Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **491,** 1709–1726 (2020)
Advance Access publication 2019 November 4

doi:10.1093/mnras/stz3075

Beyond two-point statistics: using the minimum spanning tree as a tool for cosmology

Krishna Naidoo ,[1]★ Lorne Whiteway,[1] Elena Massara,[2] Davide Gualdi,[3] Ofer Lahav,[1] Matteo Viel,[4,5,6,7] Héctor Gil-Marín[3,8] and Andreu Font-Ribera[1]

[1]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[2]*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*
[3]*ICC, University of Barcelona, IEEC-UB, Martí i Franquès, 1, E-08028 Barcelona, Spain*
[4]*SISSA – International School for Advanced Studies, Via Bonomea 265, I-34136 Trieste, Italy*
[5]*INAF – Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[6]*INFN – National Institute for Nuclear Physics, via Valerio 2, I-34127 Trieste, Italy*
[7]*IFPU – Institute for Fundamental Physics of the Universe, via Beirut 2, I-34151 Trieste, Italy*
[8]*Institute of Space Studies of Catalonia (IEEC), E-08034 Barcelona, Spain*

Accepted 2019 October 25. Received 2019 October 22; in original form 2019 July 1

ABSTRACT

Cosmological studies of large-scale structure have relied on two-point statistics, not fully exploiting the rich structure of the cosmic web. In this paper we show how to capture some of this cosmic web information by using the minimum spanning tree (MST), for the first time using it to estimate cosmological parameters in simulations. Discrete tracers of dark matter such as galaxies, *N*-body particles or haloes are used as nodes to construct a unique graph, the MST, that traces skeletal structure. We study the dependence of the MST on cosmological parameters using haloes from a suite of COmoving Lagrangian Acceleration (COLA) simulations with a box size of 250 h^{-1} Mpc, varying the amplitude of scalar fluctuations (A_s), matter density (Ω_m), and neutrino mass ($\sum m_\nu$). The power spectrum P and bispectrum B are measured for wavenumbers between 0.125 and 0.5 $h\,\mathrm{Mpc}^{-1}$, while a corresponding lower cut of $\sim$12.6 h^{-1} Mpc is applied to the MST. The constraints from the individual methods are fairly similar but when combined we see improved 1σ constraints of $\sim$ 17 per cent ($\sim$ 12 per cent) on Ω_m and $\sim$ 12 per cent ($\sim$ 10 per cent) on A_s with respect to P ($P + B$) thus showing the MST is providing additional information. The MST can be applied to current and future spectroscopic surveys (BOSS, DESI, Euclid, PSF, WFIRST, and 4MOST) in 3D and photometric surveys (DES and LSST) in tomographic shells to constrain parameters and/or test systematics.

Key words: neutrinos – methods: data analysis – cosmological parameters – large-scale structure of Universe.

1 INTRODUCTION

Over the years, a series of probes have emerged as standard tools for cosmological parameter inference. Surveys of the cosmic microwave background (CMB), large-scale structure (LSS), weak lensing (WL), and distance ladder have dominated our knowledge of cosmological parameters through measurements of the CMB angular power spectra (e.g. Planck Collaboration VI 2018), galaxy clustering (e.g. Loureiro et al. 2019), WL (e.g. Abbott et al. 2018; Hildebrandt et al. 2017), Baryonic Acoustic Oscillation (BAO) from galaxies (e.g. Alam et al. 2017) and Lyman alpha (e.g. de Sainte Agathe et al. 2019), standard candles (e.g. Riess et al. 2016) and, more recently, standard sirens (e.g. Abbott et al. 2017). These techniques are relatively mature, well understood and most importantly, reliable and trusted.

However, many of these techniques (but not all) rely on measuring the two-point correlation function (2PCF) or its Fourier space equivalent, the power spectrum. Studies that include higher order statistics, such as the three-point correlation function (e.g. Slepian et al. 2017) or bispectrum (e.g. Gil-Marín et al. 2017), have already provided interesting constraints on cosmological parameters, demonstrating the need to go beyond the 2PCF. Despite solutions to improve the speed of 2PCF and 3PCF estimators (see Scoccimarro 2015; Slepian & Eisenstein 2016), going beyond the 3PCF is currently computationally intractable. The computational cost of current N-point correlation functions (NPCF) estimators

★ E-mail: krishna.naidoo.11@ucl.ac.uk

scales by $\mathcal{O}(n^N)$; for this reason this information remains to be exploited.

The most attractive reason to explore methods that incorporate higher order statistics is their potential to break existing parameter degeneracies, to provide tighter constraints and to test systematics. Of growing interest to cosmologists is the total mass of the three neutrino species, $\sum m_\nu$. Neutrinos are massless in the standard model of particle physics; however this cannot be the case since neutrinos oscillate (Fukuda et al. 1998; Ahmad et al. 2001). Fortunately, LSS is sensitive to the mass of these elusive particles. As neutrinos are very light, they possess high thermal velocities and dampen structure formation at scales below the free streaming scale (set by when they become non-relativistic). This effect is dependent on $\sum m_\nu$ and although it can be measured, the effect is small and highly degenerate with the matter density (Ω_m) and the variance of density perturbations (e.g. as measured at $8\,h^{-1}$ Mpc (σ_8)). Currently, upper bounds of $\sum m_\nu \lesssim 0.12$–0.23 eV (95 per cent confidence limits) (Palanque-Delabrouille et al. 2015; Planck Collaboration XIII 2016; Alam et al. 2017; Loureiro et al. 2019) have been established from cosmology (specifically CMB and galaxy surveys) whilst the lower bound of $\gtrsim 0.06$ eV is given by neutrino oscillation experiments. Future experiments will be able to go further; in particular experiments such as the *Dark Energy Spectroscopic Instrument* (DESI, DESI Collaboration et al. 2016) are expected to probe below the lower bound of $\sim$0.06 eV, and are expected to make a detection of the neutrino mass (see Font-Ribera et al. 2014). However, this is to be achieved purely by a more precise measurement of the 2PCF, not by the inclusion of extra information.

