Bulletin de Méthodologie Sociologique, N. 36, September 1992, pp. 59 - 70

ESTIMATION ON THE BASIS OF SNOWBALL SAMPLES: HOW TO WEIGHT?

by

Tom A. B. Snijders
(Department of Statistics and Measurement Theory,
University of Groningen)

Paper presented at the Workshop on Generalizability Questions for Snowball Sampling and Other Ascending Methodologies, Groningen, 20-21 February, 1992.

Abstract. What are the possibilities of snowball sampling, if one desires valid statistical inference without making probabilistic assumptions on the network structure? In a critical review of the possibilities of snowball sampling for a population of vertices connected by a network of arcs, it is argued that the snowball method is much more suitable for the estimation of parameters of the network structure (or parameters of the population of arcs) than to estimate parameters of the population of vertices. Further work needs to be done to relax the assumption of randomness of the initial sample of the snowball. **Snowball Sampling, Weighting, Parameter Estimations, Social Networks.**

Résumé. Quelle possibilité donne un échantillon boule de neige pour parvenir à des inférences statistiques valides sans faire d'hypothèses probabilistes sur la structure du réseau? Dans une revue critique des apport des échantillon boule de neige pour un ensemble de points connectés par un réseau d'arcs, l'auteur montre que cette méthodes s'applique mieux pour estimer les paramètres de la structure sociale (ou paramètres de la population d'arcs) que d'estimer les paramètres de la population des points. Plus de travail est nécessaire pour amoindrir l'hypothèse de distribution aléatoire de l'échantillon initiale. **Echantillonage en boule de neige, Pondération, Estimations des Paramètres, Réseaux sociaux.**

1. INFERENCE AND SNOWBALL SAMPLES

Snowball sampling is sometimes used as a rather informal way to reach a population, and sometimes as a more or less formal sampling method with the purpose either to make inference with regard to the population of individuals or to make inference with regard to the network structure in that population. The purpose that Coleman (1958) had in mind when he introduced the idea of snowball sampling was the second one. As a formal sampling method, however, snowball sampling is known to have some serious problems because of the inherent bias. Berg (1988) states: "As a rule, a snowball sample will be strongly biased towards inclusion of those who have many interrelationships with, or are coupled to, a large number of other individuals. In the absence of knowledge of individual inclusion probabilities in different waves of the snowball sample, unbiased estimation is not possible." This paper presents some ideas about possibilities and problems related to the formal use of snowball sample designs. The leading question is as follows: to what extent is it possible to treat snowball sampling, and generalisations where nominees are sampled with known probabilities, as a probability sampling method; and to what extent can data obtained from a snowball sample be used to derive good estimators for various population parameters?

2. THE MODEL FOR SNOWBALL SAMPLING

The model assumed for snowball sampling in this paper has the following features.

* It is a model for a given, fixed but a priori unknown, network. There is one single directed relation being considered; for the sake of simplicity the relation shall be described as follows: when i is related to j, then i says that j is his friend.

* The individuals in the network are assumed to have a number of friends that varies from individual to individual; if an individual is interviewed (i.e., gets the role of respondent) then his or her number of friends can be observed and a random sample of the friends (possibly all the friends) is drawn for the continuation of the snowball.

The first feature means that the population of individuals as well as the relation among them is assumed to be unequivocally defined, and that (in terms of sampling theory) the inference may be called design-based rather than model-based. The second feature also is important. The number of nominees per respondent is not fixed (this differs from Goodman (1961), who assumes a relation where every individual has the same number of friends). This is preferable to the assumption of a fixed number of friends because (a) some respondents will have less than the given number of friends, (b) if there are so many friends that a choice has to be made among them then this choice should not be left up to the respondent but should rather be made at random; otherwise the researcher does not know what is being observed. The second feature precludes the existence of chaining processes (as treated by Erickson, 1978), except for pure chance chaining processes.

