RESEARCH ARTICLE

Sets, Graphs, and Things We Can See: A Formal Combinatorial Ontology for Empirical Intra-Site Analysis

J. Scott Cardinal

A critical aspect of analysing an archaeological site is identifying the network of relationships between the things we find and the locations where we find them. These associations are typically determined by a combination of quantitative analyses and the professional knowledge and intuition of the archaeologist, but where exactly is the boundary between what is truly empirical field data and what is inferred through our prior knowledge and field methods? How can we best support those inferences? This paper is a critical evaluation of that boundary to firmly ground, as much as possible, a quantitative analysis on only that which we can directly observe – the thing and its location – and derive associations from that basis alone. To do so, the approach described here relies on a combination of set and graph theories rather than statistical or spatial methods. This revised ontology allows a formalization, in combinatorial terms, for describing an underlying structure to contexts and assemblages that suggests a clear association between archaeological site analysis and a well-studied class of set and graph covering problems. This, in turn, points towards potential algorithmic solutions for a more holistic parsing of the total relationships between sites, contexts, assemblages, proveniences, and artefacts.

Keywords: ontology; archaeological theory; quantitative analysis; graph theory; set theory

Introduction

A fundamental aspect of archaeological work is identifying patterns within a site of interest. The data most typically used to identify interpretable patterns are derived from our unit locations and stratigraphic levels, section profiles, numerous maps and detailed plans in addition to any artefacts collected. This information constitutes the empirical archaeological record – the physical, observed, measured, counted, and mapped samples from which we will build, *by inference*, all subsequent analyses and interpretations of the site. The archaeological record, however, is inherently both incomplete due to preservation and sampling and represents a convoluted final product of various transformations through site formation processes. With every subsequent transformation, the original deposit and context become somewhat more obscured.

As we proceed through from excavation to interpretation, each step of the archaeological process entails a certain increase in abstraction from those initial empirical data. Archaeologists commonly expect, due to this incomplete nature of archaeological materials, that our inferences will reflect a certain amount of necessarily interpolated and extrapolated conclusions. Thus, we infer patterns from both the consistencies and the discontinuities between and among the data of the archaeological record. Each

interpretative step we take away from the empirical data leads to an aggregation of further inferences. Such abstractions also, necessarily, involve a corresponding degree of information loss as particulars are subsumed into generalizations.

The uncertainty introduced by moving incrementally further from the empirical basis of our data underscores the most difficult and pertinent question for interpreting the archaeological record – how can we show that our inferences, inasmuch as they are based on that empirical record, are correct? In other words, are we reasonably certain that we've correctly identified which samples belong to which contexts? Can we demonstrate that our assemblages are, in fact, related? How can we provide stronger evidence to support whether our samples are truly associated? Is there a way to penetrate the intervening layers of noise, untangle the cumulative transformations of post-depositional processes, and get a glimpse of the site as it was originally?

At the core of it, the problem is that the analyses of spatial and temporal patterns of interest are derived from the underlying structure of relationships between those empirical *samples*, rather than from the underlying spatio-temporal structures and processes of the site itself. When we speak of the “integrity” of a site, or its “stratigraphic integrity”, what is really at question is whether there is a clear and supportable chain of inference from the empirical excavated samples backwards through to the site's formation processes and their interpretive significance.

The process must start, then, not with the site's spatial patterns or assemblages, but by first establishing the underlying structural network of empirical associations within and between the site's excavated samples. In other words, there is a certain amount of pre-processing of the field data that must occur in order to have a more solid and supportable means of access to the information needed to reconstruct the other patterns of interest. The first step should be to determine *what* is being patterned by determining what within our data are empirically associated.

This paper presents a methodology, with its quantitative and computational implications, to ascertain those *initial* structural linkages from the empirical information available in the excavated samples themselves. The goal is to partition the samples into “natural” subsets – i.e., samples that are (most likely) drawn from the same population, and therefore were likely produced by the same process. If we presume that a given discrete component in the stratigraphic sequence of a site is the unique product of an underlying formation processes, then samples drawn from that context should reflect identifiable consistencies distinguishing them from other samples.

The premise of the methodology presented here is that an assemblage of artefacts from any excavated sample must be a subset of the total assemblage within its associated stratigraphic context (see **Figure 1**). This methodology necessitates a combinatorial approach, based in set and graph theory, to identify the underlying relationship between a set of artefacts (which constitute an assemblage) with a set of proveniences (which constitute a context). Ideally, each unique pairing of assemblage and context indicates a “natural” subset (i.e., one produced

by an underlying formation process) that represents a shared deposition.

Thus, this methodology entails what should be the initial stage of analysis – finding and defining the empirical interrelationships. Therefore, this step occurs *prior to* the evaluation of spatial patterning, stratigraphic sequencing, or the interpretation of formation processes and occupation activities. The emphasis here is on utilizing the empirical nature and contents of the most basic entities of archaeological excavation to allow a more holistic parsing of archaeological field data. Methods of spatial analysis are not the focus for this paper, although spatial collocation is clearly a factor. Similarly, the units of analysis are considered only with respect to their empirical expression (i.e., their observable characteristics) and not their broader interpretive connotations. Ultimately, the goal of the methodology presented here is to at least identify (and potentially provide a means to filter) the noise introduced into the archaeological record by post-deposit transformations based solely on empirically verifiable measures.

A note on nomenclature

Different regions, specializations, and sub-disciplines tend to evolve their own jargon and nomenclature for field excavation and its units (e.g., “levels” versus “splits”, “units” or “ops and sub-ops”, etc.). As a Northeastern U.S. cultural resource archaeologist, I generally work (and think) in terms of discrete gridded units and their vertical subdivisions – i.e., systematic sampling rather than complete context excavation. This does have some ramifications for the terminology I will use here. Most notably, the usage of the terms *provenience*, *context*, *component*, *association*,

Figure 1: The goals of intra-site analysis based on the relationships between the basic elements of a site.

and *assemblage* may have connotations that differ from others' regional or topical usages. In general, the usages are consistent with the definitions found in Lyman (2012).

References in this paper to *provenience* or *excavated component* are used in their most restrictive sense, and denote only the three-dimensional locator for an excavated sample. As used here, *provenience* generally denotes the identification of a vertical subdivision within an excavation unit (i.e., the soil stratum or arbitrary level division), and consists of the unique identifiers (i.e., unit and level) for a spatial reference. *Excavated component* refers to the sample itself, and entails both the attributes of the excavated sample as well as its contents.

Both *context* and *component* are used here in the sense of an associated collection of proveniences/sample, and may refer to either the collection of samples (i.e., proveniences) or to the source deposition from which the samples (i.e., excavated components) are collected.

The term *association* is not used here as a specific unit of analysis, but rather refers to the systematic connections between units of analysis. The associations between proveniences, contexts, components, and assemblages derive from the various underlying causal or formative processes, and is the structural objective of analysis.

Assemblage is used only to refer to an associated collection of artefacts or otherwise related finds (e.g., "eco-facts", features, etc.). The notion of assemblage is conceptually linked to a number of interpretive constructs, but the specific problem I want to address is the method by which we establish the initial basis for those associations. The emphasis is not on the causal mechanism that creates the associations, but on the empirical identification of patterns of association.

Motivation and Background

Reconsidering the first stages of archaeological analysis requires re-examining some of our basic assumptions about how to approach decoding the archaeological record. We must start with thinking through our initial processes of establishing the linkages between each excavated sample within a site before we consider reconstructing stratigraphic sequences, spatial distributions, or formation processes. These linkages inform many of the initial questions of interpreting field data, such as:

- How many distinct stratigraphic components are reflected by the samples?
- Which layers in one unit correspond to which layers in other units?
- Which artefact types occur together and may represent a deposited assemblage?
- Which assemblages are related, and how?
- Which units contain related layers or materials?
- How many distinct formation processes produced the observed stratigraphy?

Often, these questions and the initial relationships between excavated samples are matters of intuition and professional judgement by the archaeologist. For small or single occupation sites, this is relatively manageable by

direct assessment of artefact inventories and contingency tables, maps, field notes, and stratigraphic section profiles.

Larger or more complicated sites, however, can quickly become intractably difficult. Such sites can present numerous obstacles to assessing the stratigraphic contexts from soil section data. A variety of field and site conditions can lead to ambiguous boundaries between soil strata or other layered deposits. For example, sampling units may not be spatially contiguous. A site may have had multiple overlapping occupations and/or activities, each producing large and diverse artefact assemblages. Different contexts could intrude through or intermingle with previous or contemporaneous deposits. There may be significant post-deposition disturbance due to natural processes, later historical occupations, or modern construction. Soil strata may not have clear and obvious demarcations, or they may consist of multiple lenses of disparate soils.

As these complexities multiply, the cumulative processes affecting the spatial distributions and stratification of the soil matrix quickly obviate any simple assessment and interpretation of stratigraphic contexts. The original linkages between samples are not always apparent. In these cases, it may not be feasible to reconstruct the stratigraphic sequences and their spatial boundaries from soil descriptors or otherwise to determine associations between samples solely by their stratigraphic matrix. A supplemental or alternative diagnostic measure is required.

The artefact content of an excavated sample, though certainly not the only pertinent data, is generally considered an important diagnostic for reconstructing those associations. Whether by typology, known temporal range, or functional or morphological classification, the artefacts provide a significant source of information for identifying the related proveniences and contexts by which the sequence and spatial boundaries of a site's deposits are determined. Much of the early work in statistical archaeology (see historical retrospectives such as Ammerman 1992; Baxter 2003; Djindjian 2015) concentrated on quantitative approaches to identifying just such patterns of association within and between artefact types through seriation, typology, and spatial proximity (e.g., Cowgill 1968; Hodson 1969; Robinson 1951).

Being able to associate groups of artefacts into assemblages is an essential aspect of linking excavated samples to their contexts in conjunction with (or in the absence of) contextual stratigraphic matrices. With the advent of spatially oriented statistics in archaeology (e.g., Hietala & Stevens 1977; Hodder & Orton 1976; Whallon 1973, 1974), and later GIS (see Baxter 2003; Kvamme 1999; Richards 1998), attention shifted to spatial pattern analysis (e.g., Kintigh 1990) and predictive modelling (e.g., Brandt, Groenewoudt, & Kvamme 1992; kvamme 1990) in conjunction with various forms of association methods such as correspondence analysis and seriation. Work on intra-site spatial patterning and *in situ* assemblage analysis has largely continued to build on these foundations, albeit with far more computational power and more sophisticated modelling tools.

Quantitative approaches for assemblage analysis still tend to focus, however, on either specific artefact subsets

(e.g., Baxter, Beardah, et al. 2005; Baxter & Cool 1995; Baxter & Freestone 2003; Madgwick & Mulville 2015) or are mainly concerned with inter-site and regional comparisons of complete site assemblages (e.g., Bevan & Conolly 2009; Lockyear 2013; Peeples & Schachner 2012). The former is largely used for later-stage analyses of subsets of assemblages, while regional analyses generalize type-classes or other diagnostics as a site's complete assemblage profile. Spatial approaches to intra-site analysis have focused on point pattern analyses of artefact or assemblage distributions or on interpolation of planar stratigraphic surfaces (e.g., Barceló & Maximiano 2008; Bevan, Crema, et al. 2013; Crema, Bevan, & M. W. Lake 2010; Fieller 1993; Koetje 1991; Spikins et al. 2002), while truly three-dimensional spatial or GIS analyses are still under development (see van Leusen & Nobles 2018). Much of the work on computational stratigraphy is concerned with a site's geomorphological attributes, component seriation, or specialized applications such as automated ordering of Harris matrices (e.g., Bellanger, Tomassone, & Husi 2008; Herzog 1995, 2001; Herzog & Hansohm 2008). Fewer methods have recently been proposed for holistically evaluating the empirical relationships between both assemblages *and* their stratigraphic contexts simultaneously within a site, or for deriving the assemblages *themselves* from their stratigraphic associations (cf., recent work by Achino & Barceló 2018; Merrill 2015; Merrill & Read 2010).

I don't intend an exhaustive survey of the literature and methods, merely to highlight that determining the patterned structure of deposition within an archaeological excavation has motivated significant progress in quantitative and computational analyses. Intra-site assemblage and stratigraphic analyses, in particular, present distinct methodological challenges. Even under ideal data recovery conditions – i.e., a well-preserved site, clear stratigraphic demarcations, complete feature and component excavation, point-provenience mapping of finds, well-established typologies and chronologies, etc. – it is difficult to untangle the spatial network of relationships within a site. Often, however, it is the less than ideal conditions for which these methods are needed most. Such sites provide, however, the least suitable data for the majority of current approaches.

Most of my work is with just such “less than ideal” sites, and various attempts at adapting quantitative methods to address problematic analyses has led to the methodology presented here. In cultural resource management (CRM) archaeology, it is very common that the terms and scope of excavations are either determined or substantially constrained by non-archaeological considerations. The spatial extent of excavations may be limited by the footprint of potential construction effects, or only a certain percentage of the estimated area of a site designated for systematic excavations for example. Time, schedule, and budgets are always a concern. This, of course, frequently results in partial site excavations, limited samples, non-contiguous excavation units, and incomplete assemblages. Such sites can and should certainly still yield significant information, but the constraints and fragmentary data do limit the options for analysis. These restrictions also accentuate

certain limitations and assumptions inherent in existing methods for spatial and contextual analyses of assemblages or excavated components.

The majority of approaches used to analyse associations within assemblage data consist of a target diagnostic of interest (e.g., particular types or categories of finds) and statistical inferences based on these tabulated type frequencies. Such frequency tables effectively reflect a compositional profile that characterizes each sample and highlights similarities and/or distinctions across these samples. The reliance on contingency tables emphasizes pairwise comparisons, based on various distance measures, within and between types and their sampling unit of analysis (i.e., site, context, activity area, or excavated component).