We know from *N*-body simulations that the universe at late times appears as a cosmic web (Bond, Kofman & Pogosyan 1996). Currently this cosmic web structure is not fully incorporated into the inference of cosmological parameters. In this paper, we turn to graph theory, looking specifically at the minimum spanning tree (MST), to try to capture some of this rich information. The MST was first introduced to astronomy by Barrow, Bhavsar & Sonoda (1985). It has been typically used in cosmology for LSS classification, for example to search for cosmic web features such as filaments (see Bhavsar & Ling 1988; Pearson & Coles 1995; Krzewina & Saslaw 1996; Ueda & Itoh 1997; Coles et al. 1998; Adami & Mazure 1999; Doroshkevich et al. 1999, 2001; Colberg 2007; Balázs et al. 2008; Park & Lee 2009; Adami et al. 2010; Demiański et al. 2011; Durret et al. 2011; Cybulski et al. 2014; Alpaslan et al. 2014; Shim & Lee 2013; Shim, Lee & Li 2014; Shim, Lee & Hoyle 2015; Beuret et al. 2017; Campana, Massaro & Bernieri 2018a,b; Libeskind et al. 2018; Clarke et al. 2019). It has also been used in other contexts such as determining mass segregation in star clusters (Allison et al. 2009) and the generalized dimensionality of data points, fractals and percolation analysis (see Martinez & Jones 1990; van de Weygaert, Jones & Martínez 1992; Bhavsar & Splinter 1996). More recently, the MST was used in particle physics to distinguish between different classes of events in collider experiments (Rainbolt & Schmitt 2017). The MST's strength is in its ability to extract patterns; this is precisely why it has been used to extract cosmic web features (the type of information currently missing from most cosmological studies). The MST's weaknesses are that the statistics cannot be described analytically and that they depend heavily on the density of the tracer. This means any comparison of models via the MST will be dependent on simulations. While this makes parameter inference more challenging, the reliance on simulations is not new; in fact parameter inference through artificial intelligence (AI) and machine learning (ML) will be similarly reliant. Here, the MST may provide a bridge between the traditional 2PCF and AI/ML, allowing us to understand the information being extracted by these AI/ML algorithms.

Our goal in this paper is to understand whether the MST could be a useful tool for cosmological parameter inference for current or future photometric and spectroscopic galaxy redshift surveys. These include the *Baryon Oscillation Spectroscopic Survey*,[1] *Dark Energy Survey*,[2] DESI,[3] *Large Synoptic Survey Telescope*,[4] *Euclid*,[5] *Prime Focus Spectrograph*,[6] *Wide Field Infrared Survey Telescope*,[7] and *4-metre Multi-Object Spectroscopic Telescope*.[8] With this in mind, the paper is organized as follows. In Section 2, we describe the MST construction and statistics and we summarize the suites of simulations used in later sections. In Section 3, we demonstrate that the MST is sensitive to higher order statistics (i.e. beyond two-point). In Section 4, we explore relevant sources of systematics and methods to mitigate them. In addition, we test the sensitivity to redshift space distortions (RSDs). In Section 5, we explore the MST statistics on an unbiased tracer, and try to determine what the MST is actually measuring about the underlining density distribution. Lastly, in Section 6, we compare the MST's constraining power to that of the more traditional power spectrum and bispectrum measurements.

2 METHOD

In this section, we will describe:

(i) Some basic properties of graphs and the MST.
(ii) How the MST is constructed.
(iii) The statistics we measure.
(iv) Techniques for error estimation.
(v) The simulations used in this paper.

In mathematics, a *graph* is a set of *nodes* (points) together with a set of *edges*, where each edge joins two distinct nodes; given any two distinct nodes, there will be either zero or one edge between them. In this paper, all graphs are *undirected* and *weighted* i.e. an edge does not have an orientation, but it does have a (positive) weight (which in this paper will be the distance (defined below) between the nodes that it connects). A *path* is a sequence of nodes in which each consecutive pair of nodes is connected by an edge (and no edge is used twice); a path that returns to its starting point is a *cycle*. If there is an edge (respectively, path) between any two distinct nodes then the graph is *complete* (respectively, *connected*). Given a connected graph G (not necessarily complete), one can discard edges to obtain the MST of G. By definition this new graph is *spanning* (i.e. contains all the nodes of G), is a *tree* (i.e. is connected and contains no cycles) and is *minimal* in that the sum of the edge weights is minimal among all spanning trees. Every connected graph has a (essentially unique) MST.

In this work, we consider sets of points in various spaces, with distance between points defined to be:

(i) In two and three dimensions: Euclidean distance;

[1] http://www.sdss3.org/surveys/boss.php
[2] http://www.darkenergysurvey.org
[3] http://desi.lbl.gov/
[4] https://www.lsst.org/
[5] http://www.euclid-ec.org/
[6] https://pfs.ipmu.jp/index.html
[7] https://wfirst.gsfc.nasa.gov/
[8] https://www.4most.eu/cms/

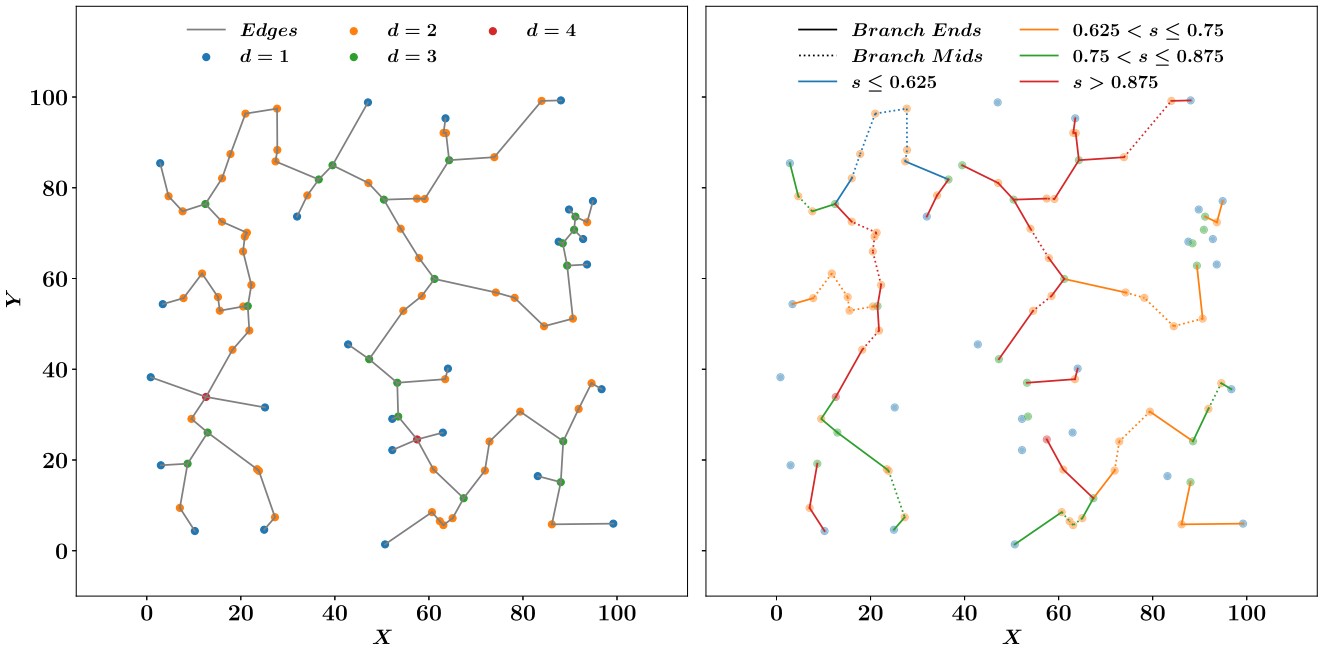

Figure 1. The MST constructed from 100 random points. Left-hand panel: the MST edges are shown. Nodes are colour coded according to their degree, i.e. the number of edges attached to them. Right-hand panel: the MST branches are colour coded according to their branch shape parameter (s). Edges that form branch ends are indicated by solid lines while edges forming the middle of branches (branch mids) are indicated by dotted lines.

