Social Networks 65 (2021) 98–109

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Assessing the performance of the bootstrap in simulated assemblage networks

John M. Roberts Jr. [a,*], Yi Yin [a], Emily Dorshorst [a], Matthew A. Peeples [b], Barbara J. Mills [c]

[a] University of Wisconsin-Milwaukee, United States
[b] Arizona State University, United States
[c] University of Arizona, United States

ARTICLE INFO

Keywords:
Assemblage data
Archaeological networks
Bootstrap
Simulation
Sampling variability
Centrality

ABSTRACT

Archaeologists are increasingly interested in networks constructed from site assemblage data, in which weighted network ties reflect sites' assemblage similarity. Equivalent networks would arise in other scientific fields where actors' similarity is assessed by comparing distributions of observed counts, so the assemblages studied here can represent other kinds of distributions in other domains. One concern with such work is that sampling variability in the assemblage network and, in turn, sampling variability in measures calculated from the network must be recognized in any comprehensive analysis. In this study, we investigated the use of the bootstrap as a means of estimating sampling variability in measures of assemblage networks. We evaluated the performance of the bootstrap in simulated assemblage networks, using a probability structure based on the actual distribution of sherds of ceramic wares in a region with 25 archaeological sites. Results indicated that the bootstrap was successful in estimating the true sampling variability of eigenvector centrality for the 25 sites. This held both for centrality scores and for centrality ranks, as well as the ratio of first to second eigenvalues of the network (similarity) matrix. Findings encourage the use of the bootstrap as a tool in analyses of network data derived from counts.

1. Introduction

Network analysis has become a prominent methodological tool in contemporary archaeology (Brughmans and Peeples, 2017; Mills, 2017; Peeples, 2019). Network analytic procedures both reflect archaeology's relational focus and complement the field's traditional emphasis on time and space (Brandes et al., 2013; Collar et al., 2015; Knappett, 2011). A wide range of substantive questions in archaeology are now being addressed with network data and analysis.

In substantive applications, analysis of archaeological networks has yielded considerable insight (e.g., Birch and Hart, 2018; Borck et al., 2015; Lulewicz, 2019; Mills et al., 2013b; Peeples and Haas, 2013). One potential concern, however, is that many network analytic measures do not easily lend themselves to an assessment of sampling variability. Note that this issue is not very serious in some settings for archaeological network analysis. For example, the idea of sampling variability may have little relevance in a network analysis of observed footpaths between remains of houses at a site (e.g., Pailes, 2014), where something reasonably close to the total original network can be recovered. But for networks based on sites' artifact assemblage similarities, it is natural to consider sampling variability in the assemblage counts underlying the network analysis. One perspective is that the excavation or other primary data collection effort resulted in what is, in effect, a sample from a larger "population" of artifacts existing at the site. Alternatively, even a data collection effort that obtained literally every artifact present at any depth at a site would still be a sample of materials that were historically in use at the site or can be considered one realization of a historical process from which the observed assemblage data emerged.

A further complication is that such research takes the network constructed from the assemblage data, not the assemblage data itself, as the object of interest. A direct formal assessment of sampling variability—as could be possible in, say, estimation of multinomial probabilities—is therefore unlikely to be feasible in this setting. For this reason, researchers have discussed the bootstrap as an approach to understanding sampling variability in measures derived from networks of assemblage similarity (Gjesfjeld, 2015; Mills et al., 2013b; Peeples et al., 2016). Bootstrapped datasets can be resampled from observed assemblage counts and then submitted to the same transformation into site-by-site

* Corresponding author.
E-mail address: jmrob@uwm.edu (J.M. Roberts).

https://doi.org/10.1016/j.socnet.2020.11.005

Available online 29 December 2020

network data as was applied to the original observed assemblage. Interpretation of network measures can then be informed by variability estimates obtained from the bootstrapped data.

Although our work is motivated by problems in archaeology, these ideas have wider applicability. TheBorgatti et al. (2009) typology includes similarity of attributes as one of the fundamental bases of network ties; one possible attribute is an observed categorical distribution for each actor from which actor-to-actor similarity can be measured. In the present paper, then, "assemblage" can be taken to mean any such categorical distribution of counts, with actual archaeological assemblage data just one example. These questions are also important for the analysis of other social networks in which ties are defined by summarizing count data. Examples include animal dominance networks in which a directed tie indicates which of a pair won the majority of their observed dominance contests (Roberts and Liedka, 1999) and animal association networks based on counts of occasions in which two animals were spatially proximate (Roberts et al., 2019).

Investigation of the bootstrap in this general data context is therefore relevant beyond archaeology and can inform research in any domain in which observed counts are the basis of network ties. We do not know of any directly similar research to date, though with the caveat that the literature involving network centrality is so extensive that we surely are not aware of all potentially related work. Perhaps the most similar research setting in which we are aware of the bootstrap being used to assess variability in network tie weights and subsequent measures is that of psychological networks of traits or other psychological elements (see Epskamp et al., 2018; Heeren et al., 2018). In that domain, a tie weight may reflect the association between two elements (variables) as observed in sample data, so the bootstrap can be performed on the underlying sample data.

For this approach to be useful, the analyst must have some confidence that the bootstrap can successfully capture underlying sampling variability in this situation. Theoretical justifications for the bootstrap (e.g., Mammen, 1992) rely on asymptotic arguments that may not be valid in finite data and are based on regularity conditions that are likely not met in the series of transformations that take assemblage data into network data and then network measures. Indeed, bootstrap approaches can fail even in seemingly more straightforward circumstances than the assemblage similarity networks of interest here. The present paper investigates this problem via simulation of assemblage networks and bootstrapping of the simulated data. Because we are then examining data generated by a known process, bootstrap assessments of sampling variability can be compared to the true variability implied by the data-generating process. This permits the evaluation of bootstrap methods' validity in the assemblage similarity context and, more generally, speaks to the utility of the bootstrap in networks of actor similarity derived from count distributions.

The paper begins with a brief review of some literature on variability in centrality measures, the archaeological networks setting, and the bootstrap as an analytic tool. We then discuss the core simulation, rooted in empirical data for a 250-year period (AD 1200–1450) from a valley in what is today southeastern Arizona. We report on various analyses of the simulated data that address the effectiveness of the bootstrap for assemblage network data. Finally, we comment on the implications of the results for the viability of the bootstrap in analysis of network data of this sort.

2. Variability in centrality measures

Analysts have long been interested in aspects of variability in network actor centrality measures; here we briefly note some of that work. Bolland (1988) conducted an early study of centrality in simulated networks produced by randomly changing elements of the 0/1 adjacency matrix of an empirical network. Among other analyses, he compared mean correlations between centrality scores in the simulated and original empirical networks under different levels of randomness (up to 20 % of network elements being changed) for different centrality measures. Mean correlations declined with increasing randomness, though remaining quite high for some measures when the simulation approach fixed network density at its observed level. For several selected nodes representing different network positions, he also reported bias and variability in centrality scores for these nodes under the various measures. Results varied by measure and the nodes' structural positions.

Costenbader and Valente (2003) assembled a collection of 59 empirical binary networks from a variety of substantive domains and created simulated data by repeatedly sampling rows of each network's adjacency matrix. They carried out this exercise for each network at a series of sampling proportions from 0.80 to 0.10, at each sampling proportion assessing the average correlation between centrality scores obtained from the simulated data and those obtained from the original network. As expected, this correlation tended to decline as a smaller proportion of a network was sampled, though at different rates for different centrality measures and with various network characteristics related to these correlations for some but not all centrality measures. More recent studies of the impact of node removal in networks from specific substantive domains include, for example, Silk et al. (2015) and Peeples et al. (2016). Research continues in this area, such as the Smith et al. (2017) examination of non-random (with respect to nodes' centrality) removal of nodes.

Borgatti et al. (2006) extended these investigations by considering various kinds of random changes to binary network data in a single analysis. Starting from randomly generated original networks, they simulated error in the form of not just node removal but also node addition, tie removal, and tie addition. As with random switching in Bolland (1988) and node removal in Costenbader and Valente (2003), the average correlations between centrality measures in the original and altered networks declined with increasing errors of any kind. Results also appeared roughly similar for all four centrality measures examined. Other studies have investigated the impact of random and non-random changes in nodes and ties on ranking of centrality scores, rather than on the scores themselves (Basu et al., 2016; Kim and Jeong, 2007).

In the present paper, we focus on eigenvector centrality. Eigenvector centrality embodies the intuition that an actor's centrality is proportional to the centralities of the alters with whom the actor is tied. That is, one is central to the extent that one's contacts are central. Writing c_j for node j's centrality score, and A_{jk} for the value of the tie between nodes j and k, this implies that $c_j \propto \sum_k A_{jk} c_k$, and suggests that centrality scores c_j be obtained from the eigenvector associated with the largest eigenvalue of the network's adjacency matrix A (Bonacich, 1972). This formulation works with either binary network data or with the weighted ties that are of interest here.

