Social Networks and Similarity of Site Assemblages

Habiba[1*] Jan C. Athenstädt[1] Barbara J. Mills[2] Ulrik Brandes[1]

[1]University of Konstanz, Germany [2]University of Arizona, USA

Abstract

There have been a number of similarity measures developed in a variety of research domains. Generally, these measures are developed for a specific context and later reused in other contexts and applications, depending on their ease of use and perceived applicability. While there might be statistical reasons to use a particular similarity index, the results of other measures should be taken into account as well, as various similarity measures do not necessarily have similar contextual meaning. Two entities can be very similar with respect to a certain similarity criterion but may be very distinct in terms of another. Thus, an understanding of the mathematical logic behind a method is crucial to the interpretation of the resulting network of similarities. We review a number of methods from the literature, for constructing similarity networks among disparate entities, regarding their applicability on data from archaeological sites. Formally, given an $N \times p$ matrix of N entities with p distinct classes of attributes, how are the entities comparable to each other with respect to the kinds of attributes they share? We distinguish three qualitatively different families of similarity measures for deducing relationships among entities that may meaningfully map onto various distinct social phenomena, such as migration, material acquisition, and movement of goods and skills, among others. Entities can be compared based on: *(a)* non–uniform weighting of attributes, *(b)* asymmetric dominance relationships, and *(c)* rank correlations. We ground the significance and distinction of these classes of measures by giving comparative and contextual examples of selected methods on a case study of archaeological collections pertaining to AD 1200-1500 from the US Southwest region. We attempt to elucidate the differences in outcomes and their meanings when choosing various similarity methods for comparing disparate entities.

1 Introduction

Researchers in the field of archaeology generally have to rely on sparse and fragmented information to understand the social behavior of the populations under study. Given the material discovered at different archaeological sites, one – but certainly not the only – way of estimating the strength of a relationship between them is by evaluating how "similar" they are to each other. Calculating pairwise similarities between site assemblages results in a network that can be seen as a proxy for social interactions and has become one popular basis for analyzing social networks in archaeology (e.g. [Hart and Engelbrecht, 2012, Mills et al., 2015, Mills et al., 2013b, Munson, 2013]).

Measuring similarity among entities is one of the most applied techniques in multivariate data analysis. Yet, similarity in and of itself has yet to be concisely defined. A simple – and slightly circular – definition of it, "is a numerical measure of the degree to which two data objects are alike" [Tan et al., 2005]. What makes two entities "alike" can vary depending on what the data represents, the type of attributes, and how the attributes are compared. In general, two entities are similar if they share many categorical attributes, or if the values of their numerical attributes are relatively "close". Dissimilarity – the complement of similarity – especially distance measures, are also frequently been used to compare entities. There have been a number of similarity/dissimilarity measures developed in a variety of domains, such as, natural language processing,

*Corresponding author. Email: habiba@uni-konstanz.de
Postal address: Department of Computer & Information Science, University of Konstanz, PO Box 67, 78457 Konstanz, Germany.

information retrieval [Manning et al., 2008, Mihalcea et al., 2006, Santini and Jain, 1999], computational biology [Heringa, 2001, Song et al., 2008], and cluster analysis [Balcan et al., 2008, Strehl et al., 2000, Tan et al., 2005], among others. Most of these measures are grounded in theoretical justifications for various distinctive types of comparisons that do not necessarily have similar contextual meanings. That is, two entities can be very similar with respect to a certain measure but may be very distinct in terms of another similarity index. This is one of the downsides of having an abundance of such methods. Many of them seemingly estimate the same general concept yet are operationalized by different procedures and on different bases. As a consequence, the results they generate, may not bear a clear correspondence to the abstract concept of similarity that they are meant to mimic.

Application of network methods in archaeology has increased considerably in the last decade [Brughmans, 2010, Collar et al., 2015]. Knappett [Knappett, 2013] provides a comprehensive state-of-the-art guide to the main themes and approaches of network analysis for archaeologists. Trends of migrations and movements [Mills, 2011, Mills et al., 2013a, Mills et al., 2013b], exchange of ideas and diffusion of technology [Golitko and Feinman, 2015, Östborn and Gerding, 2014], intra–community social and political dynamics [Munson, 2013, Munson and Macri, 2009, Scholnick et al., 2013, Paris, 2014], and transformation of social landscapes over different social and temporal scales [Mills, 2007] are some of the topics network methods have been used to address in archaeology. In recent years, multiple studies have been published on the reconstruction of networks of similarities, based on the production, consumption, and deposition of ceramic assemblages, most notably in the geographic region of the US Southwest during the late Pre–Hispanic period period [Borck et al., 2015, Mills, 2007, Mills, 2011, Mills et al., 2013a, Mills et al., 2013b, Mills et al., 2015, Peeples and Roberts, 2013]. Using the Brainerd–Robinson (BR) index [Brainerd, 1951, Robinson, 1951], networks are reconstructed that are based on similarities of consumption of ceramics among the settlements at various spatial and temporal scales. This network view of site similarities provides a supplemental approach in systematically exploring the social, political, and economical patterns of interaction among settlements in the region during that period. In other areas of the world, the BR index has also become a common way for comparing assemblages and assessing similarity including Mesoamerica [Golitko, 2015, Golitko et al., 2012, Golitko and Feinman, 2015] and the Northeast North America [Hart, 2016, Hart and Engelbrecht, 2012].

In this work we selectively review some of the more frequently used similarity measures from the literature in relation to specific concepts in archaeology. Such an approach has been outlined by Östborn and Gerding [Östborn and Gerding, 2014]. We compare these similarity measures to the BR index, which is currently most widely used in archaeological research. We argue that it is crucial to choose a method that corresponds to the specific research question and show that it is important to use and compare multiple methods. This can lead to a more nuanced picture of the historical and social contexts being explained by the type of proxy data used to represent social interactions of different kinds. Lastly, we apply some of the proposed methods to the dataset from the US Southwest that was used in [Mills et al., 2013a] and compare the resulting networks.

