From Chemistry to History via Kernel Embedding

csaaw

November 16, 2016

abstract

1 introduction

This paper explores the question: how much can the chemical composition of pottery fragments tell us about the evolving connections between settlements in Bronze Age Greece?

Some classical statistical techniques have been used to study the LBNL dataset. For instance, in [main paper], the basic approach consists of using weighted square differences between the chemical composition of two artifacts to determine their similarity, which is then used to cluster the collection of artifacts. This clustering can then be compared to the actual locations of the different sites where the artifacts were found.

The clustering of artifacts according to their chemical composition can say much about the relation between their regions of origin. For instance, if two regions contain many artifacts falling in the same chemical composition cluster, it follows that the materials used in the construction of the artifacts were similar, and perhaps that the two regions shared resources or had markets that were highly connected.

There are some issues with the classical approaches that we wish to attack. Mainly, there is no clear way to analyze connectivity among different archaeological sites based on the clustering of artifacts. A clustering of artifacts based on chemical composition that mimics well the geographical location of the archaeological sites is very illustrative, but it is not obvious how it can be used to quantify the degree of similarity between different sites.

Our contribution to the analysis of the LBNL data is to provide a well-justified similarity measure between archaeological sites based on the chemical composition of their respective artifacts. We are then able to provide quantitative description of the amount of connectivity between different archaeological sites.

2 literature review

In this section, two categories of literature are surveyed: (a) theories of Bronze Age trade relevant to the movement of ceramics, and (b) relevant archeometric and modeling methods.

Our dataset includes ceramic shreds from the 7th century BC to the Roman period, but mainly consists of works from the Late Helladic Period - III (LH III, see fig. 1 for a chronology) fig. 2 is a map with dashed lines encoded hypotheses about prominent sea routes in the late Bronze Age, which encompasses LH III.

According to the literature, as more data becomes available and methods for modeling and interpreting the data evolve questions of provenience for specific samples of ware can be traced, differentiations between pottery from the site and imported Mycenaean samples can be made, as well as new insights on relationships between trade sites and the trade routes can be mapped out. Some work has been made using the LBNL to trace provenience sites as in the work of Grave et al.[**grave2014ceramics**] who argued that focusing on geochemical precincts of regional parent geologies provides a more accurate method for locating provenience. They were able to map out the provenience of Red Lustrous Wheelmade Ware by using two datasets: the NAA results for a sample population of RLW from Bogazkoy and the LBNL Cypriot ceramic NAA dataset. In

7000	Beginning of widespread use of pottery in Mediterranean. Aegina as maritime center
3200-2000	Early Helladic
3200-2500	High point of Anatolian trade network
2000-2200	Disruption of trade between Cyclades, Mainland, and Crete. Upheaval in Mainland
2000-1550	Middle Helladic. Protopalatial and Neopalatial buildings in Crete
1650-1500	Late Helladic I / Late Cycladic I / Late Minoan IA
1500-1400	Late Helladic II
1400-1300	Late Helladic III A
1300-1200	Late Helladic III B. High point of Mycenaean influence on trade
1200-1050	Late Helladic III C
1050-	End of Bronze Age. Transitional Period: Argos, Asine and Berbati rise to prominence.

Figure 1: Timeline of Events of Interest (all dates B.C.)

instances where Mycenaean ceramics are found in other sites in the Mediterranean region such as the work by Jung, Mommsen and Pacciarelli[**jung2015west**] differentiate between vessels originating in Greece versus pottery native to the Punta di Zambrone site.

Maps like fig. 2 are of tremendous historical significance if they can be constructed accurately. A common way to map trade is to identify artifacts that have been transported from their location of origin. How do researchers infer where ceramic objects and shreds originate?

Qualitative methods include categorization by geometric features, and material features. For instance, Davis[**davis1979late**] classifies pieces from LH I Korakou by shapes, as well as surface and burnishing, measured by the eye against a color chart. The task is to sort shreds and pieces by style, and also material - these two elements are necessary to differentiate cases where a style travels but pieces are built from local materials. A qualitative study we'll discuss in detail later is that of Rutter[**rutter1975ceramic**] - he uses the shapes and visible properties of pottery found in Korakou to argue that those pieces were erroneously categorized as Mycenaean. That is: qualitative differences are used to prove a classification claim.