Because the centrality scores come from an eigenvector of the adjacency matrix, a possible perspective on variability in eigenvector centrality scores draws on eigenvalue and eigenvector perturbation theory. That theory addresses the question of how the eigenvalues and eigenvectors of a matrix A are affected when the matrix is perturbed (that is, when A is replaced by A + E for some matrix E). Results such as the classic Davis and Kahan (1970) Theorem provide bounds on the changes in the elements of the eigenvectors of A; these bounds depend on structural features of A and E, including differences between successive eigenvalues and the size (in the sense of some matrix norm) of the perturbation. Recent research has continued to refine these bounds in such settings as statistical applications (Yu et al., 2015) and random graphs (Eldridge et al., 2018). Segarra and Ribeiro (2015) considered perturbation bounds in their investigation of the stability of classic centrality measures, including eigenvector centrality, for weighted networks. As this theory may not directly address applied researchers' interest in assessing sampling variability in centrality scores from an empirical assemblage network, we do not pursue it here. Still, if sampling variability in the network is seen as producing a distribution of perturbation matrices E around a true adjacency matrix A, then the

results in this literature would imply bounds on variability in eigenvector centrality scores under this distribution.

3. Archaeological networks

Although the use of social network analysis (SNA) has grown quickly in many academic disciplines, SNA was relatively slow to gain traction in archaeology (see Brughmans and Peeples, 2017). However, this has recently changed, with archaeologists increasingly using network data and borrowing SNA methods and models to understand how people interact with one another, material things, and the natural environment (Collar et al., 2015: Fig. 5; Mills, 2017; Peeples, 2019). These applications view the structure of interactions as crucial for understanding network actors' behavior and resource distribution. SNA emphasizes an actor's (node's) position and network constraints and opportunities in explaining the node's outcomes, and facilitates research on questions such as the relationship between geographical and social distances within a network of sites. In addition, many features of the archaeological context should interest the broader network science community. For example, there are few other research settings in which it is possible to examine network change over such long timespans. Archaeological network analysis is an important arena for studying how social and geographical positions within a network act to influence future social-structural configurations.

As noted above, one important type of archaeological network is based on the measured similarity between sites' artifact assemblages (Hart and Engelbrecht, 2012; Golitko et al., 2012; Golitko and Feinman, 2015; Habiba et al., 2018; Hart et al., 2017; Mills et al., 2013a, 2013b, 2015, 2018; Östborn and Gerding, 2014; Weidele et al., 2016). In this approach, the artifacts found at each site are classified in some way, and the sites' distributions of artifacts across these classification categories can be compared. For instance, the assemblage of interest may consist of ceramics, with each artifact classified into a *ware* category based on the artifact's physical characteristics, providing raw data that report each site's ware counts (see Mills et al., 2013a, 2016). A network of sites can then be constructed by measuring the symmetric similarity of ware distributions at pairs of sites. To date, most work has relied on archaeology's Brainerd-Robinson statistic (Brainerd, 1951; Robinson, 1951), $BR_{ij} = 200 - \sum_k |P_{ik} - P_{jk}|$, or the equivalent dissimilarity index from sociology (Duncan and Duncan, 1955), $D_{ij} = \frac{\sum_k |p_{ik} - p_{jk}|}{2}$. In these expressions, i and j index sites, while k indexes wares.

The measured similarity can be considered a weight on the tie between two sites. Although this continuous information could be transformed somehow into the presence or absence of a tie between the sites, such transformation risks losing meaningful information (Peeples and Roberts, 2013). Studies of ceramic networks have used such binarization to create network displays, while still using the weighted ties in their analyses (Mills et al., 2013a, 2013b).

4. The bootstrap

The bootstrap is a general method for assessing sampling variability in statistics for which closed-form expressions for this variability are infeasible (Efron and Tibshirani, 1993). The bootstrap is one of a family of resampling methods that developed in tandem with fast computing. The nonparametric bootstrap involves sampling with replacement from the observed data: for data with N observations, a resampled dataset is created by sampling N times with replacement from the observed data. Many such samples are drawn, and the statistic of interest is calculated for each resampled dataset. Variability in the calculated statistic across the resampled (bootstrapped) datasets estimates sampling variability in the statistic, and the resulting bootstrapped distribution of the statistic can be used to construct confidence intervals for its population value. This may be as simple as using the bootstrapped distribution's standard deviation as the statistic's standard error in an otherwise typical normal- or t-based symmetric confidence interval formula. (In this paper, we usually simply refer to the bootstrap distribution's "standard deviation," but sometimes use "standard error" when convenient.) Or the resampled distribution itself could be used to construct a confidence interval, without appealing to asymptotic normality of the estimated statistic; such a confidence interval could be asymmetric and involve bias correction (Efron and Tibshirani, 1993; Manly, 1997). The parametric bootstrap differs by sampling from a parametric model rather than directly from the observed data, but again sampling variability is estimated from variability in a statistic's value in the bootstrapped datasets. (See Rosvall and Bergstrom (2010) for an example of the parametric bootstrap for weighted network data.)

Network data may not appear to fit with the framework of the nonparametric bootstrap. Of course, the growing popularity of statistical approaches to network analysis (Lusher et al., 2012) has encouraged analysts to consider observed network data in stochastic rather than deterministic terms. Still, the nonparametric bootstrap strategy of sampling with replacement does not feel very natural for the usual binary network data, at least if the problem is viewed as one of stochastic ties among a set of given nodes. However, in the present context of networks based on assemblage similarity, the underlying artifact counts provide a straightforward route into the nonparametric bootstrap, as resampling can proceed from the observed distribution of artifacts into classification categories at each site. This approach can also be used in other contexts in which similarities are derived from count distributions associated with each actor.

Assemblage data that are drawn from many sites are unlikely to have come from a single data collection effort. It is, therefore, most natural to fix the observed number N_i of artifacts at each site i in the bootstrap resampling, rather than simply fixing the overall sample size $N = \Sigma N_i$. At each site, N_i artifacts are sampled with replacement, and the resulting resampled assemblage data is transformed into a site-by-site similarity network. The network analysis of interest can then be conducted on this similarity network; repeating this many times yields a bootstrapped distribution for each network measure of interest.

4.1. Bootstrap failure

Although the bootstrap has been applied in many analytic contexts, the procedure is not guaranteed to "work" in the sense of providing a reasonable approximation to a given statistic's actual sampling distribution. Formal justifications for the bootstrap distribution's convergence to a statistic's asymptotic sampling distribution (e.g., Bickel and Freedman, 1981; Mammen, 1992) involve regularity conditions that may not always be met beyond the particular situations in which they were proposed. Also, it is possible that asymptotic results require unrealistically large sample sizes to be realized in practice. It follows that standard errors estimated from the bootstrap could differ systematically and substantially from the statistic's actual standard error. If so, confidence intervals or other summaries constructed from the bootstrap distribution may have poor coverage or otherwise be misleading. This is a serious concern because the bootstrap can fail in this sense even in seemingly straightforward situations.

A fundamental type of bootstrap failure was described by Agresti (2007) in the setting of the logit $L = log\left(\frac{\pi}{1-\pi}\right)$ for probabilities π of binary outcomes. Agresti noted that for the natural estimate of π as the number of successes divided by the number of trials, there is a non-zero probability that a sample will yield an estimated π of zero or one, as there is a non-zero probability that the sample will consist of all successes or all failures. Given this non-zero probability of the sample logit taking the value infinity (if the estimated π equals one) or negative infinity (if the estimated π equals zero), the sample logit's variance does not exist. The bootstrap estimate of the variance of the logit's sampling distribution is, therefore, in effect attempting to target something that

does not exist. Agresti contrasted this true sampling distribution with the asymptotic distribution of the sample logit; that distribution converges to normality (with finite variance) as the sample size increases.

Is an analogous situation possible when considering centrality scores from a network analysis of assemblage data? With finite assemblage data, underlying ware (category) probabilities at the various sites could be such that there is a non-zero probability of each site having an identical assemblage, but this would cause no difficulties if it simply implied that all sites have identical centrality scores. A more realistic concern in this vein is the possibility that an assemblage network be disconnected; for instance, if a site's observed assemblage consists entirely of objects that are not found anywhere else, it will have no overlap with any other site's assemblage. Then its measured similarity with all other sites will be zero, and that site will be an isolate in the network. Likewise, the assemblages of a set of sites might show no overlap with any sites outside that set, disconnecting the network even if there are no isolates. Some centrality measures are defined for both connected and disconnected networks, so disconnectedness would not introduce any difficulties. But some measures are not defined for a disconnected network, and, in turn, the sampling distribution for the scores would not be defined.

Even in light of this theoretical possibility for certain network centrality scores, we believe that bootstrap assessment of sampling variability in measures derived from assemblage networks is appropriate. In many realistic cases, a reasonable set of assemblage probabilities would, when combined with the typical numbers of artifacts in each assemblage, make a disconnected network extremely unlikely, even if this probability were not literally zero. If so, a measure's sampling variance conditional on the network being connected will still be a substantively meaningful quantity.

In practice, we assume that analysts are studying sites in particular geographic regions and time periods that have been the objects of previous theory and research, so that a pattern of complete non-overlap leading to a disconnected network would be understood to be incompatible with substantive knowledge even before looking at data. (A disconnected network may be more plausible in domains other than archaeology.) We also expect that researchers would refrain from network analysis that relied on assemblage data with very small cell counts. Note that assemblage data might record only the presence or absence of object types, rather than counts of objects, as in some burial data (Sosna et al., 2013). A similar situation would hold for animal dominance in which the data available to the analyst only indicated which animal in a pair won more contests, instead of the number of contests won by each. In that case, the similarity between sites could still be measured and used to create a network, but the approach to bootstrap assessment of sampling variability undertaken here would not be available. Further, if the substantive setting were such that a disconnected network would be possible and scientifically meaningful rather than simply a reflection of insufficient data, researchers could choose centrality measures that permit, or can be modified to permit, disconnected network input. Given all of this, the prospect suggested by the Agresti (2007) example does not seem to be an overwhelming concern in the present context.