In Table 1 we list a set of measures that we use as a base for the methods proposed in this paper. A comprehensive survey on similarity/dissimilarity measures can be found in [Choi, 2008, Choi et al., 2010, Everitt and Rabe-Hesketh, 1997].

2 Proposed methods

This work focuses on the following aspects of constructing similarity networks. First, we give an overview of methods that convert an $N \times p$ multivariate matrix of N entities represented by p attributes into an $N \times N$ similarity matrix in Section 2.1. In Section 2.2, we propose a transformation that assigns variable weights to attributes based on their assumed significance before the application of a similarity measure. Finally, in Section 2.3 we outline the approach to reconstruct cross-temporal networks of similarities. Table 2 lists the notations used in the following sections.

Sr.	Method	Informal Description	Key Characteristics
1	BR [Brainerd, 1951, Robinson, 1951]	aggregate of differences in proportions of attributes	compares differences weighted by the diversity of attributes in an entity
2	Cosine [Tan et al., 2005]	dot product of two entities normalized by the product of their magnitudes	measures the difference in the orientation of two entities
3	Euclidean [Everitt and Rabe-Hesketh, 1997]	aggregate of differences of attributes	straight line distance between two entities in a euclidean space
4	Jaccard [Jaccard, 1912]	ratio of the number of matched to the number of all non–zero attributes	compares the size of the set of attributes common between two entities to the size of the set of all non–trivial attributes of the two entities
5	l out of k [Nick et al., 2013]	if k of the top l ranked attributes match (binary)	compares attributes by their assigned ranks
6	Simple Matching [Segaran, 2007]	ratio of the number of matched to the number of all attributes	compares shared attributes to all possible attributes

Table 1: An overview of the basic methods underlying the similarity measures proposed in this paper

Notation	Definition
x or y	labels for distinct entities
N	number of entities
p	number of attributes
S	the set of p attributes
Y	the $N \times p$ multivariate matrix of N entities with p attributes/features.
$x_{i,j}$	value of the $j-th$ attribute of the $i-th$ entity
S_x	subset of S with non-zero values for entity x
V_x	binary vector of length p denoting the presence/absence of each attribute for entity x
Q_x	vector of length p denoting the value of each attribute for entity x
R_x	vector of length p denoting the value of each attribute for entity x sorted in a rank ordering

Table 2: Notations and terminology

2.1 Similarity measures

In the following we are giving detailed descriptions for the selected similarity measures we use to reconstruct networks of interactions.

Dominance relationship: An entity x dominates an entity y, if and only if, $S_y \subseteq S_x$.

Relationships among groups of people in a geographically proximal setting are not necessarily symmetric. For example, there are power, status, resources, and economic disparities that result in asymmetric dynamics among participating entities. In many such cases, the relationship can be more logically contextualized as supplier–consumer, source–sink, or political dominant–subordinate relations. The dominance relationship captures the most basic form of such an imbalanced relation among entities. Mathematically, it encodes the partial order relation among a set of entities. This method can be further refined into binarized and non–binarized dominance.

Binarized Dominance: An entity x dominates entity y if it contains all attributes of y.

$$\text{Dominance}_1(x, y) = \begin{cases} 1 & \text{if } S_y \subseteq S_x \\ 0 & \text{otherwise} \end{cases}$$

Non–binarized Dominance: An entity x dominates entity y if each attribute of x is quantitatively greater than the corresponding attribute of y.

$$\text{Dominance}_2(x, y) = \begin{cases} 1 & \text{if } Q_{x,i} > Q_{y,i} \forall i \in [1, p] \\ 0 & \text{otherwise} \end{cases}$$

Brainerd–Robinson (BR) index: The Brainerd–Robinson index compares the similarity in the proportions of values of attributes.

$$BR(x, y) = 2 - \sum_{i=1}^{p} \left| \frac{x_i}{\sum_{i=1}^{p} x_i} - \frac{y_i}{\sum_{i=1}^{p} y_i} \right|$$

This measure is specifically developed in archaeology for comparing archaeological assemblages in terms of the proportions of types of objects or other such categorical data [Brainerd, 1951, Robinson, 1951]. In this work we normalize this measure to a 0 to 1 scale.

Matching coefficient: The matching coefficient is the size of the intersection of non–trivial attributes of a pair of entities.

$$\text{Match}(x, y) = |S_x \cap S_y|$$

One of the most obvious and simple methods for gauging exchange (of artifacts) or shared ideology (cultural practices) among disparate sites is by measuring their overlap in terms of number of distinct types of artifacts found, i.e. what are common features among the two sites. Here, we do not take into account the quantitative differences in the attributes. The matching coefficient is a non–normalized version of the *Simple Matching Coefficient* [Shennan, 1997]. The matching coefficient can be extended to the k*–Common* method. That is, whether there are at least k common attributes between a pair of entities.

Difference in matches and mismatches: Difference in matches and mismatches is the quantitative difference between the shared and the mutually exclusive attributes counts.

$$\text{Match-Mismatch}(x, y) = |S_x \cap S_y| - |S_x \Delta S_y|$$

One drawback of the previous method is that it only looks at the common attributes while ignoring the attributes in which the entities differ. However, there may be cases where the differences in entities are as significant as their commonalities. In this method, we suggest setting the commonalities among entities by differences among them to get a more nuanced sense of the degree of similarity among them. Note that if we look at the differences only, this measure reduces to the *Hamming distance* [Hamming, 1950].