These approaches entail an assumption that either: 1) the *complete* composition is represented within each compared sample, or 2) fragmentary samples can be positively associated with an already established target profile. For regional or inter-site analysis, a complete profile composition for each sample (i.e., site) may be a valid assumption. For intra-site analysis, however, that would assume an unlikely homogeneity for within-site deposition or require some prior means of aggregating disparate individual samples into complete depositional components. Similarly, having a previously established target profile against which to associate independent fragmentary samples (i.e., those having only a subset of the total composition) assumes prior knowledge of a comparable assemblage. In many cases, determining that complete composition is itself the objective of the archaeological analysis.

Approaches grounded in spatial methods involve various analyses of topology or patterning in the distribution of a target feature of interest (e.g., artefact types, assemblages, soil morphology, or features), generally leading to some form of interpolation of density or probability across the area of study. For spatial analyses, there are a number of conditions and requirements for data that necessitate specific field sampling and recording strategies. Distribution analysis – whether by point patterns, lattices, or surfaces – requires sufficient areal sampling coverage and resolution for the study area to provide representative data. The shape of the distribution for the target feature of interest, and any clustering or patterning to that distribution, is highly sensitive to both the precision of the source data and selection and conceptualization of the target feature. Spatial statistics are predicated on proximity weighting of features across geographic (i.e., planar) space, so the lowest resolution of recording determines the overall quality of spatial analysis.

Most of the available methods of spatial analysis are robust for sites where it is feasible to have large area blocks of controlled excavation, with precise mapping of finds and features, or those that entail substantial regular grids of sample units. For many projects, however, that level of recording is *not* readily available or feasible whether by constraints of time or technology. For other projects, external contingencies or constraints on excavation strategies (e.g., fixed project boundaries, sampling percentages, or physical barriers) do not permit adequate spatial sampling. Without sufficient areal coverage and

resolution, though, many of the standard spatial methodologies for intra-site analysis are not viable.

For both assemblage and spatial approaches, the data requirements of current methods present a serious conundrum for either exploratory intra-site analyses or for problematic sites. Analyses in which the assemblage structure through which to associate finds, the spatial structure through which to associate samples, or both are unknown (i.e., the objective of the analysis) many of the current methods have limited application. Generally one or the other is presumed/required to be previously established by some other means.

The methodology presented here was developed specifically to address intra-site analyses in which there are significant ambiguities in the underlying spatial *and/or* assemblage structures of a site. The work originated in an attempt to identify and evaluate areas of intact stratigraphy for a project (see Davis 2015) that contained portions of a collection of multi-component sites along a highway reconstruction project. The sites were located within a heavily developed area, with substantial stratigraphic disturbances from historical land-use and modern utilities, and the excavations limited to an area within the construction corridor.

Despite these restrictions, the sites were considered historically significant and clearly retained substantial areas of intact archaeological deposits. The spatial constraints and discontinuities as well as post-deposition intermingling of assemblages, however, made their evaluation difficult. My solution at the time was essentially to conduct a rudimentary form of bi-clustering, adapted from weighted gene co-expression network analysis (Langfelder & Horvath 2008; Zhang & Horvath 2005), to identify groups of excavated components with the most similar artefact content and simultaneously identify collections of artefacts that most commonly occurred together. The basic premise was that the co-expression of artefacts was (somewhat) analogous to the co-expression of genes between sample units, with similarly interpretable “traits” (e.g., soil morphology, deposit context, sequence, or activity, etc.) correlating with specific assemblages.

Not only was this approach largely successful in identifying the areas of relatively intact deposition, it also showed that (in many cases) even the disturbed contexts retained at least some indications of their original stratigraphic coherence (Cardinal 2015).¹ Though effective for that project, further experiments showed that a broader implementation required a more generalized form – i.e., one specific to the inherently particular nature and needs of archaeological data. Archaeological field data, due to the nature of the archaeological record and our research objectives, presents distinct challenges to many of the standard assumptions of quantitative statistics as well as incorporating certain inferential constructs within the data acquisition process itself. This creates something of an illusory sense of an intrinsic structure to the data by virtue of excavation methods and recording. This *a priori* illusory structure often conflates the data for and objectives of the intended analysis of *empirical*, and therefore quantitative, structure.

What I found was that it is preferable to consider artefact types for their semantic domains of associations rather than for their independent intensities of co-presence. The information content of artefact types is in their *contextual* domains of associations, not their independent relation to other individual types within the whole collection. In probabilistic terms, the event is the simultaneous associations of all objects within a given stratigraphic context (i.e., the event of the deposit or re-deposit, as a whole), rather than the relative intensity of presence for a single type within that context.

The problem is that considering *types* as independent events or entities, we subsume an *individual* artefact's particular local associations (i.e., those within the object's specific provenience assemblage) into the total site assemblage. Not only does this lose any local specificity of the artefact's associations, but falsely conflates the associations of all objects of that particular type as identical. Therefore, compositional, social network, and other such analyses evaluate the contingent profiles of artefact types between proveniences as though the relative presence of that artefact type were the *only* determinant of its global (i.e., site-wide) associations within any deposit context.

If we consider the association between assemblages and stratigraphic contexts as *linked* networks, and prioritize their total system of associations over comparison of individual types within the site's total composition, then the local contextual information is retained. This allows assemblages and contexts to remain coherent, even if individual types are distributed through multiple contexts or individual contexts reflect multiple assemblages. Essentially, the critical contingency or adjacency matrix that should determine the network of associations between archaeological entities is intrinsically multidimensional – not pairwise. I propose that the presence of any given artefact type should not be considered as an independent variable in the composition of a single context's assemblage.

This small conceptual shift identifies a lower-level space of archaeological interpretation than compositional, type-frequency, or spatial approaches – one for which the local semantics of artefacts within an excavated context inform empirical partitioning of a site into interpretable subsets of assemblages and contexts. Not only does this retain more of the contextual information of artefact deposition, it reveals a substratum to the quantitative evaluation of both assemblages and contexts that might not otherwise be apparent. In short, this additional stage of quantitative evaluation needs to occur prior to the tabulation of contingency tables and spatial analysis that is otherwise overlooked.

Technical Ontology of Archaeological Sites

Ontology, in the broad sense, refers to the philosophical study of the essential nature of things and the relationships between them. Effectively, it is the study of what things *are*, as opposed to what they *mean*, in the sense of their fundamental and categorical existence. Ontology also has a more specific technical connotation, meaning the formal declaration and nomenclature of data entities. This includes formal definitions of properties, attributes,

and relational structure of data related to those entities. The following discussion considers certain aspects of the philosophical ontology of archaeological field data in order to understand their empirical attributes and to establish their technical ontology.

The objective of archaeological study has a certain intrinsic duality in relation to both its past and present existence. Even the most basic terms used in the description of archaeological entities reflect these dual connotations. For example, a *site* refers both to the locus of past activities and to the present disposition of their material remnants. The latter connotation, however, is not directly *equivalent* to the former – one an unobservable set of behaviours and perceptions of place in the past, and the other an observable product of subsequent processes. The fundamental premise of archaeology, of course, is that the final disposition of material evidence related to those past activities (i.e., site-present) can inform *inferences* as to the nature of them (i.e., site-past). Similarly, other common terms in archaeology (i.e., *assemblage*, *context*, etc.) entail just such parallel past/present or observed/inferred dualities.

This dualistic and analogical nature to archaeological practice is, at this point, very well established (e.g., Ascher 1961; Binford 1962, 1981; Lyman & O'Brien 2001; Schiffer 1985; Shott 1998; Sullivan 1978; Wylie 1982, 1985; Wylie & Watson 1992). In practical terms, the evidential support for archaeological analysis (particularly so for quantitative analyses) depends on prioritizing the present, material aspect in order to inform the interpretation of the unobservable past activities and place. Ultimately, however, it comes down to making a clear distinction between what can be observed and what cannot. As discussed above, the nature of archaeological interpretation is inferential and inductive, since our research objective is inherently unobservable. Accordingly, our methods rely on empirical support if those inferences are to have a solid basis for validation.

The emphasis here is on establishing a clear ontology of an archaeological site's elements that isolates the empirical (i.e., directly observable) attributes inherent in field data. We cannot observe the activities that produced the archaeological deposits, nor can we observe the various formation processes that culminate in the observable archaeological record. Basically, we need to work backwards from effects to causes. The goal is to identify which empirical attributes of the archaeological record indicate the underlying structure of its internal associations, and to isolate those attributes that are distinct from the interpretive implications of those observations.

For an archaeological site, this essentially breaks down into five distinct but interrelated entities: 1) a site, 2) the contexts that represent final-stage site formation processes, 3) the excavated components (i.e., proveniences) that constitute the sampling from the site, 4) the assemblage(s) of archaeological material collected from that site, and 5) the artefacts found within the excavated samples that constitute the elements of the assemblage. Each of these relate to the others, but each entails a specific domain of information to be analysed. As will

be discussed below, however, not all of those domains are necessarily appropriate for empirical or quantitative analyses.

Ontology of Space – sites, contexts, and proveniences

The concept of a *site* incorporates, as mentioned previously, both empirical and inferential attributes. In one sense, a site is an empirical unit with a finite spatial extent determined by the overall distribution of related archaeological materials (e.g., R. E. Dewar 1986; Dunnell & Dancey 1983; Hodder & Orton 1976; Willey & Phillips 1958). That aspect can be determined by observation – simply the spatial limits of where archaeological objects and feature are and where they are not (or at least where they cease to be related).

In another sense, the modern entity of an archaeological site is a consequence of both post-deposition formation processes and our field methodologies. An archaeological site is defined by our sampled approximation of an extant material distribution. What remains for us to define as a site is the spatial extent and distribution of artefacts and features that are comprised of the final byproducts of behavior and their subsequent transformation by historical and natural processes (Schiffer 1983, 1987). Only what remains from those processes are directly observable, not the initial deposits nor the subsequent processes themselves.

The more interpretive connotation of site is as the *location* or *place* of past activities, implying a locus of behaviour. This usage of the term is considerably more opaque. The archaeological data by which we attempt to interpret those past activities is, at best, an indirect proxy for them. Furthermore, the original disposition of materials that might be used to define such a locale (i.e., a site) is obscured by both the intervening post-deposition processes and consequences of our field methods. Our quantitative and verifiable measurements of the extent of those transformed distributions are the only empirical attributes. The others are corollary objectives of analysis and interpretation pertaining to past perceptions, use, and delineations of space.

Thus, an archaeological site is in some measure a by-product of excavation and field methodology rather than a direct expression of the processes and materials of its formation. In a very real sense, a site becomes defined as a collection of samples that constitute what is deemed the archaeological record. It is these empirical data choices that become the archaeological record and inform all interpretations that follow. Its empirical attributes are limited to those that archaeologists select to observe and record (i.e., the excavated and documented samples) and the methods used. Clearly, this constrains the empirical utility of site, as an analytical entity, to defining the overall sample space for analysis.

Similarly, *context* refers to the differential spatial associations of related deposits within a site, their formation processes, and their material assemblages. Contexts are comprised of the total spatial extents associated with a particular sequence of deposits and formation processes (see Bailey 2007; Gasche & Tunca 1983; Harris 1989; Kintigh 1990; Östborn & Gerding 2014; Schiffer 1972).

Contexts are also the spatial proxy for processes of deposition and formation, as well as for delineated areas of behavioural activities within the bounds of a site.

Contexts are also empirically problematic. Since the archaeological record of contexts is derived from the same selection of excavated samples as the site itself, contexts share the same problem of empirical ambiguities. Contexts are (empirically, at least) a related subset of the site's samples. Interpreting a context as a unified area of interest, spatially and/or behaviourally, is wholly dependent on verification that samples from it are in fact related to the same underlying formation processes and initial deposits. This renders the empirical utility of context, and the related concepts of activity areas, somewhat problematic. As discussed in the preceding sections, there can be significant difficulties in *a priori* identification of contexts in a complex site. In the empirical sense, the contexts of a site are the unknown entities to be determined from analysis.

The remaining spatial entity in a site's ontology is *provenience*, which is both a unit of spatial location and (more importantly for the current discussion) a distinct and descriptive encapsulation of the data related to an individual excavated sample. While the concept of provenience is most commonly associated with its spatial connotation (e.g., Aldenderfer 1998; Ammerman 1992; Carr 1984; Hietala & Stevens 1977; Kintigh 1990; Koetje 1991; Lyman 2012; McCoy & Ladefoged 2009; Schiffer 1983), the latter connotation as a specific sample from a site and context proves more analytically useful for this methodology.

Effectively, the idea of "provenience-as-sample" rolls together all the information associated with a given excavated component: location, stratigraphic sequence and juxtaposition, geomorphology, and material content. Each of these constituent attributes is an empirical observation, from which both context and site can then be securely built. The data referenced by provenience are the only unambiguously empirical entities related to the association and disposition of an archaeological record. Most critically, the empirical entity of provenience itself (despite its being a product of field methods) is as a unit of collocation for all of the other attributes. There are, of course, theoretical and interpretive implications to proveniences since they are constituent samples from contexts and thereby related to formation processes and activity areas. For the methodology presented below, however, the focus is on their observable content and their utility as a discrete sample observation.

Ontology of Things – assemblages and artefacts

The concept of *assemblage*² shares a similar empirical/interpretive dichotomy to that of site and context. Empirically, an assemblage is an associated collection of objects. What is empirically ambiguous is the nature of that association. There is no directly observable (i.e., empirical) measure by which any given object is positively identified as belonging to a given deposit assemblage.³ That association must instead be derived, post-excavation, from their systematic associations within and between excavated samples (i.e., their field provenience).