The (finite) population of individuals is represented as a directed labeled graph: individuals are represented by vertices, and the directed relation between the individuals by arcs. The number of vertices is denoted N; the arc indicators are denoted X :

$$X_{ij} = \left\{\begin{array}{l} 1 \\ 0 \end{array}\right. \quad \text{if} \quad \begin{array}{l} \text{there is an arc} \\ \text{there is no arc} \end{array} \quad \text{from vertex } i \text{ to vertex } j \ .$$

$(1 \leq i, j \leq N)$. It is assumed that there are no loops: $X_{ii} = 0$ for all i. For every vertex i, the out-neigbourhood of i is denoted

$$U_i = \{j \mid X_{ij} = 1\} .$$

The out-neighbourhood U_i may be called the personal network of vertex (person) i. The out-degree of i, which is the number of elements of U_i, is X_{i+}; the in-degree of i is X_{+i} .

Further, there is a vertex variable, Y, with values

$$Y_i \ , \text{ for } i = 1, \dots, N.$$

It can also be interesting to consider arc variables:

$$Z_{ij} \ , \text{ for } 1 \leq i, j \leq N,$$

for which it makes sense to assume that they are equal to 0 whenever there is no arc: $X_{ij} = 0$ implies $Z_{ij} = 0$.

The generalized snowball sampling procedure operates as follows. An initial simple random sample of size n is drawn from the population of vertices. This sample is denoted $S^{(0)}$. For every vertex i in $S^{(0)}$, the out-neigbourhood U_i of i is observed (i.e., the identities, or labels, of the vertices in U_i are observed) as well as the value of Y_i. To define the snowballing procedure, suppose that the (s-1)'th wave (s-1 = 0, 1, ...), denoted $S^{(s-1)}$, of the snowball sample has been drawn. Then the s'th wave is drawn as follows. For every $i \in S^{(s-1)}$, a simple random sample of size n_{si} $(n_{si} \leq X_{i+})$ is drawn from the out-neigbourhood U_i of i; this sample is denoted S_{si}. The union of all these samples, minus those vertices who were already contained in the earlier waves, is denoted $S^{(s)}$:

$$S^{(s)} = \bigcup_{i \in S^{(s-1)}} S_{si} \setminus (S^{(0)} \cup S^{(1)} \cup \dots \cup S^{(s-1)}) .$$

$S^{(s)}$ is the s'th wave of the snowball sample. For all i $S^{(s)}$, in the s'th wave the identities (or labels) of all vertices in the out-neighbourhoods of i are observed, together with the variables Y_i. The process stops after a given number of waves, or at the first moment where no more new vertices are observed, i.e., where $S^{(s)}$ is empty. The whole snowball sample is the union of the several waves:

$$S = \bigcup_s S^{(s)} .$$

If $X_{i+} = n_{si} = k$, then Goodman's (1961) snowball sample method is obtained; if $n_{si} = X_{i+}$, we obtain Frank's (1977) snowball sample method; if $n_{si} = 1$, then we obtain Klovdahl's (1989) random walk design. Most of this paper will be concerned with Frank's method, where no actual sampling (i.e., with sampling fractions < 1) takes place in the later waves.

This snowball method is what is called in sampling theory a *probability sampling method*: for every sample, the probability of obtaining this sample can be calculated. That is nice, but not sufficient for all purposes. The purposes we might have can be classified as follows:

(i) inference about the population of vertices, say, about the collection of values $\{Y_1, \ldots, Y_N\}$: e.g., the number of vertices, or the population average $\bar{Y}$;

(ii) inference about the population of arcs: e.g., the arc density $X_{++} /\{N(N-1)\}$, or the average of the arc variable Z_{ij} over all arcs: Z_{++} /X_{++} ;

(iii) inference about the population of personal networks $U_1, \ldots, U_N$; e.g., the number of isolates in the population, or the degree variance; (formally, this is an important special case of (i), because the personal networks U_i may be considered to be attributes of the vertices);

(iv) inference about the total network, i.e., about parameters that can be expressed as functions of the entire adjacency matrix (X_{ij} ; $1 \leq i,j \leq N$); e.g., the total number of pairs of vertices at distance d from each other, or the number of connected components (for the last parameter the snowball method does not seem a very good design).