Jaccard coefficient: The ratio of the number of common to all unique non–zero attributes for a pair of entities [Jaccard, 1912].

$$\text{Jaccard}(x, y) = \frac{|S_x \cap S_y|}{|S_x \cup S_y|}$$

This is arguably the most popular method to compare entities on binary attribute values, where co–presence of attributes means there is a positive correlation between the entities, and absence of attributes in one or the other has a negative effect on similarity. In the case of archaeological data, which in most cases is sparse, the Jaccard coefficient is especially useful as it ignores all the attributes that are mutually null for the entities being compared.

Cosine Coefficient Similarity: The Cosine Coefficient is the inner product of two attribute vectors normalized by the product of their ℓ_2-norms [Tan et al., 2005].

$$\text{Cosine similarity}(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Cosine similarity measures the angle between the orientation of two entities irrespective of their magnitudes.

Euclidean distance: The Euclidean distance of two entities is the square root of the sum of the pairwise differences between the values of their corresponding attributes [Everitt and Rabe-Hesketh, 1997].

$$\text{Euclidean distance}(x, y) = \sqrt{\sum_{i=0}^{p} (Q_{x,i} - Q_{y,i})^2}$$

$$\text{Euclidean closeness}(x, y) = \frac{1}{1 + Euclidean\ distance(x, y)}$$

This is the straight line distance between a pair of entities in euclidean space. If sites are assumed to be points in a multi–dimensional euclidean space, where each dimension represents an attribute, the length of the straight line between a pair of points portrays their dissimilarity. This measure is highly influenced by low magnitude of the attributes, as the distance to the origin of the coordinate system is smaller, leading to smaller distances to other sites with low magnitudes in attribute space.

l-out-of-top-k: This similarity method determines whether two entities match in at least l of their top k attributes, where l and k can be chosen arbitrarily [Nick et al., 2013]. It requires the attributes of each entity to be ranked based on some criteria.

$$\text{l-out-of-k}(x, y) = \begin{cases} 1 & if \ \ |V_R^x[1:l] \cap V_R^y[1:l]| \geq k \\ 0 & otherwise \end{cases}$$

One straight forward way to rank attributes is by their magnitudes, if the attributes are mutually comparable. Contextual knowledge can help in making more nuanced choices of the ranking of attributes.

This and the following method are effectively applicable in situations where attributes can be weighted by their relative importance with respect to each other. For a comprehensive understanding of alternative approaches for comparing "top–k" lists, see [Fagin et al., 2003].

Maximum Quasi–Jaccard: Maximum Quasi-Jaccard is the maximal ratio of matched attributes with respect to their ranks to all attributes in the list.

$$\text{Quasi-Jaccard}(x, y) = \arg\max_k \frac{|V_R^x[1:k] \cap V_R^y[1:k]|}{|V_R^x[1:k] \cup V_R^y[1:k]|}$$

This is the non–parameterized version of the previous method. It incrementally compares the ranked attribute lists of two entities and finds the maximal possible match with respect to the size of the lists compared so far [Nick et al., 2013]. Essentially, this index incrementally measures the Jaccard coefficient of the ordered list of attributes, of each size, for a pair of sites. It then picks the k that gives the highest match ratio, where $0 \leq k \leq p$. Measuring similarity in ranked lists is a well–studied problem in information retrieval [Webber et al., 2010] and other fields.

2.2 Non-uniform significance of wares

The similarities we have discussed so far are based on the general principle of uniform weighting of attributes of entities. This is arguably the most straight-forward way to construct such similarities. However, if it is a priori known that the attributes are not of equal importance, any measure can be further refined by placing non–uniform emphasis on the attributes. Since certain types of objects might have had a higher relevance, we can adjust the weights of the attributes based on their occurrences.

Here we use the concept of *Term Frequency – Inverse Document Frequency* (tf–idf) transformation from the field of text mining. It is used to capture the significance of individual words in a document based on their frequencies [Manning et al., 2008]. The idea behind this transformation is that the importance of a word increases proportional to the number of its occurrence in a document. However, this importance is offset by the overall occurrence of the word in the entire corpus of documents. This helps to control for the importance of very common or rarely occurring words. Once all the words in a document are ranked based on this measure, similarities among the documents can be established through any similarity index. One of the most commonly used measures of similarity in case of tf–idf is the *cosine similarity* (Section 2.1).

The tf–idf transformation is built as a product of two statistics: the *term frequency* offset by the *inverse document frequency*. The term frequency is simply the count of occurrence of a term in a document. The inverted document frequency is the inverse of the overall prevalence of the term in the entire corpus of documents.

$$\text{tf–idf}(t, D) = tf(t, D) \times idf(t) = \text{No. of occurance of t in D} \times \log\left(\frac{\text{No. of documents}}{\text{No. of documents containing t}}\right)$$

where D is a document and t is a term.

We can use this transformation for comparing sites, analogous to documents, based on their types of wares, analogous to words. High values of this index are achieved with high counts of the wares but low occurrences of the wares across sites. Thus, this measure puts less emphasis on wares that occur commonly across many sites. In archaeological terms, this means that wares that are either very rare or very frequent are weighted down to reduce their overall effect. It should be noted that this method may be counterproductive if the categories chosen for the analysis of similarity have already been pre-selected because of their significance. For example, ubiquitous cooking vessels in ceramic assemblages may already be omitted in assemblage comparisons, choosing instead the decorated service wares.

2.3 Across-Time Comparison

Migration, movement, technological transmission, and exchange patterns are inherently time related social interactions. Therefore, it is necessary to compare similarities across time. This helps to explain a number of social phenomena. For example, entities that are self-similar over time say something about their isolation and lack of innovation and transformation with time. Entities that change over time by emulating and adopting objects and adapting practices from other entities depict both the dynamics of the group within the entity as well as its practical relations with other entities. In an archaeological sense these chronological comparisons can be very useful in corroborating the migration and movement theories associated with certain regions in particular time periods and/or the adoption of innovations across areas.

