ABSTRACT

The major uncertainties in studies of the multi-scale structure of the universe arise not from observational errors but from the variety of legitimate definitions and detection methods for individual structures. To facilitate the study of these methodological dependencies, we have carried out 12 different analyses defining structures in various ways. This has been done in a purely geometrical way by utilizing the HOP algorithm as a unique parameter-free method of assigning groups of galaxies to local density maxima or minima. From three density estimation techniques (smoothing kernels, Bayesian blocks, and self-organizing maps) applied to three data sets (the Sloan Digital Sky Survey Data Release 7, the Millennium simulation, and randomly distributed points) we tabulate information that can be used to construct catalogs of structures connected to local density maxima and minima. We also introduce a void finder that utilizes a method to assemble Delaunay tetrahedra into connected structures and characterizes regions empty of galaxies in the source catalog.

Key words: catalogs – cosmology: observations – galaxies: clusters: general – large-scale structure of universe

Supporting material: machine-readable tables

1. INTRODUCTION

In the past two decades an assortment of disparate density estimation techniques has been applied to a wide variety of data sets to characterize the distribution of galaxies in the local universe. From the very beginning the purely geometrical studies were supplemented by studies of cluster luminosity functions (e.g., Holmberg 1969, and references therein to earlier work by Zwicky). While these studies have been productive, systematic intercomparison of the results continues to be problematic. The purpose here is to address this problem by presenting data for the construction of catalogs drawn from three data sources: the Sloan Digital Sky Survey (SDSS; York et al. 2000), the Millennium simulation (hereafter MS; Springel et al. 2005), and a set of randomly distributed points. Each of these data sets was analyzed in three different ways. These are the same data and analysis techniques described in the first paper in this series (Way et al. 2011, hereafter Paper I). Note that when we refer to the MS in the text we are explicitly referring to galaxies with properties derived by Croton et al. (2005) rather than the dark matter particles (see Appendix A.9 for details).

The first of these data sets allows elucidation of the structure of the actual galaxy distribution. Technology implementing fully digital, charge-coupled device (CCD) photometric and spectroscopic observations of large areas of the sky has yielded a cornucopia of surveys of the local universe in the past 15 yr: e.g., the Las Campanas Redshift Survey (LCRS; Shectman et al. 1996; although they did not actually use CCDs for their spectroscopy), the Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006), the Two Degree Field Redshift Survey (2dFGRS; Colless et al. 2001), and SDSS. A variety of density estimation techniques have been proposed and used to elicit structural information from these data compendia. Many of the methods and the catalogs they have yielded were recently discussed in Paper I. However, observational surveys continue to grow larger and more elaborate, with cumulative releases coming every 6 months to 1 yr. For example, the SDSS is at Data Release 10 as of 2013 August (Ahn et al. 2014, DR10). The DR10 is part of the SDSS 3 survey scheduled to collect data through 2014. The complexities and sizes of both surveys and analysis methods make evaluation and interpretation of results, as well as the corresponding intercomparisons, ever more difficult. Even the restricted arena of density representations is replete with different estimation techniques (of which we discuss three) and approaches to subsequent characterization of the density field (see Section 3).

To address this circumstance, we have performed spatial structure analysis of three directly comparable point data sets (measured, simulated, and random galaxy positions) using three density estimation techniques (adaptive kernels, self-organized maps, and Bayesian blocks). We hope that these analyses will be of use to researchers in making comparisons among their own methods and those described here, on a variety of redshift surveys. All elements of the corresponding nine-fold matrix (3 data sets × 3 analysis methods) were described in Paper I. Detailed characteristics of the data are described in Appendix A (Section 4).

This paper describes our procedure for converting density estimates into localized features in the spatial distribution of galaxies. As described in Paper I, this result is achieved by assembling building blocks (tessellation cells or blocks of them) into larger structures. A key point is that both the localized details and global features that result are dependent on the principles under which this assembly is carried out. Hence, one of the goals is to understand the nature of this dependence in order to elucidate the astrophysical meaning of conclusions about the origin, evolution, and current nature of the cosmic

1 Also at NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025, USA.
2 Also at Observational Astronomy, Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden.
3 http://www.sdss3.org
web. In the literature these structures are typically assigned to four classes: clusters, sheets, filaments, and voids.

Our analysis is based on the point of view that the cosmic web (e.g., van de Weygaert et al. 2009) is composed of structures of widely distributed shapes not necessarily assignable to these four classes in a straightforward one-to-one manner. Our goal is to assign galaxies in a unique and parameter-free way to coherent structures, leaving for future exploration the assembly of the resulting building blocks into larger structures including but not limited to the ones in the four conventional classes mentioned above.

2. PREVIOUS WORK

A number of recent publications have described methods for identifying and characterizing structure in redshift surveys and simulations. For some developments since the summary in Paper I see, e.g., Aragon-Calvo et al. (2010a); Cautun & van de Weygaert (2011); Sousbie et al. (2011); Sousbie (2011, 2013); Falck et al. (2013); Knebe et al. (2013); Tempel et al. (2014), and with respect to tessellation methods Schaap (2007); Pandey et al. (2013); Angulo et al. (2014). For a comparative study of density estimation schemes see Platen et al. (2011), and for an example of machine learning approaches see D’Abrusco et al. (2012).

Because voids were not discussed in Paper I, a brief review of the literature on this topic is in order. The concept of underdensities in the distribution of galaxies and the related term “void” has been around at least since the late 1970s. Not unexpectedly, this early work was characterized by vague definitions and uncertainties due to small sample sizes. Some of this confusion continues to today.

Chincarini & Rood (1976) conducted one of the first observational studies indicating the presence of voids in distribution of galaxies (for $m \lesssim 15$) in the region of the Coma supercluster. They described the effect as a “segregation in redshifts,” but it is now known that their survey was deep enough to see actual voids.

The first explicit mention of voids or holes in the galaxy distribution can probably be shared between that of Gregory & Thompson (1978) and that of Joeveer et al. (1977, 1978). The former was published in 1978, while the latter were a preprint from 1977 and its accepted version in 1978. The Joeveer et al. (1977) preprint was also distributed in the fall of 1977 among participants at IAU Symposium No. 79 in Talinn, Estonia. For more details on this time period see Einasto (2014, p. 138) and Thompson & Gregory (2011).

By the time of the 1977 IAU Symposium in Talinn, Estonia (Longair & Einasto 1978), voids or holes were common parlance among the community. Here we present a number of examples of the relevant references. Tully & Fisher (1978); Table 2 document a “void” with a volume of $\gtrsim 1000$ Mpc$^3$. Joeveer & Einasto (1978) use the words “void” and “holes” in their 1978 IAU paper, and they estimate on page 247 that “cell interiors are almost void of galaxies; they form big holes in the Universe with diameters of 100–150 Mpc.” Tifft & Gregory (1978) say on page 267 that “there are regions more than 20 Mpc in radius which are totally devoid of galaxies,” and “the foreground is again very clumpy with one major void of radius close to 40 Mpc.”

Zeldovich (1978) recognizes the large empty spaces (holes) discussed by others at the conference, while Longair (1978) in his conference summary also mentions on page 455 “holes which are about 10 Mpc in size and void of bright galaxies.” See also Schwarzschild (1982) and a more recent overview of the void phenomenon by Peebles (2001).

In a pioneering mathematical study of probabilities that a randomly placed region of given volume will contain a given number (including zero) of galaxies, White (1979) noted that the distributions of dense structures and voids are related to each other.

According to Martinez & Saar (2002, p. 368), the first (systematic) study of voids by Einasto et al. (1989) was an attempt to establish the fractal character of the galaxy distribution. These authors developed the empty sphere method, thus pioneering methods to search directly for empty or near-empty volumes.

This approach threads the series of papers by El-Ad et al. (1996); El-Ad (1997); El-Ad & Piran (1997) discussing the observational discovery of voids with the Void Finder algorithm (see also Hoyle & Vogeley (2002) for an extension of this approach). For automatic detection of voids in redshift surveys such as the IRAS catalog see El-Ad et al. (1997). The study of El-Ad & Piran (2000) is of particular interest because of its comparison of voids discovered in two independent surveys at different wavelengths. This work, starting with Einasto et al. (1989), influenced later methodological work close in spirit to that developed here, involving variations on an explicit search for volumes actually devoid of galaxies (Kaufmann & Fairall 1991; Aikio & Mahonen 1998; Elyiv et al. 2013; Tasavolli et al. 2013), via nearest neighbor techniques (e.g., Rojas et al. 2004) or “friends-of-friends” algorithms (e.g., Muñoz-Cuartas & Müller 2012).

In other work the definition of voids is tied to local minima in the density distribution, e.g., the Watershed Void Finder (WVF) (see Platen et al. 2007, and references therein), VOBOZ (Voronoi BOund Zones) (Neyrinck et al. 2005), ZOBOV (Zones Bordering on Voidness) (Neyrinck 2008), DENMAX (Bertschinger & Gelb 1991; Gelb & Bertschinger 1994; Goetz et al. 1998), and SKID (Governato et al. 1997). Many of these works are closely related to the core idea of the Hop (Eisenstein & Hut 1998) algorithm adopted here. The two classes of void finders—based on empty volumes or local density minima—have their advantages and disadvantages. The former is naturally aligned with the smoothing-free tessellation analysis introduced in our previous work in Paper I. See also the recent works by Neyrinck et al. (2014) and Nadathur & Hotchkiss (2014).

Schmidt et al. (2001) deal with voids in simulations, comparing methods based on finding empty regions of space (within observational limits and selection effects in the survey) against those based on density estimation followed by identification of density minima. They also include two different void finder algorithms with and without predefined constraints on shape. Extensive studies of the structure and dynamics of voids (Aragon-Calvo et al. 2010b; Aragon-Calvo & Szalay 2013) argue for the existence of a hierarchical distribution of voids in the context of the cosmic web and cosmic spine concepts (see also Aragon-Calvo et al. 2007, 2010a). In common usage the term “hierarchy” usually refers to discrete and ordered levels. When it is applied to distributions of voids and structures, it seems to refer to inclusion relations—e.g., smaller structures lying within larger ones. We do not perform any tests for such relationships. However, it is worth noting that the distributions of physical parameters such as size and shape, in this work or any others we know of, are continuous and multi-scale with no evidence for the

---

4 This work was apparently inspired by White’s perception of “holes” in the galaxy distribution depicted in Gregory & Thompson (1978).
presence of discrete hierarchies of any kind (see the comments in Section 4.)

A number of recent works have investigated the statistics and stacking of voids and their importance for various environmental and other cosmological issues (e.g., Hahn et al. 2007a, 2007b; D’Aloisio & Furlanetto 2007; Gaite 2009; Paranjape et al. 2012; Lavaux & Wandelt 2010; Einasto et al. 2011; Pan et al. 2012; Einasto et al. 2012; Bos et al. 2012; Lavaux & Wandelt 2012; Bolejko et al. 2012; Zaninetti 2012; Varela et al. 2012; Sutter et al. 2012; Jennings et al. 2013; Beygu et al. 2013; Krause et al. 2013; Cecarelli et al. 2013; Hamaus et al. 2014; Ricciardelli et al. 2014; Hamaus et al. 2014a). A comparison of void catalogs and detection methods applied to the MS data is found in Colberg et al. (2008); see also Knebe et al. (2011). Powerful methods of point process theory (Daley & Vere-Jones 2003; Lowen & Teich 2005), stochastic geometry (Snyder & Miller 1991), discrete Morse theory (Sousbie 2011, 2013), computational (de Berg et al. 1997; Preparata & Shamos 1985) and combinatorial geometry (Edelsbrunner 1987), and wavelet-like transforms (Leistedt et al. 2013) are currently being used to explicate multi-scale structures in the galaxy distribution (van de Weygaert et al. 2011a, 2011b, 2011c; Sousbie et al. 2011; Sousbie et al. 2011; Park et al. 2013; Hidding et al. 2014). Especially interesting are the prospects for studying voids via gravitational lensing effects (Amendola et al. 1999; Higuchi et al. 2011; Melchior et al. 2014; Clampitt & Jain 2014). A series of papers examining voids in observations and simulations using extensions of ZOBOV and other approaches were recently published (Sutter et al. 2014; Hamaus et al. 2014b; Leclercq et al. 2014; Leclercq & Wandelt 2014). They also include new publicly available software for comparing results called VIDE5 and a website for downloading their catalogs.6

Finally, Jose Gaite recently called our attention to his paper Gaite (2005) proposing the idea of constructing voids out of adjacent Delaunay tetrahedra, but with a different construction rule than that presented herein.

