DRAFT – PLEASE DO NOT CITE

Social network methods for estimating adult mortality: evidence from Rwanda[*]

Dennis Feehan[†1], Mary Mahy[‡2], and Matthew J. Salganik[§1,3]

[1]Office of Population Research, Princeton Univeristy
[2]UNAIDS, Geneva, Switzerland
[3]Department of Sociology, Princeton University

June 23, 2011

[*]This research was supported by The Joint United Nations Programme on HIV/AIDS (UNAIDS), NSF (CNS-0905086), and NIH/NICHD (R01HD062366). The opinions expressed here represent the views of the authors and not the funding agencies.

[†]dfeehan@princeton.edu, Wallace Hall, Second Floor, Princeton, NJ 08544

[‡]mahym@unaids.org, Geneva, Switzerland

[§]mjs3@princeton.edu, 145 Wallace Hall, Princeton, NJ 08544

* * * DRAFT: UAPS extended abstract, please do not cite * * *

Abstract

Measurements of adult mortality are a vital part of understanding the health and well-being of populations everywhere. In countries that lack high-quality death registration data, including most of Sub-Saharan Africa, adult death rates must be estimated using alternative strategies. This study aims to enlarge and improve our arsenal of methods for doing so. We introduce the data-augmented network scale-up method, a new, survey-based technique that can be used to estimate adult death rates from respondents' reports about their social networks. We test this method using a nationally-representative survey of 5,000 Rwandans that will take place in June-August of 2011. Although no gold-standard estimates of adult mortality are available for validation in Rwanda, we will evaluate the plausibility of our results and compare them to other estimates that are available. We will conclude with a discussion of what implications our results have for improving the method in future applications.

1 Introduction and overview

Measurements of adult mortality are a vital part of understanding the health and wellbeing of populations everywhere. Death rates above age 15 are needed to compute life expectancy, to understand socioeconomic and occupational differences in health, and to measure the impact of large health events, such as the availability of new medical treatments for certain diseases, or the impact of HIV/AIDS. In the developed world, near-complete death registration and regular censuses provide scholars and policymakers with the data necessary to understand levels and trends in adult death rates over time. In the developing world, and particularly in Sub-Saharan Africa and Asia, these sources of information are often incomplete or unavailable (Setel et al., 2007). For example, for almost all of Sub-Saharan Africa and much of Asia, figures for life expectancy at birth that are compiled by international agencies are essentially statistical guesses (Reniers et al., 2011). Two leading experts in adult mortality estimated that 85% of the people in the developing world live in areas where no gold-standard measurement of adult mortality exists (Hill and Timaeus, 2004). This lack of information makes it impossible to ascertain the population-level effectiveness of health policies, such as the distribution of antiretroviral drugs to treat HIV-positive people or efforts to curb the smoking epidemic. It also hinders efforts to accurately project the age and sex composition of national populations in the future, which makes creating policy for labor markets, educational and health systems, and other social programs very difficult.

This study aims to help enlarge and improve our arsenal of methods for estimating adult death rates by providing the first empirical application of the data-augmented network scale-up method. This new, survey-based technique is an extension of the network scale-up method, which uses respondents' reports about members of their social networks to estimate the size of a hard-to-count group (Bernard et al., 2010; Killworth et al., 1998a). The data-augmented network scale-up method collects more information about the members of respondents' social networks, permitting us to estimate occurrence-exposure rates. In this application, we collect information on the age and sex of deaths to estimate age- and sex-specific death rates. If the assumptions of the data-augmented network scale-up estimator are valid, we would expect it to improve upon some of the shortcomings of a widely employed alternative, the sibling survival method.