We use each of the similarity measures to compare sites in different time periods. Thus, we look at how similar the assemblages of sites are across time, both within a site and between different sites. This could be an indicator for the spatial propagation of certain cultural features, be it through the exchange of ideas, goods, migration, transformation of practices in a community, or another social process.

3 Case-Study: Application to US Southwest Data

We applied the proposed methods to the *Southwest Social Networks (SWSN) database* [Mills et al., 2013a, Mills et al., 2015, Mills et al., 2013b, Peeples and Roberts, 2013]. This dataset contains aggregate information on about 4.3 million ceramic artifacts found at more than 700 archaeological sites in a $334,000\,km^2$ region in the US Southwest between AD 1200 and 1500. These ceramics are classified into 42 different categories of decorated wares, for which the approximate duration of usage is known. Wares are defined by shared technological attributes and as defined in the Southwest they have geographical meaning. Within wares are sequences of ceramic types based on a number of attributes including form and surface treatment, such as painted designs that are temporally sequent. The time span has been divided into five consecutive time periods, of 50 years each. Within each period, for each site with more than 30 decorated sherds, a vector of the number of sherds of each ware is given. The finds are attributed to the periods assuming a normal distribution of popularity over the duration of usage. The number of sherds assigned to a 50-year period corresponds to the proportion of the distribution that falls into that period (see [Mills et al., 2013a] for details).

In [Mills et al., 2013a], Mills and her colleagues investigate shared consumption patterns of decorated wares among sites. Some of these patterns are conditioned by production but the overarching similarity among sites is measured though the shared use and discard of ceramics. In their work, Mills et al. apply the BR index as a similarity measure among the sites to investigate the archaeological hypothesis of demographic change in the area. They normalize the index to values between 0 and 1 for each pair of sites. For the network analytical measures in their study, they use the weighted network, but introduce a cutoff value of 0.75 (on the normalized scare) to visualize a binary network for each of the five 50-year-periods, i.e., an edge between two sites exists, if their BR index is greater than 0.75.

We have replicated this study and use selected examples to compare the various similarity measures. In order to make networks comparable, we chose the cutoff values in a way that the number of edges (i.e., the density of the network) is comparable among the networks we examine. Since the cutoffs are arbitrary, we need to control for the size of network to be able to meaningfully compare various similarity measures.

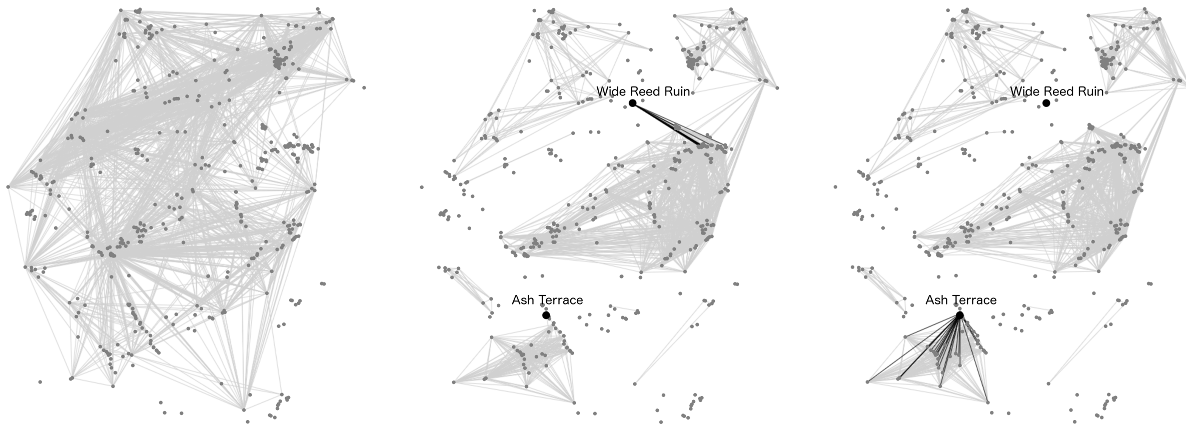
The following examples are meant to highlight the conceptual differences in the measures and briefly offer some insights into the social processes that these measures may provide.

3.1 Comparison of Selected Similarity Measures

In this section, we compare how sensitive similarity measures are to the diversity and quantities of wares on various sites.

Common Attributes Figure 1(a) depicts the similarity network of sites based on the *k-Common* wares method for AD 1200-1250. Recall that this method reproduces a network of links among sites that have at least k common wares. We set $k = 3$ for this analysis. Note that larger values of k results in non-normalized version of the jaccard index. We compare the network generated by *k-Common* method in Figure 1(a) to the network based on the BR index, reproduced in Figure 1(b) for the same time period and keeping the density ($\sim$ 2591 links) of the two networks comparable.

The structure of the *k-Common* network is radically different from the BR network. Instead of the clear local cluster structure created by the BR index, the *k-Common* network consists of many long distance links and a few nodes with a high number of links (the maximum degree of a node in the *k-Common* network is 100 and there are 39 nodes with a degree higher than 45, which is the maximum in the BR network). One reason for this is, that the *k-Common* method rewards sites that are similar in their diversity. While sites with less than three different types of wares – which are more likely to have high BR values– will not have any links. That is, very diverse yet similar sites have a higher chance of being connected. The high number of links of rather long distance in this example indicates that – despite the separate communities indicated by the BR index – a substantial number of sites with high diversity might have links to other geographically distant sites.