3. INFERENCE TO THE POPULATION OF INDIVIDUALS

Let us first discuss purpose (i). The estimation of the size of the population, N, from a snowball sample is discussed in Frank and Snijders (1992). In this section, it is assumed that N is known. To derive estimators for parameters such as population averages $\bar{Y}$, the most usual method is the Horvitz-Thompson method, which needs, however, the *inclusion probabilities*, i.e., the probabilities that the observed vertices are sampled. If the inclusion probability for vertex i is π_1, i.e.,

$$\pi_1 = P\{i \in S\},$$

then the Horvitz-Thompson estimator for Y is

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{i \in S} Y_i/\pi_i \cdot$$

Other estimators are also possible; e.g., the estimator proposed in Frank (1977, Theorem 2). It also seems sensible to use the *wavewise* inclusion probabilities:

$$\pi_{is} = P\{i \in S^{(s)}\} ;$$

it is conceivable that conditional wavewise inclusion probabilities are used,

$$\pi_{is}(S^{(0)}, \ldots, S^{(s-1)}) = P\{i \in S^{(s)} \mid S^{(0)}, \ldots, S^{(s-1)}\} .$$

Using these inclusion probabilities, unbiased estimators for $\hat{\bar{Y}}$ are, respectively,

$$\frac{1}{N} \sum_s w_s \sum_{i \in S^{(s)}} Y_i/\pi_{is}$$

and

$$\frac{1}{N} \sum_s w_s \sum_{i \in S^{(s)}} Y_i/\pi_{is}(S^{(0)}, \ldots, S^{(s-1)}) ,$$

where the w_s are weights for the waves to be chosen in a sensible way; they should sum to 1. These estimators can only be used if the wavewise inclusion probabilities, or conditional wavewise inclusion probabilities, are positive for all vertices i with $Y_i \neq 0$.

All this is very nice in theory, but *how can these inclusion probabilities be observed, computed, estimated*? Compute, yes, in principle, but observe or estimate ... For the zero-th wave, everything is fine due to the assumption of a simple random initial sample:

$$\pi_{i0} = n/N.$$

For the first wave, we have

$$\begin{aligned}\pi_{i1} &= P\{i \notin S^{(0)},\ i \in S_{1j} \text{ for some } j \in S^{(0)}\} \\ &= P\{i \notin S^{(0)}\}\, P\{i \in S_{1j} \text{ for some } j \in S^{(0)} \mid i \notin S^{(0)}\}\end{aligned}$$

$$= (1 - \frac{n}{N}) \, (1 - P(i \notin S_{1j} \text{ for all } j \in S^{(0)} \mid i \notin S^{(0)})) \, .$$

In the case described in Section 2, where S_{1j} is a simple random sample of size n_{1j} from U_j , the last conditional probability is not very attractive to calculate. In the less general case where $S_{1j} = U_j$ (all j's friends are interviewed), the result is

$$\pi_{i1} = (1 - \frac{n}{N}) \, (1 - \frac{\binom{N - x_{+i} - 1}{n}}{\binom{N-1}{n}}) \, .$$

This is a function of the in-degree x_{+i} of i; it is intuitively clear that this should be the case. When nx_{+i} / N is small, the probability that two (or more) individuals in the initial sample both have i as a friend is negligible, and ,π_{i1} can be approximated (in the general case) as follows; we write for the sampling fraction within j's personal network $n_{1j} / x_{j+} = f_j$.

$$\pi_{i1} = (1 - \frac{n}{N}) \, P(i \in S_{1j} \text{ for some } j \in S^{(0)} \mid i \notin S^{(0)})$$

$$\approx (1 - \frac{n}{N}) \sum_{j \in U_i} P(j \in S^{(0)}, \, i \in S_{1j})$$

$$= (1 - \frac{n}{N}) \sum_{j \in U_i} \frac{n}{N} f_j \, .$$

This is the *computation* for π_{i1}. We can conclude that at the very least, we need to know the *in-degree* of the respondents. This amounts to posing the question, "How many individuals would say you are their friend?" Only in very special cases may we expect to get valid and reliable answers to such a question. If the generalized snowball procedure is used, i.e., $f_j < 1$, then we need the f_j values for all individuals j who would mention respondent i as their friend ... This seems even more difficult.