As a coherent collection of (potentially) distinct but related objects, an assemblage also represents the material correlates and by-products of a discrete episode of human activities in both time and space (see Bailey 2007; Binford 1962; Clarke 1968; Hamilakis & Jones 2017; Hicks 2010; Schiffer 1987; Thomas 2015). Much like context, the associations of an assemblage are derived from a shared process of deposition. Thus an assemblage, when paired with its associated context, informs the interpretation of the range of activities of the site's occupants and the temporal sequences of those occupations and activities. Although the artefacts and features of an assemblage can be considered empirical observations through the measurement and characterization of individual components, the full extent and composition of an assemblage as a whole cannot be empirically verified prior to analysis.

Association through shared provenience or context (i.e., collocation) might be empirically measurable, but it would necessarily depend on an already determined identification for both the assemblage and its context. The individual objects and features that make up an assemblage do not, of themselves, specifically identify those associations. Therefore, the identification of assemblages, as they are recovered during excavation, is an objective of analysis rather than an empirical datum.

Like all other aspects of archaeological materials, assemblages are equally subject to various post-deposition processes that can introduce uncertainty into their concrete identification and associations. The dispersal and degradation of *in situ* assemblages by those processes create intrinsic ambiguity in the identification of assemblages from field contexts. Just as there are ambiguities in empirical identification of contexts between excavated samples, there is no unambiguous empirical association between assemblages and artefacts.

What can unambiguously be observed *in situ*, however, are the artefacts themselves – i.e., the artefacts, features, and other material remains of the site's occupations and the activities of the occupants. These may be counted, measured, weighed, and recorded in a variety of manners, but it is their physical *presence* at a given location that imparts the information of those other quantitative attributes. The artefacts, and their collocation by provenience, constitute the fundamental source of the empirical archaeological information by which associations can be determined.

The network of associations between uniquely identifiable objects indicate their relationships within both assemblages and contexts, and those associations are observable irrespective of their interpretive implications. Associations by artefact collocations, then, constitute the primary empirical indicators of association for deposit assemblages. These associations establish the analytical utility of other observable measurements of artefacts and their spatial distributions.

Empirical Ontology

Our traditional units of interpretive analysis – *site*, *assemblage*, and *context* – conflate observable data and analytic objectives under singular terms. Essentially, each can only

be considered an empirical source of data *after* those entities have been identified and validated by identifying the associations on which they depend. Although site, assemblage, and context are all derived from and convey aspects of empirical observations, they are inherently subject to secondary validations and conceptualization either by theoretical constructs or as the subject of hypotheses. It would be epistemologically problematic (to the point of logical fallacy) to incorporate the objectives of interpretation as the data of analysis. Conversely, provenience and artefact remain largely independent from supervening theoretical expectations inasmuch as they entail physically observable and measurable empirical attributes.

With respect to quantitative analyses, only unambiguously empirical data is what results from the most basic archaeological practice – i.e., the excavated component (or *provenience*) and its artefacts. These constitute the fundamental, observable data on which *all* subsequent analysis and interpretation of site, assemblage, and context are derived. The spatial unit of collocation (*provenience*) and any object at that location (*artefact*) are observable data irrespective of their interpretive connotations. Therefore, the physical intersection of these two sources of empirical data can be used to impute the descriptive attributes of context, assemblage, and site.

The question then becomes a matter of methods that can impute the appropriate associations – artefact to assemblage, and provenience to context – that can discriminate the optimal partitioning of the total site into analytically verifiable subsets. If the problem is approached in terms of identifying the networks of associations between artefact and provenience, then the answer is a matter of optimal assignment of those elements to their associated collection or set. Rather than trying to identify an unknown assemblage composition profile from type frequencies or discriminate spatial discontinuities between unknown contexts, it would be more appropriate to address the analysis as a *combinatorial* problem in terms of sets and the attributes of their empirical elements. Once those sets and subsets can be determined, then the other characteristics, such as compositions and spatial distributions, can be supported by that empirically grounded set of associations without resort to supervening *a priori* attributions.

Set Theory, Combinatorics, and their Archaeological Analogues

Set theory, or some aspect of it, has played a limited role in archaeological analysis at least since Petrie (1899) first experimented with seriation (i.e., the ordination of partially ordered sets) in the late nineteenth century. Formal applications of set theory, however, were not introduced into the archaeological literature until the 1960s, when Clarke (1968) introduced formal set theory and Venn diagrams as a way to model culture traits and areas. After an initial period of exploration during the 1960's and 70's (see King & Moll 1972, for a review of some of the early work), and a few examples since (e.g., Barceló 1996; Hermon, Niccolucci, et al. 2004; Hermon & Niccolucci 2010; Jarosław & Hildebrandt-Radke 2009; Klein et al. 1983; Merrill & Read 2010; Niccolucci & Hermon 2002; Puyol-Gruart 1999), formal applications of set theory have been relatively rare in

quantitative archaeological analysis. Set theory, although conceptually quite accessible, has more typically been esoteric and featured difficult mathematics in its implementation (see Read 1989).

The computational revolutions in quantitative scientific analyses, including those in archaeology, have both benefited from and been limited by the greatest strength of computers – the ability to quickly perform numerical calculations. To find algorithmic (i.e., computational) solutions largely entails finding a way to translate phenomena and concepts into suitable numerical terms. For archaeological problems, however, those translations are not always obvious. Certain archaeological attributes (e.g., coordinates, measurements, dates, or counts) are obviously numerical in nature. When the objectives are the associations and relationships between those entities, enumeration becomes less obvious. It is easy to forget that numbers are merely symbolic representations of *things*, and mathematics a formal codification of their logical *relationships*.

Several branches of mathematics (e.g., set theory, graph theory, combinatorics, etc.) are specifically concerned with just those sorts of relationships in which quantification is more subtle. The basis of the formalization for the archaeological ontology presented above is a set-based, combinatorial approach. A *set*, in classical set theory, is simply a collection of unique, unordered elements that itself constitutes a distinct entity. Combinatorics is the mathematical study of finite sets, which are sets with a fixed and finite number of elements. Specifically, combinatorics deals with subdivisions, structure, enumeration, and arrangements of smaller subsets within those finite sets. Using these concepts, we can easily describe a site as just such a finite set. Assemblages, contexts, and proveniences can also then be described as subsets of that set, while individual artefacts would constitute its elements.

There are numerous approaches to solving set and combinatorial problems. The following sections present a brief review of set theory covering some of the basic properties of sets, set operations, and an introduction to two generalizations of classical set theory – multisets and fuzzy sets – that have useful potential for applications to archaeological problems. In addition, a generalization on graph theory (i.e., hypergraphs) is introduced that has applications both for visualization of the complexity of these types of sets and for methods specifically designed to evaluate the intersecting areas of such graphs as an independent means of partitioning complex sets.

Table 1 summarizes some of the basic concepts, properties, and operations with their common symbolic notations. Some of this material will be familiar to those with backgrounds in computer science, data structures and queries, or probability theory but is presented here in order to provide nomenclature and context for the proposed methodology that follows.

Classical set theory

A *set*, in classical set theory, is simply a collection of unique, unordered elements that itself constitutes a distinct entity. This definition of set, intuitively, is a natural analogy to various archaeological concepts such as assemblages and typologies. In the broader sense, though, the

Table 1: Definition and notation for common set properties and operations.

Operation	Symbol	Example	Definition
Set	$\{...\}$	a, b, c, d	A collection of elements
Cardinality	$... $	$ a, b, c = 3$	# of elements in a set
Empty set	$\emptyset$	$\{\}, \emptyset =0$	A set with no elements
Universe	Ω or $\mathcal{U}$	$\Omega(\{a, b, c\}) = \{a, b, c, d, \dots, z\}$	Set of all possible elements in a set's domain
Union	$\cup$	$\{a, b, c\} \cup \{c, d\} = \{a, b, c, d\}$	Elements in either <i>A</i> or <i>B</i>
Intersection	$\cap$	$\{a, b, c\} \cap \{c, d\} = \{c\}$	Elements in both <i>A</i> and <i>B</i>
Symmetric Difference	Δ	$\{a, b, c\} \Delta \{c, d, e\} = \{a, b, d, e\}$	Elements in either <i>A</i> or <i>B</i> , but <i>not</i> in both
Multiset Sum	$\uplus$	$\{a^1, b^2\} \uplus \{a^2, b^1\} = \{a^3, b^3\}$	Sum of the multiplicities of multisets
Subset, Superset	$\subseteq, \supseteq$	$\{a, b\} \subseteq \{a, b, c\} \subseteq \{a, b, c\}$	<i>A</i> contains some, or all, elements in <i>B</i>
Strict Sub-, Superset	$\subset, \supset$	$\{a, b\} \subset \{a, b, c\} \not\subset \{a, b, c\}$	<i>A</i> contains some, but <i>not</i> all, elements in <i>B</i>
"in..."	$\in$	$a \in \{a, b, c\}$	<i>a</i> is an element <i>in</i> a set
"for all..."	$\forall$	$\forall x \in \mathbb{N} = \{0, 1, 2, 3, \dots, \infty\}$	Logical statement used in set declarations
"and"	$\wedge$	$\forall x \in \mathbb{N} \wedge x \leq 2 = \{0, 1, 2\}$	Logical statement used in set declarations
"such that..."	$ $ or $:$	$\forall x \in \mathbb{N} 0 < x \leq 2 = \{1, 2\}$	Logical statement used in set declarations
Complement	A' or A^c	$A' = \{\forall a a \notin A\}$	All elements <i>not</i> in <i>A</i>
Powerset	$\wp(A)$	$\wp(\{1, 2\}) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$	All possible subsets of a set

elements of a set can be any grouping of related entities. The system of notation and mathematical operations that have evolved to describe the properties, relationships, and interactions of sets are what allows translation of abstract elements and memberships into computationally useful algebraic forms. Of particular importance are a set's basic properties – its size or number of elements, how element membership is defined, and the domain from which the elements are drawn. Relationships between sets are determined by intersections of their shared elements, and their interactions by various operations on those elements (see **Figure 2**).

It is first necessary to define a set's *universe* or *universal set* (commonly denoted as $\mathcal{U}$ or Ω), consisting of the total domain of all possible elements. A universe can be infinite (e.g., all real numbers $\mathbb{R}$, all positive integers $\mathbb{N}$, etc.) or finite (e.g., letters of the alphabet, positive integers between one and ten), depending on the defined domain of the set's universe.

The size, or *cardinality*, of a set is simply the number of elements in the set. For example, a set defined as $A = \{a, b, c\}$ would have a cardinality of three and written as $|A| = 3$. Two sets are equal only if they share all elements in common, regardless of the order of elements (e.g., if set $A = \{1, 2, 3\}$ and set $B = \{3, 2, 1\}$, then $A = B$).

A set may also be a member or *subset* ($\subset$) of another larger set if all elements of the subset are contained by the other (e.g., if set $A = \{1, 2\}$ and set $B = \{1, 2, 3\}$ then $A \subset B$). The *complement* of a set *A*, denoted by either A' or A^c , is defined as the set of all elements *not* in that set. The union of any set and its own complement $A \cup A'$ is, by definition, its universe Ω .

Various basic operations follow from these properties – unions, intersections, and differences. The *union* ($\cup$) of sets is the set formed by the unique combination (removing

duplication) of the elements of the sets (e.g., if $A = \{a, b, c\}$ and $B = \{b, c, d\}$, then $A \cup B = \{a, b, c, d\}$). The set of shared elements between two or more sets is the *intersection* ($\cap$) of the sets (e.g., if $A = \{a, b, c\}$ and $B = \{b, c, d\}$, then $A \cap B = \{b, c\}$). Conversely, if two sets share no elements in common, then their intersection is the empty set $\emptyset$ (i.e., a set with no elements) and they are called *disjoint* (e.g., if $A = \{a, b, c\}$ and $B = \{d, e, f\}$, then $A \cap B = \emptyset$). *Difference* is simply the elements of a set that are not shared with other sets. A special form of difference, the *symmetric difference* (Δ), is the difference between the union and intersection ($A \cup B - A \cap B$), and consists of only the elements in *A* and *B* that are not common to both. **Figure 2** further illustrates several common unions and intersections for two intersecting sets *A* and *B*, their universe Ω , and their complements A' and B' .

Those are some of the basic properties of sets, but two further properties bear directly on the archaeological methodology presented below. Firstly, the elements of a set may themselves be sets. More specifically, a set of sets (or *family* of sets) may be defined as the collection of subsets across a given set. For example, given a set $A = \{a, b, c, d, e\}$, a family of sets $\mathcal{S}$ can be defined over *A* as $\mathcal{S} = \{\{a, b\}, \{c, d\}, \{e\}\}$. Although this property has numerous applications, one is particularly interesting with regard to set partitioning – the *power set*. The power set of any given set *A* – commonly denoted as $\mathfrak{P}(A)$, $\mathcal{P}(A)$, or $\wp(A)$ – is the family of sets consisting of *all possible* subsets of the given set including the empty set $\emptyset$. For example, the powerset of a set $\{1, 2\}$ would be $\wp(\{1, 2\}) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$. In this way, the power set indicates the upper bound for the number of possible partitions of the set into subsets.

Using these concepts, it becomes relatively simple to translate the archaeological site ontology described

Figure 2: A Venn diagram of unions and intersections for two sets, A and B and their complements, within a universe $\Omega_{A,B'}$

previously in terms of sets. We can define a site in terms of a set S whose elements are all artefacts collected from the site. This would be the sample space from the total domain, or universe, Ω_s of elements. We can similarly define a provenience as a subset of that domain, $P \subset S$, whose elements are the artefacts found at that location. Assemblages, naturally, would also be subsets of the site $A \subset S$. Since set elements can also be sets themselves, we can further define the site in terms a family of sets whose *elements* are P , making contexts a family of sets $\mathbb{C}$ that are also a subset of S .