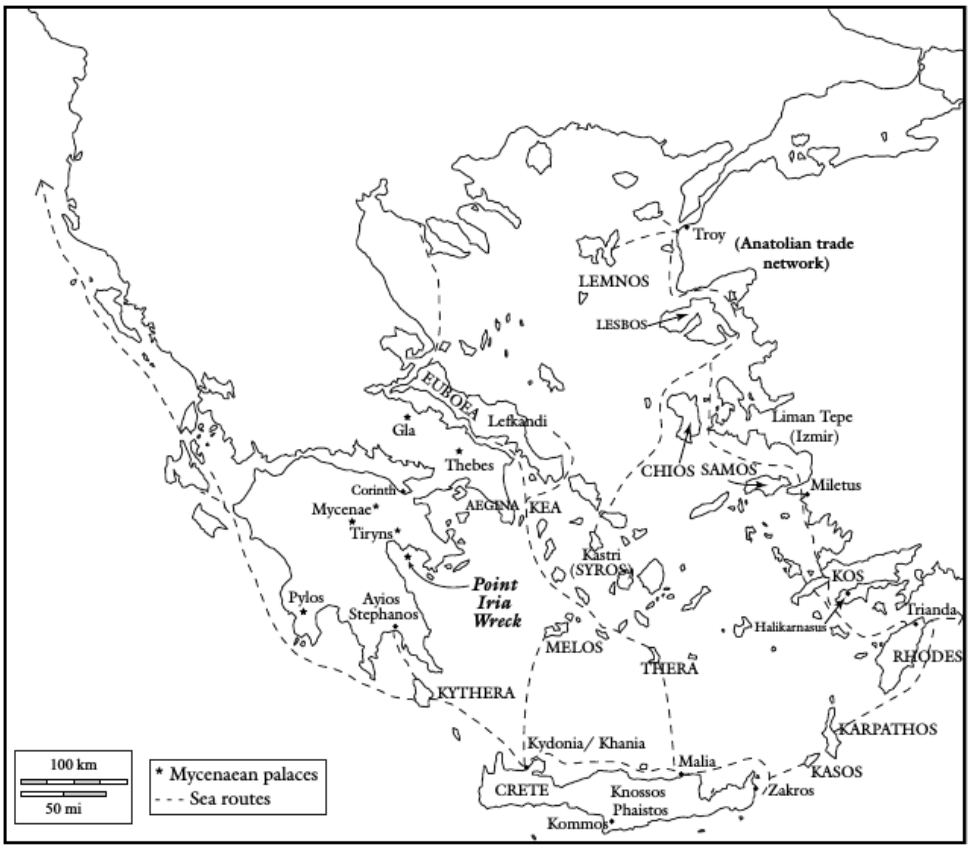
Quantitative work in classifying pottery mainly uses data from neutron activation analysis or spectral analysis of paints. This paper uses a neutron activation analysis dataset from the LBNL archeometry archives. The extant work on this dataset (which contains subsections for many geographical areas across the world) has mostly been conducted by Mommsen and colleagues. This work has two main goals: first, to be a proof of concept for quantitative analysis by reproducing known results from limited data, and second, to discover previously unknown connections as a bridge for further inquiry, both qualitative and quantitative. So for instance, in [**mommsen2002complete**], Mommsen et al take their model to succeed by matching existing predictions about early recipe variation in Argolid pottery suggested by Hoffman et al, and by adding new predictions in the case of suggesting a pattern of export from Chania to Cyprus of cream ware. Grave et al[**grave2014ceramics**] follow a similar method of combining the LBNL dataset with other information, also with Cyprus late Helladic wares.

3 Motivation/Dataset

3.1 Motivation and Summary Stats

Archaeological studies has been actively inviting the analytical methods from the entire spectrum of social science studies. However, when it comes to the application of machine-learning techniques, we have not seen much done recently. Therefore, we set out to explore the archaeological databases, searching for datasets that are rich enough, in terms of dimensions of the observables.

Baring the capability of machine-learning in mind, we found the Lawrence Berkeley National Laboratory (LBNL) Nuclear Archaeology Program Archives of particular interest, as it not only contains geographical and historical records of the artifacts, it also carry the chemical compositions for most of its records.



Map 7.3 Mycenaean palaces and Aegean maritime routes.