3. IDENTIFICATION OF STRUCTURES

The grand challenge is to produce scientifically useful characterization of a density field derived from a galaxy survey or a computational dark matter simulation. One approach is to study statistical quantities averaged over the whole data sample. Examples include estimation of correlation functions (e.g., McBride et al. 2011; Valageas & Clerc 2012; Müller et al. 2011), power spectra (e.g., Tegmark et al. 2006; Jasche et al. 2010; Neyrinck et al. 2009), and global topological information (Shandarin et al. 2004; Gott et al. 2008; James et al. 2009; van de Weygaert et al. 2010; Sousbie et al. 2011; Einasto et al. 2014). Here we develop an alternative approach, namely, identification of specific local features of the density distribution, as outlined in Figure 1.

In Paper 1 the steps on the left side of Figure 1 here yielded density surrogates for individual galaxies or small sets of them. The current paper addresses the assembly of these building blocks into structures.

It is underappreciated that the results of any such analysis are very dependent on the methodology used. Especially strong is the dependence on the assembly procedure, but all of the choices represented in Figure 1 have their effects. There are a plethora of definitions of structural classes and approaches to detecting them, using a variety of positional, photometric, and morphological information. In particular, some detection algorithms, such as those invoking prior information about galaxy colors or cluster symmetry, naturally favor detection of structures more nearly satisfying these assumed properties. This rather murky situation raises questions. Are there intrinsically distinct well-defined

---

5 http://bitbucket.org/cosmicvoids/vide_public
6 http://www.cosmicvoids.net
structural classes? If there are things such as clusters, filaments, sheets, and voids, can we uncover their nature in objective ways not unduly influenced by methodology and prior assumptions? This classification scheme has been adopted by the community, although it is clear that these categories are somewhat fuzzy: varied structures uncovered with various schemes cannot always be classified in a straightforward way. In any case it is important to exercise care in the interpretation of structural results. Comparison of observed and simulated data using identical analyses to exercise care in the interpretation of structural results. Com-
parison of observed and simulated data using identical analyses is inherently less ambiguous, but even such comparative studies depend on the nature of the analysis.

Our work seeks to avoid some of this murkiness by using purely geometrical information derived from galaxy positions, and in ways not tuned to emphasize any particular shape. The approach here is to characterize structure over a range of scales, i.e., *multi-scale structure* (a more precise term than the commonly used *large-scale structure*). We aim to make maximal use of the information in the data, but with neither prior shape constraints nor account of geometrically extraneous systematics such as the *red sequence* in clusters (Gladders & Yee 2000; Rykoff et al. 2014; Rozo & Rykoff 2014; Rozo et al. 2014). We adopt what is arguably the simplest possible definition: a *structure is the watershed of a single critical point*—that is, of a local density maximum or minimum (see Section 3.1 for specifics). This choice rules out structures with two peaks, for example, but if desired these could be sought during a post-processing with some kind of merger criterion. Identification of voids via local density minima is supplemented with a novel void-finding procedure in Section 3.2. In a further bid toward objectivity and parameter freedom we use tessellation techniques so that the scale on which these maxima are determined is automatic, data adaptive, and not predefined.

The full definition of structures then requires a prescription for what to attach to the local maxima or minima. Here we assemble structures out of elementary building blocks as outlined in Section 4.3.3 of Paper I, avoiding arbitrary choices through the use of the parameter-free HOP algorithm of Eisenstein & Hut (1998).

### 3.1. The HOP Algorithm

The rest of this paper is devoted to the process of assembling structures out of building blocks (see the right-hand side of Figure 1). The next subsection describes a general algorithm for assembling local structures in any data representation consisting of these three elements:

1. a set of discrete entities, called *objects*;
2. the value of a function *f* for each object;
3. adjacency information among the objects.

We refer to *f* as the *HOP function*. For the computations only the last two items matter, as the algorithm makes no reference to the identity of the objects. In continuous Morse theory (Milnor 1969) the *Morse function*—the analog of our *f*—must be infinitely differentiable; in the discrete theory of Forman (2002) the corresponding function must satisfy some similarly delicate conditions. Here, in contrast, the HOP function is essentially arbitrary since it only needs to provide an ordering of the objects. Hence, the only condition on *f* is that no two objects are assigned the same value; violations of this constraint can be fixed in a trivial way.

In some of the cases reported here the objects are individual galaxies with each one being assigned an HOP function given by a density estimate: e.g., reciprocals of Voronoi cell volumes, kernel density estimates (KDE), or self-organizing map (SOM) class identifiers related to density. SOMs Kohonen (1984) are a neural network scheme that maps samples in an *N*-dimensional data space to a “classification space” of smaller dimensionality (usually 2D) in a way that preserves topology—i.e., samples with similar characteristics are mapped to adjacent regions. The resulting map can then be partitioned to perform unsupervised classification.

In another case—that of Bayesian blocks (BB; Scargle et al. 2013)—the objects are connected sets of adjacent galaxies with the appropriate density estimate (number of galaxies in the block divided by its volume). In the final example considered here, where the objects are the Delaunay tetrahedra in the tessellation of galaxy positions, the definition of *f* requires some thought, as elaborated in Section 3.2.

For the adjacencies referred to in item 3 above we use those defined by the Voronoi tessellation itself, as follows: two objects, either individual Voronoi cells or blocks thereof, are deemed adjacent if they intersect at a common face. This construct can be viewed as defining nearest neighbors in a data-adaptive fashion, with no a priori restriction on the number of neighbors. It thus conveys local information regarding the distribution of galaxies more efficiently than, say, nearest neighbors with a predefined number of neighbors. Similarly, two Delaunay tetrahedra are considered adjacent if they share a common triangular face.

Identification of watershed structures in a spatial distribution of objects is conveniently implemented using the group-finding algorithm HOP (Eisenstein & Hut 1998). Here by the term “HOP” we mean only the parameter-free group-finding step (Section 2.1 in Eisenstein & Hut 1998) distinguished from smoothing and merging procedures used elsewhere in their paper to find bound structures (halos). These procedures do involve parameters. Aubert et al. (2004) describe a variant of this post-processing, called “AdaptaHOP”; their approach differs in many ways from ours, such as regarding the sampling as noise to be smoothed over, and has procedures that generate hierarchical leaves in a tree. See also Springel (1999) for a discussion of a related algorithm called *SUBFIND*. Motivated by the requirements for analysis of massive cosmological data sets, Skory et al. (2010) deals with computational parallelization issues. Turk et al. (2011) provide a toolkit that contains an implementation of HOP.

This algorithm is quite general. For any given HOP function and adjacencies defined for each object in set *S*, it yields a partition of *S* into groups with the following properties.

1. The elements of the partition, here called “groups,” are sets of objects from *S*.
2. There is one such connected group for each local maximum of *f*.
3. In a given group *f* decreases monotonically away from the maximum.
4. Every object is in one and only one of the groups (i.e., the groups partition the space).
5. The partition is unique and parameter-free.

---

7 Objects being joined by an edge in the Delaunay tessellation are equivalent to this condition (but not to other possible definitions such as sharing Voronoi edges or vertices in lieu of faces). These adjacencies are easily established from information supplied by most data analysis systems, such as the *n*-dimensional MatLab routine *voronoin*, namely, identities of the vertices of each Voronoi cell. In finding adjacencies it is very useful to first compile a list of all galaxies whose Voronoi cells touch each vertex.
In short, HOP identifies all of the local maxima of \( f \) and the connected structures flowing from them; together these are the discrete analog of the so-called descending manifolds—mountain peaks plus their watersheds. It can identify structures of any shape—containing arbitrary mixtures of convexities and concavities, possibly even failing to be simply connected. The underlying idea of HOP is a simple hill-climbing prescription. It iteratively associates each object with neighbors that have larger values of \( f \) according to the HOP algorithm (Eisenstein & Hut 1998).

[Given.] A set \( S \) of \( N \) spatially distributed data objects \( o_i, i = 1, 2, \ldots, N \)

1. Establish an index array \( I = 1, 2, 3, \ldots, N \) for any convenient ordering of the objects.
2. Assign a value of \( f \) to each object.
3. For each object identify all objects adjacent to it, i.e., its spatial neighbors as defined earlier in this section.
4. For each set consisting of an object and all of its neighbors, find the object with the largest value of \( f \).
5. Iterate as follows:
   a. For \( i = 1, 2, \ldots, N \):
      i. Let \( j_i \) be the current index value in position \( i \) of \( I \).
      ii. In \( I \) replace \( j_i \) with that found in step 4 for object \( j_i \) (not that for object \( i \)).
   b. Repeat (a) until no index value changes.
6. Set \( K = I \) with duplicate values removed.

Eliminating the duplicate values in the converged \( I \) yields a set of indices \( K \) pointing one-to-one to each of the local density maxima—objects denser than all of their neighbors. Each of the objects ends up pointing via the converged \( I \) to one and only one of these maxima (i.e., to one of the values in \( K \)). This property generates for each local maximum a connected structure, consisting of a set of paths connecting adjacent galaxies along which the density is monotonic. These structures are equivalent to the watershed defined in image processing and many of the cosmic web algorithms referenced above.

In short, this algorithm uses a simple hill-climbing procedure\(^8\) to find a unique partition of the objects into groups, one associated with each of the local maxima. Alternatively, by jumping instead to the neighbor with the smallest value of \( f \) in step 4, HOP can find basins of attraction for all of the local minima instead. We usually refer to structures associated with local maxima as groups, rather than clusters, since they may, for example, be filamentary or sheet-like. With a similar freedom with standard terminology, drainage basins associated with local minima can be called “voids,” although we distinguish these structures from the empty collections of Delaunay tetrahedra discussed in Section 3.2.

The following MatLab code fragment uses vector operations to implement the iteration, given two arrays initialized as follows:

\[
\text{index} = [1, 2, \ldots N] \quad \text{contains the initial indices of the objects (taken in arbitrary order)}
\]

\[
\text{id}_\text{max_neighbor} \quad \text{contains indices of the neighbors found in step 4}
\]

\[\text{while 1}
\]

\[\text{index}_\text{new} = \text{id}_\text{max_neighbor}(\text{index})\]

\[\% \text{ Each object hops to highest neighbor}
\]

\[\text{id}_\text{change} = \text{find}(\text{index}_\text{new} = \text{index})\]

\[\% \text{ Locate index changes}
\]

\[\text{if isempty(\text{id}_\text{change})}
\]

\[\text{break};\text{end}
\]

\[\% \text{ If no index changes escape while loop}
\]

\[\text{index} = \text{index}_\text{new};
\]

\[\% \text{ Implement changes due to hops}
\]

\[\text{end}
\]

We take the objects in \( S \) to be individual galaxies, blocks containing several galaxies, or Delaunay tetrahedra—attached to which are values of \( f \). This function is given by the corresponding KDE or BB density estimates, given by classes from the SOM method, or derived from the sizes of the Voronoi cells or Delaunay tetrahedra.

The following points elaborate some details of the algorithm and our application of it.