Multisets and repeating elements

One significant limitation of the classical set concept is that the elements of a set must be *unique*, meaning that an element must appear only once in a set without repetition. Going back to our archaeological ontology, this presents a problem in that there are often multiple artefacts of any given type with the set of a site or provenience. There is, however, a generalization of classical sets that does allow for repeated elements. This kind of set is called a *multiset*. The concept of a multiset is merely

a set defined such that each element may be repeated (Blizard 1988, 1991). In a multiset, each element of the set is actually composed of two values – the element and its multiplicity or number of repetitions (see Blizard 1988; Singh, Ibrahim, et al. 2007; Syropoulos 2001; Tzouvaras 1998; and Hickman 1980, for thorough background). Also called an *mset* or *bag* in some literature, it is often used (albeit implicitly) in textual analysis (e.g., a “bag of words” analysis analyses the association rules of terms and each term’s frequency).

Each element of a multiset $\mathcal{A}$ is comprised of an ordered pair consisting of an element a and its multiplicity $m(a)$. Notation for elements in a multiset is rather variable, so the paired values (more formally known as a 2-tuple) is commonly rendered either as an ordered pair in the form $a, m(a)$ or as scripts in the form $a^{m(a)}$. A multiset $\mathcal{A} = \{a, a, a, b, c, c\}$, for example, could be denoted as either $\mathcal{A} = \{(a, 3), (b, 1), (c, 2)\}$ or $\mathcal{A} = \{a^3, b, c^2\}$. Each multiset also has an associated set, known as the *support* or *root*, which is a classical (i.e., non-repeating) set consisting of the unique elements without their multiplicities. The support set is formally defined as all x , where x is an element in the

universal set Ω_A of multiset $\mathcal{A}$, such that the multiplicity m_A at x is equal to or greater than one. In formal notation, this would be written as $\text{Supp}(\mathcal{A}) = \{\forall x \in \Omega_A | m_A(x) \geq 1\}$. For our example multiset $\mathcal{A}$ above, the support set would be $\text{Supp}(\mathcal{A}) = \{a, b, c\}$.

The cardinality (i.e., size) of a multiset is simply the sum of all $m(a)$, so in the previous example the cardinality $|\mathcal{A}|$ would be $3 + 1 + 2 = 6$. In addition to a multiset's cardinality, it also has an attribute of *dimension*, which is simply the cardinality of its support set (e.g., $|\text{Supp}(\mathcal{A})| = 3$). In other words, a multiset is determined by both the number of unique elements *and* the total multiplicity of those elements.

As a generalization of classical sets, multisets have equivalent (although also generalized) forms of the various set operations described in the previous section (i.e., union, intersection, powerset, etc.). Since multisets need to consider element multiplicities, however, the operations and their expressions are somewhat different. For example, unions and intersections need to consider a multiset's repeated elements and to specify subset relationships explicitly. This is needed in order to distinguish whether a union operation results in the sum of multiplicities (if the two multisets are declared disjoint) or the greater multiplicity (if one multiset is a subset of the other). In this case, the use of the multiset sum ($\uplus$) is generally used to indicate the former and the standard union ($\cup$) the latter (e.g., $a^1 \uplus a^1 = a^2$, but $a^1 \cup a^1 = a^1$, see **Figure 3**).

For intra-site analysis, typically we are not starting with the entire collection of all artefacts contained within the

site (i.e., the unknown universe Ω_s from which the site's collection is sampled). The site's *collected* assemblage is only a subset (i.e., the site sample $\mathcal{S}$) of Ω_s comprised of the *union* of all samples from the excavated proveniences. Each provenience contains some even smaller subset of Ω_s . Obviously, then, the assemblage data to be assessed (in terms of sets) is merely the sample space $\mathcal{S}$, drawn from the unknown total population Ω_s , comprised of each individual sample.

The objective of intra-site assemblage analysis is to find the natural combinations of these proveniences (i.e., sample subsets) to find whether there is a natural partitioning of the site's assemblage $\mathcal{S}$ into archaeologically interpretable subsets. It is possible (even likely) that those natural assemblage subsets could intersect within $\text{Supp}(\mathcal{S})$, since certain artefact types may be part of more than one such assemblage.

By addressing archaeological assemblages as multisets, however, these interrelationships between site, assemblage, and provenience can be formally specified for algorithmic expression while remaining archaeologically intuitive. If we consider site assemblages as multisets, then the mapping from the archaeological entities to multisets further simplifies our set-based ontology. Furthermore, a multiset's multiplicity and support set introduce a clear method for relating artefact to typology.

We can now define a site's assemblage as a multiset $\mathcal{S}$ consisting of all collections from the site. Each artefact type and its frequency could be represented as an element $a_i^{m(a_i)}$ in $\mathcal{S}$. Each type a_i is a member of the

Figure 3: Difference between set union $\cup$, multiset union $\cup$, and multiset sum $\uplus$. Note the handling of intersecting elements (element C).

support set $Supp(S)$, and the artefact frequencies as each type's multiplicity $m_s(a_i)$ within S for all a_i . Similarly, we can also define each provenience's collection of artefacts as multisets in the same way. This is, in practice, not much different from common type-frequency inventories, and can easily be translated from one format to the other.

Fuzzy sets and partial memberships

Another limitation of classical set theory is that an element's membership in any given set is binary – i.e., an element either *is* or *is not* a member of that set. There are a number of contexts, however, in which an element's set membership is not necessarily so well-defined. For an archaeological example, consider a particular artefact type that may have been in use across multiple periods of occupation. Such an artefact type could potentially be associated with more than one assemblage or context in a multicomponent site. In that case, the object's membership would be ambiguous, so could not reasonably be assigned to a single set. This is a common problem when conducting most standard forms of cluster analysis.

Yet another generalization of classical sets specifically addresses this issue of partial, multiple, or ambiguous membership for an element in a set – a *fuzzy set*. Fuzzy sets are superficially similar to multisets in that each element is comprised of a pair of values, but are generalized in order to address an element's *partial* membership in a set rather than its repetition (see Dubois & Prade 2015; J. Lake 1976; Zadeh 1965, 1996, 1997). In the case of the fuzzy set, instead of a multiplicity there is a *membership function* $\mu(x)$. A membership function has a value between zero and one that indicates the degree of membership of that element to the set – zero meaning no membership in the set, and one meaning complete membership. A fuzzy set, similarly to a multiset, has a related support set consisting of a classical set containing all elements of the fuzzy set for which $\mu(x) > 0$.

The membership function allows for indistinct boundaries between sets. This type of set was derived specifically to deal with circumstances in which an element's membership in any given classical set is not determined by a simple binary (0,1) or Boolean (True, False) decision boundary. In classical set theory, an element either is or is not a member of a given set. In fuzzy set theory, such binary membership is called a “crisp” set. An element's membership in a fuzzy set, by contrast, consists of a gradient *between* zero and one.

The membership function for a fuzzy set should not, however, be simply viewed as a probability or proportion of membership. These would, in terms of fuzzy theory, be simple determinants derived from the composition and distribution of the sample space itself. Instead, membership is an ascribed and contextual zonal gradient, indicating the relative “truth” of the element's assignment to that set. The degree of “truth” is determined relative to other defined sets. The combined *field of discourse* of the sets, defined as the input space consisting of the total set of possible memberships, effectively maps to a partitioning of the universe of the sets Ω (as defined earlier).

In this way, fuzzy sets can be defined to provide a set membership determinant in circumstances where the decision boundary entails some ambiguity, such as an element's “proximity” (so to speak) within the mapping of the universal set's sample space. Since the membership function is not necessarily a function of the elements, instead being imposed based on an ascribed distribution within the field of discourse, the set of output values of the membership function may itself be a fuzzy set.

Common examples of simple fuzzy set memberships are the assignment of certain continuous values to nominal classifications such as “short, tall” or “small, medium, large” that have indistinct boundaries but concrete connotations. In the short-tall example, a person that is four feet tall is generally considered “short” compared to a person that measures six feet or more in height. A “crisp” membership cut-off of “ $tall \geq 6ft$ ” would, however, nonsensically render even an individual measured at a thousandth of an inch below six feet as “short”. A fuzzy membership function, on the other hand, could be assigned based on the proximity of values to that six-foot cut-off.

Although the membership function $\mu(x)$, explicitly, is not a probability or proportion on set membership, it could be possible to derive from the membership function a contingent probability for an entity's membership given other elements of the set. Going back to archaeological analyses, the logics of fuzzy sets share a certain intuitive association with traditional archaeological classifications for occupation periods (e.g., “Early Woodland” or “Late nineteenth century”), phases (e.g., “Formative”, “Classic”, and “Post-Classic”), and activities (e.g., “Subsistence”, “Domestic”, “Ritual”, etc.).

Whereas multisets address repetition of artefact types, fuzzy sets address ambiguous categorical boundaries (see Barceló 1996, 2008; Baxter 2009; Niccolucci & Hermon 2002). The logic of fuzzy sets allows for the ambiguous determinacy that could arise from potential membership between sets. Multiset multiplicities *could* be apportioned between sets, but they would remain “crisp” in the sense that their membership in a set is a binary true-false proposition. Fuzzy sets, while dependent on choices of membership function, more closely parallel the often vague boundary conditions inherent to archaeological assemblages and spaces (e.g., Barceló 1996; Hermon, Niccolucci, et al. 2004; Jarosław & Hildebrandt-Radke 2009; Klein et al. 1983; Niccolucci & Hermon 2002).

Fuzzy multisets – multiple partial membership

While multisets allow repetition of elements and fuzzy sets allow gradient boundaries of element membership within a given set, There are also situations in which a collection of elements might exhibit *both* multiplicity and partial membership. Examples of this sort of ambiguity commonly arise in fields such as textual analysis and natural language processing. Consider a scenario in which a word may appear multiple times in a text but not always have the same connotation, such as the case with homonyms (e.g. “bat”, “cup”, “crop”, or “right”). In a sense, not all

instances of these words can truly be considered the same word. Such words are completely distinct in both meaning and function depending on their context. In that case, it is not merely a matter of the word's relative membership to a particular set, but also distinguishing between the pertinent fields of discourse for each instance of the word. In short, there would need to be separate membership functions for each possible field of discourse associated with each repetition of the word.

In archaeological applications, this may occur for artefact types that have multiple possible categorizations depending on their context. Take, for example, the possible categorizations of a porcelain sherd in historical deposits. A small porcelain sherd could be refined tableware, or a piece from a toy doll or child's tea set. A large chunk could be an electrical insulator or piece of a toilet. Each of these instances of 'porcelain' would have a completely different field of discourse depending on assemblage context. Such artefact types would require a multiplicity of membership functions to each context.

There is an additional generalization of fuzzy sets that allows multiple membership functions for an element – i.e., fuzzy multisets (see Miyamoto 2001, 2004; Syropoulos 2001, 2012; Yager 1986; and Singh, Alkali, & Ibrahim 2013, for thorough treatment of fuzzy multisets and their history). A fuzzy multiset permits a multiplicity of contextually independent membership functions. The idea of fuzzy multisets was developed specifically to deal with issues of multiple relations in the classification of set elements, where those elements may have natural associations within *multiple* fields of discourse.

A fuzzy multiset element has an associated membership function depending on its local embedding, but also has additional and contextually differentiated memberships in other separate and distinct fields of discourse. In that way, a word for which the context determines the specific meaning or connotation is associated with the word itself in the lexical support set of the corpus' multiset, but additionally has separate relative gradients of membership associations dependent on its contextual embedding.

In many ways, the concept of fuzzy multisets could address the various archaeological caveats noted in the preceding sections. They allow for both the multiplicity required to assert site assemblages as sets of independent discrete artefacts with related types, and to assert the vague boundaries of assemblage and context membership for any given artefact type. Multisets address the repetition of artefacts within a type, but do not allow for ambiguous categorical membership. Fuzzy sets address categorical ambiguity for types, but do not allow for multiplicity. Fuzzy multisets provide a generalized framework that incorporates both, allowing classification schema that potentially account for more contextually sensitive categories.

Comprehension of complex sets with hypergraphs

In discrete mathematics, set theory and graph theory are closely related. A graph is, in its most general sense, just a representation of relationships and associations. In its

most typical form, a graph is a collection of nodes (also called vertices), which are connected by edges indicating a relationship between each pair of nodes. Graphs can also, however, be considered in terms of sets – a given graph can be defined as a system of sets $G = (V, E)$ consisting of a set of the graph's vertices $V = \{v_1, \dots, v_n\}$, and a set of edges $E = \{e_1, \dots, e_n\}$ comprised of paired subsets $e_n = \{v_i, v_j\}$ of the vertices. While graphs are often employed as a means of visualizing these relationships, certain mathematical properties of graphs allow for highly efficient algorithmic solutions for otherwise difficult computational problems.

The recent archaeological interest in social network analysis is, in part, a consequence of this algorithmic efficiency. The major advantage of network analysis is in using these innate properties of graphs to determine edge adjacency between nodes, rather than more computationally intensive methods for determining proximity and relationships between data points. These efficiencies can also be applied to combinatorial problems, so offer a potential suite of tools for addressing the archaeological ontology as well. The limitation here, however, remains the issue concerning the *paired* subsets of vertices. As discussed previously, the relationships of interest for determining associations within the archaeological ontology are not necessarily limited to their pairwise correspondence. Graphs of the form described are not easily suited to more complex sets of relations.

The concept of a *hypergraph* is a generalization on graphs that permits graph edges to simultaneously connect an arbitrary number of vertices (see **Figure 4**), whereas a standard graph edge allows the relation of exactly two vertices. Hypergraphs largely came to prominence along with developments in computer science. The increasing need to model more complex information relationships, such as data structures and computer networks, necessitated a more general formulation of traditional graph theory in discrete mathematics to model the more complex systems (see Berge 1973, 1984; Berge & Ray-Chaudhuri 2006; Bretto 2013).