Figure 2: image from [**demand2011mediterranean**]

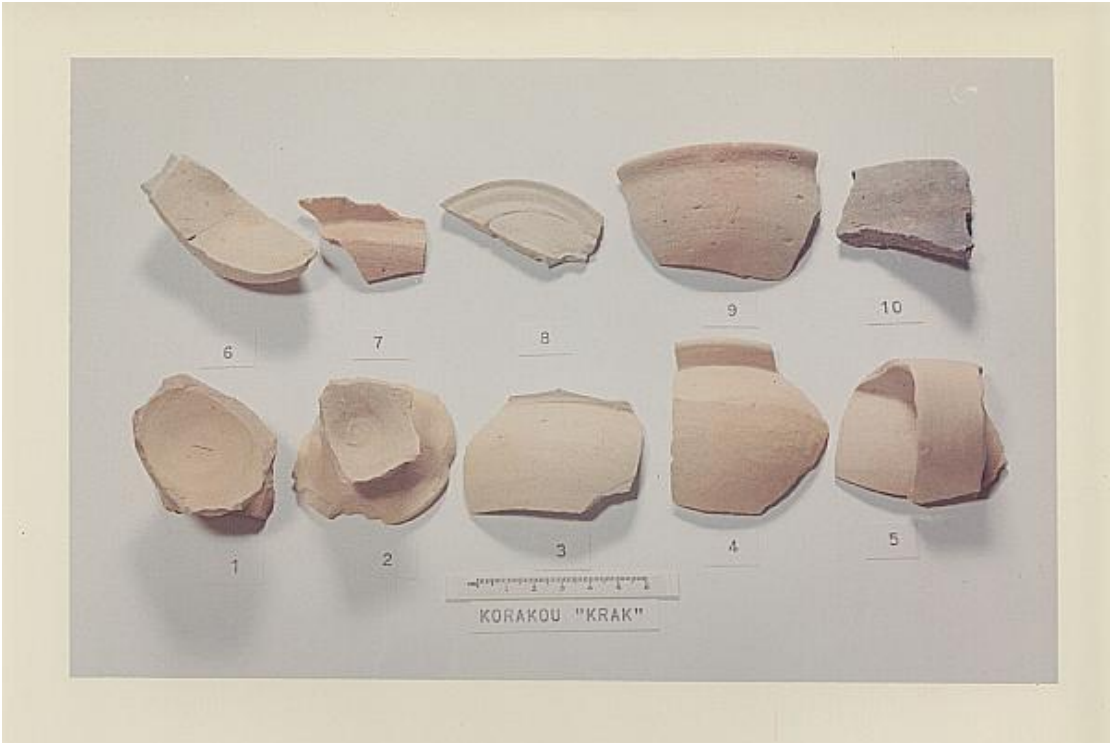


Figure 3: Example of shreds in the database

Due to limited time and resources, we focused exclusively on the Greece records, totaling 886 pieces of artifacts from ancient Greece, discovered from 31 archaeological sites[1]. The time frame in our sample covers from the Early Helladic era (around 3200 BC) to Roman Republic.

3.2 Data manipulation

For the records of artifacts from Greece, the dataset came with 1,198 observations in the begining, pulled directly from `http://core.tdar.org/`[2] The variables include the discovery site, the associating era, geo-coordinates and the chemical compositions of 33 elements[3]. However, the chemical composition data is not complete. Thus, we dropped a subset of observations as well as insignificant variables of elements ($Sb\ Ba\ As\ Sr\ V$) to obtain a sample of 886 observations.

4 Problem Formulation

5 problem formulation

Our goal in this project is to understand relations of evolving similarity between archeological sites over time. In this section, we define this problem more exactly.

First, a brief intuitive gloss on our technique. In our model, similarity reflects a relationship between our estimates of the underlying functions which are 'generating' the artifacts. By modeling it this way, we reduce the site information to an underlying probability function. Then, these functions can be compared, and a relationship computed pairwise between all the sites. This allows us to construct a network. Finally, we can look at the network among sites

[1] 30 in Greece, and one (Perati in Attica) in Turkey.

[2] To take a quick look at the artifacts, please refer to **??**.

[3] Elements are: *Al Ca V Dy Mn Na K Sr As U Eu Ba Sm La Ti Lu Nd Co Sc Fe Ce Yb Cs Ta Sb Cr Th Ni Rb Tb Hf Zn*

over the entire period, or segmented by time. In the latter analysis, we separate the artifacts into eras, and look at the similarities in each era separately.