(a) Common Attributes (2591 links) (b) Brainerd-Robinson (2598 links) (c) BR with tf–idf (2598 links)

Figure 1: Symmetric networks for AD 1200-1250 with and without tf–idf

In archaeological terms, however, this may not always be desirable. For example, there can be sites that have only a few occurrences of many different wares. In general, this does not get at communities of practice like the BR index does, rather it might be highly biased towards very small scale exchanges since a single pot (or even a single sherd) from a distant source could help create a link. In typical archaeological context, it is not always about consumption but about sets of presence/absence categories. *k-Common* method does not capture consumption except by presence/absence and will therefore not be as useful for interpreting differences in communities of practice that depend on redundant use and discard [Mills et al., 2015]. The *k-Common* method is one extreme that indicates all possible material connections, but can be influenced by very small samples.

tf-idf Figure 1(c) shows the BR network from AD 1200-1250 with wares weighted by the tf-idf values. Although the macro–level structure of the tf-idf based network visually looks similar to the BR network from Figure 1(b), there are micro–level differences that allude to alternative properties of the two weighting schemes. For example, the *Wide Reed Ruin* site in the central northern area gets completely disconnected in the tf-idf network whereas the otherwise unconnected site of *Ash Terrace* in the south gains a considerable amount of new connections. It is worth noting here that this presence/absence of links is not the direct result of thresholding. Recall that for comparing different methods we control for the network density only. That is, the networks are generated by selecting approximately same number of top ranked edges based on weights assigned by the method. Hence, in the above example, *Ash Terrace* may acquire more links at a lower threshold. However globally those edges are not highly ranked and thus drop out in the BR network sooner than the tf-idf weighted BR network.

Wide Reed Ruin shared proportionally comparable amount of `Cibola White Ware` and `Early White Mountain Red Ware` with its neighbors towards the east, which results in strong BR index based links. However, these two types of wares seem to be highly common overall during that period. Thus, the tf-idf weighting renders them less significant compared to other infrequent types, which results in lower weight links between *Wide Reed Ruin* and other sites with which it shared this ware type.

Ash Terrace has proportionally similar amounts of `Cibola White Ware` and `Tucson Basin Brown Ware` to those of *Big Pot* and *Flieger* sites. These three sites have exactly the same two types of wares in proportionally similar, albeit quantitatively different, amounts. Resulting in very high BR values between *Ash Terrace* and the other two sites, such that it is exclusively connected to just these two sites. On the other hand with the tf-idf weighting, *Ash Terrace* gains a remarkable number of connections. This is primarily due to the `Tucson Basin Brown Ware` which seems to have been typical to only a small, regionally confined, part of the studied area, and therefore gets a higher ranking by the tf-idf transformation.

Comparison of *Wide Reed Ruin* and *Ash Terrace* for BR with and without the tf-idf transformation is an instructive example. The tf-idf weighting resets the significance of wares relative to their abundance in overall region before applying BR or another similarity index for site comparison.

BR Index In addition to resulting in stronger links among less diverse sites, shown in example from Figures 1(a) and 1(b), the BR measure is less sensitive to the actual quantities of the different types of wares. The *Griffen Wash Complex* and *Indian Point Complex* sites (Figure 2) share seven different types of wares in the period of AD 1200-1250, in slightly varying quantities. However, they have a BR value of less than 0.65. On the other hand, sites with only one type of ware have the maximum BR value to sites that have only and exactly the same type of ware irrespective of the quantities. Hence, more diverse sites are inherently penalized by the BR index.

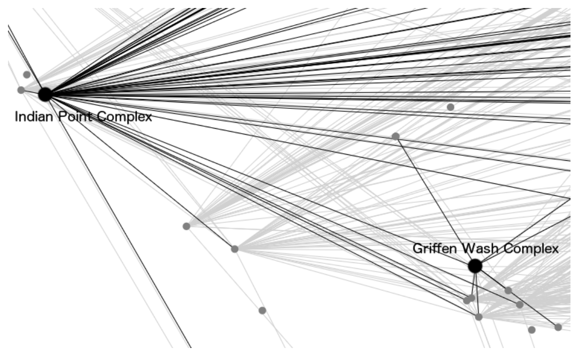
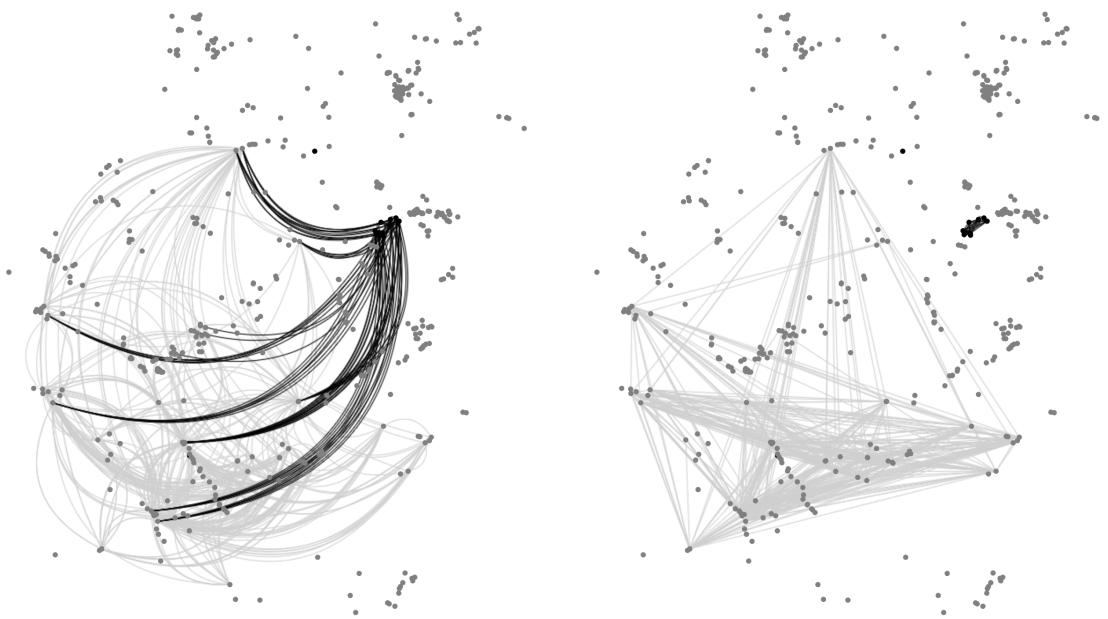