The remainder of this extended abstract begins by describing the new estimator and its relation to the network scale-up method in more detail. Next, we present the first empirical estimates of adult death rates from the new method, using data from a nationally-representative survey of about 5,000 people in Rwanda. Since no gold standard measurements of adult death rates are available in Rwanda, we will use a range of demographic techniques to evaluate the plausibility of our estimates. If the results of the 2010 Rwanda DHS are available in time, we will also be able to compare our estimates to the ones produced by sibling survival module on that survey; additionally, we will compare our estimates to other sources, like the estimates produced by the UN Population Division, and the estimates produced by the National Institute of Statistics of Rwanda. Finally, we conclude with a discussion of how these preliminary empirical results suggest that the data-augmented network scale-up method be modified or improved in future studies.

2 Social network methods for estimating the size of a target population

We begin with a description of the network scale-up method, upon which our new estimator is based. We then explain why the network scale-up method cannot produce estimates of the quantities most relevant for the study of adult mortality, and proceed to introduce the data-augmented network scale-up method, which can.

2.1 The network scale-up method

The network scale-up method is based on the assumption that members of people's social networks are, on average, representative of the general population (Bernard et al., 2010). On a standard survey, respondents are asked to report about themselves, and inferences about the general population are made from those responses. On a network scale-up survey, respondents report about members of their networks, and these reports are used to estimate the sizes of hard-to-count groups. In the literature, these hard-to-count groups are often called target populations.

The scale-up method proceeds in two steps, which we will elaborate in more detail below. The first step is to estimate how large each respondent's social network is, and the second step is to estimate what fraction of each respondent's network is in the target population.

The intuition is given in Figure 1, where the members of the general population are represented by circles. We wish to estimate the size of the target population which, here, is the circles colored grey. A quick count shows that, in reality, there are 30 members of the whole population, six of whom are grey. To produce an estimate, we begin by taking a sample of the population members. In this simple example, our sample has size 1 and the selected respondent is colored black. We ask the respondent a series of questions that estimate the number of people in her social network; here, this is 10. We also ask her how many of the people in her network are grey. In this case, the response is 2. The result is that we estimate that the size of the grey population is $\frac{2}{10} \times 30 = 6$.

Mathematically, the scale-up estimator can be written as

$$\hat{N}_t = \frac{\sum_i y_i}{\sum_i \hat{d}_i} \times N, \tag{1}$$

where N_t is the size of the target population to be estimated, N is the size of the total population, y_i is the number of members of the target population reported to be known by the ith survey respondent, and $\hat{d}_i$ is our estimate for the network size of the ith survey respondent (Killworth et al., 1998a).

By asking about the people in each respondent's network, we learn about many more people than we interview. This is a big advantage when we are trying to measure rare populations, whose small size means that a direct approach to measuring them would require very large samples. In this way, the NSUM estimator is similar to the sibling survival estimator.

Definition of 'to know' In order to employ the network scale-up estimator, we must first to establish is what it will mean for a respondent to know someone (Bernard et al., 2010; Killworth et al., 1998b). In Figure 1, this is the definition that determines whether or not a pair of circles is connected with an line. A commonly used definition in the network scale-up literature is to tell the survey respondent that she knows someone if:

- you know the person and the person knows you, by sight or by name
- you have been in contact with the person over the past two years, either in person, on the phone, by mail, or on line
- you could contact the person if you needed to