What is this analysis meant to accomplish? Archaeologists have looked at Greek pottery artifacts for centuries in an attempt to understand trade, and each individual site we consider has been described in elaborate detail (for example, see [**davis1979late**]). Our project is meant to be *focused and precise* in its use of data, while aiming at a *general and holistic target* - simplicity of element abundance profile. This might seem like an odd mismatch, but it also has the potential to either turn up novel connections or confirm old results from a fairly independent source.

So the problem we are attempting to solve is to take a well-defined but limited set of features (element abundance) and generate a similarity network using only those features. Then, we will compare the relations in our network to theories and findings from archaeologists and historians about the trade and migration patterns that might effect the similarity of pottery artifacts. In the most general case, we consider the progression of trade power (among our sites, a progression from Aegina in the early Bronze Age, to Festos in the Minoan period, then Mycenae in the late Helladic, and finally Berbati). These theories of trade relations also imply similarity relations - just of a more qualitative kind, and based a far wider body of evidence. Finding significant points of agreement between our model and these theories would be an indication that the similarity relation we've calculated based on these simple chemical properties is latching on to something deeper. This could be because of similarities in raw materials, or firing techniques, but most likely a combination of the two. Alternately, finding unexpected similarity relations or relations that run contrary to extant archeological theories might be a call for explanation: could there be a genuine trade connection between these regions? Or if not, could the similarity in composition of the shards be reflection some other kind of connection?

In short, we view the application of machine learning techniques to archeology as aiding in a discovery process. The discovery process in question in this paper is comparing quantitative similarity of chemical composition of pottery to qualitative similarity of regions based on extant theories of influence, migration and trade.

6 Methodology

We are interested in analyzing artifacts from $N = 20$ archaeological sites. We denote by X^i the collection of artifacts in the i-th site, that is $X^i = \{x_1^i, x_2^i, \ldots, x_{n_i}^i\}$, where n_i is the number of artifacts in site i and x_m^i refers to the m-th artifact in the i-th site. The chemical composition data available for each artifact consists of the concentration of $L = 27$ different chemicals. Artifact x_m^i then consists of the (normalized) vector containing these chemical concentrations. The simplest way to compare the chemical composition of two artifacts is through their square distance

$$\|x_m^i - x_l^j\|^2 = \sum_{k=1}^{L} ((x_m^i)_k - (x_l^j)_k)^2$$

where $(x_m^i)_k$ refers to the k-th (normalized) chemical concentration of artifact x_m^i. This in turn is used to construct many similarity measures, for instance

$$k(x_m^i, x_l^j) = \exp(-h\|x_m^i - x_l^j\|^2)$$

where $h > 0$ is a fitting parameter. This corresponds to the widely used Gaussian kernel. Other similarity measures k between artifacts based on their chemical composition are possible. The question of which similarity measure between chemical composition of artifacts is more appropriate for statistical inference is an important one, but we do not consider it further. The challenge is then to construct a similarity measure between different sites based on the similarities $k(x_m^i, x_l^j)$ between their respective artifacts. We propose constructing such similarity function by looking at the average similarity between all pair of artifacts (x_m^i, x_l^j) belonging to sites X^i

and X^j, respectively, given by

$$S(X^i, X^j) = \frac{1}{n_i n_j} \sum_{m=1}^{n_i} \sum_{l=1}^{n_j} k(x_m^i, x_l^j).$$

The justification for this similarity measure between archaeological sites is given by the theory of kernel mean embedding, as described below. We now wish to illustrate how well the similarity measure S performs in finding similarities between archaeological sites based on the chemical composition of their respective artifacts. We do this by providing a network plot of the archaeological sites, where two sites X^i and X^j are connected if their similarity $S(X^i, X^j)$ is larger than a prescribed threshold. The results are illustrated in the figures..

7 Similarity

In this section we detail the way in which similarities are computed. We draw inspiration from a well known set of methods in the machine learning community, and acknowledge that the mathematical machinery might not be familiar to the intended audience. Therefore the reader is invited to skip this section on the first read and come back to it when necessary. We think of the chemical composition vectors $\{x_m^i\}_{m=1}^{n_i}$ corresponding to the artifacts in the i-th site X^i as consisting of samples from some distribution P_i.