Figure 2: Sites *Griffen Wash Complex* and *Indian Point Complex* in the BR-network (AD 1200-1250)

3.2 Directed Methods

In Figure 3, we compare the asymmetric similarity method of (strict) Binary Dominance to the symmetric BR index in the AD 1400-1450 period. Recall that a site dominates another if all of the wares at the latter site are also present at the former. Additionally, it is strict dominance, if the former site has at least one type of ware not found at the latter site. In order to visualize the networks generated by the directed methods, we chose to draw the edges in a clockwise bending curve, from the dominating site to the dominated one.

Figure 3(a) shows that some north-eastern sites, especially from the small densely connected cluster – referred to as the *Zuni* sites – dominate large parts of the network, indicating high level of ware diversity on these sites. All these sites have at least four different wares. This relation is not visible in the BR network in Figure 3(b), even though the cutoff of 0.53 resulting in the 496 edges is rather low. Historically, the *Zuni* region is one of the few northern regions not to have been depopulated during the migration towards the south in the late thirteenth century [Mills, 2007]. It acts as a melting pot, adopting wares from different regions, which is captured by this dominance relation. The term *dominating* must be interpreted cautiously here, though. Depending on the archaeological context, it could mean a number of different social phenomena, including migration, which is the basis of one of the hypothesis about the *Hawikku* site [Mills, 2007]. There can be a number of other issues of equifinality– for example, pilgrimages or other (multiple) kinds of interactions that could be the potential explanation for the diversity and relative stability of these sites [Peeples and Haas, 2013]. Also, it should be pointed out that non–producers might import everything for other reasons. Hence, this dominance relation can capture a variety social trends.



(a) Binary Dominance (right bending edges indicat dominant to dominated relation)

(b) Brainerd-Robinson

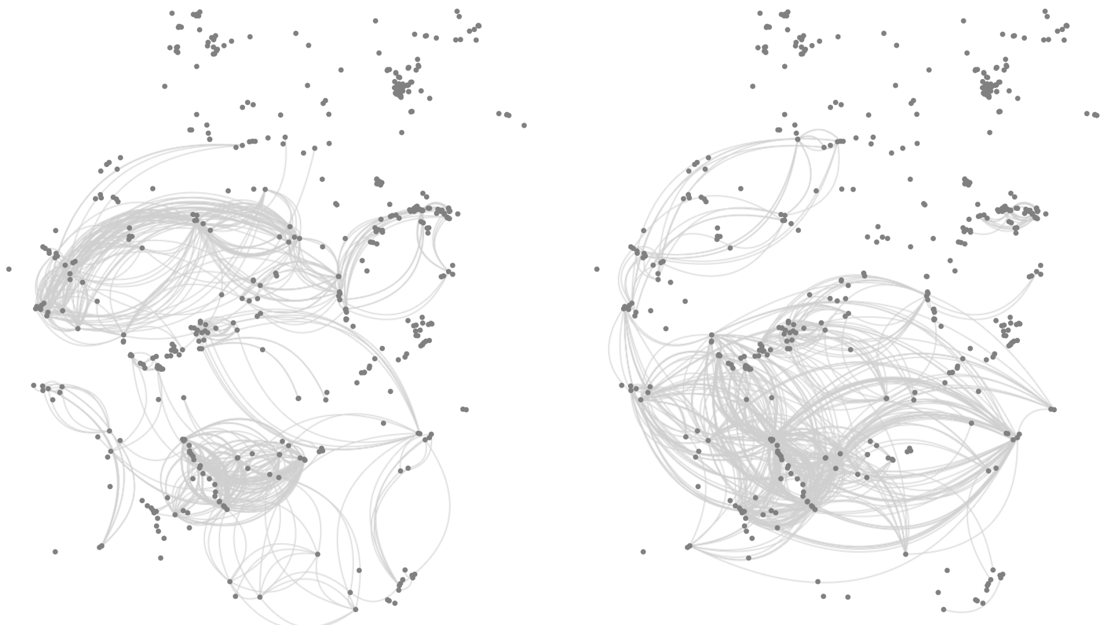
Figure 3: Directed and symmetric networks for AD 1400-1450 (both with 496 links)

3.3 Across-Time Comparison

In order to compare sites between different time periods (for example as an indicator for migration), we generated directed networks by calculating the similarity of sites from earlier to later period. For ease of readability, we limit our examples to symmetric measures. While across-time comparison is not limited to these, directed methods would generate four different values per dyad, making visualizations of the networks challenging.

Figure 4 shows the outcome of two similarity measures, between the periods AD 1300-1350 and AD 1350-1400. The links are clockwise-bend from the earlier to the later period. Figure 4(a), depicts differences in matches and mismatches of attributes with a cutoff value of 2. The network conveys two general patterns. In the north, many links are directed from the western sites towards the more central northern area. In the south, there is a concentration of links towards a chain of sites located along the *San Pedro River*. One explanation for this increasing similarity is due to the shared practice of production and consumption of wares in the southern Southwest region. Mills et al. already observed a progressively higher connectedness in these areas starting in the fourteenth century [Mills et al., 2015]. Deeper exploration of the data from these two consecutive periods reveal that sites in the west shared a high number of wares with the sites to the west of them in the later period. The major difference is that in the AD 1300-1350 period, sites in the river valley have two wares (`Cibola White Ware` and `Tucson Basin Brown Ware`) that are not present in the next period, leading to lower similarity scores for the links pointing away from the river. This observation leads to the open question, whether the disappearance of wares on a site can be an indicator of migrations away from that site to other sites with less diversity or the opposite?

