Metodološki zvezki, Vol. 1, No. 2, 2004, 395-405

The Network Scale-Up Method: A Simulation Study in Case of Overlapping Sub-Populations

Silvia Snidero, Roberto Corradetti[1] and Dario Gregori[2]

Abstract

The network scale-up method is a social network estimator for the size of hidden or hard-to-count subpopulations. These estimators are based on a simple model which have however strong assumptions. The basic idea is that the proportion of the mean number of people known by respondent in a subpopulation E of T of size e is the same of the proportion that the subpopulation E forms in general population T of size t: $\frac{m}{c} = \frac{e}{t}$, where c is the number of persons known by each respondent and m is the mean number of persons known by each respondent in the subpopulation E. The persons known by every subject is called the "social network", and its size is c, estimated by several estimators proposed in the recent literature. In this paper we present a Monte Carlo simulation study aimed at understanding the behavior of the scale-up method type estimators under several conditions. The first goal was to understand what would be the ideal number of subpopulations of known size to be used in planning the research. The second goal was to analyze what happens when we use overlapped subpopulations. Our results showed that with the scale-up estimator we always obtain biased estimates for any number of subpopulations employed in estimates. With the Killworth's ML estimator, the improvement of scale-up method, we have substantially unbiased estimates under any condition. Also in case of overlapping, and increasing the degree of it among subpopulations, bias raises with scale-up method, instead it remains close to zero with ML estimator.

[1] Dept. of Mathematics and Statistics applied to Human Sciences, University of Torino, Italy
[2] Department of Public Health and Microbiology, University of Torino, Italy
This work has been done with a 2003 grant of the Franca and Diego the Castro Foundation, Torino

1 Introduction

During the last few years, Bernard (Bernard, 1991; Bernard, 1989; Killworth, 1998a; Killworth, 1998b; Bernard, 2001) proposed some new approaches aimed at estimating the size of hidden or hard-to-count populations. These methods are largely based on the concept of social network: a loose definition of it is the set of people that a person knows. The width of the network thus dependents on the definition of "knowing" someone. Beside the difficulties on defining exactly this concept, the main advantage of the technique consists in asking people about problems indirectly, i.e.: problems that in principle are not regarding him/her directly, asking instead on how many people he/she knows with the specific characteristic. The resulting samples are smaller as compared to those obtained using common estimators for events with small prevalence.

The general set up of these methods assumes a total population T of size t and a subpopulation E of T of size e. Thus, the basic assumption underlying the scale-up method's class of estimators is

$$\frac{m}{c} = \frac{e}{t} \tag{1.1}$$

where m is the mean number of persons known in E and c is a constant representing the social network size of any member of T. Hence, the proportion of subjects in E known to any member of T is the same as the proportion of members of E belonging to general population T, i.e.: e/t.

In order to estimate the unknown size e we need the number c of people that a person knows (the social network). In the past, several estimators of the social network size were proposed (Freeman, 1989; McCarty, 2000; Bernard, 1990; Killworth, 1990). Some of these estimators propose to estimate c using the number of persons known by each respondent in several subpopulations of the known size (Killworth, 1998a). Thus, the basic idea is to ask respondents how many people they know in the target subpopulation and how many people they know in a certain number of subpopulations of the known size. For example, one could ask: "How many people do you know who are seropositive?" (the unknown size subpopulation) and "How many people do you know that owing a VISA card?" (the known size subgroup).

Killworth et al. (Killworth, 1998a; Killworth, 1998b) presented two estimators for the dimension of e: a revision of the original scale-up estimator and another one based on maximum likelihood.

Several strong assumptions are underlying the scale-up method approach to the estimate of unknown population sizes, which can lead to several problems in estimates (McCarty, 2000): (i) each subject in T has the same probability to know a person in subgroup E, (ii) everyone in T knows all about his/her acquaintance and (iii) the negligible difficulty that is to recall in short time all people known in a certain subpopulation.

The violation of these assumptions can conduct to some problems, spotted in the literature and known as the barrier, transmission and estimation effects. Indeed, some social and geographical characteristics can create a barrier in knowing some specific groups of persons. The second effect faces us when information about a person is not transmitted with the same probability to his/her acquaintance. The

difficulty to recall people belonging to a certain subpopulation can lead to estimation effects.

