

# A Note on Network Sampling in Drug Abuse Research

Marinus Spreen<sup>1</sup> & Moniek Coumans<sup>2</sup>

*Maastricht University & Addiction Research Institute Rotterdam*

*In this article we discuss a network sampling design that can be applied in drug abuse research at the community level. At this level often some partial sampling frame such as the register of a drug aid agency is available. This partial sampling frame can be used as the start of a network sample. Each selected registered drug abuser mentions his relationships with other drug abusers, and from those newly mentioned drug abusers who are not registered a second probability sample is drawn. Using this network sampling design the mean contact rates between clients, between clients and non registered drug abusers and between non registered drug abusers can be estimated despite the unknown total number of drug abusers. The design is illustrated by an analysis of the network data of the Heerlen Drug Monitoring System.*

## INTRODUCTION

In studies of drug abuse populations such as heroin, cocaine and methadone users, standard probability sampling designs are often impractical due to imperfect sampling frames which decrease the possibilities of formal inference (Van Meter, 1992). Some practical problems are the unknown size of the population, the geographical clustering of groups of users, the identification of the target group, the establishment of contact, etc. To cope with these problems a frequently applied data collection procedure is a link-tracing procedure. A link-tracing procedure is a data collection method that follows social relations in the study population by using the contact patterns that exist between the drug abusers. The classical link-tracing procedure is the snowball sampling technique (Goodman, 1961), in which persons are asked to mention a fixed number  $k$  of other persons, who, in turn,

---

<sup>1</sup>Department of Methodology & Statistics / Maastricht University / P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: Marinus.Spreen@stat.unimaas.nl.

<sup>2</sup>Addiction Research Institute (IVO), Rotterdam, The Netherlands. E-mail: coumansm@GGDOZL.NL.

are selected for extending the initial sample by mentioning  $k$  other persons, and so on .... Some other link-tracing techniques are the snowball design of Frank (1977a), the random walk design of Klovdahl (1989), and the adaptive cluster sampling design of Thompson (1991). An overview of link-tracing data collection techniques is given in Spreen (1992).

In drug abuse studies link-tracing methods are mainly used as a tool to find a substantial amount of respondents in order to describe the population in terms of individual characteristics. Standard nonprobability sampling designs such as targeted sampling (Watters & Biernacki, 1989) have been elaborated with the intention to mirror an initial simple random sample from the total population. Subsequently the link-tracing procedure starts from this initial sample. Data obtained by link-tracing sampling methods can also be used to describe the population in terms of structural characteristics (Snijders & Frank, 2000; Thompson & Frank, 2000). Until now the analysis of link-tracing data from a structural perspective has been largely ignored in drug abuse studies.

The purpose of this article is to introduce a network sampling design that can be used in community-based drug abuse studies. Often some registers of aid agencies are available from which to start a network sample with an initial probability sample. We will illustrate this design with an analysis of the network data from the Heerlen Drug Monitoring System (Coumans, et al., 2000). The tentative question we explore in this article is whether community oriented prevention/intervention strategies could be applied using the networks of the clients of the aid agencies. Therefore estimators are needed for the mean number of contacts within and between the clients of the aid agencies and those drug abusers who are not client. Another structural study of the client population only is discussed in Spreen & Coumans (2000). First some graph theoretical definitions and notation must be introduced.

### Graph theoretical problem definition

We consider an undirected graph  $G$  with vertex set  $V = \{1, 2, 3, \dots, N\}$  and adjacency matrix  $\mathbf{Y}$ , representing a set of social actors and some relationship between them. The adjacency matrix is defined on the set  $V^2$  of the ordered pairs of vertices;  $Y_{ij} = 1$  if there is an edge between vertices  $i$  and  $j$ , and  $Y_{ij} = 0$  otherwise ( $Y_{ij} = 0$  for all  $i$ ). Since the graph is undirected,  $Y_{ij} = Y_{ji}$  for all  $i, j$ . Based on some binary auxiliary variable  $Z$ , vertex set  $V$  can be partitioned into two disjoint vertex subsets  $\alpha$  and  $\beta$  ( $\alpha \cap \beta = \emptyset$ ), i.e.