Essentially, a hypergraph provides a generalized algebra for representing the various interactions between sets of sets. A hypergraph H is a topological representation of both a set of vertices $V(H) = \{v_1, \dots, v_n\}$ and the family of sets of edges $E(H) = \{e_1, \dots, e_n\}$, so that every edge e_i is a subset of $V(H)$. This is no different than the graph definition above, with the exception that the hypergraph's edges may consist of an *arbitrary* number of elements. Since all edges consist of a subset of the set of vertices, every edge is a member of the power set of the set of vertices. Importantly, these edges may be multisets for more complex interactions in higher dimensions.

Aside from the mathematical methods and implications, hypergraphs also provide a powerful visualization tool for sets of complex relationships (see Bahmanian & Šajna 2015; Bretto 2013; Paquette & Tokuyasu 2011; Zhou, Huang, & Schölkopf 2007). Since vertices and edges (also known as *links*) represent set elements and set relationships between groups of those elements, a hypergraph can visually convey complexities in the organizing structures

Figure 4: Example hypergraph H with seven vertices/nodes $V(H) = \{v_1, \dots, v_7\}$ and four edges $E(H) = \{e_1, \dots, e_4\}$, where each edge e_i connects a subset of vertices v_i (User:Kilom691 2010).

within intersecting data. They are also versatile enough as algebraic formalizations to provide a full complement of analytical techniques. Based in the geometric study of graphs and their topologies, they constitute their own specialization in discrete mathematics.

Of particular interest, in regards to site and assemblage analyses, are methods of decomposing⁴ large and highly complex hypergraphs into simplified or reduced subsets of the hypergraph (e.g., Agrawal et al. 1998; Ahlswede et al. 1997; Bárány 2005; Björklund, Husfeldt, & Koivisto 2009; M. Dewar et al. 2016; Keevash & Mycroft 2015). These partitioning methods are related to *semi*-supervised machine learning procedures, in which only a subset of classifiers (i.e., the identifiers of set or group membership) within a sample space are known (e.g., a site assemblage in which only a small subset are temporally diagnostic). These reduced (or *contracted*) hypergraphs are sub-graphs, partitioning the hypergraph's edges, that retain the information on the overall topology and structure of the source graph. This allows structural inferences for the full, non-reduced hypergraph by analysis of the reduced graphs without significant information loss and yet involving less computational complexity.

In terms of its application to the archaeological ontology as defined as a system of sets, consider a hypergraph of an archaeological site constructed in two possible ways – one in which the vertices are the artefacts, and one in which the vertices are proveniences. If we construct a site hypergraph consisting of artefact vertices, then the hypergraph edges constitute the relationships of artefacts within assemblages (i.e., the associated sets of artefacts), with the edges determined by their collocations through shared sets of proveniences. Conversely, a site hypergraph of provenience vertices would render the edges as contexts (i.e., associated sets of proveniences), with the edge

sets determined by shared artefact content. What remains to be determined, then, is the criteria for finding optimal solutions for assigning those edge memberships.

Combinatorial Formalizations of Archaeological Ontology

The previous discussion touched on some of the parallels between the ontology of archaeological field data and several aspects of sets. Conceptually, at least, the traditional views and treatment of archaeological assemblages, as collections of discrete objects, fit well with the logic of set theory. We can begin to formalize a mathematical structure of archaeological assemblages, using standard set notations from these initial premises, to describe the content of excavated proveniences on its own.⁵ More importantly, the formalization establishes a set of very specific entities, and their partitioning criteria, from which algorithmic methods can reasonably be derived. Since the empirical content of the site assemblage ontology is located in the artefacts and proveniences, these are the basis for formalization on which assemblage, context, and site are dependent.

Formalization as sets

First, let's stipulate that the total sampled collection of all artefacts at a site constitutes a fixed, finite multiset $\mathcal{S}$. This is the sample space for the complete site assemblage. Each element in $\mathcal{S}$ corresponds to an artefact drawn from an unknown universal set $\Omega_{\mathcal{S}}$. Archaeologically, $\Omega_{\mathcal{S}}$ is the unknown total set of all artefacts deposited during the site's occupations. Site excavations are typically a sampling of the total site area and assemblage. The collected sample $\mathcal{S}$ from those excavations is therefore expected to constitute a representative subset of that total site deposition assemblage $\Omega_{\mathcal{S}}$ (i.e., an unknown 100% sample).

Furthermore, each excavated provenience contains an individual sample of artefacts drawn from that unknown total site population Ω_S . The total sampled collection $\mathcal{S}$ is merely the sum of those provenience samples. So we can define a provenience's collected sample to be an indexed⁶ multiset $\mathcal{V}_i$ (for $i = 1$ to n proveniences), containing artefact types a_i with frequency $m_{\mathcal{V}_i}(a_i)$. The multiset $\mathcal{V}_i$ defined to be "for all ($\forall$) artefacts of a type a_i in ($\in$) the site's total artefacts (Ω_S), occurring $m_{\mathcal{V}_i}(a_i)$ times, where ($|$) the number of occurrences of that type $m_{\mathcal{V}_i}(a_i)$ is an integer ($\mathbb{N}$) greater than or equal to one." The total site sample collection $\mathcal{S}$ is then the multiset sum ($\uplus$) of all $\mathcal{V}_i$. In formal notation:

$$\begin{aligned} \text{Provenience } \mathcal{V}_i &= \{a_1^{m_{\mathcal{V}_i}(a_1)}, \dots, a_n^{m_{\mathcal{V}_i}(a_n)}\} \\ &= \{\forall a_i \in \Omega_S \mid m_{\mathcal{V}_i}(a_i) \in \mathbb{N} \geq 1\} \\ \text{Site } \mathcal{S} &= \{a_1^{m_S(a_1)}, \dots, a_n^{m_S(a_n)}\} \\ &= \uplus_{i=1}^n \mathcal{V}_i \end{aligned}$$

Since $\mathcal{S}$ is the just the combination of all $\mathcal{V}_i$, then each $\mathcal{V}_i$ would naturally be a member of the power set of $\mathcal{S}$. The total collection of proveniences therefore describes a specific family of sets $\mathbb{V}$ over the site $\mathcal{S}$ (i.e., a family of sets that is a subset of the power set). Every $\mathcal{V}_i$ that exists⁷ ($\exists$) is a member of the site's powerset $\wp(\mathcal{S})$, but cannot be the empty set $\emptyset$. The multiset sum ($\uplus$) of all members $\mathcal{V}_i$ of $\mathbb{V}$ must be equal to $\mathcal{S}$, and ($\wedge$) each $\mathcal{V}_i$ must be a disjoint subset of artefacts from $\mathcal{S}$ (i.e., an artefact cannot be in two locations). Therefore, the intersection ($\cap$) of all $\mathcal{V}_i$ would have to be the empty set $\emptyset$:

$$\begin{aligned} \text{All proveniences } \mathbb{V} &= \{\mathcal{V}_1, \dots, \mathcal{V}_n\} \\ &= \{\exists \mathcal{V}_i \subseteq \wp(\mathcal{S}) - \emptyset \mid \uplus_{i=1}^n \mathcal{V}_i = \mathcal{S} \wedge \bigcap_{i=1}^n \mathcal{V}_i = \emptyset\} \end{aligned}$$

The goal of the intra-site analysis is to find a natural partitioning of $\mathcal{S}$ into subsets of artefacts (i.e., assemblages $\mathcal{A}$) that have a one-to-one correspondence with a natural partitioning of $\mathbb{V}$ into subsets of proveniences (i.e., contexts $\mathcal{C}$). In other words, each assemblage $\mathcal{A}_i$ consists of a subset of artefact types a_i from $\mathcal{S}$ that are related through a set of proveniences $\mathcal{V}$ that are a subset of $\mathbb{V}$. Simultaneously, each context $\mathcal{C}_i$ consists of a subset of proveniences $\mathcal{V}$ that are related through their artefact types a_i . Neither $\mathcal{A}$ nor $\mathcal{C}$ are known, so the objective is to determine whether such natural pairings can and do exist within the observed collections $\mathcal{S}$ and $\mathbb{V}$.

If we consider an assemblage to be a subset of a site's artefacts indicative of an occupation and/or activity, and consider a context to be a subset of proveniences containing the assemblage deposited by that activity, then we're looking at evaluating two specific families of sets over the two related domains – the total domain $\mathcal{S}$ of artefacts, and the total domain $\mathbb{V}$ of proveniences. The site's assemblages would be a family of sets $\mathbb{A}$ over $\mathcal{S}$ for which there also exists an equivalent (or nearly equivalent) family of

sets $\mathbb{C}$ over $\mathbb{V}$ comprising the site's contexts. Each assemblage $\mathcal{A}_i$ and each context $\mathcal{C}_i$ is (ideally, but not always) disjoint from all others. Perfect or ideal partitioning of a site is indicated when, for all $\mathcal{V}_i$ in $\mathbb{C}$, each $\mathcal{V}_i$ is a subset of $\mathcal{A}_i$ and the multiset sum $\uplus \mathcal{V}_i \in \mathcal{C}_i$ equal to a member $\mathcal{A}_i$ in $\mathbb{A}$.

More explicitly, the objective is to find a member of the powerset $\wp(\mathcal{S})$ that is equivalent to the multiset sum ($\uplus$) of a member of the powerset $\wp(\mathbb{V})$. In this case, it is a matter of defining a system of sets $(\mathcal{A}, \mathcal{C})$ for which there exists ($\exists$) some unknown set $\mathcal{X}_i$ in ($\in$) $\wp(\mathcal{S})$ and also exists ($\exists$) a related unknown set $\mathcal{Y}_i$ in ($\in$) $\wp(\mathbb{V})$ such that ($|$) $\mathcal{X}_i$ is equal to (or, more commonly, approximately equivalent $\approx$ to) the multiset sum ($\uplus$) of all $\mathcal{V}_j$ in ($\in$) $\mathcal{Y}_i$, which can be written:

$$\begin{aligned} (\mathcal{A}, \mathcal{C}) &= \{\exists \mathcal{X}_i \in \wp(\mathcal{S}) - \emptyset, \\ &\quad \exists \mathcal{Y}_i \in \wp(\mathbb{V}) - \emptyset \\ &\quad \mid \mathcal{X}_i \approx \uplus_{j=1}^n \mathcal{V}_j \in \mathcal{Y}_i\} \end{aligned}$$

The more common cases are ones in which artefacts from different natural assemblages are found within the same excavated component, stratigraphic layers are mixed or inverted, later features cut through older deposits, and the spatial distribution of materials is anything but discrete. In these cases, there would be no complete solution in which assemblage or context families of sets would be internally disjoint. Instead, they could instead be *almost disjoint* (i.e., the cardinality of the intersection is much less than the cardinality of either set so that given two sets A and B then $|A \cap B| \ll |A|, |B|$), suggesting that an algorithmic solution should search for an optimally minimum intersection rather than the empty set $\emptyset$.

It is important to consider that a site, as excavated, reflects the cumulative spatial disposition of various sequences of deposit and re-deposit. All of the original assemblages will be represented in the samples, but contexts may become conflated by later intrusions and/or post-occupation disturbances. The formalization presented above is assessing the natural partitions within that final *cumulative* disposition. By further analysing the discrepancies between the final partitioned contexts and assemblages, the degree of assemblage admixture – i.e., the integrity of the site – can be quantitatively determined.

Formalization as hypergraph

With the natural partitioning of a site defined as the system of sets $(\mathcal{A}, \mathcal{C})$, it is also possible to represent these same definitions as the elements of a hypergraph $H(V, E)$. If we take the elements of $\mathcal{S}$ (as defined above) as the set of all vertices V of the graph, and the edges E to be the elements of each provenience $\mathcal{V}_i \in \mathbb{V}$ (i.e., the family of sets $\mathbb{V}$), then the hypergraph $H(\mathcal{S}, \mathbb{V})$ fully describes the total collected samples of the site and their empirical relationships of collocation.

Construed as a graph, the partitioning of a site's assemblages and contexts entails the identification of discrete communities or *cliques* determined by their network of

edges. This raises interesting implications for a computational solution, since graph problems tend to be significantly less difficult to solve. Given the definition of $(\mathcal{A}, \mathcal{C})$ in terms of $\mathcal{S}$ and $\mathbb{V}$ as shown above, then each $(\mathcal{A}_r, \mathcal{C}_r)$ would by definition be equivalent to a maximum disjoint clique and sub-graph of the hypergraph $H(\mathcal{S}, \mathbb{V})$.

Implications of Combinatorial Formalization

The basic premise of archaeological field practice is to record all necessary information to associate archaeological features and artefacts with their spatial contexts. Similarly, the basic premise of intra-site assemblage analysis is to discern the patterning of classes of artefacts within those spatial contexts in order to infer patterns of activities. The goal of each, essentially, is to provide a means of identifying coherent groups that provide interpretable information within both the total artefact assemblage and spatial units. Since the ideal group membership is an unknown, this is effectively a problem of simultaneous clustering across spatial units and assemblage artefact types – which artefact types consistently occur together in space, and which spatial units consistently contain similar artefact types (or appear to be drawn from the same source assemblage). There are a number of existing methods that can be (or have been) applied to this problem, but the specific nature of archaeological assemblages do present certain difficulties.