1. Choices for the definition of the neighbor relation in step 3, and indexed by \( \text{id}_\text{max_neighbor} \), include densest neighbor (to find manifolds descending from local maxima) and least dense neighbor (to find manifolds ascending from local minima).
2. Throughout this discussion there is no explicit mention of the dimensionality of the data. One of the beauties of the HOP algorithm is that it applies to spaces of any dimension. Here contact with the dimension of the data arises only in the definition of adjacency, which we compute from the Voronoi tessellation of the 3D galaxy positions. But once the adjacencies are assigned, dimension is completely irrelevant.
3. The unique output of this algorithm is independent of the order of the initial indexing (step 1) or the order in which the objects are considered in step 5, modulo an inconsequential re-ordering of the output groups.
4. The iteration can be carried out in other ways than shown explicitly above, e.g., by following individual objects to their final destinations, rather than the parallel procedure in step 5a and the MatLab code fragment listed above. Such path tracking is of use for constructing analogs of topological saddle points, not discussed here.
5. If two or more objects are assigned identical values, rare except in the case of the discrete SOM class identifiers, there may be a dependence on the way the resulting ambiguity is resolved.
6. After the preprocessing represented by the initial steps 1–4, iteration 5 is guaranteed to converge rapidly because of the monotonic nature of the bounded upward jumps, which are the source of the name HOP.
7. All that matters is the ordering of the function values, so \( f \) can be replaced with ordinal numbers, i.e., integers indexing the array \( f \) in increasing order.
8. The first step of the ZOBOV algorithm (Neyrinck 2008) and most Morse theory algorithms is based on what is essentially the same as HOP.
9. There is no loss of information due to smoothing in the process of assembly of objects into structures, although Bayesian blocks can be thought of as a form of smoothing (more properly “chunking”) in preprocessing.
10. HOP is a major simplification, sidestepping much of the complexity of Morse theory (continuous or discrete) and
Persistence concepts that characterize modern topological data analysis (see references in Section 2).

(1) Nonetheless, the results presented here roughly match in number, size distribution, and visual appearance those from more elaborate algorithms, e.g., based on discrete Morse theory. The essential difference is that small structures, discarded by others because they are not persistent as some parameter is varied, we regard as conveying important information and are therefore retained in our analysis.

(2) Any one application of HOP finds maxima or minima (not both at the same time). Each galaxy is assigned to a nearby maximum (minimum) even if it lies in a sparse region clearly associated with a local minimum (maximum). A possible concern therefore is that the outer reaches of our structures around density maxima are contaminated by galaxies properly associated with density minima (and vice versa). Furthermore, some galaxies must lie in the “no man’s land” regions between voids and clusters. Some analysis methods address these issues by truncating structures, utilizing saddle points and intersections of ascending and descending manifolds. Even though we do not impose any spatial cutoffs, the visual appearance of our structures does not show any obvious problems of this sort.

Before showing examples of structures determined with HOP, the next section describes some considerations relevant to another way to detect voids.

3.2. Delaunay Tetrahedra as Void Tracers

As described in Section 1, the so-called voids in the galaxy distribution have been the subject of considerable study. These features are informative regarding the multi-scale structure of the universe, just as are dense structures. A variety of definitions of voids and detection methods keyed to the defining characteristics provide different views of both individual and overall structures. As discussed above in Section 2, some detection methods focus on local density minima; others locate volumes of space empty of galaxies within the limits of the survey or simulation, with no explicit reference to local minima in a continuous density representation. We here develop an approach of the latter kind that we believe is novel in its use of HOP to collect Delaunay cells into voids.

Delaunay and Voronoi tessellations are duals to one another, each partitioning the data space into small subvolumes in different ways but elucidating similar spatial information. We have seen that cells in a Voronoi tessellation of galaxy positions are good building blocks for constructing a representation of the corresponding density field. However, Delaunay tessellation is much more effective than Voronoi tessellation for void detection. The toy example in Figure 2 compares their responses to an artificial empty region in a set of otherwise random 2D points. The strategy is to find an objective way to identify a set of cells approximating the void, for example, by selecting those cells larger than some adopted size threshold.

Several problems beset Voronoi tessellation’s partial success in the left panel. Identification of void cells is relatively complicated. Both upper and lower thresholds are required, the putative void cells shown in gray lying in between. The cells above the upper threshold (black) are oversized owing to edge effects (see Paper I), and those below the lower threshold (white) are in the denser non-void region. Typically there is no range of cell areas that includes all the cells in the void and only those cells—that is, rejecting both edge and extra-void cells. In the left panel of Figure 2 three obvious edge cells are incorrectly denoted as void cells (gray). Adjusting the upper area threshold to correct this mistake eliminates some true void cells. Even the best Voronoi coverage does not provide a very exact representation of the void’s shape and extent. Finally, since they must contain a data point, the identified void cells not only extend outside the true circular void but carry a nontrivial density, namely, unity divided by the cell area. On the other hand, in the Delaunay triangulation (Figure 2, right panel) the circular void is well represented using a single area threshold.
Figure 3. Delaunay triangulation of 132 random points in the unit square. The set of shaded triangles constitutes a connected region empty of points but with a shape dictated by the rather arbitrary choice of which triangles to paste together.

yielding a small number of triangles, each of which is empty and can be interpreted as carrying zero density.

Furthermore, sets of several contiguous Delaunay tetrahedra typically make up structures devoid of galaxies. Figures 2 and 3 show quirky 2D examples of this fact. In the right panel of Figure 2 the dashed line running diagonally between the lower-left and upper-right corners does not obviously define a structure of any interest, but in fact the set of Delaunay triangles that it intersects is a void in the form of a jagged polyhedron not containing any of the points. Figure 3 demonstrates the same thing for triangles intersected by a continuous curve (not shown) in the plane. Neither of these shapes is what one thinks of as a reasonable void, but as discussed in Section 3.3, this does not mean that they are somehow not real. In principle, collections of adjacent tetrahedra are not necessarily empty of galaxies (see Section 3.5), but in the analyses presented here they always are. It is conceptually pleasing to identify truly empty volumes in a principled way and avoid the somewhat oxymoronic construction “void galaxies.” Of course, there are galaxies defining the surface of our polyhedral voids, so this feature should not be overemphasized.

The 2D toy examples in Figures 2 and 3 were chosen for ease of visualization, but these conclusions are even more definitive in 3D: many paths similarly define snake-like connected configurations of empty Delaunay tetrahedra (see Figure 3). Detected void structures are very sensitive to the assembly process, as was evident in the Aspen-Amsterdam void finder comparison project. The void volume fractions with the various void finders reported in Column 4 of Table 2 of Colberg et al. (2008) range from 0.13 to 1.0. The last number, due to Platen and Van de Weygaert’s “Watershed Delaunay Tessellation Field Estimator,” means that the entire data space is represented as a single highly convoluted but empty polyhedron. With a goal of identifying coherent low-density structures, an obvious scheme is to start from locally largest Delaunay tetrahedra and apply criteria for attaching one or more of the four face-sharing neighbors, perhaps based on something like size, shape, or distance. To avoid subjectivity of such ad hoc procedures, we use HOP’s prescription for partitioning the set of tetrahedra composing the Delaunay tessellation (Section 3.4). To do this, we need a new definition of the function $f$ in algorithm 7, since a surrogate for galaxy density is not appropriate for void finding. In order to represent the degree of emptiness, we take $f$ equal to tetrahedron volume. In much the same way that small Voronoi cells correspond to large density, large Delaunay cells correspond to a large degree of emptiness.$^{10}$

3.3. Ambiguity, Uncertainty, Noise, and Persistence

Quantifying uncertainty in data analysis requires careful assessment of any process affecting the signal of interest, either randomly or systematically, anywhere along the entire chain leading from raw measurements to the final estimate. In addition, subsequent interpretation must allow for dependencies of the results on the analysis method. Hence at a minimum, the following issues need to be considered in assessing the cosmological significance of the present analysis.

(1) Errors in measured sky positions and redshifts of individual galaxies.

10 Tessellation ameliorates dependencies on the size of the region sampled: “If you want to measure the density of biomass in a treetop, you have to choose a window of maybe a cubic foot. Ten times less and you sample either a single leaf or a blob of air. Ten times more and you have almost reduced the tree to an operational point” (Koenderink 1990). Tessellation turns this dilemma on its head: the data points adaptively fix both the size and location of the windows.
(2) The effect of random motions on estimated distances.
(3) Sampling bias connected with fiber collisions.
(4) Distortion of tessellation cells near edges of the data space.
(5) The initial random field of density perturbations.
(6) The finite number of galaxies forming randomly in the evolving density field.
(7) Sampling a small subset of the galaxies that have formed in the given volume.
(8) The variety of distinct but equally justifiable definitions of structures.
(9) The variety of different analysis methods.
(10) The variety of different selections of input data.

Many of these items are either noise (to be removed, diminished, or otherwise accounted for) or signal, depending on the context. The following discussion addresses this distinction, for the listed items, as dictated by our goals. The direct observational errors in issue 1 are described in Blanton et al. (2005); Abazajian et al. (2009) and include both small approximately normally distributed errors and larger outliers. In a nutshell, the sky-positional errors are quite small on the scale of interest here, and the random distance uncertainties derived from redshift errors are on average much less than $\approx 0.5$ Mpc. Figure 4 validates this statement via a direct comparison of the redshift and sky-position errors with the lengths of the edges of the cells in the Delaunay tessellation of the data. All measurements errors are small compared to even the smallest tetrahedra edge lengths, justifying neglect of these errors for even the smallest of the structures. Note that we do not try to remove redshift distortions (issue 2), reserving their treatment to post-factor examination of the shapes of dense clusters. The sampling bias issues connected with fiber collisions (issue 3) and Voronoi cells near edges (issue 4) were discussed in Paper I. For the present purposes all of these errors can be assumed to be small in magnitude and, owing to the inclusion of many individual galaxies, should not substantially impact the overall results.

The remaining entries in the above list require more discussion. Some are part of the astrophysical signal of interest. A large amount of research has been devoted to issues of uncertainty in computational topology (Edelsbrunner & Harer 2010) and topological data analysis (Zomorodian 2005). This work has focused on simplifications realized by discarding or consolidating less important features yielded by various analysis schemes. Examples of goals of this approach are amelioration of noise, especially discreteness noise; reduction of complexity and memory requirements; and promotion of better visualization and understanding of structure revealed by removing extraneous details. This methodology requires quantification of importance, ideas for which range from simple size criteria to the rather complex notions of topological persistence (Edelsbrunner 1987; Edelsbrunner et al. 2002; Gyulassy & Natarajan 2005; Edelsbrunner & Harer 2010; Gerber et al. 2010; Chen et al. 2013; see also http://math.stanford.edu/~gunnar/actanumericathree.pdf). Persistence methods postulate that the importance of a feature is measured by how long it is present as some parameter is scanned over a range of values. A quantitative link from persistence to probability may be obtained using bootstrap methods (Marzban & Yurtsever 2011; Chazal et al. 2013; Fasy et al. 2014). More recently, the persistence concept has made its way into astronomical applications (Sousbie 2011; Sousbie et al. 2011; Sousbie 2013; Cisewski et al. 2014) resulting in use of the term “the persistent cosmic web.”

However, with our goal of characterizing the complete range of multi-scale Structure, these methods discard some of the very information we seek. Modern cosmology posits that structure in the universe started as spatially random density fluctuations. Our universe evolved deterministically from this single set of
initial conditions; this process involves nothing like an ensemble of realizations of a random process. Accordingly, we consider items 5 and 6 signal, not noise. However, in other contexts, such as dark matter simulations, discarding small structures as unimportant consequences of initial spatial randomness may be useful. Item 7, sometimes called “discreteness noise,” is inherent to data consisting of a limited number of points drawn from an unknown distribution. Appendix B of Liivamägi et al. (2012) gives a detailed error analysis of this concept based on the Poisson model of Peebles (1980). But for reasons similar to those discussed with regard to items 5 and 6, we also regard item 7 as part of the astrophysical signal of interest. Nevertheless, the random Poisson data we have included may be of use in other contexts where noise abatement may be useful. Our Appendix B contains some further remarks about potential effects of what is often called topological noise.

