Network sampling hard drug users:

a structural analysis of the clients of aid agencies in Heerlen

Marinus Spreen[1]
Department of Methodology & Statistics
Maastricht University

Moniek Coumans
Addiction Research Institute
Rotterdam

Abstract

Link-tracing methods are data collection methods that follow social relations in the target population. In studies of hard drug users populations link-tracing methods are often applied because formal sampling frames are lacking. Individual as well as relational information is collected with link-tracing methods. In most studies the focus is mainly on describing the population by individual characteristics despite the available structural information. In this paper we discuss a structural analysis of the clients of the drug aid agencies in Heerlen, The Netherlands. The relational data are obtained from a random sample of the client registers of the drug aid agencies in Heerlen. For each selected client his/her hard drug using contacts in the total hard drug users population were observed. Using the concept of a graph total one may assess the effectivity of community-based prevention/intervention strategies with as purpose to reach hard drug users inside and outside the aid agencies.

[1] Corresponding author: Marinus Spreen, Maastricht University, Department of Methodology & Statistics, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Phone: +31433882277. E-mail: Marinus.Spreen@Stat.Unimaas.nl.

1. Introduction

The need for accurate statistical information in society applies also to populations of heroin, cocaine and methadone users, hereafter referred to as hard drug users. The use of probability samples is widely accepted to be the most efficient and accurate way for studying populations using partial information. However, in studies of hard drug 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 target population by using the contact patterns that exist between individuals, here hard drug users. 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, 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 (1977), 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 studies of hard drug users 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. However, data obtained by link-tracing methods can also be used to describe the population in terms of structural or relational characteristics (Snijders & Frank, 2000). For instance, one may estimate the total number of Intravenous Drug Users (IDU's) and the total number of relations between IDU's. In this paper we focus on the estimation of relational characteristics. Until now the analysis of link-tracing data from a structural perspective has been largely ignored in studies of hard drug users. The purpose of this paper is to introduce some known design-based inference methods that can be used when describing a hard drug

users population by means of a link-tracing procedure. To express the emphasis on relational characteristics, we prefer the term network sampling.

In this paper we consider the situation in which the members of some population G are either in subpopulation α or in β. Furthermore members of G are supposed to have relations with other members of G. For subpopulation α we have a perfect sampling frame, for subpopulation β the sampling frame is lacking. From subpopulation α a one-wave snowball sample is drawn, i.e., a simple random sample is drawn from α, and each selected respondent mentions his relations with other hard drug users in G. Relational data collected this way can be used to estimate by design the number of relations within subpopulation α, and between subpopulations α and β. Also the number of indirectly connected pairs of members in α can be estimated. As an illustration we use relational data from the Heerlen Drug Monitoring System (Coumans, Neve & van de Mheen, 2000). Here subpopulation α are the clients of the aid agencies and subpopulation β are hard drug users outside the aid agencies (hereafter referred to as nonclients). Section 2 introduces some necessary notation and definitions. Section 3 briefly discusses the estimation principles of link-tracing samples. As an illustration in section 4 a structural analysis of the population of hard drug users in Heerlen is described. Finally section 5 gives some conclusions.

2. Notation and definitions

An undirected graph G with vertex set $V = \{1,2,...,N\}$ and adjacency matrix $\mathbf{Y}$, representing a set of social actors and some relationship between them, is considered. 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 ($Y_{ii} = 0$ for all i). Since the graph is undirected $Y_{ij} = Y_{ji}$ for all i, j. The vertices in G are allowed to have p vertex characteristics $X_i \in \Re^p \left(i \in \{1,...,N\}\right)$. Vertex set V is composed of two subsets α and β based on an auxiliary binary variable Z. In the Heerlen research (see section 4) Z is an indicator variable for client registration, i.e.

1. $\alpha = \{i \in V | Z = 1\}$, i.e. the clients of the aid agencies in G

2. $\beta = \{j \in V | Z = 0\}$, i.e. the nonclients of the aid agencies in G.

The number of vertices in α is denoted N_α, the number of vertices in β N_β.