As previously noted, archaeological data is rarely “clean” in the sense of being discretely compartmentalized by occupation assemblage or context. Assemblages become mixed within stratigraphic contexts both through occupation activities and post-deposition disturbances. If we suppose, though, that these overlapping interactions of assemblages and contexts may be represented by the membership functions of fuzzy multisets within the site’s field of discourse,⁸ then the formal methods for this type of set entail mechanisms for ascribing the appropriate degree of set membership. Initial partitioning of the site’s total assemblage multiset into assemblages and contexts would inherently identify these interactions via the topological overlap (Li & Horvath 2007; Song, Langfelder, & Horvath 2012; Yip & Horvath 2007; Zhang & Horvath 2005) of its hypergraph edges. By comparing the empirical overlaps between assemblages within contexts, indicated by the overall context associations of proveniences and the natural partitioning of the assemblage, one should be able to discern the patterns of relative memberships from stratigraphic dislocations.

Approaching assemblage analysis as a combinatorial problem, the discrimination of a site’s assemblages and their contexts of deposition become a matter of optimal set partitioning. As described above, this entails an optimal partitioning of the site’s total assemblage multiset $\mathcal{S}$ such that there is an equivalent partitioning of a family of sets $\mathbb{V}$ over $\mathcal{S}$ of proveniences. In both cases, the optimization criteria is that there exist equivalent (or roughly equivalent) subsets within both the total site assemblage and proveniences, and that these subsets be disjoint or (minimally) almost disjoint from any other such subsets. Combinatorial optimization of this form is a well-studied problem in discrete mathematics

known as *set packing*, or more generally an *exact cover* problem, for which there is a similar graph problem known as the *clique* problem. The objective of the set packing problem is to find how many (if any) optimally independent substructures exist within a set or graph. There are a number of variants on this family of combinatorial problems, given various constraints and particular domains of study, and numerous computational approaches have been explored for this common application.

Similarly, the objective of clique problems is to find an optimal subset of vertices (or *subgraphs*) that meet some particular criteria. For the assemblage analysis, the problem may be viewed as a variant on maximal cliques – i.e., finding the largest subgraph within an arbitrary graph (in this case, a hypergraph) that cannot be included in any larger subgraph. As with set packing, clique problems are well studied in both computer science and mathematics. Although the general forms of these problems are NP-hard or NP-complete⁹ (i.e., have no ideal closed-form solution that is feasibly computable, see Karp 1972), constrained forms with finite parameters are either computationally solvable or have computable estimations (e.g., Erdős, Rothschild, & Singhi 1977; Segundo, Tapia, & Lopez 2013; Tang et al. 2013; Yuster 2006). This class of problems occurs frequently in real-world applications such as operations research and computer science, so has received considerable attention in other disciplines (see Arkin & Hassin 1998; Caprara et al. 2000; Colbourn, Kocay, & Stinson 1986; Gargano & Hammar 2009; Halldórsson & Losievskaja 2009; Kapovich & Nagnibeda 2013; Murawski & Bossaerts 2016; Ordman 1995). With the expansion of data science as its own field of study, particularly in regard to large and complex data with unknown classifications, significant advances have been made in algorithmic solutions to such problems over the last couple of decades.

Discussion

The formalization presented above is only the initial step toward the problem of evaluating the structuring relationships within archaeological deposits, and an algorithmic implementation of the framework is currently under development. This is still a preliminary effort with a number of issues remaining to be resolved. In particular, further work is needed towards establishing the relationship between the network of associations derived from the set- and graph-based ontology and the network of stratigraphic and spatial associations. I believe, however, that the formal ontology presented here provides a more empirically grounded and semantically sensitive approach to those networks of archaeological relationships than is available by strictly statistical or spatial analyses. By readdressing such a fundamental concept as the underlying empirical basis of intra-site associations, my hope is to reinvigorate and broaden the discussion regarding the nature of data for quantitative analyses.

Correctly defining a problem and specifying its components are always necessary steps towards finding a solution. If the goal is a quantitative or computational solution, however, these steps are absolutely critical in the process.

Formalizing the problem and specifying its evaluative criteria, typically in a mathematical or formal logical structure, delineates both the possibilities for an algorithmic solution and the form and content of its output (Barceló & Bogdanovic 2015; Bunge 1963; Read 1974, 1989; Turchin 1993). Methodologically, the manner in which these entities and criteria are defined imposes certain categorical structures onto otherwise unstructured observational data – a potential source of significant selection biases (Guarino 1995; Sowa 2000; Turchin 1993).

Intuitively, archaeologists recognize and correct for such biases in their considerations of sites and assemblages and in their interpretations. Our training and insight allow us to identify patterns and their mismatches (e.g., artefacts out of context, disturbed proveniences, etc.). Quantitative and computational solutions, conversely, leave no room for the luxury of intuition. Under-specification and categorical errors in these definitions result in incoherent and erroneous algorithmic outputs. The biggest danger in quantitative and computational approaches is that they will nearly always output *an* answer – regardless of whether or not such output is reasonably the *correct* answer.

The isolation of discrete assemblages and contexts within a site forms the basis of nearly all archaeological interpretation, and the most prevalent form of empirical data are artefacts and the proveniences from which they are recovered. These field data are central to archaeological methodology. The complex nature of those field data, however, present significant problems for quantitative inference. The standard procedures of normalization, feature selection, and dimensionality reduction are certainly able to coerce the sparse and high-variance distributions, typical of archaeological assemblages, into suitable conditions to allow the application of any number of statistical techniques. In doing so, however, the complex interrelationships of assemblages and contexts are necessarily simplified with some corresponding degree of information loss. I believe these problems are only further exacerbated by reliance on compositional profiles and artefact type-frequency approaches.

I have presented a different approach that I believe to be more appropriate to the nature of archaeological field data. The combinatorial approach captures more of the internal complexities of assemblages and contexts by viewing them holistically rather than as discrete independent events or variables. A clear structural ontology of site, assemblage, and context suggests that the necessary empirical units of analysis devolve to the most basic elements – proveniences and artefacts. By addressing those elements in terms of combinatorial sets rather than compositional frequencies, the conceptual mappings between the standard archaeological units and various aspects of set theory render an obvious and relatively simple formalization. That formalization, in turn, leads to well-defined and studied classes of combinatorial problems in discrete mathematics and graph theory, but remains consistent with an archaeologically grounded ontology. The specific archaeological implementation of these concepts will, however, require further research.

Notes

- ¹ A summary of the methods was presented at CAA2017 in Atlanta, GA as “Unscrambling the Egg – Quantitative, assemblage-based component con-sociation methods for densely mixed or disturbed contexts”.
- ² Note that I am not referring here to the concepts of assemblage theory in the sense of Deleuze & Guattari (1987) or DeLanda (2006), or the current theoretical debates regarding a “symmetrical” archaeology (e.g., Fowler 2017; Hamilakis & Jones 2017; Hicks 2012; Normark 2010; Witmore 2014), but rather the common archaeological connotation of the term.
- ³ Typically, there are *a priori* associations that may be inferred by an object’s typological, temporal, or functional classifications. These are imputed rather than empirical classes, however, and are necessarily reliable indicators of association (e.g., see critique of functional and typological classifications in Loughmiller-Cardinal 2018). The objective of this article is defining an empirical linkage of association through purely observable measures.
- ⁴ *Decomposing* a graph means creating a set of sub-graphs by partitioning the edges of the graph. For hypergraphs, decomposition entails partitioning the hyperedge sets. There are numerous approaches and types of graph decomposition, but these are well beyond the scope of this paper.
- ⁵ For the time being, the other attributes of proveniences as stratigraphic units (e.g., their spatial locations, soil morphology, etc.) are set aside as a separate analysis. Proveniences as spatial objects, with particular stratigraphic attributes, are better addressed by other quantitative methods. The formalization described here provides means to impute the related assemblage and context information to each artefact and provenience, which can then provide data for subsequent analyses.
- ⁶ An indexed set, indicated by the subscript i , denotes that it is itself a member of a family of sets.
- ⁷ It is necessary in this case to specify that a given $\mathcal{V}_i$ exists (i.e., is part of the empirically observed set), or the definition for $\mathbb{V}$ would include all *possible* proveniences. In other words, every possible combination of artefacts that *could* have been found together given the total collected artefacts.
- ⁸ i.e., the sample space of the site’s universe Ω_s with respect to the assemblage and context families of artefacts and proveniences, respectively.
- ⁹ “NP” stands for non-deterministic polynomial time. NP-complete problems are those for which there is no efficient algorithmic solution. This means that the time to compute such a problem would increase exponentially such that the time to compute all possible solutions would not be computationally feasible. Verifying a *single* solution is feasible, however, so practical solutions to NP-complete decision problems focus on heuristic approximations, estimates with constrained parameters, or stochastic searches.

Acknowledgements

I extend my sincere gratitude for the support and thoughtful insights of the numerous people that have made this work possible, and to the editors and reviewers for their considered and constructive comments. My thanks to Christina Rieth for the opportunity and encouragement that began this research as well as comments through earlier drafts, and to John Hart for encouraging my independent research. My thanks also to Miquel Colobran, Oliver Nakoinz, Grégoire van Havre, and Georg Roth for their conversations and encouragement at recent CAA conferences. A special thanks to Peter Biro, James Cardinal, and Martin Pickands for commenting on earlier drafts. Most of all, I am thankful for my family's support, especially for my brilliant wife Jennifer Loughmiller-Cardinal – without whom I would not have believed my ideas worth presenting, and whose keen insights always inspire.

Competing Interests

The author has no competing interests to declare.

References

- Achino, KF and Barceló, JA.** 2018. Spatial Prediction: Reconstructing the Spatiality of Social Activities at the Intra-Site Scale. *Journal of Archaeological Method and Theory*. DOI: <https://doi.org/10.1007/s10816-018-9367-1>
- Agrawal, R, Gehrke, J, Gunopulos, D and Raghavan, P.** 1998. Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD Record*, 27(2): 94–105. DOI: <https://doi.org/10.1145/276305.276314>
- Ahlsweede, R, Alon, N, Erdős, PL, Ruszinkó, M and Székely, LA.** 1997. Intersecting systems. *Combinatorics, Probability and Computing*, 6(2): 127–137. DOI: <https://doi.org/10.1017/S0963548397003003>
- Aldenderfer, MS.** 1998. Quantitative methods in archaeology: A review of recent trends and developments. *Journal of Archaeological Research*, 6(2): 91–120. DOI: <https://doi.org/10.1007/BF02446161>
- Ammerman, AJ.** 1992. Taking Stock of Quantitative Archaeology. *Annual Review of Anthropology*, 21(1): 231–249. DOI: <https://doi.org/10.1146/annurev.an.21.100192.001311>
- Arkin, EM and Hassin, R.** 1998. On Local Search for Weighted k -Set Packing. *Mathematics of Operations Research*, 23(3): 640–648. DOI: <https://doi.org/10.1287/moor.23.3.640>
- Ascher, R.** 1961. Analogy in Archaeological Interpretation. *Southwestern Journal of Anthropology*, 17(4): 317–325. DOI: <https://doi.org/10.1086/soutjanth.17.4.3628943>
- Bahmanian, MA and Šajna, M.** 2015. Hypergraphs: connection and separation. *arXiv ePrints*, 1–31. arXiv: 1504.04274.
- Bailey, G.** 2007. Time perspectives, palimpsests and the archaeology of time. *Journal of Anthropological Archaeology*, 26(2): 198–223. DOI: <https://doi.org/10.1016/j.jaa.2006.08.002>
- Bárány, I.** 2005. Applications of Graph and Hypergraph Theory in Geometry. *Combinatorial and Computational Geometry*, 52: 31–50.
- Barceló, JA.** 1996. Heuristic classification and fuzzy sets. New tools for archaeological typologies. In: Kamermans, H and Fennema, K (eds.), *Interfacing the Past: Computer Applications and Quantitative Methods in Archaeology 1995*, 155–164. *Analecta Praehistorica Leidensia* 28. Leiden University Press.
- Barceló, JA.** 2008. Towards a True Automatic Archaeology: Integrating Technique and Theory. In: Posluschny, A, Lambers, K and Herzog, I (eds.), *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology, Berlin, Germany, April 2–6, 2007*, 413–417. Bonn: Dr. Rudolf Habelt.
- Barceló, JA and Bogdanovic, I.** (eds.) 2015. *Mathematics and Archaeology*. Boca Raton: CRC Press. DOI: <https://doi.org/10.1201/b18530>
- Barceló, JA and Maximiano, A.** 2008. Some notes regarding distributional analysis of spatial data. In: Posluschny, A, Lambers, K and Herzog, I (eds.), *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology, Berlin, Germany, April 2–6, 2007*, 1–6. June 2016. Bonn: Dr. Rudolf Habelt.
- Baxter, MJ.** 2003. *Statistics in archaeology*. London: Arnold.
- Baxter, MJ.** 2009. Archaeological data analysis and fuzzy clustering. *Archaeometry*, 51(6): 1035–1054. DOI: <https://doi.org/10.1111/j.1475-4754.2008.00449.x>
- Baxter, MJ, Beardah, CC, Cool, HEM and Jackson, CM.** 2005. Compositional analysis of archaeological glasses. *Mathematical Geology*, 37(2): 183–196. DOI: <https://doi.org/10.1007/s11004-005-1308-3>
- Baxter, MJ and Cool, HEM.** 1995. Notes on Some Statistical Aspects of Pottery Quantification. *Medieval Ceramics*, 19: 89–98.
- Baxter, MJ and Freestone, IC.** 2003. Log-ratio analysis in archaeometry Glass compositional data. *Archaeometry*, 48(3): 511–531. DOI: <https://doi.org/10.1111/j.1475-4754.2006.00270.x>
- Bellanger, L, Tomassone, R and Husi, P.** 2008. A statistical approach for dating archaeological contexts. *Journal of Data Science*, 6: 135–154.
- Berge, C.** 1973. *Graphs and hypergraphs*, 6. North-Holland mathematical library. Amsterdam: North-Holland.
- Berge, C.** 1984. *Hypergraphs: combinatorics of finite sets*, 45. North-Holland mathematical library. Amsterdam: North Holland.
- Berge, C and Ray-Chaudhuri, D.** (eds.) 2006. *Hypergraph seminar: Ohio state university, 1972*, 411. Lecture notes in mathematics. Berlin: Springer-Verlag.
- Bevan, A and Conolly, J.** 2009. Modelling spatial heterogeneity and nonstationarity in artifact-rich landscapes. *Journal of Archaeological Science*, 36(4): 956–964. DOI: <https://doi.org/10.1016/j.jas.2008.11.023>

