents}} + \underbrace{\bar{\sigma} + \frac{\text{cov}_F(y_i, \sigma_i)}{\bar{y}_{F,U}}}_{\text{within respondents}} - 1 \right].$$

Thus, Equation **??** shows that when respondents do not choose detailed alters uniformly at random, the resulting bias can be decomposed into a term that is related to how much variation there is between respondents and a term that is related to how much deviation from simple random sampling there is within each respondent.

C Additional results

country	Conversational contact	Meal
Brazil	13.1 (12.5, 13.6)	6.3 (5.9, 6.6)
Colombia	10.5 (10, 11.1)	7.2 (6.9, 7.6)
Great Britain	12.7 (11.6, 13.9)	4.4 (3.7, 5.3)
Indonesia	11 (10.4, 11.6)	7.5 (7, 8)
United State	12.1 (11.6, 12.5)	5 (4.6, 5.4)

Table 2: Estimated average degree and 95% confidence interval, by type of personal network

Country	$\mathbb{E}[\Delta_{IC}^{cc} - \Delta_{IC}^{meal}]$
Brazil	0.18 (0.02, 0.46)
Colombia	0.21 (-0.04, 0.56)
Great Britain	0.18 (-0.05, 0.65)
Indonesia	0 (-0.2, 0.19)
United States	0.05 (-0.05, 0.19)

Table 3: Estimated sampling distribution of the difference in internal consistency check squared error for the conversational contact network minus internal consistency check squared error for the meal network. Positive values mean that the conversational contact network was less internally consistent than the meal network, as measured by squared error.