For illustrative purposes consider the population graph G with vertex set $V = \{1,2,...,6\}$ in Figure 1 where $\alpha = \{1,2,3\}$ and $\beta = \{4,5,6\}$.

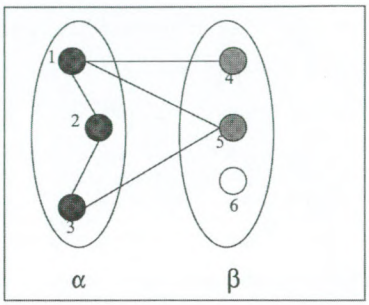

Figure 1 Illustration graph G

The population degree of vertex i with other $j \in V$ is denoted $d_i = \sum_{j \in V} Y_{ij}$; the degree of vertex i with other $j \in \alpha$ by $Y_{i+} = \sum_{j \in \alpha} Y_{ij}$. The number of relations between vertices of α is denoted

$$R_\alpha = \frac{1}{2} \sum_{i \in \alpha} \sum_{j \in \alpha} Y_{ij}; \tag{2.1}$$

and between vertices of α and β,

$$R_{\alpha\beta} = \sum_{i \in \alpha} \sum_{j \in \beta} Y_{ij}. \tag{2.2}$$

In Figure 1 $R_\alpha = 2$ and $R_{\alpha\beta} = 3$. Note that we do not consider relations within population β in this paper.

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

$$\mu_\alpha = \frac{2}{N_\alpha} R_\alpha ; \tag{2.3}$$

the mean number of relations for $i \in \alpha$ with $j \in \beta$ is

$$\mu_{\alpha\beta} = \frac{1}{N_\alpha} R_{\alpha\beta} . \tag{2.4}$$

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

The sociometric distance is defined as the length of the shortest path between two vertices. In this paper we focus only on sociometric distances between the vertices of subset α. For the ordered pairs of vertices in α^2 a sociometric distance *1* between a pair of vertices (i, j) is the existence of a relation, and defined by

$$\{d(i,j) = 1\} = \{(i,j) \in \alpha^2 \mid Y_{ij} = 1\}, \tag{2.5}$$

and sociometric distance *2* by

$$\{d(i,j) = 2\} = \{(i,j) \in \alpha^2 \mid (1 - Y_{ij}) \max_{k \in V} Y_{ik} Y_{jk} = 1\} . \tag{2.6}$$

If no path is observed between *i* and *j*, i.e. they are vertices in different components of the graph, the sociometric distance is defined infinite, denoted $d(i,j) = \infty$; by convention $d(i,i) = 0$. In Figure 1 the sociometric distances between pairs (1,2) and (2,3) is one, i.e. $d(1,2) = d(2,3) = 1$, and the sociometric distance between vertex pair (1,3) is two, i.e. $d(1,3) = 2$. Note that vertex pair (1,3) is indirectly related via vertex 2 but also via vertex 5. We return to this in section 4.

100

The number of sociometric distances *1* and *2* can alternatively be described using the concept of graph totals (Frank, 1977b). A graph total is defined as the sum of the values of some real-valued function F of vertex labels and arc frequencies in the ordered pairs of vertices denoted by

$$T = \sum_{(i,j)\in V^2} F_{ij}\,, \tag{2.7}$$

where $F_{ij} = f(F_{ii}, F_{jj}, F_{ij}, F_{ji})$ is *1* if the dyad (a pair of vertices) has a certain property. Graph total (2.7) can be understood as a general graph parameter that can be used to describe various graph characteristics. For instance, the number of relations between IDU's who know each other longer than 5 years is an example of a graph total. In this graph total F_{ii} and F_{jj} determine whether i and j are IDU's and $F_{ij} = F_{ji}$ whether they know each other. Moreover, if they do, whether they know each other longer than 5 years. In this paper we consider only graph totals based on the observation of one relation, i.e. whether there is a relation between i and j. Consequently throughout this paper $F_{ij} = Y_{ij}$. The number of dyads for which $F_{ij} = 1$ is also called a dyad count (Frank, 1978). The specific graph totals of the form (2.7) we use in this paper are the total number of sociometric distances *1* and *2*.

