Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

From theory to practice: Lessons learned from implementing the Network Sampling with Memory method

by Aurélie Santos and Géraldine Charrance

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada's Symposium 2022
Data disaggregation: Building a more representative data portrait of society

From theory to practice: Lessons learned from Implementing the Network Sampling with Memory method

Aurélie Santos and Géraldine Charrance[1]

Abstract

To overcome the traditional drawbacks of chain sampling methods, the sampling method called "network sampling with memory" was developed. Its unique feature is to recreate, gradually in the field, a frame for the target population composed of individuals identified by respondents and to randomly draw future respondents from this frame, thereby minimizing selection bias. Tested for the first time in France between September 2020 and June 2021, for a survey among Chinese immigrants in Île-de-France (ChIPRe), this presentation describes the difficulties encountered during collection—sometimes contextual, due to the pandemic, but mostly inherent to the method.

Keywords: Sampling; collection; hard-to-reach population.

1. NSM: an innovative, ambitious method for studying hidden populations

1.1 Chinese immigrants in Île-de-France, a hard-to-survey population

Schiltz identifies the 1990s as the time when social surveys covering "marginal" populations began appearing in public statistics (Schiltz, 2005). The need for data is important for these "marginalized" populations, often in precarious situations or exposed to social or health risks. However, these populations are hard to survey because they are small, their members cannot be identified in a survey frame, and they may even want to conceal the fact that they are a member of the population of interest to avoid being stigmatized (Marpsat and Razafindratsima, 2010). Using general population surveys to capture these populations usually does not produce a large enough sample to conduct statistical analyses. These constraints therefore rule out the use of probabilistic surveys in the general population to capture and characterize these populations.

Chinese immigrants in Île-de-France fall into this category of hard-to-reach populations: in the *Trajectoires et Origines* survey (Beauchemin et al., 2016) conducted in 2008, only 69 respondents were born in China. However, data sources such as the French census, the *échantillon démographique permanent* (EDP) or *Fidéli* contain little information to characterize these groups of immigrant populations and can comprise a lack of coverage (missing migrants in irregular situations).

1.2 Network-based survey methods particularly suited to this type of population

Chain sampling methods involve surveying within networks by having respondents recruit new respondents. Therefore, they do not require a survey frame. There are also other rationales for using these methods: a belief that respondents are more likely to participate when they are invited to respond by acquaintances rather than being contacted at random; and a desire to understand the structure of networks and the role of social relationships.

[1]Aurélie Santos, Institut national d'études démographiques, 9 cours des Humanités, CS 50004, 93322 Aubervilliers, France (aurelie.santos@ined.fr); Géraldine Charrance, Institut national d'études démographiques, 9 cours des Humanités, CS 50004, 93322 Aubervilliers, France (geraldine.charrance@ined.fr).

These methods have long suffered from the reputation of producing biased estimates. However, recent studies based on the respondent-driven sampling (RDS) method have shown that, under certain assumptions, the likelihood of being sampled could be estimated by the number of connections each respondent has with other population members. The assumptions are: (1) The population consists of a single network: all individuals are connected to one another, more or less directly (essential assumption for mobilizing Markov chain theory and thus arguing in favour of convergence of the estimators toward unbiased estimators). (2) The connections between individuals are reciprocal. (3) Individuals randomly recruit from among their acquaintances. Despite this, the accuracy of the method remains problematic: while the sampling variance of a simple random sample is inversely proportional to the sample size, the accuracy of RDS depends on both the sample size and network structure.

Therefore, the objective of this paper is to mobilize the network sampling with memory (NSM) method. The basic random walk approach is improved by incorporating information about the network's local topography and recently sampled cases in order to improve efficiency during sampling (Avin and Krishnamachari, 2008). The NSM method supposedly improves sampling efficiency from a network by collecting data on the network from the respondents, who are used to gradually reveal the list of population members. Over time, the list of designated people in the survey tends to resemble the full list of population members, which theoretically helps NSM to be more like a simple random survey.