While any of these last three factors, 5–7, are possible justifications for simplification using topological persistence or related measures, none are actually a source of uncertainty about the reality and nature of multi-scale structure in the current universe derived from a given redshift survey. The distribution of structures derived from discrete samples provides information about initial fluctuations and their subsequent evolution. Therefore, removing or smoothing away small-scale structures is at worst discarding useful cosmological information; at best, it makes the conclusions dependent on postulated models for the relevant physical processes. For this reason, and because the goal here includes geometrical and not just topological analysis, we do not employ any of the simplification procedures cited above. But in other contexts such as global analysis (e.g., estimation of a few summary topological statistics, such as genus, Minkowski functionals, or Betti numbers), it may be reasonable to regard scatter about a smooth correlation function or within realizations from different initial data as noise. In such cases, countermeasures such as topological persistence techniques may be justified.

Several of the potential noise sources listed above, namely, items 3, 6, 7, and 10, can be classed as sampling errors: issues that affect which galaxies are actually in our sample. These are not noise in the sense of measurement error for a parameter, but in some contexts these issues are relevant for characterizing the overall uncertainty of the analysis. Neither these errors nor the final analysis results (sets of clusters and voids) are simple parameters, so characterizing the uncertainty of the latter due to the former is difficult. A full and rigorous treatment of this form of uncertainty is beyond the scope of this paper. However, some insight can be gained from a simulation study related to the jackknife statistical procedure (Efron & Tibshirani 1993). The “leave-one-out” feature of jackknife resampling is equivalent to preventing a galaxy from entering the sample and can therefore shed light on the sensitivity of the cluster analysis to sampling error. The three tables below report straightforward analysis of a small subset of the full SDSS sample, namely, the 940 galaxies within ±0.01 redshift units of the centroid of the data. This restricted sample was analyzed with the same tessellation and Bayesian block analysis used on the full data set, as was each of 940 subsamples of 939 galaxies obtained by removing one from the restricted sample. Table 1 shows that all but four of the 940 cases had the same number of levels (namely, six) as the full sample (denoted “true” here). Table 2 elucidates the distribution of the blocks among these levels—which can be seen to be closely similar in the subsamples to that in the full restricted sample. Finally, Table 3 addresses the detailed correspondence between the identities of the galaxies in given blocks. As expected, in the overwhelming majority (96.44%) of cases the populations of the blocks are identical in the true and resampled cases, the main exceptions being those blocks containing the galaxy left out of that particular sample. As in the previous table, the statistical summaries omit the four cases where the number of levels is not equal to the true value, since identification of corresponding blocks is then problematic.

The largest source of ambiguity in multi-scale structure is the strong dependence of analysis results on analysis methodology, and the fact that there is no one correct methodology or definition of structures—see items 8–10. For example, we saw that by merely adjusting the halting criterion for assembling Delaunay tetrahedra into voids (Section 3.2), the output of voids ranges from a single void encompassing the entire space (see Figure 3) to a void for each tetrahedron. The question is not where between these extremes the truth lies but what representations provide the most useful information—for example, in the comparison of observations and simulations. Better yet, it can be very fruitful to study structural representation as a function of methodological assumptions and values of parameters of the analysis.

The HOP results that follow are examples of convenient representations using a simple notion of attaching to an elementary structure the neighbors that it dominates—as in the definitions of Voronoi cells, Bayesian blocks, and groups of building blocks that thread this paper. While a fairly natural construct, this is by no means claimed to be better or more fundamental than any others.

### 3.4. Structures Obtained with the HOP Algorithm

Now turn to some examples of the identification of spatial structures using the HOP algorithm to assemble the elementary objects or building blocks (i.e., Voronoi cells, Bayesian blocks, or Delaunay tetrahedra) into a unique set of connected
structures. Each such structure descends or ascends monotonically from one of the critical objects—local maxima or minima of the adopted density or voidness function. These peaks—for example, each Bayesian block denser than all its face-adjacent neighbors—can be easily identified by direct search but are also automatically produced by the HOP algorithm. The structures attached to the peaks are analogous to their watersheds. Such structures could be classified in one way or another (e.g., in the four customary classes: clusters, filaments, sheets, and voids, macroscopically of dimensionality 0, 1, 2, and 3, respectively), but their shapes are widely distributed in shape-space and do not fit cleanly in discrete clusters of shape parameters.

The sole information needed for each galaxy consists of two items, the first being the value of the HOP function \( f \)—typically a density estimate or its surrogate. It is natural for tessellation-based studies to take as the density of an object the number of galaxies in it divided by its volume. The reasoning for individual Voronoi cells is straightforward: small cells occur in crowded regions where the cell size is small. This relation intuitively supports the idea that the reciprocal of a cell volume is a reasonable surrogate for local density at or near that cell. In addition, this construct can provide an unbiased estimate of local density (Platen et al. 2011). The fact that only the relative order of the densities matters (item 7 in the list of properties of HOP in Section 3.1) is further protection against bias effects. Correspondingly, the density we assign in the case of Bayesian blocks is the number of galaxies in the block divided by its volume, the latter defined as the sum of the volumes of the cells making up the block. In the KDE case we evaluate the estimated continuous density field at the position of each galaxy. One of the SOM parameters is taken as a rough density surrogate (Paper I).

The other item necessary is a list of adjacent neighbors for each galaxy. As indicated earlier in Section 3.1 for the Voronoi-based tessellations, we take two objects \( A \) and \( B \) (cells or blocks) to be adjacent to each other if and only if there is at least one pair of Voronoi cells, one member of the pair in \( A \) and the other in \( B \), that share a common face. Here we take advantage of the natural definition of a data-adaptive number of near neighbors that Voronoi tessellation provides. The KDE algorithm does not determine neighbors, and we simply impose the adjacency information copied from the Voronoi tessellation. Two Delaunay tetrahedra are considered adjacent if and only if they share a common triangular face.

Note that Voronoi cell volume is not a property of a single galaxy but is determined by its propinquity to its neighbors; hence, information from distances to other galaxies is represented in both the cell volumes and the identities of neighboring cells.

Table 4 summarizes some of the basic properties of the collections of structures resulting from five choices for the building blocks for structures in the SDSS. In higher-dimensional Bayesian blocks (Jackson et al. 2010) one constructs a 1D array consisting of ordered values of a cell variable. In Paper I this quantity was taken to be the volume of the Voronoi cell. Since HOP more naturally operates on density, the whole Bayesian block analysis of Paper I was redone. This time density was used as the cell variable in addition to that of volume. These two analyses are listed in the first two rows of the table, showing that there is not a large difference in the number of structures identified. The nature of the HOP input for the other three cases is described by the corresponding entries in the first two columns of the table. The second column indicates the definition of \( f \), taken to be the density of galaxies within a Bayesian block or Voronoi cell, KDE density, or SOM class. The objects fed to the HOP algorithm (Column 1) are individual galaxies except in the first two cases, where they are collections of galaxies in blocks. The third column indicates the number of objects input to the algorithm. The last two columns give the number of structures, or groups of galaxies, associated with density maxima and minima.

The following two tables record similar information for the other two data sets, the MS and independently and randomly distributed points, respectively.

For a representative selection of the cases in Tables 4, 5, and 6, Figure 5 plots normalized distributions of the structures’ effective radii, defined in terms of its volume \( V \) by

\[
R_{\text{eff}} = H_0 \left( \frac{3V}{4\pi} \right)^{1/3}.
\]
Figure 5. Distributions of the effective radii of groups connected with density maxima (left) and density minima (right) found with HOP for the three standard data samples: SDSS (top), MS (middle), and Poisson (bottom). Circles: BB (volume); squares: BB (density); solid: KDE; thick solid: $1/V_{\text{Voronoi}}$; and dashed: SOM class. In each case the base-10 log of the number distribution is plotted against the effective radius in Mpc.

Here for $V$ we use the sum of the volumes of the Delaunay tetrahedra in the structure, but alternatively one could use the equal or slightly larger volume of the convex hull. The distributions obtained with direct local density estimates (BB, Voronoi, and KDE) are similar, with broad peaks in the range 10–20 Mpc. The SOM distributions are based on an HOP function that is discrete and only indirectly expresses density, so it is not surprising that they are rather different from the others. These distributions are quite similar to that shown in Figure 2 of Pan et al. (2012).

The region of the Sloan Great Wall is perhaps the richest region of the nearby universe. Figure 6 compares our group structures from Voronoi cells alone (corresponding to row 4 of Table 4 labeled $1/V_{\text{Voronoi}}$; colored polygons) with superclusters in this region (crosses inside circles) taken from Table 1 of Einasto et al. (2011). The bulk of the Sloan Great Wall is in the lower-left quadrant. This figure is limited to galaxies in the redshift range 0.045–0.085 ascribed to the Great Wall. To eliminate some clutter, only HOP groups with 25 or more galaxies and projected areas of more than 16 square degrees are shown. The correspondence with previously cataloged superclusters is not one-to-one, as expected because of the very different detection principles involved.

3.5. Some Properties of Delaunay Voids

We now discuss voids in more detail, pointing out some potential problems that need to be addressed, including the possibility of galaxies lying inside Delaunay voids, including edge effects similar to those mentioned above and in Paper I for dense structures.

The Delaunay tetrahedron method described above yields volumes almost completely devoid of galaxies in the sample. From the way these void structures are constructed, the galaxies at the vertices of the component tetrahedra for the most part lie on the void surface, leaving the inside empty. A surface galaxy is one from which there is a path to the outside that does not intersect any of the tetrahedra. With this definition a galaxy is inside a Delaunay void if and only if all of the void’s tetrahedra that have it as a vertex cover the full solid angle ($4\pi$ sr) as seen from that galaxy. A simple but effective procedure to identify Delaunay voids and possible interior galaxies is as follows.
Figure 6. Sky distribution of galaxies for the region of the Sloan Great Wall. The HOP groups are delineated by polygons filled with random colors. These are the projected 2D convex hulls of the sky positions of galaxies contained in the group, but slightly expanded to improve the visualization. The opacities of the polygons are linear in the redshift for the group: close darker ones thus appear to be in front of more distant lighter ones. The heavy black circles with plus signs are nominal positions of the 13 superclusters in this region given in Table 1 of Einasto et al. (2011).

1. Compute the Delaunay tessellation of the galaxy positions.
2. Identify groups of tetrahedra making up voids using HOP with $f =$ tetrahedral volume.
3. For each such Delaunay void, containing $N_{\text{void}}$ tetrahedra:
   
   (a) Collect a list of all $4N_{\text{void}}$ triangular faces of the tetrahedra making up the void.
   
   (b) Identify the faces that appear in this list only once.
   
   (c) The vertices of such faces are on the surface.
4. Identify as internal galaxies any that are not on the surface.

In summary, we define surface galaxies as those that populate the hull (not to be confused with the convex hull) of the galaxies circumscribing the void; any void galaxies not on the hull are then internal. Our analysis method strongly disfavors internal galaxies: not a single one of the many HOP-found Delaunay voids reported here contains even a single one. The mechanism behind this numerical result is unclear, but under HOP internal galaxies are extremely rare or possibly impossible.

