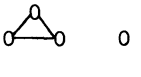
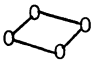
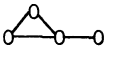
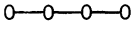
THE VARIANCE OF AN ESTIMATE OF LINKAGE DENSITY FROM A SIMPLE RANDOM SAMPLE OF GRAPH NODES

C. H. Proctor, North Carolina State University

In order to estimate the density of linkages in a finite graph one may select a simple random sample of nodes and determine for each pair of nodes selected whether or not there is a link between them. The sample proportion of linked pairs is an unbiassed estimate of the population proportion or linkage density. The variance of this estimate can be expressed as a function of certain graph moments and an unbiassed estimate of the variance can be found. The purpose of this paper is to describe the variance formulae.

The results were worked out to aid in interpreting some data on an acquaintanceship network in Wake Forest, North Carolina. The research was supported by the Institute of Statistics, Raleigh Section at North Carolina State University and is described rather informally in a mimeographed paper [4]. For such data the nodes are called actors and the graph is referred to as a social network [1]. This somewhat sociologically specialized terminology will be retained in favor of the more general graph theory one. It should be noted that only one link may join a pair of actors and it is supposed that both actors must be interviewed in order to determine whether a link is present.

In the following discussion ℓ's will denote sample proportions and λ's population proportions; n is written for the sampled number of actors and N is the population number. The quantity ℓ_{21} is the proportion of linked actor-pairs in the sample. The quantity ℓ_{32} is the sample proportion of actor-triples with two links. The quantity λ_{41} is the population proportion of actor-quadruples with one link. That the first subscript refers to the number of actors and the second to links can be inferred from these examples. When dealing with actor-quadruples containing two, three and four links a third, alphabetic, subscript is added to distinguish the cases as follows:

Proportion	Structure	Proportion	Structure
λ_{42a}	0-0-0 0	λ_{43c}	
λ_{42b}	0-0 0-0	λ_{44a}	
λ_{43a}		λ_{44b}	
λ_{43b}			

The expressions for the variance of ℓ_{21} and an unbiassed estimate of this variance turn out as follows:

$$V(\ell_{21}) = E(\ell_{21}^2) - [E(\ell_{21})]^2$$

$$= \{\lambda_{21}/\binom{n}{2} + 2(n-2)(\lambda_{32} + 3\lambda_{33})/3\binom{n}{2} + (n-2)(n-3)[\lambda_{42b} + \lambda_{43b} + \lambda_{44b} + 2(\lambda_{44a} + \lambda_{45}) + 3\lambda_{46}]/6\binom{n}{2}\} - \lambda_{21}^2 \quad (1)$$

$$= \lambda_{21}\left[\frac{1}{\binom{n}{2}} - \frac{1}{\binom{N}{2}}\right] + \frac{2}{3}\{(\lambda_{32} + 3\lambda_{33})\left[\frac{n-2}{\binom{n}{2}} - \frac{N-2}{\binom{N}{2}}\right]\} + \frac{1}{6}[\lambda_{42b} + \lambda_{43b} + \lambda_{44b} + 2(\lambda_{44a} + \lambda_{45}) + 3\lambda_{46}]\left[\frac{(n-2)(n-3)}{\binom{n}{2}} - \frac{(N-2)(N-3)}{\binom{N}{2}}\right]. \quad (2)$$

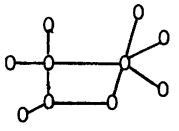
The expression within curly brackets in (1) was found by first writing ℓ_{21}^2 as $[2\sum_{t>u}^{n} a_{tu}/n(n-1)]^2$ in which a_{tu} equals one or zero according as to whether or not the t^{th} drawn and u^{th} drawn actors are linked. Then the square of the summation was expanded and terms of three kinds were collected. The types were: a_{tu}^2, $a_{tu}a_{tu'}$, and $a_{tu}a_{t'u'}$, where a prime denotes a subscript unequal to the unprimed one. The expected value was then taken using the facts that $E(a_{tu}^2) = \lambda_{21}$, $E(a_{tu}a_{tu'}) = \lambda_{32}/3 + \lambda_{33}$, and $E(a_{tu}a_{t'u'}) = (\lambda_{42b} + \lambda_{43b} + \lambda_{44b})/3 + 2(\lambda_{44a} + \lambda_{45})/3 + \lambda_{46}$, while the numbers of the three kinds of terms are $n(n-1)/2$, $n(n-1)(n-2)$, and $n(n-1)(n-2)(n-3)/4$ respectively.