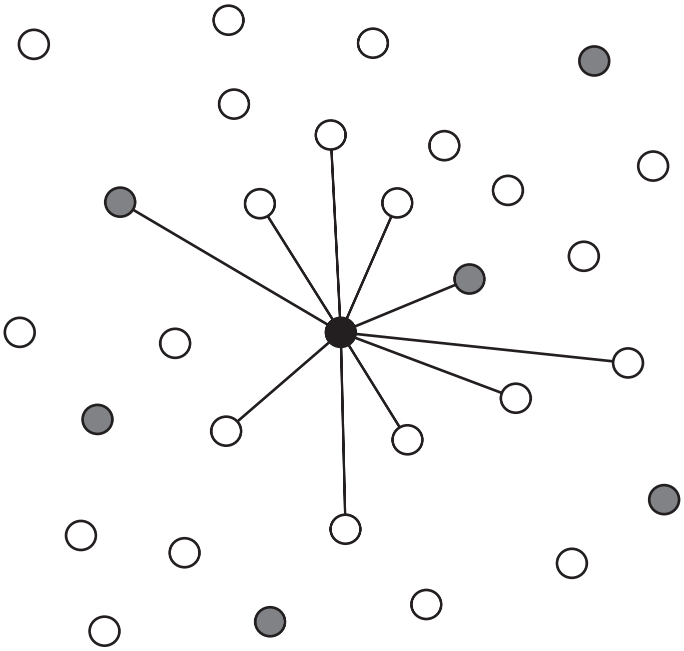

Figure 1: Illustration of the network scale-up estimator for the size of a target population. In the figure, each circle represents a member of the population. Those colored grey are members of the target population whose size we'd like to measure. The black circle is our survey respondent. Each of the people known by the respondent is indicated by a line. The respondent reports knowing two members of the target population; her total degree is 10, so the scale-up estimate of the total population size is $\frac{2}{10} \times 30 = 6$.

Many other definitions are clearly possible. The NSUM study in Rwanda will explore two different definitions in an attempt to determine which one is most effective, and how sensitive estimates are to the choice of definition. In general, we expect a good definition of 'to know' to be clear, easy to remember, and likely to be applicable in many different settings.

Estimating the size of respondents' networks Once we have settled upon a definition of to know, there are several methods available for estimating the size of each respondent's personal network; the one we consider here is called the known population method (Bernard et al., 2010). We begin by identifying several populations whose total size is known; for example, administrative records may give us a total count of the number of mailmen in the population. We then ask each survey respondent how many mailmen she knows; we expect the number she reports to be proportional to the size of her network. If there are 10,000 mailmen in a country with a population of 1 million, and a respondent reports knowing one of them, then we estimate that the size of her personal network is

$$\frac{1}{10,000} \times 1,000,000 = 100.$$

By asking questions about many populations of known size (typically about 20), we can obtain a more precise estimate of the size of each respondents network. Mathematically, we compute our estimate of the size of respondent i's personal network, $\hat{d}_i$, with

$$\hat{d}_i = \frac{\sum_k y_{ik}}{\sum_k N_k} \times N, \tag{2}$$

where k indexes the known populations, y_{ik} is the number of people respondent i reports knowing in known population k, N_k is the total size of known population k, and N is the total size of the whole population (Killworth et al., 1998a).

Once we have estimated the size of each respondent's network, the application of the NSUM estimator is straightforward, and requires only a question on the number of members of the target population the respondent knows.

There are many advantages and disadvantages to the NSUM approach; a thorough discussion of them can be found in Bernard et al. (2010). Below, we will briefly review some key points in the context of the modified estimator that we propose for mortality rates.

Applications to date To date, the NSUM approach has been used mainly to estimate the size of groups of interest to the public health community – for example, the size of populations most at risk of HIV/AIDS: injecting drug users, men who have sex with men, and female sex workers; Bernard et al. (2010) has a complete description of the various application of the method to public health. Although in many cases the estimates from the scale-up approach have been in broad agreement with other strategies for estimating the sizes of these groups, there have been few opportunities to validate the method by using it to try and measure the size of a group whose total is known (but not used in estimating the social network sizes).

2.2 The data-augmented network scale-up estimator and its application to adult mortality rates

The NSUM estimator could measure crude death rates simply by taking the target population to be the number of people who have died over a given time period[1]. Unfortunately, this quantity alone is typically of little use to researchers and policymakers, since the number of deaths in a population is dramatically affected by its age structure. For example, recent UN figures put the crude death rate in Sweden to be about 10 deaths per 1,000 people, while the same figure for Mexico is about 5 deaths per 1,000 people (United Nations Population Division, 2008). However, it would be misleading to conclude from those numbers that Mexicans are, on average, healthier than Swedes. Instead, Mexico's population is much younger than Sweden's, and younger people everywhere are less likely to die than older people. In general, scholars and policymakers need estimates of death rates to be broken down by age and sex to be useful.