While edge effects in Delaunay tessellations are less serious than in Voronoi tessellations, a second potential problem is that tetrahedra at the edges of the data space are systematically larger than they would be if not located there. Owing to the complexity of the SDSS boundaries in three dimensions, an automated test for whether or not a tetrahedron is at or near an edge is difficult. Here we compute the minimum distance of the four galaxies in a tetrahedron from the nearest point on the convex hull of the data. A complication arises when the outward-facing triangles of tetrahedra at the edge are exceptionally large, for then this minimum intergalactic distance is not actually representative of the distance from the edge. This difficulty is easily circumvented by adding points just outside these triangles, thus creating an augmented convex hull, faithful to the actual one but with no large faces. Figure 7 displays, via 2D histograms, the joint distributions of tetrahedron effective radius and minimum distance from the augmented convex hull. These plots illuminate the “edge effect”: within approximately 0.01 redshift units of the hull the Delaunay tetrahedra are anomalously large. This result motivates the cut on hull distance described below.

Figure 8 shows 3D plots of the four largest SDSS voids that are farther than 0.007 redshift units from the augmented hull. The largest of these voids (upper left panel) is centered at \( \text{R.A.} = 14.5 \pm 0.2^\circ, \text{decl.} = 39.4 \pm 2.2^\circ, \text{and} \ z = 0.1041 \pm 0.0034 \) and is therefore near and possibly associated with the so-called Boötes or giant void, given various positions by different authors---e.g., R.A. \( \approx 13^\circ (11.5–14.3), \text{decl.} \approx 40: (26.5–52.0), \) and \( z \approx 0.11 \) by Kopylov & Kopylova (2002). Theirs is a very different kind of structure, an order of magnitude larger in linear size and containing, according to these authors, not just galaxies but 17 clusters in the ranges shown. It is clear that we are finding very different void structures than those obtained with other methods.

Voids are typically considered to be regions where the density of galaxies is \( \approx 10\% \) of the mean density, with sizes mostly less than \( \approx 10 \text{ Mpc} \) (e.g., Pan et al. 2012; Coil 2012; Patiri et al. 2006). By definition the Delaunay voids derived here have no galaxies within the analyzed data set and hence have close to 0\% of the mean density.

Of course, the density in the voids is not exactly zero, and we now derive some crude upper limits on densities implied by empty voids. Define the dimensionless parameter

\[
\Lambda \equiv \lambda V,
\]

where \( \lambda \) is the density (galaxies per unit volume) and \( V \) is the volume under consideration. Then the likelihood for \( n \)
Figure 7. 2D histograms: counts of the Delaunay tetrahedra as a function of effective radius (Mpc, ordinate) and minimum distance from the circumscribing galaxies to vertices of the augmented convex hull of the full data set (redshift units, abscissa). Left: SDSS; middle: MS; right: Poisson. The density scale and contours represent the base-10 logarithm of the number of cases (per unit area in the plot, Mpc $\times$ redshift units, although only relative values are of interest). Most of the tetrahedra in the tails with hull distances less than about 0.01 and radii greater than about 10 are assumed to be artificially enlarged by edge effects.

Figure 8. Four largest SDSS Delaunay voids judged to be free of edge effects. The two integers in the legends are the number of circumscribing galaxies and tetrahedra, respectively—followed by the effective radius $R$ in Mpc and in parentheses the convexity (the sum of volumes of the tetrahedra divided by the volume of the convex hull of the galaxies).

Points independently distributed in $V$ is given by the Poisson distribution

$$P_n(\Lambda) = \frac{\Lambda^n e^{-\Lambda}}{n!}$$

since there is no background in this exercise. Thus, the likelihood for an empty volume is just

$$P_0(\Lambda) = e^{-\Lambda}.$$  \hspace{1cm} (3)

The density upper limit at significance level $p_0$ corresponds to the dimensionless parameter $\Lambda_{\text{limit}}$ satisfying

$$F_{\text{exp}}(\Lambda_{\text{limit}}) = \int_{\Lambda_{\text{limit}}}^{\infty} e^{-\Lambda} d\Lambda = e^{-\Lambda_{\text{limit}}} = p_0,$$  \hspace{1cm} (5)

where $F_{\text{exp}}$ is the cumulative distribution function of the exponential distribution.
and the corresponding upper limit on the actual density is

$$\lambda_{\text{ul}} = -\frac{\log p_0}{\nu}$$

(6)
at the significance level $p_0$.

This simple analysis does not take into account the fact that the voids do not cover random positions but in fact are circumscribed by a set of galaxies at their edges. This environmental effect can be approximately included by using the probability distribution of tetrahedron volumes in a Delaunay tessellation. Muche (1996) studied the distributional properties of Delaunay cells in a 3D Poisson distribution. Rewritten slightly in terms of our dimensionless density variable $\Lambda$, Equation (3.1) of Muche (1996) is

$$f(\Lambda) = 70 \int_0^\pi \int_0^{\pi-x} \int_0^\pi \Lambda \sin z \times \exp \left[ -\frac{2\pi \Lambda}{g(x, y)(1 + \cos z)\sin^2 z} \right] dz \, dy \, dx,$$

(7)

where

$$g(x, y) = \sin(x) \sin(y) \sin(x + y).$$

(8)

Replacing Equation (5) is

$$F_{\text{tess}}(\Lambda_{\text{limit}}) \equiv \int_{\Lambda_{\text{min}}}^{\Lambda_{\text{limit}}} f(\Lambda) \, d\Lambda = p_0.$$  

(9)

Straightforward numerical evaluation of this triple integral, plotted in Figure (9), shows that the simple exponential function underestimates density upper limits by about a factor of five.

Figure (10) displays the upper limits based on Equation (5), but as just mentioned, those from Equation (9) can be obtained approximately by applying a factor of five. Figure (10) plots the distribution of density upper limit for $p_0 = 0.05$ (equivalent to a 95% confidence). In order to avoid cells with inflated volumes due to edge effects, a cut of 0.0055 redshift units (205 Mpc) was applied on the minimum distance between the void vertices and the nearest face of the convex hull of the full data set. In addition, most of the large number of upper limits larger than the overall mean density (shown as a vertical dashed line) are simply not shown.

Neither of the results discussed above are rigorous upper limits. For example, they do not take into account that the underlying density is not constant. What is more reliable is the comparison of the upper limits based on the three data samples: the SDSS and MS samples yield a number of voids with lower limits considerably smaller than the mean density and somewhat lower than the random sample. Further density information about the voids detected may come from a census of SDSS galaxies not included in our volume-limited (VL) sample. Some of those not included in the VL sample will have spectroscopic redshifts to the limit of the Main-Like galaxy sample ($m_r < 18$) (see Appendix A.3), while others may have photometric redshifts to the limit of the photometric sample ($m_r < 21$) (see York et al. 2000; Strauss et al. 2002). A future paper will explore density limits in voids, as well as other descriptors of structures.

4. CONCLUSIONS

We have provided tools and data products to explore the multi-scale structure of the distribution of galaxies, using the SDSS DR7 redshift survey data, and in comparison with simulated and random data. The procedure demonstrated here departs from other work in use in several ways, starting with structural building blocks in the form of tessellation elements or small collections of them. The HOP algorithm was used to identify structures connected with density maxima and minima. Density estimators described in Paper I for the HOP function were used, coupled with adjacencies defined by Voronoi-Delaunay tessellation. Our HOP-based procedure is much simpler than those based directly on discrete Morse theory. Nevertheless, it identifies all local density maxima and minima plus their descending and ascending manifolds. Furthermore, we eschew methods such as topological persistence because, by eliminating structures based on a notion of importance, they unnecessarily discard valuable information. Such methods may allow one to concentrate on certain salient features, but the results are then dependent on the choices of importance quantifier and methods for discarding, combining, or otherwise modifying features.

Note that nature does not single out any one definition of structural elements or procedures for identifying and characterizing them. Methods invoking other than the purely geometrical information utilized here, such as colors or gravitational binding, undoubtedly yield very different structural descriptions. This dependence on methodology is not an uncertainty, statistical or otherwise, but an inevitable and useful feature of the diversity of analysis approaches.

A few summary statistics are presented here, and further details will be presented in future publications. The 3D distributions of multi-scale structure are not easily displayed in a paper. We encourage the reader to explore visualizations of the data given as electronic-only material. Such displays might be compared to the density map shown in Figure 2 of Gott et al. (2008), which suggests a visual similarity of the distribution of the highest 7% and lowest 7% of a smoothed and pixelated density distribution. Such displays, of course, cannot convey a complete visualization; rather, they show that these two spatial distributions of very different quantities (one of galactic density, the other of degree of local sparsity) are at least superficially quite similar. A future paper will explore this similarity by investigating auto-correlation functions, cross-correlation functions, and other statistical techniques and make detailed comparison with...
similar collections of multi-scale results such as dense structures (Park et al. 2012) and voids (Sutter et al. 2012).

All of the data needed to construct structure catalogs are contained in electronic-only files described fully in Appendix A. We provide files containing information to construct catalogs of multi-scale density and void structures based on several modes of analysis of the data, with a special eye toward comparing SDSS structures with those in simulation and purely random data sets. Each galaxy is assigned to a structure defined by a local maxima, and the collection of these for a given maximum defines a max-structure (also called a group or cluster). In addition, each galaxy is also assigned to a structure defined by a local minimum, with the collections defining min-structures (or voids). One can apply cutoffs in order to limit the outskirts of individual structures, procedures to eliminate structures that are not significant according to a given criterion, and possibly other post-processing techniques.

The geometrical structures in the spatial distribution of galaxies, here termed the multi-scale structure of the universe, have been said to form a hierarchy (van de Weygaert et al. 2009; Neyrinck 2008; Aragon-Calvo et al. 2010b; Aragon-Calvo & Szalay 2013), often without a precise definition of what this means. As mentioned in Section 2, this nomenclature might be taken to imply the existence of a set of discrete levels ordered to yield a hierarchy. These ideas seem to be derived from concepts such as merger trees in theoretical or phenomenological models of structure formation; see the discussion in Knebe et al. (2013). It is possible that such effects reveal themselves as a statistical tendency for small structures to lie within larger ones rather than at random, in the time-honored notion of clusters of clusters of galaxies (popularly referred to as superclusters), or similar constructs (but see Yu & Peebles 1969). Detection of such relationships would require further analysis of the structures reported and will be considered in a future publication. However, studies of characteristics of multi-scale structure in the universe, including the present work, demonstrate only continuous distributions and fail to show evidence of discreteness in either qualitative characteristics or quantitative observables. A continuous or self-similar distribution (e.g., Einasto et al. 1989) is, if anything, the opposite of a discrete hierarchy. A similar point has been made by Peebles (1974, 1984). Note that the discrete density levels of Bayesian blocks or KDE described here and in Paper I are a contrivance for density representation and have nothing to do with a discrete hierarchy of any kind in the actual distributions.

The catalogs and other data products given here can be utilized by any group to compare the structures found by any technique and make them immediately comparable to those of another. Future papers in the current series will describe more detailed statistical summaries of geometric and topological properties of overdense, underdense, and empty structures and carry out various comparisons with previous cluster and void catalogs.

We are grateful to the NASA-Ames Director’s Discretionary Fund and to Joe Bredekamp and the NASA Applied Information Systems Research Program for support and encouragement. Thanks go to Ani Thakar and Maria Nieto-Santisteban for their help with our many SDSS casjobs queries. Michael Blanton’s help with using his SDSS NYU-VAGC catalog is also very much appreciated. We are grateful to Patrick Moran, Christopher Henze, Changbom Park, Paul Sutter, Mark Neyrinck, Thierry Sousbie, Tom Abel, Pratyush Pranav, Peer-Timo Bremer, Attila Gyulassy, James (“B.J.”) Bjorken, and Jessi Cisewski. Special thanks go to Slobodan Simic and members of the CAMCOS project at San Jose State University, Joseph Fitch, David Goulette, Jian-Long Liu, Mathew Litrus, Brandon Morrison, Hai Nguyen Au, and Catherine (Boersma) Parayil for useful comments and for an ongoing collaboration on developments of
the HOP algorithm for topological data analysis. None of these acknowledgments should be construed to imply agreement with the ideas expressed here.