1.3 Contributions of the network sampling with memory variant

The designers of the network sampling with memory method wanted to develop an asymptotically unbiased method (like the RDS approach when its assumptions are met) and whose sampling variance and therefore the mean absolute error decreases rapidly with the size of the sample collected. For this, the NSM method uses two sampling modes operating in tandem to improve the accuracy of estimates: a "list" mode that guarantees unbiased asymptotic estimates (naive list mode and even sampling) and a "search" mode. The sampling process is used to uncover the list of population members (search), then sampling is done within that list with replacement (list). At the end of the questionnaire, each respondent is asked to provide the contact information of their friends who are members of population A (this list is referred to here as a "roster"). At each stage, new members of the network are added to the list (total roster), which contains all population members identified by respondents. The modes come into play successively throughout collection: first the naive list, then search, and finally even sampling. To determine the mode to mobilize, indicators that reveal the network exploration level are used.

Naive list mode: To start collection randomly
Collection starts with identifying seeds, i.e., the first survey respondents. Their responses generate an initial network. At the start of collection, an individual is randomly selected from among the first ones identified in the network (called "nodes"). With very little knowledge of the network at that point, the simplest, most egalitarian survey technique should be used: the simple random survey.

Search mode: To explore the network
The objective is to survey unexplored parts of the network. This involves identifying respondents who are most likely to lead us there: these individuals are called "node-bridges" (bridges between subparts of the network). For each respondent, their likelihood of being a bridge is calculated, based on the proportion of people referred once among their "friends." After identifying five individuals most likely to be node-bridges, one is selected (in proportion to their likelihood of being a node-bridge), then one of their friends is randomly drawn from among those identified only once and not surveyed. This is two-stage selection.

Even sampling mode (List mode): To homogenize exposures to selection
To mitigate the effect of oversampling of the nodes that appear early, the even sampling mode homogenizes the cumulative sampling rates (CSRs). For each selection, regardless of the mode, the CSR of individuals exposed to sampling is incremented by their likelihood of being selected. Updated when the individual is a candidate for sampling, the CSR measures individuals' exposure to successive selections. Thus, homogenizing the CSRs involves exposing new individuals or those never selected in the past to selection. This is done by excluding from the draw the nodes most exposed to selection previously and that have a significant CSR. Here, a simple random draw is used among nodes with a CSR lower than the even sampling rate (ESR) (or the 100 lowest CSRs if the volume of nodes who meet the condition is less than 100).

The output of the search mode is based on three criteria: (1) have collected more than 50 questionnaires, (2) have a network size greater than 200 or if the last 5 interviews did not uncover new individuals and finally (3) when the parameter P1 (number of individuals referred only once and not surveyed) falls below threshold A1 (established by the method's designers). When P1 is low, it means that many individuals have been referred multiple times, in other words, that a certain exploration level has been reached. It is then appropriate to use even sampling to submit previously ignored individuals to the draws and to balance the CSRs. In practice, the first two criteria were met fairly quickly. It was the P1<A1 criterion that seemed to be the most discriminating, the one that forced us to remain in search mode until the end of collection.

2. From theory to practice: Complex implementation of the NSM method

Data collection for the ChIPRe survey ran from September 2020 to June 2021 and involved 10 interviewers and a team of 5 collection managers (computer scientist, statisticians, design engineer). The next part covers our experience of putting the NSM into production. Despite intense preparation and some subsequent automation of the collection processes (matching and selection algorithms, web-based collection and tracking application designed ad hoc), many adjustments during collection were required.

2.1 Finalizing the sampling and collection protocol "on the spot"

2.1.1 Developing of the selection principles by "trial and error"

The first difficulty that the management team encountered was developing rules governing selection. The method involves tracking the gradual revealing of a network, and therefore requires regularly renewing the draws in order to populate the list of individuals "to be surveyed" and progress within that network. Since the method had undergone very little testing, we did not have any recommendations about the optimal frequency and size of the draws. Regarding the frequency, we started by producing very small draws (between 5 and 10 individuals) twice a week. That strategy made it possible to closely monitor the evolution of the network at the start of collection, but soon proved to be too time-consuming. We then decided to do only one draw per week.