Recall that the age-specific death rate is defined as

$$M_a = \frac{D_a}{N_a}, \tag{3}$$

D_a is the number of deaths occurring to members of the population who were age a, and N_a

[1] In this case, strictly speaking, the definition of 'to know' should not require that the respondent be able to get in contact with the people she knows. In practice, experience leads us to predict that this will not have an impact on respondents' reports, especially if interviewers are well trained. Nonetheless, this will obviously be a consideration in choosing the definition of 'to know' that will be used in the study.

is the number of person-years of exposure at age a in the population. The NSUM approach can be adapted to estimate this quantity: for each death the respondent reports, we can collect the age and sex of the decedent. This information will permit us to estimate the total number of deaths and also their distribution by age and sex. Putting this in terms of Equation 1:

$$\widehat{D_a} = \frac{\sum_i y_{i,a}}{\sum_i \widehat{d_i}} \times N \tag{4}$$

Where $\widehat{D_a}$ is the estimated number of deaths in age group a in the population, and $y_{i,a}$ is the number of deaths in age group a reported known by respondent i. Equation 4 is thus an estimator for the numerator of the age specific death rate in Equation 3.

The difficulty is then to estimate the exposures that are in the denominators of the death rates we wish to compute from Equation 3. The total size of the population we learn about is easily estimated as the sum of the respondents' network sizes, $\sum_i \widehat{d_i}$, but in order to compute death rates, we need to know the age-sex distribution of those people. Put another way, in order to produce estimates of death rates, we need to be able to estimate the age and sex distributions of the acquaintances that respondents report about. This would be hard to measure directly, because it is difficult to obtain a random sample of a respondent's acquaintances, and it is implausible that a respondent would be able to accurately list all of them. A first approach would to assume that the age-sex distribution of people's acquaintances is the same as in the population as a whole; then, we could simply use the age-sex distribution of the survey respondents as an estimate of the age-sex distribution of their acquaintances.

A better estimate of the age-sex distribution of the set of acquaintances being reported about by all of the respondents makes use of the estimated degrees of the respondents themselves. If the average 20 year old knows twice as many people as the average 60 year old, we can assume that the average 20 year old shows up in the set of acquaintances that respondents report about twice as often as the average 60 year old. Mathematically, we can estimate the size of the set of acquaintances in age group a, P_a, using:

$$\widehat{P_a} = \frac{\sum_{j \in S_a} \widehat{d_j}}{\sum_{i \in S} \widehat{d_i}} \times N, \tag{5}$$

where S_a is the survey respondents in age group a, and S is all of the survey respondents. $\widehat{P_a}$ is thus the product of an estimate of the fraction of respondents' acquaintances in age group a and the total population size. This simple estimator for P_a we currently propose might be improved upon through the incorporation of information about the age-sex distribution of the known populations; if that information is available, the model proposed by McCormick and Zheng (2011) could be explored as a method for estimating the denominator of death rates.

Put together, the new estimator for age-specific death rates is

$$\widehat{m_a} = \frac{\widehat{D_a}}{\widehat{P_a}}, \tag{6}$$

where $\widehat{D_a}$ is estimated using Equation 4, and $\widehat{P_a}$ is estimated using Equation 5. Since this estimator is based on the idea of the network scale-up method, but requires us to collect more information about the members of respondents networks, we call it the data-augmented network scale-up method (DANSUM). It permits us to estimate occurrence-exposure rates from a network scale-up survey.