In the BR network, Figure 4(b), with a comparable number of edges, those patterns are not visible. This again suggests that it is important to apply multiple methods of similarity and see how they might help to answer specific research questions.



(a) Match – Mismatch (792 links)

(b) Brainerd-Robinson (793 links)

Figure 4: Aross-time networks comparing AD 1300-1350 to AD 1350-1400 (edges bend clockwise from earlier to later time period)

4 Discussion

In this work, we gave an overview of a variety of methods for building similarity networks among entities represented through a set of attributes. We distinguished three qualitatively different classes of similarity measures that may meaningfully map to various distinct social phenomena such as migration, movement, exchange, and skill and material transfer, among others. We compared entities based on: *(a)* weighted, unweighted, symmetric, and asymmetric similarities among uniformly weighted attributes, *(b)* non–uniform weighting of attributes (tf-idf), and *(c)* rank correlations of attributes. Moreover, entities can be compared to themselves or to others across different temporal scales, if such data is available. We grounded the significance and distinction of these classes of measures by giving comparative and contextual examples of these methods on a case study of archaeological collections pertaining to AD 1200-1500 from the US Southwest. We attempted to elucidate the differences in outcomes and their meanings when choosing various similarity methods through this dataset.

The methods presented in this work were selected based on their theoretical and conceptual variations that can potentially be mapped to a wide array of social processes. Through this work, our aim was to emphasize that similarity is an abstract concept that has conceptually subjective and imprecise meaning. However, it can be operationalized through various methods which can result in arbitrarily different outcomes of similarities among the compared entities. Of course, not every scientific research applies these various measures arbitrarily. But due to lack of surveys in application of similarities measures in various archaeological contexts, it is necessary to objectively compare these variety of measures to facilitate the choice of which measure to work with in different contexts. Therefore, choosing the appropriate method for measuring similarity is a crucial first step. Two important factors we considered are: the specific research question that needs to be answered through this exercise as well as the inherent features of the data and their quality.

Overall, we have touched upon the following three themes with regards to similarity measures: the generic nature of the similarity concept, the profusion of methods for measuring similarity, and the theoretical and conceptual differences among various methods. A natural next step would be to map the various concepts of similarity to the social process they embody. Hence, here we are laying the groundwork for systematically and objectively choosing the appropriate method with respect to the research question. Moreover, the methods, based on non–uniform and ranked weighting of attributes, allude to taking into consideration the characteristics of attributes while choosing the appropriate similarity measure for a given research question. Furthermore, the quality of the data is another factor that cannot be overlooked while applying a certain method, as various methods are sensitive to different characteristics of data such as sparsity, skewness in sampling, and differences in scales of quantities, among others.

Lastly, once the appropriate measure for operationalizing similarity has been picked and networks among entities are built accordingly, the next step is to tap into the vast number of methods developed in the network analysis literature to systematically study various global and local properties of the network itself [Brandes et al., 2013]. For example, network structure, group cohesiveness, roles and centralities of certain entities for various social processes, are some of the well–developed concepts in network analysis [Hennig et al., 2012, Wasserman and Faust, 1994]. These concepts encode a number of social phenomena and can help reveal patterns that are hard to discern through non–network based techniques.

Acknowledgments

We thank the members of the Southwest Social Networks Project for their input, especially Jeffery Clark, Matthew Peeples, and John Roberts. The SWSN database was created through NSF grants #0827007 to the University of Arizona and #0827011 to Archaeology Southwest.

Funding

The research leading to these results is part of the ERC-Synergy project NEXUS1492 which has received funding from the European Research Council (https://erc.europa.eu/funding-and-grants) under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 319209. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[Balcan et al., 2008] Balcan, M.-F., Blum, A., and Vempala, S. (2008). Clustering via similarity functions: Theoretical foundations and algorithms.

[Borck et al., 2015] Borck, L., Mills, B. J., Peeples, M. A., and Clark, J. J. (2015). Are social networks survival networks? An example from the late pre–hispanic US Southwest. *Journal of Archaeological Method and Theory*, 22:33–57.

[Brainerd, 1951] Brainerd, G. W. (1951). The place of chronological ordering in archaeological analysis. *American Antiquity*, 16:301–3.

[Brandes et al., 2013] Brandes, U., Robins, G., McCranie, A., and Wasserman, S. (2013). What is network science? *Network Science*, 1(01):1–15.

[Brughmans, 2010] Brughmans, T. (2010). Connecting the dots: towards archaeological network analysis. *Oxford Journal of Archaeology*, 29(3):277–303.

[Choi, 2008] Choi, S.-S. (2008). Correlation analysis of binary similarity and dissimilarity measures.

[Choi et al., 2010] Choi, S.-S., Cha, S.-H., and Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Systemics, cybernetics and informatics*, 8(1).

[Collar et al., 2015] Collar, A., Coward, F., Brughmans, T., and Mills, B. J. (2015). Networks in archaeology: Phenomena, abstraction, representation. *Journal of Archaeological Method and Theory*, 22(1):1–32.

[Everitt and Rabe-Hesketh, 1997] Everitt, B. and Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*. Kendall's advanced theory of statistics. Arnold.