The total number of sociometric distances *1* between the vertex pairs in α^2 is identical to the number of relations (2.1), and can be alternatively defined as

$$T_1 = \frac{1}{2}\sum_{i=1}^{N_\alpha}\sum_{j=1}^{N_\alpha} F_{ij}^{(1)}\,, \tag{2.8}$$

where $F_{ij}^{(1)} = \begin{cases} 1 \text{ if } d(i,j) = 1 \\ 0 \text{ otherwise} \end{cases}$

The total number of sociometric distances *2* between the vertex pairs in α^2 is denoted

$$T_2 = \frac{1}{2}\sum_{i=1}^{N_\alpha}\sum_{j=1}^{N_\alpha} F_{ij}^{(2)} \tag{2.9}$$

where $F_{ij}^{(2)} = \begin{cases} 1 \text{ if } d(i,j) = 2 \\ 0 \text{ otherwise} \end{cases}$.

In Figure 1 $T_1 = R_\alpha = 2$ and the number of pairs at distance *2* is $T_2 = 1$ (vertex pair $(1,3)$).

3. Basic graph sampling and estimation principles

Suppose simple random sample without replacement S of size n is drawn from $\alpha \subset V$. For each $i \in S$ vertex characteristic X_i is observed. The standard Horvitz-Thompson estimator (Horvitz & Thompson, 1952) or the π-estimator (Särndal, Swensson & Wretman, 1992) of the total $T_X = \sum_{i \in \alpha} X_i$ is

$$\hat{T}_X = \sum_{i \in S} \frac{X_i}{\pi_i}, \tag{3.1}$$

where $\pi_i = n/N_\alpha$ is the probability for i to be included in the sample.

This estimator is unbiased provided that $\pi_i > 0$ for all $i \in S$. The variance is defined as

$$Var(\hat{T}_X) = \sum_{i \in \alpha} \sum_{j \in \alpha} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) X_i X_j \tag{3.2}$$

An unbiased estimator of this variance is given by Horvitz and Thompson (1952), provided that $\pi_{ij} > 0$ for all $i, j \in S$,

$$\hat{Var}(\hat{T}_X) = \sum_{i \in S} \sum_{j \in S} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}} \right) \frac{X_i}{\pi_i} \frac{X_j}{\pi_j}, \tag{3.3}$$

Another frequently used variance estimator in HT-estimation theory is that of Yates and Grundy (1953) which we do not consider in this paper.

Using (3.1) we can only infer to individual characteristics of population α, because the method of measurement is restricted to individual characteristics. Additional relational

information is obtained by asking each $i \in S$ to give his relations with other vertices who are member of vertex set V. When the focus is on structural aspects between the vertices in α only, we may ask each $i \in S$ to give his relations with other vertices in α. When the focus is also on relations between vertices in α and β we may ask each $i \in S$ to give his relations with other vertices in G.

In this paper we consider two well known methods of graph measurement: the partial graph and the subgraph measurement method (Frank, 1971). The partial graph measurement method is defined as: for each $i \in S$ his relations in G are observed (asked). For instance, in the Heerlen research each sampled user is asked to indicate his relations with other users in Heerlen. Consequently, the final dyad sample consists of the adjacencies Y_{ij} for $i \in S$ and $j \in S$, $i \in S$ and $j \in \alpha$, or $i \in S$ and $j \in \beta$. In this situation there is information about pairs of vertices both drawn in the sample and about pairs in which one of the vertices is drawn in the sample. The inclusion probability π_{ij} is defined as the probability that the sample contains vertex pair (i, j). This may differ from the definition usually applied in survey sampling because the observability of vertex pair (i, j) is dependent on the applied method of measurement. The inclusion probability of the partial graph method of measurement (Frank, 1971; Spreen, 1999) is

$$\begin{aligned} \pi_{ij} &= 1 - \frac{\binom{N_\alpha - n}{2}}{\binom{N_\alpha}{2}} \\ &= 1 - \frac{(N_\alpha - n - 1)(N_\alpha - n)}{(N_\alpha - 1)N_\alpha} \end{aligned} \tag{3.4}$$

i.e. one minus the probability that neither of the *2* specified vertices are in the sample.