Several issues have been raised in the past literature on the statistical behavior of these estimators. Most of the critics are related to the loose definition of network, which is rarely consistent among applications (Freeman, 1989; Bernard, 1989; McCarty, 2000), possibly leading to very different estimates for the same phenomenon.

On the other hand the proposed estimators have been presented and justified using mostly heuristic arguments, missing a strict definition of the statistical properties, which affect the capability of implementing a statistical plan in designing a field research. For instance, Killworth et al. (Killworth, 1998a; Killworth, 1998b) suggested, in a very limited setting, that using at least 20 subpopulations of a known size to estimate social network dimension leads with the scale-up method to unbiased estimates of the target subgroup size.

Moreover, the problem of overlapping subpopulations has not been addressed explicitly. For overlapping subpopulations we intend that a person can belong to one or more subgroups, e.g. a subject can stay in a subpopulation definite by "persons calling Michael" and simultaneously belong to "people possessing a swimming pool". This problem face us potentially in all studies using subpopulations. The exception is when we use subpopulation based on names, where overlapping is not possible.

In the present paper we investigate, using a Monte Carlo simulation experiment, the statistical behavior of the scale-up method and the ML estimator under several conditions with populations with different degrees of subpopulation overlapping.

2 The social network based estimators

Both the scale-up and the ML estimators are extensions of the basic model described in Section 1.

The scale-up method Assuming that E_0 is a subpopulation of T of unknown size e_0, then the estimator (Killworth, 1998a) of e_0 is:

$$ {}_1\hat{e}_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{m_{i0}\cdot t}{c_i} \tag{2.1} $$

where m_{i0} is the number of persons known by respondent i in the unknown subpopulation E_0 and c_i is the network size of i-th respondent.

If we have a sufficient large random sample of n members of T, the distribution of the number known by each respondent in the target subpopulation m_{i0} is virtually normal (Johnsen, 1995). In this case the standard error of ${}_1\hat{e}_0$ is:

$$ s.e.({}_1\hat{e}_0) = \sqrt{\frac{t}{n}}\sum_{i=1}^{n}\frac{s.e.(m_{i0})}{\sqrt{c_i}}. \tag{2.2} $$

The maximum likelihood estimator The ML estimator (Killworth, 1998b) is found by maximizing the probability

$$\text{Prob}(c_i, m_{i0}, e_0) = \prod_i \left(\begin{array}{c} m_{i0} \\ c_i \end{array} \right) p_0^{m_{i0}} (1-p_0)^{c_i - m_{i0}} \quad \text{where } p_0 = \frac{e_0}{t}, \tag{2.3}$$

by varying e_0, obtaining

$$_2\hat{e}_0 = t \frac{\sum_i m_{i0}}{\sum_i c_i}. \tag{2.4}$$

It is demonstrated that the maximum likelihood estimator is unbiased, i.e.: with $E(_2\hat{e}_0) = e_0$ and standard error given by

$$S.E.(_2\hat{e}_0) = \sqrt{\frac{t \cdot {}_2\hat{e}_0}{\sum_i c_i}}. \tag{2.5}$$

The main difference between the two estimators is that the scale-up estimator requires the sum of the ratios for each respondent i of m_{i0} (the number of persons known by respondent i in the target subpopulation) and c_i (the network size of i-th respondent). The maximum likelihood estimator computes the ratio of the sum of m_{i0} over all respondents and the sum of c_i social networks size over all respondents.

The estimation of social network size The only unknown variable in the above mentioned estimators is the social network size of each respondent, and therefore we have to substitute c_i with a good estimate of it.

In order to estimate social network sizes there are two estimators that use subpopulations of known sizes (Killworth, 1998a). The first is the *subgroup estimator*, which can be obtained by maximizing:

$$\text{Prob}(i \text{ knows } m_{ij}, j = 1, 2, \cdots, L) = P_{c_i} = \prod_{j=1}^{L} \left(\begin{array}{c} m_{ij} \\ c_i \end{array} \right) p_j^{m_{ij}} q_j^{c_i - m_{ij}} \tag{2.6}$$

where m_{ij} is the number of persons known by respondent i in the j-th subpopulation of known size ($j = 1, \ldots, L$) and $p_j = \frac{e_j}{t}$ is the fraction of subpopulation E_j in T.

