On graphical representations of similarity in geo-temporal frequency data

Daniel Weidele a,*, Mereke van Garderen a, Mark Golitko b, c, Gary M. Feinman c, Ulrik Brandes a

a Department of Computer & Information Science, University of Konstanz, Germany
b Department of Anthropology, University of Notre Dame, United States
c Department of Anthropology, Field Museum of Natural History, United States

ARTICLE INFO

Article info:
Article history:
Received 27 November 2015
Received in revised form 27 May 2016
Accepted 30 May 2016

Keywords:
Network visualization
Maya
Obsidian
MDS
Spatio-temporal data

ABSTRACT

Its focus on dependencies and patterns in relational data makes network science a promising addition to the analytic toolbox in archaeology. Despite its tradition in a number of other fields, however, the methodology of network science is only in development and its scope and proper usage are subject to debate. We argue that the historical linkage with graph theory and limitations in commonly available software form an obstacle to leveraging the full potential of network methods. This is illustrated via replication of a study of Maya obsidian (Golitko et al. Antiquity, 2012), in which it seemed necessary to discard detailed information in order to represent data in networks suitable for further processing. We propose means to avoid such information loss by using methods capable of handling valued rather than binarized data. The resulting representations corroborate previous conclusions but are more reliable and thus justify a more detailed interpretation of shifting supply routes as an underlying process contributing to the collapse of Maya urban centers. Some general conclusions for the use of network science in archaeology are offered.

© 2016 Published by Elsevier Ltd.

1. Introduction

1.1. Theoretical background

Network Science is the study of the collection, management, analysis, interpretation, and presentation of relational data (Brandes et al., 2013). It combines statistical, combinatorial, algorithmic, and graphical methods to address research questions amenable to a network perspective. As for any science, a precise understanding of the potentials and interrelations as well as limitations of network science methods is vital in order to apply them appropriately and obtain meaningful results.

Network approaches are becoming increasingly commonplace. A range of examples demonstrate that also in archaeology new insight can be obtained. A network perspective was used to analyze the use of raw materials and knapping techniques in the pre-colonial Caribbean (Mol, 2014), to understand the collapse of inland Maya urban centers (Golitko et al., 2012; Golitko and Feinman, 2015), to study the transformation of social networks in the late pre-Hispanic US Southwest (Mills et al., 2013, 2015), to explore the co-occurrence and trade routes of Roman table wares (Brughmans, 2010; Brughmans and Poblome, 2012), to study information diffusion through Roman space (Graham, 2006), to model maritime interaction in the Aegean Bronze Age (Knappett et al., 2008), and to identify social and cultural boundaries in Papua New Guinea (Terrell, 2010), to name but a few examples.

However, the methodology of network science is only in development and proper usage standards are the subject of debate. Brughmans (2013) identifies two critical issues regarding the current status in this domain: (1) a lack of awareness and understanding of the broad range of formal network methods within the archaeological discipline has led to a limited methodological scope; (2) the application of network methods in archaeology has been driven mostly by possibility, rather than by specific archaeological research questions. As a result of these two issues, network science applications in archaeology have been dominated by a few popular
methods.

One such popular method is binarization, replacing valued data with zeroes and ones. This converts a weighted network, in which each pair of nodes is connected with a link of some value, into a binary network, in which links can only be present (1) or absent (0). This technique, though very useful in principle, should be applied only with care and double-checking of conclusions, as was illustrated by Peeples and Roberts Jr. (2013) using a number of case studies. Due to the strong link of network science with graph theory, networks are often represented as binary and methods designed to handle valued data are less commonly used in current network science applications. Since, however, binarization incurs information loss, it should be avoided where possible.

1.2. Our contribution

We consider a chain of operations which obtained a prominent place among network methods used in archaeology. In this approach (see for example Mills et al., 2013; Golitko et al., 2012; Golitko and Feinman, 2015), a network is built from similarities between site assemblages. The network is then binarized using some threshold value. Unless sites are shown at the geographic locations, a layout of the graph is computed, typically using a spring-embedder algorithm. While this often serves to visually communicate results, exploration of the network diagram can also lead to new conclusions for the authors themselves.