Disadvantages There are several disadvantages to the DANSUM approach to estimating death rates. The first is that it requires us to estimate the size of each respondent's network[2]. As we discussed above, our preferred way of measuring the size of respondents' networks requires a set of population groups whose total size is known (Bernard et al., 2010). Studies in the US and similar countries have used high quality administrative data to locate subgroups whose size is known; in the example above, we used the number of mailmen in the country. In some settings, especially in Sub-Saharan Africa, a list of groups whose total size is known may be hard to identify. However, it is worth pointing out that these quantities could be estimated from the same survey used to administer the NSUM questions. For example, a nationally representative household survey could ask female respondents whether or not they gave birth in the past year; the responses to that question allow us to estimate a national total number of women who gave birth in the past year. We could then use the estimated total number of women who gave birth in the past year

[2] Actually, a closer look at Equation 1 reveals that we only need to be able to estimate the sum (or, equivalently, the mean) of the sizes of the respondents' networks.

as one of the known populations in estimating the size of each respondent's network. No studies to date have used known populations derived from the same survey as the scale-up questions. However, some of the known populations to be used in the Rwanda survey will have totals derived from the Rwanda DHS. These totals will be estimated, with known sampling error, unlike many of the other values used for the known populations.

Another important disadvantage in the DANSUM approach is the assumption that respondents are able to perfectly report on all of their acquaintances. This is especially problematic if we are interested in measuring the size of populations whose behavior is highly stigmatized, like injecting drug users. It is likely the case that not all of the people who know injecting drug users are in fact aware of that fact. Although this is an important concern for some target populations, it is less of a problem for mortality, so we will not discuss it in more detail here; however, there are strategies for measuring and adjusting NSUM estimates due to this problem (Salganik et al., 2010).

Differences in the networks of people who die and those who do not are also a potential challenge to the DANSUM estimator. For example, if the definition of 'to know' requires acquaintances to have seen each other in the past 12 months, then someone who died 9 months prior to the survey and who did not go to work in the 3 months before his death would not count as having been known by his work colleagues. This could be exacerbated in situations where deaths are preceded by long illnesses. Methodologically, the key issue here is that people who die might have, on average, smaller social networks than people who do not. Empirical results from the DANSUM method will help us understand how much of a problem this might be. Carefully chosen definitions of 'to know' could help to ameliorate this issue; also, if we collect additional data, we can use the methods of Salganik and Feehan (2011) to produce DANSUM estimates that account for differences in network size.

Advantages On the other hand, there are several important advantages to the DANSUM approach. The first is that it can produce estimates for the size of a target population from a household survey. Furthermore, we expect it to be less expensive than alternative methods: since each interview yields information about many more than one person, the size of the sample required to produce accurate estimates will be considerably smaller than a direct estimation approach. Put another way, as long as its assumptions hold, DANSUM can produce quite precise estimates for a given sample size. Another attractive feature is that it is possible to use the method to estimate size of the groups whose total sizes are known. Recall, from Equation (2), that we use several groups whose total size is known to estimate the network size of each respondent. We can take each of these known populations in turn, pretend it is not known, use the remainder of the known populations to estimate the respondents' network sizes, and then the scale-up method to estimate the size of the held out group. We can therefore check to see how accurately we can estimate the size of each of the known groups. Mathematically, if we are holding out known population j, the modified version of Equation (2) would be:

$$\widehat{d_{i,-j}} = \frac{\sum_{k \neq j} y_{ik}}{\sum_{k \neq j} N_k} \times N, \tag{7}$$

and the scale-up estimate of the size of known group j, $\widehat{N_j}$, would be

$$\widehat{N_j} = \frac{\sum_i y_{ij}}{\sum_i \widehat{d_{i,-j}}} \times N. \tag{8}$$

An example of this internal validation exercise, from the study in Curatiba, Brazil, is shown in Figure 2 (Salganik et al., 2010).