To determine the number of individuals drawn in each one, we used the following parameters: the remaining survey load for each interviewer; their load of relaunching and identifying incomplete rosters; the size of the network at the time of selection compared with its size in the previous one. The downside is that duplicates must first be identified in order to determine the number of new individuals added to the total roster. After various experiments, we decided to produce draws never exceeding half of the new entrants in the network since the previous draw.

Once the draw was completed, there was still the sampled individuals to be divided among the interviewers. We started out with an objective of equity among interviewers: giving each one the same number of new cases in order to level the playing field for making progress in the network. That choice soon proved to be untenable because it was at odds with adhering to "affiliations." The interviewers wanted to follow "their" network, in other words, survey the referrals of "their" respondents (if sampled). However, equitably distributing the sampled people among them involved far too costly "affiliation transfers" (negotiation with a respondent was required to have their network be surveyed by another unknown interviewer!) We therefore opted for a principle of strict following of "affiliations" when dividing up the sampled people. With these various constraints, it was possible to program a total number of sampled people, but it was impossible to anticipate the new load that each interviewer would receive with each draw. This led to a high level of complexity with managing the team in the field.

2.1.2 Changing the sampling algorithm along the way

Putting the algorithm into production during actual collection brought various surprises that simulations on synthetic networks did not predict. We adapted the calculation criteria to determine the list of rosters eligible for each draw in search mode, excluding the nodes previously sampled in order to, among other things, limit artificial overexposure to the drawing of the rosters that were already candidates for the previous draw.
Faced with the very slow pace of progress in the network and in terms of our objectives and the duration of the survey, we decided to increase threshold A1 (see part 1) in order to have a greater chance of reaching it and of trying out the

"output" from search mode. That threshold was not arbitrarily decided on; it was drawn from the NSM method's founding article, in which various versions of that threshold are presented (Mouw, 2012).

2.1.3 Adaptations of the protocol to address realities in the field

Due to the slow progress in the network, we had to extend until June 2021 the duration of the surveying, which was originally supposed to end in late March 2021. That extension was accompanied by a partial renewal of the team of interviewers and, therefore, new training during the survey.

Due to the difficulties accessing certain social and regional groups, and in order to offset the gradual extinction of certain branches of the network, new seeds were recruited and added during collection. People from the Wenzhou region, students, and people in the most precarious situations were given special attention in this regard. The reality in the field soon required us to be more flexible in awarding the gift vouchers to thank the respondents. The protocol provided for a €15 voucher for the questionnaire and a €20 one for filling out a roster with 6 informed contacts. We felt it was more realistic to distribute the roster voucher using three contacts referred.

Finally, implementing the NSM method, without a practical guide for transitioning it to production, led to a lot of trial and error to determine how the algorithm works, to design a strategy of proportionate draws, adequate tracking of the seeds, and more realistic protocols.

2.2 Discovering an algorithm with sometimes counterintuitive and disincentivizing effects

2.2.1 The paradoxical preference for small rosters with no duplicates

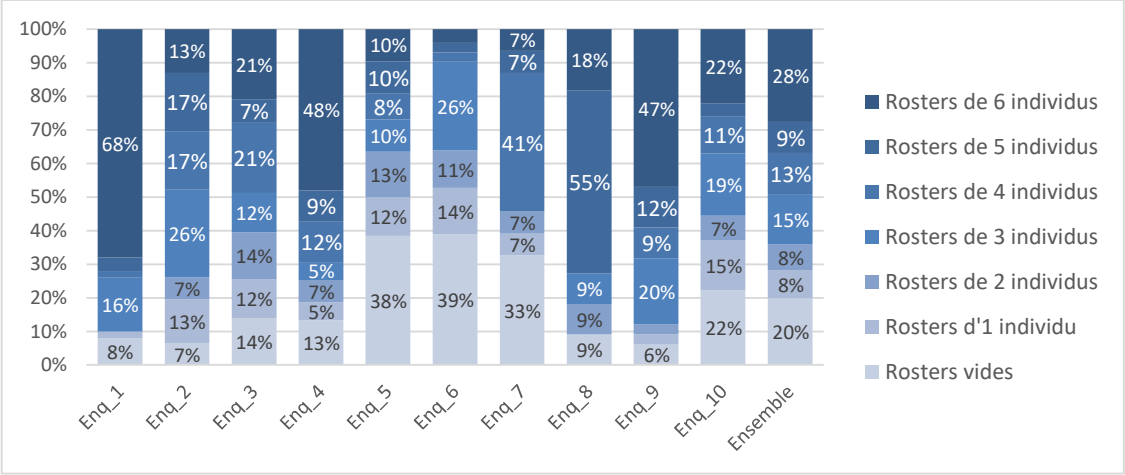
While our instructions to the interviewers stressed the importance of collecting large rosters, we gradually realized that the algorithm instead favoured small rosters during the draws. Based on our interpretation, this "perverse effect" is due to the fact that the algorithm considers small rosters as likely describing networks of more isolated people, or potential node-bridges to other harder-to-access subnetworks.