(ii) On the sphere (i.e. RA, Dec.): subtended angle;
(iii) Using RA, Dec., redshift: convert redshift to comoving distance (using the fiducial cosmology), then use Euclidean distance.

Given a set of points S we wish to investigate the MST of the complete graph on these points (i.e. there is an edge between every pair of points and all these edges are candidates for inclusion in the MST); we refer to this as the MST of S. See Fig. 1 for an example of such an MST. Now Kruskal's (1956) algorithm (described below) takes as input a connected graph (not necessarily complete) and discards certain edges so as to find its MST. In theory, we should input to this algorithm the complete graph on S. However this is inefficient as the complete graph contains many edges (e.g. between widely separated points) that are very unlikely to appear in the output MST; it is sufficient to input to Kruskal's algorithm a pruned graph that retains only shorter edges.

To this end, we use as input to Kruskal's algorithm the k nearest neighbours graph (kNN), i.e. the graph in which each point has an edge to its kNN. Here k is a free parameter (and should not be confused with the wavenumber used in harmonic analysis). We calculate this graph using the `kneighbours_graph` function from `scikit-learn`.[9] Note that if k is too small then the kNN graph need not be connected (it might consist of several isolated islands); in most cases considered, $k > 10$ ensures that kNN will be connected (but when applying scale cuts (see Section 4.2) a larger k is needed).

We then apply the `scipy minimum_spanning_tree`[10] function, which implements Kruskal's algorithm. This algorithm removes all the edges from the graph, sorts these removed edges by length (shortest to longest), and then sequentially re-inserts them, omitting an edge if its inclusion would create a cycle. This continues until all points are connected into a single tree. The Kruskal algorithm can be shown to scale as $\mathcal{O}(N_{\rm E} \log N_{\rm V})$ (see Cormen et al. 2009, section on Kruskal's algorithm) where $N_{\rm E}$ is the number of edges in the supplied spanning graph and $N_{\rm V}$ is the number of nodes. At most $N_{\rm E} \simeq N_{\rm V}^2$ but this can be greatly reduced by using the kNN graph, which changes the scaling from $\mathcal{O}(n^2 \log n)$, where n is the number of nodes, to $\mathcal{O}(kn \log n)$. Since usually $k \ll n$ this greatly reduces computation time.

We tested the sensitivity to the choice of k by using a graph with 256^3 points (HZ = High σ_8 and zero $\sum m_\nu$ simulations at $z = 0$ explained later in Section 5). We compared the total length of the MST when $k = 50$ (a proxy for $k = \infty$) and found a fractional difference of $\sim 2 \times 10^{-6}$ for $k = 20$, $\sim 2 \times 10^{-7}$ for $k = 30$, and $\sim 3 \times 10^{-8}$ for $k = 40$. It appears that $k = 20$ gives a good balance between computation time and an accurate estimation of the MST, so we use this value except where stated otherwise.

2.1 Statistics from the minimum spanning tree

Any given MST is a complex structure with many interesting features. In this study, we are not interested in these individual features but rather the overall properties and their relation to cosmological parameters. Taking inspiration from Rainbolt & Schmitt (2017) and Krzewina & Saslaw (1996) we measure the probability distribution (i.e. histograms) of the following:

(i) Degree (d): the number of edges attached to each node.
(ii) Edge lengths (l): the length of edges.
(iii) From branches, which are chains of edges connected with intermediary nodes of $d = 2$, we measure:

(a) Branch lengths (b): the sum of edges that make up the branch.
(b) Branch shape (s): the straight line distance between the branch ends divided by the branch length.

[9] http://www.scikit-learn.org
[10] https://scipy.org/

These statistics are displayed in Fig. 1. Of course one could consider other statistics to extract from the MST (see Alpaslan et al. 2014) but we choose to explore these as they have been shown to successfully aid in the classification of particle physics interactions (see Rainbolt & Schmitt 2017). The MST will have a total of $n - 1$ edges (Kruskal 1956), where n is the number of nodes. Since each edge has a node on either end, each edge contributes twice to the total degree of the MST. Hence the expectation value for d will be:

$$\langle d \rangle = \frac{2(n-1)}{n} \simeq 2. \tag{1}$$

By definition the branch shapes satisfies $0 \leq s \leq 1$. Often s is near 1, so to facilitate visual comparison we frequently plot $\sqrt{1-s}$ instead of s. Straighter branches correspond to $\sqrt{1-s}$ closer to zero.

Additionally, it is useful in certain circumstances, particularly when comparing MSTs that contain different number of nodes, to look at the dimensionless parameters of:

(i) $\ln\left(\bar{l}\right)$, where $\bar{l} = l/\langle l \rangle$ and $\langle l \rangle$ is the average edge length.
(ii) $\ln\left(\bar{b}\right)$, where $\bar{b} = b/\langle b \rangle$ and $\langle b \rangle$ is the average branch length.

Comparing the distribution of these dimensionless parameters is only appropriate if the distribution of points is scale-independent. In cosmology, this is not necessarily the case for higher order statistics, so these should be used sparingly.

2.1.1 Computational issues for finding branches

Once the MST is constructed, we know the edge lengths (l) and the indices of the nodes at either end of the edges. These can be trivially used to find the degree (d) of each node and edge end. To find branches, we search for edges joining a $d = 2$ node to a $d \neq 2$ node (i.e. 'branch ends') and edges joining two $d = 2$ nodes (such edges, which form the middle parts of branches, are referred to as 'branch mids'). To find the branches we begin with a branch end, search for a branch mid that is connected to it, and continue to grow the branch until no more branch mids can be added. At this point, we then search for the branch end that finishes it. This is a computationally expensive procedure but can be trivially made faster by dividing the entire tree into sections and running the algorithm on the sections independently. Branches straddling the boundaries will be left incomplete, but can be completed by matching any remaining incomplete branches.

`MiSTree` (Naidoo 2019), the PYTHON package to construct the MST and derive its statistics, is made publicly available.[11]

2.2 Error estimation

Uncertainties for the MST statistics are generated in two ways.

(i) In the cases where many realizations of a data set can be generated easily we will estimate the mean and standard deviation from an ensemble of realizations.