In this paper, we consider the steps during this process at which information loss occurs. We demonstrate that binarization, which may sometimes appear necessary to be able to apply the intended methods, can actually be avoided. To do so we suggest methods able to handle valued data at each step of the analytic pipeline. We also note that common spring-embedder algorithms do not result in layouts that can be interpreted reliably. With the nature of archaeological research questions in mind, we introduce a variant method for visualizing and analyzing geo-temporal frequency data that gives a more accurate representation of the raw data. We illustrate that this new method can lead to slightly different results than the network mapping of Golitko et al. (2012) on Maya obsidian. We stress that this case study replication is only an example to illustrate the techniques we introduce. Due to the omnipresence of geo-spatial frequency data in the archaeological discipline, the method is in fact widely applicable.

The present contribution should not be understood as a competing analysis of particular archaeological hypotheses. Instead, our contribution is methodical: we point out a strategy to obtain more reliable visual representations and use the archaeological case study on Maya obsidian as a concrete example.

1.3. Data and case study

We identify a class of data that regularly constitutes the basis for archaeological studies. We refer to this class as geo-temporal frequencies, which can be defined given:

- a set of geographic locations \( L \),
- a set of discrete time points \( T \),
- a set of classes of artifacts \( C \)

as a three-dimensional tensor \( X : L \times T \times C \). So \( X_{l,t,c} \) represents the number of, for instance, pottery sherds of ware \( c \) found at site \( l \) dated to time \( t \).

As a case study we consider the work of Golitko et al. (2012) on Maya trade relations in eastern Mesoamerica between 250 CE and 1520 CE. In this study, network methods are applied to archaeological data on material culture, which in turn is used as a proxy for trade. We evaluate the methods used and suggest a number of improvements and extensions. We replicate the case study together with an application of the suggested method which leads to a more precise visualization of the data that allows some new observations.

The data set consists of obsidian assemblages from 121 archaeological sites. Obsidian is considered an ideal material to use for the reconstruction of trade relations since the original source of an obsidian artifact can be chemically determined with high confidence. The three main sources of obsidian in the eastern Mesoamerican Maya area are San Martin Jilotepeque (SMJ), El Chayal (ELC), and Ixtepeque (IXT), all currently located in Guatemala. For ease of viewing and analysis, all Mexican obsidian sources have been compiled into one category (MEX), and all non-major sources in Honduras and Guatemala have been grouped into one category (OTHER).

Fig. 1 shows a map of the study area on which the sites and sources are indicated. The node area corresponds linearly to the absolute number of sourced obsidian objects found at this site, which makes clear how large the differences really are. For ease of viewing, we will use a logarithmic scaling in the remainder of this paper, which makes the differences in node sizes a lot smaller as compared to this figure. Sites are colored according to their geographical zone after Adams and Culbert (1977). We will use the same encoding throughout this paper.

The assemblages have been dated to four time intervals: the Classic period (~250 CE/300–800), the Terminal Classic period (~800–1050 CE), the Early Postclassic period (~1050–1300 CE), and the Late Postclassic period (~1300–1520 CE). Fig. 2 shows the geographical distribution of obsidian from the different sources throughout the four periods as small multiples: a matrix with a column for each period and a row for each obsidian source. Sites [sources] are represented by dots [triangles] in their geographical locations. The node sizes correspond to the logarithmically scaled absolute number of sourced obsidian objects found at this site for a given source and period. The color intensities represent the proportion of the obsidian found at this site for this period that came from this source. A small, black node in the Classic-ELC cell means that for this site, (almost) all of the material found for the Classic period came from source ELC, but that there were not many pieces in total. A large, medium grey node in the Terminal Classic-IXT cell means that for this site only about half of the objects found for the Terminal Classic period came from source IXT, but that this was still quite a large number of objects.

1.4. Preliminaries

In the following we describe how to build a network out of the data described above. Following Brandes et al. (2013) we represent a network variable from geo-temporal frequency data as a mapping \( x : \mathcal{L} \rightarrow \mathcal{Y} \) of dyads from a finite domain \( \mathcal{L} \subseteq \mathcal{N} \times \mathcal{C} \) comprised of ordered pairs of nodes \( \mathcal{L} \) and affiliations \( \mathcal{A} \) to values in a range \( \mathcal{Y} \).

Of the possible combinations with \( \mathcal{L}, \mathcal{A} \subseteq \{L,T,C\} \) we focus on site-site interaction domains \( \mathcal{LL} \) where \( \mathcal{L} = \mathcal{A} = L \). These provide a natural way of directly preserving the geographical context, and are presumably therefore frequently subject of study in archaeological research. We consequently define the network mapping \( x^{LL} \) on the interaction domain \( L \times L \) as

\[
x^{LL} : L \times L \rightarrow \mathcal{Y}.
\]

This means that we look at all possible combinations of two sites (the nodes in our network), and assign a weight to the link between each of these pairs.