- $\alpha = \{i \in V \mid Z = 1\}$  with order  $N_\alpha$
- $\beta = \{u \in V \mid Z = 0\}$  with order  $N_\beta$ .

For the sake of clarity throughout the paper, vertices  $i$  and  $j$  refer to subset  $\alpha$ ;  $u$  and  $v$  to subset  $\beta$ .

Based on vertex sets  $\alpha$  and  $\beta$  population graph  $G$  can be decomposed into three subgraphs:

1. subgraph  $G_\alpha$  with arcs between the vertices of set  $\alpha$ ,
2. subgraph  $G_\beta$  with arcs between the vertices of set  $\beta$ , and
3. subgraph  $G_{\alpha\beta} = G_{\beta\alpha}$  with arcs between the vertices of sets  $\alpha$  and  $\beta$ .

Figure 1 is an illustration of population graph  $G$  with vertex set  $V = \{1, 2, \dots, 7\}$  of order  $N = 7$  and size  $R = 10$ , i.e.  $G$  consists of  $N$  vertices and  $R$  arcs. Based on auxiliary variable  $Z$  vertex set  $V$  is partitioned into subset  $\alpha = \{1,2,3,4\}$  and subset  $\beta = \{5,6,7\}$ .

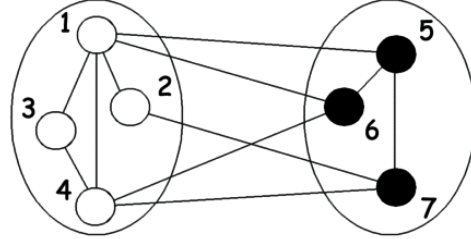


Figure 1 Population  $G$  with vertex set  $V$

The number of relations between vertices of  $\alpha$ , i.e. the size of subgraph  $G_\alpha$ , is denoted

$$R_\alpha = \frac{1}{2} \sum_{i \in \alpha} \sum_{j \in \alpha} Y_{ij} ; \quad (1)$$

between the vertices of  $\beta$ , i.e. the size of subgraph  $G_\beta$ , is denoted

$$R_\beta = \frac{1}{2} \sum_{u \in \beta} \sum_{v \in \beta} Y_{uv} ; \quad (2)$$

and between vertices of  $\alpha$  and  $\beta$ , i.e. the size of subgraph  $G_{\alpha\beta} = G_{\beta\alpha}$  is denoted

$$R_{\alpha\beta} = R_{\beta\alpha} = \sum_{i \in \alpha} \sum_{u \in \beta} Y_{iu} . \quad (3)$$

In Figure 1  $R_\alpha = 4$ ,  $R_\beta = 2$  and  $R_{\alpha\beta} = R_{\beta\alpha} = 5$ .

The mean number of relations for  $i \in \alpha$  with other  $j \in \alpha$  is

$$\mu_\alpha = \frac{2}{N_\alpha} R_\alpha ; \quad (4.a)$$

for  $i \in \alpha$  with  $u \in \beta$  is

$$\mu_{\alpha\beta} = \frac{1}{N_\alpha} R_{\alpha\beta} . \quad (4.b)$$

The mean number of relations for  $u \in \beta$  with other  $u \in \beta$  is

$$\mu_\beta = \frac{2}{N_\beta} R_\beta ; \quad (5.a)$$

for  $u \in \beta$  with  $j \in \alpha$  is

$$\mu_{\beta\alpha} = \frac{I}{N_\beta} R_{\beta\alpha}. \quad (5.b)$$

In Figure 1,  $\mu_\alpha = \frac{4}{4}$ ,  $\mu_\beta = \frac{2}{3}$ ,  $\mu_{\alpha\beta} = \frac{5}{4}$  and  $\mu_{\beta\alpha} = \frac{5}{3}$ .