A graph total of the form (2.7) can be estimated by

$$\hat{T} = \frac{1}{2} \sum_{i,j \in S} \frac{F_{ij}}{\pi_{ij}} + \sum_{\substack{i \in S \\ j \in \alpha \setminus S}} \frac{F_{ij}}{\pi_{ij}}. \tag{3.5}$$

Note that for $F_{ij} = F_{ij}^{(1)} = Y_{ij}$ the number of relations (2.1) or (2.8) is estimated and for $F_{ij} = F_{ij}^{(2)}$ the number of sociometric distances *2* (2.9).

The definition of the subgraph method of measurement in sample S is that each $i \in S$ gives information about other sampled vertices $j \in S$. For instance, a selected drug user is instructed to indicate his relations with other sampled drug users. Consequently the final dyad sample consists of the adjacencies Y_{ij} for $i \in S$ and $j \in S$. The subgraph inclusion probability in the final dyad sample is

$$\pi_{ij}^{B} = \frac{\binom{n}{2}}{\binom{N_\alpha}{2}} = \frac{\binom{N_\alpha - 1}{n-1}}{\binom{N_\alpha}{n}} = \frac{n(n-1)}{N_\alpha(N_\alpha - 1)} \tag{3.6}$$

Note that the inclusion probability (3.6) is the conventional second-order inclusion probability of standard survey sampling (see Särndal, Swensson & Wretman, 1992). A graph total of the form (2.7) can be estimated by

$$\hat{T} = \frac{1}{2}\sum_{i,j \in S} \frac{F_{ij}}{\pi_{ij}^{B}}. \tag{3.7}$$

Note that for $F_{ij} = F_{ij}^{(1)} = Y_{ij}$ the number of relations (2.1) or (2.8) is estimated and for $F_{ij} = F_{ij}^{(2)}$ the number of sociometric distances *2* (2.9).

To decide whether a graph parameter allows a design-unbiased HT-estimator, it is necessary to know the inclusion probabilities of the vertex pairs as well as the observation of the structural aspect (the F_{ij}-value) in the sample data. This way the F_{ij}-value can be expanded for inference purposes.

To find the variance of graph total estimators Frank (1971) noted that in a simple random sample all dyads are equally likely to be included. By defining the second-order probabilities π_{ijkl} as the probability that vertex pairs (i,j) and (k,l) will be included in the sample, it can be shown that the inclusion of two, three, or four specified vertices in the sample depend only on the vertex frequencies and not on the identities of the vertices.

Recall that in this paper $F_{ij} = Y_{ij}$, $Y_{ij} = Y_{ji}$, and $F_{ii} = 0$ by convention. For estimators of graph totals based on the subgraph method, i.e. estimator (3.7), it can be shown that the variance is (theorem 5 of Frank (1977a)

$$Var\,(\hat{T}) = \frac{(\pi_4 - \pi_2^2)}{\pi_2^2} R_\alpha^2 + \frac{(\pi_3 - \pi_4)}{\pi_2^2} Q + \frac{(\pi_2 - 2\pi_3 + \pi_4)}{\pi_2^2} R_\alpha, \tag{3.8}$$

In (3.8) $Q = \sum_{i=1}^{N_\alpha} (Y_{i+})^2$ is the sum of squares of the degrees and π_H is the inclusion probability for 4-tuple (i,j,k,l) in the final dyad sample consisting of H distinct vertices denoted

$$\pi_H = \frac{\binom{n}{H}}{\binom{N_\alpha}{H}} \tag{3.9}$$

for $H = 2,3,4$. For clarification of (3.9): π_4 is the inclusion probability of the dyads (i,j) and (k,l), π_3 is for instance the inclusion probability of dyads (i,j) and (i,l), and π_2 is for instance the inclusion probability of dyads (i,j) and (i,j).