Let us now consider two more specific situations, where difficulties are a bit smaller.

(1) *Multiplicity sampling*. In multiplicity sampling (Sirken, 1970, and later publications), individuals are asked about themselves as well as about household members or about others having a specific (family) relationship with the respondent. This can be considered as a one-wave snowball sample where the relation is defined as "being part of the same household", etc. In such applications of multiplicity sampling, the network structure is obtained from a partitioning of the population (e.g., into households). The crucial circumstance here is that it is well possible for the respondent to give the in-degrees of his/herself and of the nominees.

(2) An *undirected* graph, i.e., a relation that is by definition mutual; combined with a snowball design with $f_i = 1$, i.e., where all friends belong to the next wave of the snowball (except those who were encountered already at earlier waves). For this case, it is not too difficult to compute π_{is}. Define

d_{ij} = graph-theoretic distance between vertices i and j,
i.e., the length of the shortest path from i to j.

(E.g., $d_{ii} = 0$ for all i; if i and j are friends, then $d_{ij} = 1$; if i and j are not friends but do have a mutual friend then $d_{ij} = 2$; if there is no path from i to j, then $d_{ij} = \infty$.) Then

$$i \in S^{(s)} \text{ if and only if } \min\{d_{ij} \mid j \in S^{(0)}\} = s \, .$$

Define

$$D_{is} = \#\{j \mid d_{ij} \leq s\} \, ,$$

the number of vertices (including i) at distance at most s from i; note that $D_{i0} = 1$. Then the wavewise inclusion probabilities can be computed from

$$\pi_{i0} + \pi_{i1} + \ldots + \pi_{is} = 1 - \frac{\binom{N - D_{is}}{n}}{\binom{N}{n}} \, .$$

The value of D_{is} can be observed by extending the snowball sample s-1 waves beyond the wave where vertex i was observed. This implies that we could take a 2s-1 - wave snowball sample, use the first s waves to observe the numerical values of Y_i and (of course) the relations, and the last s-1 only to observe the relations so that D_{it} will be known for t = 1, ..., s.

What does the literature say? Frank (1977, 1979) treats one-wave snowball samples with $f_j = 1$, and assumes that in-degrees are observable. (He also treats in the 1977 paper some other sampling designs: network sampling a la Granovetter (1976), sampling of random edges, and also estimation problems in the spirit of purpose (ii) mentioned above.) Frank (1979) goes on (in his Section V) to discuss a specific probabilistic model for contact processes which enables him to get some further results, but this is model-based rather than design-based inference, so it is not relevant for the questions of this paper. Schmeidler (1990, p. 21) proposes a weighting method that is not fully clear to me, and for which a rationale is not given. I must confess that my impression is that his proposal does not make much sense.

My conclusions are the following with respect to the use of snowball samples for purpose (i), if estimators based on the idea of the Horvitz-Thompson estimator are used.

(1) In-degrees for non-symmetric relations are often unobservable (an exception is multiplicity sampling). Therefore, in order to get statistically valid estimates from snowball designs, it is advisable to use *symmetric relations only*. In practice, this means that the inclusion criterion has to be defined in such a way that the relation may be considered to be mutual; this can be checked by noting whether nominees, when interviewed, mention among their "friends" the respondent who mentioned them. There remain serious problems of data reliability, however; the literature review by Sudman (1985) is relevant here.

(2) Stick to a low number of waves. Valid statistical analysis of a one-wave snowball sample is difficult enough. In a 2s-1 - wave snowball sample for an undirected (mutual) relation, use the last s-1 waves just to estimate the D_{is} needed to compute the inclusion probabilities.