A classic example of a more typical sort of bootstrap failure was introduced by Bickel and Freedman (1981) and discussed by many others since. The example concerns the estimation of θ in data drawn from a $[0, \theta]$ uniform distribution. The greatest observed value estimates theta, but the bootstrap distribution of the maximum value in a finite sample will quite poorly approximate the maximum's true distribution for a known theta. In addition, results that in other situations demonstrate the convergence of the bootstrap distribution to the desired asymptotic distribution do not hold in this case.

Various conditions can make the bootstrap more prone to failure (Chernick, 2007). For instance, a very small observed sample provides little information on the underlying distribution of values from which it came, and so resampling from it will be unlikely to usefully approximate

Table 1a
Ceramic sherd data for simulations; decorated wares.

Sites	Decorated Wares																						Decorated	Total
	1	2	3	5	6	8	9	10	11	12	13	15	18	19	20	21	22	23	25	26	27			
1	2	0	0	0	0	25	0	0	3	0	0	0	0	329	0	5	14	0	0	0	0		378	1219
2	0	0	0	0	0	18	0	0	2	0	0	0	0	69	0	0	0	0	0	0	0		89	279
3	0	0	0	0	1	95	0	0	2	1	0	0	0	12	0	9	30	0	0	0	4		154	911
4	0	0	1	0	0	539	0	1	3	2	0	0	0	102	1	17	64	0	0	0	5		735	2275
5	1	0	13	0	0	2783	1	0	19	2	0	0	0	280	0	143	415	0	0	0	5		3662	8622
6	0	0	0	0	0	26	0	0	0	0	0	0	0	322	0	1	10	0	0	0	0		359	1244
7	0	0	1	0	3	90	0	0	8	0	0	0	0	127	0	24	15	0	0	0	0		268	988
8	0	0	0	0	0	93	0	0	9	0	0	0	0	7	0	6	17	0	0	0	1		133	458
9	3	0	1	0	0	165	0	0	82	1	0	0	0	173	0	4	20	0	0	0	19		468	2345
10	11	2	0	0	0	49	0	0	12	0	0	0	0	566	0	14	2	0	0	0	0		656	1818
11	0	0	2	0	0	71	0	0	0	25	0	0	0	0	0	5	42	0	0	0	9		154	1601
12	0	0	4	0	0	5	0	0	13	4	0	0	0	252	0	9	6	0	0	0	104		397	2422
13	0	0	3	0	0	458	0	0	13	0	0	0	0	201	0	15	29	0	0	0	5		724	2346
14	0	0	1	0	0	84	0	0	2	2	0	0	0	33	0	8	36	0	0	0	14		180	1045
15	2	0	1	0	0	353	0	0	36	4	0	0	0	27	0	16	112	0	0	0	5		556	2544
16	258	0	0	264	0	1285	0	0	396	476	16	0	0	9543	0	35	6	59	2	13	263		12,616	42,656
17	19	1	1	0	0	5	0	0	404	1	0	0	0	2950	0	9	4	0	1	0	32		3427	5457
18	0	0	85	0	0	15	0	0	0	16	0	0	0	25	0	0	15	5	0	0	80		241	10,893
19	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	4	6	0	0	0	0		65	220
20	1	0	0	1	0	0	0	0	11	4	0	0	0	89	0	4	6	0	0	0	0		116	478
21	4	0	0	0	0	0	0	0	1	28	0	0	0	140	0	1	0	2	0	0	15		191	1967
22	0	0	0	0	0	7	0	0	1	1	0	0	0	78	0	2	0	0	0	0	0		89	371
23	0	0	6	0	1	76	0	0	1	1	0	0	0	410	0	39	76	0	0	0	6		616	1831
24	545	0	39	159	0	1298	0	0	647	1077	11	1262	10	5872	0	357	3540	104	0	0	903		15,824	73,496
25	0	0	0	0	0	8	0	0	1	0	0	0	0	0	0	0	16	0	0	0	0		25	301

the distribution of some statistic. Other conditions involve much deeper statistical theory, but the overall message is still that one cannot assume that the bootstrap is an effective tool in all settings. Given these possible pitfalls, it is important for network analysis in archaeology or other settings with analogous data that the effectiveness of the bootstrap be evaluated in the specific context of centrality scores derived from assemblage similarity networks. Because this network analysis involves complicated transformations of the assemblages' ware distributions, direct investigation of the bootstrap's viability in assessing variability in site centrality scores is needed.

5. Current study

In the present paper, we use simulation to explore the bootstrap for network analysis of archaeological assemblage data. By simulating assemblage datasets under a realistic probability structure for the wares at the sites, we can determine the "true" distribution of network statistics implied by the probability structure. Applying the bootstrap to each of many simulated datasets allows a comparison of estimated sampling variability from the bootstrap against the true variability in a statistic. This permits assessment of the accuracy of bootstrap estimates of sampling variability in a meaningful substantive context.

5.1. Simulation

We simulated artifact assemblages under a probability structure derived from actual ceramic assemblage data from the San Pedro Valley in Arizona, in the North American Southwest (Clark and Lyons, 2012; Mills et al., 2013a, 2013b). At each of 25 archaeological sites in this region, ceramic fragments obtained in excavations have been classified into 35 ware categories. Ceramics can also be categorized more finely into types (see Clark and Lyons, 2012), but analysis of ceramic similarity networks to date has focused on the higher-level classification into wares, due to the likely greater reliability of this classification (Mills et al., 2013a, 2016). Note that the simulation does not reflect all archaeologically relevant features of the San Pedro data. For instance, not all 25 sites were occupied simultaneously, and substantive analyses have distinguished between ceramic similarity networks at different periods of the region's occupation (Mills et al., 2013a, 2013b; Roberts et al., 2012). Nonetheless, the data provide a realistic basis for simulation of ceramic similarity networks, with sample sizes ranging from 220 to 73,496 across the 25 sites. For comparison, mean sample size for the larger Southwest database covering Arizona and parts of New Mexico is 1,731 with a standard deviation of 16,431, so the sample sizes here are fairly typical.

Tables 1a and 1b gives observed ware counts at each of the 25 sites. Below we discuss the distinction between decorated and undecorated wares; for the moment, simply note that Table 1a reports on decorated ware categories, and Table 1b on undecorated wares. We have relabeled site names and ware categories to emphasize that these data are being used here to provide a realistic basis for the simulation, rather than to draw substantive conclusions about the actual archaeological case-study. (The actual San Pedro site and ware names are given in Tables S1, S2A, and S2B in the Supplementary Material.) For simulation, these counts were converted into probabilities at each site by dividing by the site total, with data simulated independently at each site. The observed site totals N_i are preserved in all simulated datasets (and all bootstrap resampling from the simulated data). Differences in observed site totals likely reflect both the amount of deposited material at the site and the extent of the data collection effort. We view, therefore, the combination of different sites' data as akin to product multinomial data composed of a set of distinct multinomials (Bishop et al., 1975).

5.2. Network data and measures

We measured the similarity between sites i and j via the dissimilarity index D_{ij} between their assemblages; because the network should reflect similarity rather than dissimilarity—that is, there should be a greater weight on ties between sites with more similar assemblages—we used $(1 - D_{ij})$ as the weight on tie (i, j). Although simulations and bootstrap resampling involved all wares present in an assemblage, in constructing the network data we assessed similarity between sites using only a subset of wares. Wares can be described as decorated or undecorated, with decoration involving the addition of colored coatings, called slips, or paint of various colors made from organic and mineral materials. This designation can be complicated in practice, as sometimes a given ware contains both decorated and undecorated types. Because decorated wares are thought to have symbolic rather than purely utilitarian

Table 1b
Ceramic sherd data for simulations; undecorated wares.