(ii) If only a single realization is available we will use jackknife errors. Here, we divide up our data set into n regions and run the analysis n times, each time removing a single different region from the analysis yielding an output $\theta_{\rm i}$. The errors, $\Delta\theta_{\rm jack}$, are estimated using

$$\Delta\theta_{\rm jack} = \left[\frac{n-1}{n}\sum_{i=1}^{n}\left(\theta_{\rm i} - \bar{\theta}\right)^2\right]^{1/2}, \tag{2}$$

where $\bar{\theta}$ is the average of $\theta_{\rm i}$.

2.3 Simulation summary

We use several simulations suites; these are summarized in Table 1. We discuss these simulations in greater detail in the relevant sections of the paper where they are used.

3 SENSITIVITY OF MST TO COSMIC WEB PATTERNS

3.1 Heuristic argument

There are compelling reasons to believe the MST should be sensitive to cosmic web patterns. Consider how the Kruskal algorithm constructs the MST (see Section 2). An edge is added only if this does not create a cycle; this means that the very construction of the MST requires an awareness of neighbouring edges or more generally the environment each edge inhabits. More generally this means the inclusion of a single edge is not defined solely by the 2PCF but by its local environment. Therefore, we should expect the MST to contain more information than is present in the 2PCF.

3.2 Illustris versus adjusted Lévy flight

Testing whether the MST is sensitive to higher order statistics is rather challenging since at present there are no analytical descriptions of the MST statistics.

To go around this theoretical limitation we instead carry out an analysis similar to that of Hong et al. (2016), comparing the Illustris[12] (Nelson et al. 2015; Vogelsberger et al. 2014) simulations (see Section 3.2.1) to an adjusted Lévy flight (ALF) simulation that is tuned to have almost identical 2PCF but different higher order information.

Lévy flights (Mandelbrot 1982) are random walk simulations where the step size (the distance between one point and the next) is given by a fat-tailed power-law probability distribution function (PDF). This ensures that its 2PCF will follow a power law (see Mandelbrot 1982) similar to that found for galaxies. However, although a standard Lévy flight scheme may be able to replicate the 2PCF at large scales, at small scales, the 2PCF eventually plateaus (see Hong et al. 2016). Since the MST is sensitive to small scales, it is important that the Lévy flight simulation match that of the Illustris sample at small scales. We are able to match the 2PCF of the Illustris sample at all scales using an ALF simulation as explained below.

3.2.1 Illustris galaxy sample

We use the subhalo catalogue of the Illustris-1 snap 100 sample and follow Hong et al. (2016) to include only subhaloes which are large

[11] https://github.com/knaidoo29/mistree

[12] http://www.illustris-project.org

Table 1. A summary of the simulation suites used in this study. For each simulation suite we list its name, the method used to produce it, the point distribution used, and the use to which it is put.

Name	Method	Points	Usage
Illustris	Hydrodynamic	Subhaloes	Testing the sensitivity of the MST to higher order statistics (i.e. beyond two-point)
MICE	*N*-body	Galaxies	Exploring the sensitivity to RSDs
*ν*N-body	*N*-body	Dark matter particles and haloes	Using an unbiased tracer we look to find what the MST is actually measuring
PICOLA	COLA	Haloes	Comparing sensitivity of the MST to traditional methods

and dark-matter-dominated:

$$M_* \geq 10^8\,\mathrm{M}_\odot,$$
$$M_* < 0.63 M_{\mathrm{DM}}, \tag{3}$$

where M_* and M_{DM} are the stellar and dark matter mass of the subhaloes respectively. We will refer to this as the Illustris galaxy sample.

3.2.2 *Adjusted Lévy flight*

We generate an ALF simulation with the same number of 'galaxies' as our Illustris sample and (almost) the same 2PCF. For comparison with Illustris we enforce periodic boundary conditions. The standard Lévy flight has step sizes t with cumulative distribution function (CDF),

$$\mathrm{CDF}(t) = \begin{cases} 0 & \text{for} \quad t < t_0, \\ 1 - \left(\frac{t}{t_0}\right)^{-\alpha} & \text{for} \quad t \geq t_0, \end{cases} \tag{4}$$

where t_0 and α are free parameters. This yields a simulation with a power-law 2PCF of the form $C(t_0, \alpha)t^{3-\alpha}$ at scales larger than t_0 (where $C(t_0, \alpha)$ is a constant determined by the free parameters), below this scale the 2PCF plateaus (see Hong et al. 2016). To have control of the 2PCF below scales of t_0 we introduce an ALF model with the following CDF:

$$\mathrm{CDF}(t) = \begin{cases} 0 & \text{for} \quad t < t_{\mathrm{s}}, \\ \beta \left(\frac{t - t_{\mathrm{s}}}{t_0 - t_{\mathrm{s}}}\right)^{\gamma} & \text{for} \quad t_{\mathrm{s}} \leq t < t_0, \\ (1-\beta)\left[1 - \left(\frac{t}{t_0}\right)^{-\alpha}\right] + \beta & \text{for} \quad t \geq t_0. \end{cases} \tag{5}$$

This introduces three new parameters: β, γ, and t_{s}. Rather than having a step size PDF that jumps from zero to a maximum at t_0, the ALF is constructed to have a slow rise to the maximum at t_0. The second piece of the CDF describes a transfer function that operates between t_{s} and t_0 (where by definition $t_{\mathrm{s}} < t_0$). Here, γ allows us to control the gradient of this rise and β allows us to define the fraction of step sizes below t_0.

3.2.3 *Comparison*

The Illustris sample contains 63 453 galaxies. We create a sample of the same size using an ALF model with parameters $\alpha = 1.5$, $t_0 = 0.325$, $t_{\mathrm{s}} = 0.015$, $\beta = 0.45$, and $\gamma = 1.3$ (where length-scales t_0 and t_{s} are given in h^{-1} Mpc). The two samples have approximately equal 2PCFs down to scales of $0.01\,h^{-1}$ Mpc by construction. The 2PCF was calculated on a single realization of the ALF model with varying β, γ, t_{s}, and t_0 ($\alpha = 1.5$ was kept constant, see Hong et al. 2016). We then chose the parameters that produced the closest match, i.e. by minimizing the sum of difference between the 2PCF in log space. The Illustris and ALF sample show widely different MST statistics (see Fig. 2), thereby demonstrating the sensitivity of the MST to higher order statistics. The bimodal distribution of edge and branch lengths shown in Fig. 2 occurs in over- and underdensities (explored in more detail in Section 5). Note also that we see differences in the shape of branches and the distribution of degrees to a statistically significant level, although these differences are not as striking as the difference in edge and branch length distributions.