Another finding: the algorithm also discriminates against rosters containing duplicates (a duplicate is an individual previously referred by another respondent). However, the search for duplicates is among the objectives of the survey: the more duplicates there are, the lower the network's saturation indicator (P1, see part 1) and the more likely it is that to go below the output threshold of the Search mode (A1) to switch to even sampling. Therefore, the interviewers are trained to collect enough information to make identifying duplicates possible thanks to the matching algorithm designed specifically for the survey.

2.2.2 … and leading to significant interviewer strategies and biases

These trends by the algorithm are counterintuitive and counterproductive because the interviewers eventually become aware of them and adjust their behaviour to minimize the collection efforts. One interviewer quickly understood these two effects and began collecting small rosters containing 1 to 3 referrals, thereby being sure to limit the presence of duplicates and maximize her chances of having her rosters selected in the next draw (Figure 2.2.2-1, Enq_6). As such, she was certain of always having a full list "to survey," while providing the minimum effort to negotiate the rosters. However, progress in the network, and ultimately switching from one sampling mode to the other, is based on a collective effort for revealing as many connections as possible in the surveyed network. In addition, the structure of the network turns out to be biased; its respondents likely had larger networks than what their roster ultimately showed, and its respondents on the empty roster likely had a network, more difficult to negotiate, but not non-existent.

Figure 2.2.2-1
Distribution of the collected rosters by size (excluding seeds) per interviewer



Without going as far as this stage of rationalization, the "anti-duplicate" bias is still a disincentive: the presence of a duplicate is not perceived as good news by the interviewers because it implies that the roster will not be given priority in the draw. Therefore, the work of collecting identifying information (the most difficult to negotiate) to detect duplicates is not valued. The algorithm does not encourage interviewers to "produce" duplicates.

2.2.3 A sampling timeline that is far from the in-field reality

Finally, the algorithm's time line is not the same as the one in the field. Some of the rosters collected in December may have been eligible for sampling in March, requiring difficult negotiation work in the field for the interviewers.

These sampling results are even less accepted by the interviewers because they are combined with others, as described above: if the algorithm mobilizes older, less used rosters, it is also because the current supply of rosters is less conducive to being sampled, due in particular to duplicates. In addition to the discouragement associated with the non-selection of rosters of significant size with well-identified duplicates, there is the difficulty of negotiating the continuation of collection with respondents called upon several months earlier.

In conclusion, the algorithm produces counterintuitive effects making collection management difficult. An ongoing instruction effort by the management teams was required to make the collection process acceptable and, to the extent possible, consistent with the principles of the NSM method. Finally, the method requires an in-person operating framework, which we were deprived of by the COVID-19 pandemic. The suspending of the in-person aspect brought on by the second lockdown and the return of the epidemic in the fall significantly impacted collection.

2.3 Dealing with a public health context very unfavourable to in-person surveys

2.3.1 A population all the more difficult to reach in a public health crisis

Early on, the Chinese immigrants in Île-de-France showed a great deal of distrust associated with the health risks, having been informed of the unfolding of the epidemic in China. The returns from the field showed that some of them had other concerns in mind, after suffering heavy losses with the closing of businesses and restaurants, or because they had to maintain a connection with family members in China, who themselves were often struggling because of the epidemic.