Another network size estimator is the *proportional estimator*

$$\hat{c}_i = t \cdot \frac{\sum_{j=1}^{L} m_{ij}}{\sum_{j=1}^{L} e_j} \tag{2.7}$$

with standard error

$$s.e.(\hat{c}_i) = \sqrt{\frac{t\hat{c}_i}{\sum_{j=1}^{L} e_j}}. \tag{2.8}$$

Killworth et al. (Killworth, 1998a) proved that for small values of m_{ij}/c and p_j, the estimate of c_i is almost the same of that obtained by the *subgroup estimate*.

Every estimate of c_i has an error that is transmitted to the unknown subpopulation size estimate. It is proven (Killworth, 1998b) that the effect of this error on estimate is negligible if $\sum_{ij} m_{ij}$ is sufficiently large.

3 The Monte Carlo experiment

The simulations are organized in two parts: the first was intended to determine the optimal number of subpopulations to be used in the estimate and to understand the behavior of the estimators in case of small samples. In the second part we considered the case of overlapped subgroups. In both parts of simulations we used both the scale-up method and the maximum likelihood estimator.

For the simulations we used a total population T composed by a certain number of subpopulations of a known size, the social network size of each element of population being c and the subpopulation to estimate having a sample size of n.

We simulated the network size of each member of the general population T with random generation from an exponential distribution, with mean based on published results (Killworth, 1998a; Killworth, 1998b). On this basis we generated two different network sizes, one with mean 105 (Killworth, 1998a) and another with mean 160:

$$c_1 \sim \exp\left(\frac{1}{105}\right) \quad \text{and} \quad c_2 \sim \exp\left(\frac{1}{160}\right).$$

The social network size were estimated using the *proportional estimator*. Then, we constructed four target subpopulations, with different relative sizes ($p_0 = \frac{e_0}{t}$): 0.001, 0.002, 0.01 and 0.1.

Thousand simulations were carried out in the following way:

1. we extracted the sample from the total population;
2. for each element in the sample we estimated its network size ($\hat{c}_i$);
3. then we estimated the target subpopulation size for each member ($\hat{e}_{i0}$) using $\hat{c}_i$;
4. eventually, we averaged the results obtained for each element in the sample and calculated the standard error.

3.1 Non-overlapped subpopulations

We implemented the simulations with two different kind of subpopulations: homogeneous and non-homogeneous.

Homogeneous subpopulations

We constructed three general populations T of size 50.000, respectively owning 5, 25 and 100 subpopulations. In each of these subpopulations the subgroups were homogeneous, i.e. they had the same size ($e_h = e_j$, for each subpopulation h and j). In all cases, the total size of the subpopulations ($\sum_j e_j$) added up to 20.000.

We extracted a sample of 2.000 members (n) from the general population and we estimated the sizes of the four subpopulations. Both social network sizes were simulated in the case of the network scale-up method; with the ML estimator we used only the social network size distributed as $\exp\left(\frac{1}{160}\right)$.

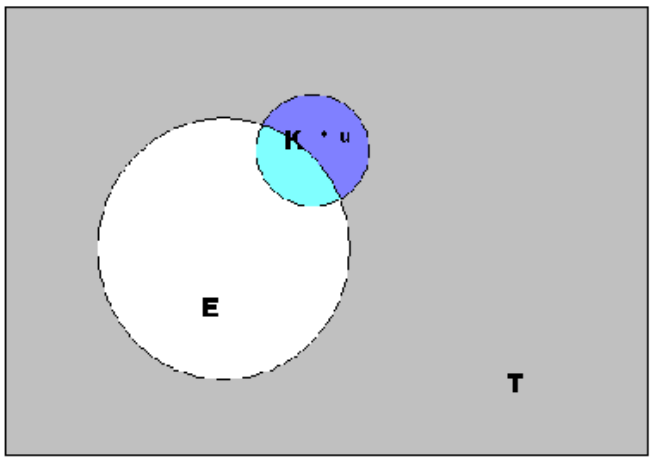

Figure 1: Representation of the idea of the scale-up method estimators.

Non-homogeneous subpopulations

We constructed two populations of dimension $t = 50.000$ with non homogeneous subpopulations, one with subpopulations total size of 20.000 and another one with 30.200.

We considered only a subpopulation with relative size 0.001, extracting a sample with size n equal to 2.000. For this simulation, the only social network size considered was the one distributed as $\exp\left(\frac{1}{160}\right)$.