A concern often raised with the sibling survival approach is that sibships who have entirely died out will not show up in the sample. The analogous risk here – that a set of people who all know each other will die out – is presumably so small that it is negligible, except in very extreme cases. Another concern is that siblings may live far apart or generally be removed from one another's experience, leading to less accurate reporting. This is something that DANSUM can address by choosing definitions of 'to know' which ensure that respondents are asked to report about people they are most acquainted with. In general, the tradeoff between the advantages and disadvantages of the sibling survival and DANSUM estimates of mortality is an empirical question.

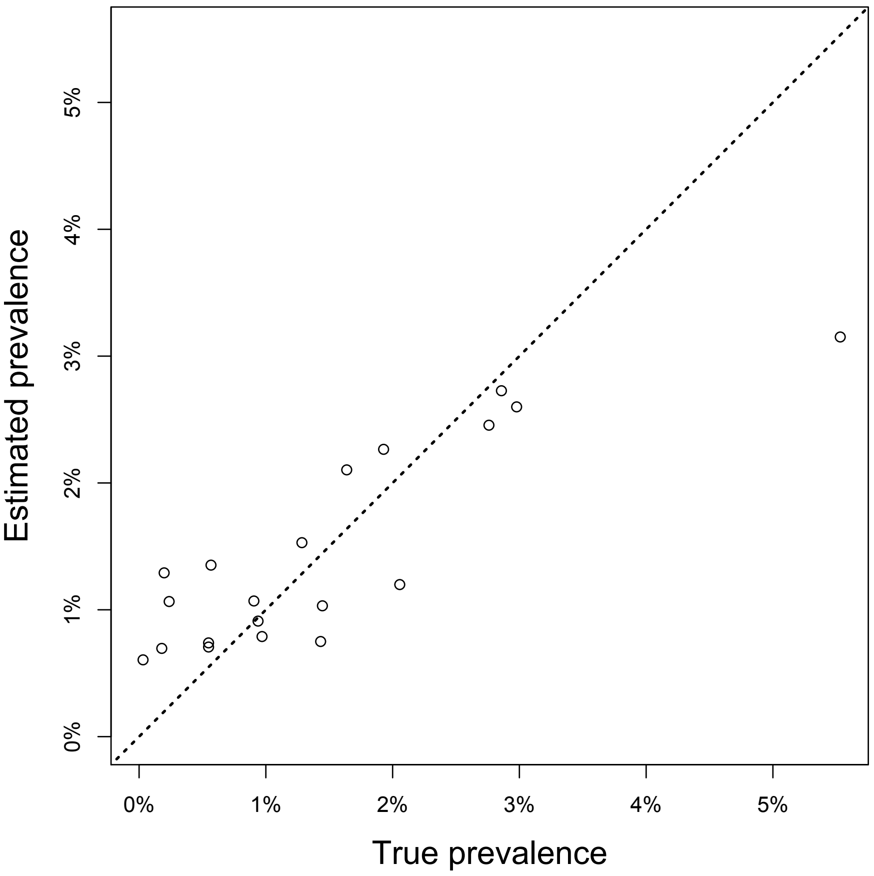

Figure 2: An example, from Salganik et al. (2010), of an internal validation check that the NSUM approach allows us to perform. Each point is a subgroup of known size in the general population. Each survey respondent was asked how many of the members of each known subgroup she knew, and the responses were used to estimate the size of each respondent's personal network. By removing each of the known subgroups, one at a time, and treating it as unknown, we can use the NSUM approach to estimate its size (see Equations 7 and 8). The relationship between the true subgroup sizes (x axis) and the NSUM estimates (y axis) are shown here. If the NSUM estimate is exactly correct, it would lie on the dashed line. Points below the dashed line are underestimated, while those above are overestimated. We see that the estimated values are clustered closely around the true ones, with the exception of the outlier whose true size of about 5.5% is underestimated as about 3%. This group is middle school students, and it would make sense that relatively few of the adult respondents' acquaintances would be in this group, leading to an underestimate.