Provided that $\pi_4 > 0$, i.e. $n \geq 4$, an unbiased variance estimator of (3.8) is obtained using the corresponding sample statistic divided by the appropriate inclusion probability π_H (Frank 1977b: theorem 4, corollary 5), i.e.

$$\hat{Var}\,(\hat{T}) = \left(\frac{1}{\pi_2^2} - \frac{1}{\pi_4}\right) R_\alpha^2(S) + \left(\frac{1}{\pi_4} - \frac{1}{\pi_3}\right) Q(S) + \left(\frac{2}{\pi_3} - \frac{1}{\pi_2} - \frac{1}{\pi_4}\right) R_\alpha(S), \tag{3.10}$$

where (S) indicates the structural information that is observed with the sample.

For estimators of graph totals based on the partial graph method, i.e. estimator (3.5), Frank (1977a: theorem 8) showed that the variance is

$$Var(\hat{T}) = \frac{(q_4 - q_2^2)}{(1-q_2)^2} R_\alpha^2 + \frac{(q_3 - q_4)}{(1-q_2)^2} Q + \frac{(q_2 - 2q_3 + q_4)}{(1-q_2)^2} R_\alpha . \tag{3.11}$$

In (3.11) 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.

$$q_H = \frac{\binom{N_\alpha - n}{H}}{\binom{N_\alpha}{H}} . \tag{3.12}$$

An unbiased variance estimator is given by

$$\begin{aligned} \hat{Var}(\hat{T}) = & \frac{(q_4 - q_2^2)}{(1-q_2)^2(1-2q_2+q_4)} R_\alpha^2(S) + \frac{(q_3 - q_4)}{(1-q_2)^2(1-2q_2+q_4)} Q(S) \\ & + \frac{(q_2 - 2q_3 + q_4)}{(1-q_2)^2(1-2q_2+q_4)} R_\alpha(S) \end{aligned} \tag{3.13}$$

An alternative is to approximate (3.9) by $p^H = \left(\frac{n}{N_\alpha}\right)^H$. For instance (3.8) is then simplified to

$$Var(\hat{T}) \approx \frac{q}{p} Q + \frac{q^2}{p^2} R_\alpha \tag{3.14}$$

and (3.10) to

$$\hat{Var}(\hat{T}) \approx \frac{q}{p^4} Q(S) - \frac{q^2}{p^4} R_\alpha(S) , \tag{3.15}$$

where $q = 1 - p$ (Frank, 1977a, corollary 7). Capobianco and Frank (1982) note that the same variance expressions would be obtained for a Bernoulli sampling scheme with selection

probability p with estimator $\hat{T} = \frac{1}{2}\sum_{i,j \in S} \frac{Y_{ij}}{p^2}$. They argue that (3.14) and (3.15) are usually sufficient for the degree of approximation needed in practice for sample surveys.

A remark is in order here. Variance estimators (3.10) and (3.13) may have negative values (Frank, 1971; Karlberg, 1997). This holds for most HT variance estimators, but also for most Yates-Grundy variance estimators. In a simulation study of several triad count estimators Karlberg (1997) showed that the obtained Yates-Grundy variance estimators more frequently had negative estimates than HT variance estimators for some observation schemes. He also remarked that the more structural information is collected in the sample, the less frequent are negative variance estimates. In (3.10) and (3.13) there are some negative terms (note that $\pi_4 < \pi_2^2$, and $\frac{2}{\pi_3} < \frac{1}{\pi_2} + \frac{1}{\pi_4}$ may occur). Especially when the amount of collected structural information is small, (3.10) and (3.13) are sensitive for very minor differences (see for instance example 4.3.4 (p.94) in Frank (1971)).

4. Illustration of a structural analysis

To illustrate the use of graph total estimators in population studies of hard drug users, relational data obtained from the Heerlen Drug Monitoring System (DMS) is analysed. 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.

In this paper we analyse data from the network sample. We focus on the estimation of the number of contacts between clients of the aid agencies and between clients and hard drug users that are not registered (hereafter called nonclients). From a health promotion perspective these figures 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 nonclients. The distribution of contacts between clients and nonclients of the aid agencies are analysed using the introduced graph total concepts. Note that a relation between clients or between clients and nonclients is viewed in this section 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. $\alpha = \{1,2...,435\}$, as an initial sampling frame. A simple random sample without replacement S of size $n = 38$ 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. In other words the partial graph method was applied. The criteria for the alters to be included in the sample were:

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

This way not only individual and relational information about clients of the aid agencies was collected but also about other users directly connected to the clients.