Sites	Undecorated Wares														
	4	7	14	16	17	24	28	29	30	31	32	33	34	35	Undecorated
1	16	0	0	775	0	0	0	0	0	0	50	0	0	0	841
2	1	0	1	152	0	0	0	0	1	0	35	0	0	0	190
3	54	0	0	685	0	0	0	0	0	0	17	1	0	0	757
4	16	0	0	1514	0	0	0	0	5	0	3	2	0	0	1540
5	125	0	3	4724	0	0	0	0	5	0	76	14	13	0	4960
6	23	0	0	812	0	0	0	0	0	0	50	0	0	0	885
7	65	0	0	642	0	0	0	0	1	0	10	2	0	0	720
8	65	0	0	242	0	0	0	0	0	0	15	0	1	2	325
9	125	0	1	1660	0	0	0	0	7	0	50	34	0	0	1877
10	13	0	0	1099	0	0	0	0	1	0	42	0	7	0	1162
11	74	0	0	1333	0	0	0	0	1	0	37	2	0	0	1447
12	333	0	0	1605	0	0	0	0	0	0	56	30	1	0	2025
13	10	0	2	1537	0	0	1	0	0	0	70	2	0	0	1622
14	484	0	0	363	0	0	1	0	1	0	14	2	0	0	865
15	858	0	0	1085	0	0	0	0	3	0	40	2	0	0	1988
16	308	2	1	29,024	1	4	8	0	0	4	549	0	139	0	30,040
17	23	14	0	1839	0	0	0	0	3	0	98	18	35	0	2030
18	1180	0	1	9344	0	0	1	0	15	0	109	2	0	0	10,652
19	3	0	0	149	0	0	0	0	0	0	3	0	0	0	155
20	3	0	0	323	0	0	0	0	1	0	22	10	3	0	362
21	42	0	0	1599	0	0	0	1	2	0	65	60	7	0	1776
22	0	0	0	277	0	0	0	0	0	0	5	0	0	0	282
23	27	0	0	1154	0	0	0	0	0	0	32	2	0	0	1215
24	3118	0	0	54,158	0	0	8	388	0	0	0	0	0	0	57,672
25	148	0	0	118	0	0	0	0	2	0	8	0	0	0	276

importance, and are more consistently used categories in archaeological research, assessing site similarity based on only the subset of decorated wares more closely represents how researchers have used ceramic assemblages to construct networks (Mills, 2016; Mills et al., 2013b). In other domains, it may likewise be appropriate to assess actors' similarity using only a subset of the observed assemblage categories. Decorated wares are indicated in Table 1a, along with the total number of decorated artifacts at each site.

In principle, any network measure that is appropriate for valued, symmetric data can be investigated in an analysis of assemblage similarity networks. To date, much attention has focused on the interpretation of sites' eigenvector centrality (Mills et al., 2013a, 2013b, 2015, 2018), as defined in Section 2. (Note that bootstrap methods have been used in the eigenvector setting, as in Efron and Tibshirani's (1993) example of principal components.) Eigenvector centrality has been popular in archaeological applications in part due to Borgatti's (2005) typology of the nature of network "flow" and the logics of different centrality measures. In that typology, eigenvector centrality is discussed as appropriate when the network flow process involves the potential for simultaneous rather than the sequential influence of all of a node's contacts, in which, as Borgatti (2005: 62) notes, walks rather than trails, paths, or shortest paths are relevant. Archaeologists consider assemblages to be the cumulative result of consumption activities, which are systematically sampled through excavation and/or surface surveys (Mills et al., 2016; Peeples et al., 2016). The presence of different sets of vessels at a site is affected by the flows of imported ceramics to the site, the degree of on-site production, and the socially constrained choices that the residents make in what to use for specific purposes such as cooking, serving, and storage. The activities in which ceramic vessels are used, and their frequencies of use, directly affect container breakage resulting in different proportions of ceramics at each site.

We focused on sites' eigenvector centrality here. Throughout the analyses, eigenvectors were normalized to have sum of squares equal to the number of sites (here 25), so that 1 represents a typical centrality score. We also considered ratios of first to second eigenvalues as assessments of the network structure's unidimensionality and centrality scores' adequacy as descriptions of that structure. Note that eigenvector centrality analysis of a disconnected network will yield a set of centrality scores for each component, so that scores from different components cannot be legitimately compared. In the present case, however, a disconnected network would require that all of the assemblages at some component's sites have no overlap with those at all sites in other components. Even when restricting assessment of similarity to the subset of decorated wares, this situation did not occur in any of the simulations (or bootstrap resamples) analyzed here.

There may be greater substantive interest in sites' ranking by centrality than in sites' literal centrality scores. Perhaps differences in rank will simply seem more interpretable than differences in some sort of normalized centrality score, especially in conveying the results to audiences that have limited familiarity with network methods. But with respect to the theme of the current paper, the discrete nature of ranks may make analysts less confident in the bootstrap's validity, as well as introducing the possibility of edge effects at the maximum and minimum ranks. Ranks have also been examined in the larger bootstrap literature (for instance, Hall and Miller, 2009). Therefore, we considered site centrality rank along with centrality scores in our analyses.

5.3. Data and analysis

To identify the true sampling distribution of centrality scores under the San Pedro-based probability structure above, we simulated 500,000 assemblage datasets. Each simulated dataset was converted into a weighted site-by-site network, as described above. We examined dyad-level variability in the resulting networks, with assessment of this variability setting the stage for our next step in which we obtained a sampling distribution of eigenvector centrality for each of the 25 sites over 500,000 simulations. (This information could also be used to consider joint distributions of site centrality for pairs or larger subsets of sites, but we do not pursue that here.) Although technically these will only approximate the true sampling distributions of site centrality implied by the probability structure, an approximation resulting from 500,000 simulations will be very accurate. These distributions can be characterized by descriptive statistics and histograms, and variability in the centrality scores can be considered in light of dyad-level variability in the networks.

To assess the performance of the bootstrap, we simulated 5000 additional assemblage datasets, and for each created 10,000 bootstrap replicate datasets. We chose 10,000 as a typical number of bootstrap replicates that a practicing researcher might use, though of course increasing computing power means that ever-larger numbers of bootstrap replicates are feasible in practice. We transformed each bootstrap replicate into a weighted network, and calculated eigenvector centrality for the sites.

We then investigated the bootstrap data in order to understand how well bootstrap estimates of sampling variability in site centrality track the true variability in these measures. We first examined pooled bootstrapped centrality scores from all 5000 simulated datasets and compared these distributions to the true distributions from the 500,000 simulations above via descriptive statistics and histograms. This is informative, but it does not directly correspond to an analysis that a researcher would perform in practice. As discussed above, the standard deviation of a site's eigenvector centrality across the 10,000 bootstrapped datasets is an estimate of the sampling variability (standard error) in that measure. We therefore especially focused on the distributions of these bootstrap estimates of sampling variability in site centrality across the 5000 simulated datasets. We examined the means and standard deviations of these distributions of bootstrap variability estimates and compared the estimates to the centrality measures' true sampling variability (as obtained from 500,000 simulated datasets). This helps show how well a single bootstrap analysis can be expected to capture the true sampling variability in site centrality.

An alternative that we did not pursue here would be to compare, for each site and each of the 5000 simulated datasets, the distribution of 10,000 bootstrapped centrality scores to the true sampling distribution (from 500,000 simulations). This could involve a direct measure, such as Kolmogorov-Smirnov, of the difference between the two distributions. Such an analysis is potentially interesting given that bootstrap confidence intervals for a parameter may directly rely on the bootstrap distribution's tails, rather than simply using the bootstrap standard deviation as the standard error in a confidence interval formula (Efron and Tibshirani, 1993). However, the standard deviation (standard error) is an appealing and straightforward summary, even if it does not capture all aspects of the relevant distribution. We took the bootstrap estimate of the sampling standard deviation (standard error) as the main quantity of interest.

We also examined distributions of first to second eigenvalue ratios of the network (similarity) matrix, comparing bootstrap estimates of variability in this ratio to its true sampling variability. This ratio is one traditional means of assessing the dimensionality of a matrix. In factor analysis, whether this ratio (for eigenvalues of an adjusted correlation matrix) exceeds three is a long-standing informal test of unidimensional structure; Slocum-Gori and Zumbo (2011) and others have investigated such tests. In the eigenvector centrality context, this ratio might be taken to indicate the adequacy of the (unidimensional) centrality scores as a description of network structure. A low ratio would not invalidate the scores, as the original motivation for eigenvector centrality does not require a unidimensional structure. But it would point to the presence of additional meaningful structure that is not represented in the unidimensional centrality scores. This ratio is, therefore, another structural measure for which the bootstrap can assess sampling variability.

6. Results

In the following subsections, we discuss (i) the dyad-level variability in the networks constructed from 500,000 simulated assemblages. Next, we examine (ii) the true sampling distribution of the sites' eigenvector centralities under the San Pedro probability structure, based on 500,000 simulated assemblages. We briefly note (iii) the combined distributions of 50,000,000 centrality scores for each site (from 5000 simulations and 10,000 bootstrap replications), but mainly focus on (iv) the bootstrap estimates of sampling variability obtained in the 5000 simulated datasets. We describe these distributions and discuss their implications for the quality of the bootstrap's performance in a single dataset. In addition, we consider similar comparisons of bootstrap estimates and true variability for network dimensionality (as expressed via ratios of eigenvalues of the simulated network data).

6.1. Dyad-level variability in networks from simulated assemblage data

We examined descriptive information on variability in tie weights—that is, the measured site-to-site assemblage similarities— for the 300 pairs of sites across 500,000 simulated datasets. These similarities are not, in their own right, our main direct interest, but their descriptive statistics help in understanding the setting in which the centrality scores were generated. For the most part, there was not great variability in the weights, although in individual simulations they could depart from their means to a substantial extent. The average standard deviation over the 300 pairs was only 0.016, but in many dyads the range of minimum and maximum weights exceeded 0.25. Also, the difference between 5th and 95th percentile values was greater than 0.08 for many pairs. Descriptive information for all dyads is given in the Supplementary Material (Table S3). In many pairs, raw histograms suggested roughly normally distributed similarities, although for some pairs these histograms appeared rather unsmooth when using small bins. Fig. S1–S15 in the Supplementary Material give examples of these histograms from several pairs of sites among those with the most and least variability in their tie weights. Note that variability in the weights across simulations seemed to reflect assemblage size, with greater variability associated with smaller assemblages.