4. INFERENCE TO THE TOTAL NETWORK

Purposes (ii) and (iv) are concerned with the network structure. E.g., what Frank (1977, p. 247) calls *graph totals* are the kind of parameters that are estimated under purpose (ii). It is not without reason that Goodman (1961) focuses his paper on the estimation of the frequencies of various kinds of chains and cycles. Snowball sampling is a method where such frequencies are estimated in a very natural way. For example, consider the transitivity parameter

$$T = \frac{N_2}{C_3} ,$$

where N_2 is the number of chains of length 2,

$$N_2 = \#\{i,\ j,\ k \mid X_{ij} = X_{jk} = 1,\ i \neq k\},$$

and C_3 is the number of triangles,

$$C_3 = \{i,\ j,\ k \mid X_{ij} = X_{jk} = X_{ki} = 1\} .$$

N_2 can be estimated from a single-wave snowball sample, e.g., as follows. Let N_2 ($S^{(0)}$) be the number of chains of length 2, of which at least one of the end points is an element of $S^{(0)}$:

$$N_2(S^{(0)}) = \#\{(i,\ j,\ k) \mid X_{ij} = X_{jk} = 1,\ i \neq k,\ i \in S^{(0)} \text{ or } k \in S^{(0)}\} .$$

This is a statistic in the single-wave snowball design. For every pair (i, k) holds that

$$P\{i \in S^{(0)} \text{ or } k \in S^{(0)}\} = 1 - \frac{\binom{N-2}{n}}{\binom{N}{n}} = 1 - \frac{(N-n)(N-n-1)}{N(N-1)}$$

$$= \frac{n(2N-n-1)}{N(N-1)} .$$

This implies that

$$E\{N_2(S^{(0)})\} = N_2 \frac{n(2N-n-1)}{N(N-1)} ,$$

so that an unbiased estimator for N_2 is

$$\hat{N}_2 = \frac{N(N-1)}{n(2N-n-1)} N_2(S^{(0)}) .$$

Similarly, let C_3 ($S^{(0)}$) be the number of triangles of which at least one the vertices is in $S^{(0)}$:

$$C_3(S^{(0)}) = \{i,\ j,\ k \mid X_{ij} = X_{jk} = X_{ki} = 1,$$
$$\text{at least one of } i,\ j, \text{ or } k \text{ is in } S^{(0)}\} .$$

Then C_3 ($S^{(0)}$) is a statistic in the 2-wave snowball design. For all triples (j, k) it holds that

$$P\{i,\ j, \text{ or } k \in S^{(0)}\} = 1 - \frac{\binom{N-3}{n}}{\binom{N}{n}} = 1 - \frac{(N-n)(N-n-1)(N-n-2)}{N(N-1)(N-2)}$$

$$= \frac{n\{n^2 - 3n(N-1) + (3N^2-6N+2)\}}{N(N-1)(N-2)} ,$$

so that an unbiased estimator for C_3 is

$$\hat{C}_3 = \frac{N(N-1)(N-2)}{n\{n^2 - 3n(N-1) + (3N^2-6N+2))\}} C_3(S^{(0)}) .$$

It can be concluded that, from the 2-wave snowball design, we can construct $\hat{N}_2 / \hat{C}_3$ as a reasonable estimator for the transitivity parameter N_2 / C_3 . This is not meant to imply that this estimator is the most efficient one possible; but it does demonstrate that for network parameters that are functions of chain and cycle frequencies, the snowball method yields rather natural ways to construct reasonable estimators.

From this discussion, which can be underpinned further by Goodman (1961) and Frank (1977), it can be concluded that for the estimation of certain network parameters, such as are meant under purposes (ii) and (iv), snowball sampling may be much more adequate than for the estimation of the vertex-population parameters of purpose (i).

5. NON-RANDOMNESS OF THE INITIAL SAMPLE?

The weak underbelly of the snowball method is the assumption of a random initial sample. Snowball samples mostly are taken from populations for which a sampling frame is not available. There are even cases where the number of vertices is to be estimated from the snowball sample - a problem definition which is contradictory to the availability of a sampling frame. Without a sampling frame, how can we draw a simple random initial sample?

