CONNECTIONS 22(2): 161-70

Non-Parametric Standard Errors and Tests for Network Statistics

Tom A.B. Snijders
ICS, University of Groningen[1]

Stephen P. Borgatti
Carroll School of Management, Boston College

Two procedures are proposed for calculating standard errors for network statistics. Both are based on resampling of vertices: the first follows the bootstrap approach, the second the jackknife approach. In addition, we demonstrate how to use these estimated standard errors to compare statistics using an approximate t-test and how statistics can also be compared by another bootstrap approach that is not based on approximate normality.

In social network analysis, we are used to calculating descriptive statistics for networks, but not so used to accompanying these statistics with standard errors. Yet the general arguments for the benefits of standard errors do apply to social network analysis: it is useful to have an indication of how precise a given description is, particularly when making comparisons between groups. The problem is that there are no established, widely applicable, ways of calculating standard errors for network statistics. Our objective in this paper is to develop some procedures for doing so.

In the general (non-network) case, there are, roughly speaking, two approaches to calculating standard errors. The first is to take some descriptive statistic as the point of departure and find a way to calculate a standard error that requires a minimum of assumptions – the simplest example is the commonly calculated standard error of the mean of a simple random sample, which is based only on the assumption that the sample is simple random. The second approach is to formulate some statistical model for the observations, estimate the parameters of this model, and calculate the standard error of these parameter estimates. This paper presents an elaboration of the first of these two approaches for the case of network data. We assume that a researcher is interested in some descriptive statistic – the density of the network, an index for transitivity, network centralization, or any other network property – and wishes to have a standard error for this descriptive statistic without making implausibly strong assumptions about how the network came about.

Two general-purpose non-parametric methods have been proposed in the statistical literature to construct standard errors for complicated or poorly understood statistics: the jackknife (Tukey, 1958)

[1]Address mail to Tom A.B. Snijders, Department of Statistics and Measurement Theory, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. Email: T.A.B.Snijders@ppsw.rug.nl.

and the bootstrap (Efron, 1979). Both methods are based on resampling, i.e., constructing many artificial data sets out of the observed data set, and using the variability between these artificial data sets. These methods have been theoretically elaborated and have led to a number of applications. A number of useful reviews exist in the literature, including ones by LePage and Billard (1992) and Shao and Tu (1995). A user-friendly DOS program for bootstrap and jackknife analysis, *BOJA*, is available (Boomsma, 1991).

In the literature, these methods are developed for the usual random samples with rectangular data matrices but not for network data with their typical square data matrices and empty diagonal. This note proposes extensions of these methods to the network situation. It is assumed, at least initially, that a network data set *Y* with N nodes is available, there is interest in a statistic *Z* that is calculated from *Y*, and we would like to have a standard error for *Z*. All of our examples are based on network density, but it is important to remember that the procedures are very general, and could be used with virtually any network statistic, including centralization measures, number of cliques, fit to a core/periphery model, and so on.

THE JACKKNIFE

The basic idea of the jackknife is that, given a dataset of N sample elements, N artificial datasets are created by deleting each sample element in turn from the observed dataset. The new datasets are quite similar, but the variability among them does give an indication of the variability that may be expected between independent replicates of the data set.[2] The jackknife standard error for statistics *Z* defined for rectangular data matrices corresponding to simple random samples is defined by

$$\sqrt{\frac{N-1}{N}\sum_{i=1}^{N}(Z_{-i}-Z_{-\_})^2}$$

where Z_{-i} is the statistic obtained for the data set from which case *i* is deleted, and $Z_{-\cdot}$ is the average of Z_{-1}, ..., Z_{-N} . The fact that the sum of squares is multiplied by a factor close to 1, instead of being divided by *N*-1, reflects the fact that these *N* artificial data sets are much more similar than would be *N* independent replicates.