6.2. Distributions of centrality scores from simulated data

We next discuss the distributions of site centrality in 500,000 datasets simulated from the San Pedro probability structure. Table 2 displays summary information on the distribution of eigenvector centrality scores for each site across 500,000 simulations. These distributions represent the true variability in each site's centrality under the probability structure and site assemblage totals used for the simulations. Table 2 also shows each site's true centrality, calculated via the similarities derived from the sites' ware probabilities.

We note several features of these distributions. First, the distributions' mean or median values were generally close to the sites' true centrality scores. The greatest absolute difference between a mean of these distributions and the corresponding true score was 0.014, for Site 25, and on average across the sites the absolute difference was only about 0.004. Second, standard deviations or variances of these distributions of site centrality were relatively small compared to the means or medians. It was rare for a site's centrality score in a particular simulation to depart dramatically from its true value; recall from the previous section that there was also relatively little dyad-level variability in network tie weights. Third, variability in centrality scores (represented by the standard deviations reported in the table) was somewhat more associated with the sites' true centralities than with the amount of data at each site. Across the 25 sites, this standard deviation correlated –0.70 with the true centrality score, and –0.41 with the logged total number of artifacts. In general, the true variability in a site's centrality scores was greater for sites with fewer artifacts and with lower centrality.

6.3. Sampling variability as assessed by the bootstrap

We next turn to the bootstrap assessment of sampling variability. For each of 5000 simulated datasets, we executed the bootstrap using 10,000 resampled datasets. We can first consider, for each site, the resulting 50,000,000 (from 5000 × 10,000) bootstrapped eigenvector centrality scores. As mentioned above, this is a somewhat artificial construction, because an actual analysis would involve bootstrapping from a single dataset, not from 5000. Still, examining this set of 50,000,000 scores for each site helps give a sense of how variability in bootstrapped centrality scores compares to the true variability implied

Table 2
Descriptive statistics for eigenvector centrality scores from 500,000 simulated datasets.

Site	From Probs.	Mean	Std. Dev.	Min	Max	Percentile				
						5th	25th	50th	75th	95th
1	1.203	1.205	0.018	1.108	1.271	1.173	1.193	1.206	1.218	1.233
2	1.220	1.223	0.009	1.125	1.260	1.207	1.217	1.223	1.229	1.237
3	0.801	0.794	0.035	0.634	0.960	0.736	0.770	0.794	0.818	0.852
4	0.830	0.829	0.024	0.718	0.942	0.789	0.813	0.829	0.845	0.869
5	0.754	0.753	0.018	0.674	0.838	0.724	0.741	0.753	0.765	0.783
6	1.177	1.182	0.021	1.055	1.261	1.146	1.168	1.182	1.196	1.214
7	1.140	1.141	0.021	1.036	1.227	1.106	1.127	1.141	1.155	1.174
8	0.760	0.754	0.037	0.588	0.929	0.695	0.729	0.754	0.779	0.815
9	1.044	1.045	0.022	0.935	1.141	1.008	1.030	1.045	1.060	1.080
10	1.193	1.194	0.015	1.113	1.258	1.168	1.184	1.195	1.205	1.219
11	0.636	0.630	0.023	0.514	0.740	0.592	0.615	0.631	0.646	0.668
12	1.024	1.025	0.023	0.911	1.122	0.987	1.010	1.026	1.041	1.063
13	0.988	0.987	0.022	0.878	1.083	0.951	0.972	0.987	1.001	1.022
14	0.924	0.918	0.035	0.747	1.075	0.860	0.895	0.918	0.941	0.974
15	0.770	0.765	0.022	0.666	0.878	0.729	0.750	0.765	0.780	0.801
16	1.187	1.190	0.010	1.135	1.234	1.174	1.184	1.190	1.197	1.206
17	1.100	1.106	0.014	1.034	1.164	1.083	1.097	1.106	1.115	1.128
18	0.452	0.449	0.038	0.278	0.625	0.386	0.423	0.449	0.475	0.512
19	1.112	1.116	0.023	0.996	1.186	1.074	1.101	1.118	1.133	1.150
20	1.146	1.142	0.019	0.992	1.215	1.109	1.131	1.144	1.156	1.171
21	1.043	1.046	0.025	0.908	1.146	1.003	1.030	1.047	1.063	1.084
22	1.192	1.185	0.027	1.039	1.265	1.134	1.168	1.188	1.205	1.224
23	1.208	1.212	0.011	1.153	1.260	1.193	1.205	1.213	1.220	1.231
24	0.961	0.961	0.009	0.923	1.001	0.947	0.955	0.961	0.966	0.975
25	0.502	0.488	0.066	0.129	0.708	0.375	0.445	0.491	0.534	0.590

by the underlying probability structure. The Supplementary Material includes a table (S4) of descriptive information for each site's 50,000,000 scores in this combined bootstrap distribution.

Fig. 1 compares histograms of the combined bootstrap distribution of centrality scores and the true distribution for Site 12, which is a rather typical site in terms of its artifact abundance and true centrality. The distributions are similar, but variability in the site's centrality implied by the bootstrap replicates is greater than the true variability (as indicated by the flatter and wider histogram). The Supplementary Material provides a table (S5) comparing variability in the sites' combined bootstrap distributions with that in the true distributions from 500,000 simulations, as well as graphs like Fig. 1 for each site (Fig. S16–S40). However, note again that these comparisons do not speak to the results that a researcher would be likely to obtain from an analysis of a single dataset.

We therefore examined the distributions, across 5000 simulated datasets, of bootstrap estimates of variability in the site centrality scores. For each simulated dataset, we estimated the sampling variability in each site's centrality score via its standard deviation over 10,000 bootstrapped datasets. We then examined the resulting distribution of 5000 bootstrap variability estimates for each site. To graphically display these distributions of 5000 estimated standard deviations of site centrality, we constructed the boxplots shown in Fig. 2. The "box" displays the 25th percentile, the median, and the 75th percentile of the distribution of standard deviation (standard error) estimates for each site. The 25th to 75th percentile inter-quartile range (IQR) is used to create the "whiskers". The upper whisker shows the highest observed value that is within 1.5 IQR of the 75th percentile, while the lower whisker shows the lowest observed value that is within 1.5 IQR of the 25th percentile. The boxplots are positioned higher for sites in which there is more inherent variability in centrality scores, with these differences closely tracking the true differences in variability of the sites' centrality scores. Over the 25 sites, mean bootstrap standard deviations correlated almost perfectly ($r = 0.998$) with the true sampling variability in site centrality, and the true standard deviation was on average just 1.003 times greater than the mean bootstrap standard deviation. Table 3 compares each site's true standard deviation of site centrality to its mean bootstrap standard deviation.

The boxplots are, in general, shorter for sites in which the bootstrap standard deviation estimates vary less across the 5000 simulated datasets. There was only a moderate relationship between the IQR of a site's bootstrap standard deviations and its logged true centrality ($r = -0.30$), but a stronger relationship with its logged abundance ($r = -0.62$), suggesting a pattern of less variability in bootstrap standard deviations for sites with greater abundance. These relationships are illustrated in Fig. 2, in which sites are ordered left to right in the figure by highest to lowest abundance and colored according to their centrality (with the ordering red, orange, yellow, green, and blue representing highest to lowest centrality in groups of five).

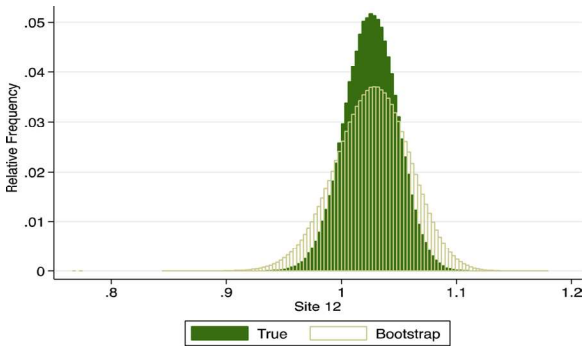

Fig. 1. True sampling distribution of centrality scores and combined bootstrap distribution for Site 12.

To understand the magnitude of this variability in the bootstrap assessments, we compared the IQR of each site's bootstrap standard deviations to its true sampling variability (its true standard deviation). Across the 25 sites, the IQR of the bootstrap standard deviations was, on average, only one-ninth of the true standard deviation, and in no instance did this ratio exceed 0.30; see Table 3. This suggests that the bootstrap standard deviation obtained from analysis of a single dataset is very likely to be a reasonable estimate of the true sampling variability.

6.4. *Dimensionality*

For the similarity matrix (with diagonal zeros) implied by the artifact probabilities used in our simulations, the ratio of first to second eigenvalues was 2.45. This suggests that the underlying network structure is not strictly unidimensional, at least under the aforementioned classic factor-analytic cutoff, but with the departure from unidimensionality not too dramatic in light of this traditional criterion. Fig. 3 shows that the sampling distribution of eigenvalue ratios was concentrated at roughly the ratio obtained from the similarity matrix implied by the artifact probabilities. The mean ratio in the true sampling distribution was 2.44, with standard deviation 0.074. We compared the true sampling variability in this ratio with the corresponding bootstrap estimates, and, on average, the bootstrap was quite successful at recovering this sampling variability. Across the 5000 simulated datasets, the mean bootstrap standard deviation was 0.073. Also, as the standard deviation over the 5000 simulations of this bootstrap assessment of variability was only 0.005, large departures from this average were rare. Fig. 4 displays the distribution of 5000 bootstrap standard deviation estimates of sampling variability in the eigenvalue ratio.