4 BOUNDARY EFFECTS AND REDSHIFT SPACE DISTORTIONS

We study possible sources of systematic errors that could affect the MST. In particular we would like to establish to what extent simulations need to replicate survey properties.

4.1 Boundary effects

Galaxy surveys often contain complex survey footprints with regions masked due to stars and varying completeness and it is important to understand how such footprints will affect the MST. Imposing a mask on the data set results in two effects:

(i) Additional edges are included to join nodes near the boundaries. These would have otherwise been joined by nodes outside the boundary in a larger MST.
(ii) New edges are located near the centre whose purpose appears to be to unify the structure as a single spanning tree. In a larger spanning tree, these separated regions would be connected through routes that extend beyond the boundary.

The net result of these effects is to create a slight bias towards longer edges and slightly longer branches. Interestingly, all edges in the larger MST (within the boundary) are present in the smaller MST. This property always holds, as can easily be proven using the 'cycle property' of the MST (see Katriel, Sanders & Träff 2003).

We investigate the effects of a realistic mask by using the BOSS CMASS MD-Patchy mocks North mask (Rodríguez-Torres et al. 2016), which includes masking for bright stars, bad fields, centrepost, and collision priority.[13] In Fig. 3, we demonstrate the effects of this mask on random points placed within the CMASS footprint (with the same density as the CMASS galaxies) with and without a mask. The MST is then calculated on 1000 realizations tomographically (i.e. on the sphere). The degree and branch shape show little change but the distribution of edge lengths show a significant tendency towards longer edges when a mask is used. This is mirrored by a similar effect in the distribution of branch

[13] See http://www.sdss3.org/dr9/algorithms/boss_tiling.php#veto_masks

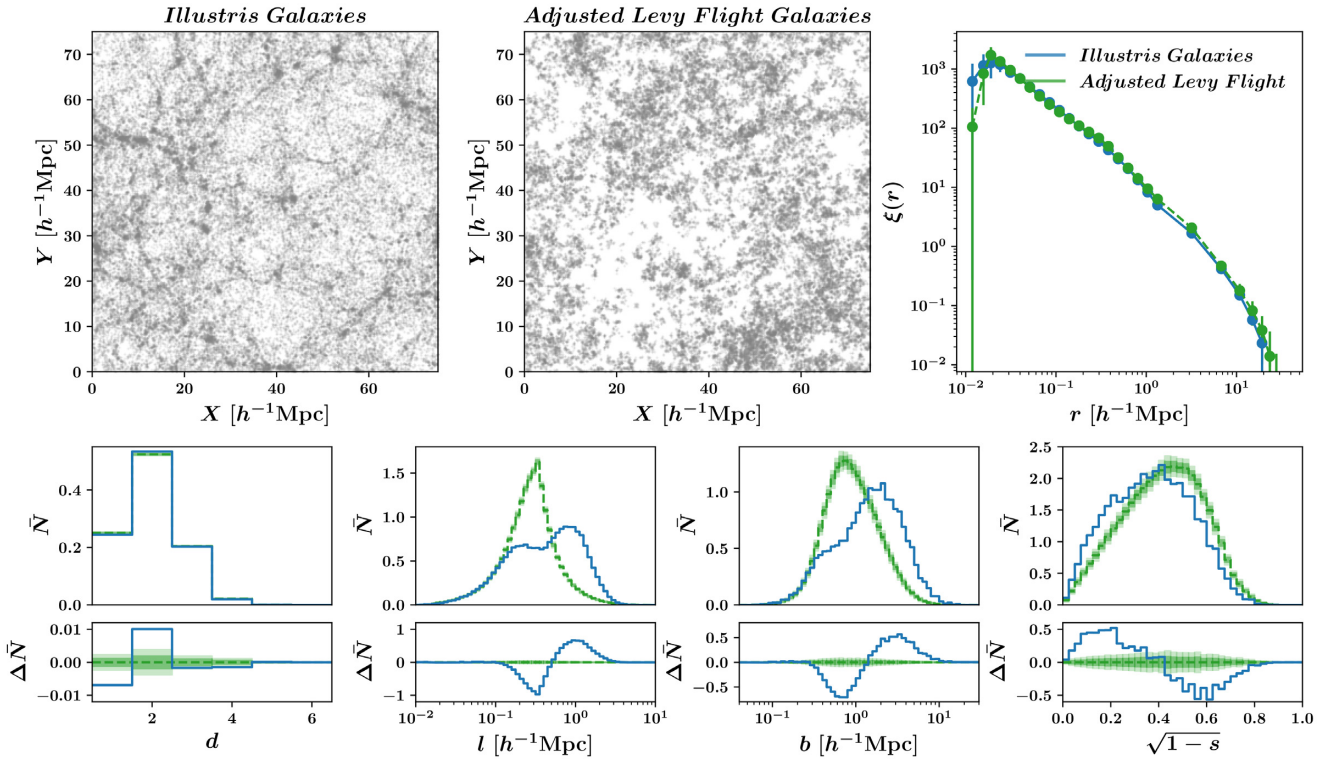

Figure 2. Top panels: the left shows the Illustris galaxy sample and the middle panel shows one realization of the ALF. Visually these two simulations are different in their distribution of galaxies. However they have virtually identical 2PCFs by construction (right-hand panel). Illustris measurements are shown in blue and the mean for 100 realizations of the ALF is shown by the green dashed line; and green envelopes show the 1σ (darker) and 2σ regions. Bottom panels: the histogram distributions of the MST statistics (from left to right): degree (d), edge length (l), branch length (b), and branch shape (s; note we plot the $\sqrt{1-s}$ value instead because the distribution peaks towards 1 and it is easier to see the difference in this projection). The difference between the PDF is displayed in the bottom subplots where zero on the y-axis corresponds to the mean counts for the ALF PDF. The measurements from the MST are significantly different for each of these simulations. In particular, the distributions of edge lengths and branches show some bimodality for the Illustris sample which is not present in the ALF. This demonstrates the sensitivity of the MST to patterns in the cosmic web as the bimodal distribution appears to be driven by void and cluster environments (explored in Section 5.2.2).