- Bevan, A, Crema, ER, Li, X and Palmisano, A.** 2013. Intensities, interactions and uncertainties: some new approaches to archaeological distributions. In: Bevan, A and Lake, MW (eds.), *Computational Approaches to Archaeological Spaces*. Walnut Creek: Left Coast Press.
- Binford, LR.** 1962. Archaeology as Anthropology. *American Antiquity*, 28(2): 217–225. DOI: <https://doi.org/10.1007/s10816-010-9080-1>
- Binford, LR.** 1981. Behavioral Archaeology and the 'Pompeii Premise'. *Journal of Anthropological Research*, 37(3): 195–208. DOI: <https://doi.org/10.1086/jar.37.3.3629723>
- Björklund, A, Husfeldt, T and Koivisto, M.** 2009. Set Partitioning via Inclusion-Exclusion. *SIAM Journal on Computing*, 39(2): 546–563. DOI: <https://doi.org/10.1137/070683933>
- Blizard, WD.** 1988. Multiset Theory. *Notre Dame Journal of Formal Logic*, 30(1): 36–66. DOI: <https://doi.org/10.1305/ndjfl/1093634995>
- Blizard, WD.** 1991. The development of multiset theory. *Modern Logic*, 1(4): 319–352.
- Brandt, R, Groenewoudt, BJ and Kvamme, KL.** 1992. An experiment in archaeological site location: Modeling in the Netherlands using GIS techniques. *World Archaeology*, 24(2): 268–282. DOI: <https://doi.org/10.1080/00438243.1992.9980207>
- Bretto, A.** 2013. *Hypergraph Theory: An Introduction*. Heidelberg: Springer. DOI: <https://doi.org/10.1007/978-3-319-00080-0>
- Bunge, M.** 1963. A General Black Box Theory. *Philosophy of Science*, 30(4): 346–358. DOI: <https://doi.org/10.1086/287954>
- Caprara, A, Kellerer, H, Pfersch, U and Pisinger, D.** 2000. Approximation algorithms for knapsack problems with cardinality constraints. *European Journal of Operational Research*, 123(2): 333–345. DOI: [https://doi.org/10.1016/S0377-2217\(99\)00261-1](https://doi.org/10.1016/S0377-2217(99)00261-1)
- Cardinal, JS.** 2015. Analysis of Data Recovery Assemblages. In: Davis, NL (ed.), *Fort Edward Village Site, Fort Edward Feeder Canal Bridge Site, and Hilfinger Pottery Site*, 38–86. Cultural Resource Survey Program 8. The New York State Education Department. Chap. 3.
- Carr, C.** 1984. The Nature of Organization of Intra-site Archaeological Records and Spatial Analytic Approaches to Their Investigation. In: *Advances in Archaeological Method and Theory*, 7: 103–222. Elsevier. DOI: <https://doi.org/10.1016/B978-0-12-003107-8.50008-6>
- Clarke, DL.** 1968. *Analytical archaeology*. London: Methuen & Co Ltd.
- Colbourn, CJ, Kocay, WL and Stinson, DR.** 1986. Some NP-complete problems for hypergraph degree sequences. *Discrete Applied Mathematics*, 14(3): 239–254. DOI: [https://doi.org/10.1016/0166-218X\(86\)90028-4](https://doi.org/10.1016/0166-218X(86)90028-4)
- Cowgill, GL.** 1968. Archaeological Applications of Factor, Cluster, and Proximity Analysis. *American Antiquity*, 33(3): 367. DOI: <https://doi.org/10.2307/278705>
- Crema, ER, Bevan, A and Lake, MW.** 2010. A probabilistic framework for assessing spatio-temporal point patterns in the archaeological record. *Journal of Archaeological Science*, 37(5): 1118–1130. DOI: <https://doi.org/10.1016/j.jas.2009.12.012>
- Davis, NL.** (ed.) 2015. *Fort Edward Village Site, Fort Edward Feeder Canal Bridge Site, and Hilfinger Pottery Site*. Cultural Resource Survey Program 8. The New York State Education Department.
- DeLanda, M.** 2006. *A new philosophy of society. Assemblage theory and social complexity*. London: Continuum.
- Deleuze, G and Guattari, F.** 1987. *A thousand plateaus. Capitalism and schizophrenia*. Trans. by Massumi, B, 2. Minneapolis: University of Minnesota Press.
- Dewar, M, Healy, J, Pérez-Giménez, X, Prałat, P, Proos, J, Reiniger, B and Ternovsky, K.** 2016. Subgraphs in Non-uniform Random Hypergraphs. In: Bonato, A, Graham, FC and Prałat, P (eds.), *Algorithms and Models for the Web Graph*, 140–151. Lecture Notes in Computer Science 10088. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-49787-7_12
- Dewar, RE.** 1986. Discovering settlement systems of the past in New England site distributions. *Man in the Northeast*, 31: 77–88.
- Djindjian, F.** 2015. A short history of the beginnings of mathematics in archaeology. In: Barceló, JA and Bogdanovic, I (eds.), *Mathematics and Archaeology*, 65–85. CRC Press. DOI: <https://doi.org/10.1201/b18530>
- Dubois, D and Prade, H.** 2015. The legacy of 50 years of fuzzy sets: A discussion. *Fuzzy Sets and Systems*, 281: 21–31. DOI: <https://doi.org/10.1016/j.fss.2015.09.004>
- Dunnell, RC and Dancey, WS.** 1983. The Siteless Survey: A Regional Scale Data Collection Strategy. *Advances in Archaeological Method and Theory*, 6(1983): 267–287. DOI: <https://doi.org/10.1016/B978-0-12-003106-1.50012-2>
- Erdős, P, Rothschild, BL and Singhi, NM.** 1977. Characterizing cliques in hypergraphs. *Ars Combinatoria*, 4: 81–118. January 1977.
- Fieller, NRJ.** 1993. Archaeostatistics: Old Statistics in Ancient Contexts. *The Statistician*, 42(3): 279. DOI: <https://doi.org/10.2307/2348807>
- Fowler, C.** 2017. Relational Typologies, Assemblage Theory and Early Bronze Age Burials. *Cambridge Archaeological Journal*, 27(1): 95–109. DOI: <https://doi.org/10.1017/S0959774316000615>
- Gargano, L and Hammar, M.** 2009. A note on submodular set cover on matroids. *Discrete Mathematics*, 309(18): 5739–5744. DOI: <https://doi.org/10.1016/j.disc.2008.05.019>
- Gasche, H and Tunca, Ö.** 1983. Guide to Archaeostratigraphic Classification and Terminology: Definitions and Principles. *Journal of Field Archaeology*, 10(3): 325. DOI: <https://doi.org/10.2307/529547>
- Guarino, N.** 1995. Formal ontology, conceptual analysis and knowledge representation. *International journal*

- of human-computer studies, 43(5–6): 625–640. DOI: <https://doi.org/10.1006/ijhc.1995.1066>
- Halldórsson, MM** and **Losievskaja, E.** 2009. Independent sets in bounded-degree hypergraphs. *Discrete Applied Mathematics*, 157(8): 1773–1786. DOI: <https://doi.org/10.1016/j.dam.2008.11.013>
- Hamilakis, Y** and **Jones, AM.** 2017. Archaeology and Assemblage. *Cambridge Archaeological Journal*, 27(1): 77–84. DOI: <https://doi.org/10.1017/S0959774316000688>
- Harris, EC.** 1989. *Principles of Archaeological Stratigraphy*, 2nd ed. San Diego: Academic Press.
- Hermon, S** and **Niccolucci, F.** 2010. A fuzzy logic approach to typology in archaeological research. Proceedings of CAA2004, Prato 13–17 April 2004. In: Niccolucci, F and Hermon, S (eds.), *Beyond the Artifact. Digital Interpretation of the Past*, 28–35. Oxford: Archaeopress.
- Hermon, S, Niccolucci, F, Alhaique, F, Iovino, M-R** and **Leonini, V.** 2004. Archaeological Typologies – an Archaeological Fuzzy Reality. In: Fischer-Ausserer, K, Börner, W, Gorany, M and Karlhuber-Vöckl, L (eds.), *Enter the Past. The E-way into the four Dimensions of Cultural Heritage*, 30–34. CAA 2003. Computer Applications and Quantitative Methods in Archaeology. BAR International Series 1227.
- Herzog, I.** 1995. Combining stratigraphic information and finds. In: Wilcock, J and Lockyear, K (eds.), *CAA 93 (1993)*, 109–114. BAR International Series 598. Computer Applications and Quantitative Methods in Archaeology. Oxford: Tempus Reparatum.
- Herzog, I.** 2001. Possibilities for Analysing Stratigraphic Data. In: *Archäologie und Computer*, 11. Workshop 6. Vienna: Phoibos Verlag.
- Herzog, I** and **Hansohm, J.** 2008. Monotone Regression—A Method for Combining Dates and Stratigraphy. *Tech. rep. Paper presented at the Workshop 12 “Archäologie und Computer. Kulturelles Erbe und neue Technologien” held in Vienna 2007*, 1–11. Bonn, Germany: The Rhineland Commission for Archaeological Monuments and Sites.
- Hickman, JL.** 1980. A note on the concept of multiset. *Bulletin of the Australian Mathematical society*, 22(2): 211–217. DOI: <https://doi.org/10.1017/S000497270000650X>
- Hicks, D.** 2010. The Material-Cultural Turn: Event and effect. In: Hicks, D and Beaudry, MC (eds.), *The Oxford Handbook of Material Culture Studies*, 25–99. Oxford: Oxford University Press. Chap. 2.
- Hicks, D.** 2012. *The Material-Cultural Turn*. In: Hicks, D and Beaudry, MC (eds.), 1: 25–98. Oxford University Press. DOI: <https://doi.org/10.1093/oxfordhb/9780199218714.013.0002>
- Hietala, HJ** and **Stevens, DE.** 1977. Spatial analysis: multiple procedures in pattern recognition studies. *American Antiquity*, 42(4): 539–559. DOI: <https://doi.org/10.2307/278927>
- Hodder, I** and **Orton, C.** 1976. *Spatial analysis in archaeology*. New studies in archaeology 1. Cambridge: Cambridge University Press.
- Hodson, FR.** 1969. Searching for structure within multivariate archaeological data. *World Archaeology*, 1(1): 90–105. DOI: <https://doi.org/10.1080/00438243.1969.9979430>
- Jarosław, J** and **Hildebrandt-Radke, I.** 2009. Using multivariate statistics and fuzzy logic system to analyse settlement preferences in lowland areas of the temperate zone: an example from the Polish Lowlands. *Journal of Archaeological Science*, 36(10): 2096–2107. DOI: <https://doi.org/10.1016/j.jas.2009.06.004>
- Kapovich, I** and **Nagnibeda, T.** 2013. Subset currents on free groups. *Geometriae Dedicata*, 166(1): 307–348. DOI: <https://doi.org/10.1007/s10711-012-9797-y>
- Karp, RM.** 1972. Reducibility among Combinatorial Problems. In: Miller, RE, Thatcher, JW and Bohlinger, JD (eds.), *Complexity of Computer Computations*, 85–103. Boston, MA: Springer. DOI: https://doi.org/10.1007/978-1-4684-2001-2_9
- Keevash, P** and **Mycroft, R.** 2015. A geometric theory for hypergraph matching. *Memoirs of the American Mathematical Society*, 233(1098). DOI: <https://doi.org/10.1090/memo/1098>
- King, JL** and **Moll, RG.** 1972. Set Theory Models: an Approach to taxonomic and locational relationships. In: Clarke, DL (ed.), *Models in archaeology*, 735–55. Methuen and Co. Ltd. London.
- Kintigh, KW.** 1990. Intrasite spatial analysis: A commentary on major methods. In: Voorrips, A (ed.), *Mathematics and Information Science in Archaeology: A Flexible Framework*, 3: 165–200. Bonn: Holos.
- Klein, F, Norambuena, N, Rosselot, S, Vaccaro, H, Cádiz, H, Martínez, E** and **Gana, J.** 1983. Using fuzzy logic to manage uncertain multi-modal data in an archaeological GIS. *Revista chilena de obstetricia y ginecología*, 48(4): 255–262.
- Koetje, TA.** 1991. Simulated Archaeological Levels and the Analysis of Le Flageolet II, the Dordogne, France. *Journal of Field Archaeology*, 18(2): 187–198. DOI: <https://doi.org/10.1179/009346991792208399>
- Kvamme, KL.** 1990. The fundamental principles and practice of predictive archaeological modeling. In: Voorrips, A (ed.), *Mathematics and information science in archaeology: A flexible framework*, 257. Bonn: Holos.
- Kvamme, KL.** 1999. Recent Directions and Developments in Geographical Information Systems. *Journal of Archaeological Research*, 7(2): 153–201. DOI: <https://doi.org/10.1007/s10814-005-0002-9>
- Lake, J.** 1976. Sets, fuzzy sets, multisets and functions. *Journal of the London Mathematical Society*, s2–12(3): 323–326. DOI: <https://doi.org/10.1112/jlms/s2-12.3.323>
- Langfelder, P** and **Horvath, S.** 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1): 559. DOI: <https://doi.org/10.1186/1471-2105-9-559>
- Li, A** and **Horvath, S.** 2007. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics*, 23(2): 222–231. DOI: <https://doi.org/10.1093/bioinformatics/btl581>