We can not. The best we can do is to draw the respondents, as much as possible, from independent sources. E.g., if a snowball sample of drug users is to be taken and "bars" is one of the "social milieux" where initial respondents can be sought, not more than one initial respondent is to be sought in one bar or in one small-scale "social environment" of any kind. This physical approximation to independence will hopefully lead to something approximating random sampling in the sense that for individuals to be together in the initial sample is uncorrelated with the direct and indirect (i.e., larger distance) ties between them. However, practically all "ethnographic" methods to get initial respondents will lead to bias in the sense that the more widely known individuals (i.e., those with higher in-degrees) are over-represented, even in the initial sample. It would be interesting to try some simulation examples to see how badly this affects estimation results, and to try to find correction methods for this bias.

A related point is that there are often several separate social sources of initial respondents; e.g., in a study of drug users: bars, police contacts, socio-medical institutions, and educational institutions. The initial sample is then stratified, and about the sizes of the various subpopulations corresponding to these social sources, there often is not enough information to determine whether the initial sampling fractions in the subpopulations are anywhere near each other. In such cases, the assumption that the initial sample of the snowball is a stratified random sample may be much closer to reality than the assumption that it is a simple random sample. It could be worthwile to elaborate estimation methods that are valid under the assumption of a stratified random initial sample; or, more generally, under the assumption of known but varying probabilities of inclusion in the initial sample.

As a last comment, a warning is in order with respect to the relation between initial sample size and size of the population (i.e., the number of vertices) from which the snowball sample is drawn. With regard to the information it gives about population size and network structure, the snowball method lives on the chains that return to vertices observed earlier in the snowball. The observed number of returning chains should be sufficiently large so that the relative error in this number is not too high; say, at least 50 returning chains should be observed. For a single-wave snowball design in an undirected graph, the expected number of chains i - j - k where i and k are in the initial sample, is pn(n - 1)/2, where p is the fraction of pairs of vertices which are at distance 2 from each other. In a large undirected graph where degrees are about 12, the order of magnitude of out-neigbourhoods at distance 2 is 100 (a bit less than 12^2), so p 100/N. This implies $pn(n-1)/2 \approx 50n(n-1)/N$. If this is to be at least 50, then n^2 N. This seems to imply the rule of thumb that, for relations where we can ask respondents about something like 12 nominees at the most, the initial sample size n of a one-wave snowball sample should not be much smaller than the square root of the population size, in order to make precise statistical inferences from snowball samples.

REFERENCES

Berg, S. (1988), "Snowball sampling". In: S. Kotz and N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences*, vol. 8, 528-532.

Coleman, J.S. (1958), "Relational analysis: The study of social organizations with survey methods". *Human Organization*, 17, 28-36.

Erickson, B.H. (1978), "Some problems of inference from chain data". In: K.F. Schuessler (ed.), *Sociological Methodology-1979*, 276-302. San Francisco: Jossey-Bass.

Frank, O. (1977), "Survey sampling in graphs". *Journal of Statistical Planning and Inference*, 1, 235-264.

Frank, O. (1979), "Estimation of population totals by use of snowball samples". In: P. Holland and S. Leinhardt (eds.), *Perspectives on Social Network Research*, 319-347. New York: Academic Press.

Frank, O., and Snijders, T.A.B. (1992), "Estimating the size of a hidden population by snowball samples". In preparation.

Goodman, L.A. (1961), "Snowball sampling". *Annals of Mathematical Statistics*, 32, 148-170.

Granovetter, M. (1976), "Network sampling: some first steps". *American Journal of Sociology*, 81, 1287-1303.

Klovdahl, A.S. (1989), "Urban social networks: Some methodological problems and possibilities". In: M. Kochen (ed.), *The Small World*, Chapter 10. Norwood, N.J.: Ablex.

Schmeidler, J. (1990), "Weighting procedures for ethnographic random samples". *Bulletin de Methodologie Sociologique*, 29, 15-24.

Sirken, M.G. (1970), "Household surveys with multiplicity". *Journal of the American Statistical Association*, 65, 257-266.

Sudman, S. (1985). "Experiments in the measurement of the size of social networks". *Social Networks*, 7, 127-151.

==