In this first part, 40 simulations were implemented: 26 with the scale-up method (4 different unknown subpopulations to estimate $\times$ 2 social network sizes $\times$ 3 different groups of homogeneous subpopulations with known size, and 2 with non-homogeneous subgroups) and 14 with the ML estimator (4 different subpopulations to estimate $\times$ 3 different homogeneous groups of known size $\times$ 1 social network size and 2 with non-homogeneous subgroups).

3.2 Overlapped subpopulations

For this part of Monte Carlo simulations, we constructed five general populations with different degrees of overlapping among subpopulations.

These populations had a dimension t of 50.000 subjects, organized in 25 subgroups which sizes are in the range 100-2.500: first subpopulation was 100, the second 200, and so on until the 25th with a size of 2.500.

In the first generated population, the subgroups are not overlapped, i.e. each population's element can belong at most to one subpopulation, e.g. names subgroups (see Figure 1).

The second population is constructed as follows: 25% of elements of the smallest subpopulation (with size 100) belongs to all biggest subpopulations, 25% of elements of second smallest subgroup belongs to all biggest subpopulations, and so on. Hence, 25% of all subpopulations has the character of the biggest subpopulation, that with

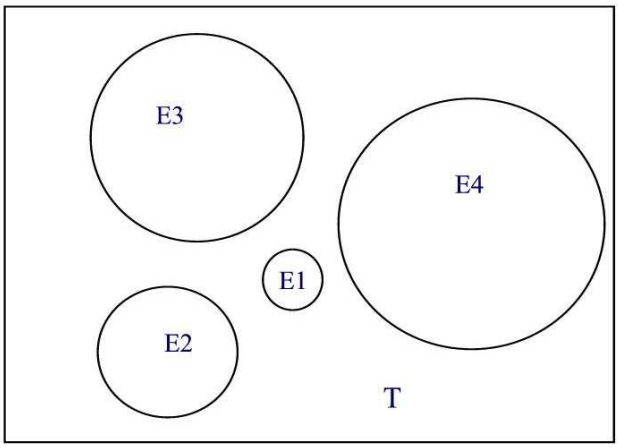

Figure 2: Non overlapping subpopulations

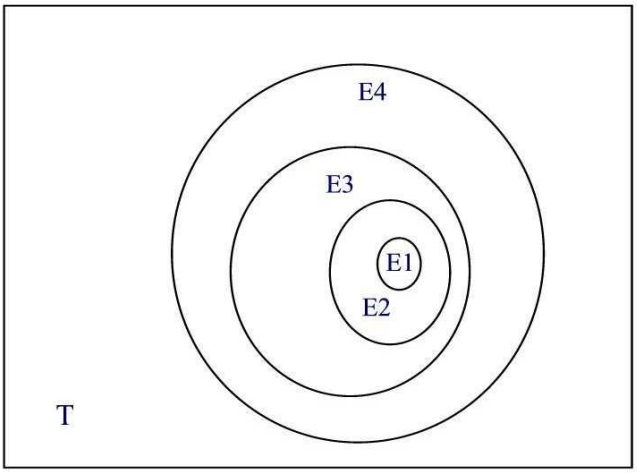

Figure 3: Overlapping subpopulations.

2.500 members.

Other populations were constructed in the same manner but with overlapping degrees of 50%, 75% and 100%. The latter represents the extreme case of overlapping: all elements of smaller subpopulations are contained in the biggest subgroup (see Figures 2 and 3).

Also in this case we used both estimators, but only with a reference to the subpopulation with a relative size 0.001. Moreover, social network dimensions were distributed as exp(1/160) and we extracted a sample of 2.000 elements (n).

Table 1: Relative bias in non-overlapped homogeneous subpopulations. Values are expressed × 100.

actual p_0	*network size*	*Scale-up*			*ML*		
		5	*25*	*100*	*5*	*25*	*100*
0,10	c_1	9,25	8,62	7,84			
	c_2	6,06	6,43	5,89	-0,04	-0,02	-0,35
0,20	c_1	11,40	10,01	9,96			
	c_2	7,51	7,39	6,41	-0,02	0,08	0,08
1,00	c_1	10,42	9,63	9,71			
	c_2	7,38	6,88	6,77	-0,05	-0,04	-0,03
10,00	c_1	10,46	9,55	9,80			
	c_2	7,36	6,79	7,05	-0,02	0,00	0,00

Table 2: Standard error in non-overlapped homogeneous subpopulations. Values are expressed × 100.

actual p_0	*network size*	*Scale-up*			*ML*		
		5	*25*	*100*	*5*	*25*	*100*
0,10	c_1	0,05	0,05	0,05			
	c_2	0,05	0,05	0,05	0,01	0,01	0,01
0,20	c_1	0,07	0,07	0,07			
	c_2	0,07	0,07	0,07	0,01	0,01	0,01
1,00	c_1	0,17	0,16	0,16			
	c_2	0,16	0,16	0,16	0,02	0,02	0,02
10,00	c_1	0,50	0,49	0,43			
	c_2	0,49	0,49	0,49	0,06	0,06	0,06

4 Results

4.1 Non-overlapped simulations

The relative bias (Table 1) varies from 11.40% to 6.41% with the scale-up estimator and from -0.05% to 0.08% with the ML estimator, which is known to be unbiased, slightly smaller in case of network sizes with the mean 160 (2-3%) as compared to those with a mean of 105. Both with the scale-up method and the ML estimator (Table 2), increasing the size leads to increased standard errors of the estimates, passing from 0.0005 to 0.005 with the scale-up estimator and from 0.0001 to 0.0006 with the ML estimator. On the contrary, increasing the number of used subpopulations does not affect the standard errors, which remain constant.

In case of non-homogeneous subgroups (Table 3), when the degree of coverage of the total subpopulations sizes increases, the relative bias become smaller. Indeed, raising the total size of the subpopulations from 20.000 to 30.200 subjects, we observe that the relative bias becomes about one half with the scale-up estimator

Table 3: Bias and standard error in non-overlapped and non-homogeneous subpopulations. Values are expressed × 100.

total	*Scale-up*			*ML*		
size	$\hat{p}_0$	*rel. bias*	*s.d.*	$\hat{p}_0$	*rel. bias*	*s.e.*
20.000	0,11	8,53	0,05	0,10	0,10	0,01
30.200	0,10	4,38	0,05	0,10	-0,03	0,01

Table 4: Relative bias and standard error in overlapped subpopulations. Values are expressed × 100.

overlap.	*Scale-up*			*ML*		
rate	$\hat{p}_0$	*rel. bias*	*s.e.*	$\hat{p}_0$	*rel. bias*	*s.e.*
0	0,10	1,00	0,05	0,10	-0,03	0,01
25	0,11	9,27	0,05	0,10	-0,28	0,01
50	0,12	22,56	0,06	0,10	-0,24	0,01
75	0,15	52,97	0,06	0,10	0,12	0,01
100	0,31	213,81	0,09	0,10	-0,04	0,01

and about one third with the ML estimator. Comparing the results with 25 non-homogeneous subgroups (leading to a total size of the subpopulations of 20.000) with those obtained with homogenous subpopulations, the relative bias is higher in the non homogeneous case. Indeed, considering the non homogeneous case and the homogenous one, the relative bias is about 8.53% and 6.43% respectively for the scale-up estimator and about 0.10% and 0.02% for the ML estimator.

4.2 Overlapped subpopulation

With the network scale-up method, as shown in Table 4, the relative bias increases from about 1.00% with no overlapping rate of 0 up to 213.81% with an overlapping rate among subpopulations of 100%. Also the standard errors increase when the degree of overlapping increases. On the contrary, with the ML estimator, estimates are substantially unbiased for any degree of overlapping and standard errors are almost constant at the 0.0001 level.

5 Discussion

The first goal of our analysis was to identify the optimal number of subpopulations of the known size to employ in planning investigations based on the scale-up method or the ML estimator. Killworth (Killworth, 1998a; Killworth, 1998b) suggested, based on a heuristic reasoning on a concrete data set, that the "optimal" number of subpopulations is about 20, obtaining in this way substantially unbiased results. Our simulations show that the scale-up method, whatever number of subgroups is used, leads to biased estimates. On the contrary, the results for the maximum

likelihood estimator, are clearly indicating that the estimates are unbiased for any number of subpopulation considered. This makes the Killworth's suggestions too conservative for the latter estimator, where the biggest issue seems to be the sizes of the subpopulations. Our results indicate a marked loss in efficiency of the ML estimator as the size of the subpopulations tends to increase. Also the bias in the scale-up estimator seems to depend on the size of the social network considered, reducing markedly in presence of bigger social networks. However, this result is of little usefulness in planning the research, since it is not under direct control of the researcher, who usually doesn't know the network of the people being interviewed a priori.

The second goal was to analyze the behavior of both estimators with homogeneous or non-homogeneous subpopulations. From our results there is a slight evidence that it would be preferable to have homogeneous subpopulations, both for reducing the bias and to increase efficiency. Although under control of the researcher, this parameter is often very difficult to obtain in concrete situations, when the amount of data available on the sizes of the subpopulations are not a great number.

The third goal was to see if the statistical behavior of the estimators improved increasing the subpopulation total size, i.e.: the degree of coverage of the target population T. Our simulations show that increasing the total size of known subpopulations leads to an increased bias in both estimators, markedly for the scale-up method and only slightly for the ML estimator. Published results indicate that as the coverage increases, the standard errors of the social network size estimates (Equation 2.8, (Killworth, 1998a)) lower down. This is not the outcome of our simulations, where the standard error of target subpopulation size does not seem to decrease: this is most probably due to the specific setting of our simulations, where the standard error of $\hat{c}_i$ is negligible in the estimate of e_0.

The fourth goal was aimed at understanding the behavior of both estimators in case of using subpopulations with some degree of overlapping. In this case, the scale-up estimator is showing a marked bias, increasing impressively as the degree of overlapping increases. On the contrary, the ML estimator leads to substantially unbiased estimates, with an efficiency independent from the degree of overlapping.

The impact of our findings is of some direct interest in epidemiological and social research. Based on the above considerations, all the estimates of HIV and AIDS populations (Killworth, 1998a) based on the network scale-up method are most probably biased upward, with a magnitude which depends on the degree of overlapping in subpopulations, which cannot be excluded in the cited researches. These results confirm Killworth's (Killworth, 1998b) findings, which, based on a reanalysis of the previous data, showed a marked difference in the estimates obtained with the ML estimator.

The results of our work are however far from being conclusive. First, we did not investigate how the precision of estimate of c affects the estimate of e, which could be of direct interest in field work, also because of the possible errors in estimating the social network. From a statistical point of view, the two step estimate, which requires the estimate of c to be plugged in the formulas for estimating e is clearly unsatisfactory. More research is needed to get a joint estimate and joint standard

errors for both quantities. Finally, but not of less importance, unbiasedness is surely not by itself an assurance of the correctness of the final estimates. Other factors, such for instance transmission, barrier and estimation effects are a potential greater source of bias, which is difficult to be evaluated and is specific for each problem studied.

References

[1] Bernard, H.R., Johnsen, E.C., Killworth, P.D., and Robinson, S. (1991): Estimating the size of an average personal network and of an event subpopulation: some empirical results. *Social Science Research*, **20**, 109-121.

[2] Bernard, H.R., Johnsen, E.C., Killworth, P.D., and Robinson, S.(1989): Estimating the size of an average personal network and of an event subpopulation. In M. Kochen (Ed.): *The Small World*, 159-175.

[3] Killworth, P.D., Johnsen, E.C., McCarty, C., Shelley, G.A., and Bernard, H.R. (1998a): A social network approach to estimating seroprevalence in the United States. *Social Networks*, **20**, 23-50.

[4] Killworth, P.D., McCarty, C., Bernard, H.R., Shelley, G.A., and Johnsen, E. C. (1998b): Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluation Review*, **22**, 289-308.

[5] Bernard, H.R., Killworth, P.D., Johnsen, E.C., Shelley, G.A., and McCarty, C. (2001): Estimating the ripple effect of a disaster. *Connections*, **24**, 18-22.

[6] Freeman, L.C. and Thompson, C.R. (1989): Estimating acquaintanceship volume. In M. Kochen (Ed.): *The Small World*, 147-158.

[7] McCarty, C., Killworth, P.D., Bernard, H.R., Johnsen, E.C., and Shelley, G.A. (2000): Comparing two methods for estimating network size. *Human Organization*, **60**, 28-39.

[8] Bernard, H.R., Johnsen, E.C., Killworth, P.D., McCarty, C., Shelley, G.A. and Robinson, S. (1990): Comparing four different methods for measuring personal social networks. *Social Networks*, **12**, 179-215.

[9] Killworth, P.D., Johnsen, E.C., Bernard, H.R., Shelley, G.A., and McCarty, C. (1990): Estimating the size of personal networks. *Social Networks*, **12**, 289-312.

[10] Johnsen, E.C., Bernard, H.R., Killworth, P.D., Shelley, G.A., and McCarty, C. (1995): A social network approach to corroborating the number of AIDS/HIV victims in the US. *Social Networks*, **17**, 167-187.