Using graph total estimator (3.5) the estimated total of relationships between the clients of the aid agencies was estimated to be about 975 with a standard error of 240. The average number of contacts an arbitrary client had with other clients was estimated to be

$$\hat{\mu}_\alpha = \frac{2}{N_\alpha}\hat{R}_\alpha = 4.48 \ (s.e.\ 1.1) .$$

To estimate the number of relations between the clients and the nonclients, estimator (3.1) and variance estimator (3.3) were applied, resulting in an estimate of

$$\hat{R}_{\alpha\beta} = \frac{\sum_{i\in S}\sum_{j\in\beta} Y_{ij}}{\pi_i} = 1053 \ (s.e.\ 138).$$

The average number of relations an arbitrary client had with hard drug users not registered by any aid agency is estimated as

$$\hat{\mu}_{\alpha\beta} = \frac{1}{N_\alpha}\hat{R}_{\alpha\beta} = 2.42 \ (s.e.\ 0.32).$$

The implication of this structural analysis is that an arbitrary client of the aid agencies tend to have relatively more contacts (65%) with other clients than with nonclients (35%). Consequently, if $\hat{\mu}_\alpha$ and $\hat{\mu}_{\alpha\beta}$ are understood as the mean numbers of communication channels per client with other hard drug users, we could assess that the diffusion of a client-based community prevention/ intervention strategy tend to reach for a substantial part other clients.

The number of pairs of indirectly connected clients may also be relevant for a client-based community prevention/intervention strategy because two clients who are not directly related to each other, may have a mutual friend. If this mutual friend happens to be another client, the diffusion of prevention/intervention strategies will be restricted to the client population. If this mutual friend happens to be a nonclient, the diffusion of prevention/intervention strategies will have a chance to reach also the nonclient population. In other words, the more indirect contacts pairs of clients have via nonclients, the more likely a client-based community prevention/intervention strategy would reach a substantial part of the non-registered hard drug users.

To estimate the number of pairs of indirectly connected clients, we use graph total estimator (3.7). To observe sociometric distance *2* between a pair of clients the following conditions must be satisfied:

- a direct relation between clients $i \in \alpha$ and $j \in \alpha$ does not exist, i.e., $Y_{ij} = 0$;
- clients $i \in \alpha$ and $j \in \alpha$ have at least one mutual hard drug user $k \in V$ directly connected to both of them in G, i.e. $max_{k \in V} Y_{ik} Y_{jk} = 1$.

To decide which measurement method we may use, consider Figure 2 which illustrates two hypothetical population networks of order *4*. In both networks a simple random sample of order $n = 3$ is drawn resulting in $\{i, j, h\} \in S$ and $\{k\} \notin S$.

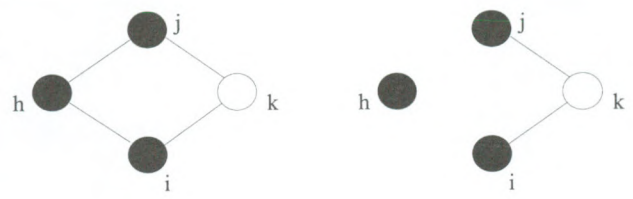

Figure 2 *Two hypothetical situations for* $\{i, j, h\} \in S$ *and* $\{k\} \notin S$

We focus on vertex pair (i, j), and denote the observed distance between i and j in the sample by $d^*(i, j)$. In both graphs the actual sociometric distance between client pair (i, j) is $d(i, j) = 2$. Suppose the subgraph measurement method is applied. The sampled structural information leads in graph A to $d^*(i, j) = 2$ but suggests $d^*(i, j) = \infty$ in graph B. This implies that for the subgraph measurement method a sociometric distance *2* (the true F_{ij}-value) cannot always be ascertained from the sample data. Thus for the subgraph measurement method an observed within-sample distance greater than *2* can hide a real distance equal to *2*. For the partial graph measurement method, however, for each pair of clients in the final dyad sample the actual sociometric distance is always observed in both population graphs, i.e. $d^*(i, j) = 2$ for all sampled pairs with $d(i, j) = 2$.