The interviewers also witnessed a shrinking effect in the socializing network: respondents were not seeing friends and acquaintances as much as before the crisis, and it was less appropriate for them to be calling upon them for a questionnaire survey.

Some subpopulations proved to be particularly inaccessible virtually. The Wenzhou seem to belong to very closed networks, with limited availability and little interest in the research problem expressed. Undocumented immigrants

are a vulnerable subpopulation and not quick to refer acquaintances, especially remotely. They often live in shared dwellings where privacy is lacking in order to properly respond to the survey by phone. Strong involvement in the field to build a relationship of trust with these respondents was necessary at the start of collection, but the interruption of face-to-face interviews made it impossible to explore this network as expected.

2.3.2 Negotiating tools considerably impaired without face-to-face interviews

Collecting roster lines is based on the trust established when running the questionnaire, and the negotiation phase for obtaining the contact information of the respondents' friends is much easier to undertake in person than virtually. The interviewer can ask the respondent to call their friends right then and support them in the pitch. Although this "control" over the respondent's negotiating with their friends is feasible in person, it is no longer when remote. As a result, obtaining a full roster required the interviewer to follow up with the respondent at a later date, and the entire pace of collection was slowed. Finally, not being able to give the respondent the gift voucher right away at the end of the survey made the financial incentive less effective and the recommendation to peers less motivated.

The pandemic added complexity to our survey among a population that was already hard to access at the outset and branded by the stigma of a virus identified as originating from their native country. Collecting rosters was impacted by the weakness of the relationships established during that period and the inability to negotiate gift vouchers and other supporting strategies in person. Finally, the virtual mode is not well suited to this type of protocol that relies primarily on recommendations by peers, and thus the trust and connections established in the field.

3. Conclusion

Despite these many difficulties, 501 questionnaires were collected, resulting in a response rate of 60% (844 individuals sampled). In addition to the questionnaires, 1,689 "connections" were identified through the rosters, in other words, 3.37 individuals referred per interview (out of a maximum of 6 referrals), which represents 1,523 different individuals in the network. We never met the conditions for switching out of the Search mode. Indicator P1 (proportion of individuals not surveyed and referred only once) never went below 0.6, whereas we had set threshold A1 at 0.4. We wondered about the consequences of not running in even sampling mode on the quality and representativeness of the sample and the validity of the method in such a situation.

It is hard to predict what our results would have been if our collection had not been affected by the health context that strongly impacted the acceptability of the survey, the peer recommendations, and the pace of collection. However, some difficulties seem inherently related to the NSM method. The unpredictability of the "interviewer load", the preference for small rosters, and the aversion to duplicates are effects that work against the initial objective of collecting large rosters in order to move the survey along, and that alter the acceptability of the method by those who must experience it on a daily basis. NSM thus appears to be a method designed using theoretical or already revealed networks, but one that struggles to measure the practical cost of the actual process of revealing the network. Our article is a contribution in this regard: it provides concrete insights into the application of such a methodology. Ultimately, the method seems promising to us, but probably more suited to experimenting with a better known and smaller population. That would make it possible to better report on the benefits of the method in terms of the quality of the sample collected, compared with just a simple RDS.

References

Avin, C. and B. Krishnamachari (2008), "The Power of Choice in Random Walks: An Empirical Study." *Computer Networks*, 52(1):44 -60.

Beauchemin C. et al. (2016), « Trajectoires et Origines. Enquête sur la diversité des populations en France", Ined Editions.

Marpsat, M. and N. Razafindratsima (2010), "Les méthodes d'enquêtes auprès des populations difficiles à joindre," *Introduction to the special issue. Methodological Innovations Online.* , 5 (2) pp. 3-16.

Mouw, T., and A.M. Verdery (2012), “Network Sampling with Memory: A proposal for more efficient sampling from social networks,” *Sociological Methodology*, ; 42(1), pp. 206-256.

Schiltz, M.-A. (2005), “Making and Unmaking Groups: Numerical Information on “Difficult-Access Populations”,” *Bulletin of Sociological Methodology*, 86(1), pp.30-54.