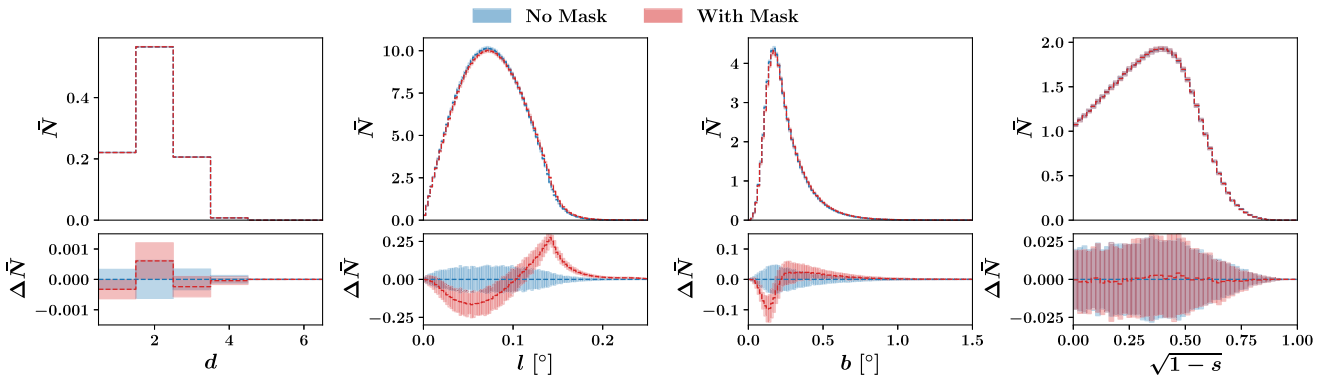

Figure 3. The MST statistics, calculated tomographically on random points placed in the BOSS CMASS North footprint (placed with the same density as the BOSS CMASS galaxies), with (red) and without (blue) using the CMASS mask. We see a significant shift towards longer edges in the MST performed with the mask, with a similar effect seen in the distribution of branch lengths. For the degree and branch shape, the masking has no statistically significant effect.

lengths. This is because the mask eliminates shorter paths, forcing the MST to include longer edges that would (without the mask) have been excluded. This demonstrates that realistic masks with holes do have an impact on the MST and must be included in any future analysis.

4.2 Scale cuts

In cosmology, there is often a need to apply scale cuts in real space. This can occur for a variety of reasons: theoretical uncertainty at small scales both from simulation and from analytic formulae

and also practically from fibre collisions in spectroscopic surveys. For the 2PCF, this is rather simple to mitigate; you simply restrict the domain of the 2PCF to exclude separations below the scale cut. With the MST this is more complicated. Unfortunately, there does not appear to be a way to deal with this after the MST has been constructed; this is because the problematic smallest edges will by construction be incorporated in the graph. To ensure that problematic small scales are removed from the MST we alter the *k*NN graph that is the input to the Kruskal algorithm by removing edges whose length is below the desired scale cut.

4.3 Redshift space distortion on MICE galaxies

RSDs (Kaiser 1987), caused by the Kaiser and Fingers of God effects, will distort the measured redshift of galaxies and thus will impact the inferred comoving distance. Since this effect alters the 3D distribution of galaxies, it will inevitably affect the MST statistics.

We explore this effect by comparing the MST performed on a subset of the MICE galaxy catalogue (Crocce et al. 2015) in real and redshift space (i.e. with RSD). Here, we randomly draw 10 realizations of 500 000 galaxies with real comoving distances between 1000 to 1500 h^{-1} Mpc. We ensure that the density of galaxies is constant so that the number of galaxies $\propto D_c^3$, where D_c is the radial comoving distance from the observer.

Fig. 4 shows the MST statistics with and without the RSD effect. We see significant results in all the MST statistics demonstrating the importance of including this effect in any future MST study.

5 WHAT DOES THE MINIMUM SPANNING TREE MEASURE?

This section considers the following questions:

(i) What do the MST statistics look like on an unbiased tracer (i.e. *N*-body dark matter particles)?
(ii) What does the MST statistics tell us about the underlining density distribution?
(iii) What is the relation of MST statistics to 2PCF?
(iv) What happens when we change simulation resolution?
(v) How do the MST statistics change when measured on haloes (i.e. a more galaxy-like tracer)?

5.1 *ν*N-body simulations

Five *N*-body simulations (see Massara et al. 2015) were made by running the TREEPM code GADGET-III (Springel 2005). The following cosmological parameters were common to all simulations: $\Omega_m = 0.3175$, $\Omega_b = 0.049$, $\Omega_\Lambda = 0.6825$, $h = 0.6711$, and $n_s = 0.9624$. See Table 2 for a list of the simulations used and their respective cosmological parameters, particle numbers and box sizes. The cold dark matter energy density is set to $\Omega_c = \Omega_m - \Omega_b - \Omega_\nu$ where $\Omega_\nu h^2 \simeq \sum m_\nu/(94.1\ \text{eV})$. Cold dark matter and neutrinos are both treated as collisionless particles. They differ in their masses and in their initial conditions, where the initial conditions for neutrinos receive an extra thermal velocity obtained by randomly sampling the neutrino Fermi–Dirac momentum distribution (Viel, Haehnelt & Springel 2010). These are evolved from an initial redshift of $z = 100$. Table 2 summarized the simulations used.

5.2 MST application to dark matter particles

An MST was constructed on the dark matter particles from the HZ, LZ, and LN simulations (see Table 2), where errors were calculated using the jackknife method (Section 2.2). Figs 5, 6, 7, and 9 use the same colour scheme: HZ in blue, LZ in orange, and LN in green. We boost the speed of the MST calculation by allowing this to be done in parallel, breaking the *N*-body snapshots into 64 cubes. We then implement the scale cut strategy discussed in Section 4.2 and partition the data set into four groups (to dilute the sample to look at larger sales) and apply a scale cut of $l_{\rm min} = 2\ h^{-1}$ Mpc.

5.2.1 *Features in the minimum spanning tree statistics*

In Fig. 5, we plot the MST statistics for these different simulations at redshifts $z = 2$, 1, 0.5, and 0. The plots display how the MST statistics evolve over cosmological time, as discussed below:

(i) *Degree:* the distribution of degree remains relatively similar in all simulations and does not appear to evolve greatly over redshift, although differences between the simulations become more pronounced at lower redshifts.
(ii) *Edge length:* overall we see that the distribution shows a high sensitivity to redshift, evolving from a single distribution into a bimodal one at smaller redshift.

(a) $l \geq 3\ h^{-1}$ Mpc: a broad peak is seen in the distribution at $l \simeq 4\ h^{-1}$ Mpc. This feature dampens at lower redshift with the peak consistently highest for LN, followed by LZ, and then HZ.
(b) $l < 3\ h^{-1}$ Mpc: a secondary peak emerges and dominates at lower redshift, which rises against the scale cut limit of $l_{\rm min} = 2$.
(c) $l \sim 3\ h^{-1}$ Mpc: between the two peak features is a region where seemingly all three distributions appear to converge and the orderings of the peaks above and below this point switch.

(iii) *Branch length:* the evolution appears virtually identical to the edge length distribution except at larger scales.
(iv) *Branch shape:*

(a) A broad peak at $\sqrt{1-s} = 0.6$ which is present in all simulations. This peak is always highest for LN followed by LZ and HZ.
(b) A subpeak at $\sqrt{1-s} \sim 0.05$ which dampens at lower redshift. This suggests that some branches at low redshift are fairly straight. Since the simulation we use are fairly low in resolution we suspect that this feature is more an indication that the particles have not undergone much mixing and are still very close to their initial perturbed grid layout. This could be used as a diagnostic to test whether *N*-body simulations have moved from their perturbed gridded initial conditions.
(c) Lastly, we see the emergence of two bumps between $\sqrt{1-s} \sim 0.7 - 1$ at low redshift. Comparison of the branch shape statistics with and without a scale cut show this is caused by the introduction of the scale cut, which forces some branches to be more curved. Branch shapes without a scale cut rarely see $\sqrt{1-s} > 0.8$.

5.2.2 *Exploring the minimum spanning tree relation to density*

To gain a greater physical intuition of what these statistics are telling us about cosmology, we subdivide the 1 h^{-1} Gpc cube into smaller

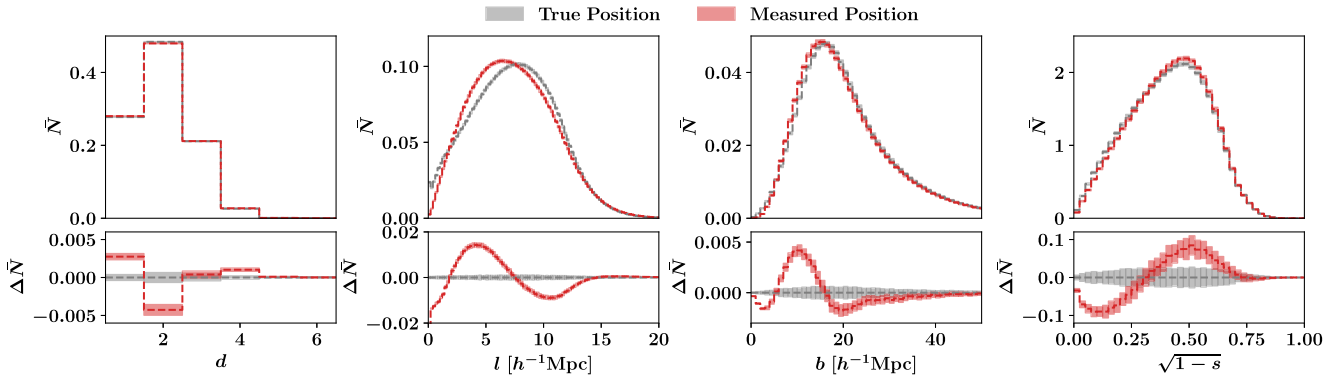

Figure 4. The effects of RSDs on the MST statistics. From left to right: the MST statistics degree (d), edge length (l), branch length (b), and branch shape (s). Bottom panels show the differences. Ten realizations of 500 000 MICE galaxies were generated and the MST were constructed on their true positions (grey) and then the measured positions (red), i.e. the inferred positions based on their redshifts including RSD. The envelopes correspond to 1σ uncertainties. Significant differences between the MST statistics show that the MST is sensitive to the RSD effect.

Table 2. Simulation and cosmological parameters for the N-body simulations. Massara et al. (2015) use different names, which we list here.

Name	Reason for name	Massara et al. (2015)	Box size (h^{-1} Mpc)	$N_{\rm cdm}$	N_ν	$\sum m_\nu$ (eV)	σ_8	$10^9 A_{\rm s}$
HZ	High σ_8, zero $\sum m_\nu$	L0	1000	256^3	0	0	0.834	2.13
LZ	Low σ_8, zero $\sum m_\nu$	L0s8	1000	256^3	0	0	0.693	1.473
LN	Low σ_8, non-zero $\sum m_\nu$	L60	1000	256^3	256^3	0.6	0.693	2.13
HZHR	High σ_8, zero $\sum m_\nu$, high resolution	H0	500	512^3	0	0	0.834	2.13
LNHR	Low σ_8, zero $\sum m_\nu$, high resolution	H60	500	512^3	512^3	0.6	0.693	1.473

25 h^{-1} Mpc cubes. In these cubes, we calculate the density contrast δ,

$$\delta = \frac{N_{\rm DM}}{\langle N_{\rm DM}\rangle} - 1, \tag{6}$$

where $N_{\rm DM}$ is the number of dark matter particles in a particular cube and $\langle N_{\rm DM}\rangle$ is the average across all cubes. Fig. 6 illustrates the relationship between the average degree ($\langle d\rangle$), edge length ($\langle l\rangle$), branch length ($\langle b\rangle$), and branch shape ($\langle s\rangle$) and the density contrast inside these cubes.

(i) d versus δ: we see that the mean of the degree, d, is relatively constant at $d \simeq 2$ as a function of density. The variance shows a strong dependence on density, with overdensities having very low variance, i.e. predominantly $d = 2$, and underdensities showing a much larger variance and a slight tilt towards $d = 1$. Of course, we should expect high-density environment to form the main 'backbone' of the MST, since these are the areas where the edges are shortest.

(ii) l and b versus δ: both the edge and branch length distribution show a very similar relation to density. Shorter edges and branches are mostly associated with overdensities and vice versa. Furthermore as the simulations evolve in redshift this relation becomes more pronounced. In both these statistics, we see that HZ appears consistently to have more overdense and underdense regions than the other two simulations. We also see that LN appears to have marginally but consistently higher overdense and underdense regions than LZ.

(iii) s versus δ: the mean of the branch shape appears centred at 0.75 and shifts slightly to a mean of 0.7 for higher densities. Furthermore, as with the degree, the biggest relation to density is with the variance, which increases as the density lowers.

This analysis demonstrates a clear relation between MST statistics and environment (i.e. the local density).

5.2.3 *Relation to the matter power spectrum*

In Fig. 7, we calculate the matter power spectra, $P(k)$, measured from these simulations. The dependence on redshift can be characterized by a simple shift in amplitude. We see that (at all k) HZ has more power, followed by LZ and then LN. At small k, LZ converges to HZ while at large k, LZ converges to LN. Notice, that the strength in $P(k)$ at large k is matched by a tendency for shorter edges in the MST, demonstrating the MST expected dependence on clustering.