The jackknife principle was studied by Frank and Snijders (1994) for network statistics. They found that the multiplication factor (*N*-1)/*N* is not adequate for network statistics. The reason is that this multiplication factor is based on the property, valid for most statistics based on simple random samples, that their variance is inversely proportional to the sample size (or approximately so). This is not so for network statistics. The number of relevant elements of an $N \times N$ adjacency matrix with no reflexive ties is $N(N-1)$, and the variance of *Z* will more likely be inversely proportional to $N(N-1)$ than to *N*. Accordingly, Frank and Snijders (1994) proposed for network statistics the jackknife variance estimated defined by

[2] There are generalizations based on deleting more than one sample element, but these are not considered in this note.

$$s.e._J(Z) = \sqrt{\frac{N-2}{2N}\sum_{i=1}^{N}(Z_{-i} - Z_{-\_})^2}$$

where Z_{-i} is the network statistic obtained for the data set from which vertex i is deleted (so that a network on N-1 vertices remains), and $Z_{-\cdot}$ is again the average of Z_{-1}, ..., Z_{-N} . For the statistics studied by Frank and Snijders (estimates for the number of vertices in an unknown graph, based on a snowball sample), the jackknife estimate of standard error performed quite well.

THE BOOTSTRAP

The basic idea of the bootstrap is that the observed data are treated as a population in itself, and that artificial samples of size N are drawn with replacement from the observed data. Thus, each artificial sample will contain multiple copies of some elements of the observed data, whereas other observed data points will be missing from the artificial sample. For network data, the obvious analogy is to draw a sample with replacement from the *vertices*.

To specify this more explicitly, suppose that the data consist of a network on N vertices denoted i = 1, ..., N, where the tie between vertices i and j is denoted Y_{ij} . (The network could be a graph, a directed graph, or a graph with valued edges.) A large number M of bootstrap samples is to be drawn. Each single bootstrap sample is drawn in the following way. A random sample with replacement is drawn from the vertices, and denoted $i(1)$, ..., $i(N)$. This means that all these $i(k)$ are independent draws from the numbers 1, ..., N. The artificial network Y^* is the network induced by these vertices $i(1)$ to $i(N)$. If vertices k and h in the artificial network correspond to different original vertices $i(k)$ and $i(h)$, this means simply that

$$Y^*_{kh} = Y_{i(k)i(h)}, \ \textit{for } i(k) \neq i(h)$$

i.e., in the artificial network the tie between vertices k and h is the same as the tie between vertices $i(k)$ and $i(h)$ in the observed network.

It is not obvious what to do for the ties between those artificial vertices that correspond to the same real vertex, at least in networks where reflexive ties are not defined. The idea of resampling vertices is that the procedure is meant to leave the basic network structure intact, and any scheme for filling in ties between artificial vertices corresponding to the same real vertex runs the risk of mixing up this structure. (One source of comfort is that, as the number of vertices grows larger, the expected fraction of such doubtfully determined ties will get closer to 0.) As an expedient solution, we propose a dyad-based bootstrap for these ties. This means that the values for the dyads defined by artificial vertices corresponding to the same real vertex are chosen, with replacement, from the set of all $N(N-1)/2$ dyads. The order of the two elements within the dyad is also determined randomly.

The bootstrap standard error is then determined as follows. The described procedure of generating an artificial network is repeated M times independently, where M is large, e.g., 1,000. For each artificial network drawn in this way, the statistic of interest is calculated. Denote these artificial statistics by $Z^{*(1)}$ to $Z^{*(M)}$. This means that $Z^{*(m)}$ is calculated on the basis of the m'th artificially generated network Y^*. The artificial networks are regarded as networks that might have been observed instead of the actually observed one, so that $Z^{*(1)}$ to $Z^{*(M)}$ is regarded as a synthetic sample from the distribution of Z. Accordingly, the bootstrap standard error is

$$s.e._B(Z) = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(Z^{*(m)} - Z^{*(\cdot)})^2}$$

where $Z^{*(\cdot)}$ is the mean of the $Z^{*(m)}$.

The main assumption of this bootstrap standard error is that it makes sense to regard the vertices as interchangeable, since the observed vertices are indeed treated interchangeably in the sampling process. Thus, for a network composed of a class of school children this might be more reasonable than for a network existing within a hierarchically structured organization (given that the hierarchy matters for the observed network).

EVALUATING A NETWORK STATISTIC

In this section we consider the problem of comparing an observed network statistic Z with a theoretical value μ. For example, consider a social relation that serves as a conduit for the transmission of a virus.[3] The greater the density of the network, the faster and more certainly the virus is spread. Based on theoretical considerations, we might postulate the existence of a threshold value of network density below which the infection cannot sustain itself, and above which there is danger of developing an epidemic.[4] For any given network, the question is whether the density of ties falls within the safe range.

A standard approach in this situation is to define a null hypothesis stating that the network density is less than the parameter μ, and reject the hypothesis if the observed statistic is sufficiently larger than the parameter, relative to the standard error of the sampling distribution. Hence, we calculate

$$t = \frac{Z - m}{s}$$

and reject the null if t is larger than 1.65, which is the critical value associated with a maximum Type 1 error of 0.05 in a 1-tailed test.

As an empirical example, consider the network of friendship ties among 67 prison inmates collected by Gagnon in the 1950s, reported by MacRae (1960), and available as part of the UCINET 5 software package (Borgatti, Everett and Freeman, 1999). Let us assume that the theoretical "tipping point" that separates epidemic from extinction occurs at density 3%. The observed density for this network is 0.0412. The standard error, as estimated by the bootstrap method with 1,000 samples, is 0.0060 (it is 0.0036 using the jackknife method). Converting to standard error units, we obtain (0.0412 - 0.03)/0.0060 = 1.87. This value is larger than 1.65 and in fact corresponds to a 1-tailed significance level of 0.03. Therefore, we reject the null hypothesis and provisionally conclude that the population is in danger.

It should be noted that in this approach we have assumed that the shape of the sampling distribution is approximately normal, and therefore use the bootstrap sampling distribution only to estimate the variance of this distribution. An alternative approach that does not make this assumption is to use the bootstrap sampling distribution directly to calculate the probability of obtaining an observed density as large as actually observed given the null hypothesis. Hence we would like to simply count the

[3] Equally, we can consider the transmission of a rumor or the adoption of an innovation.

[4] For example, the R0 parameter described by Anderson (1982).

proportion of bootstrap samples that have a test statistic larger than the observed. However, the bootstrap distribution is centered on (or near) the observed statistic, Z, rather than the theoretical parameter. This is because the bootstrap samples from the data rather than from the null distribution. Therefore, as Noreen (1989) suggests, we need to subtract the mean of the bootstrap sampling distribution ($Z^{*(.)}$) from each value of $Z^{*(m)}$ and then add back the theoretical value. We then count the proportion of $Z^{*(m)}$ values that are larger than the observed Z. In effect, we assume the shape of the bootstrap distribution is correct, but simply mis-centered, and we use the center that corresponds to our null hypothesis.

In our example, the mean bootstrap density was 0.0406. We therefore subtracted 0.0406 from each bootstrap density, and added 0.03. We then counted the number of samples in which $Z^{*(m)}$-0.0406+0.03 was greater than or equal to the observed value of 0.0412. Adding 1 to this count and dividing by M+1 gives an estimate of the proportion of samples (including the observed) which would equal or exceed the observed value – in short, the significance level. In this case, we obtained a significance of 0.038, which agrees well with our previous estimate.

Table 1
Comparison of Approaches for the One Sample Case

	Classical Estimate ($s/\sqrt{n}$)	Bootstrap-Assisted SE*	Bootstrap Direct Method*
SE	0.0030	0.0060	NA
T-Statistic	3.73	1.87	NA
1-Tailed Significance	< 0.001	0.031	0.038

* Using 5,000 bootstrap samples

Table 1 compares the significance levels obtained via the two bootstrap methods, and the (inappropriate) classical approach, which estimates the standard error of the sampling distribution from standard deviation of the sample variable. Note that the two bootstrap methods agree closely, while the classical method, whose assumptions are violated by network data, yields very different values.

COMPARING TWO NETWORKS

Another important application area is the comparison of a network statistics for two different groups. For example, Ziegler et al (1985) report the corporate interlocks among the major German business entities (15 in total). Stokman et al (1985) report interlocks among the major Dutch business entities (16 in total). The data are available as part of the UCINET 5 (Borgatti, Everett & Freeman, 1999) software package. The question we pose is this: is the level of interlock (i.e., the density of ties) different in the two countries?

The observed density of the Dutch network was 0.5, while the density of the German network was 0.6381, for an observed difference of 0.1381.To test the significance of this difference, we can construct a bootstrap or jackknife-based t-test. Assuming a null hypothesis of no difference, the

standard approach[5] is to calculate

$$t = \frac{Z_1 - Z_2}{\sqrt{SE_1^2 + SE_2^2}}$$

where *SE1* and *SE2* are standard errors usually estimated from the standard deviation of the measured variable in each sample. In the network case, we substitute the jackknife or bootstrap-derived standard errors as outlined above. Selecting the bootstrap as our method of choice, the standard errors for the Dutch and German networks are 0.0902 and 0.1083 respectively. The t-statistic works out to

$$t = \frac{0.5 - 0.6381}{\sqrt{0.0902^2 + 0.1083^2}} = \frac{-0.1381}{0.140943} = 0.9798$$

which is clearly not significant. Thus, we cannot conclude that the Dutch and German economies have developed different levels of corporate interlock.

PAIRED SAMPLES

Next we consider the case of the same network observed at two points in time. For example, Jean Bartunek[6] collected work relationships at two different times among faculty and staff at a school that included elementary, middle, and high school sub-units. The school had had a history of autonomous action within these levels and inadequate coordination among them. Time 1 was at the beginning of the school year, just before a new staff position was implemented with the explicit purpose of increasing coordination, and therefore work relationships, among the faculty staff. The data at Time 2 were collected at the end of the school year, 9 months after the position was created and a person hired to fill it. One of the questions posed by the school was whether the new person was successful in increasing coordination among the administration and faculty (in network terms, whether work ties were increased). In short, we would want to know whether the density at Time 2 is significantly greater than the density at Time 1.

This situation is different from the last one in that the two samples here are not independent: instead, we have two sets of measurements of the same relation on the same set of players, and the relationship between a pair of players at Time 1 is unlikely to be independent of their relationship at Time 0. Therefore, to conduct a t-test, we must construct a different approach, analogous to the classical paired-sample t-test, for estimating the standard error of the difference.

Using the bootstrap, we propose two approaches to the paired sample case, just as we did with the independent samples case. The first is as follows. Given the set of *N* actors in the observed network(s), a random sample of size *N* is drawn with replacement. For this artificial set of actors, two separate networks, one for each time period, are then constructed using the procedures outlined earlier. The density of each is computed, and the difference between them recorded. This is repeated *M* times, and the S.E. of the difference (SE_d) is computed as in Equation 4, where Z refers to the

[5] For independent samples and unknown population variances.

[6] The data are described in Stevenson and Bartunek (1996).

difference in density. This is then used to calculate a t-statistic as shown in Equation 7.

$$t = \frac{Z_1 - Z_2}{s.e._B} \quad \textbf{(7)}$$

This approach is convenient but as noted before assumes that the shape of the sampling distribution is approximately normal. The second approach dispenses with this assumption by directly counting the proportion of bootstrap samples that yield a difference as extreme as actually observed. Once again, however, the bootstrap distribution is (asymptotically) centered on the observed difference rather than the theoretical expectation (usually, zero). Therefore, we subtract the mean of the bootstrap distribution from each bootstrap difference score, and then count the proportion of mean-centered bootstrap differences are as large as the difference actually observed.

As an empirical example[7], we use Kapferer's (1972) tailor shop data. He recorded "sociational" (friendship, emotional) data on 39 members of a tailor shop in Zambia at two points in time. After the first set of observations there was a failed strike attempt. After the second set, there was a successful strike. A theory of collective movements might suggest that strikes cannot be successfully organized unless members of the group are well-enough connected to enable a single view to emerge (rather than a multiplicity of views held in disparate corners of the network). The hypothesis one would want to test then is that the density at Time 2 is higher than at Time 1.

Table 2.
Bootstrap-Assisted Paired Sample T-Test

	Time 2	Time 1	Difference
Density	0.3009	0.2132	0.0877
Bootstrap SE (5000 samples)	0.0306	0.0271	0.0245
T-Statistic			3.5773
Significance			< 0.001

As shown in Table 2, the observed density at Time 2 is clearly higher than at Time 1, and the difference is significant when we run a paired sample t-test using the bootstrap-derived *SE* for the difference. The null hypothesis is rejected, and the research hypothesis is supported.[8]

[7] We abandon the Bartunek dataset because, on examining the data, we found that the density at Time 2 actually went down slightly, obviating the need for a one-tailed test that density increased.

[8] Of course, the results support a number of theories, including the one that states that successful strikes have the effect of creating cohesion among the workers.

As an aside, it is interesting to note that, had we assumed independent samples, the t-statistic would have been 2.15, which is much smaller than the 3.58 we obtain with the paired samples formula, though still significant.

Table 3.
Direct Bootstrap Approach to Comparing
Difference in Densities for Two Measurements
On the Same Actors

	Difference in Density
Observed	0.0877
Avg. of bootstrap distribution	0.0841
Prop. of bootstrap samples with mean-adjusted difference as large as observed	0.0004

Table 3 gives the results of using the direct bootstrap approach. As shown in the table, the average density difference in the bootstrap sampling distribution was 0.0841. Subtracting this quantity from each bootstrap density and counting the proportion of samples with mean-centered differences greater than the observed difference of 0.0877, we obtain a significance value of 0.0004, which agrees well with the t-test computation.

DISCUSSION

We have proposed non-parametric standard errors, based on an extension of the bootstrap and jackknife principles to network data. In addition we have examined direct methods of evaluating specific hypotheses that are also based on the resampling principle, but which do not assume normality of the artificial sampling distribution. All of these techniques are computer-intensive but, in principle, easy to apply.

Standard errors and statistical tests are inevitably based on considerations that the data – in our case, the network – "could have been different". These differences could occur because of observation errors, unreliability of measurement, the contingent – or probabilistic – nature of the processes that gave rise to the observed relations, sampling of vertices, choice of the observation moment (i.e., sampling in time), or whatever. Even in the study of entire networks, such considerations often are realistic. You cannot have a statistical test without the assumption that the data could have been different; the question is, which differences would have been likely?

The main basis for both the bootstrap and jackknife approaches is the assumption that the vertices are interchangeable. Expressing this more intuitively, it is assumed that for different vertices *i* and *j*, the *i*th row and column of the adjacency matrix are "just as good" as the *j*th row and column, and that the

network in which the *i*th row and column was replaced by the *j*th "could also have been observed" – ignoring, for the moment, the (*i, j*) and (*j, i*) elements. In still other words, the essential structure of the network would not be changed to an important extent by this replacement. If it is reasonable to take this viewpoint, then the bootstrap standard error and the bootstrap tests are reasonable also.

For the jackknife standard error, there is an additional assumption: the variance of the statistic is assumed to be approximately inversely proportional to $N(N-1)$, where N is the number of vertices. This assumption is not always well-founded, and it can be checked only if the distribution of the network is known, i.e., in theoretical cases. It seems more realistic to assume that, in general, the variance of network statistics is inversely proportional to something between N and $N(N-1)$. This makes the bootstrap standard error more reliable than the jackknife standard error, except in those cases where evidence for the reliability of the jackknife standard error has specifically been established.

As such, the proposed methods bear some relation to the permutation technique for testing relations between networks, also known as QAP ("Quadratic Assignment Procedure") correlation, proposed by Hubert and Baker (1978) and elaborated and popularized, e.g., by Krackhardt (1988). In contrast to the bootstrap and jackknife procedures, this permutation technique requires that two (or more) networks with the same vertex set are available, and a null hypothesis about their statistical independence (possibly partialling out other variables) is being tested. This null hypothesis is understood as follows: the identity of the N vertices is immaterial for the relation between the two networks, implying that a data set with permuted vertices could just as well have been observed. Similar to the bootstrap and jackknife procedures, the permutation technique is based on artificial data sets: in the simplest case where conformity of two adjacency matrices is tested, one matrix is left as it is while in the other the rows and columns are permuted correspondingly. The QAP technique provides the distribution and standard errors under the null hypothesis of independence. Thus, it is non-parametric because it does not make assumptions about the probability distribution of the networks considered separately, but it is restricted to the null hypothesis that the two networks are statistically independent.

It should be noted that there exist some simple (and usually unrealistic) null models under which the standard errors of certain statistics have been calculated. For distributions that imply exchangeable vertices (in other words, permutation invariance), it makes sense to compare the standard errors derived under such a distribution to the bootstrap and jackknife standard errors. Examples are the $U|L$ distribution, where the graph or digraph is random under the condition of a fixed number of arcs; and the $U|M,A,N$ distribution, where the dyad count of a digraph is supposed to be given but for the rest the digraph is random (see Wasserman and Faust, 1994, Chapter 13). For some network statistics, standard errors under such distributions have been calculated. E.g., Holland and Leinhardt (1975) give the standard sampling variance of arbitrary linear combinations of the triad census under the $U|M,A,N$ distribution and Snijders (1981) gives the sampling variance of the degree variance under the $U|L$ distribution. For the majority of graphs generated under such a distribution with exchangeable vertices, it may be expected that the bootstrap and jackknife standard errors are of the same order of magnitude as the standard errors calculated by such formulae, because all three are appropriate in such cases. However, for graphs that have a low probability under these simple distributions – in other words: for which such a simple distribution does not give a good fit – the formulae are untrustworthy and the bootstrap standard error is more reliable, since it requires only the exchangeability of the vertices and not the large degree of randomness that is inherent to these simple distributions. It may be expected, because of this large degree of randomness assumed by the simple distributions, that the bootstrap standard errors will tend to be larger than the standard errors derived for the simple distributions.

Let us end by mentioning that the basis for these non-parametric standard errors and probabilities is

mainly intuitive, and that it would be interesting to see research devoted to their reliability. In the meantime, we suggest that it is reasonable to use the techniques we propose, since (a) there seem to be no alternatives in the general case, and (b) it is better to have a rough impression of the uncertainty or variability associated with observed network statistics than none at all. Therefore we hope that especially the bootstrap standard error will be applied widely by network analysts.[9]

REFERENCES

Boomsma, A. (1991). *BOJA. A Program for Bootstrap and Jackknife Analysis*. User's guide. Groningen: ProGAMMA (information at http://www.gamma.rug.nl)

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7, 1-26

Holland, P.W., and Leinhardt, S. (1975). The Statistical Analysis of Local Structure in Social Networks. In Heise, D.R. (ed.), *Sociological Methodology, 1976*, 1-45. San Francisco: Jossey-Bass

Hubert, L.J., and Baker, F.B. (1978). Evaluating the Conformity of Sociometric Measurements. *Psychometrika*, 43, 31-41

Krackhardt, D. (1988). Predicting with Networks: Nonparametric Multiple Regression Analysis of Dyadic Data. *Social Networks*, 10, 359 - 381

LePage, F., and Billard, L. (eds.) (1992). *Exploring the Limits of Bootstrap*. New York: Wiley

Noreen, E.W. (1989). *Computer-Intensive Methods for Testing Hypotheses*. New York: Wiley

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer

Snijders, T.A.B. (1981), The Degree Variance: An Index of Graph Heterogeneity. *Social Networks, 3*, 163-174.

Stevenson, W., & Bartunek, J. (1996). Power, interaction, position, and the generation of cultural agreement in organizations. Human Relations, 49, 75-104.

Tukey, J. (1958). Bias and Confidence in Not Quite Large Samples (abstract). *Annals of Mathematical Statistics*, 29, 614

Wasserman, S., and Faust, K. (1994). *Social Network Analysis. Methods and Applications*. Cambridge etc.: Cambridge University Press.

[9] Versions of these tests, as applied to network density only, have been incorporated as an undocumented feature in Ucinet 5 (starting with version 5.2.0.3).