Thanks to Jose Gaite for pointing out his tetrahedra collection scheme, which we were unaware of until we disseminated a preprint of this publication.

Funding for the SDSS has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is http://www.sdss.org/

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, the Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy, the Max-Planck-Institute for Astrophysics, New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

The Millennium Run simulation used in this paper was carried out by the Virgo Supercomputing Consortium at the Computing Centre of the Max-Planck Society in Garching. The semianalytic galaxy catalog is publicly available at http://www.mpa-garching.mpg.de/galform/agnpaper.

This research has made use of NASA’s Astrophysics Data System Bibliographic Services. This research has also utilized the viewpoints (Gazis et al. 2010) software package.

APPENDIX A

DATA DETAILS

Of the three data sets studied, the first is a VL sample of 146,112 galaxies drawn from the SDSS. The second catalog is drawn as similarly as possible from the MS, yielding a VL sample of 171,388 galaxies. Third is a set of 144,700 points mimicking the SDSS VL sample but randomly and independently distributed so that there is no spatial structure beyond that imposed by the SDSS sampling. All conversions from redshift coordinates to Mpc are based on a Hubble constant of 73 km s$^{-1}$ Mpc$^{-1}$.

A.1. The SDSS NASA/AMES Value Added Galaxy Catalog (AMES–VAGC)

This section provides details in addition to those given in Paper I. The NASA Ames Research Center SDSS Value Added Catalog (NASA–AMES–VAGC) is based on the New York University Value Added Catalog (NYU–VAGC Blanton et al. 2005), which is in turn derived from Data Release 7 of the SDSS (Abazajian et al. 2009). We now describe the stages in the catalog creation.

A.2. Stage 1: Extracting Tables from the SDSS NYU-VAGC

The contents of a number of NYU-VAGC fits table files (described below) were extracted and used to create stage 1 of the catalog. An index of those fits files is listed below. At the time the catalog was created only the NYU-VAGC had SDSS K-corrected absolute magnitudes readily available, and hence we did not originally use the catalogs available via the excellent SDSS casjobs server.12

Selections were applied to each of the following three NYU-VAGC fits files:

1. object_sds5_spectro.fits:
   a. SDSS_SPECTRO_TAG: Galaxy Spectrum exists
   b. PRIMTARGET: Select Main Galaxy Sample targets
   c. OBJTYPE: Select type GALAXY
   d. CLASS: Select type GALAXY
   e. Z: Estimated redshift
   f. Z_ERR: Estimated redshift error. Only allowed to be greater than zero since negative values indicated an invalid estimate
   g. ZWARNING: Must be equal to zero to indicate no warning flags in the redshift estimation procedures

2. object_sds5_imaging.fits:
   a. RA: Right Ascension
   b. DEC: Declination
   c. NCHILD: Must be zero, indicating that it is not part of a blended parent or blended itself (!BLENDED)
   d. RESOLVE_STATUS: Used to obtain only one instance of each object
   e. VAGC_SELECT: To satisfy the Main-Like criteria of the NYU-VAGC
   f. FLAGS: Include only !BRIGHT, !BLENDED, !SATURATED
   g. MODELFUX: Model Magnitude fluxes (extinction corrected)
   h. MODELFUX_IVAR: Inverse variance of the fluxes (flux errors)
   i. PETROR50: 50% Petrosian Radius
   j. PETROR90: 90% Petrosian Radius

3. kcorrect.noonemodel.z0.10.fits:
   a. ABSMAGS: Absolute magnitudes in U, G, R, I, Z, J, H, K using a 0.1 blueshift of the bandpasses for k-corrections.

The outputs of these selections were concatenated into a single stage 1 NYU-VAGC Main-Like Galaxy Sample catalog containing 561,421 galaxies. See Figure 11 for a plot of the points in right ascension (R.A.) and declination (decl.). The catalog at this stage contained an internally assigned identification number, R.A., decl., apparent magnitudes (u, g, r, i, z), apparent magnitude errors, absolute magnitudes (U, G, R, I, Z, J, H, K), absolute magnitude errors, redshift, redshift error, and Petrosian 50% and 90% radii.

A.3. Stage 2: Obtaining a Contiguous and VL Sample

The maximum number of galaxies in our VL sample consistent with common practices in using the SDSS turned out to be 163,157. These selections (e.g., Choi et al. 2010) are redshift $z < 0.12$ and absolute magnitude in the r bandpass $M_r^{0.1} < 20.0751$. These are consistent with a red-band apparent magnitude upper limit defined by the Strauss et al. (2002) Main Galaxy Sample as $r < 17.77$, although the NYU-VAGC Main-Like sample goes down to $r = 18$.

Next, samples were removed outside a defined contiguous region avoiding several irregular features extending beyond the smooth outer (2D) shape of the distribution of points, as well as disconnected and isolated patches lying entirely outside.

---

11 Such identically and independently distributed (IID) processes are often called Poisson processes (here with a spatially constant event rate) because the counts in fixed volumes obey the Poisson distribution.

12 http://casjobs.sdss.org
The Astrophysical Journal, 799:95 (24pp), 2015 January 20
Way, Gazis, & Scargle

Figure 11. Plot of the entire SDSS DR7 Main-Like Galaxy Sample from the NYU-V AGC catalog (both black and gray points). The points in gray are those of the VL subsample derived from stage 2 of the catalog as described in Appendix A.1. The black points were eliminated from the VL sample as a result of redshift and absolute luminosity cuts ($z < 0.12$ and $M^B_0 = -20.0751$) and the desire for a contiguous geometric sample.

This region was centered on the north galactic cap roughly corresponding to $100 < R.A. < 270$ and $-7 < \text{decl.} < 65$. The contiguous region contains 146,112 objects and is defined by the gray area in Figure 11.

A.4. Stage 3: 55$''$ Fiber Placement Issue and Coordinate Transform

The angular separation in arcseconds to the six nearest neighbors for every point was estimated. This allows one to quickly identify any neighbor within 55$''$. This was necessary because the fiber plug plate of the SDSS does not allow fibers to be placed closer than 55$''$ to each other. However, there are a large number of overlapping plates, which means that there are some galaxies with spectra within this 55$''$ fiber limit. Since these overlaps cover only part of the full area, it represents a systematic bias that must be eliminated in order to consistently sample the true underlying galaxy distribution. To do so, we removed a randomly chosen member of any pair found within 55$''$ of each other. This process eliminates 6314 galaxies from the sample.

To use Euclidean coordinates with units the same in all three dimensions, the right ascension ($\alpha$), declination ($\delta$), and redshift ($z$) were transformed into Cartesian coordinates according to

\begin{align}
 x &= z \cos(\delta) \cos(\alpha) \\
 y &= z \cos(\delta) \sin(\alpha) \\
 z &= z \sin(\delta)
\end{align}

(equivalent to the MatLab© function sph2cart), thus yielding rectangular coordinates, each with units of redshift and convertible to physical units by multiplying by $c/H_0$, with $c$ the speed of light and $H_0$ the Hubble constant. A nonlinear conversion can also be made for a given cosmological model, but it will yield only a small correction over the low-redshift range of these data.

A.5. Stage 4: Voronoi-related Calculations

The Voronoi tessellation (e.g., Okabe et al. 2000) of the remaining galaxies was calculated (see Paper I for more details). From this tessellation a number of additional parameters are derived:

1. Cell volume: $V$
2. The distance between each galaxy and the center of its Voronoi cell: $d_{CM}$
3. The minimum and maximum dimension of each Voronoi cell: $R_{\min}, R_{\max}$
4. Cell radius: $R_{\text{Voronoi}} = (3V/4\pi)^{1/3}$
5. A measure of cell elongation: $(E = R_{\min}/R_{\max})$
6. A measure of the magnitude of the local density gradient: $d_{CM}/R_{\text{Voronoi}}$
7. A scaling parameter for distances: the average density of the VL SDSS data raised to the minus 1/3 power: $d_{\text{uniform}} = 3.2 \times 10^{-3}$ in units of redshift.

The first three are fundamental properties of the Voronoi cells. They are defined for individual cells but are dependent on neighboring galaxies by virtue of the way the Voronoi tessellation is defined. In turn, they are used to derive useful properties 4, 5, and 6. The first two of these are summary descriptions of the size and shape of the cell. The separation

In Paper I it was claimed that 6540 galaxies were eliminated, but this is incorrect.
between each galaxy and the center of its Voronoi cell is a vector that approximates the magnitude and direction of the local gradient in the density of galaxies. It is here represented by its magnitude in item 6.

The average distance in item 7, a property of the full sets of galaxies in the catalogs, is not used in the assignment of individual galaxies to classes. Instead, it is used as a scaling factor to make distance parameters such as $d_{CM}$ and $R_{Voronoi}$ dimensionless. The average distance here is computed as the average spacing, $(V/N)^{1/3}$, between samples. The actual value for the MS was very near that of the SDSS, while the Random was set to this value when the data set was created. This quantity was chosen because it is well defined, straightforward to calculate, and insensitive to details such as the usage of the Voronoi tessellation algorithm.

A.6. Stage 5: FlaggingBoundary Points

The cells near the boundaries of the tessellated volume are distorted to one degree or another. Depending on the distance of the cell from the boundary, this effect ranges in importance from small to large. The most distortion happens when the tessellation algorithm assigns to a cell one or more vertices well outside the data volume, or even leaves a vertex undefined because it formally lies at infinity. One could attempt to correct for such distortion, but as described in Paper I, we feel that it is better to simply eliminate galaxies whose Voronoi cells appear to have been significantly distorted by boundary effects. Our criteria for identifying such cells, as detailed in the section titled “The Voronoi Cell Boundary Problem” of Paper I, led to the rejection of 5807 boundary cells, leaving 133,991 galaxies in the sample to be used for the SDSS density estimations reported here.

A.7. Stage 6: Building a Table for Casjobs

In order to make the sample useful for users of casjobs (where most SDSS users obtain their data), we have attempted to obtain SDSS object identification numbers from the PhotoObjAll.ObjID table for all of the objects in the final density sample. This was necessary because the NYU-VAGC DR7 catalog does not contain the same object identification numbers as those found in the SDSS DR7 casjobs catalog. To obtain the object identifications, the fGetNearestObjAllEq function of casjobs was used. Objects were matched within $1^\circ$ of the R.A. and decl. of the NYU-VAGC-derived objects. From 146,112 points (see Section 3.2), 145,875 PhotoObjAll.ObjID identifications were found (known simply as the ObjID in SDSS casjobs parlance), meaning that 237 points did not exist in the casjobs catalog. This 0.16% loss should not be a major inconvenience for casjobs-based procedures. Those 237 objects in the final NASA–AMES–VAGC catalog without casjobs ObjID numbers will still be in the publically released catalog but will instead contain an 18-character string (the same length as the unique SDSS ObjID) with each object numbered from 00000000000000001 to 00000000000000237.

A.8. The AdaptiveKernel Map Classes

In Table 3 of Paper I the Bayesian blocks (BB) and adaptive kernel map (AKM) methods had a number of classes that ranged from low density to high. The class structure for the self-organizing map (SOM) method was more complex (see Table 2 in Paper I). The AKM method produces a continuous range of densities rather than specific classes. In order to mimic the BB and SOM class methods, a filter was applied to the AKM densities to produce the 11 classes found in Table 3 of Paper I:

$$AKM_{class} = 12 - \text{round}((\log_{10}(AKM_{density})/0.6947) \times 20) - 7,$$

(A4)

Figure 12 shows the resulting correspondence between AKM density and class.$^{15}$

A.9. The Millennium Simulation AMES Value Added Catalog

To create a VL sample from the MS, a similar procedure was followed to that described in Appendix A.1. This is possible since one can obtain the absolute magnitude estimates in the same bandpasses as the SDSS for the galaxies in the MS (Croton et al. 2005). First, one must convert the MS Cartesian coordinates and velocities ($x$, $y$, $z$, $v_x$, $v_y$, $v_z$) to right ascension, declination, and redshift using $H_0 = 73$ km s$^{-1}$ Mpc$^{-1}$, $\Omega_m = 0.25$, $\Omega_{\Lambda} = 0.75$, and $\Omega_k = 0$. The apparent magnitudes were derived from the given absolute magnitudes using the luminosity distance. The luminosity distance requires the redshift and radial distances derived from the MS.$^{16}$ The same redshift

---

$^{14}$ Query: select a.*, b.objid as matchObjID into mydb.nyuvagccross from MyDB.densitycatalog a cross apply dbo.fGetNearestObjAllEq(a.ra, a.dec, 0.0167) b

$^{15}$ A better method to segment the data might have been to utilize the unique strengths of Bayesian blocks, but that was not done herein.

$^{16}$ See Cheng (2005); Peacock (1999); Hogg (1999) for more on the luminosity distance.
| No. | Name | Description | Value* |
|-----|------|-------------|--------|
| 1   | objid(high)b | NYU-VAGC Identifier | 588848898 |
| 2   | objid(low)b | NYU-VAGC Identifier | 834497650 |
| 3   | id2 | Running index | 1 |
| 4   | vagc_specobjid | NYU-VAGC Spectrum ID | 8270 |
| 5   | x | x-Coordinate | −0.103657 |
| 6   | y | y-Coordinate | 0.018571 |
| 7   | z | z-Coordinate | −0.001935 |
| 8   | ra | Right ascension | 169.842883 |
| 9   | dec | Declination | −1.052567 |
| 10  | redshift | Observed redshift | 0.105326 |
| 11  | redshift_err | Redshift error | 0.000021 |
| 12  | u | Apparent u magnitude | 19.411664 |
| 13  | g | Apparent g magnitude | 17.451290 |
| 14  | r | Apparent r magnitude | 16.461469 |
| 15  | i | Apparent i magnitude | 16.045569 |
| 16  | z | Apparent z magnitude | 15.674621 |
| 17  | u | Absolute u magnitude | −18.443730 |
| 18  | g | Absolute g magnitude | −20.302310 |
| 19  | r | Absolute r magnitude | −21.225930 |
| 20  | i | Absolute i magnitude | −21.588990 |
| 21  | z | Absolute z magnitude | −21.916710 |
| 22  | U | Absolute U magnitude | −22.972160 |
| 23  | G | Absolute G magnitude | −22.972160 |
| 24  | R | Absolute R magnitude | −22.972160 |
| 25  | I | Absolute I magnitude | −22.972160 |
| 26  | Z | Absolute Z magnitude | −22.972160 |
| 27  | p50_u | Petrosian 50% u radius | 4.096841 |
| 28  | p50_g | Petrosian 50% g radius | 5.219957 |
| 29  | p50_r | Petrosian 50% r radius | 5.338530 |
| 30  | p50_i | Petrosian 50% i radius | 5.066374 |
| 31  | p50_z | Petrosian 50% z radius | 5.263294 |
| 32  | p90_u | Petrosian 90% u radius | 9.189430 |
| 33  | p90_g | Petrosian 90% g radius | 16.446740 |
| 34  | p90_r | Petrosian 90% r radius | 17.238600 |
| 35  | p90_i | Petrosian 90% i radius | 16.903650 |
| 36  | dCM/R_Voronoi | Centroid → point (normalized) | 0.265860 |
| 37  | R_Voronoi/dUniform | Cell volume/total volume | 0.796684 |
| 38  | R_Max | Distance from sample to farthest vertex | 1.067450 |
| 39  | R_Min | Distance from sample to nearest vertex | 0.356638 |
| 40  | R_Max / R_Min | Elongation | 0.558307 |
| 41  | cnWinners | Class ID with most votes | 5 |
| 42  | volume | Cell volume | 1.6569451e-08 |
| 43  | bb_vol_lev | Level ID BB(vol)d | 11 |
| 44  | bb_den_lev | Level ID BB(den)e | 3 |
| 45  | bb_den_blk | Block ID BB(den)e | 344 |
| 46  | f55 | 0; but 1 if cell collision test fails | 0 |
| 47  | fbad | 0; but 1 if boundary test fails | 0 |
| 48  | density_akm | KDE density | 2687.7676 |
| 49  | bandwidth_akm | KDE bandwidth | 0.00058819564 |
| 50  | levels_akm | KDE density level | 7 |
| 51  | ID(Vor,+) | Max-structure ID; HOP f = 1/volume | 294 |
| 52  | ID(Vor,−) | Min-structure ID; HOP f = 1/volume | 65 |
| 53  | ID(AKM,+) | Max-structure ID; HOP f = 1/density_akm | 179 |
| 54  | ID(AKMD,+) | Min-structure ID; HOP f = 1/density_akm | 503 |
| 55  | ID(SOM,+) | Max-structure ID; HOP f = cnWinners | 1 |
| 56  | ID(SOM,−) | Min-structure ID; HOP f = cnWinners | 158 |
| 57  | ID(BB(volume),+) | Max-structure ID; HOP f = n(blk)/volume; BB(vol)d | 578 |
| 58  | ID(BB(volume),−) | Min-structure ID; HOP f = n(blk)/volume; BB(vol)d | 163 |
| 59  | ID(BB(density),+) | Max-structure ID; HOP f = n(blk)/volume; BB(den)e | 3 |
| 60  | ID(BB(density),−) | Min-structure ID; HOP f = n(blk)/volume; BB(den)e | 9 |

Notes.

* Value from the first row of the online table sdss_master.txt.

b These long integer identifiers are divided into two parts: the most significant nine digits (high) and least significant (low). Using Matlab, after executing load sdss_master.txt+, the string [ int2str(sdss_master(:, 1)) int2str(sdss_master(:, 2)) ]+ rejoins the two parts.

c dUniform, the average distance between objects in the sample, equals 3.2e-3 redshift units.

d Bayesian block analysis based on Voronoi cell volume.

e Bayesian block analysis based on Voronoi cell density.

(This table is available in its entirety in machine-readable form.)
Table 9
Column Identifiers: 171,388 MS Galaxies (ms_master.txt)

| No. | Name       | Description                        | Value a |
|-----|------------|------------------------------------|---------|
| 1   | objid      | NYU-VAGC Identifier                | 1       |
| 2   | id2        | Running index                      | 1       |
| 3   | x          | x-Coordinate                        | 7.180451|
| 4   | y          | y-Coordinate                        | 11.865457|
| 5   | z          | z-Coordinate                        | 6.817040|
| 6   | ra         | Right ascension                     | 63.682026|
| 7   | dec        | Declination                         | 58.840448|
| 8   | redshift   | Observed redshift                   | 0.006335|
| 9   | u          | Apparent u magnitude                | 10.343420|
| 10  | g          | Apparent g magnitude                | 8.671343|
| 11  | r          | Apparent r magnitude                | 7.864622|
| 12  | i          | Apparent i magnitude                | 7.498015|
| 13  | z          | Apparent z magnitude                | 7.173348|
| 14  | U          | Absolute u magnitude                | −20.612991|
| 15  | G          | Absolute g magnitude                | −22.285069|
| 16  | R          | Absolute r magnitude                | −23.091789|
| 17  | I          | Absolute i magnitude                | −23.458397|
| 18  | Z          | Absolute z magnitude                | −23.783064|
| 19  | dCM/R_Voronoi | centroid → point (normalized)     | 0.210046|
| 20  | R_Voronoi/dUniformb | cell volume/total volume               | 0.271421|
| 21  | R_{Max} | Distance from sample to farthest vertex | 0.380684|
| 22  | R_{Min}  | Distance from sample to nearest vertex | 0.186301|
| 23  | R_{Max}/R_{Min} | Elongation                          | 0.463751|
| 24  | cnWinners | Class ID with most votes            | 1       |
| 25  | volume    | Cell volume                         | 0.057023573|
| 26  | bb_vol_{lev} | Level ID BB(vol)^{f}               | 1       |
| 27  | bb_vol_{blk} | Block ID BB(vol)^{f}              | 46      |
| 28  | bb_den_{lev} | Level ID BB(den)^{d}               | 26      |
| 29  | bb_den_{blk} | Block ID BB(den)^{d}              | 56626   |
| 30  | f55       | 0; but 1 if cell collision test fails | 0       |
| 31  | fbad      | 0; but 1 if boundary test fails     | 0       |
| 32  | density_{akm} | KDE density                       | 9739.7164|
| 33  | bandwidth_{akm} | KDE bandwidth                     | 0.0007490186|
| 34  | levels_{akm} | KDE density level                   | 7       |
| 35  | ID(Vor,+) | Max-structure ID; HOP \( f = 1/\text{volume} \) | 2       |
| 36  | ID(Vor,−) | Min-structure ID; HOP \( f = 1/\text{volume} \) | 4       |
| 37  | ID(AKM,+) | Max-structure ID; HOP \( f = 1/\text{density}_{\text{akm}} \) | 2       |
| 38  | ID(AKM,−) | Min-structure ID; HOP \( f = 1/\text{density}_{\text{akm}} \) | 1       |
| 39  | ID(SOM,+) | Max-structure ID; HOP \( f = \text{cnWinners} \) | 1       |
| 40  | ID(SOM,−) | Min-structure ID; HOP \( f = \text{cnWinners} \) | 1       |
| 41  | ID(BB(volume),+) | Max-structure ID; HOP \( f = n(\text{blk})/\text{volume}; BB(vol)^{f} \) | 2481 |
| 42  | ID(BB(volume),−) | Min-structure ID; HOP \( f = n(\text{blk})/\text{volume}; BB(vol)^{f} \) | 897   |
| 43  | ID(BB(density),+) | Max-structure ID; HOP \( f = n(\text{blk})/\text{volume}; BB(den)^{d} \) | 10602 |
| 44  | ID(BB(density),−) | Min-structure ID; HOP \( f = n(\text{blk})/\text{volume}; BB(den)^{d} \) | 43    |

Notes.

a Value from the first row of the online table ms_master.txt.
b dUniform, the average distance between objects in the sample, equals 3.2e-3 redshift units.
c Bayesian block analysis based on Voronoi cell volume.
d Bayesian block analysis based on Voronoi cell density.

(This table is available in its entirety in machine-readable form.)

and absolute magnitude cuts as in the AMES–VAGC were applied, leaving 171,388 out of \( \sim 9 \) million points in the original MS catalog. A total of 16,283 points were eliminated to emulate the SDSS 55″ fiber collision issue, while 6178 were eliminated because of boundary effects. This leaves 148,927 points.

The same distance scaling factor as used for the SDSS data, as described in Appendix A.5, item 7, namely, \( d_{\text{uniform}} = 3.2 \times 10^{-3} \), was used to derive the same Voronoi quantities found for the SDSS in Appendix A.5.