- Lockyear, K.** 2013. Applying bootstrapped Correspondence Analysis to archaeological data. *Journal of Archaeological Science*, 40(12): 4744–4753. DOI: <https://doi.org/10.1016/j.jas.2012.08.035>
- Loughmiller-Cardinal, J.** 2018. Distinguishing the Uses, Functions, and Purposes of Classic Maya ‘Chocolate’ Containers: Not All Cups Are for Drinking. *Ancient Mesoamerica*, 1–18. DOI: <https://doi.org/10.1017/S0956536117000359>
- Lyman, RL.** 2012. A Historical Sketch on the Concepts of Archaeological Association, Context, and Provenience. *Journal of Archaeological Method and Theory*, 19(2): 207–240. DOI: <https://doi.org/10.1007/s10816-011-9107-2>
- Lyman, RL and O’Brien, MJ.** 2001. The direct historical approach, analogical reasoning, and theory in Americanist archaeology. *Journal of Archaeological Method and Theory*, 8(4): 303–342. DOI: <https://doi.org/10.1023/A:1013736416067>
- Madgwick, R and Mulville, J.** 2015. Reconstructing depositional histories through bone taphonomy: Extending the potential of faunal data. *Journal of Archaeological Science*, 53: 255–263. DOI: <https://doi.org/10.1016/j.jas.2014.10.015>
- McCoy, MD and Ladefoged, TN.** 2009. New Developments in the Use of Spatial Technology in Archaeology. *Journal of Archaeological Research*, 17(3): 263–295. DOI: <https://doi.org/10.1007/s10814-009-9030-1>
- Merrill, M.** 2015. Lattice Theory to Discover Spatially Cohesive Sets of Artifacts. In: Barceló, JA and Bogdanovic, I (eds.), *Mathematics and Archaeology*, 383–401. June 2015. CRC Press. DOI: <https://doi.org/10.1201/b18530-24>
- Merrill, M and Read, DW.** 2010. A New Method Using Graph and Lattice Theory to Discover Spatially Cohesive Sets of Artifacts and Areas of Organized Activity in Archaeological Sites. *American Antiquity*, 75(3): 419–451. DOI: <https://doi.org/10.7183/0002-7316.75.3.419>
- Miyamoto, S.** 2001. Fuzzy multisets and their generalizations. *Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View*, Calude, CS and Rozenberg, G (eds.), 2235: 225–235. May. DOI: https://doi.org/10.1007/3-540-45523-X_11
- Miyamoto, S.** 2004. Multisets and Fuzzy Multisets as a Framework of Information Systems. In: Torra, V and Narukawa, Y (eds.), *Modeling Decisions for Artificial Intelligence, Proceedings*, 27–40. Berlin: Springer. DOI: https://doi.org/10.1007/978-3-540-27774-3_4
- Murawski, C and Bossaerts, P.** 2016. How Humans Solve Complex Problems: The Case of the Knapsack Problem. *Scientific Reports*, 6: 1–10. September. DOI: <https://doi.org/10.1038/srep34851>
- Niccolucci, F and Hermon, S.** 2002. Estimating subjectivity of typologists and typological classification with fuzzy logic logic. *Archeologia e Calcolatori*, 13: 217–232.
- Normark, J.** 2010. Involutions of Materiality. *Journal of Archaeological Method and Theory*, 17(2): 132–173. DOI: <https://doi.org/10.1007/s10816-010-9079-7>
- Ordman, ET.** 1995. Cliques in hypergraphs and mutual exclusion using tokens. *Journal of Combinatorial Mathematics and Combinatorial Computing*, 19: 209–224.
- Östborn, P and Gerding, H.** 2014. Network Analysis of Archaeological Data: a Systematic Approach. *Journal of Archaeological Science*, 46(1): 75–88. DOI: <https://doi.org/10.1016/j.jas.2014.03.015>
- Paquette, J and Tokuyasu, T.** 2011. Hypergraph visualization and enrichment statistics: how the EGAN paradigm facilitates organic discovery from big data. In: Rogowitz, BE and Pappas, TN (eds.), *Human Vision and Electronic Imaging XVI*, 7865: 78650E. DOI: <https://doi.org/10.1117/12.890220>
- Peeples, MA and Schachner, G.** 2012. Refining correspondence analysis-based ceramic seriation of regional data sets. *Journal of Archaeological Science*, 39(8): 2818–2827. DOI: <https://doi.org/10.1016/j.jas.2012.04.040>
- Petrie, WME.** 1899. Sequences in Prehistoric Remains. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 29(3/4): 295. DOI: <https://doi.org/10.2307/2843012>
- Puyol-Gruart, J.** 1999. Computer science, artificial intelligence and Archaeology. In: Barceló, JA, Briz, I and Vila, A (eds.), *New Techniques for Old Times*, 1–15. CAA98. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 26th Conference, Barcelona, March 1998 (BAR International Series 757). Archaeopress.
- Read, DW.** 1974. Some Comments on the Use of Mathematical Models in Anthropology. *American Antiquity*, 39(1): 3. DOI: <https://doi.org/10.2307/279215>
- Read, DW.** 1989. Statistical methods and reasoning in archaeological research: A review of praxis and promise. *Journal of Quantitative Anthropology*, 1: 5–78.
- Richards, JD.** 1998. Recent trends in computer applications in archaeology. *Journal of Archaeological Research*, 6(4): 331–382. DOI: <https://doi.org/10.1007/BF02446083>
- Robinson, WS.** 1951. A Method for Chronologically Ordering Archaeological Deposits. *American Antiquity*, 16(4): 293–301. DOI: <https://doi.org/10.2307/276978>
- Schiffer, MB.** 1972. Archaeological Context and Systemic Context. *American Antiquity*, 37(2): 156–165. DOI: <https://doi.org/10.2307/278203>
- Schiffer, MB.** 1983. Toward the Identification of Formation Processes. *American Antiquity*, 48(4): 675–706. DOI: <https://doi.org/10.2307/279771>
- Schiffer, MB.** 1985. Is There a ‘Pompeii Premise’ in Archaeology? *Journal of Anthropological Research*, 41(1): 18–41. DOI: <https://doi.org/10.1086/jar.41.1.3630269>
- Schiffer, MB.** 1987. *Formation processes of the archaeological record*, 1st ed. Albuquerque, NM: University of New Mexico Press.
- Segundo, PS, Tapia, C and Lopez, A.** 2013. Watching subgraphs to improve efficiency in maximum clique

- search. In: Ali, M, Bosse, T, Hindriks, K, Hoogendoorn, M, Jonker, C and Treur, J (eds.), *Contemporary Challenges and Solutions in Applied Artificial Intelligence*, 489: 115–122. Springer. DOI: https://doi.org/10.1007/978-3-319-00651-2_16
- Shott, MJ.** 1998. Status and Role of Formation Theory in Contemporary Archaeological Practice. *Journal of Archaeological Research*, 6(4): 299–329. DOI: <https://doi.org/10.1007/BF02446082>
- Singh, D, Alkali, A and Ibrahim, AM.** 2013. An Outline of the Development of the Concept of Fuzzy Multisets. *International Journal of Innovation, Management and Technology*, 4(2): 2–5. DOI: <https://doi.org/10.7763/IJIMT.2013.V4.393>
- Singh, D, Ibrahim, AM, Yohanna, T and Singh, JN.** 2007. An Overview of the Applications of Multisets. *Novi Sad Journal of Mathematics*, 37(3): 73–92.
- Song, L, Langfelder, P and Horvath, S.** 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1): 328. DOI: <https://doi.org/10.1186/1471-2105-13-328>
- Sowa, JF.** 2000. *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole.
- Spikins, P, Conneller, C, Ayestaran, H and Scaife, B.** 2002. GIS Based Interpolation Applied to Distinguishing Occupation Phases of Early Prehistoric Sites. *Journal of Archaeological Science*, 29(11): 1235–1245. DOI: <https://doi.org/10.1006/jasc.2001.0752>
- Sullivan, AP.** 1978. Inference and evidence in archaeology: a discussion of the conceptual problems. *Advances in archaeological method and theory*, 1(1978): 183–222. DOI: <https://doi.org/10.1016/B978-0-12-003101-6.50012-3>
- Syropoulos, A.** 2001. Mathematics of Multisets. In: Calude, CS, Paun, G, Rozenberg, G and Salomaa, A (eds.), *Multiset Processing SE – 17*, 2235: 347–358. August. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/3-540-45523-X_17
- Syropoulos, A.** 2012. On Generalized Fuzzy Multisets and their Use in Computation. *Iranian Journal of Fuzzy Systems*, 9(2): 113–125. arXiv: 1208.2457v1.
- Tang, Q, Peng, Y, Zhang, X and Zhao, C.** 2013. On hypergraph cliques and polynomial programming. *ArXiv ePrints*. arXiv: 1312.6973.
- Thomas, J.** 2015. The future of archaeological theory. *Antiquity*, 89(348): 1287–1296. December. DOI: <https://doi.org/10.15184/aqy.2015.183>
- Turchin, V.** 1993. On cybernetic epistemology. *Systems Research and Behavioral Science*, 10(1): 1–28. DOI: <https://doi.org/10.1002/sres.3850100102>
- Tzouvaras, A.** 1998. The linear logic of multisets. *Logic Journal of IGPL*, 6(6): 901–916. DOI: <https://doi.org/10.1093/jigpal/6.6.901>
- User:Kilom691.** 2010. *A hypergraph with 7 vertices and 4 edges*. Hypergraph.svg: Kilom691, derivative work: Pgdx [CC BY-SA 3.0 or GFDL]. Wikimedia Commons. URL: <https://commons.wikimedia.org/wiki/File:Hypergraph-wikipedia.svg> (visited on 05/07/2018).
- Van Leusen, M and Nobles, G.** 2018. 3D Spatial Analysis: the Road Ahead. In: Matsumoto, M and Uleberg, E (eds.), *Oceans of Data. Proceedings of the 44th Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, 471–478. Oxford: Archaeopress Publishing, Ltd.
- Whallon, R.** 1973. Spatial Analysis of Occupation Floors I: Application of Dimensional Analysis of Variance. *American Antiquity*, 38(3): 266–278. DOI: <https://doi.org/10.2307/279714>
- Whallon, R.** 1974. Spatial Analysis of Occupation Floors II: The Application of Nearest Neighbor Analysis. *American Antiquity*, 39(1): 16–34. DOI: <https://doi.org/10.2307/279216>
- Wiley, GR and Phillips, P.** 1958. *Method and Theory in American Archaeology*. Chicago: University of Chicago Press.
- Witmore, C.** 2014. Archaeology and the New Materialisms. *Journal of Contemporary Archaeology*, 1(2): 203–246. DOI: <https://doi.org/10.1558/jca.v1i2.16661>
- Wylie, A.** 1982. An analogy by any other name is just as analogical a commentary on the Gould-Watson dialogue. *Journal of Anthropological Archaeology*, 1(4): 382–401. DOI: [https://doi.org/10.1016/0278-4165\(82\)90003-4](https://doi.org/10.1016/0278-4165(82)90003-4)
- Wylie, A.** 1985. The Reaction against Analogy. In: *Advances in Archaeological Method and Theory*, 63–111. Elsevier. DOI: <https://doi.org/10.1016/B978-0-12-003108-5.50008-7>
- Wylie, A and Watson, RA.** 1992. On scepticism, philosophy, and archaeological science. *Current Anthropology*, 33(2): 209–214. DOI: <https://doi.org/10.1086/204052>
- Yager, RR.** 1986. On the theory of bags. *International Journal of General System*, 13(1): 23–37. DOI: <https://doi.org/10.1080/03081078608934952>
- Yip, AM and Horvath, S.** 2007. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8: 22. DOI: <https://doi.org/10.1186/1471-2105-8-22>
- Yuster, R.** 2006. Finding and counting cliques and independent sets in r-uniform hypergraphs. *Information Processing Letters*, 99(4): 130–134. DOI: <https://doi.org/10.1016/j.ipl.2006.04.005>
- Zadeh, LA.** 1965. Fuzzy sets. *Information and Control*, 8(3): 338–353. DOI: [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)
- Zadeh, LA.** 1996. Fuzzy sets and information granularity. In: Klir, GJ and Yuan, B (eds.), *Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A. Zadeh*, 6: 433–448. Advances in Fuzzy Systems – Applications and Theory. World Scientific. DOI: <https://doi.org/10.1142/2895>
- Zadeh, LA.** 1997. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*,

90(2): 111–127. DOI: [https://doi.org/10.1016/S0165-0114\(97\)00077-8](https://doi.org/10.1016/S0165-0114(97)00077-8)

Zhang, B and **Horvath, S.** 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1). DOI: <https://doi.org/10.2202/1544-6115.1128>

Zhou, D, Huang, J and **Schölkopf, B.** 2007. Learning with Hypergraphs: Clustering, Classification, and Embedding. In: Schölkopf, B, Platt, J and Hofmann, T (eds.), *Advances in Neural Information Processing Systems. 20th Annual Conference on Neural Information Processing Systems (NIPS)*, 19: 1601–1608. Max-Planck-Gesellschaft. Cambridge: MIT Press.

How to cite this article: Cardinal, JS. 2019. Sets, Graphs, and Things We Can See: A Formal Combinatorial Ontology for Empirical Intra-Site Analysis. *Journal of Computer Applications in Archaeology*, 2(1), pp. 56–78. DOI: <https://doi.org/10.5334/jcaa.16>

Submitted: 22 June 2018

Accepted: 22 February 2019

Published: 16 April 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

]u[

Journal of Computer Applications in Archaeology, is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS