

Disordered Heterogeneous Universe: Galaxy Distribution and Clustering across Length Scales

Oliver H. E. Philcox^{1,2,3,4,*} and Salvatore Torquato^{2,5,†}¹*Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08540, USA*²*School of Natural Sciences, Institute for Advanced Study, 1 Einstein Drive, Princeton, New Jersey 08540, USA*³*Center for Theoretical Physics, Department of Physics, Columbia University, New York, New York 10027, USA*⁴*Simons Foundation, New York, New York 10010, USA*⁵*Department of Chemistry, Department of Physics, Princeton Materials Institute, and Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey 08540, USA* (Received 16 July 2022; revised 2 January 2023; accepted 10 January 2023; published 14 March 2023)

The studies of disordered heterogeneous media and galaxy cosmology share a common goal: analyzing the disordered distribution of particles and/or building blocks at microscales to predict physical properties of the medium at macroscales, whether it be a liquid, colloidal suspension, composite material, galaxy cluster, or entire Universe. The theory of disordered heterogeneous media provides an array of theoretical and computational techniques to characterize a wide class of complex material microstructures. In this work, we apply them to describe the disordered distributions of galaxies obtained from recent suites of dark matter simulations. We focus on the determination of lower-order correlation functions, void and particle nearest-neighbor functions, certain cluster statistics, pair-connectedness functions, percolation properties, and a scalar order metric to quantify the degree of order. Compared to analogous homogeneous Poisson and typical disordered systems, the cosmological simulations exhibit enhanced large-scale clustering and longer tails in the void and particle nearest-neighbor functions, due to the presence of quasi-long-range correlations imprinted by early Universe physics, with a minimum particle separation far below the mean nearest-neighbor distance. On large scales, the system appears hyperuniform, as a result of primordial density fluctuations, while on the smallest scales, the system becomes almost antihyperuniform, as evidenced by its number variance. Additionally, via a finite-scaling analysis, we compute the percolation threshold of the galaxy catalogs, finding this to be significantly lower than for Poisson realizations (at reduced density $\eta_c = 0.25$ in our fiducial analysis compared to $\eta_c = 0.34$), with strong dependence on the mean density; this is consistent with the observation that the galaxy distribution contains voids of up to 50% larger radius. However, the two sets of simulations appear to share the same fractal dimension on scales much larger than the average intergalaxy separation, implying that they lie in the same universality class. We also show that the distribution of galaxies is a highly correlated disordered system (relative to the uncorrelated Poisson distribution), as measured by the τ order metric. Finally, we consider the ability of large-scale clustering statistics to constrain cosmological parameters, such as the Universe's expansion rate, using simulation-based inference. Both the nearest-neighbor distribution and pair-connectedness function (which includes contributions from correlation functions of all order) are found to considerably tighten bounds on the amplitude of quantum-mechanical fluctuations from inflation at a level equivalent to observing 25 times more galaxies. The pair-connectedness function in particular provides a useful alternative to the standard three-particle correlation, since it contains similar large-scale information to the three-point function, can be computed highly efficiently, and can be straightforwardly extended to small scales (though likely requires simulation-based modeling). This work provides the first application of such techniques to cosmology, providing both a novel system to test heterogeneous media descriptors and a tranche of new tools for cosmological analyses. A range of extensions are possible, including implementation on observational data; this will require further study on various observational effects, necessitating high-resolution simulations.

DOI: [10.1103/PhysRevX.13.011038](https://doi.org/10.1103/PhysRevX.13.011038)

Subject Areas: Astrophysics, Complex Systems, Statistical Physics

*ohep2@cantab.ac.uk

†torquato@electron.princeton.edu

I. INTRODUCTION

From condensed phases of matter to ecological systems to the primordial distribution of matter in the Universe, nature abounds with examples of disordered arrangements of interacting entities that form structures with diverse geometries and topologies. To understand the collective behavior of such phenomena, it is vital to have a mathematical formalism that enables a stochastic description of the constituent objects, particularly with regards to their spatial distribution and clustering, whether they be carbon atoms, concrete conglomerates, or individual galaxies. The theory of disordered heterogeneous media [1,2], which includes techniques from statistical mechanics [3], provides a natural and powerful machinery with which to equip ourselves in this venture. In particular, its primary objective is to connect the properties of the interacting constituents to their large-scale attributes, such as a material’s bulk transport and mechanical and electromagnetic properties. This is rigorously done by generally relating the bulk properties to an infinite set of diverse types of statistical correlation functions that characterize the microstructures [1], including those that contain topological information, such as phase connectivity and percolation characteristics. While this theoretical machinery has been primarily applied to Earth-bound materials, their applicability is far from terrestrial: these techniques work similarly for any phenomenon that can be treated as complex disordered heterogeneous media [1], including the spatial distribution and clustering of galaxies.

While the bulk of cosmological research in the past two decades has focused on analysis of the cosmic microwave background (the radiation signature of physics in the first $\sim 300\,000$ years, which provides a snapshot of the early Universe), that concerning the distribution of galaxies using statistical descriptors has become progressively more important [4–6], particularly with the advent of large-scale surveys, such as the forthcoming Dark Energy Spectroscopic Instrument (DESI) [7] and the Euclid satellite [8]. The distribution of galaxies traces the distribution of matter in the early Universe (see, e.g., Ref. [4]); as such, it encodes information on a wealth of cosmological parameters, such as the density of matter. An open question is how best to analyze the data: most works focus on measuring the *correlation functions* of the galaxy distribution and comparing them to physical models (either explicitly derived or numerically simulated) (see, e.g., Ref. [9]), though this is known to be suboptimal in terms of information content. While a number of alternative statistics have been proposed (including void statistics [10,11], marked density fields [12–15], Gaussianized fields [16–19], reconstructed density fields [20] field-level inference [21–23], Minkowski functionals and other topological descriptors [24–33], and beyond), there is little consensus on which have practical utility (with most having been applied only to the dark matter distribution), and few are

natural from a theoretical standpoint. An important insight is that the galaxy distribution is simply a set of irregularly arranged pointlike particles in three dimensions; this is mathematically identical to the structure of many terrestrial materials, including atomic systems, colloids, and sphere packing. As such, both scenarios can be treated with the same mathematical formalisms; i.e., heterogeneous media and statistical mechanical techniques designed to quantify the clustering of particles in materials can be used to provide a practical and well-motivated manner in which to understand the galaxy distribution.

This work considers the application of a number of statistical descriptors from the theory of disordered heterogeneous media to characterize structurally the distribution of galaxies, which we treat as discrete point configuration. We ask two main questions. (1) What can we learn about cosmology through the lens of disordered heterogeneous media and statistical mechanics? (2) What condensed matter physics lessons, more broadly, can we learn from cosmological structures? As a proof of concept, we will consider a number of descriptors [1,4], including the two- and three-particle correlation functions, “void” and “particle” nearest-neighbor functions, certain cluster statistics, pair-connectedness functions, percolation properties, and scalar order metrics to quantify the degree of order. We show how their behaviors in the cosmic landscape, probed through cosmological dark matter simulations, differs substantially from that expected from a simple Poissonian distribution of points as well as well-known homogeneous models of correlated disordered point patterns, showing that cosmology challenges general expectations of standard heterogeneous media and statistical mechanical models. In particular, we will find stronger clustering on large scales, giving an enhancement in the pair correlation and pair-connectedness functions and an excess of large-scale voids; these effects arise due to early Universe physics, which source quasi-long-range correlations in the galaxy distribution and create a hyperuniform system. Particular interest will be paid to the question of clustering and phase “percolation”; this is a well-understood phenomenon for many models in condensed matter but can be similarly extended to galaxy distributions, and yields interesting results. The cosmological case will be found to percolate faster, but asymptote to the Poisson case if the density is low, and both scenarios share the same set of critical exponents. Furthermore, we will consider the utility of descriptors from the theory of disordered heterogeneous media in cosmological settings, quantifying how they can add additional information regarding the early and late Universe, finding that the pair-connectedness function adds significant cosmological information at minimal additional computational cost. We caution that further work will be required before the statistics can be applied to observational data: this must include discussion of redshift-space effects (arising from the conversion of galaxy redshifts to distance, creating

anisotropy with respect to the sample line of sight) and the dependence of descriptors on galaxy properties, such as luminosities or masses.

Our conclusions to the above questions asked will be the following: (1) the theory of heterogeneous media and statistical mechanics provides an array of useful tools that can enhance the utility of galaxy survey datasets, strengthening the constraints on cosmological parameters and probing novel features of the distribution, (2) due to its quasi-long-ranged correlations, galaxy samples exhibit very different behavior to most terrestrial media, and thus provide an important sandbox for applying and understanding condensed matter techniques. While we restrict ourselves to galaxy surveys in this work, they are by no means the only cosmological application of such statistics: a number of other phenomena could be described by such approaches. These include two-phase media such as the distribution of cosmic voids (empty regions of gargantuan extent) and the statistics of ionized hydrogen bubbles during the “reionization” phase of the Universe. Such areas provide a bountiful mine from which to derive future work.

The remainder of this work is structured as follows. In Sec. II, we provide an overview of the statistics used in this work, before presenting the proposed testing ground (simulated galaxy samples) in Sec. III. A comparison of the statistics on galactic and Poisson data is shown in Sec. II, with Sec. V providing a discussion of percolation physics in the two systems. Finally, Sec. VI considers the utility of a specific statistic, the pair-connectedness function, in cosmological contexts, before we conclude in Sec. VII.

II. STATISTICAL DESCRIPTORS OF POINT CONFIGURATIONS

In this section, we define the various statistical descriptors of point configurations that will be used in the remainder of this work. We principally adopt notation from the statistical mechanics community (particularly following Ref. [1]), though we connect this to the cosmological terminology, when relevant. Although we principally work in \mathbb{R}^3 , most of the following discussion remains relevant in other metric spaces. A schematic illustrating the various statistical descriptors considered in this work is shown in Fig. 1.

A. Correlation functions

The fundamental quantity describing a discrete set of N points in some large region in \mathbb{R}^d of volume V is the N -particle probability density functions P_N , which is defined such that $P_N(\mathbf{r}_1, \dots, \mathbf{r}_N) d\mathbf{r}_1 \dots d\mathbf{r}_N$ is the probability of finding the first particle within $d\mathbf{r}_1$ of \mathbf{r}_1 , the second within $d\mathbf{r}_2$ of \mathbf{r}_2 , etc. [1,4]. Of more practical use is the n -particle probability density function, marginalized over

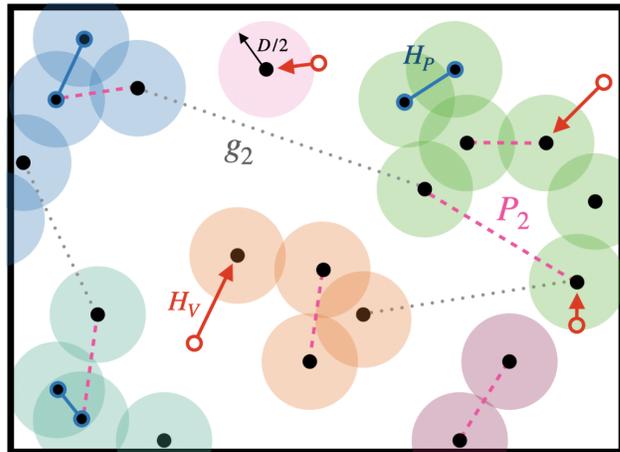


FIG. 1. Schematic depicting the various statistical descriptors used in this work. The black points indicate the positions of random particles (here visualized in \mathbb{R}^2), with the groups of colored circles (each of diameter D) demonstrating the clusters. The pair correlation function g_2 counts pairs of particles belonging both to the same cluster (pink lines) and to different clusters (gray lines), whereas the pair-connectedness function P_2 contains only particles within the same cluster (pink lines). We show also the nearest-neighbor functions: H_V encodes the distance between a randomly positioned point and the nearest galaxy (red arrows), while H_P gives the separation between a galaxy and its closest neighbor (blue lines). Further details on these statistics are given in Sec. II.

the positions of the other $(N - n) \geq 0$ particles: this is defined as

$$\rho_n(\mathbf{r}_1, \dots, \mathbf{r}_n) = \frac{N!}{(N-n)!} \int d\mathbf{r}_{n+1} \dots d\mathbf{r}_N P_N(\mathbf{r}_1, \dots, \mathbf{r}_n, \mathbf{r}_{n+1}, \dots, \mathbf{r}_N), \quad (1)$$

with $\rho_n(\mathbf{r}_1, \dots, \mathbf{r}_n) d\mathbf{r}_1 \dots d\mathbf{r}_n$ being proportional to the probability of finding one indistinguishable particles within $d\mathbf{r}_1$ of \mathbf{r}_1 , $d\mathbf{r}_2$ of \mathbf{r}_2 , etc.

For a statistically homogeneous medium, the one-particle density function is a constant, i.e., $\rho_1(\mathbf{r}) \equiv \bar{\rho}$, which is the mean *number density* (number of points per unit volume in the infinite-volume or thermodynamic limit), commonly labeled \bar{n} in cosmological contexts (with $\bar{\rho}$ often used to refer to the mean energy density of the Universe). More generally, for statistically homogeneous systems, $\rho_n(\mathbf{r}_1, \dots, \mathbf{r}_n)$ is translationally invariant, enabling us to reexpress it as follows:

$$\rho_n(\mathbf{r}_1, \dots, \mathbf{r}_n) = \bar{\rho}^n g_n(\mathbf{r}_{12}, \dots, \mathbf{r}_{1n}), \quad (2)$$

where $g_n(\mathbf{r}_{12}, \dots, \mathbf{r}_{1n})$ is the n -particle *correlation function* (closely related to the cosmologists’ n -point correlation function), which depends on the relative positions \mathbf{r}_{12}, \dots , where $\mathbf{r}_{ij} \equiv \mathbf{r}_j - \mathbf{r}_i$. The two-particle or pair correlation

function g_2 is particularly important in applications, and is schematically illustrated in Fig. 1. For translationally invariant point configurations without *long-range order*, $g_n(\mathbf{r}_{12}, \dots, \mathbf{r}_{1n}) \rightarrow 1$ when the points (or particles) are mutually far from one another, i.e., as $|\mathbf{r}_{ij}| \rightarrow \infty$ ($1 \leq i < j < n$), $\rho_n(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) \rightarrow \rho^n$. Thus, the deviation of g_n from unity provides a measure of the degree of spatial correlations (positive or negative) between the particles. Note that for a translationally invariant *Poisson* (spatially uncorrelated) point configurations, $g_n = 1$ is unity for all values of its argument. If the point configuration is in addition statistically isotropic, the functions g_n are invariant under joint rotations of \mathbf{r}_{ij} , such that g_2 is a function only of $|\mathbf{r}_{12}| \equiv r_{12}$, and g_3 depends only on r_{12} , r_{13} and r_{23} [34].

In cosmology it is commonplace to consider not the probability density function of the full ensemble, but a set of realizations of the microscopic density (often known as the density field), each denoted by $\hat{\rho}(\mathbf{r})$. This gives the probability that there is a particle within $d\mathbf{r}$ of \mathbf{r} for a specific field (for sufficiently small $d\mathbf{r}$), and can be represented as a sum of N Dirac deltas: $\hat{\rho}(\mathbf{r}) = \sum_{i=1}^N \delta_D(\mathbf{r} - \mathbf{x}_i)$. Averaging over realizations (denoted by the expectation operator \mathbb{E}), we can relate $\hat{\rho}(\mathbf{r})$ to the n -particle probability density functions:

$$\mathbb{E}[\hat{\rho}(\mathbf{r}_1) \dots \hat{\rho}(\mathbf{r}_n)] \equiv \rho_n(\mathbf{r}_1, \dots, \mathbf{r}_n). \quad (3)$$

If the field is statistically homogeneous, and the volume (i.e., $\int d\mathbf{r}$) sufficiently large, this is equivalent to a spatial average via the ergodic principle. In cosmological contexts, Eq. (3) is usually adopted, with $\hat{\rho}(\mathbf{r})$ often referred to as $\bar{\rho}[1 + \delta(\mathbf{r})]$ for *overdensity* field δ . Additionally, it is conventional to work with *disconnected* correlation functions $\xi^{(n)}$, often known as n -point correlation functions: the first few satisfy

$$\begin{aligned} \xi^{(2)}(\mathbf{r}_{12}) &\equiv \mathbb{E}[\delta(\mathbf{r}_1)\delta(\mathbf{r}_2)] \equiv g_2(\mathbf{r}_{12}) - 1, \\ \xi^{(3)}(\mathbf{r}_{12}, \mathbf{r}_{13}, \mathbf{r}_{23}) &\equiv \mathbb{E}[\delta(\mathbf{r}_1)\delta(\mathbf{r}_2)\delta(\mathbf{r}_3)] \\ &\equiv g_3(\mathbf{r}_{12}, \mathbf{r}_{13}, \mathbf{r}_{23}) - g_2(\mathbf{r}_{12}) - g_2(\mathbf{r}_{13}) \\ &\quad - g_2(\mathbf{r}_{23}) + 2, \end{aligned} \quad (4)$$

see Ref. [4], and are all zero under Poisson statistics. We will principally work with the full g_n functions in this work, adopting statistical mechanics conventions.

A particularly important descriptor is the *structure factor* $\mathcal{S}(\mathbf{k})$, which is related to the Fourier transform of the total correlation function $h(r) \equiv g_2(r) - 1$:

$$\mathcal{S}(\mathbf{k}) \equiv 1 + \bar{\rho} \tilde{h}(\mathbf{k}) \equiv 1 + \bar{\rho} \int d\mathbf{r} e^{i\mathbf{k}\cdot\mathbf{r}} [h(\mathbf{r})]. \quad (5)$$

For a Poisson point distribution, $\mathcal{S}(\mathbf{k}) = 1$ for all \mathbf{k} . The structure factor and the cosmologists' *power spectrum* [4] $P(\mathbf{k})$ are trivially related to one another via $\mathcal{S}(\mathbf{k}) = \bar{\rho} P(\mathbf{k})$.

The structure factor provides a useful way to quantify large-scale (low wave number $k \equiv |\mathbf{k}|$) correlation and fluctuation properties of a point configuration and plays a central role in the hyperuniformity concept. Hyperuniform states of matter are correlated systems that are characterized by an anomalous suppression of long-wavelength (i.e., large length scale) density fluctuations compared to those found in garden-variety disordered systems, such as ordinary fluids and amorphous solids [35,36]. A hyperuniform (or superhomogeneous [37]) many-particle system in d -dimensional Euclidean space \mathbb{R}^d is one in which (normalized) density fluctuations are completely suppressed at very large length scales, implying that the structure factor $\mathcal{S}(\mathbf{k})$ tends to zero in the infinite-wavelength limit, i.e.,

$$\lim_{|\mathbf{k}| \rightarrow 0} \mathcal{S}(\mathbf{k}) = 0. \quad (6)$$

Equivalently, a hyperuniform system is one in which the number variance $\sigma_N^2(R) \equiv \langle N(R)^2 \rangle - \langle N(R) \rangle^2$ of particles within a spherical observation window of radius R grows more slowly than the window volume in the large- R limit, i.e., slower than R^d . Typical disordered systems, such as liquids and structural glasses, have the standard asymptotic volume scaling $\sigma_N^2(R) \sim R^d$ and hence are not hyperuniform. For general translationally invariant point configuration in \mathbb{R}^d , the local number variance $\sigma_N^2(R)$ is determined exactly by the pair statistics [35]:

$$\begin{aligned} \sigma_N^2(R) &= \bar{\rho} v_1(R) \left[1 + \rho \int_{\mathbb{R}^d} h(\mathbf{r}) \alpha(r; R) d\mathbf{r} \right] \\ &= \bar{\rho} v_1(R) \left[\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \mathcal{S}(\mathbf{k}) \tilde{\alpha}(k; R) d\mathbf{k} \right], \end{aligned} \quad (7)$$

where $v_1(R) = \pi^{d/2} R^d / \Gamma(1 + d/2)$ is the volume of a d -dimensional sphere of radius R , and $\alpha(r; R)$ is the *scaled intersection volume*, the ratio of the intersection volume of two spherical windows of radius R whose centers are separated by a distance r to the volume of a spherical window, known analytically in any space dimension [1,38]. Its Fourier transform is the non-negative function given by

$$\tilde{\alpha}(k; R) = 2^d \pi^{d/2} \Gamma(1 + d/2) \frac{[J_{d/2}(kR)]^2}{k^d}, \quad (8)$$

where $J_\nu(x)$ is the Bessel function of order ν .

Consider translationally invariant point configurations that are characterized by a structure factor with a radial power-law form in the vicinity of the origin, i.e.,

$$\mathcal{S}(\mathbf{k}) \sim |\mathbf{k}|^\alpha \quad \text{for } |\mathbf{k}| \rightarrow 0. \quad (9)$$

For hyperuniform systems, the exponent α is positive ($\alpha > 0$) and its value determines three hyperuniformity

classes corresponding to different large- R scaling behaviors of the number variance [35,36,39]:

$$\sigma^2(R) \sim \begin{cases} R^{d-1} & \alpha > 1 \text{ (class I)} \\ R^{d-1} \ln R & \alpha = 1 \text{ (class II)} \\ R^{d-\alpha} & \alpha < 1 \text{ (class III)}. \end{cases} \quad (10)$$

Classes I and III describe the strongest and weakest forms of hyperuniformity, respectively. States of matter that belong to class I include all perfect crystals [35,39], many perfect quasicrystals [39–41], “randomly” perturbed crystal structures [42–45], classical disordered ground states of matter [35,46,47], as well as systems out of equilibrium [48,49]. Class II hyperuniform systems include some quasicrystals [41], the positions of the prime numbers [50], and many disordered classical [48,51–54] and quantum [55–57] states of matter. Examples of class III hyperuniform systems include classical disordered ground states [58], random organization models [59] and perfect glasses [48]. Certain disordered hyperuniform systems are poised at an “inverted” critical point in which the volume integral of the total correlation function $h(\mathbf{r})$ is quasi-long-ranged but its volume integral is bounded [35,36].

By contrast, for any nonhyperuniform system, the local variance has the following large- R scaling behaviors [60]:

$$\sigma^2(R) \sim \begin{cases} R^d & \alpha = 0 \text{ (typical nonhyperuniform)} \\ R^{d-\alpha} & -d < \alpha < 0 \text{ (antihyperuniform)}. \end{cases} \quad (11)$$

For a typical nonhyperuniform system, the structure factor $\mathcal{S}(0)$ is bounded [36]. In antihyperuniform systems, $\mathcal{S}(0)$ is unbounded, i.e.,

$$\lim_{|\mathbf{k}| \rightarrow 0} \mathcal{S}(\mathbf{k}) = +\infty, \quad (12)$$

and hence are diametrically opposite to hyperuniform systems. Antihyperuniform systems include fractals, systems at thermal critical points (e.g., liquid-vapor and magnetic critical points) [61–65], as well as certain substitution tilings [66].

B. Order metric

Given the richness of the spectrum of possible microstructures that can arise in condensed phase systems, an outstanding challenging task has been the quantification of their degree order or disorder. Scalar order or disorder metrics have been profitably employed to quantify the degree of order in many-particle systems, including sphere packings; see Refs. [1,67] and references therein. Any

scalar order metric $\Psi(\mathbf{R})$ is a well-defined non-negative scalar function of a many-particle configuration \mathbf{R} and if, for any two configurations \mathbf{R}_A and \mathbf{R}_B , $\Psi(\mathbf{R}_A) > \Psi(\mathbf{R}_B)$, we say that configuration \mathbf{R}_A is to be considered more ordered than configuration \mathbf{R}_B . It has been suggested that a good scalar order metric should have the following additional properties [68]: (1) sensitivity to any type of ordering without bias toward any reference system, (2) ability to reflect the hierarchy of ordering between prototypical systems given by common physical intuition (e.g., perfect crystals with high symmetry should be highly ordered, followed by quasicrystals, correlated disordered packings without long-range order, and finally spatially uncorrelated or Poisson distributed particles), (3) capacity to detect order at any length scale, and (4) incorporation of both the variety of local coordination patterns and the spatial distribution of such patterns.

The recently introduced τ order metric [47] fulfills these requirements and has been fruitfully employed to characterize the degree of order across length scales of a diverse set of disordered media [47,69–71]. This metric, which we compute here for the first time for the galaxies, is defined as

$$\begin{aligned} \tau &= \frac{1}{D^d} \int_{\mathbb{R}^d} d\mathbf{r} h^2(\mathbf{r}) \\ &= \frac{1}{(2\pi)^d D^d \bar{\rho}^2} \int_{\mathbb{R}^d} d\mathbf{k} [\mathcal{S}(\mathbf{k}) - 1]^2, \end{aligned} \quad (13)$$

where D is a characteristic “microscopic” length scale. This scalar metric measures deviations of two-particle statistics from that of the Poisson distribution. Since both positive and negative correlations contribute to the integral, due to the fact that $h(\mathbf{r})$ is squared, τ measures the degree of translational order across length scales. It clearly vanishes for the uncorrelated Poisson distribution, diverges for an infinite crystal, and is a positive bounded number for correlated disordered systems without long-range order (i.e., Bragg peaks). It is interesting to note that the τ order metric is closely related to the negative of the excess two-particle entropy of the system [72].

C. Nearest-neighbor functions

Another well-known set of statistical descriptors that arise in rigorous bounds on the macroscopic physical properties of disordered heterogeneous media, such as suspensions of spheres, and employed in the statistical mechanics of many-particle systems are nearest-neighbor functions [1,57,73]. There two types of such functions: void and particle quantities. The void and particle nearest-neighbor probability density functions $H_V(r)$ and $H_P(r)$, respectively, are defined as follows:

$$H_V(r)dr = \text{probability that a point of the point configuration lies at a distance between } r \text{ and } r + dr \text{ from an arbitrary point in the space,} \quad (14)$$

$$H_P(r)dr = \text{probability that a point of the point configuration lies at a distance between } r \text{ and } r + dr \text{ from an other point of the point configuration.} \quad (15)$$

The associated dimensionless “exclusion” probabilities $E_V(r)$ and $E_P(r)$ are defined as follows:

$$E_V(r) = \text{probability of finding a spherical cavity of radius } r \text{ empty of any points in the point configuration.} \quad (16)$$

$$E_P(r) = \text{probability of finding a spherical cavity of radius } r \text{ centered at an arbitrary point of the point configuration empty of any other points.} \quad (17)$$

It follows that the exclusion probabilities are complementary cumulative distribution functions associated with the density functions and thus are related to the latter via

$$E_V(r) = 1 - \int_0^r H_V(x)dx, \quad H_V(r) = -\partial_r E_V(r), \quad (18)$$

and

$$E_P(r) = 1 - \int_0^r H_P(x)dx, \quad H_P(r) = -\partial_r E_P(r). \quad (19)$$

The moments of $H_V(r)$ and $H_P(r)$, defined by

$$\ell_V^{(k)} = \int_0^\infty r^k H_V(r)dr, \quad (20)$$

$$\ell_P^{(k)} = \int_0^\infty r^k H_P(r)dr \quad (21)$$

are particularly useful integral nearest-neighbor measures, with the $k = 1$ version of the latter representing the mean nearest-neighbor distance between particles. The void nearest-neighbor function H_V has received some attention in cosmology, both historically [74–76] and in recent works, in particular via the “ k NN” statistics [77–80], generalizing the above to the k th nearest neighbor. This has been shown to yield strong constraints on cosmological parameters (cf. Sec. VI) and can be modeled semianalytically. It is noteworthy that k NN statistics and related quantities have been studied and fruitfully applied in the field of statistical mechanics [81–84].

Both the void and particle nearest-neighbor functions generally involve integrals over all the n -particle correlation functions, $\{g_n\}$ ($n = 2, 3, 4, \dots$) [73,77]. While the void and particle nearest-neighbor functions are identical to one another for a Poisson point configuration [e.g., $E_V(r) = E_P(r) = \exp(-4\pi r^3 \bar{\rho}/3)$ in three dimensions],

they are generally different from one another for correlated systems, as manifested by their different series representations [73]. Both the void and particle quantities arise in rigorous bounds on the effective transport and mechanical properties of heterogeneous media [1]. It is noteworthy that the void nearest-neighbor functions play a deep role in the covering problem of discrete geometry [85]. The covering problem asks for the point configuration that minimizes the radius of overlapping spheres circumscribed around each of the points required to cover d -dimensional Euclidean space \mathbb{R}^d [86]. The above void statistic also bears some similarities to the “void size function” used in cosmology [10,87]. The latter quantity is usually constructed from smoothed density fields, with voids identified as spherical regions with a mean density below some critical threshold, usually 30% of the system’s mean density. This differs from H_V in two key ways: (a) the voids defined by Eq. (14) contain no particles, thus they are equivalent to requiring a critical density of zero, (b) the cosmologists’ void contains no subvoids: i.e., any empty area of space within a void cannot be classified as a smaller void, unlike for H_V .

D. Clustering and connectedness functions

To quantify the geometrical and topological properties of the class of disordered heterogeneous media consisting of particles distributed throughout a matrix phase, it is often useful to statistically characterize *particle clusters* that are defined according to some connectivity criterion [1,88–92]. For point (zero-dimensional) particles, this can be achieved by circumscribing each point by spheres of diameter D , which generally may overlap with one another. Such a decoration of the points by possibly overlapping spheres divides the space into two disjoint regions or “phases,” encompassing points that do and do not lie within a distance $D/2$ of at least one point. Two spheres are deemed to be connected if they overlap. Defining the *reduced density* $\eta \equiv \bar{\rho}\pi D^3/6$ (in \mathbb{R}^3), it is clear that as the diameter

D at fixed mean density $\bar{\rho}$ increases from zero, η and the fraction of space occupied by the spheres will increase and clusters of various sizes will form and grow [1].

Once clusters have been identified, one can determine the pair-connectedness function $P_2(r, \eta)$, where $P_2(r, \eta) \times 4\pi r^2 dr$ is the conditional probability of finding a particle in a shell of radius dr at radial distance r from another particle in the same cluster (assuming statistical homogeneity and isotropy). Equivalently, this quantity gives the probability that there exists a path from the first to the second point that never leaves the particle phase, i.e., one that is always within a distance $D/2$ of at least one particle (cf. Fig. 1) [93]. This is the *connected* contribution to the full pair correlation function $g_2(r)$:

$$g_2(r) \equiv P_2(r, \eta) + B_2(r, \eta), \quad (22)$$

where the pair-blocking function $B_2(r, \eta)$ gives the correlation between pairs of particles which do not lie in the same cluster. For $r < D(\eta)$, any pair of points must be within the same cluster; thus $P_2(r < D, \eta) = g_2(r)$.

A related quantity is the direct-connectedness function $C_2(r, \eta)$ (also known as the non-nodal correlation function) [88]. This is the probability that two points separated by a distance r are connected by a path through the set of random particles that does not involve nodes (i.e., one that cannot be broken by a single cut, as in Fig. 1). A general path between two points contains either zero or at least one node: this permits the Ornstein-Zernike (OZ) decomposition [88,89],

$$P_2(\mathbf{r}_{12}, \eta) = C_2(\mathbf{r}_{12}, \eta) + \bar{\rho} \int d\mathbf{r}_3 C_2(\mathbf{r}_{13}, \eta) P_2(\mathbf{r}_{32}, \eta), \quad (23)$$

labeling $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$. The first quantity on the rhs contains paths with no nodes, thus it involves the first factor of C_2 , while for the second, we integrate over the position of the node closest to \mathbf{r}_1 (assuming a statistically homogeneous field $\bar{\rho}$), noting that the path from \mathbf{r}_1 to \mathbf{r}_3 contains no nodes by definition, yielding another function of C_2 . Finally, the path from \mathbf{r}_3 to \mathbf{r}_2 can contain nodes, leading to the final factor of P_2 . In Fourier space this gives a simple relation between the pair-connectedness and direct-connectedness functions:

$$\tilde{C}_2(k, \eta) = \frac{\tilde{P}_2(k, \eta)}{1 + \bar{\rho} \tilde{P}_2(k, \eta)}, \quad \tilde{P}_2(k, \eta) = \frac{\tilde{C}_2(k, \eta)}{1 - \bar{\rho} \tilde{C}_2(k, \eta)}. \quad (24)$$

To gain intuition for the pair-connectedness function (and related statistics), it is instructive to consider its form for a Poissonian system (noting that there is no Gaussian limit, given that we are dealing with discrete systems). At low densities, it can be computed as a perturbation series in $\bar{\rho}$ (or, more strictly, in η/η_c), first considering pairs of particles that are directly linked by the covered phase [i.e., their centers lie within $D(\eta)$], then moving to pairs linked via a third particle and so on. This leads to the decomposition

$$P_2^{\text{Poiss}}(\mathbf{r}_{12}, \eta) = \Theta_D(\mathbf{r}_{12}) + \bar{\rho} [1 - \Theta_D(\mathbf{r}_{12})] \int d\mathbf{r}_3 \Theta_D(\mathbf{r}_{13}) \Theta_D(\mathbf{r}_{32}) + \bar{\rho}^2 [1 - \Theta_D(\mathbf{r}_{12})] \int d\mathbf{r}_3 d\mathbf{r}_4 \Theta_D(\mathbf{r}_{13}) \Theta_D(\mathbf{r}_{34}) \Theta_D(\mathbf{r}_{42}) [1 - \Theta_D(\mathbf{r}_{14})] [1 - \Theta_D(\mathbf{r}_{23})] + \dots, \quad (25)$$

where the Heaviside function $\Theta_D(\mathbf{r}) \equiv \Theta_H(r - D)$ selects pairs with separations below D . In this expansion, successive terms integrate over progressively more particle positions with, for example, the second term averaging over the position of \mathbf{r}_3 , which must be within a distance of D from both \mathbf{r}_1 and \mathbf{r}_2 . As such, this expression is difficult to compute beyond second order (which is convolutional) and thus rarely used in practice, unless $\eta \ll \eta_c$ and we restrict to small scales. In practice, approximate treatments

are usually adopted, such as via the OZ equation (23) combined with heuristic ‘‘closure’’ relations such as the Percus-Yevick form [94]. These give accurate predictions for $P_2(r, \eta)$ in low-density regimes at relatively small r . This stands in contrast to the case familiar from cosmology, when the modeling of $g_2(r)$ becomes progressively more accurate as r increases.

For a general system, a similar decomposition to Eq. (25) is possible, and takes the form

$$P_2(\mathbf{r}_{12}, \eta) = g_2(\mathbf{r}_{12}) \Theta_D(\mathbf{r}_{12}) + \bar{\rho} [1 - \Theta_D(\mathbf{r}_{12})] \int d\mathbf{r}_3 g_3(\mathbf{r}_{13}, \mathbf{r}_{32}) \Theta_D(\mathbf{r}_{13}) \Theta_D(\mathbf{r}_{32}) + \bar{\rho}^2 [1 - \Theta_D(\mathbf{r}_{12})] \int d\mathbf{r}_3 d\mathbf{r}_4 g_4(\mathbf{r}_{13}, \mathbf{r}_{34}, \mathbf{r}_{42}) \Theta_D(\mathbf{r}_{13}) \Theta_D(\mathbf{r}_{34}) \Theta_D(\mathbf{r}_{42}) [\Theta_D(\mathbf{r}_{14})] [1 - \Theta_D(\mathbf{r}_{23})] + \dots \quad (26)$$

In this case, the expansion depends on the correlation functions g_n , since there exists background correlations in addition to that induced by the circumscribed spheres around points. As expected, this implies that $P_2(r < D, \eta) = g_2(r)$, with the n th-order term involving correlators of the form g_{n+2} . Formally, this expression may be extended to all orders, via the relation

$$\bar{\rho}^2 P_2(r, \eta) = \prod_{k=2}^{\infty} \bar{\rho}^k \int \left(\prod_{i=1}^k dr_i \right) \delta_D(|\mathbf{r}_1 - \mathbf{r}_k| - r) g_k(\mathbf{r}_1, \dots, \mathbf{r}_k) \prod_{i=1}^{k-1} \Theta_D(\mathbf{r}_{i(i+1)}) \prod_{m=1}^{k-2} \left[\prod_{j=m+2}^k [1 - \Theta_D(\mathbf{r}_{mj})] \right]. \quad (27)$$

While this form is not particularly useful for analytic treatments, due to the difficulty inherent in performing the high-dimensional integrals present for $k > 3$ it illustrates how the pair-connectedness function is composed of all possible correlation functions, and thereby partly resums information found at all orders. At very low densities, $P_2(r)$ tends to $\Theta_D(r)g_2(r)$, thus we do not expect this statistic to add information; however, as η increases, the fraction of information contributed by the higher-order g_n increases, until the system becomes nonperturbative at $\eta \approx \eta_c$ (whence the notion of connectedness breaks down). One may ask whether an OZ-like equation with some closure relation can be used to provide an approximate analytic form for $P_2(r, \eta)$ in the general case. Unfortunately, this is far from trivial, since any scheme only involving g_2 will miss any contributions to P_2 from g_3 and above, which are of particular cosmological interest, especially when performing parameter inference in conjunction with g_2 (as is the case below). We leave further treatment of this problem to future work.

E. Continuum percolation

Percolation describes the appearance of a *phase transition* in the system, which, in the above case, corresponds to the emergence of long-range connectivity in the point cloud due to arbitrarily large clusters of points, again defined by spheres of some diameter. This phenomenon is of relevance in a wide variety of physical settings such as the transport of fluid in porous systems, the appearance of fractures in geological formations, spread of diseases, and the collapse of gas into stars [1,2,90,92,95]. This crossover from nonpercolating clusters to the appearance of the incipient sample-spanning cluster (infinite in the thermodynamic limit) is characterized by a *critical reduced density* (also known as a percolation density), $\eta_c = \bar{\rho}\pi D_c/6$ (in \mathbb{R}^3), where D_c is the critical sphere diameter at fixed mean density $\bar{\rho}$ with the system said to have percolated for $\eta > \eta_c$. For a Poissonian system in \mathbb{R}^3 , numerical simulations find $\eta_c \approx 0.34$ [1], implying that the connected phase fills about 29% of the space. In addition to the connectedness functions described above, percolation theory utilizes a number of other statistical descriptors, which we outline below.

The *mean cluster size* $S(\eta)$ gives a simple manner in which to characterize clustering and percolation at some

reduced density η [88]. This quantity is simply the mean number of particles in a cluster containing a randomly chosen particle. It is directly related to the pair-connectedness function by the following relation:

$$S(\eta) = 1 + \bar{\rho} \int d\mathbf{r} P_2(\mathbf{r}, \eta) = [1 - \bar{\rho} \tilde{C}_2(0, \eta)]^{-1}, \quad (28)$$

where the second equality follows from Eq. (24). When $\eta > \eta_c$, clusters of infinite extent appear, thus $S(\eta) \rightarrow \infty$, and the volume integral of the pair-connectedness function diverges. Given a form for $C_2(r, \eta)$, the second relation provides a useful manner in which to estimate η_c , by solving $\bar{\rho} \tilde{C}_2(0, \eta_c) = 1$. Furthermore, the behavior close to the phase transition can be expressed in terms of *critical exponents* of the field: in particular, $S(\eta) \sim (\eta - \eta_c)^{-\gamma}$ and $C_2(r, \eta) \sim r^{-\alpha}$ at large r , for $\eta \rightarrow \eta_c^-$, where α and γ are found to be universal for a broad class of physical models [92,96].

The mean cluster size can also be written in terms of so-called s -mer cluster statistics for densities below the percolation threshold η_c [1,97]; namely,

$$S(\eta) = \frac{\sum_{s=1}^{\infty} s^2 n_s}{\sum_{s=1}^{\infty} s n_s}, \quad \eta < \eta_c. \quad (29)$$

Here n_s is the average number of s -mers, clusters containing s particles, per unit number of particles. This representation will be employed to estimate $S(\eta)$ from simulations.

For a finite (aperiodic) system of size L , percolation may be studied by considering the existence of *sample-spanning clusters*, i.e., clusters of connected points which reach from the top to the bottom of the system (in some dimension). In the $L \rightarrow \infty$ limit, these clusters will appear only for $\eta > \eta_c$: for finite systems, the behavior can be characterized using the *percolation probability*, $\Pi(\eta, L)$, which is the probability that a realization of size L will contain a sample-spanning cluster. This will be used to compute the percolation threshold η_c in Sec. V, via a finite-scaling analysis. If such a cluster exists, its mass [i.e., the number of constituent particles, denoted $M(L, \eta)$] can be used to ascertain the effective mass *fractal dimension* d_F of the field. In particular,

$$M(L, \eta) \sim L^{d_F}, \quad \eta \rightarrow \eta_c^-, \quad (30)$$

where $d_F \approx 2.42$ for Poisson systems, and any other processes in the same universality class (cf. Refs. [1,98,99]).

III. GALAXY SURVEYS AS POINT CLOUDS

Although the statistics described in Sec. II have been principally applied to study the properties of physical materials, their applicability extends far beyond the terrestrial regime. In this work, we consider their application to spectroscopic galaxy surveys, such as those of the upcoming Dark Energy Spectroscopic Instrument (DESI) and Euclid projects [7,8]. Such projects will measure the angular positions and redshifts of $\mathcal{O}(10^7)$ bright galaxies from ground- and space-based telescopes, providing a three-dimensional map of the Universe with unprecedented resolution. Fundamentally, galaxy surveys measure a set of N galaxy positions with some associated weights, representing experimental effects. In many typical analyses [100–102], these are assigned to some coarse grid in \mathbb{R}^3 , and the associated field taken to be an inhomogeneous Poisson sample of an underlying continuous field. This is itself modeled as a nonlinear transformation of the underlying *dark matter* distribution, whose correlation functions (particularly g_2 and g_3) encode early Universe physics, with the field obeying Gaussian statistics on sufficiently large scales, before a perturbative (and well-understood [103]) regime takes hold. Explicitly, the microscopic density $\hat{\rho}(\mathbf{x})$ satisfies

$$\hat{\rho}(\mathbf{x}) \sim \text{Poisson}(\bar{\rho}[1 + \delta_g(\mathbf{x})]), \quad \tilde{\delta}_g(\mathbf{k}) \sim \mathcal{N}(0, \tilde{h}_{2,g}(k)), \quad (31)$$

on sufficiently large scales, where δ_g is a normally distributed continuous background field with variance $\tilde{h}_{2,g}(k)$. On small, nonlinear, scales, a variety of galaxy formation processes become important and the above approach is known to be insufficient. This has led to a flurry of interest in additional statistics beyond the simple correlation functions.

An alternative to the standard approach is to consider the point cloud traced by the galaxies as the fundamental object, facilitating direct application of the clustering techniques described in Sec. II. Rather than working with observational data directly, this work will make use of simulated data drawn from the publicly available QUIJOTE suite [104], which is a collection of 40 000 realizations of the Universe, each contained within a cubic volume of size $L = 1000h^{-1}$ Mpc. (Following cosmologists convention, we work in h^{-1} Mpc units, where 1 Mpc $\equiv 10^6$ parsec and $h^{-1} \approx 1.4$ is used to remove a leading scaling.) In particular, we use dark matter simulations that have been evolved down to redshift zero (today), and contain a set of $\sim 10^5$ *dark matter halos*: spheroidal agglomerations of matter in which galaxies are known to form. (In this work, we use only halos containing at least 64 dark matter particles to

avoid discreteness effects; these have masses $M \gtrsim 5 \times 10^{12}h^{-1}M_\odot$ in our baseline simulations.)

Rather than dealing with the complexities of assigning galaxies to dark matter halos as a function of their mass (for example, using a halo occupation distribution [105]), we use the positions of the dark matter halos as a direct proxy for the galaxy positions, which is sufficient for this initial study. As such, we do not require the Poisson-Gaussian assumptions of Eq. (31), and will utilize the galaxy catalog only as a discrete point cloud. In most scenarios, we will use the galaxy catalogs extracted from 1000 QUIJOTE high-resolution simulations, each run with the same underlying physical model, but with varying realizations of the (stochastic) initial conditions. A section of a typical simulation is shown in Fig. 2. We caution that these simulations do not fully represent observational data, in particular, due to their limited mass resolution and lack of (magneto)hydrodynamic effects. However, their simplified nature makes them ideal for the proof-of-concept study considered herein, since it allows for a large number of simulations (and thus determination of accurate covariances). Further work will necessarily require application of the above tools to higher-resolution simulations, though these are fewer in number.

IV. PHENOMENOLOGICAL CLUSTERING STATISTICS

A. Pair correlation function

We begin by considering the pair correlation function $g_2(r)$ of the QUIJOTE simulation suite. This is estimated from the array of galaxy positions using the CORRFUNC code [106], which computes the statistic in a set of bins with centers $\{r_a\}$ via

$$\hat{g}_2(r_a) = \frac{1}{N\bar{\rho}v_a} \sum_{i,j=1}^N \begin{cases} 1 & \text{if } |\mathbf{r}_i - \mathbf{r}_j| \text{ in bin } r_a \\ 0 & \text{else,} \end{cases} \quad (32)$$

where v_a is the volume of bin a and \mathbf{r}_i is the 3D position of galaxy i (accounting for periodic wrapping). In Fig. 3(a), we display the obtained $g_2(r)$ functions, alongside corresponding results from a Poisson random sample with the same number density ($\approx 1 \times 10^{-4}h^3 \text{ Mpc}^{-4}$) and volume. As expected, the latter is simply unity everywhere, while the former shows considerable structure, and is quite different to that expected from most simple heterogeneous media [1]. On small scales (with $r \lesssim 0.2h^{-1}$ Mpc, considerably less than the average galaxy separation of $\approx 8h^{-1}$ Mpc), $g_2(r)$ decays to zero; this is as expected, since the galaxies are of finite size and cannot overlap, enforcing some minimum separation. (Physically, galaxies *can* overlap; however, they would be classed as a single object in this paradigm.) The fact the mean galaxy (particle) separation is about an order of magnitude greater than the minimum pair separation is atypical behavior for most

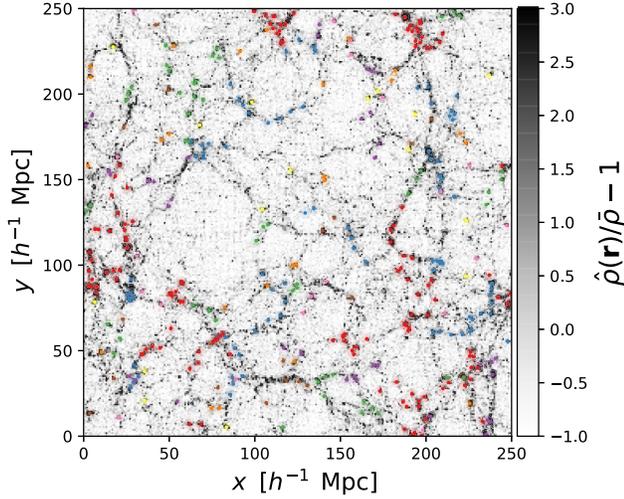
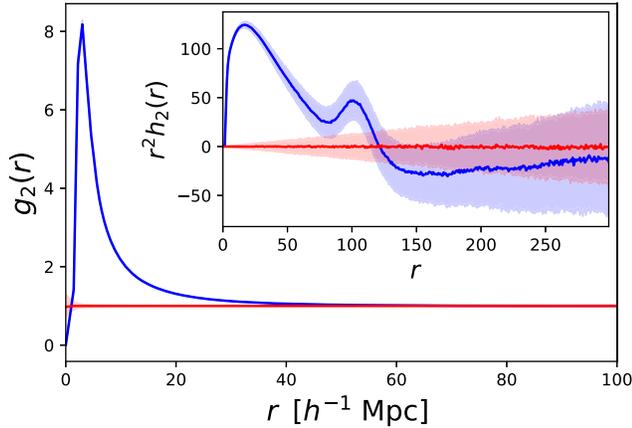


FIG. 2. Example of the cosmological simulations used in this work. This shows a $(250 \times 250 \times 50)h^{-3} \text{ Mpc}^3$ slice of the dark matter distribution from a single QUIJOTE simulation, with the color bar indicating the fractional density of simulation particles in each pixel. The colored points show the positions of dark matter halos; these are used as a proxy for galaxies. To facilitate analysis of percolation and connectedness, galaxies are assigned to clusters (indicated by the various colors), via a burning (or “friends-of-friends”) algorithm with a separation corresponding to reduced density η , here set to 0.2. Throughout this work, length is given in cosmologists’ units of megaparsec divided by the reduced expansion rate $h \approx 0.7$, in which the mean pairwise particle separation is $\approx 8h^{-1} \text{ Mpc}$.

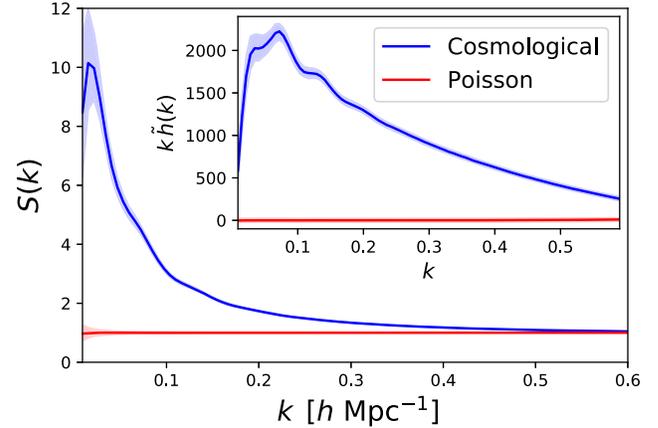
condensed phase systems in which these two length scales are comparable to one another (see Refs. [1,3]). At large scales, $g_2(r)$ decays to zero slightly slower than r^{-4} (in fact, $r^{-3.96}$), implying that large-scale correlations are suppressed; since the decay is between r^{-d} and $r^{-(d+1)}$ for dimension $d = 3$, the system is said to have *quasi-long-range* correlations. This has a physical origin: the large-scale behavior of $g_2(r)$ arises from correlations in the Universe’s quantum initial conditions, imprinted before cosmological inflation. Because of the dynamics of expansion (and slight breaking of time invariance in “slow-roll” inflation), these are suppressed on the largest scales. From the inset of Fig. 3(a), we note that $g_2(r)$ has considerable structure on intermediate scales, with a prominent peak at $r \approx 100h^{-1} \text{ Mpc}$ sourced by acoustic oscillations 14 billion years ago [107]. Clearly, the correlation properties of the galactic point cloud are very different to those for most media; this arises due to the combination of Poisson-like placements of galaxies and an underlying background stochastic field from the early Universe.

B. Structure factor

The structure factor $\mathcal{S}(k)$ tells a similar story as the pair correlation function. This is computed by first assigning the galaxies to a grid, then computing $\tilde{h}(k)$ via fast Fourier transforms, here implemented using NBODYKIT [108]. From Fig. 3(b), we observe a super-Poissonian signature on all scales, with a characteristic decline following a peak at $k \approx 0.05h \text{ Mpc}^{-1}$. This peak (known as the equality peak)



(a) Pair correlation function



(b) Structure factor

FIG. 3. Measurements of the pair correlation function (a) and structure factor (b) from 1000 QUIJOTE cosmological simulations (blue) alongside 1000 Poisson realizations (red). Both sets of simulations have the same number density ($\approx 1 \times 10^{-4} h^3 \text{ Mpc}^{-3}$) and a volume of $1h^{-3} \text{ Gpc}^3$, with a mean pairwise particle separation of $\approx 8h^{-1} \text{ Mpc}$. The shaded regions show the statistical variance between realizations (which grows large on large scales), and we note that the two statistics are related by a Fourier transform. In (a), the inset shows the pair correlation function in the typical cosmologists’ normalization, plotting $r^2 h_2(r) \equiv r^2 [g_2(r) - 1]$; this clearly brings out the structure imprinted by early Universe physics. At large r , we find $h_2(r) \sim r^{-(3+n_s)}$, where $n_s \approx 0.96$: this indicates the presence of quasi-long-range correlations. Similarly, the inset of (b) shows $k\tilde{h}(k)$, equal to the cosmologists’ $kP(k)$ (with a slope of k^{n_s} on large scales). The oscillatory features at $k \sim 0.1h \text{ Mpc}^{-1}$ arise from acoustic waves in the early Universe.

corresponds to a change in the Universe’s expansion rate at early times, with matter starting to drive the expansion rather than radiation pressure. At larger k , we again see the characteristic acoustic features, here shown by oscillations in $\tilde{h}(k)$. On the largest scales, the power spectrum or structure factor $\mathcal{S}(k) \sim k^{0.96}$ in the infinite-wavelength limit $k \rightarrow 0$ (just visible in this plot), with a slope set by the physics of inflation, and hence because the exponent α in (9) is 0.96, the Universe belongs to class III hyperuniformity, as defined in relation (10). If the Universe was scale invariant according to the Peebles-Harrison-Zeldovich spectrum with $\mathcal{S}(k) \sim k$ [4], then, because $\alpha = 1$, it would be hyperuniform of class II. Of course, either scenario implies that the structure factor vanishes in the limit $k \rightarrow 0$, the system is *hyperuniform* [109,110]. In contrast to many terrestrial media, the large-scale behavior is well understood, and can be predicted using a variety of cosmological codes; this occurs since it is an imprint of underlying dark matter physics, rather than a true pairwise interaction.

C. τ order metric

To quantify the degree of order or disorder of the galaxies, we compute the metric τ , which is defined by Eq. (13), as discussed in Sec. II B. We take $d = 3$ and the *unclustered* interparticle separation $D = \bar{\rho}^{-1/3} \approx 20h^{-1}$ Mpc to be the characteristic length scale. Here $\tau = 4.85$ for the cosmological sample, which is to be compared to $\tau = 8.37 \times 10^{-6}$ for the (finite-volume) Poisson realizations, close to the infinite-volume expectation of $\tau = 0$. This result supports the well-known results that the galaxy distribution is not purely random (uncorrelated), but instead is a correlated disordered system. To place the magnitude of τ for the galaxies in the context of other models of correlated disordered media, we compute τ for the random sequential addition (RSA) process, which is a time-dependent (nonequilibrium) procedure that generates disordered sphere packings in \mathbb{R}^d [111,112]. Starting with an empty but large volume in \mathbb{R}^d , the RSA process is produced by randomly, irreversibly, and sequentially placing nonoverlapping spheres into the volume. If a new sphere does not overlap with any existing spheres, it will be added to the configuration; otherwise, the attempt is discarded. This procedure is repeated for ever-increasing volumes; then, an appropriate infinite-volume limit is obtained. One can stop the addition process at any time t , obtaining RSA configurations with a range of packing fractions $\phi(t)$ up to the maximal “saturation” value $\phi(\infty)$ in the infinite-time limit, which for three dimensions is about 0.3812 [112]. Using the data for pair statistics given in Ref. [112], we find $\tau = 6.17$ for saturated RSA packings in \mathbb{R}^3 , which is close in value to that of the galaxies.

D. Local number variance

As discussed in Sec. II, the local number variance $\sigma_N^2(R)$ can also be computed from the measured correlation

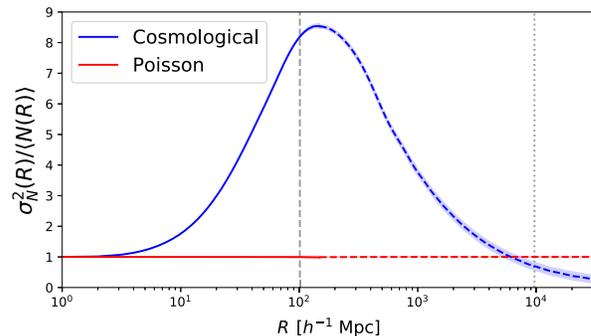


FIG. 4. Local number variance for the cosmological (blue) and Poisson (red) simulations as a function of scale R . This is defined as the variance of the number of particles found in spheres of radius R and computed directly from the correlation function shown in Fig. 3. The dashed lines show an extrapolation to large R , using the theoretical asymptotic structure factor form [with $\mathcal{S}(k) \sim k^{0.96}$]. The vertical lines show two characteristic scales: the sound horizon at recombination (dashed), which sources acoustic wave in the early Universe, giving the bump in $h(r)$, and the size of the Universe as it transitioned from radiation to matter dominated (dotted).

function, and provides a useful tool with which to assess the system’s order. Here, this is computed from the measured $g_2(r)$ values via Eq. (7), and plotted in Fig. 4, alongside its extrapolation to large R , using the well-known large-scale limit $\mathcal{S}(k) \sim k^{0.96}$. Notably, we find the number variance to increase faster than the Poisson case at small R , roughly up to the scale corresponding to the second peak in $r^2h(r)$ (arising from the imprint of acoustic oscillations from the early Universe), then fall to sub-Poisson values by scales corresponding to the peak in $\mathcal{S}(k)$. A variance that increases much faster than that for Poisson systems at small R is unusual for typical correlated disordered systems that have been investigated in condensed matter physics. Of course, that the large-scale variance approaches zero indicates that the system is hyperuniform; however, these scales are difficult to measure with most cosmological surveys.

E. Void and particle nearest-neighbor functions

In Fig. 5, we depict the nearest-neighbor functions of the two sets of simulations, which provide an alternative description of the system’s geometrical and topological properties, as discussed in Sec. II C (see also Ref. [77] for a previous discussion of the void function of galaxies, yielding similar results). These are obtained from the simulations by histogramming the minimum distance between each pair of particles (for H_p) or a pair of particles and a Poisson random particle (for H_v , determining if this particle lies within a void). For the Poisson system, we find identical results for the void and particle nearest-neighbor density functions, as expected, but significant differences for the QUIJOTE simulations. While the cosmological case has a similar distribution of small

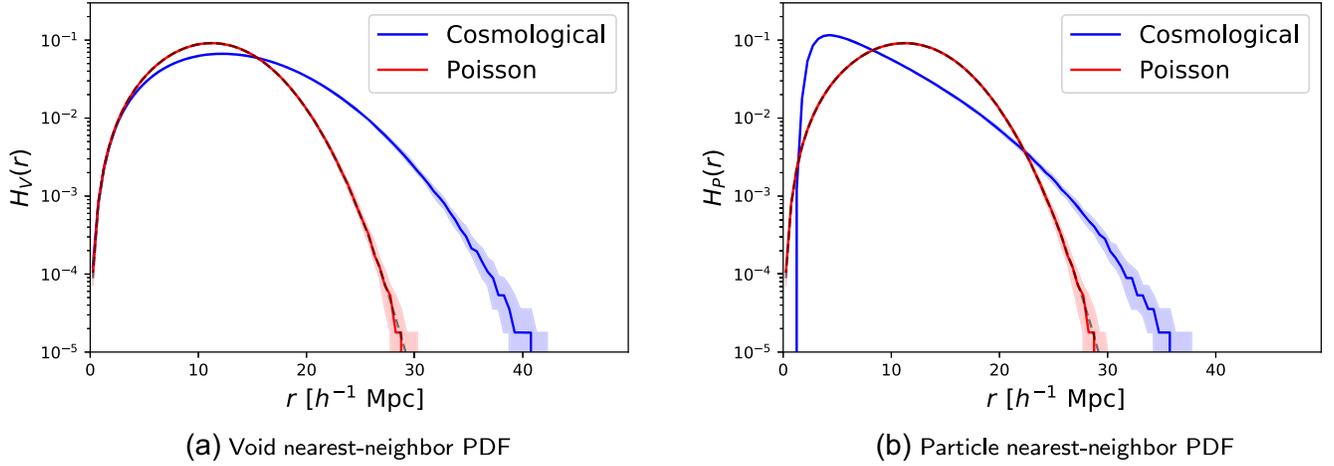


FIG. 5. Comparison of the void and particle nearest-neighbor probability density functions for the cosmological and Poissonian dataset, as defined in Eqs. (14) and (15). Panel (a) shows the probability distribution of finding a void of radius r in the cosmological (blue) and Poisson (red) datasets, averaged over 100 realizations, while panel (b) gives the distribution function of the distance of a given particle from its nearest neighbor. For the Poisson case, the dashed lines show theory curves (defined in Sec. II C), which are in excellent agreement with the simulations. The void distribution in the QUIJOTE simulations follows the (mean-density-matched) Poisson distribution at small r , but has an excess of large voids (shown by much broader tails), due to the quasi-long-range correlations. In contrast, the particle distribution $H_P(r)$ differs between the cosmological and Poissonian simulations on all scales, notably with an absence of small separations (due to halo exclusion effects) and an enhancement on large scales.

($r \lesssim 5h^{-1}$ Mpc) voids to that found in the Poisson realizations, it boasts significantly broader tail toward large r , and thus a somewhat larger mean void size. Specifically, the first moment of $H_V(R)$, $\ell_V^{(1)}$, defined by Eq. (20), is equal to $14h^{-1}$ Mpc ($11h^{-1}$ Mpc) for the cosmological (Poisson) simulations. Interestingly, the maximal void size (averaged over realizations) for the QUIJOTE simulations is $42h^{-1}$ Mpc, which is almost 50% larger than that for the Poisson system with a maximal void size of $30h^{-1}$ Mpc. Furthermore, the variance of H_V , defined as $\ell_V^{(2)} - (\ell_V^{(1)})^2$, is much larger for the cosmological case: $2.7h^{-2}$ Mpc² instead of $1.5h^{-2}$ Mpc². In particular, the above results suggest that the galaxies will also boast a lower percolation threshold, foreshadowing what we describe below.

For the particle distribution, we note that (a) the cosmological simulations have enhanced large-scale clustering, and thus a broad tail to the nearest-neighbor distance at large r , and (b) there is a sharp cut at low r , with no galaxies found within a separation of $\sim 1h^{-1}$ Mpc. This is a consequence of “halo exclusion”; a pair of galaxies cannot be arbitrarily close, else they would be identified as a single object in the simulation code. Between these two effects, we find a reduced mean particle nearest-neighbor distribution in QUIJOTE, indicating that galaxies are more likely to be found in large-scale clusters. This matches theoretical

expectations. Specifically, the mean nearest-neighbor distance between particles $\ell_p^{(1)}$, defined by Eq. (21), is equal to $8h^{-1}$ Mpc ($11h^{-1}$ Mpc) for the cosmological (Poisson) simulations, with a minimum distance of $1.1h^{-1}$ Mpc ($0.29h^{-1}$ Mpc). In addition, the variance of the cosmological H_P is again larger than the Poisson case, finding $2.5h^{-2}$ Mpc² instead of $1.5h^{-2}$ Mpc².

F. Pair-connectedness and direct-connectedness functions

The astrophysical pair-connectedness function $P_2(r)$ has not been previously studied in the literature, and is of particular interest to both cosmology and condensed matter physics. To construct this, we first take the set of $N \sim 10^5$ galaxy positions in each QUIJOTE (or Poisson) simulation, and assign clusters via a “burning” algorithm (often known as “friends-of-friends” in cosmology) [113,114], here using the NBODYKIT implementation [108]. This finds sets of points for which each member is connected to each other member via a path through the clustered phase formed of spheres of radius $D(\eta) = [6\eta/\pi\bar{\rho}]^{1/3}$ around each point, where η is the reduced density. Given the set of particles and cluster memberships (visualized in Fig. 2), we compute the pair-connectedness function in bins with centers $\{r_a\}$ via

$$\hat{P}_2(r_a, \eta) = \frac{1}{N\bar{\rho}v_a} \sum_{i,j=1}^N \begin{cases} 1 & \text{if } |\mathbf{r}_i - \mathbf{r}_j| \text{ in bin } r_a \text{ and } i, j \text{ in same cluster} \\ 0 & \text{else,} \end{cases} \quad (33)$$

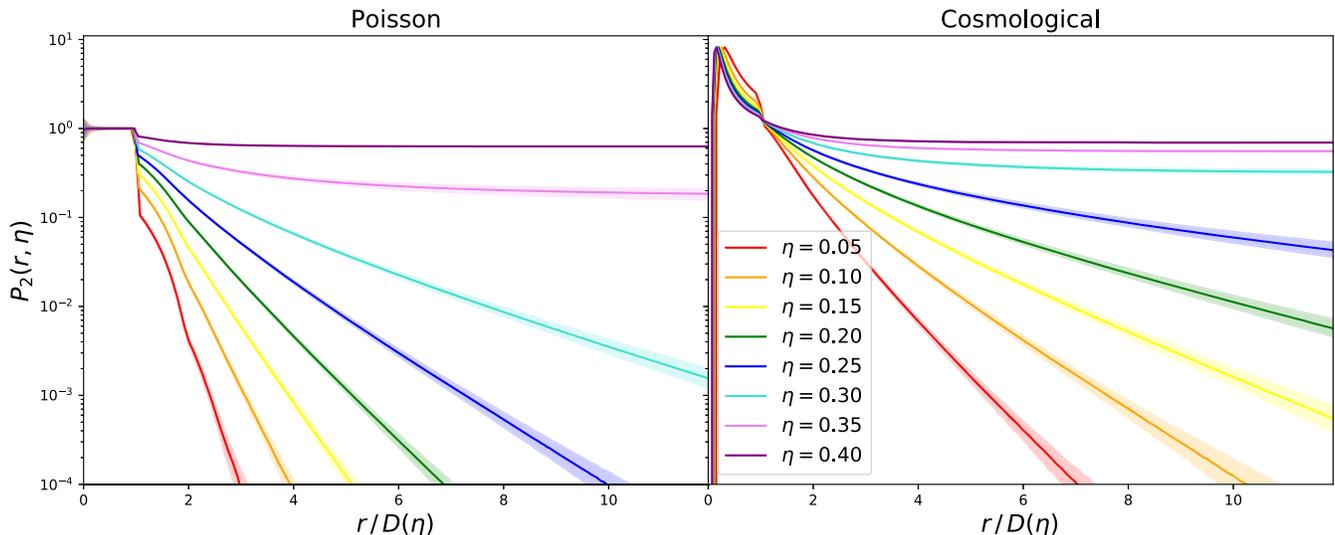


FIG. 6. Pair-connectedness function $P_2(r, \eta)$ for the Poisson (left) and cosmological (right) simulations. Results are shown for a variety of reduced densities η , corresponding to clustering distances $D(\eta)$ in the range $[10, 20]h^{-1}$ Mpc, and shaded regions show the 1σ deviations expected from statistical fluctuations. The QUIJOTE simulations show significantly enhanced correlations on large scales, due to the underlying correlations of matter imprinted in the early Universe. This additionally suggests that the galaxy sample will percolate at lower η : this will be explored in Sec. V.

analogous to Eq. (32). This is achieved using a custom modification of the CORRFUNC code [106], which accepts pairs only if they have the same cluster index.

Figure 6 displays the pair correlation functions from the QUIJOTE simulations alongside the more familiar Poisson case. The latter match our expectations: $P_2(r) = 1$ for $r < D$ (since all particles with this separation must be in the same cluster), and $P_2(r)$ falls sharply with r for $r > D$ (due to an absence of large-scale clusters), with an enhanced decline at low η . For large η , the volume integral of $P_2(r)$ appears to diverge (at least in the infinite-volume limit), indicating percolation. [Note that the simulations are computed in periodic boxes, which are known to be suboptimal for computing $P_2(r)$ on the largest scales [96]. This will be addressed in Sec. V in the context of finite-scaling analyses.] For the cosmological simulations, we first note that $P_2(r) = g_2(r)$ for $r < D$, as expected. At larger r , we find that $P_2^{\text{sim}}(r) > P_2^{\text{poiss}}(r)$ for all choices of η , indicating that our galaxy catalogs contain more long-range correlations than a Poisson random field of the same density, and suggesting that the system will also percolate quicker. In the large η limit (i.e., above percolation), $P_2(r) \rightarrow g_2(r)$, since all points belong to the same cluster. The cosmological utility of $P_2(r, \eta)$ will be discussed in Sec. VI.

The direct-connectedness function also plays an important role in the analysis of connected systems, in part due to its appearance in the Ornstein-Zernike equation (23). Given $P_2(r)$, this can be computed using Eq. (24), performing the Fourier transforms numerically via the FFTLog prescription [115]. Figure 7 shows $C_2(r, \eta)$ for both the cosmological and Poisson simulations, alongside the analytic

Percus-Yevick (PY) model, which solves the OZ equation by asserting that $C_2(r > D, \eta) = 0$ and $P_2(r < D, \eta) = 1$ [94]. [This is computed for point objects by using the correspondence with the known (cubic) form for hard spheres via $C_2^{\text{PY, Poiss}}(r, \eta) = -C_2^{\text{PY, hard sphere}}(r, -\eta)$ [116].] For the Poisson case, we find good agreement between theory and simulations for small η (far from the percolation threshold of $\eta_c \approx 0.34$), particularly away from the boundary at $r = D(\eta)$. We observe very little power from the region with $r > D(\eta)$, since most intracluster pathways with $r > D(\eta)$ contain at least one node, and thus do not contribute to $C_2(r)$. The cosmological simulations show a very different behavior, with two peaks observed, with one in similar location to the Poisson system and one at smaller r . This statistic represents the complexities of the clustering on smaller scales than that typically seen in P_2 [with $D(\eta) \sim 10h^{-1}$ Mpc], and the differences arise primarily due to small-scale physics, such as the restriction that galaxies cannot be arbitrarily close together. We also note that $C(r, \eta)$ was found to be ill-behaved for $\eta \gtrsim 0.3$ [due to $P_2(r, \eta)$ not being square integrable], indicating that the cosmological simulations have percolated by around this value of η (cf. Sec. V). In practice, we expect the percolation threshold to depend on the peculiarities of the galaxy sample in question: this will be discussed further below.

V. PERCOLATION AND FRACTAL DIMENSIONS

We now turn to the issue of percolation, following the discussion in Sec. II E. As noted earlier, determining the mean cluster size $S(\eta)$ in a system as a function of η is a

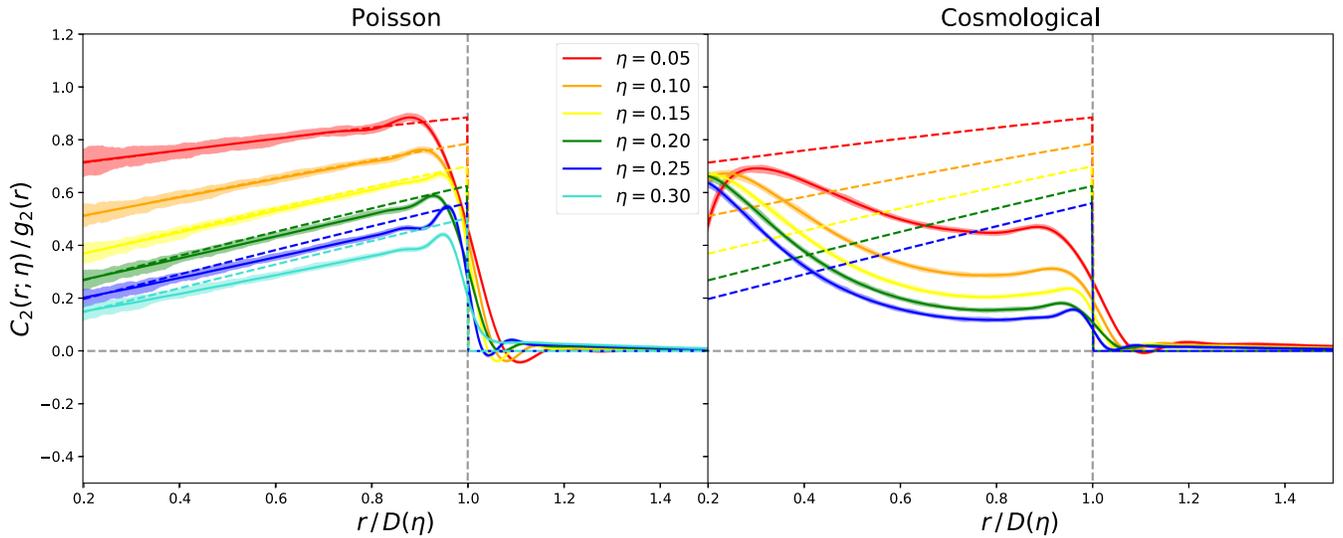


FIG. 7. Direct-connectedness function $C_2(r, \eta)$ for the Poisson (left) and cosmological (right) simulations. This follows Fig. 6, but focuses on smaller scales, and additionally includes analytic predictions from the Percus-Yevick model. We additionally normalize all quantities by $g_2(r)$, and exclude values of η for which $C_2(r, \eta)$ is not well-behaved (beyond the percolation threshold).

useful way in which to test whether a system has reached percolation. We utilize the representation of $S(\eta)$ in terms of s -mer cluster statistic n_s , as defined by Eq. (29). In principle, we expect $S(\eta) \rightarrow \infty$ as $\eta \rightarrow \eta_c$; in practice, $S(\eta) \leq N$, where N is the total number of particles in the box. To account for this, it is useful to analyze a number of different configurations with different box sizes L (and thereby $N \equiv \bar{\rho}L^3$). Here, we construct (aperiodic) subboxes from the QUIJOTE simulations, with L in the range $[400, 800]h^{-1}$ Mpc (noting that the majority of our analyses are restricted to $r < 200h^{-1}$ Mpc), and construct

analogous Poisson realizations for each. To examine percolation at each choice of box size, we generate clusters for various values of η by varying the sphere radius $D(\eta)$ and utilizing burning (friends-of-friends) algorithms, as described above.

Figure 8 shows the mean cluster size for the two datasets as a function of L and η . In both cases, we observe that $S(\eta, L)$ begins to approach its asymptotic limit as η increases, and, moreover, the limit is approached faster as the box size increases. Extrapolating the Poissonian results to large L , the percolation threshold [whence

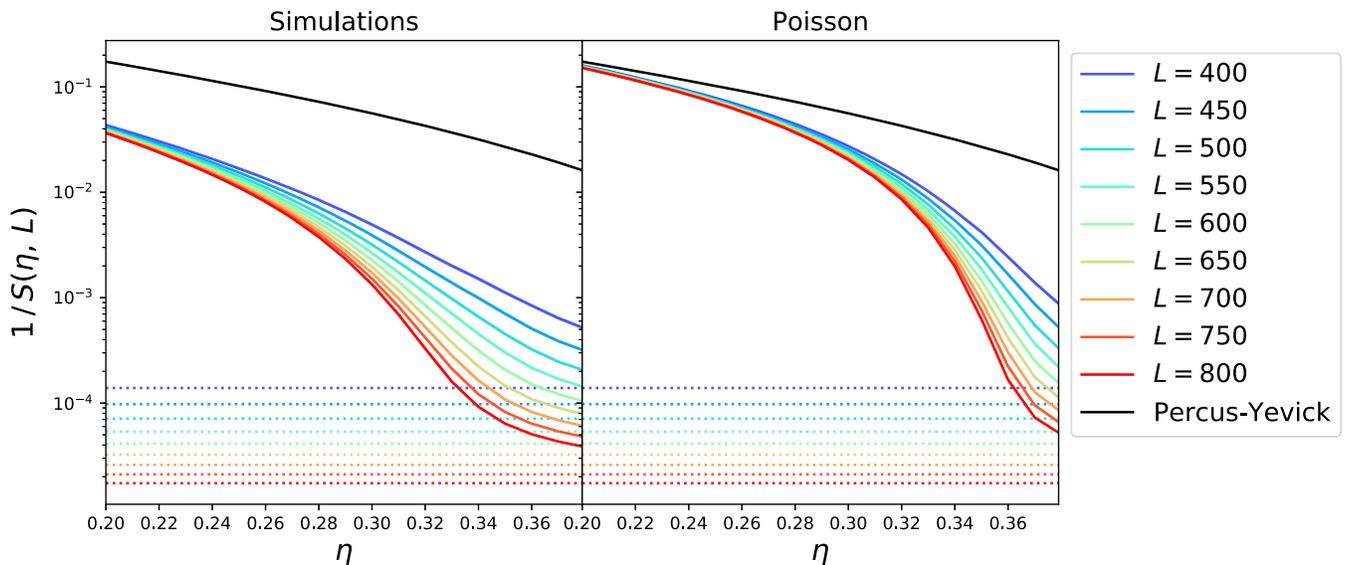


FIG. 8. Mean number of particles per cluster $S(\eta, L)$, as a function of the reduced density η and (cubic) simulation box size L . We show results for both cosmological data and Poisson realizations, plotting the reciprocal $1/S(\eta, L)$ averaged over 1000 realizations. Note that $S(\eta)$ is bounded by the number of particles in the dataset (N), shown by dotted lines. We additionally show the Percus-Yevick prediction in black, which is inaccurate for all but the smallest η .

$S(\eta, L) \approx N]$ appears to be around $\eta_c = 0.35$, matching previous studies [97]. For the cosmological simulations, we find a generally slower approach to η_c (corresponding to a different critical exponent), and additionally a lower percolation threshold, around $\eta_c = 0.28$ for $L \rightarrow \infty$. As above, this arises since the galaxy sample contains a stochastic background inhomogeneity, leading to various areas being super- or sub-Poisson populated in a correlated manner. Even at low η : for $\eta = 0.1$, clusters in the full-volume cosmological simulation contain an average of ≈ 15 particles, while those in the Poisson realizations contain ≈ 2 only.

It is further instructive to consider the size distribution of clusters, via the average number of s -mers n_s , as defined in Eq. (29). This is shown in Fig. 9 for a suite of cosmological and Poissonian boxes at $L = 800$ with a variety of values of the reduced density η . At low η , we find that the ratio of cosmological and Poissonian simulations is a strongly increasing function of s , with the largest slopes seen for small reduced densities. In this limit, the system is far from percolation; thus large clusters are rare in both systems. The enhanced correlations in the galaxy distribution seen in the cosmological case increase the probability of an s -mer forming (at fixed s), giving this stark difference in behavior. As η approaches the percolation threshold, the n_s ratio becomes roughly constant with s ; this indicates that the additional galaxy correlations impact only the largest s -mers, as we are dominated by the clustering signal imprinted by the circumscribed spheres, rather than any intrinsic effects.

To measure the percolation threshold of the cosmological simulations in a robust fashion, we perform a finite-scaling analysis, following the approach of Ref. [117],

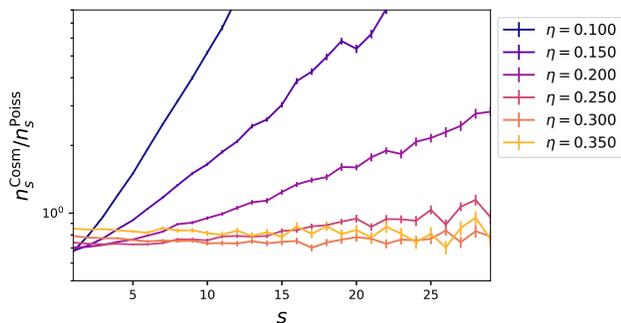


FIG. 9. Comparison of the cluster size distribution in the cosmological and Poisson simulations. We plot the mean number of s -mers per unit particle (defined as clusters containing s member) as a function of s , normalizing to the Poisson prediction. Results are shown for various values of the reduced density η and we assume an aperiodic simulation volume of size $L = 800h^{-1}$ Mpc, averaging over 100 realizations. At large η , the two sets of simulations have a similar n_s distribution (at least for small s), while at low η , the enhanced clustering in the cosmological simulations leads to significantly more s -mers, with the mean number of particles per cluster increasing with η (cf. Fig. 8).

originally formulated in Ref. [118]. In essence, this computes the percolation probability $[\Pi(\eta, L)$, defined as the fraction of realizations containing a cluster for which the circumscribed spheres overlap with both the top and bottom of the box] for the simulations at various values of L and η and extrapolates using asymptotic scaling relations to find the $L \rightarrow \infty$ limit. Figure 10 shows the obtained percolation probability distribution for both sets of simulations as a function of the volume filling fraction ϕ . This is computed numerically for each simulation from the probability that a randomly chosen point within the box is within a distance $D/2$ from the nearest particle, i.e., whether it is within the sphere phase; for the Poisson case, this is asymptotically equal to $1 - e^{-\eta}$. The behavior seen in Fig. 10 is qualitatively similar for the Poisson and cosmological system: the percolation probability is small for low ϕ (whence the typical extent of the cluster is far below L), and asymptotes to unity at large ϕ . As the box size increases, the transition becomes sharper, and asymptotes to a Heaviside function in the $L \rightarrow \infty$ limit. It is also clear that the cosmological simulations percolate at smaller values of η than the Poisson realizations; this is as expected, and indicates their enhanced clustering due to background inhomogeneities.

To extract the percolation thresholds from $\Pi(\phi, L)$, we fit the data to the phenomenological sigmoid model of Ref. [117], as shown in Fig. 10:

$$\Pi(\phi, L) \approx \frac{1}{2} \left[1 + \tanh \left(\frac{\phi - \phi_c(L)}{\Delta(L)} \right) \right], \quad (34)$$

where $\phi_c(L)$ and $\Delta(L)$ are the percolation volume fraction and width at box size L and ϕ is the volume filling fraction obtained as described above. Asymptotically, $\Delta(L) \sim L^{-1/\nu}$ and $\phi_c(L) - \phi_c \sim L^{-1/\nu}$ for critical exponent [1]; by fitting for ν from the obtained values of $\Delta(L)$, we can thus obtain $\phi_c \equiv \lim_{L \rightarrow \infty} \phi_c(L)$. Here, we find a critical exponent of $\nu = 0.85 \pm 0.03$ (0.88 ± 0.03) for the QUIJOTE (Poisson) simulations, with a corresponding percolation threshold of $\phi_c = 0.223$ (0.290) or $\eta_c = 0.252$ (0.343), each with a statistical error around 0.001. The Poissonian case matches standard results [1], and, as foreshadowed in Figs. 6 and 8, the cosmological system percolates at lower densities, due to the additional clustering signature imprinted by early Universe and galaxy formation physics. In addition, the fact that the two sets of simulations appear to share the same critical exponent ν suggests that they belong to the same universality class, as do other correlated disordered systems [119].

It is important to note that the percolation thresholds found herein are not a universal property of galaxy distributions; rather, they depend on the galaxy sample in question. To explore this, we have repeated the analysis using a galaxy sample with half the density of the fiducial sample, and another including dark matter halos (i.e.,

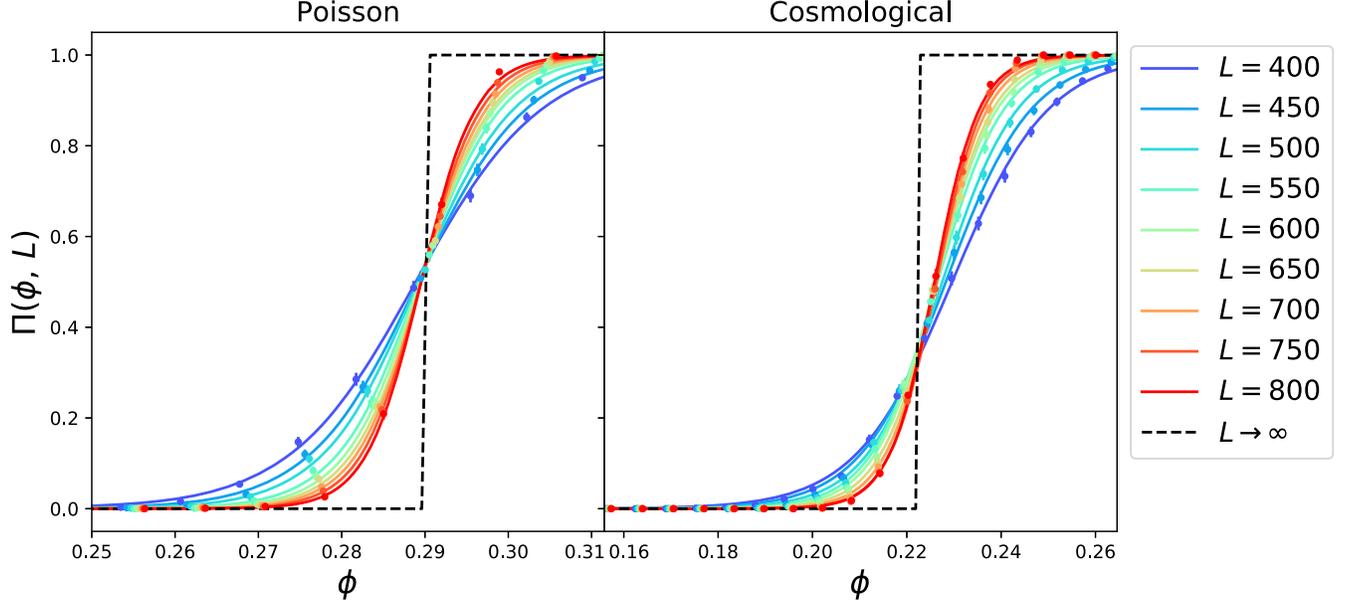


FIG. 10. Percolation probability $\Pi(\phi, L)$ obtained from 1000 QUIJOTE and Poisson simulations, for various nonperiodic box sizes L and volume filling fractions ϕ . This is defined as the fraction of realizations containing a sample spanning cluster, and asymptotes to a Heaviside function centered at the percolation threshold for $L \rightarrow \infty$ (shown as a dotted line). Points represent the values computed from simulations, while the solid lines show a fit using the sigmoid function of Eq. (34). Using a finite-scaling analysis, we find the percolation thresholds of $\eta_c = 0.252$ (0.343) for the QUIJOTE (Poisson) simulations, though we caution that the cosmological result depends on the sample density.

galaxies) down to half of the aforementioned minimum size. For the former case (with $\bar{\rho} \sim 0.5 \times 10^{-4} h^3 \text{ Mpc}^{-3}$, we find that the percolation threshold for the cosmological sample increases to $\eta_c = 0.285$, while it remains the same for the density-matched Poissonian sample (as expected). This can be rationalized by noting that the galaxies roughly follow Poisson statistics above a stochastic background, caused by the matter density; if $\bar{\rho}$ is reduced, the Poisson part of the stochasticity becomes more dominant, thus η_c tends toward its Poisson limit. In the second scenario, we find $\eta_c = 0.207$, significantly lower than the fiducial analysis. In this case, we have both a sample of almost twice greater density and one that is more biased with respect to the continuous dark matter density (such that $\hat{\rho}_{\text{galaxy}}/\hat{\rho}_{\text{dark matter}}$ is larger, smoothed on sufficiently large scales). In both cases, however, we find a similar critical exponent (0.82 ± 0.03) to the above.

Finally, we consider the effective fractal dimension of the system d_F . As discussed in Sec. II, this may be computed from the dependence of the sample-spanning cluster mass $M(\eta, L)$ (i.e., its number of constituent particles) on the simulation box size L at the percolation threshold η_c [Eq. (30)]. To explore this, we repeat the above analysis for the fiducial sample, computing the mass of the sample-spanning cluster (when it exists) for five box sizes in the range $[700, 800]h^{-1} \text{ Mpc}$ and five reduced densities in the range $\eta_c \pm 0.1$. For the Poisson system, fitting for the relationship $M(\eta, L) \sim L^d$ and

interpolating to η_c gives $d_F \equiv d(\eta_c) \approx 2.40 \pm 0.07$, matching that predicted from theory [1]. For the galaxy sample, we find $d_F = 2.36 \pm 0.08$, which is consistent with the Poisson realizations, even though the percolation threshold differs. This is an important result: the cosmological sample lies in the same universality class as simple Poisson realizations, for the range of scales considered: $r \sim 500\text{--}1000h^{-1} \text{ Mpc}$. This is broadly consistent with previous results on smaller scales; Refs. [98,99,120] describe a variety of methods to ascertain the effective fractal index, with galaxy counts yielding $d_F = 2.2 \pm 0.2$ on $\lesssim 10h^{-1} \text{ Mpc}$ scales, and correlation functions finding the same on $\lesssim 100h^{-1} \text{ Mpc}$ scales.

VI. PAIR-CONNECTEDNESS FUNCTION AS A COSMOLOGICAL DESCRIPTOR

A. Background

A crucial problem in modern-day cosmology is the extraction of physical parameters from observed statistics, such as the distribution of galaxies. In the standard paradigm (dubbed $\nu\Lambda\text{CDM}$), six parameters are of relevance: (1) the Universe's current expansion rate H_0 , (2) the density of baryonic matter ω_b , (3) the combined density of dark matter and baryonic matter Ω_m , (4) the amplitude of clustering in the Universe σ_8 , (5) the slope of the primordial power spectrum (i.e., structure factor) n_s , and (6) the sum of the neutrino masses $\sum m_\nu$. While n_s and

Ω_b are well constrained by observations of the cosmic microwave background [121], the remaining parameters are a key target for upcoming galaxy surveys. Traditionally, they are constrained through summary statistics such as the two-particle and three-particle correlation functions. Such an analysis proceeds by the fitting measured statistics to analytic models depending on the above physical parameters [102].

In this section, we consider the utility of alternative statistics (described above) in this effort. Three metrics by which we judge a statistic to be useful are (a) ease of computation, (b) dimensionality, and (c) information content. Here, we will principally concentrate on the pair-connectedness function $P_2(r)$, since this has not previously been used in cosmology, unlike void probability or nearest-neighbor functions [80]. As shown above, the statistic is simple and fast to measure from the data and has a low-dimensional form, satisfying two of the above criteria. We now proceed to quantify its ability to constrain cosmological parameters.

An alternative approach to the above prescription is to model the entire galaxy distribution directly (without compressing to statistics such as the correlation functions), with either perturbative methods [122–125] or machine learning approaches [126–131]. In principle, this approach enables one to obtain optimal constraints on all parameters of interest, though it is nontrivial to implement in practice, due to the huge dimensionality of the galaxy distribution and the necessity to run a large number of expensive simulations.

B. Quantifying information content

Standard cosmological analyses proceed by measuring a set of statistics from a dataset, then comparing them to accurate physical models depending on cosmological parameters, including those discussed above. If the noise properties of the statistics are known (for example, if we assert that the distribution of g_2 is a multivariate Gaussian), this comparison can be used to place constraints on the underlying parameters via Bayes theorem. A useful estimate of the constraining power of some statistic X (e.g., g_2) can be obtained using a Fisher matrix [132], defined as

$$F_{\alpha\beta}^X = \left(\frac{dX}{d\theta_\alpha} \right)^T \mathbf{C}_X^{-1} \left(\frac{dX}{d\theta_\beta} \right), \quad (35)$$

where $\{\theta_\alpha\}$ are the set of cosmological parameters of interest, and \mathbf{C} is the covariance matrix of X (treated as a vector), i.e., $\mathbf{C}_X = \mathbb{E}[XX^T]$, averaging over realizations of the underlying microscopic density at fixed θ . According to the Cramér-Rao theorem, $(F^{-1,X})_{\alpha\alpha}$ gives the best possible constraint on θ_α from a measurement X , i.e., $\text{var}(\theta_\alpha) \geq (F^{-1,X})_{\alpha\alpha}$. [This limit is saturated if X obeys Gaussian statistics, i.e., $\hat{X} \sim \mathcal{N}(X(\theta), \mathbf{C})$.] To assess the utility of statistics such as P_2 and g_2 , we need simply

compute the covariance matrix and the parameter derivatives appearing in Eq. (35), both of which can be done using a set of simulations. Explicitly, given a set of n realizations $\hat{X}^{(i)}$ with varying initial conditions, the two can be computed via

$$\hat{\mathbf{C}}_X = \frac{1}{n-1} \sum_{i=1}^n (\hat{X}^{(i)} - \bar{X})(\hat{X}^{(i)} - \bar{X})^T, \\ \widehat{\frac{dX}{d\theta_\alpha}} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{X}^{(i)}(\theta_\alpha + \delta\theta_\alpha) - \hat{X}^{(i)}(\theta_\alpha - \delta\theta_\alpha)}{2\delta\theta_\alpha}, \quad (36)$$

using finite difference for the parameter derivatives, and denoting $\bar{X} = (1/n) \sum_{i=1}^n \hat{X}^{(i)}$ [104].

While the Fisher forecast is appealing in its simplicity, it is not without limitations. Firstly, it gives accurate bounds on cosmological properties only if the statistics are Gaussian distributed, which can break down in the case of large correlations between bins, and the parameter posterior is Gaussian, which fails for non-negative parameters, for example. Secondly, a large number of simulations may be required to compute the quantities in Eq. (36), and, if too few are used, the constraining power of a given statistic will be overestimated. (This occurs since noise in the parameter derivatives add a positive definite contribution to the Fisher matrix, and thus reduce the size of the inverted matrix, i.e., the output parameter variances.) An alternative approach is to use *simulation-based inference* (also known as likelihood-free analysis) [133–135]. In essence, this draws a set of cosmological parameters θ from some input prior, computes a realization $\hat{X}(\theta)$ for each, and compares a “true” dataset to the empirical distribution from the simulations. This does not make assumptions on the statistics’ noise properties, and, in the case of too few simulations, will only underestimate the cosmological utility.

Here, we examine the constraining power of various summary statistics using both the Fisher matrix formalism (which has become commonplace in cosmology) and simulation-based inference (which is far less common, though more accurate). In particular, we consider the cosmological parameters H_0 , Ω_m , σ_8 , and $\sum m_\nu$ and the following descriptors: g_2 , g_3 , P_2 , and H_V , all of which can be defined for discrete point clouds, such as the galaxy density used in this work. For the pair-connectedness function, we fix the reduced density to $\eta \equiv \bar{\rho}\pi D^3/6 = 0.2$, which is a useful balance between the uninformative case ($\eta = 0$) and the percolated limit discussed in Sec. V, though we note that other choices may yield somewhat different results. Additional statistics could be straightforwardly added, though we caution that descriptors such as the number variance are fully described by g_2 , and will thus not add additional information. Using the FASTPM code [136], we run $n = 512$ QUIJOTE-like dark matter simulations with the following fiducial parameters:

$\{H_0 = 68 \text{ km s}^{-1} \text{ Mpc}^{-1}, \Omega_m = 0.31, \omega_b = 0.0227, n_s = 0.96, \sigma_8 = 0.8178, \sum m_\nu = 0 \text{ eV}\}$, and compute the four statistics for each realization.

In all cases, we consider only scales above $20h^{-1} \text{ Mpc}$ (where the simulations are accurate, given the mean pairwise separation of $\approx 8h^{-1} \text{ Mpc}$), and choose the radial bin sizes to keep the dimensionality fixed to $\mathcal{O}(100)$ elements. g_2 , P_2 , and H_V are computed as before, with g_3 computed using the approach of Ref. [137], involving a decomposition into a Legendre multipole basis, with components $g_{3,\ell}(r_1, r_2)$ for $\ell \in \{0, 1, 2, 3\}$. In accordance with Sec. III, we compute this statistic in configuration space (rather than as a Fourier-space bispectrum), which obviates the need to grid the particles. These simulations are used to compute the covariance matrix \mathbf{C} of the statistics using Eq. (36) (initial testing demonstrated that this is a sufficient number of simulations to estimate \mathbf{C} and its inverse robustly, after including the correction factor of Ref. [138]), the structure of which is visualized in Fig. 11. We find significant correlations both within and between a number of statistics. In particular, the individual bins of g_3 and P_2 are highly correlated, indicating that their noise properties may not be Gaussian. In contrast, the void

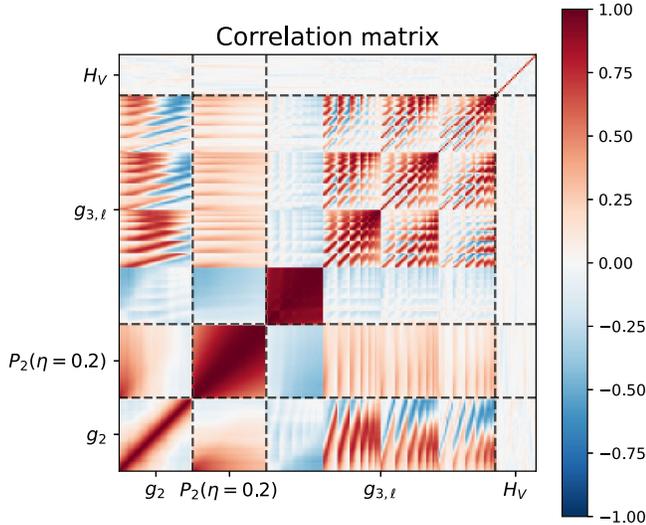


FIG. 11. Correlation matrix of the pair correlation function (g_2), the pair-connectedness function (P_2 , with reduced density $\eta = 0.2$), the three-particle correlation function (g_3), and the void nearest-neighbor function (H_V) obtained from 512 cosmological simulations run with a fiducial set of cosmological parameters. The correlation matrix is defined as the covariance matrix normalized by its leading diagonal, i.e., $C_{ij}/\sqrt{C_{ii}C_{jj}}$, with all values lying in the range $[-1, 1]$. The labels show the statistic of interest, which are demarcated by the dotted lines. In each statistic, the bins are ordered from small scale (bottom left) to large scale (top right). For g_3 , we use four Legendre multipoles $g_{3,\ell}$ with $\ell \in \{0, 1, 2, 3\}$. Notably, the individual bins of $g_{3,0}$ and P_2 are highly correlated, and there are a number of nontrivial correlations between observables, though few with H_V .

nearest-neighbor function has an almost diagonal correlation matrix, and is seen to be largely independent from other statistics. This suggests that it can add significant information compared to analyses using g_2 alone.

The other ingredient required for Fisher forecasting is the set of parameter derivatives. These are computed using Eq. (36), with $n = 512$ simulations (again computed using FASTPM, with a total cost of $\sim 10^4$ CPU hours), utilizing finite difference in each of the eight sets of parameters. For the neutrino mass, we have the bound $\sum m_\nu > 0$, thus we instead utilize one-sided derivatives, following Ref. [104], and use the method of Ref. [139] to emulate the effects of massive neutrinos by modifying the initial conditions. Following this, we compute the Fisher matrix via Eq. (35) for various combinations of statistics. We caution that this result appears to retain some dependence on n due to residual noise in the parameter derivatives. This will lead to the constraints being artificially tightened somewhat; however, it is computationally impractical to increase n by a significant amount.

For the simulation-based inference (SBI), we utilize a set of 8192 galaxy simulations computed using FASTPM with the method of Ref. [139] at random locations in parameter space, according to the flat priors: $H_0 \in [0, 100] \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m \in [0, 0.3]$, $\sigma_8 \in [0.4, 1.2]$, $\sum m_\nu \in [0, 4] \text{ eV}$. (The neutrino mass limit is significantly weaker than the bound from the latest probes [121], but is appropriate given the small volume of the simulations.) Summary statistics for each are computed as before, and fed into the SBI code, which uses neural networks (via the sequential neural posterior estimation method) to compute the parameter posterior, given a true observation, taken from the mean of the fiducial simulations discussed above.

C. Results

Figure 12 shows the constraints on cosmological parameters from analyses using the pair correlation function and pair-connectedness function, both via the Fisher and SBI forecasts. From the Fisher forecast, we observe that P_2 is a poor predictor of the expansion rate and matter density, but gives much tighter constraints on the clustering amplitude and neutrino mass than g_2 . This is unsurprising: the slope of $P_2(r)$ is strongly sensitive to the galaxy clustering properties (set by σ_8 , and, on small scales, $\sum m_\nu$), but (being a monotonic function) contains little information on other properties such as early Universe physics. The SBI forecasts give qualitatively similar results, with the P_2 constraints on H_0 and Ω_m being largely dominated by the prior, with the combined $g_2 + P_2$ constraints reproducing those of g_2 alone. For σ_8 , P_2 is again shown to be of considerable use, with a significant (factor of a $\approx 2.6\times$, equivalently to observing a 7 times greater volume of space) tightening in the one-dimensional posterior found by adding P_2 , driven by the differing degeneracy directions in the $\sigma_8 - H_0$ and $\sigma_8 - \Omega_m$ planes. In contrary to the Fisher result, the SBI

forecast suggests that the pair-connectedness function does not give significant additional information on the neutrino mass; however, it is shown to change the $\sigma_8 - \sum m_\nu$ degeneracy direction considerably.

The disagreement between Fisher and SBI forecasts both quantitatively (in terms of the reduction in width of the σ_8 posterior) and qualitatively (whether the $\sum m_\nu$ posterior is affected) may appear a little unsettling. We attribute this to a number of reasons: (1) as mentioned above, the Fisher forecast will give artificially narrow constraints if insufficient simulations have been run, (2) the Fisher forecast is inaccurate for parameters whose posterior is non-Gaussian (such as the neutrino mass, due to the $\sum m_\nu > 0$ constraint), and (3) the SBI forecasts can be artificially *broadened* by insufficient simulations being run. However, the results of Fig. 12 are enough to convince us that P_2 contains significant information regarding the clustering amplitude σ_8 , and its inclusion greatly aids cosmological analyses, including via degeneracy breaking with $\sum m_\nu$. Although we present results only for reduced density $\eta = 0.2$ here, a similar story holds also for $\eta = 0.1$; in this case, the improvements in cosmological parameters are somewhat weaker, due to the higher-order correlator contributions to $P_2(r, \eta)$ being suppressed [Eq. (27)]. We expect that combining

measurements of the pair-connectedness function with multiple values of η could further increase the constraining power, again at little computational cost.

It is interesting to compare the cosmological utility of the pair-connectedness function to that of other higher-order statistics. Before doing so, let us briefly outline our predictions. In this test, we are limited to relatively large scales ($r \gtrsim 20h^{-1}$ Mpc, due to simulation resolution effects), where the galaxy distribution (if treated as a continuous field) is close to Gaussian. As such, we expect the majority of the information content on cosmological parameters to be encapsulated by the two- and three-point functions, g_2 and g_3 , with only a small amount leaking into higher-order statistics. In this case, the combination of g_2 with alternative statistics will likely perform worse than that of g_2 and g_3 ; our question is whether there are statistics that are able to recoup most of the information present in g_3 in a simpler form (for example, in the unidimensional H_V and P_2 statistics). If such a statistic exists, it is likely that it also contains significant information on small scales (as probed by future surveys such as that of the Subaru Prime Focus Spectrograph and the DESI Bright Galaxy Survey [140,141]), where the perturbative hierarchy described above breaks down. Indeed, statistical physics provides examples of small-scale systems where $g_2 + P_2$

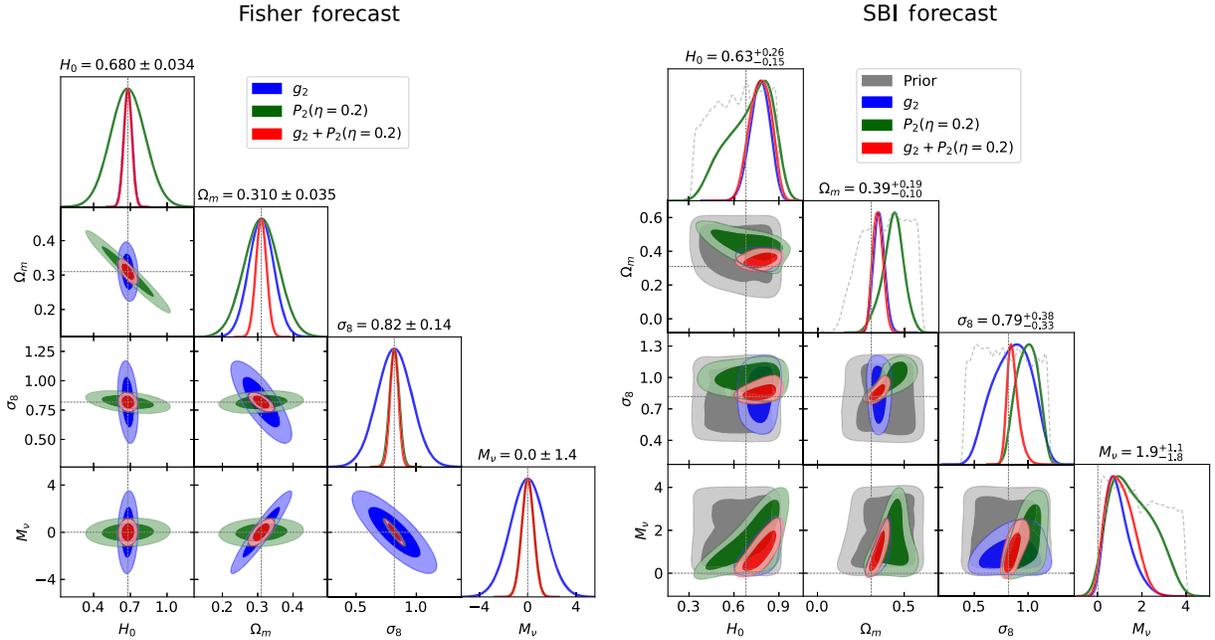


FIG. 12. Forecasted constraints on key cosmological parameters using the pair correlation function g_2 and the pair-connectedness function P_2 , using a Fisher forecast (left) and simulation-based inference (right). The dark (light) ellipses represent our 68% (95%) confidence intervals on the various parameters possible from a single galaxy simulation (as in Fig. 2), with the gray regions showing the prior used in the simulation-based approach (from 8192 simulations). We show results for four parameters: H_0 , giving the Universe's expansion rate, the matter density Ω_m , the clustering amplitude σ_8 , and the total neutrino mass $\sum m_\nu$. The diagonal figures give the marginalized constraints on single parameters, while the off diagonals show the correlations; e.g., the bottom left-hand panel shows the correlation between h and σ_8 . We find that P_2 is a poor predictor of H_0 and Ω_m (shown by the broad contours), but can tightly constrain σ_8 . The combination $g_2 + P_2$ significantly increases the precision by which this parameter can be measured, and we find similar results from the Fisher forecasts (which are, in general, overoptimistic), and the simulation-based inference (which is usually conservative).

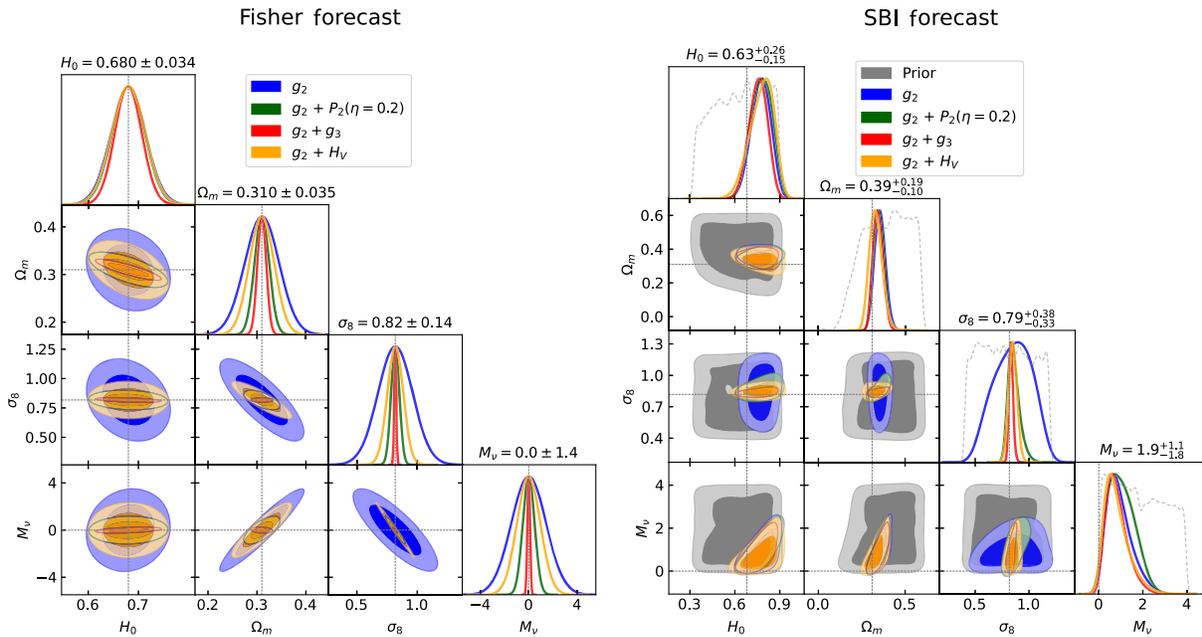


FIG. 13. As Fig. 12 but comparing the constraining power from the combination of the pair correlation function with the pair-connectedness function (green), the three-particle correlation function (red), and the void nearest-neighbor function (yellow). The pair-connectedness function and void nearest-neighbor functions are found to perform almost as well as the three-particle function in this scenario (in terms of constraining σ_8) and are much faster to compute and analyze.

outperforms $g_2 + g_3$ [91]; it will be interesting to study such effects further in the future.

In Fig. 13 we show the Fisher and SBI constraints on the same parameter set as above for g_2 in combination with g_3 , P_2 , and H_V [142]. From both the Fisher and SBI forecasts, we find that no additional statistics lead to significant improvements in the expansion rate constraints, except for a slight tightening from g_3 . This is not surprising: H_0 is primarily measured from an oscillatory feature in g_2 arising from acoustic waves in the early Universe, which is generally absent in other statistics. In the SBI forecast, the same is true for the matter density, Ω_m , and the neutrino mass, though the Fisher forecasts disagree on this aspect, as above, and should therefore caution is warranted (especially given the larger dimensionality of g_3). For the clustering amplitude σ_8 , we find similar improvement when combining g_2 with any other statistic, with a slight preference for g_3 in the SBI analysis (or a significant one for the Fisher forecast). This matches the above predictions.

Our conclusion from this exercise is the following: if one wishes to constrain the Universe’s clustering amplitude (a key target of modern-day cosmology), the addition of P_2 or H_V into cosmological analyses provides an excellent route (cf. Ref. [77]), and contains similar information to g_3 . Importantly, the alternative statistics are of much lower dimension than g_3 and P_2 is much less computationally expensive to measure (requiring ~ 5 CPU minutes per simulation, instead of ~ 1 CPU hour for g_3 or H_V). If one performs the analysis using a combination of pair-connectedness functions with different values of η , the

results may be stronger still. While this analysis is necessarily simplistic and limited to comparatively large scales (due to the nature of the simulation suite), it nevertheless suggests that the pair-connectedness function is a new statistic of significant potential, and could carry important information also on small scales. In contrast to correlators such as g_2 and g_3 [145], this is difficult to model analytically even at large r , due to its inherent dependence on short-scale physics including the connection between galaxies and dark matter. For this reason it will likely prove useful to adopt a simulation-based methodology to analyze P_2 , such as the SBI techniques discussed above, and marginalize over parameters controlling galaxy formation. On small scales, a similar approach is required for any statistic, due to the breakdown of perturbative modeling.

VII. SUMMARY

In this work, we have considered the application of the theory of disordered heterogeneous media and statistical mechanics to cosmology, and of cosmology to the former. By treating the distribution of galaxies in the present-day Universe as a point process, we can analyze the data using techniques developed to characterize heterogeneous media, such as the correlation functions and nearest-neighbor distributions. Furthermore, by augmenting the dataset with some concept of “connectedness” (here defined by circumscribing the galaxies with spheres), we may utilize various clustering diagnostics and pair-connectedness functions, which encode a different subset of the information present

within the distribution and additionally allows the percolation properties to be determined. Such a framework (a) provides a novel method for understanding the galaxy distribution, whose importance will only grow in the next decade with the plurality of upcoming telescopes, and (b) demonstrates the applicability of heterogeneous media and statistical mechanical techniques in a very different regime to that usually explored.

Our main conclusions are the following.

- (i) The galaxy distribution exhibits very different physical properties to those of conventional materials, leading to distinct signatures in a wide variety of clustering and correlation descriptors. On the largest scales, the system approaches hyperuniformity, while on the smallest, it becomes almost antihyperuniform and strongly inhomogeneous; this dichotomy arises from the fact that the minimum separation between galaxies is much smaller than the mean interparticle distance, with localized groups of galaxies separated by vast cosmic distance.
- (ii) Physically, the cosmological system has two peculiarities: (a) although we treat the galaxies as point objects, they have some physical scale in practice, and cannot overlap, and (b) the distribution carries the signatures of a large-scale stochastic background that modulates the quasi-Poissonian distribution; this is sourced by early Universe physics and gravitational evolution,
- (iii) The galaxy pair correlation function shows this behavior clearly, with the expected hyperuniform tail appearing only at gargantuan scales ($r \gtrsim 200h^{-1}$ Mpc), and with a sharp peak at the mean pairwise particle separation of $r \approx 8h^{-1}$ Mpc. This scale separation induces a large number variance, which is unusually super-Poissonian on small scales, yet sub-Poissonian on the largest. These results are consistent with the order metric τ , which we determine for the galaxy sample for the first time: its value ($\tau = 4.85$) implies that the system is strongly correlated and disordered. The nearest-neighbor functions are again consistent with this picture, with considerably extended tails, a lack of particle pairs below some critical galaxy size, and much enhanced variance relative to the Poissonian case (and most other common scenarios).
- (iv) Analyses of pair-connectedness functions, mean particle numbers, and sample-spanning clusters indicate that the galaxy sample percolates at significantly lower reduced densities than corresponding Poisson realizations. For the fiducial galaxy simulations, finite-scaling analysis gives $\eta_c = 0.25$ in the former case compared to $\eta_c = 0.34$ in the latter, a difference which is amplified by increasing the sample density. This is again supported by the above

evidence: the scale separation is a consequence of the extra small-scale clustering in the galaxy distribution which leads to faster percolation. Both scenarios appear to have the same critical exponents and fractal dimensions, implying that they live in the same universality class, despite very different physics operating.

- (v) The pair-connectedness function is a conceptually straightforward and easy-to-measure statistic that carries useful and accessible large-scale information about the underlying physical parameters of the Universe, and can be trivially extended to small scales. This could enhance the cosmological utility of future galaxy surveys, in combination with conventional techniques. Using simulation-based analysis techniques, we forecast that constraints on amplitude of clustering improve by a factor of ≈ 5 (or ≈ 25 in terms of survey volume) when performing inference using the large-scale pair-connectedness and pair correlation functions as opposed to the pair correlation function alone (which is standard in cosmology). This provides a useful alternative to the three-particle correlation function g_3 , which is of significantly lower dimension and much faster to model, and is shown to be a resummation of correlation functions of all order. Unlike the large-scale three-particle function, it seems unlikely that $P_2(r)$ can be modeled analytically; simulation-based treatments will likely be required in this case.

The galaxy samples used in this work are purposefully simplified, in order to provide a proof-of-concept study capturing the essential physical attributes of the cosmological setup. More work is needed before the statistics can be applied to real data, and will require the following: (a) higher resolution simulations containing more particles, allowing smaller scales to be probed, (b) inclusion of real galaxies in the simulations, rather than dark matter halos, and the associated physical uncertainties with their formation [146,147], (c) anisotropic distortions in the Universe created by transforming from redshifts to physical coordinates [148], and (d) inhomogeneities in the field induced by observational effects, such as the limited field of view of the telescope. However, all of these complexities have been overcome a number of times before for other statistics (such as the correlation functions [102]) and we expect can be similarly surmounted in this case. Furthermore, it is important to characterize how the statistics depend on the galaxy sample: for the correlation functions, this is well understood (and encapsulated by “bias parameters,” which depend on galaxy mass and luminosity), but should be explored further for nearest-neighbor and pair-connectedness functions, as well as the percolation threshold.

Finally, we consider the broader extensions of this work. Although we have restricted our gaze to galaxy

distributions, this is far from being the only stochastic distribution in the Universe. One additional application could include a more principled treatment of cosmic voids [10,11]: these are low-density regions in the galaxy distribution that form a partition of the space, and could be described by the same mathematics as that invoked for percolation. Even more relevant is the distribution of “bubbles” of ionized gas around the first galaxies [149,150]. The growth of such bubbles likely led to the Universe’s reionization approximately one billion years after the big bang, the time of which is set by percolation itself. Finally, we note that there are a wealth of techniques from the theory of disordered heterogeneous media that have not been considered in this work. It would be interesting to consider the utility of the various descriptors in the context of “simulated annealing” [91,151], to understand the extent to which any statistic can capture the full complexities of the field, though we caution that conventional approaches will likely need to be modified to account for the peculiarities of the galaxy distribution, in particular its significant scale separation. Further still, we may consider how annealing techniques allow us to recover “effective pair interactions” between individual galaxies [152], and thus learn more about the Universe’s average dynamics.

ACKNOWLEDGMENTS

We thank Jim Peebles, Paul Steinhardt, Robert Scherrer, Arka Banerjee, Tom Abel, and Haina Wang for insightful comments on this manuscript. O. H. E. P. is additionally grateful to Alice Pisani and Will Coulton for useful discussions regarding cosmic voids and Fisher forecasts, respectively. We additionally thank the anonymous referees for an insightful report. O. H. E. P. is supported by a fellowship from the Simons Society, and thanks the Institute for Advanced Study for their hospitality. S. T. thanks the Institute for Advanced Study for their hospitality during his sabbatical leave there. The authors are pleased to acknowledge that the work reported in this paper was substantially performed using the Princeton Research Computing resources at Princeton University, which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology’s Research Computing Division. Additional computations were performed on the Helios cluster at the Institute for Advanced Study, Princeton.

-
- [1] S. Torquato, *Random Heterogeneous Materials: Microstructure and Macroscopic Properties* (Springer-Verlag, New York, 2002).
 - [2] M. Sahimi, *Heterogeneous Materials I: Linear Transport and Optical Properties* (Springer-Verlag, New York, 2003).
 - [3] J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 4th ed. (Academic Press, New York, 2013).

- [4] P. J. E. Peebles, *The Large-Scale Structure of the Universe* (Princeton University Press, Princeton, NJ, 1980).
- [5] W. C. Saslaw, *The Distribution of the Galaxies* (Cambridge University Press, Cambridge, England, 2000).
- [6] A. Gabrielli, F. S. Labini, M. Joyce, and L. Pietronero, *Statistical Physics for Cosmic Structures* (Springer-Verlag, New York, 2005).
- [7] A. Aghamousa *et al.* (DESI Collaboration), *The DESI Experiment Part I: Science, Targeting, and Survey Design*, arXiv:1611.00036.
- [8] R. Laureijs, J. Amiaux, S. Arduini, J. L. Auguères, J. Brinchmann, R. Cole *et al.*, *Euclid Definition Study Report*, arXiv:1110.3193.
- [9] F. Bernardeau, S. Colombi, E. Gaztanaga, and R. Scoccimarro, *Large Scale Structure of the Universe and Cosmological Perturbation Theory*, *Phys. Rep.* **367**, 1 (2002).
- [10] R. K. Sheth and R. van de Weygaert, *A Hierarchy of Voids: Much Ado about Nothing*, *Mon. Not. R. Astron. Soc.* **350**, 517 (2004).
- [11] A. Pisani *et al.*, *Cosmic Voids: A Novel Probe to Shed Light on Our Universe*, arXiv:1903.05161.
- [12] R. K. Sheth, *The Halo-Model Description of Marked Statistics*, *Mon. Not. R. Astron. Soc.* **364**, 796 (2005).
- [13] M. White, *A Marked Correlation Function for Constraining Modified Gravity Models*, *J. Cosmol. Astropart. Phys.* **11** (2016) 057.
- [14] E. Massara, F. Villaescusa-Navarro, S. Ho, N. Dalal, and D. N. Spergel, *Using the Marked Power Spectrum to Detect the Signature of Neutrinos in Large-Scale Structure*, *Phys. Rev. Lett.* **126**, 011301 (2021).
- [15] O. H. E. Philcox, E. Massara, and D. N. Spergel, *What Does the Marked Power Spectrum Measure? Insights from Perturbation Theory*, *Phys. Rev. D* **102**, 043516 (2020).
- [16] D. H. Weinberg, *Reconstructing Primordial Density Fluctuations. I—Method*, *Mon. Not. R. Astron. Soc.* **254**, 315 (1992).
- [17] M. C. Neyrinck, I. Szapudi, and A. S. Szalay, *Rejuvenating the Matter Power Spectrum: Restoring Information with a Logarithmic Density Mapping*, *Astrophys. J. Lett.* **698**, L90 (2009).
- [18] X. Wang, M. Neyrinck, I. Szapudi, A. Szalay, X. Chen, J. Lesgourgues, A. Riotto, and M. Sloth, *Perturbation Theory of the Cosmological Log-Density Field*, *Astrophys. J.* **735**, 32 (2011).
- [19] H. Rubira and R. Voivodic, *The Effective Field Theory and Perturbative Analysis for Log-Density Fields*, *J. Cosmol. Astropart. Phys.* **03** (2021) 070.
- [20] D. J. Eisenstein, H.-j. Seo, E. Sirko, and D. Spergel, *Improving Cosmological Distance Measurements by Reconstruction of the Baryon Acoustic Peak*, *Astrophys. J.* **664**, 675 (2007).
- [21] F. Elsner, F. Schmidt, J. Jasche, G. Lavaux, and N.-M. Nguyen, *Cosmology Inference from a Biased Density Field Using the EFT-Based Likelihood*, *J. Cosmol. Astropart. Phys.* **01** (2020) 029.
- [22] G. Cabass and F. Schmidt, *The EFT Likelihood for Large-Scale Structure*, *J. Cosmol. Astropart. Phys.* **04** (2020) 042.

- [23] F. Schmidt, *Sigma-Eight at the Percent Level: The EFT Likelihood in Real Space*, *J. Cosmol. Astropart. Phys.* **04** (2021) 032.
- [24] I. Gott, J. Richard, D. H. Weinberg, and A. L. Melott, *A Quantitative Approach to the Topology of Large-Scale Structure*, *Astrophys. J.* **319**, 1 (1987).
- [25] T. Matsubara, *Analytic Expression of the Genus in Weakly Non-Gaussian Field Induced by Gravity*, *Astrophys. J. Lett.* **434**, L43 (1994).
- [26] T. Matsubara, *Nonlinear Evolution of Topology of Large Scale Structure, in Cosmological Constant and the Evolution of the Universe*, edited by K. Sato, T. Sugihara, and N. Sugiyama (1996), p. 45, <https://adsabs.harvard.edu/full/2003MNRAS.344..761M>.
- [27] J. Schmalzing, *Minkowski Functionals in Cosmology: An Overview*, in *Proceedings of the 12th Potsdam Cosmology Workshop: Large Scale Structure: Tracks and Traces: International Workshop, 1997* (World Scientific, Singapore, 1998), pp. 195,196, <https://arxiv.org/abs/astro-ph/9508154>.
- [28] C. Hikage, J. Schmalzing, T. Buchert, Y. Suto, I. Kayo, A. Taruya, M. S. Vogeley, F. Hoyle, J. Richard Gott, and J. Brinkmann (SDSS Collaboration), *Minkowski Functionals of SDSS Galaxies. I. Analysis of Excursion Sets*, *Publ. Astron. Soc. Jpn.* **55**, 911 (2003).
- [29] P. Parihar, M. S. Vogeley, I. Gott, J. Richard, Y.-Y. Choi, J. Kim, S. S. Kim, R. Speare, J. R. Brownstein, and J. Brinkmann, *A Topological Analysis of Large-Scale Structure, Studied Using the CMASS Sample of SDSS-III*, *Astrophys. J.* **796**, 86 (2014).
- [30] M. Biagetti, J. Calles, L. Castiblanco, A. Cole, and J. Noreña, *Fisher Forecasts for Primordial Non-Gaussianity from Persistent Homology*, *J. Cosmol. Astropart. Phys.* **10** (2022) 002.
- [31] R. Meester, R. Roy, and A. Sarkar, *Nonuniversality and Continuity of the Critical Covered Volume Fraction in Continuum Percolation*, *J. Stat. Phys.* **75**, 123 (1994).
- [32] K. R. Mecke, T. Buchert, and H. Wagner, *Robust Morphological Measures for Large-Scale Structure in the Universe*, *Astron. Astrophys.* **288**, 697 (1994).
- [33] T. Sousbie, C. Pichon, S. Colombi, D. Novikov, and D. Pogosyan, *The Three Dimensional Skeleton: Tracing the Filamentary Structure of the Universe*, *Mon. Not. R. Astron. Soc.* **383**, 1655 (2008).
- [34] Because of the conversion of cosmological redshifts into distances, the observed galaxy density is not isotropic and but distorted along the line of sight to the galaxy survey \hat{n} . As such, the pair correlation function depends on an additional angle, i.e., $g_2(r) \rightarrow g_2(r, \hat{r} \cdot \hat{n})$. We neglect this dependence in this work, but note that it will be important when the statistics discussed herein are applied to observational data.
- [35] S. Torquato and F. H. Stillinger, *Local Density Fluctuations, Hyperuniform Systems, and Order Metrics*, *Phys. Rev. E* **68**, 041113 (2003).
- [36] S. Torquato, *Hyperuniform States of Matter*, *Phys. Rep.* **745**, 1 (2018).
- [37] A. Gabrielli, M. Joyce, and F. S. Labini, *Glass-like Universe: Real-Space Correlation Properties of Standard Cosmological Models*, *Phys. Rev. D* **65**, 083523 (2002).
- [38] S. Torquato and F. H. Stillinger, *New Conjectural Lower Bounds on the Optimal Density of Sphere Packings*, *Exp. Math.* **15**, 307 (2006).
- [39] C. E. Zachary and S. Torquato, *Hyperuniformity in Point Patterns and Two-Phase Heterogeneous Media*, *J. Stat. Mech.* (2009) P12015.
- [40] C. Lin, P. J. Steinhardt, and S. Torquato, *Hyperuniformity Variation with Quasicrystal Local Isomorphism Class*, *J. Phys. Condens. Matter* **29**, 204003 (2017).
- [41] E. C. Oğuz, J. E. S. Socolar, P. J. Steinhardt, and S. Torquato, *Hyperuniformity of Quasicrystals*, *Phys. Rev. B* **95**, 054119 (2017).
- [42] A. Gabrielli, *Point Processes and Stochastic Displacement Fields*, *Phys. Rev. E* **70**, 066131 (2004).
- [43] A. Gabrielli and S. Torquato, *Voronoi and Void Statistics for Superhomogeneous Point Processes*, *Phys. Rev. E* **70**, 041105 (2004).
- [44] A. Gabrielli, M. Joyce, and S. Torquato, *Tilings of Space and Superhomogeneous Point Processes*, *Phys. Rev. E* **77**, 031125 (2008).
- [45] J. Kim and S. Torquato, *Effect of Imperfections on the Hyperuniformity of Many-Body Systems*, *Phys. Rev. B* **97**, 054105 (2018).
- [46] O. U. Uche, F. H. Stillinger, and S. Torquato, *Constraints on Collective Density Variables: Two Dimensions*, *Phys. Rev. E* **70**, 046122 (2004).
- [47] S. Torquato, G. Zhang, and F. H. Stillinger, *Ensemble Theory for Stealthy Hyperuniform Disordered Ground States*, *Phys. Rev. X* **5**, 021020 (2015).
- [48] G. Zhang, F. H. Stillinger, and S. Torquato, *The Perfect Glass Paradigm: Disordered Hyperuniform Glasses Down to Absolute Zero*, *Sci. Rep.* **6**, 36963 (2016).
- [49] Q.-L. Lei and R. Ni, *Hydrodynamics of Random-Organizing Hyperuniform Fluids*, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22983 (2019).
- [50] S. Torquato, G. Zhang, and M. de Courcy-Ireland, *Hidden Multiscale Order in the Primes*, *J. Phys. A* **52**, 135002 (2019).
- [51] A. Donev, F. H. Stillinger, and S. Torquato, *Unexpected Density Fluctuations in Disordered Jammed Hard-Sphere Packings*, *Phys. Rev. Lett.* **95**, 090604 (2005).
- [52] C. E. Zachary, Y. Jiao, and S. Torquato, *Hyperuniform Long-Range Correlations Are a Signature of Disordered Jammed Hard-Particle Packings*, *Phys. Rev. Lett.* **106**, 178001 (2011).
- [53] Y. Jiao and S. Torquato, *Maximally Random Jammed Packings of Platonic Solids: Hyperuniform Long-Range Correlations and Isostaticity*, *Phys. Rev. E* **84**, 041309 (2011).
- [54] S. Atkinson, G. Zhang, A. B. Hopkins, and S. Torquato, *Critical Slowing Down and Hyperuniformity on Approach to Jamming*, *Phys. Rev. E* **94**, 012902 (2016).
- [55] R. P. Feynman and M. Cohen, *Energy Spectrum of the Excitations in Liquid Helium*, *Phys. Rev.* **102**, 1189 (1956).
- [56] L. Reatto and G. V. Chester, *Phonons and the Properties of a Bose System*, *Phys. Rev.* **155**, 88 (1967).
- [57] S. Torquato, A. Scardicchio, and C. E. Zachary, *Point Processes in Arbitrary Dimension from Fermionic Gases, Random Matrix Theory, and Number Theory*, *J. Stat. Mech.* (2008) P11019.
- [58] C. E. Zachary and S. Torquato, *Anomalous Local Coordination, Density Fluctuations, and Void Statistics in*

- Disordered Hyperuniform Many-Particle Ground States*, *Phys. Rev. E* **83**, 051133 (2011).
- [59] D. Hexner and D. Levine, *Hyperuniformity of Critical Absorbing States*, *Phys. Rev. Lett.* **114**, 110602 (2015).
- [60] S. Torquato and J. Kim, *Nonlocal Effective Electromagnetic Wave Characteristics of Composite Media: Beyond the Quasistatic Regime*, *Phys. Rev. X* **11**, 021002 (2021).
- [61] B. Widom, *Equation of State in the Neighborhood of the Critical Point*, *J. Chem. Phys.* **43**, 3898 (1965).
- [62] L. P. Kadanoff, *Scaling Laws for Ising Models Near T_c* , *Physics* **2**, 263 (1966).
- [63] M. E. Fisher, *The Theory of Equilibrium Critical Phenomena*, *Rep. Prog. Phys.* **30**, 615 (1967).
- [64] K. G. Wilson and J. Kogut, *The Renormalization Group and the ϵ Expansion*, *Phys. Rep.* **12**, 75 (1974).
- [65] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman, *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* (Oxford University Press, Oxford, England, 1992).
- [66] E. C. Oğuz, J. E. S. Socolar, P. J. Steinhardt, and S. Torquato, *Hyperuniformity and Anti-Hyperuniformity in One-Dimensional Substitution Tilings*, *Acta Crystallogr. Sect. A* **A75**, 3 (2019).
- [67] S. Torquato, *Perspective: Basic Understanding of Condensed Phases of Matter via Packing Models*, *J. Chem. Phys.* **149**, 020901 (2018).
- [68] A. R. Kansal, S. Torquato, and F. H. Stillinger, *Diversity of Order and Densities in Jammed Hard-Particle Packings*, *Phys. Rev. E* **66**, 041109 (2002).
- [69] S. Atkinson, F. H. Stillinger, and S. Torquato, *Static Structural Signatures of Nearly Jammed Disordered and Ordered Hard-Sphere Packings: Direct Correlation Function*, *Phys. Rev. E* **94**, 032902 (2016).
- [70] M. A. Klatt, J. Lovrić, D. Chen, S. C. Kapfer, F. M. Schaller, P. W. A. Schönhofer, and S. Torquato, *Universal Hidden Order in Amorphous Cellular Geometries*, *Nat. Commun.* **10**, 811 (2019).
- [71] F. Martelli, S. Torquato, N. Giovambattista, and R. Car, *Large-Scale Structure and Hyperuniformity of Amorphous Ices*, *Phys. Rev. Lett.* **119**, 136002 (2017).
- [72] E. Lomba, J.-J. Weis, and S. Torquato, *Disordered Hyperuniformity in Two-Component Non-Additive Hard Disk Plasmas*, *Phys. Rev. E* **96**, 062126 (2017).
- [73] S. Torquato, B. Lu, and J. Rubinstein, *Nearest-Neighbor Distribution Functions in Many-Body Systems*, *Phys. Rev. A* **41**, 2059 (1990).
- [74] B. S. Ryden and E. L. Turner, *A Statistical Comparison of Voids in the Galaxy Distribution and n -Body Simulations*, *Astrophys. J. Lett.* **287**, L59 (1984).
- [75] S. D. M. White, *The Hierarchy of Correlation Functions and Its Relation to Other Measures of Galaxy Clustering*, *Mon. Not. R. Astron. Soc.* **186**, 145 (1979).
- [76] R. Balian and R. Schaeffer, *Scale-Invariant Matter Distribution in the Universe*, *Astron. Astrophys.* **220**, 1 (1989).
- [77] A. Banerjee and T. Abel, *Nearest Neighbour Distributions: New Statistical Measures for Cosmological Clustering*, *Mon. Not. R. Astron. Soc.* **500**, 5479 (2021).
- [78] A. Banerjee and T. Abel, *Cosmological Cross-Correlations and Nearest Neighbour Distributions*, *Mon. Not. R. Astron. Soc.* **504**, 2911 (2021).
- [79] A. Banerjee, N. Kokron, and T. Abel, *Modelling Nearest Neighbour Distributions of Biased Tracers Using Hybrid Effective Field Theory*, *Mon. Not. R. Astron. Soc.* **511**, 2765 (2022).
- [80] Y. Wang, A. Banerjee, and T. Abel, *Detection of Spatial Clustering in the 1000 Richest SDSS DR8 redMaPPer Clusters with Nearest Neighbor Distributions*, *Mon. Not. R. Astron. Soc.* **514**, 3828 (2022).
- [81] D. J. Vezzetti, *A New Derivation of Some Fluctuation Theorems in Statistical Mechanics*, *J. Math. Phys.* **16**, 31 (1975).
- [82] R. M. Ziff, *On the Bulk Distribution Functions and Fluctuation Theorems*, *J. Math. Phys.* **18**, 1825 (1977).
- [83] T. M. Truskett, S. Torquato, and P. G. Debenedetti, *Density Fluctuations in Many-Body Systems*, *Phys. Rev. E* **58**, 7639 (1998).
- [84] S. Torquato, J. Kim, and M. A. Klatt, *Local Number Fluctuations in Hyperuniform and Nonhyperuniform Systems: Higher-Order Moments and Distribution Functions*, *Phys. Rev. X* **11**, 021028 (2021).
- [85] S. Torquato, *Reformulation of the Covering and Quantizer Problems as Ground States of Interacting Particles*, *Phys. Rev. E* **82**, 056109 (2010).
- [86] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups* (Springer-Verlag, New York, 1998).
- [87] A. Pisani, E. Massara, D. N. Spergel, D. Alonso, T. Baker, Y.-C. Cai *et al.*, *Cosmic Voids: A Novel Probe to Shed Light on Our Universe*, *Bull. Am. Astron. Soc.* **51**, 40 (2019).
- [88] A. Coniglio, U. D. Angelis, and A. Forlani, *Pair Connectedness and Cluster Size*, *J. Phys. A* **10**, 1123 (1977).
- [89] G. Stell, *Exact Equation for the Pair-Connectedness Function*, *J. Phys. A* **17**, L855 (1984).
- [90] S. Torquato, J. Beasley, and Y. Chiew, *Two-Point Cluster Function for Continuum Percolation*, *J. Chem. Phys.* **88**, 6540 (1988).
- [91] Y. Jiao, F. H. Stillinger, and S. Torquato, *A Superior Descriptor of Random Textures and Its Predictive Capacity*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 17634 (2009).
- [92] S. Torquato, *Effect of Dimensionality on the Continuum Percolation of Overlapping Hyperspheres and Hypercubes*, *J. Chem. Phys.* **136**, 054106 (2012).
- [93] Formally, this can be defined as $P_2(r, \eta) \equiv \mathbb{E}[\hat{\rho}(\mathbf{x})\hat{\rho}(\mathbf{x} + \mathbf{r})\hat{\Phi}(\mathbf{x}, \mathbf{x} + \mathbf{r}, \eta)]$, for clustering function

$$\Phi(\mathbf{x}, \mathbf{x} + \mathbf{r}, \eta) = \begin{cases} 1 & \text{if } \exists \Gamma: [0, 1] \rightarrow \mathbb{R}^3 \text{ s.t. } \{\Gamma(0) = \mathbf{x}, \Gamma(1) = \mathbf{x} + \mathbf{r}, \int_0^1 d\gamma \varphi_D(\Gamma(\gamma)) \geq \ell(\Gamma)\} \\ 0 & \text{else,} \end{cases}$$

where we consider all paths Γ connecting \mathbf{x} and $\mathbf{x} + \mathbf{r}$ for which the integral of $\varphi(\mathbf{r})$ [defined as the particle phase; i.e., $\varphi(\mathbf{x}) = \Theta_H[\int d\mathbf{y} \hat{\rho}(\mathbf{y}) \Theta_H(D - |\mathbf{x} - \mathbf{y}|)]$ for Heaviside Θ_H] is at least the line length $\ell(\Gamma)$, i.e., those passing only through connected regions.

- [94] J.K. Percus and G.J. Yevick, *Analysis of Classical Statistical Mechanics by Means of Collective Coordinates*, *Phys. Rev.* **110**, 1 (1958).
- [95] H. Hadwiger, *Über Treffanzahlen bei Translationsgleichen Eikörpern*, *Arch. Math.* **8**, 212 (1957).
- [96] S.B. Lee and S. Torquato, *Monte Carlo Study of Correlated Continuum Percolation: Universality and Percolation Thresholds*, *Phys. Rev. A* **41**, 5338 (1990).
- [97] S.B. Lee and S. Torquato, *Pair Connectedness and Mean Cluster Size for Continuum-Percolation Models: Computer-Simulation Results*, *J. Chem. Phys.* **89**, 6427 (1988).
- [98] A. Gabrielli, F. Sylos Labini, M. Joyce, and L. Pietronero, *Statistical Physics for Cosmic Structures* (Springer-Verlag, New York, 2005).
- [99] Y. Baryshev and P. Teerikorpi, *Fractal Approach to Large-Scale Galaxy Distribution*, [arXiv:astro-ph/0505185](https://arxiv.org/abs/astro-ph/0505185).
- [100] M. Tegmark, A.J.S. Hamilton, M.A. Strauss, M.S. Vogeley, and A.S. Szalay, *Measuring the Galaxy Power Spectrum with Future Redshift Surveys*, *Astrophys. J.* **499**, 555 (1998).
- [101] S. Alam, M. Ata, S. Bailey, F. Beutler, D. Bizyaev, J.A. Blazek *et al.*, *The Clustering of Galaxies in the Completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological Analysis of the DR12 Galaxy Sample*, *Mon. Not. R. Astron. Soc.* **470**, 2617 (2017).
- [102] O.H.E. Philcox and M.M. Ivanov, *BOSS DR12 Full-Shape Cosmology: Λ CDM Constraints from the Large-Scale Galaxy Power Spectrum and Bispectrum Monopole*, *Phys. Rev. D* **105**, 043517 (2002).
- [103] D. Baumann, A. Nicolis, L. Senatore, and M. Zaldarriaga, *Cosmological Non-Linearities as an Effective Fluid*, *J. Cosmol. Astropart. Phys.* **07** (2012) 051.
- [104] F. Villaescusa-Navarro *et al.*, *The Quijote Simulations*, *Astrophys. J. Suppl. Ser.* **250**, 2 (2020).
- [105] Z. Zheng, A.A. Berlind, D.H. Weinberg, A.J. Benson, C.M. Baugh, S. Cole, Romeel Dave, C.S. Frenk, N. Katz, and C.G. Lacey, *Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies*, *Astrophys. J.* **633**, 791 (2005).
- [106] M. Sinha and L.H. Garrison, *Corrfunc—A Suite of Blazing Fast Correlation Functions on the CPU*, *Mon. Not. R. Astron. Soc.* **491**, 3022 (2020).
- [107] D.J. Eisenstein and W. Hu, *Baryonic Features in the Matter Transfer Function*, *Astrophys. J.* **496**, 605 (1998).
- [108] N. Hand, Y. Feng, F. Beutler, Y. Li, C. Modi, U. Seljak *et al.*, *nbodykit: An Open-Source, Massively Parallel Toolkit for Large-Scale Structure*, *Astrophys. J.* **156**, 160 (2018).
- [109] S. Torquato and F.H. Stillinger, *Local Density Fluctuations, Hyperuniformity, and Order Metrics*, *Phys. Rev. E* **68**, 041113 (2003).
- [110] S. Torquato, *Hyperuniform States of Matter*, *Phys. Rep.* **745**, 1 (2018).
- [111] B. Widom, *Random Sequential Addition of Hard Spheres to a Volume*, *J. Chem. Phys.* **44**, 3888 (1966).
- [112] S. Torquato, O.U. Uche, and F.H. Stillinger, *Random Sequential Addition of Hard Spheres in High Euclidean Dimensions*, *Phys. Rev. E* **74**, 061308 (2006).
- [113] M. Davis, G. Efstathiou, C.S. Frenk, and S.D.M. White, *The Evolution of Large Scale Structure in a Universe Dominated by Cold Dark Matter*, *Astrophys. J.* **292**, 371 (1985).
- [114] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 2018).
- [115] A.J.S. Hamilton, *Uncorrelated Modes of the Non-Linear Power Spectrum*, *Mon. Not. R. Astron. Soc.* **312**, 257 (2000).
- [116] G. Stell, *Exact Equation for the Pair-Connectedness Function*, *J. Phys. A* **17**, L855 (1984).
- [117] W. Xu, Z. Zhu, Y. Jiang, and Y. Jiao, *Continuum Percolation of Congruent Overlapping Polyhedral Particles: Finite-Size-Scaling Analysis and Renormalization-Group Method*, *Phys. Rev. E* **99**, 032107 (2019).
- [118] M.E. Fisher, *Critical Probabilities for Cluster Size and Percolation Problems*, *J. Math. Phys.* **2**, 620 (1961).
- [119] S.B. Lee and S. Torquato, *Monte-Carlo Study of Correlated Continuum Percolation: Universality and Percolation Thresholds*, *Phys. Rev. A* **41**, 5338 (1990).
- [120] R. Balian and R. Schaeffer, *Galaxies: Fractal Dimensions, Counts in Cells, and Correlations*, *Astrophys. J. Lett.* **335**, L43 (1988).
- [121] N. Aghanim *et al.* (Planck Collaboration), *Planck 2018 Results. VI. Cosmological Parameters*, *Astron. Astrophys.* **641**, A6 (2020).
- [122] G. Cabass, *The EFT Likelihood for Large-Scale Structure in Redshift Space*, *J. Cosmol. Astropart. Phys.* **01** (2021) 067.
- [123] G. Cabass and F. Schmidt, *The EFT Likelihood for Large-Scale Structure*, *J. Cosmol. Astropart. Phys.* **04** (2020) 042.
- [124] F. Schmidt, G. Cabass, J. Jasche, and G. Lavaux, *Unbiased Cosmology Inference from Biased Tracers Using the EFT Likelihood*, *J. Cosmol. Astropart. Phys.* **11** (2020) 008.
- [125] M. Schmittfull, M. Simonović, V. Assassi, and M. Zaldarriaga, *Modeling Biased Tracers at the Field Level*, *Phys. Rev. D* **100**, 043514 (2019).
- [126] C. Modi, F. Lanusse, U. Seljak, D.N. Spergel, and L. Perreault-Levasseur, *CosmicRIM: Reconstructing Early Universe by Combining Differentiable Simulations with Recurrent Inference Machines*, [arXiv:2104.12864](https://arxiv.org/abs/2104.12864).
- [127] U. Seljak, G. Aslanyan, Y. Feng, and C. Modi, *Towards Optimal Extraction of Cosmological Information from Nonlinear Data*, *J. Cosmol. Astropart. Phys.* **12** (2017) 009.
- [128] B. Dai and U. Seljak, *Translation and Rotation Equivariant Normalizing Flow (TRENFlow) for Optimal Cosmological Analysis*, *Mon. Not. R. Astron. Soc.* **516**, 2362 (2022).
- [129] D. Jamieson, Y. Li, S. He, F. Villaescusa-Navarro, S. Ho, R.A. de Oliveira *et al.*, *Simple Lessons from Complex Learning: What a Neural Network Model Learns about Cosmic Structure Formation*, [arXiv:2206.04573](https://arxiv.org/abs/2206.04573).
- [130] D. Jamieson, Y. Li, R.A. de Oliveira, F. Villaescusa-Navarro, S. Ho, and D.N. Spergel, *Field Level Neural Network Emulator for Cosmological N-Body Simulations*, [arXiv:2206.04594](https://arxiv.org/abs/2206.04594).
- [131] R. Alves de Oliveira, Y. Li, F. Villaescusa-Navarro, S. Ho, and D.N. Spergel, *Fast and Accurate Non-Linear Predictions of Universes with Deep Learning*, in *Proceedings of the 34th Conference on Neural Information Processing Systems, 2020*, [arXiv:2012.00240](https://arxiv.org/abs/2012.00240).
- [132] R.A. Fisher and E.J. Russell, *On the Mathematical Foundations of Theoretical Statistics*, *Phil. Trans. R. Soc. A* **222**, 309 (1922).

- [133] G. Papamakarios and I. Murray, *Fast ϵ -Free Inference of Simulation Models with Bayesian Conditional Density Estimation*, in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), Vol. 29, [10.5555/3157096.3157212](#).
- [134] J. Alsing, T. Charnock, S. M. Feeney, and B. D. Wandelt, *Fast Likelihood-Free Cosmology with Neural Density Estimators and Active Learning*, *Mon. Not. R. Astron. Soc.* **488**, 4440 (2019).
- [135] K. Cranmer, J. Brehmer, and G. Louppe, *The Frontier of Simulation-Based Inference*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055 (2020).
- [136] Y. Feng, M.-Y. Chu, U. Seljak, and P. McDonald, *FastPM: A New Scheme for Fast Simulations of Dark Matter and Haloes*, *Mon. Not. R. Astron. Soc.* **463**, 2273 (2016).
- [137] Z. Slepian and D. J. Eisenstein, *Computing the Three-Point Correlation Function of Galaxies in $\mathcal{O}(N^2)$ Time*, *Mon. Not. R. Astron. Soc.* **454**, 4142 (2015).
- [138] J. Hartlap, P. Simon, and P. Schneider, *Why Your Model Parameter Confidences Might Be Too Optimistic. Unbiased Estimation of the Inverse Covariance Matrix*, *Astron. Astrophys.* **464**, 399 (2007).
- [139] A. E. Bayer, A. Banerjee, and U. Seljak, *Beware of Fake ν 's: The Effect of Massive Neutrinos on the Nonlinear Evolution of Cosmic Structure*, *Phys. Rev. D* **105**, 123510 (2022).
- [140] A. Aghamousa *et al.* (DESI Collaboration), *The DESI Experiment Part I: Science, Targeting, and Survey Design*, [arXiv:1611.00036](#).
- [141] M. Takada *et al.* (PFS Team Collaboration), *Extragalactic Science, Cosmology, and Galactic Archaeology with the Subaru Prime Focus Spectrograph*, *Publ. Astron. Soc. Jpn.* **66**, R1 (2014).
- [142] We recall that our H_V void statistic follows a somewhat different definition to the void size function often used in cosmology, and is restricted to comparatively large scales ($r > 20h^{-1}$ Mpc, as for the other statistics), with only spherical voids. This differs from the approach used in several cosmological studies [11,143,144] and explains the reduced utility found herein, and the different correlation properties seen in Fig. 11.
- [143] C. D. Kreisch, A. Pisani, C. Carbone, J. Liu, A. J. Hawken, E. Massara, D. N. Spergel, and B. D. Wandelt, *Massive Neutrinos Leave Fingerprints on Cosmic Voids*, *Mon. Not. R. Astron. Soc.* **488**, 4413 (2019).
- [144] C. D. Kreisch, A. Pisani, F. Villaescusa-Navarro, D. N. Spergel, B. D. Wandelt, N. Hamaus, and A. E. Bayer, *The GIGANTES Dataset: Precision Cosmology from Voids in the Machine Learning Era*, *Astrophys. J.* **935**, 100 (2022).
- [145] M. M. Ivanov, O. H. E. Philcox, T. Nishimichi, M. Simonović, M. Takada, and M. Zaldarriaga, *Precision Analysis of the Redshift-Space Galaxy Bispectrum*, *Phys. Rev. D* **105**, 063512 (2022).
- [146] V. Desjacques, D. Jeong, and F. Schmidt, *Large-Scale Galaxy Bias*, *Phys. Rep.* **733**, 1 (2018).
- [147] R. H. Wechsler and J. L. Tinker, *The Connection between Galaxies and Their Dark Matter Halos*, *Annu. Rev. Astron. Astrophys.* **56**, 435 (2018).
- [148] N. Kaiser, *Clustering in Real Space and in Redshift space*, *Mon. Not. R. Astron. Soc.* **227**, 1 (1987).
- [149] A. Loeb and R. Barkana, *The Reionization of the Universe by the First Stars and Quasars*, *Annu. Rev. Astron. Astrophys.* **39**, 19 (2001).
- [150] K.-G. Lee, R. Cen, J. R. Gott, III, and H. Trac, *The Topology of Cosmological Reionization*, *Astrophys. J.* **675**, 8 (2008).
- [151] C. L. Y. Yeong and S. Torquato, *Reconstructing Random Media*, *Phys. Rev. E* **57**, 495 (1998).
- [152] S. Torquato and H. Wang, *Precise Determination of Pair Interactions from Pair Statistics of Many-Body Systems In and Out of Equilibrium*, *Phys. Rev. E* **106**, 044122 (2022).