Again, in order to mimic the BB and SOM class methods, a filter was applied to the AKM densities to produce the 13 MS classes found in Table 3 of Paper I:

\[
\text{AKM}_{\text{class}} = 14 - \text{round}((\log_{10}(\text{AKM}_{\text{density}})/5.6947) \times 20) - 7.
\]
Table 10
Column Identifiers: 144,700 Random Points (poiss_master.txt)

| No. | Name          | Description                  | Value  |
|-----|---------------|------------------------------|--------|
| 1   | id2           | Running index                | 1      |
| 2   | x             | Coordinate                   | -0.032697 |
| 3   | y             | Coordinate                   | -0.006482 |
| 4   | z             | Coordinate                   | 0.081098  |
| 5   | ra            | Right ascension              | 191.212624 |
| 6   | dec           | Declination                  | 67.656439  |
| 7   | redshift      |                              | 0.093803  |
| 8   | dCM/R_Voronoi | Centroid → point (normalized) | 0.851645  |
| 9   | R_Voronoi/dUniform | Cell volume/total volume | 0.981894  |
| 10  | R_Max         | Distance from sample to farthest vertex | 3.133988 |
| 11  | R_Min         | Distance from sample to nearest vertex | 0.348597 |
| 12  | R_Max/R_Min   | Elongation                   | 0.292865  |
| 13  | cnWinners     | Class ID with most votes     | NaN     |
| 14  | volume        | Cell volume                  | NaN     |
| 15  | bb_vol_lev    | Level ID BB(vol)c            | 0      |
| 16  | bb_vol_blk    | Block ID BB(vol)c            | 0      |
| 17  | bb_den_lev    | Level ID BB(den)d            | 0      |
| 18  | bb_den_blk    | Block ID BB(den)d            | 0      |
| 19  | f55           | 0; but 1 if cell collision test fails | 0      |
| 20  | fbad          | 0; but 1 if boundary test fails | 1      |
| 21  | density_akm   | KDE density                  | -9999   |
| 22  | bandwidth_akm | KDE bandwidth                | -9999   |
| 23  | levels_akm    | KDE density level            | 0      |
| 24  | ID(Vor,+))    | Max-structure ID; HOP f = 1/volumel | 1      |
| 25  | ID(Vor,-)     | Min-structure ID; HOP f = 1/volumel | 1      |
| 26  | ID(AKM,+))    | Max-structure ID; HOP f = 1/density_akm | 1477    |
| 27  | ID(AKM,-)     | Min-structure ID; HOP f = 1/density_akm | 1      |
| 28  | ID(SOM,+))    | Max-structure ID; HOP f = cnWinners | 1      |
| 29  | ID(SOM,-)     | Min-structure ID; HOP f = cnWinners | 1      |
| 30  | ID(BB(volume),+) | Max-structure ID; HOP f = n(blk)/volume; BB(vol)c | 0      |
| 31  | ID(BB(volume),-) | Min-structure ID; HOP f = n(blk)/volume; BB(vol)c | 0      |
| 32  | ID(BB(density),+) | Max-structure ID; HOP f = n(blk)/volume; BB(den)d | 0      |
| 33  | ID(BB(density),-) | Min-structure ID; HOP f = n(blk)/volume; BB(den)d | 0      |

Notes.

a Value from the first row of the online table poiss_master.txt.

b dUniform, the average distance between objects in the sample, equals 3.2e−3 redshift units.

c Bayesian block analysis based on Voronoi cell volume.

d Bayesian block analysis based on Voronoi cell density.

(This table is available in its entirety in machine-readable form.)

Table 11
Zwicky Morphological Matrix: Effect of Noise on Numbers and Sizes of Structures

| (a) Noise Effect | (b) δN | (c) Nonexistent | (d) Small | (e) Medium | (f) Large |
|------------------|-------|-----------------|-----------|------------|----------|
| A Create         | +1    | Small           | •••       | ••••       | •••••    |
| B Separate into Two | +1  | Small           | •••       | ••••       | •••••    |
| C Reduce content | 0     | Small           | Small     | Medium     | Medium   |
| D Increase content | 0   | Medium          | •••       | ••••       | •••••    |
| E Merge          | -1    | Medium          | Large     | Large+     | •••••    |
| F Destroy        | -1    | Null            | Null      | Null       | •••••    |

Notes. The size indicators here are only rough and nominal. The plus and minus signs are to be interpreted as “even larger” or “even smaller,” respectively.

A.10. The Randomly Distributed Point Catalog

The creation of the randomly distributed data point catalog was outlined in detail in Paper I. The initial data set contains a similar number of points (144,700) as both the AMES–VAGC and derived MS catalogs. The final catalog, after removing 6219 points corresponding to the 55" issue discussed previously and 6649 boundary points discovered after the Voronoi tessellation, yields 131,832 points.17

The galaxy positions were then converted to rectangular Cartesian coordinates according to the same formulas used for the SDSS data, namely, Equations (A1), (A2), and (A3). As above, any transformation that would require picking a value for the Hubble constant or a cosmological model was avoided. The same distance scaling factor used for the SDSS and MS data, as described in Appendix A.5, item 7, namely, $d_{\text{uniform}} = 3.2 \times 10^{-3}$, was used to derive the same Voronoi quantities found for the SDSS in Appendix A.5.

As in the previous two cases, in order to mimic the BB and SOM class Methods, a filter was applied to the AKM densities to produce the 10 uniform classes found in Table 3 of Paper I:

$$\text{AKM}_{\text{class}} = 11 - \text{round}((\text{log}_{10}(\text{AKM}_{\text{density}})/5.6947) \times 40) - 13.$$  

(A6)

17 The final number of points described in Paper I is incorrect. The 144,700 quoted was before the removal of the 55" and boundary value points, not after.
A.11. Structure Catalog Information: Electronic-only Files

For each of the three data sets we have constructed a flat ASCII file, the columns of which contain information, one row for each galaxy, of use for assembling many kinds of structure catalogs of interest. Many of the entries echo data from the NYU-VAGC data archive, for the reader’s convenience in constructing and exploring structure catalogs derived from the new results. The names of these electronically accessible files are sdss_master.txt, ms_master.txt, and poiss_master.txt, and the rest of this section describes their contents and provides a few notes on their use in constructing structure catalogs. The terms “Max-structure” and “Min-structure” mean HOP groups associated with local maxima and local minima, respectively.

The last 10 columns of structure IDs can be used to construct catalogs as follows. Let the MatLab variable index_structures denote an array containing the integers in one of these columns. This contains, for each galaxy G, the index of the structure to which the galaxy is assigned by the converged HOP iteration. Then one can construct an array containing these structure IDs using the MatLab command

\[
\text{ids} = \text{unique(index_structures)},
\]

which is also just the array 1, 2, \ldots, $M$, where $M$ is the number of structures HOP has identified. Then for any structure ID $m$ satisfying $1 \leq m \leq M$ the indices of the galaxies in that structure (indexed in the original raw data array, including galaxies that later failed the f55/fbad tests) can be found from

\[
\text{galaxy_indices} = \text{find(index_structures == ids(m))};
\]

This allows one to compute many things for that structure, such as the $xyz$-coordinates of all the galaxies in it, the volume of the structure (as the sum of the Voronoi volumes), the number of galaxies in it, and the density in galaxies per unit volume—using the corresponding data in the other columns of the master file.

The following two tables give similar identifications for the MS and Poisson data files. Fewer entries have been defined for these data sets, but the meanings of the parameters that are in common are the same.

A.12. Void Catalog Information: Electronic-only Files

For each of the three data sets we have also constructed a flat ASCII file containing identities and descriptions of the HOP voids. The names of these electronically accessible files for the SDSS, MS, and Poisson data are delaunay_voids_sdss.txt, delaunay_voids_ms.txt, and delaunay_voids_poiss.txt, respectively. All three files have the format given in Table 7. The identification number is a running index for the 458,173, 503,832, and 465,357 voids in the three cases, respectively. The next two columns contain the number of tetrahedra and galaxies in the void, followed by the void effective radius in Equation (1), convexity (see the caption to Figure 8), and distance from the augmented hull of the full data set. The columns beginning with Column 7 give the identities of the galaxies circumscribing the void; these are the running index values id2 in Tables 8, 9, and 10, respectively. These rows contain a variable number of galaxy IDs. All rows with less than the maximum number (25, 29, and 23 in the three cases, respectively) of galaxy IDs are padded with zeros to yield fixed record-length files. These lengths are 31, 35, and 29 in the three cases, respectively.

APPENDIX B

ZWICKY MORPHOLOGICAL ANALYSIS OF TOPOLOGICAL NOISE EFFECTS

The effect of sampling or measurement imperfections on data used to estimate a distribution, especially in higher dimensions, is always much more complicated than, say, for the case of simple parameter estimation. In the current spatial statistics context, the various processes discussed in Section 3.3 (and here, in the abuse of terminology described in that section, termed “noise”) may have several effects on the estimated structures. Table 11 is a morphological box, a device pioneered by Zwicky ([1948], 1957) to facilitate complete investigation of parameter spaces. The first column (a) of this chart lists all six possible effects noise of any kind may have on a specific structure, including creation of a new structure and modification or destruction of an existing one. Column (b) gives the accompanying net change in the number of structures. Given the “before” size of a structure shown in the headings columns (c)–(f) the “after” size (with noise) is entered in the rows below. The whole point of this construct is to bring to light effects that might not be obvious at first thought. For example, noise may actually bring about the apparent merger of two structures into one (row E), e.g., turning two small structures into one medium structure as in box E(d). A conclusion derivable from this matrix is that the common procedure of eliminating the smallest structures in the size distribution may be only partially effective at de-noising.

REFERENCES

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Ahn, C., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17
Akio, J., & Mahonen, P. 1998, ApJ, 497, 534
Amendola, L., Frieman, J., & Waga, I. 1999, MNRAS, 309, 465
Angulo, R., Chen, R., Hilbert, S., & Abel, T. 2014, MNRAS, 444, 2925
Aragon-Calvo, M., Jones, B., van de Weygaert, R., & van der Hulst, J. 2007, A&A, 474, 315
Aragon-Calvo, M., Platen, E., van de Weygaert, R., & Szalay, A. 2010a, ApJ, 723, 364
Aragon-Calvo, M. A., & Szalay, A. S. 2013, MNRAS, 428, 3409
Aragon-Calvo, M. A., van de Weygaert, R., Araya-Melo, P. A., Platen, E., & Szalay, A. S. 2010b, MNRAS, 404, L89
Aubert, D., Pichon, C., & Colombi, S. 2004, MNRAS, 352, 376
Bertschinger, E., & Gelb, J. M. 1991, ComPh, 5, 164
Beygu, B., Kreckel, K., van de Weygaert, R., van der Hulst, J. M., & van Gorkom, J. H. 2013, AJ, 145, 120
Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005, AJ, 129, 2562
Bolejko, K., Clarkson, C., Maartens, R., et al. 2012, PhRvL, 110, 021302
Bo, E., van de Weygaert, R., Dolag, K., & Pettorino, V. 2012, MNRAS, 426, 440
Cautun, M., & van de Weygaert, R. 2011, record ascl:1102.020
Cecerecelli, L., Paz, D., Laires, M., Padilla, N., & Lumb, D. G. 2013, MNRAS, 434, 1435
Chazal, F., Fasy, B., Lecci, F., et al. 2013, Modeling and Analysis of Information Systems, 20, 96
Chen, Y.-C., Genovese, C., & Wasserman, L. 2013, arXiv:1312.2098v1
Cheng, T. P. 2005, Relativity, Gravitation, and Cosmology: A Basic Introduction (Oxford: Oxford Univ. Press)
Chincarini, G., & Rood, H. J. 1976, ApJ, 206, 30
Choi, Y., Park, C., Kim, J., et al. 2010, ApJS, 109, 181
Ciesielski, J., Croft, R., Freeman, P., et al. 2014, MNRAS, 440, 2599
Clampitt, J., & Jain, B. 2014, arXiv:1404.1834
Colli, A. 2012, in Planets, Stars and Stellar Systems, ed. T. D. Oswalt (New York: Springer), http://ned.ipac.caltech.edu/level5/March12/Coil/Coil09.html
Colberg, J., Pearce, F., Foster, C., et al. 2008, MNRAS, 387, 933
Colless, M. M., Dalton, G. M., Maddox, S., et al. 2001, MNRAS, 328, 1039
Croton, D. J., Farrar, G. R., Norberg, P., et al. 2005, MNRAS, 356, 1155
D’Abrusco, R., Fabbian, G., Djorgovski, G., et al. 2012, ApJ, 755, 92