7. Ranked centrality scores

As noted earlier, substantive researchers may be more focused on rankings of site centrality than on the centrality scores themselves. Ranks are, of course, inherently discrete, raising suspicions that the bootstrap may be less effective in estimating ranks' sampling variability than it is for variability in the numerical centrality scores. Our discussion of results for centrality rank parallels the discussion for numerical centrality scores in the previous section.

Table 4 reports on the true sampling distribution of centrality ranks (based on 500,000 simulated datasets). For many sites, there was relatively little sampling variability in their ranked centrality, with the middle 90 % of the distribution spanning only a few values. In the sites with the greatest sampling variability in centrality rank, the distribution's standard deviation equaled about two places in the 25-site ranking.

Turning to the bootstrap assessments of this variability, Fig. 5 compares the combined (50,000,000 observations, from 5000 simulations x 10,000 bootstrap replications) bootstrap distribution of centrality rank to the true sampling distribution for Site 9. Site 9 was typical in its difference in these two distributions' standard deviations; other sites' figures are shown in the Supplementary Material (Fig. S41–S65), which also includes descriptive tables (S6 and S7). The bootstrap distribution was more variable than the true distribution, but otherwise quite similar, as in the comparison above using centrality scores rather than ranks. As before, however, this examination of the combined bootstrap distribution does not represent the situation of a researcher working with a single bootstrap distribution obtained from an analysis of one dataset.

For the 5000 simulated datasets, Fig. 6 shows boxplots of bootstrap standard deviations estimating sampling variability in site centrality rank; it is analogous to Fig. 2, but for ranks rather than centrality scores. As in Fig. 2, sites are ordered from highest to lowest abundance, and centrality quintiles are indicated by color (with highest to lowest centrality ordered as red, orange, yellow, green, and blue). The vertical axis is in the centrality rank scale, so variability can be interpreted in terms of

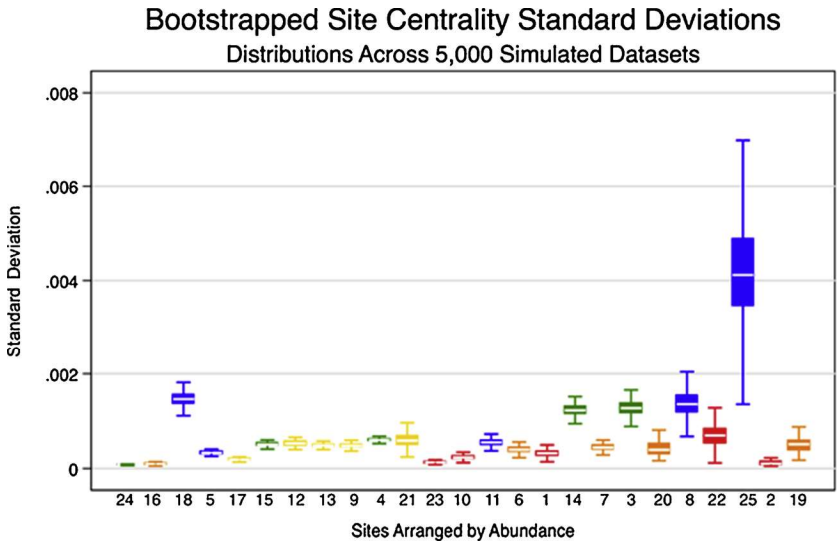

Fig. 2. Boxplots of bootstrapped estimates of standard deviations of site centrality.

Table 3
Comparison of true standard deviation of centrality score to bootstrap estimates.

Sites	True SD	Mean	IQR	True to Mean SD Ratio
1	0.0184	0.0178	0.0025	0.9637
2	0.0093	0.0109	0.0027	1.1746
3	0.0352	0.0356	0.0027	1.0105
4	0.0243	0.0245	0.0008	1.0075
5	0.0179	0.0182	0.0010	1.0140
6	0.0206	0.0197	0.0021	0.9582
7	0.0207	0.0210	0.0019	1.0125
8	0.0367	0.0367	0.0047	0.9990
9	0.0218	0.0219	0.0013	1.0027
10	0.0153	0.0152	0.0018	0.9920
11	0.0232	0.0234	0.0019	1.0072
12	0.0229	0.0229	0.0014	0.9994
13	0.0217	0.0220	0.0010	1.0141
14	0.0346	0.0351	0.0021	1.0134
15	0.0222	0.0224	0.0011	1.0120
16	0.0099	0.0098	0.0011	0.9940
17	0.0138	0.0137	0.0010	0.9948
18	0.0384	0.0383	0.0023	0.9954
19	0.0231	0.0219	0.0045	0.9447
20	0.0192	0.0204	0.0049	1.0654
21	0.0245	0.0243	0.0039	0.9903
22	0.0275	0.0257	0.0058	0.9365
23	0.0115	0.0114	0.0011	0.9911
24	0.0086	0.0087	0.0005	1.0133
25	0.0660	0.0645	0.0110	0.9781

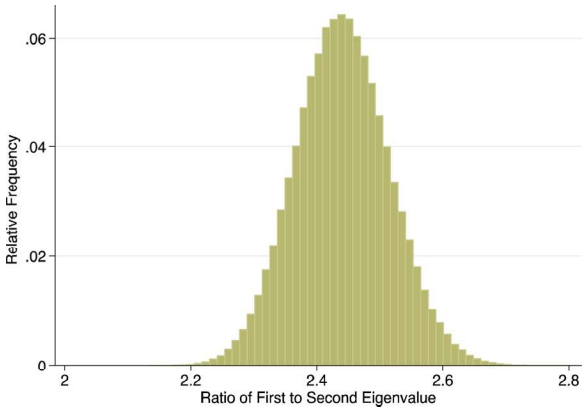

Fig. 3. True sampling distribution of ratio of first to second eigenvalues.

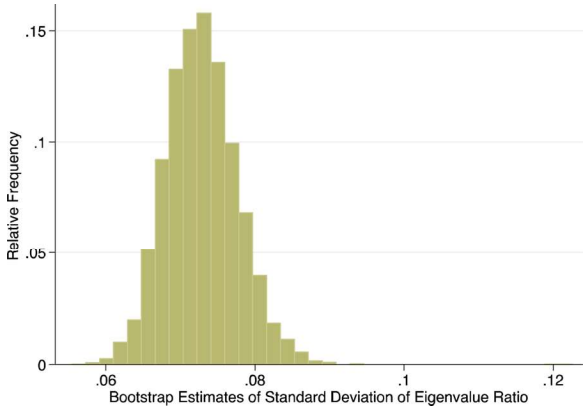

Fig. 4. Distribution of bootstrapped estimates of the standard deviation of ratio of first to second eigenvalues.

deviations in ranks.

On average, bootstrap standard deviations successfully recovered the true sampling variability in centrality rank; see Table 5. The ratio of the true standard deviation to the mean bootstrap estimate averaged 1.002 across sites, and these two quantities were correlated 0.975. Also, the IQRs for distributions of bootstrap estimates were typically small compared to the corresponding mean bootstrap estimate. Site 11 was an outlier in that comparison because its mean bootstrap standard deviation was so small, but for the remaining sites, the IQR of the bootstrap estimate distribution was, on average, roughly one-third of the true standard deviation.

While there were exceptions, the general pattern was one of greater inherent variability in the centrality rank estimates (the boxplots positioned higher) for the more central sites: the mean bootstrap standard deviation correlated –0.72 with the true centrality rank. Further, the bootstrap estimates of variability tended to be most similar across the 5000 simulated datasets (shorter boxplots) for sites ranked lower in true centrality; r = –0.57 between the IQR and the true centrality rank. There was also a general decrease in variability of bootstrap standard deviation estimates as site abundance increased (r = –0.54 between the IQR and the logged abundance). This was similar to the pattern in Fig. 2.

Table 4
Descriptive statistics for ranked eigenvector centralities from 500,000 simulated datasets.