In order to get the final form of $V(\ell_{21})$ in (2) the quantity λ_{21}^2 was written $[2\sum_{i>j}^{N} A_{ij}/N(N-1)]^2$ where A_{ij} equals 1 (0) if actor i is linked (not linked) to j and expanded as was done for ℓ_{21}^2. In this case the i and j subscripts refer to population identification numbers and the A_{ij}'s are not random variables.

An estimate of $V(\ell_{21})$ may be calculated using the corresponding ℓ quantities in place of the λ quantities in (2). This quantity will be denoted $v(\ell_{21})$. The fact that $E(\ell_{ks}) = \lambda_{ks}$ for

any number of actors k and any structural subscript s insures that $E(v(\ell_{21})) = V(\ell_{21})$. To prove that $E(\ell_{ks})$ equals λ_{ks} one first writes ℓ_{ks} as the sum of $\binom{n}{k}$ indicator-of-structure-s variables divided by $\binom{n}{k}$ and then notes that the expected value of each and every indicator variable is λ_{ks}. This is an "argument of symmetry" [3]. Other properties of the estimate $V(\ell_{21})$ are not yet known but the study of higher moments of the distributions of both ℓ_{21} and $v(\ell_{21})$ will undoubtedly be greatly facilitated by the work of D. E. Barton and F. N. David on graph moments [2].

A small scale numerical example may help to illustrate the computation of $v(\ell_{21})$. A questionnaire was sent to a simple random sample of 20 names from the about 2,000 names in the North Carolina State University Staff Directory in 1964, and pairs of persons were said to be linked if each reported they had "spoken" to the other. The sociogram of linkages (note the 10 isolates) was as follows:



0 0 0 0
0 0 0
0 0 0

From these data one can calculate the following:

$\ell_{21} = 10/190 = .052632$

$\ell_{32} = 20/1140 = .017544$

$\ell_{33} = 0$

$\ell_{42b} = 3/4845 = .00061920$

$\ell_{43b} = 20/4845 = .00412797$

$\ell_{44b} = 0$

$\ell_{44a} = 1/4845 = .00020640$

$\ell_{45} = 0$, $\lambda_{46} = 0$

Thus $v(\ell_{21}) = (.052632)(.0052627)$
$+ .0116960(.0937373)$
$+ (.000859995)(-.385475)$
$= .00104.$

The estimate of density thus suffers a estimated coefficient of variation of 61%. Since the effect on the variance of an increase in sample size is roughly inversely proportional to $\binom{n}{2}$, it follows that in order to reduce the coefficient of variation to 6% would require an increase in sample size from 20 to over 200. For n = 200 the estimated coefficient of variation is still about 9%.

If a sample of 190 pairs of names, involving 380 or somewhat fewer persons, had been drawn as a simple random sample of the 1,999,000 pairs in the population the variance of the estimated density could be estimated as

$pq/n = (.052632)(.947368)/190 = .00026$

In so far as the simple random sample of actors (actor-SRS) also contains data on 190 pairs there appears to be a loss in precision for this population when using the actor-SRS rather than what may be called a pair-SRS of the same number of pairs. Of course, the cost of making the 190 observations would normally be greatly increased in the pair-SRS method over the actor-SRS.

[1] Chapters by Eric R. Wolf, J. Clyde Mitchell and Adrian C. Mayer in Banton, M. (ed.) The Social Anthropology of Complex Societies, A.S.A. Monographs 4, London: Travistock Publications; Frederick A. Praeger, Publishers, New York, 1965.

[2] Barton, D. E. and F. N. David, "The Random Intersection of Two Graphs," F. N. David (ed.), Research Papers in Statistics, John Wiley and Sons, Inc., New York, 1966, pp. 445-459.

[3] Cochran, W. G., Sampling Techniques, second edition, John Wiley and Sons, Inc., New York, 1963, page 22.

[4] Proctor, C. H., "Two Measurement and Sampling Methods for Studying Social Networks," Institute of Statistics Project Report - 1966, N. C. State University, Raleigh, North Carolina.