Like Golitko et al. (2012), we rely on the assumption that the
Fig. 1. Overview of the sites and sources in the study area, colored according to their geographical zone after Adams and Culbert (1977); node area corresponds linearly to the number of obsidian artifacts found at the site.
One way to measure the strength of the connection between two sites is to look at the similarity of their material culture. Brainerd-Robinson similarity (Brainerd, 1997; Robinson, 1951) is a prominent index designed for this purpose. It relates sites to each other by computing their similarity based on the relative frequencies of observed classes of artifacts. We adapt the Brainerd-Robinson similarity of two sites \( s_{BR}(i, j, t) \) for a given time \( t \in T \) to our notation as

\[
\frac{1}{C_0} \sum_{c \in C} |D_{l,c} - D_{j,c}|
\]

where \( D_{l,c} \) is the relative frequency of material from class \( c \in C \) at site \( l \in L \) for period \( t \in T \). The relative frequency is computed by dividing the absolute number of objects from class \( c \) for one site and period by the total number of objects for this site and period; in the archaeological literature, this is often referred to simply as frequency. Note that the sum of the relative frequencies of all classes is always one, and therefore \( s_{BR} \in [0,1] \). By relativizing the absolute frequencies, this measure ensures that larger sites are not emphasized over smaller ones. Since sites are excavated with different temporal and monetary efforts, a measure that weighs the influence of sites by their absolute frequencies might lead to an undesired bias. However, at this point we remark it would be worthwhile to evaluate alternatives to this measure.

Some network layout methods require distances between nodes rather than similarities. We can transform similarity \( s_{BR} \) into a dissimilarity \( d_{BR} \) by subtracting \( s_{BR} \) from its maximum possible value,

\[
d_{BR}(i, j, t) = 1 - s_{BR}(i, j, t),
\]

and we refer to it also as the Brainerd-Robinson distance (cf. Section 2.2).

Fig. 2. Spatio-temporal view on the obsidian distribution as a matrix of small multiples: each cell shows the number of obsidian objects from a particular source (rows) for a particular period (columns) found at each site. Node area corresponds logarithmically to the number of objects, color intensity corresponds to the relative number of material for this period that came from this source, placement corresponds to geographical location. The numbers in the top left corner of each cell show the total number of objects from this source found for this period.
We slice the data by time period to obtain a separate weighted, complete similarity or dissimilarity network for each period. Let $s_{BR}^t$ and $d_{BR}^t$ denote the mappings by which an every pair of sites is assigned its value $s_{BR}^t(i,j) \mapsto s_{BR}^t(i,j,t)$ and $d_{BR}^t(i,j) \mapsto d_{BR}^t(i,j,t)$.

## 2. Representation of Brainerd-Robinson networks

For ease of comparison, we first review the original approach of Golitko et al. (2012) in Section 2.1. We then move on to discuss the decisions that were made in processing the data and propose an alternative approach with extensions in Section 2.2. In Section 2.3, we compare the results of the different approaches for the data of our case study.

### 2.1. Spring embedding (reproduction)

To visually represent similarity networks $s_{BR}^t$, Golitko et al. (2012) used spring-embedding as a layout method. A concise introduction is given in Brands (2014), but conceptually, a spring layout is obtained from an equilibrium state of a simulated physical system that consists of repelling nodes connected by springs instead of edges. While repulsion helps unfold the graph, the springs keep connected nodes close to each other. Various spring systems have been proposed, and some of them (Fábes, 1984; Fruchterman and Reingold, 1991) are among the most widely used graph layout algorithms today.

The main reasons for choosing spring-embedder algorithms are their intuitiveness and their flexibility in integrating additional layout objectives. However, among others, a major problem is the iterative nature of implemented simulations: iterations can get stuck in local minima that correspond to less desirable layouts, and since implementations typically start from random initial configurations to avoid systematic biases, the layout obtained can be different in each run of the algorithm.

To make use of standard graph visualization techniques, Golitko et al. (2012) perform three steps. First, the mini-max graph (Cochrane and Lipo, 2010) of the $s_{BR}^t$-network is determined for each of the periods. This means that all edges with $s_{BR}^t$ below a certain threshold are removed from the graph. The threshold is chosen such that the maximum number of edges is removed without disconnecting the graph. Since similarity networks tend to be (almost) complete, filtering is a way to reduce clutter in the layout. The second step is a binarization during which all similarities that disconnect the graph. Since similarity networks tend to be

\[ \sum_{ij} \frac{1}{d_{LL}^t(i,j)^2} \left( \left\| p_i - p_j \right\| - d_{LL}^t(i,j) \right)^2 \]

which quantifies the representation error of layout coordinates $p_i \in \mathbb{R}^2$, $i \in \mathbb{W}$, with respect to the given distances. Note that the inverse squared weights deliberately reduce the contribution of errors in the representation of large dissimilarities. Iterative optimization of this function is sensitive to local minima as well but becomes more robust when initialized with coordinates obtained from metric scaling (Brandes and Pich, 2009).

We therefore propose to determine coordinates for sites based on a metric scaling of Brainerd-Robinson dissimilarities, and increase the influence of local details by subjecting these coordinates to stress minimization afterwards.

The visualizations in the centers of Figs. 3–6 are the result of applying this approach to dissimilarity networks $d_{LL}^t$ from the four periods. To display relative magnitudes, line thickness and intensity correspond inversely to dissimilarity values $d_{LL}^t$, i.e., a thicker and darker line represents a higher similarity.

As an extension we propose to also add the principal axes of the $d_{LL}^t$-space (the five-dimensional space used in the MDS computation) to the transformed Euclidean space of the representation (the 2D figure). The $d_{LL}^t$-space is defined by the five different material sources. Adding them into the layout provides us with visual landmarks that make it easier to interpret the rest of the network. This can be achieved by adding an artificial site location $l_c$ for each obsidian source $c \in C$ to the data, using degenerate frequencies $D_{c,t}$, where we set

\[ D_{c,t,c'} = \begin{cases} 1, & \text{if } c = c' \\ 0, & \text{otherwise.} \end{cases} \]

Each source thus represents a site that contains 100% of its own obsidian, but no obsidian from other sources. Again we run Classic MDS and Stress Majorization as described in Section 2.2. In Figs. 3–6 the image on the right shows the results of this extended method, where source locations are red and have capitalized labels. The source locations can be seen as landmarks that support in the interpretation of the results. They can be considered as the fixed points of a frame that repels or attracts (depending on their assemblages) the actual site locations.

### 2.2. Multidimensional scaling

Multidimensional Scaling (MDS) is a family of techniques for dimension reduction. It has been discussed in opposition to graph layout algorithms (DeJordy et al., 2007) but can in fact by used as a layout algorithm itself. Indeed, doing so combines the quantitative advantages of expressing dissimilarities in terms of distances with the qualitative guidance of explicit connectivity representation in node-link diagrams.

Metric MDS (Torgerson, 1952) is a technique suitable for metric data and known to favor large dissimilarities. In our case, each node corresponds to a position in a five-dimensional space spanned by the sources, with coordinates defined by relative frequencies of obsidian from the corresponding source. Since Brainerd-Robinson dissimilarity defines a pseudometric in that space (Shuchat, 1984), metric MDS is suitable to obtain a two-dimensional representation in which Euclidean distances resemble Brainerd-Robinson dissimilarities most closely with respect to a certain error function. The two main advantages of this approach are:

- that the entire data is utilized (rather than a binarization obtained from thresholding), and
- that the solution is essentially unique (rather than changing with every execution).

As a graph layout technique to be used in visualization, however, metric MDS is inferior in the representation of small dissimilarities. Distance scaling (Gansner et al., 2005), on the other hand, is the use of non-metric MDS (Kruskal, 1964) for graph layout and can be viewed as a special type of spring-embedder with springs of various length (Kamada and Kawai, 1989). The objective is to minimize a so-called stress function

\[ \sum_{ij} \frac{1}{d_{LL}^t(i,j)^2} \left( \left\| p_i - p_j \right\| - d_{LL}^t(i,j) \right)^2 \]

that the solution is essentially unique (rather than changing with every execution).
Fig. 3. Network layouts for the Classic period (~250 CE/300–800) computed by spring-embedding (left, reproduction of the result by Golitko et al. (2012)), multidimensional scaling (middle) and multidimensional scaling including the obsidian sources (right). Node sizes correspond logarithmically to the number of objects found at a site for this period, colors correspond to geographical zones after Adams and Culbert (1977). For the MDS pictures, link width and intensity correspond to similarity.
Fig. 4. Network layouts for the Terminal Classic period (~800–1050 CE) computed by spring-embedding (left, reproduction of the result by Golitko et al. (2012)), multidimensional scaling (middle) and multidimensional scaling including the obsidian sources (right). Node sizes correspond logarithmically to the number of objects found at a site for this period, colors correspond to geographical zones after Adams and Culbert (1977). For the MDS pictures, link width and intensity correspond to similarity.
Fig. 5. Network layouts for the Early Postclassic period (~1050–1300 CE) computed by spring-embedding (left, reproduction of the result by Golitko et al. (2012)), multidimensional scaling (middle) and multidimensional scaling including the obsidian sources (right). Node sizes correspond logarithmically to the number of objects found at a site for this period, colors correspond to geographical zones after Adams and Culbert (1977). For the MDS pictures, link width and intensity correspond to similarity.
FIG. 6. Network layouts for the Late Postclassic period (~1300–1520 CE) computed by spring-embedding (left, reproduction of the result by Golitko et al. (2012)), multidimensional scaling (middle) and multidimensional scaling including the obsidian sources (right). Node sizes correspond logarithmically to the number of objects found at a site for this period, colors correspond to geographical zones after Adams and Culbert (1977). For the MDS pictures, link width and intensity correspond to similarity.
sites in similar routes of transportation. The present analysis mostly retains the large scale positioning found in the previous approach and continues to offer support for an increasing connection of Maya area sites to coastal routes of transportation from the Classic period onwards. Additionally the deterministic and more accurate graphical representation of the data allows interpreting node positions on a level of detail that was avoided in the original publication. However, the analysis is still highly limited by the sample size as many individual nodes in the actual prehistoric transportation system are omitted.

We find that in some cases the proposed method results in a layout where geographically proximal sites are close to each other, which supports the assumption that there exists a relation between Brainerd-Robinson and geographic distances.

2.3.1. Classic

In the proposed MDS approach we find that sites along the Belizean coast are positioned in a way that more closely respects their geographical relationships— for instance, identical positioning of Moho Cay and Chac Balam, both located on the Belizean Caribbean coast, and Ek Xul and Uxbenka, which are proximal sites in southern Belize. In the original spring-embedded graphs these sites did not appear to be much more similar to one another than to most of the rest of the central Maya area. The level of detail guaranteed by the MDS visualization may allow for a suggestion of particular inland routes up river drainages in Belize such as the relationship between Ek Xul-Uxbenaka-Chan-Tikal, which could suggest a route through southern Belize and around the Maya mountains to more inland sites on the Belize/Guatemalan border.

2.3.2. Terminal classic

The overall structure of both networks for the Terminal classic period is roughly similar, although the retention of weaker links suggests connections between the site of Huacanacastl (Soconusco region) and sites further north in the Guatemalan Highlands and along the Belizean coast, possibly reflecting the location of Huacanacastl at Pacific end of a riverine path through the highlands ending near the Belizean/Honduran border. A direct link between Isla Cerritos and Chichén Itzá and Copan (a probable access point of IXT obsidian) is retained while keeping the distinct clustering of northern Yucatan sites intact. This is consistent with our interpretation of the increasing importance of trade along the eastern coast of Belize that would have linked Copán (exporting IXT obsidian) with Chichén Itzá, a major center of distribution for obsidian and the key bridge between central Mexico and the Maya area during this time (Golitko and Feinman, 2015). In contrast, doing the same with the spring-embedder and no threshold results in little interpretable structure across the study area. As for the classic period, proximal sites such as Labna and Xkipche (northern Yucatan) that appear relatively far apart in the spring-embedder approach are closely positioned in the new representations.

2.3.3. Early and Late Postclassic

The limited number of sites available makes any differences in structure less evident for the Early postclassic period, however, the new visualization places Xelha and Colha, two sites located along the eastern Yucatan coast, in close proximity, and further away from the nearby site of San Gervasio, better representing the differences in assemblages present among these particular sites. In contrast the original spring-embedder visualization places these sites equidistant from one another. This may suggest variable routes of supply along the Yucatan coast during this time period. Node positioning in the LPC period seems more related to geographical locations when compared to the same data visualized using the spring-embedding algorithm. For instance, regional clustering is more evident for Highland Guatemalan sites, particularly those connected to the SMJ source. The role of the coastal sites Laguna de On and Caye Coco in linking northern Yucatan to the rest of the study area is far more evident, again an expected feature of network structure given knowledge of geography and probable routes of movement in eastern Mesoamerica, further demonstrating the likely importance of coastal Yucatecan sites in obsidian transport.

3. Evaluation

In the following, the qualitative insights into the case study above are backed by more quantitative evidence on threshold sensitivity and the accuracy of distance representation in network layouts.