Site	From Probs.	Mean	Std. Dev.	Min	Max	Percentile				
						5th	25th	50th	75th	95th
1	3	3.400	1.688	1	10	1	2	3	4	7
2	1	1.445	0.774	1	10	1	1	1	2	3
3	19	19.316	1.022	17	22	18	19	19	20	22
4	18	18.184	0.443	17	22	18	18	18	18	19
5	22	21.166	0.770	18	22	20	21	21	22	22
6	7	5.857	1.618	1	12	3	5	6	7	8
7	9	8.603	1.283	1	14	7	8	9	9	11
8	21	20.906	1.163	17	23	19	20	21	22	22
9	12	12.774	0.891	8	17	12	12	13	13	14
10	4	4.570	1.357	1	10	2	4	4	6	7
11	23	23.004	0.063	22	24	23	23	23	23	23
12	14	13.605	0.880	10	17	12	13	14	14	15
13	15	14.941	0.706	10	17	14	15	15	15	16
14	17	16.859	0.492	12	20	16	17	17	17	17
15	20	20.414	0.873	18	22	19	20	20	21	22
16	6	5.280	0.913	2	8	4	5	5	6	7
17	11	10.586	0.597	7	13	10	10	11	11	11
18	25	24.702	0.457	23	25	24	24	25	25	25
19	10	9.919	1.000	6	15	8	9	10	11	11
20	8	8.492	0.891	4	15	7	8	8	9	10
21	13	12.799	0.901	9	17	12	12	13	13	14
22	5	5.339	2.247	1	14	2	3	6	7	9
23	2	2.569	1.182	1	8	1	2	2	3	5
24	16	15.975	0.477	13	17	15	16	16	16	17
25	24	24.294	0.464	22	25	24	24	24	25	25

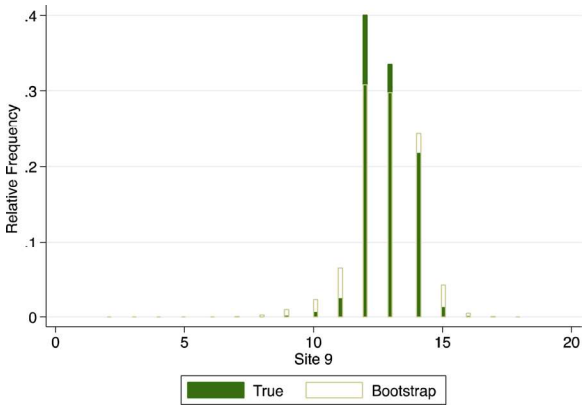

Fig. 5. True sampling distribution of ranked centrality and combined bootstrap distribution for Site 9.

8. Conclusion

Taken as a whole, the simulations reported here support the use of the bootstrap in assessing sampling variability in measures resulting from analysis of assemblage networks. Estimates of sampling variability in site centrality measures obtained from the bootstrap were typically quite similar to the measures' true variability. Further, variability in these bootstrap estimates was usually rather modest in comparison to true sampling variability for either centrality scores or ranked centrality. While simulation work like this cannot explore the full gamut of possible conditions that archaeological network researchers may encounter in practice, the simulations are strongly rooted in real data and likely provide appropriate guidance for a wide range of empirical situations. These findings therefore bolster the use of bootstrap estimates of sampling variability in analyses of networks derived from assemblage data and similar data in other settings.

Additional investigation can expand the range of simulation conditions under which the bootstrap's effectiveness is evaluated, including the use of probability matrices derived from empirical examples in other data contexts. One path for such research would be to consider larger or smaller assemblage sizes than those examined here. Our intuition is that the San Pedro assemblage sizes are reasonably representative of typical ceramic assemblage data, so that these data provide an informative foundation for our simulation study. But larger or smaller assemblages can certainly occur as well, and, moving beyond ceramics, typical assemblage sizes may be quite different for other kinds of artifacts or for similar data from other scientific fields. In particular, smaller assemblage sizes may lead to a good deal more variability in tie weights and centrality scores than was the case here, and this may affect bootstrap performance. As sample size is one of the factors determining the bootstrap's effectiveness in general, simulations focusing on contexts in which relatively small assemblages are the norm will be of special interest.

Further research can examine additional site centrality measures. Eigenvector centrality has been prominent in archaeological network research to date, but other measures may be more informative in other settings. A more fundamental issue is whether the analyst will use the measured similarity between sites as the tie weights or instead binarize the similarities in some way, creating an unweighted network in which ties are either present or absent. Peeples and Roberts (2013) suggested that even if such binarization helped in producing a legible graphical display of an assemblage network, it would be best to conduct analyses on the original weighted network data, not the binarized network. If analysis instead also included such a binarization step, it could affect the bootstrap's ability to assess sampling variability in network measures. Along with the generic impact on the bootstrap of the discretized data, binarized assemblage networks also will be much more prone to disconnectedness. If the network measures being employed cannot be calculated for disconnected networks, the analyst will need to decide whether it is appropriate to simply discard those bootstrap replications that resulted in a disconnected network.

Finally, note that the approach used here is that of the simple nonparametric bootstrap, based on direct resampling from the observed artifact assemblage. In the ceramic setting here, one consequence is that if a site had no sherds of a particular ware, it cannot have any in the resampled datasets either. This may suggest use of some parametric bootstrap approach instead; in the archaeological setting, this would allow all wares at least some probability of appearing at a given site in the bootstrap replicates, subject to logical constraints imposed by known

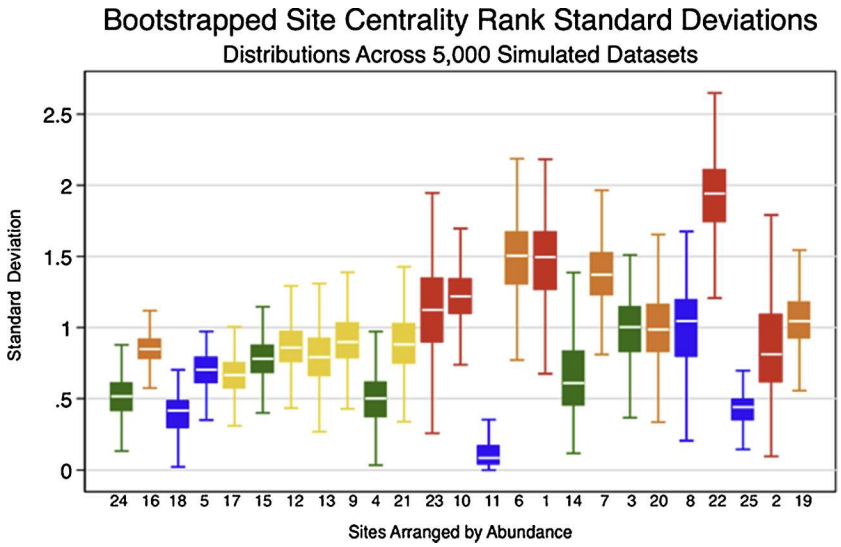

Fig. 6. Boxplots of bootstrapped estimates of standard deviations of site centrality rank.

Table 5
Comparison of true standard deviation of ranked centralities to bootstrap estimates.

Sites	True SD	Mean	IQR	True to Mean SD Ratio
1	1.6876	1.4564	0.3988	1.1587
2	0.7745	0.8933	0.4707	0.8670
3	1.0221	0.9784	0.3123	1.0447
4	0.4433	0.4996	0.2396	0.8872
5	0.7697	0.6892	0.1735	1.1167
6	1.6176	1.4852	0.3645	1.0892
7	1.2826	1.3914	0.2936	0.9218
8	1.1630	0.9717	0.3962	1.1969
9	0.8910	0.9129	0.2412	0.9760
10	1.3571	1.2138	0.2413	1.1181
11	0.0625	0.1210	0.1243	0.5165
12	0.8799	0.8634	0.2154	1.0192
13	0.7062	0.7952	0.2654	0.8880
14	0.4925	0.6639	0.3733	0.7417
15	0.8731	0.7702	0.1889	1.1335
16	0.9132	0.8417	0.1365	1.0849
17	0.5972	0.6579	0.1758	0.9077
18	0.4574	0.3760	0.1836	1.2163
19	0.9997	1.0498	0.2487	0.9523
20	0.8906	1.0039	0.3303	0.8872
21	0.9009	0.8857	0.2743	1.0172
22	2.2467	1.9154	0.3623	1.1730
23	1.1821	1.1154	0.4426	1.0598
24	0.4775	0.5066	0.1894	0.9426
25	0.4639	0.4123	0.1385	1.1250

dates for wares' production or use and sites' occupation. The empirical ware proportions for each site could be replaced by some smoothing that drew on theoretical ware distributions or information from ware distributions at other sites. While we do not pursue such an approach here, this may be especially appealing when sites' assemblage sizes are relatively modest. Such an extension may also be particularly useful for domains in which there is no analogue to the ware and site dates here, and therefore no logical reason why some category would necessarily have zero probability for a given actor. This sort of parametric bootstrap could be evaluated in a similar manner to our investigation of the nonparametric bootstrap.

Acknowledgements

This work was supported by (United States) National Science Foundation grants 1758606 and 1758690 to the University of Wisconsin-Milwaukee and Arizona State University. The Co-Editor and anonymous reviewers made helpful comments and suggestions on an earlier version of the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.socnet.2020.11.005.