In Figure 1 we also observe a seemingly trivial relationship

$$\frac{\mu_{\alpha\beta}}{\mu_{\beta\alpha}} N_\alpha = N_\beta, \quad (6)$$

which can, however, be used to get an indication of the total number of vertices.

### Sampling design and estimation

We consider a simple random sample  $S$  of  $n$  vertices from  $\alpha$ . Each  $i \in S$  indicates his relations in  $G$ . This implies that  $i$  mentions relations with other vertices  $j \in \alpha$  and  $u \in \beta$ . The number of vertices vertex  $i \in \alpha$  mentions in  $G_\alpha$  is denoted  $a_{i|\alpha}$ , the number of vertices  $i \in \alpha$  mentions in  $G_\beta$  is denoted  $a_{i|\beta}$ . The total number of relations the  $n$  selected vertices mention in  $G_\alpha$  is denoted  $R_\alpha(S)$ ; the total number of relations the  $n$  selected vertices mention with vertices  $u \in \beta$  is denoted  $R_{\alpha\beta}(S)$ .

Define by  $\beta(S)$  the set of newly mentioned  $u \in \beta$  of order  $N_\beta(S)$ ; a sample  $T$  of size  $m$  according to a known probability design is drawn from  $\beta(S)$ . Each  $u \in T$  indicates his relations in  $G$ . This implies that  $u$  may mention relations with other vertices  $i \in \alpha$  and  $w \in \beta$ . The number of vertices vertex  $u \in \beta$  mentions in  $G_\beta$  is denoted  $a_{u|\beta}$ ; the number of vertices  $u \in \beta$  mentions in  $G_\alpha$  is denoted  $a_{u|\alpha}$ . The total number of relations the  $m$  selected vertices mention in  $G_\beta$  is denoted  $R_\beta(T)$ ; the total number of relations the  $m$  selected vertices mention with vertices  $i \in \alpha$  is denoted  $R_{\beta\alpha}(T)$ .

To estimate the average number of contacts vertex  $i \in \alpha$  has with other vertices  $j \in \alpha$  the well-known graph total estimators (Frank, 1971, 1977a,b, 1978; Capobianco & Frank, 1982) can be applied, i.e.

$$\begin{aligned} \hat{\mu}_\alpha &= \frac{2}{N_\alpha} \hat{R}_\alpha \\ &= \frac{2}{N_\alpha} \left( \frac{1}{2} \sum_{i,j \in S} \frac{Y_{ij}}{\pi_{ij}} + \sum_{\substack{i \in S \\ j \in \alpha \setminus S}} \frac{Y_{ij}}{\pi_{ij}} \right) \end{aligned} \quad (7)$$

where

$$\pi_{ij} = 1 - \frac{\binom{N_\alpha - n}{2}}{\binom{N_\alpha}{2}} = 1 - \frac{(N_\alpha - n - 1)(N_\alpha - n)}{(N_\alpha - 1)N_\alpha}$$

An unbiased variance estimator is given by

$$\widehat{Var}(\mu_a) = \left( \frac{2}{N_a} \right)^2 \left( \frac{(q_4 - q_2)}{(1 - q_2)^2 (1 - 2q_2 + q_4)} R_\beta^2(S) + \frac{(q_3 - q_4)}{(1 - q_2)^2 (1 - 2q_2 + q_4)} Q(S) + \frac{(q_2 - 2q_3 + q_4)}{(1 - q_2)^2 (1 - 2q_2 + q_4)} R_a(S) \right) \quad (8)$$

where  $q_H$  is defined as the inclusion probability that  $H$  specified distinct vertices are in the complement  $\bar{S}$  for  $H = 2, 3, 4$ , i.e.

$$q_H = \frac{\binom{N_\alpha - n}{H}}{\binom{N_\alpha}{H}} \quad (9)$$

and  $Q(S) = \sum_{i=1}^n (Y_{i+})^2$  is the sum of squares of the degrees of the sampled vertices.

Estimator (7) can also be used to estimate the average number of vertices at distance 2 in  $G_\alpha$  (see Spreen & Coumans, 2000).

The mean degree of relations  $i \in \alpha$  has with vertices from  $\beta$  can be estimated by the conventional HT-estimator (Särndal et al, 1992), i.e.

$$\begin{aligned} \hat{\mu}_{\alpha\beta} &= \frac{1}{N_\alpha} \sum_{i \in S} \sum_{u \in \beta} \frac{Y_{iu}}{\pi_i} \\ &= \frac{R_{\alpha\beta}(S)}{n} \end{aligned} \quad (10)$$

where  $\pi_i = n/N_\alpha$  and  $\pi_{ij} = n(n-1)/N_\alpha(N_\alpha - 1)$

The variance estimator is defined as

$$\widehat{Var}(\hat{\mu}_{\alpha\beta}) = \frac{1}{N_\alpha^2} \sum \sum_s \left( 1 - \left( \frac{\pi_i \pi_j}{\pi_{ij}} \right) \right) \frac{a_{i|\beta}}{\pi_i} \frac{a_{j|\beta}}{\pi_j} \quad (11)$$

To estimate the mean degree of relations  $u \in \beta$  has with vertices from  $\alpha$  and  $\beta$ , the unknown order of  $G_\beta$  will provide some problems because of the computation of design-based inclusion probabilities. A strategy to avoid these problems is to use the weighted sample mean as an estimator of  $N_\beta$  (Särndal et al, 1992), i.e.  $\hat{N}_\beta = \sum_T (1/\pi_u)$ , where  $N_\beta$  does not need to be known. We propose to approximate the inclusion probability of  $u \in T$  by the relative frequency of the observed amount of relations with  $i \in \alpha$ . Denote the number of vertices  $u \in T$  that have mentioned  $d$  vertices in  $G_\alpha$  by  $M_d$  for  $d = 1, 2, \dots$ . Then we may define

$$\pi_{u|d} = \frac{M_d}{m} \quad (12)$$

as the first-order inclusion probability of vertex  $u \in \beta$ .

To compute the second-order inclusion probability that vertices  $u$  and  $w$  are included in sample  $T$  we use the same approach. Denote the number of pairs of vertices with the same sum of degrees  $e$  by  $M_e$  for  $e = 2, 3, \dots$ . The pairwise inclusion probability is defined as the relative frequency

$$\pi_{uw|e} = \frac{M_e}{(m-1)} \quad (13)$$

An estimator for the average number of relations  $u \in \beta$  has with other  $w \in \beta$  can be defined as

$$\hat{\mu}_\beta = \frac{2\hat{R}_\beta}{\hat{N}_\beta} = \frac{\sum_{u \in T} \frac{a_{u|\beta}}{\pi_u}}{\sum_{u \in T} \frac{1}{\pi_u}} \quad (14)$$

with variance estimator

$$\hat{V}ar(\hat{\mu}_\beta) = \frac{1}{\hat{N}_\beta^2} \sum_T \sum_T \left( 1 - \left( \frac{\pi_u \pi_w}{\pi_{uw}} \right) \right) \left( \frac{a_{u|\beta} - \hat{\mu}_\beta}{\pi_u} \right) \left( \frac{a_{w|\beta} - \hat{\mu}_\beta}{\pi_w} \right) \quad (15)$$

The average number of relations vertex  $u \in \beta$  has with other vertices  $i \in \alpha$  can be estimated by

$$\hat{\mu}_{\beta\alpha} = \frac{\hat{R}_{\beta\alpha}}{\hat{N}_\beta} = \frac{\sum_{u \in T} \frac{a_{u|\alpha}}{\pi_u}}{\sum_{u \in T} \frac{1}{\pi_u}} \quad (16)$$

with variance estimator

$$\hat{V}ar(\hat{\mu}_{\beta\alpha}) = \frac{1}{\hat{N}_\beta^2} \sum_T \sum_T \left( 1 - \left( \frac{\pi_u \pi_w}{\pi_{uw}} \right) \right) \left( \frac{a_{u|\alpha} - \hat{\mu}_{\beta\alpha}}{\pi_u} \right) \left( \frac{a_{w|\alpha} - \hat{\mu}_{\beta\alpha}}{\pi_w} \right) \quad (17)$$

## ILLUSTRATION

As an illustration we analyse network data obtained from the Heerlen Drug Monitoring System (DMS). The purpose of the DMS is to describe the population of marginalised (nearly) daily users of opiates and/or other drugs (like cocaine) in terms of prevalence, patterns of use, problems (with use), social relationships and contacts with aid agencies. The system is based on three pillars, knowing:

1. information collected by a group of key informants who regularly report on phenomena and developments in and involving drug use;
2. ethnographic qualitative information about the natural context in which drug use takes place is collected by community field workers;

3. quantitative information about distributions and associations of various individual and relational characteristics in the population is collected by a network sample.

We focus on the contact rates between clients of the aid agencies, between clients and hard drug users that are not registered (hereafter called NR's), and between the NR's. From a health promotion perspective these patterns are relevant because they give an impression of the extent to which aid agencies could employ their clients for community prevention/ intervention strategies to reach also a substantial amount of NR's. Note that a relation between two drug abusers is viewed as a channel of communication.

In Heerlen 435 hard drug users were registered as a client of the aid agencies (at June 1 1999). Local experts assessed this figure to be a substantial part of the total unknown population. Because the purpose of the DMS is to draw on a regular base samples it was decided to use the client list, i.e.  $\alpha = \{1, 2, \dots, 435\}$ , as an initial sampling frame. A simple random sample without replacement  $S$  of size  $n = 39$  was drawn from  $\alpha \subset V$  and for each  $i \in S$  his relations with other users (alters) were observed and individual characteristics about the alters were collected. The criteria for the alters to be included in the sample were:

- respondent and alter must meet each other on a daily or regular base in Heerlen,
- respondent and alter must know each others sur- and family name,
- alter must know the respondent as a hard drug (heroin, cocaine, etc.) user.

The 39 selected clients mentioned 110 other clients with whom they reported 164 relations, i.e.  $R_\alpha(S) = 164$ . They also mention 81 drug abusers who were not registered, i.e.  $N_\beta(S) = 81$ , with whom they reported 93 relations, i.e.  $R_{\alpha\beta}(S) = 93$ .

The next step of the sample was to draw a random selection of these  $N_\beta(S) = 81$  NR's. Due to all kinds of practical problems, we managed a random selection of  $m=18$ . These 18 non registered drug abusers mentioned 66 other users who were client of the aid agencies with whom they reported 91 relations, i.e.  $R_{\beta\alpha}(T) = 91$ . A total of 24 other NR's were mentioned by this group with whom they reported 26 relations, i.e.  $R_\beta(T) = 24$ .

From this sample we estimated the following averages (between brackets are the standard errors):

- an arbitrary client of the aid agencies has about 4 close relationships with other clients and about 2 with non registered hard drug users, i.e.  $\hat{u}_\alpha = 4.18(0.33)$  and  $\hat{u}_{\alpha\beta} = 2.38(0.38)$ .
- an arbitrary non registered hard drug user has about 1 to 2 close relations with other non registered drug abusers and about 5 with clients, i.e.  $\hat{u}_\beta = 1.67(1.20)$  and  $\hat{u}_{\beta\alpha} = 4.81(1.58)$ .

Thus both groups of drug abusers mentioned about 7 other drug abusers. The standard errors of the means of the non registered hard drug users are higher than those of the clients. A part of this difference is due to the small sample size of  $T$ . However because  $T$  was drawn randomly, the high standard errors give also an indication that the non registered hard drug users vary more in their contact rates with clients and NR's.

Clients mention about 64% in their own group while the NR's mention about 25% within their own group. From a community based health intervention strategy it is important to notice that the clients

have the tendency to have relations with other clients, although the non registered drug abusers also have a high proportion relationships with clients. Consequently to diffuse policy measures via communication channels of the clients only will be debatable, because they have the tendency to have contacts with other clients. However the non registered drug abusers could be 'easily' reached with some extra effort, as is shown by the average contact rates.

Another feature of this network design is that it is possible to get an indication of the total size of the drug abuse population in Heerlen. If we accept the assumption that each heroin abuser that is not registered knows at least 1 other heroin abuser that is a client of the aid agencies, then by using equation (1.6) we can define a simple ratio estimator for the total:

$$\hat{N} = N_{\alpha} + \frac{\hat{\mu}_{\alpha\beta}}{\hat{\mu}_{\beta\alpha}} N_{\alpha} = 435 + \frac{2.38}{4.81} 435 \approx 650$$

Local experts could not imagine that there was a large group of drug abusers in Heerlen not meeting this restrictive assumption. An advantage of this simple estimator is that in principle respondents do not need to give the identities of the people they mention. However, this implies that a random sample from the NR's must be replaced by a nonprobability sample. For monitoring purposes it will be a more easy and cheaper way to estimate the population of this difficult survey group.

## REFERENCES

- Capobianco, M. & Frank, O. (1982). Comparison of statistical graph-size estimators. *Journal of Statistical Planning and Inference*, 6, 87-97.
- Coumans, A.M., Neve, R.J.M. & Mheen, H. van de. (2000). Het proces van marginalisering en verharding in de drugsce van Parkstad Limburg.(The process of marginalisation and hardening in the Heerlen drugsce). IVO, Rotterdam.
- Frank, O. (1971). *Statistical inference in graphs*. FOA Repro, Stockholm.
- Frank, O. (1977a) Survey sampling in graphs . *Journal of Statistical Planning and Inference*, 1, 235-264.
- Frank, O. (1977b) Estimation of Graph Totals . *Scandinavian Journal of Statistics*, 4, 81-89.
- Frank, O. (1978) Sampling and estimation in large social networks . *Social Networks*, 1, 91-101.
- Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32, 148-170.
- Klodahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In: M. Kochen(ed.), *The Small World*, Norwood, N.J.:Ablex.

- Meter van, K.M. (1990). Methodological and design issues: Techniques for assessing the representatives of snowball sampling. In: *The collection and interpretation of data from hidden populations*, NIDA Research Monograph 98, Rockville.
- Särndal, C. E., Swensson, B. & Wretman, J. (1992). Model assisted survey sampling. New York: Springer-Verlag.
- Snijders, T.A.B. & Frank, O. (2000). Estimation of population characteristics from one-wave snowball samples in structured populations. *Paper presented at the Second International Workshop on Network Sampling, Maastricht (The Netherlands), March 2-4.*
- Spreen, M. (1992) Rare populations, hidden populations, and Link-Tracing Designs: What and Why? *Bulletin de Methodologie Sociologique*, 36, 34-58.
- Spreen, M. & Coumans, M. (2000) Network sampling hard drug users. A structural analysis of the clients of aid agencies in Heerlen. To appear: *Kwantitatieve Methoden*.
- Thompson, S.K. (1991) Stratified adaptive cluster sampling, *Biometrika*, 78, 2, 389-397.
- Thompson, S.K & Frank, O. (2000). Model-based estimation with link-tracing sampling designs, *Survey Methodology*, 26, 1, 87-98.
- Watters, J.K. & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.