3.1. Protocol

We are interested in how well distances in a layout represent the input distances obtained via the Brainerd-Robinson index. Since the previously used spring-embedder approach involves thresholding and binarization, we also want to assess the representation error introduced in these preprocessing steps prior to the layout.

For each period we therefore compute:

- the target matrix \( d_{ij} \) of distances from Brainerd-Robinson index
- filtered versions of that matrix for thresholds including and above the one defining the mini-max graph
- binarized versions of the filtered matrices
- spring-embedder layouts of the graphs corresponding to the filtered and binarized matrices
- distance matrices for both the filtered and the filtered and binarized matrices (using an all-pairs shortest-paths algorithm)
- an MDS layout of the original BR distance matrix

To avoid testing on the inherent optimization criterion of MDS itself, the stress from Equation (4), we quantify the difference between the original BR distance matrix and the Euclidean distances in a layout or the distance matrix of the reduced networks using the root-mean-square error (RMSE) which is defined as follows. For distances \( d(i,j), i,j \in L \), obtained either as the Euclidean distances in a layout or directly from a transformed matrix, the representation error with respect to the BR distance matrix \( d_{ij} \) is defined by

\[
RMSE(d, d_{ij}) = \min_\alpha \sqrt{\frac{1}{|L|^2} \sum_{i,j \in L} (\alpha \cdot d(i,j) - d_{ij}(i,j))^2},
\]

where parameter \( \alpha \) ensures that differences are independent of scaling. To facilitate comparison across periods, we normalize the RMSE with the number of sites.

3.2. Results

In Fig. 7, representation errors are shown as a function of the degree to which the original data has been distorted.

The x-axis is defined by the threshold values below which edges have been filtered out. A low threshold value results in many edges being filtered out, a high threshold value means most edges are kept in. RMSE scores are mapped to the y-axis.

Since the proposed MDS approach uses the complete, non-transformed BR distance matrix, it is independent of the threshold. The mismatch between distances in the layout and the desired distance is therefore depicted as a straight line (dark orange).
Fig. 7. Representation error of distances in MDS (dark orange) and spring-embedder layouts (dark blue), and of distances incurred by filtering (light orange) and filtering and binarizing (light blue) BR distances for threshold values at which connectivity is maintained. The spring-embedder was run 25 times on each graph and the distribution of RMSE scores is indicated by a dot for the median and vertical lines connecting the minimum and maximum value with the first and third quartile. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(a) Classic (~AD 250/300-800)

(b) Terminal Classic (~AD 800-1050)

(c) Early Postclassic (~AD 1050-1300)

(d) Late Postclassic (~AD 1300-1520)
The matrix of shortest-path distances in filtered networks is an indication of the error introduced by thresholding (light orange). For the minimum threshold that leaves the network connected, this error is even larger than the one introduced by the MDS layout. The more of the original matrix is retained, i.e., the higher the threshold, the closer the distances are to the desired ones.

The error introduced by thresholding and binarization (light blue) is given by the shortest-path distances in the resulting graph. Here, the trend is reversed because binarization of increasingly complete matrices yields increasingly cliquish graphs in which there is low variation in distances.

Spring embedding of these graphs (dark blue) introduce further distortion because these distances cannot be represented accurately in two-dimensional layouts. Moreover, there is a degree of randomness in spring-embedding so that the figures show the distribution of representation errors rather than a single value.

The experiment clearly shows that MDS yields more accurate representation for any threshold. While spring-embedders actually perform best near the lowest possible threshold value, the error is halved by avoiding the transformation and using MDS directly.

4. Conclusion

We proposed to use valued graph representations and MDS techniques to visualize archaeological similarity networks. These do not incur the information loss from quantization that is unavoidable for visualization techniques requiring binarized network data. Unlike many other spring-embedder approaches, properly initialized stress minimization yields interpretable layouts rather than complete layouts based on non-distorted original data can be interpreted with higher confidence and in more detail.

In the case study re-analysis we observed, in particular, that for many sites the association between geographic and layout distance was actually stronger than suggested by previous visualizations. Assuming that geographic distance is reflected in the distribution of obsidian this adds further evidence to the proposition that the proposed technique represents these data more accurately, as does our quantitative experiment.

The additional modification of incorporating sources as artificial sites throughout the process yields even more informative visualizations. The relative frequencies of obsidian at a site can be inferred qualitatively from the position relative to sources, and as part of the network sources also exert an influence on the relative positioning of sites.

Acknowledgment

This research was funded in part by the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013), ERC grant agreement no 319209.

References