References

Agresti, A., 2007. An Introduction to Categorical Data Analysis. John Wiley & Sons, Hoboken, NJ.
Basu, S., Maulik, U., Chatterjee, O., 2016. Stability of consensus node orderings under imperfect network data. IEEE Trans. Comput. Soc. Syst. 3, 120–131.
Bickel, P.J., Freedman, D.A., 1981. Some asymptotic theory for the bootstrap. Ann. Stat. 9, 1196–1217.
Birch, J., Hart, J.P., 2018. Social networks and Northern Iroquoian Confederacy dynamics. Am. Antiq. 83, 13–33.
Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, MA.
Bolland, J.M., 1988. Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks. Soc. Networks 10, 233–253.
Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. 2, 113–120.
Borck, L., Mills, B.J., Peeples, M.A., Clark, J.J., 2015. Are social networks survival networks? An example from the late pre-Hispanic US Southwest. J. Archaeol. Method Theory 22, 33–57.
Borgatti, S.P., 2005. Centrality and network flow. Soc. Networks 27, 55–71.
Borgatti, S.P., Carley, K.M., Krackhardt, D., 2006. On the robustness of centrality measures under conditions of imperfect data. Soc. Networks 28, 124–136.
Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G., 2009. Network analysis in the social sciences. Science 323, 892–895.
Brainerd, G.W., 1951. The place of chronological ordering in archaeological analysis. Am. Antiq. 16, 301–313.
Brandes, U., Robins, G., McCrainie, A.N.N., Wasserman, S., 2013. What is network science? Netw. Sci. 1, 1–15.
Brughmans, T., Peeples, M.A., 2017. Trends in archaeological network research: a bibliometric analysis. J. Hist. Network Res. 1, 1–24.
M.R. Chernick . Bootstrap Methods: A Guide for Practitioners and Researchers, Wiley Interscience Hoboken, NJ 2007.
Clark, J.J., Lyons, P.D., 2012. Migrants and Mounds: Classic Period Archaeology of the Lower San Pedro Valley. Anthropological Papers 45. Archaeology Southwest, Tucson, AZ.
Collar, A., Coward, F., Brughmans, T., Mills, B.J., 2015. Networks in archaeology: phenomena, abstraction, representation. J. Archaeol. Method Theory 22, 1–32.
Costenbader, E., Valente, T.W., 2003. The stability of centrality measures when networks are sampled. Soc. Networks 25, 283–307.
Davis, C., Kahan, W.M., 1970. The rotation of eigenvectors by a perturbation. III. SIAM J. Numer. Anal. 7, 1–46.
Duncan, O.D., Duncan, B., 1955. Residential distribution and occupational stratification. Am. J. Sociol. 60, 493–503.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall/CRC, New York, NY.

Eldridge, J., Belkin, M., Wang, Y., 2018. Unperturbed: spectral analysis beyond Davis-Kahan. Proc. Machine Learning Res. 83, 321–358.
Epskamp, S., Borsboom, D., Fried, E.I., 2018. Estimating psychological networks and their accuracy: a tutorial paper. Behav. Res. Methods 50, 195–212.
Gjesfjeld, E., 2015. Network analysis of archaeological data from hunter-gatherers: methodological problems and potential solutions. J. Archaeol. Method Theory 22, 182–205.
Golitko, M., Feinman, G.M., 2015. Procurement and distribution of pre-Hispanic Mesoamerican obsidian 900 BC–AD 1520: a social network analysis. J. Archaeol. Method Theory 22, 206–247.
Golitko, M., Meierhoff, J., Feinman, G.M., Williams, P.R., 2012. Complexities of collapse: the evidence of Maya obsidian as revealed by social network graphical analysis. Antiquity 86, 507–523.
Habiba, Athenstädt, Jan, C., Mills, B.J., Brandes, U., 2018. Social networks and similarity of site assemblages. J. Archaeol. Sci. 92, 63–72.
Hall, P., Miller, H., 2009. Using the bootstrap to quantify the authority of an empirical ranking. Ann. Stat. 37, 3929–3959.
Hart, J.P., Engelbrecht, W., 2012. Northern Iroquoian ethnic evolution: a social network analysis. J. Archaeol. Method Theory 19, 322–349.
Hart, J.P., Birch, J., St-Pierre, C.G., 2017. Effects of population dispersal on regional signaling networks: an example from northern Iroquoia. Sci. Adv. 3, e1700497.
Heeren, A., Bernstein, E.E., McNally, R.J., 2018. Deconstructing trait anxiety: a network perspective. Anxiety Stress Coping 31, 262–276.
Kim, P.J., Jeong, H., 2007. Reliability of rank order in sampled networks. Eur. Phys. J. B 55, 109–115.
Knappett, C., 2011. An Archaeology of Interaction: Network Perspectives on Material Culture and Society. Oxford University Press, Oxford.
Lulewicz, J., 2019. The social networks and structural variation of Mississippian sociopolitics in the southeastern United States. Proc. Natl. Acad. Sci. 116, 6707–6712.
Lusher, D., Koskinen, J., Robins, G., 2012. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications. Cambridge University Press, New York, NY.
Mammen, E., 1992. When Does Bootstrap Work?: Asymptotic Results and Simulations. Springer, New York, NY.
Manly, B.F., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman and Hall, New York, NY.
Mills, B.J., 2016. Communities of consumption: cuisines as constellated networks of situated practice. In: Roddick, A.P., Stahl, A.B. (Eds.), Knowledge in Motion: Constellations of Learning Across Time and Place. University of Arizona Press, Tucson, AZ, pp. 247–270.
Mills, B.J., 2017. Social network analysis in archaeology. Annu. Rev. Anthropol. 46, 379–397.
Mills, B.J., Clark, J.J., Peeples, M.A., Haas Jr., W.R., Roberts Jr., J.M., Hill, J.B., Huntley, D.L., Borck, L., Breiger, R.L., Clauset, A., Shackley, M.S., 2013a. Transformation of social networks in the late pre-Hispanic US Southwest. Proc. Natl. Acad. Sci. 110, 5785–5790.
Mills, B.J., Roberts Jr., J.M., Clark, J.J., Haas Jr., W.R., Huntley, D., Peeples, M.A., Borck, L., Ryan, S.C., Trowbridge, M., Breiger, R.L., 2013b. The dynamics of social networks in the prehispanic US Southwest. In: Knappett, C. (Ed.), Network Analysis in Archaeology: New Approaches to Regional Interaction. Oxford University Press, Oxford, pp. 181–202.
Mills, B.J., Peeples, M.A., Haas Jr., W.R., Borck, L., Clark, J.J., Roberts Jr., J.M., 2015. Multiscalar perspectives on social networks in the late prehispanic Southwest. Am. Antiq. 80, 3–24.
Mills, B.J., Clark, J.J., Peeples, M.A., 2016. Migration, skill, and the transformation of social networks in the pre-Hispanic Southwest. Econ. Anthropol. 3, 203–215.
Mills, B.J., Peeples, M.A., Aragon, L.D., Bellorado, B.A., Clark, J.J., Giomi, E., Windes, T. C., 2018. Evaluating Chaco migration scenarios using dynamic social network analysis. Antiquity 92, 922–939.
Östborn, P., Gerding, H., 2014. Network analysis of archaeological data: a systematic approach. J. Archaeol. Sci. 46, 75–88.
Pailes, M., 2014. Social network analysis of early classic Hohokam corporate group inequality. Am. Antiq. 79, 465–486.
Peeples, M.A., 2019. Finding a place for networks in archaeology. J. Archaeol. Res. 27, 451–499.
Peeples, M.A., Haas Jr., W.R., 2013. Brokerage and social capital in the prehispanic U.S. Southwest. Am. Anthropol. 115, 232–247.
Peeples, M.A., Roberts Jr., J.M., 2013. To binarize or not to binarize: relational data and the construction of archaeological networks. J. Archaeol. Sci. 40, 3001–3010.
Peeples, M.A., Mills, B.J., Haas Jr., W.R., Clark, J.J., Roberts Jr., J.M., 2016. Analytical challenges for the application of social network analysis in archaeology. In: Brughmans, T., Collar, A., Coward, F. (Eds.), The Connected Past: Challenges to Network Studies in Archaeology and History. Oxford University Press, Oxford, pp. 59–84.
Roberts Jr., J.M., Liedka, R.V., 1999. On summary measures of binarized dominance data. Soc. Networks 21, 23–35.
Roberts, A.I., Chakrabarti, A., Roberts, S.G., 2019. Gestural repertoire size is associated with social proximity measures in wild chimpanzees. Am. J. Primatol. 81, e22954.
Roberts Jr., J.M., Mills, B.J., Clark, J.J., Haas Jr., W.R., Huntley, D.L., Trowbridge, M.A., 2012. A method for chronological apportioning of ceramic assemblages. J. Archaeol. Sci. 39, 1513–1520.
Robinson, W.S., 1951. A method for chronologically ordering archaeological deposits. Am. Antiq. 16, 293–301.
Rosvall, M., Bergstrom, C.T., 2010. Mapping change in large networks. PLoS One 5, e8694.
Segarra, S., Ribeiro, A., 2015. Stability and continuity of centrality measures in weighted graphs. IEEE Trans. Signal Process. 64, 543–555.
Silk, M.J., Jackson, A.L., Croft, D.P., Colhoun, K., Bearhop, S., 2015. The consequences of unidentifiable individuals for the analysis of an animal social network. Anim. Behav. 104, 1–11.
Slocum-Gori, S.L., Zumbo, B.D., 2011. Assessing the unidimensionality of psychological scales: using multiple criteria from factor analysis. Soc. Indic. Res. 102, 443–461.
Smith, J.A., Moody, J., Morgan, J., 2017. Network sampling coverage II: the effect of non-random missing data on network measurement. Soc. Networks 48, 78–99.
Sosna, D., Galeta, P., Šmejda, L., Sladek, V., Bruzek, J., 2013. Burials and graphs: relational approach to mortuary analysis. Soc. Sci. Comput. Rev. 31, 56–70.
Weidele, D., van Garderen, M., Golitko, M., Feinman, G.M., Brandes, U., 2016. On graphical representations of similarity in geo-temporal frequency data. J. Archaeol. Sci. 72, 105–111.
Yu, Y., Wang, T., Samworth, R.J., 2015. A useful variant of the Davis–Kahan theorem for statisticians. Biometrika 102, 315–323.