3 Data and Methods

After the data for the next Rwanda Demographic and Health Survey had finished being collected, a stratified, two-stage cluster sample of approximately 5,000 people was drawn and, between June and August of 2011, those people will be interviewed as a pilot of the network scale-up method for measuring the size of populations most at risk of HIV/AIDS. The study is intended to mimic a DHS, so it uses parts of the same survey team and survey process that the 2011 Rwanda DHS did. Each sampled household will randomly be assigned to one of two possible definitions of 'to know,' permitting a study of the extent to which varying the definition has an impact on estimates. This experimental test of the role of the specific definition of 'to know' in the implementation of network scale-up studies will add to our understanding of how the method works in practice, which will begin to build a base of knowledge about various potential definitions of 'to know' that will be an important part of the design of future network scale-up studies.

The survey instrument includes the items necessary for our network-based estimator for adult death rates. Since the 2011 Rwanda DHS collected the sibling history module, we will eventually be able to compare the estimates of adult death rates produced by the two methods. Although we will not have gold-standard measurements of adult death rates to compare the sibling and network-based estimates to, it will be interesting to see if the relationship between the two sets of estimates is, and also to compare them to other estimates for adult mortality in Rwanda.

4 Conclusion

This extended abstract outlines a study that is intended to help expand and refine our understanding of how we can estimate adult mortality rates in settings where gold-standard measurements are not available. First, we describe the data-augmented network scale-up method, a new technique that can be used to estimate adult mortality rates based on survey questions about respondents' social networks. We then present the first empirical application of the new method, using data collected in a survey of about 5,000 people in Rwanda. We will be able to compare the new estimator to estimates from other sources, eventually including estimates from the sibling survival method. We will conclude with a discussion of how our results suggest that we develop and refine the method for use in future studies.

References

Bernard, H. R., Hallett, T., Iovita, A., Johnsen, E. C., Lyerla, R., McCarty, C., Mahy, M., Salganik, M. J., Saliuk, T., Scutelniciuc, O., Shelley, G. A., Sirinirund, P., Stroup, D. F., and Weir, S. (2010), "Counting hard-to-count populations: the network scale-up method for public health," *Sexually Transmitted Infections*, Forthcoming.

Hill, K. and Timaeus, I. M. (2004), "Unconventional Approaches to Mortality Estimation," in *The First Human Mortality Database Symposium*, Max Planck Institute for Demographic Research.

Killworth, P. D., McCarty, C., Bernard, H. R., Shelly, G. A., and Johnsen, E. C. (1998a), "Estimation of seroprevalence, rape, and homelessness in the U.S. using a social network approach," *Evaluation Review*, 22, 289–308.

Killworth, P. D., Johnsen, E. C., McCarty, C., Shelly, G. A., and Bernard, H. R. (1998b), "A social network approach to estimating seroprevalence in the United States," *Social Networks*, 20, 23–50.

McCormick, T. and Zheng, T. (2011), "Latent demographic profile estimation in at-risk populations," Working paper.

Reniers, G., Masquelier, B., and Gerland, P. (2011), "Adult Mortality Trends in Africa," in *International Handbook on Adult Mortality*, eds. R. G. Rogers and E. M. Crimmins, Springer.

Salganik, M. and Feehan, D. (2011), "Generalized network scale-up method for estimating the size of hard-to-count populations," Tech. rep., Office of Population Research, Princeton University.

Salganik, M., Mello, M., Abdo, A., Bertoni, N., Fazito, D., and Bastos, F. (2010), "The game of contacts: estimating the social visibility of groups," *Social Networks*.

Setel, P., Macfarlane, S., Szreter, S., Mikkelsen, L., Jha, P., Stout, S., and AbouZahr, C.

(2007), “A scandal of invisibility: making everyone count by counting everyone,” *The Lancet*, 370, 1569–1577.

United Nations Population Division (2008), “The World Population Prospects: The 2008 Revision,” .