7.1 Reproducing Kernels

To build a network we need an appropriate notation of similarity between sites (probability densities). We desire the probability densities to inhabit a well behaved space of functions (a Hilbert Space) where a common notion of similarity exists. We propose to embed each pd into the so called RKHS of a kernel K and use iner product in this space as the similarity measure. The similarities we propose capture similarities among probability densities as elements of a function space. To determine these similarities we need first determine the function space the probability densities inhabit. The function space we will use is the so called Reproducing Kernel Hilbert Space (RKHS) associated to a reproducing kernel. We now define these mathematical objects. A *reproducing kernel* is a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a unique Hilbert space of functions $\mathcal{H}$ such that $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and the reproducing property

$$f(x) = \langle f, k(\cdot, x) \rangle \tag{1}$$

holds for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$

7.2 Kernel Mean Embedding

Let k be a reproducing kernel as in section, the *kernel mean embedding* of the probability density P in the RKHS H of k is

$$\phi_0(P) = \int k(\cdot, x) P(x) dx$$

Since the true form of P is unknown we use the available data to estimate the KME. Let $x_l{}_{l=1}^n$ be an iid sample from P, then the empirical kme of O is

$$\phi(P) = \frac{1}{n} \sum_{l=1}^{n} k(\cdot, x_l).$$

Although $\phi(P)$ depends on the sample we will omit this dependencies to simplify notation. With this tool at hand, we define the similarity between site i and site j as the inner product of their

corresponding KME's.

$$\begin{aligned} S(X^i, X^j) &= \langle \phi(P_i), \phi(P_j) \rangle_H \\ &= \frac{1}{n_i n_j} \sum_{l=1}^{n_j} \sum_{m=1}^{n_j} \langle k(\cdot, X_l^i), k(\cdot, x_m^j) \rangle_H \\ &= \frac{1}{n_i n_j} \sum_{l,m} k(x_l^i, x_m^j) \end{aligned}$$

where the last equality follows from the reproducing property.

8 Results/explanations

In **??**, similarity results are displayed for each of four eras, whereas **??** similarity over all of the artifacts is displayed. For the overall results, the standard historical narrative would predict Mycenae to be fairly central; a vast majority of the artifacts are from the Late Helladic III period, in which Mycenae is the dominant power in the Aegean [**demand2011mediterranean**]. Our results replicate this: Mycenae is the most connected node in the graph.