[Fagin et al., 2003] Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, pages 28–36, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Golitko, 2015] Golitko, M. (2015). *LBK Realpolitik: An archaeometric study of conflict and social structure in the Belgian early Neolithic*. Archaeopress Archaeology Series. Archaeopress, Oxford.

[Golitko and Feinman, 2015] Golitko, M. and Feinman, G. M. (2015). Procurement and distribution of pre–hispanic mesoamerican obsidian 900 BC–AD 1520: a social network analysis. *Journal of Archaeological Method and Theory*, 22(1):206–247.

[Golitko et al., 2012] Golitko, M., Meierhoff, J., Feinman, G. M., and Williams, P. R. (2012). Complexities of collapse: the evidence of maya obsidian as revealed by social network graphical analysis. *Antiquity*, 86(332):507–523.

[Hamming, 1950] Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 26(2):147–160.

[Hart, 2016] Hart, J. P. (2016). New Trends in Prehistoric North-eastern North American Agriculture EvidenceA View from Central New York. In Lee-Thorp, J. and Katzenberg, M. A., editors, *The Oxford Handbook of the Archaeology of Diet*. Oxford University Press, Oxford.

[Hart and Engelbrecht, 2012] Hart, J. P. and Engelbrecht, W. (2012). Northern iroquoian ethnic evolution: A social network analysis. *Journal of Archaeological Method and Theory*, 19(2):322–349.

[Hennig et al., 2012] Hennig, M., Brandes, U., Pfeffer, J., and Mergel, I. (2012). *Studying Social Networks - A Guide to Empirical Research*. Campus-Verlag.

[Heringa, 2001] Heringa, J. (2001). *Sequence Similarity*. John Wiley & Sons, Ltd.

[Jaccard, 1912] Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50.

[Knappett, 2013] Knappett, C., editor (2013). *Network Analysis in Archaeology: New Approaches to Regional Interaction*. Oxford University Press.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Mihalcea et al., 2006] Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

[Mills, 2007] Mills, B. J. (2007). A regional perspective on ceramics and zuni identity, AD 200–1630. In Gregory, D. and Wilcox, D., editors, *Zuni origins: toward a new synthesis of Southwestern archaeology*, pages 210–238. University of Arizona Press.

[Mills, 2011] Mills, B. J. (2011). Themes and models for understanding migration in the southwest. In Nelson, M. C. and Strawhacker, C., editors, *Movement, Connectivity, and Landscape Change in the Ancient Southwest*, pages 345–360. University Press of Colorado.

[Mills et al., 2013a] 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., and Shackley, M. S. (2013a). Transformation of social networks in the late prehispanic US Southwest. *Proc Natl Acad Sci U S A*.

[Mills et al., 2015] Mills, B. J., Peeples, M. A., Haas Jr, W. R., Borck, L., Clark, J. J., and Roberts Jr, J. M. (2015). Multiscalar perspectives on social networks in the late prehispanic Southwest. *American Antiquity*, 80(1):3–24.

[Mills et al., 2013b] 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., and Breiger, R. L. (2013b). The dynamics of social networks in the late prehispanic US Southwest. In Knappett, C., editor, *Network analysis in archaeology: New approaches to regional interaction*, pages 181–202. Oxford University Press.

[Munson, 2013] Munson, J. (2013). From metaphors to practice. Opertationlizing network conepts for archaeological stratigraphy. *Journal of Archaeological Method Theory*.

[Munson and Macri, 2009] Munson, J. L. and Macri, M. J. (2009). Sociopolitical network interactions: A case study of the classic maya. *Journal of Anthropological Archaeology*, 28(4):424 – 438.

[Nick et al., 2013] Nick, B., Lee, C., Cunningham, P., and Brandes, U. (2013). Simmelian backbones: Amplifying hidden homophily in facebook networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 525–532, New York, NY, USA. ACM.

[Östborn and Gerding, 2014] Östborn, P. and Gerding, H. (2014). Network analysis of archaeological data: A systematic approach. *Journal of Archaeological Science*, 46(1):75 – 88.

[Paris, 2014] Paris, E. H. (2014). Cross-valley communities: Identity and interaction in early postclassic period highland chiapas. *Journal of Anthropological Archaeology*, 34(0):78 – 99.

[Peeples and Haas, 2013] Peeples, M. A. and Haas, W. R. (2013). Brokerage and social capital in the pre–hispanic U.S. Southwest. *American Anthropologist*, 115(2):232–247.

[Peeples and Roberts, 2013] Peeples, M. A. and Roberts, J. M. (2013). To binarize or not to binarize: relational data and the construction of archaeological networks. *Journal of Archaeological Science*, 40(7):3001–3010.

[Robinson, 1951] Robinson, W. S. (1951). A method for chronologically ordering archaeological deposits. *American antiquity*, pages 293–301.

[Santini and Jain, 1999] Santini, S. and Jain, R. (1999). Similarity measures. *IEEE Transactions on pattern analysis and machine Intelligence*, 21(9):871–883.

[Scholnick et al., 2013] Scholnick, J. B., Munson, Jessica, L., and Macri, M. J. (2013). Positioning power in a multi-relational framework: a social network analysis of Classic Maya political rhetoric. In Knappett, C., editor, *Network Analysis in Archaeology*, pages 95–124. Oxford University Press.

[Segaran, 2007] Segaran, T. (2007). *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O'Reilly, Beijing.

[Shennan, 1997] Shennan, S. (1997). *Quantifying archaeology*. University of Iowa Press.

[Song et al., 2008] Song, N., Joseph, J. M., Davis, G. B., and Durand, D. (2008). Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput Biol*, 4(4).

[Strehl et al., 2000] Strehl, A., Ghosh, J., and Mooney, R. (2000). Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64.

[Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

[Webber et al., 2010] Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38.