In formula (2.9) the sociometric distance *2* is regarded as a special case of a graph total. Applying (3.7) an estimator for the total number of indirectly connected client pairs is

$$\hat{T}_2 = \left(\frac{1}{2} \sum_{i,j \in S} \frac{F_{ij}^{(2)}}{\pi_{ij}^{B}} \right) = 10741 \ (s.e.\ 3354)$$

where

$$F_{ij}^{(2)} = \begin{cases} 1 \ if \ d(i,j) = 2 \ for \ i \in S \ and \ j \in S \\ 0 \ otherwise \end{cases}.$$

Note that π_{ij}^{B} is defined in (3.6), i.e. the probability that clients i and j are both drawn in the sample.

The proportion of indirectly connected clients is estimated by

$$\frac{2\hat{T}_2}{N(N-1)} = 0.11 \ (s.e.\ 0.04),$$

and can be interpreted as the proportion of pairs of hard drug users not directly connected to each other but who have at least one mutual contact. To assess the effectness of the diffusion of an intervention starting from the client group it is relevant to know whether these mutual contacts are other clients, nonclients, or other clients as well as nonclients. In Figure 3a,b,c the different types are graphically displayed:

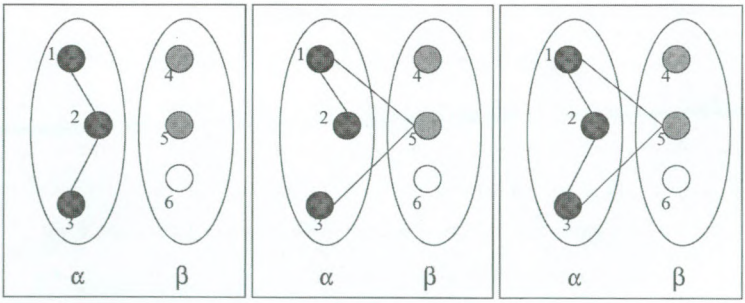

Figure 3a **Figure 3b** **Figure 3c**

Figure 3. Three different types of mutual contacts of pair {1,3}

In Figure 3a *d(1,3) =2* via vertex $\{2 \in \alpha\}$, in Fig. 3b *d(1,3)=2* via vertex $\{5 \in \beta\}$, and in Fig. 3c *d(1,3) =2* via vertices $\{2 \in \alpha\}$ and $\{5 \in \beta\}$.

From the estimated total of indirect connected clients *87,5%* have mutual contacts only with other clients (Fig.3a), about *7,5%* only with nonclients (Fig.3b), and about *5%* have mutual contacts with other clients and nonclients. This implies that a client-based community prevention/intervention strategy in Heerlen will have the tendency to reach mainly other clients.

5. Concluding remarks

The results of the structural analysis of the hard drug users population in Heerlen showed that applying a network measurement method to a random sample of clients of the aid agencies provides valuable structural information. This information was not restricted to the clients only but also to hard drug users who are not clients of the aid agencies, and directly connected to clients. Using the concept of graph totals we got an indication that clients on the average have the tendency to be linked with other clients. For community-based interventions this is important information because it shows that one must do an extra effort to reach those hard drug users that are not clients.

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 drugscene van Parkstad Limburg. (The process of marginalisation and hardening in the Heerlen drugscene). 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.

Horvitz, D.G. & D.J. Thompson (1952).A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

Karlberg, M (1997*). Triad count estimation and transitivity testing in graphs and digraphs.* Doctoral dissertation, Department of Statistics, Stockholm University, *Akademitryck AB, Edsbruk.*

Klovdahl, 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 repesentatives 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. (1999) Sampling personal network structures. Statistical inference in ego-graphs. *ICS disseration series.* Amsterdam: Thesis Publishers.

Thompson, S.K. (1991) Stratified adaptive cluster sampling, *Biometrika*,78, 2, 389-397.

Watters, J.K. & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.

Yates, F. & P. M. Grundy (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B 15, 253-261.