In **??**,

9 Conclusion

10 method

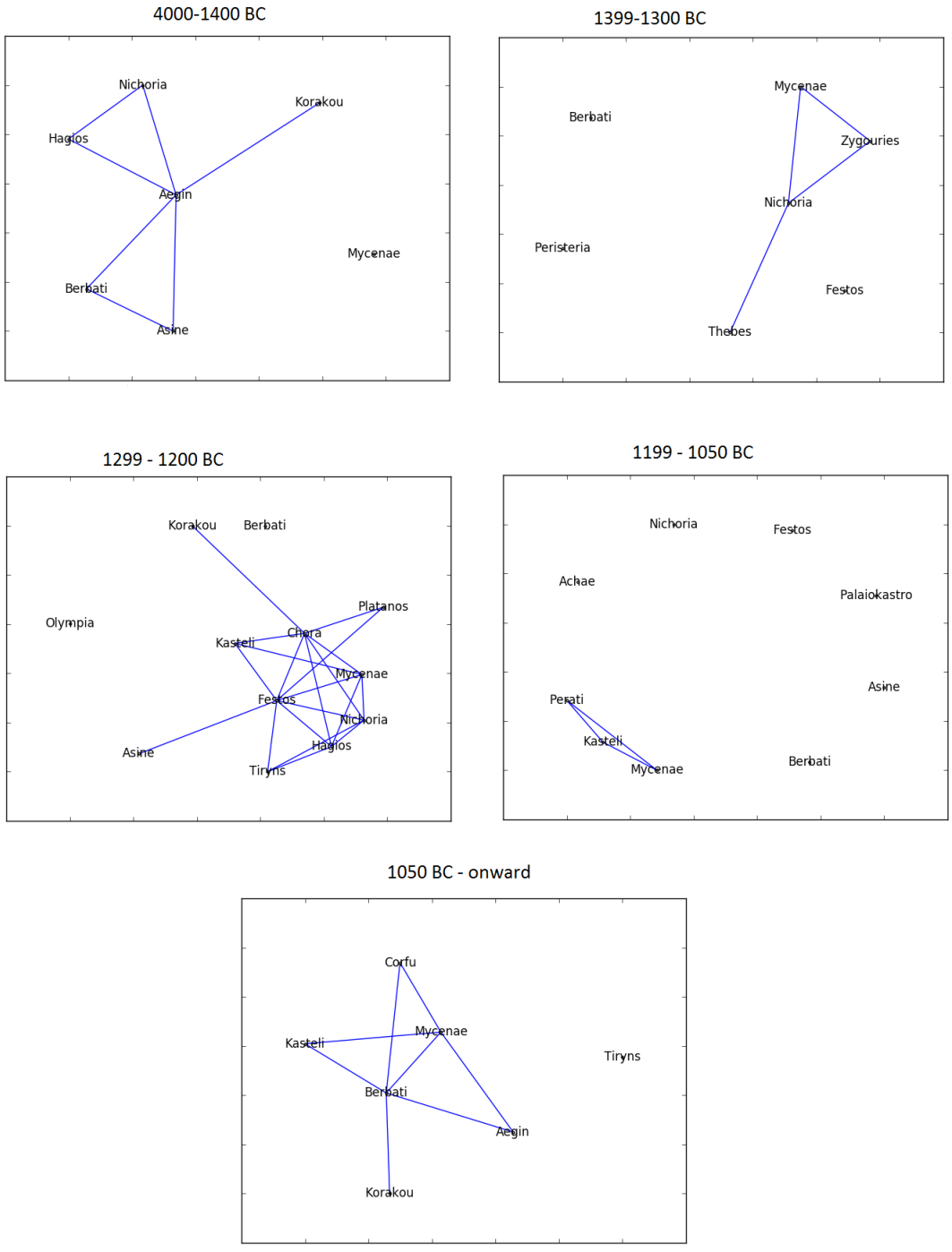


Figure 4: Evolution of networks of the Archaeological sites

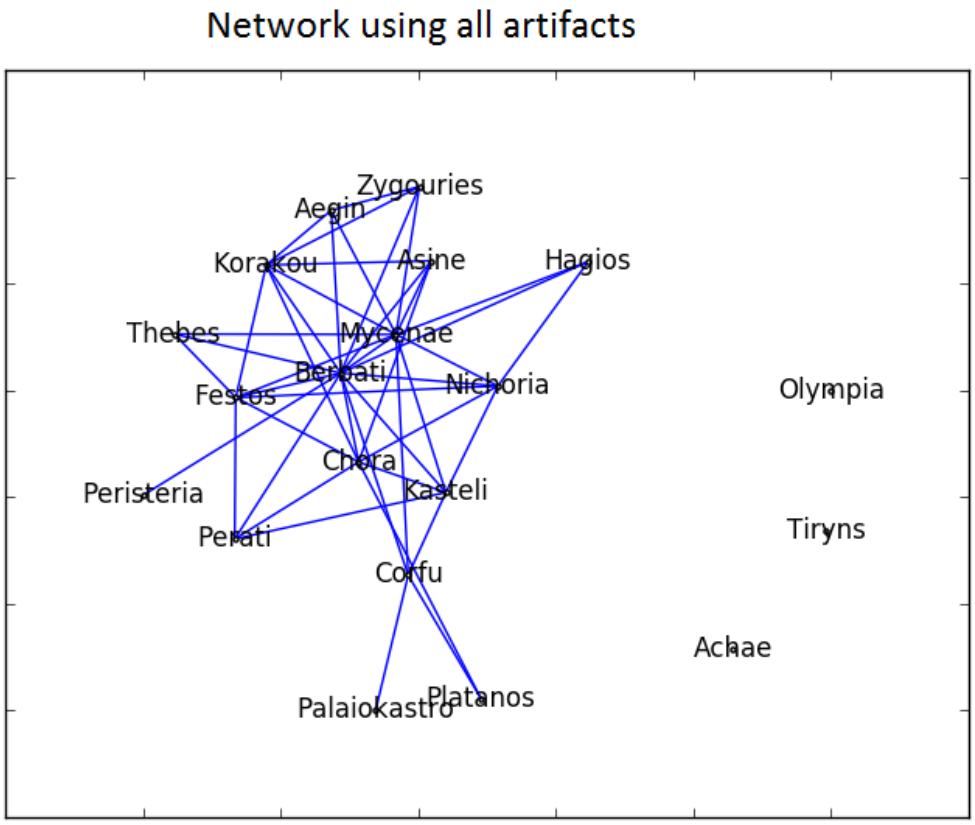


Figure 5: Overall network of the Archaeological sites