5.2.4 *Simulation resolution*

The MST of N-body simulations will be affected by the resolutions used. To measure the sensitivity of the MST statistics to the simulation resolution, we calculate the MST on higher resolution versions of HZ and LN called HZHR and LNHR (see Table 2 for details of simulation properties). The resulting distributions of the MST statistics are shown in Fig. 8. For comparison, we additionally subsample these two simulation boxes by randomly selecting particles in the simulation with equal number of particles. In the more sparsely sampled version of HZHR and LNHR, the more resolved extreme high- and low-density environments are still imprinted. This can be seen by the fact that in the bottom panels of Fig. 8 there appears to be more features at high and low values of l. This illustrates the importance of high-resolution simulations on the MST profiles inferred. We could also use high-resolution simulations to calibrate the scale cut for low-resolution simulations

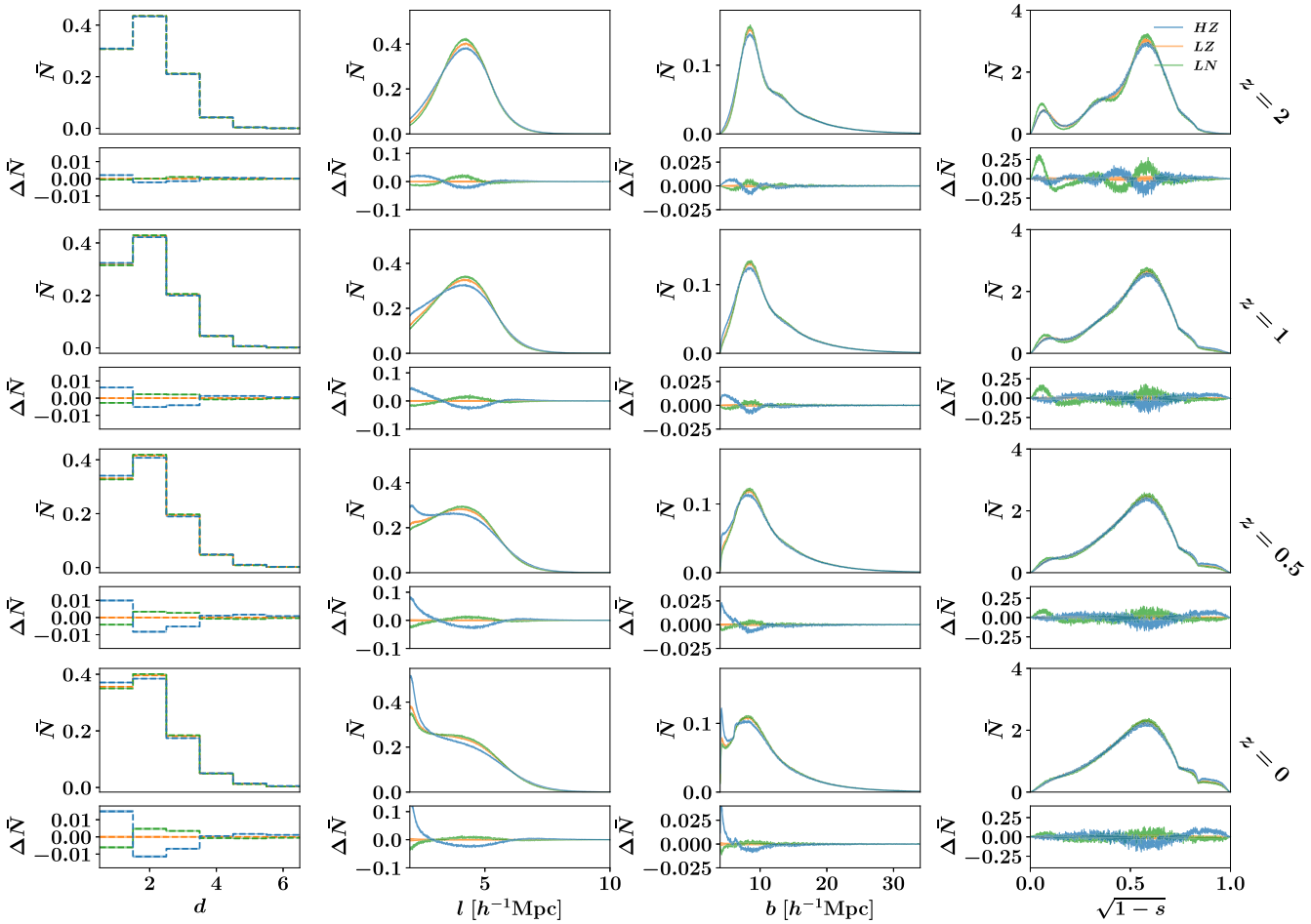

Figure 5. From left to right: the distribution of degree (d), edge length (l), branch length (b), and branch shape (s). These are obtained by dividing the full 1 $(h^{-1}\,\mathrm{Gpc})^3$ box into 250 $(h^{-1}\,\mathrm{Mpc})^3$ cubes for speed. These are then partitioned into four groups to minimize the effect of applying a scale cut of 2 h^{-1} Mpc. From top to bottom: distributions are shown with respect to redshift 2, 1, 0.5, and 0. These are further subdivided into a top subplot of the distributions and a bottom subplot of the differences. Simulations shown are HZ (blue), LZ (orange), and LN (green). See Section 5.2.1 for a detailed explanation of the distribution features, differences, and evolution.

by allowing the scale cut to vary until the MST statistics reach agreement between the high- and low-resolution simulations. We additionally measure the MST on a completely random set of points (shown in grey) illustrating how the more sparsely subsampled data set appears to be asymptotically approaching these profiles.

5.3 MST application to haloes

Halo catalogues were derived from the HZ, LZ, and LN simulation snapshots. We study these to get a sense of what the MST statistics will look like when performed on a biased tracer, such as galaxies. We dropped the $z = 2$ snapshots as they contained too few haloes to be meaningful. Unlike the N-body simulation, we do not apply a scale cut since the density of haloes is quite low and the fraction of edges below $l_{\mathrm{min}} = 2\ h^{-1}$ Mpc is very low. The MST statistics derived from the haloes is shown in Fig. 9. The number of haloes varies both across simulations and across redshift snapshots (see Table 3) – this is different from dark matter particles whose number count is constant across redshift and simulations.

To mitigate this issue, for each redshift we match the number of haloes to the lowest number found in the simulations (thus always matching the number of haloes found in the LZ simulations). For those with more haloes, we simply select the most massive haloes. In Fig. 9, we find no real noticeable difference in the statistics suggesting the degeneracies of the MST may be similar to that found for $P(k)$.

6 COMPARING THE SENSITIVITY TO COSMOLOGY OF POWER SPECTRUM, BISPECTRUM, AND THE MINIMUM SPANNING TREE

In this section, we compare the sensitivities to cosmological parameters of power spectrum $P(k)$, bispectrum $B(k_1, k_2, k_3)$, and MST, measured on the same halo catalogues, to establish whether the MST can improve parameter constraints. Specifically, we compare the constraints on A_s, Ω_m, and $\sum m_\nu$ for 10 sets of mock simulations. To obtain reliable posterior distributions for the three methods and their joint constraints, we would normally run a Markov Chain Monte Carlo (MCMC) using an analytic expression for the data vector. However, the MST statistics cannot be obtained analytically and hence have to be obtained from simulations. $P(k)$, $B(k_1, k_2,$

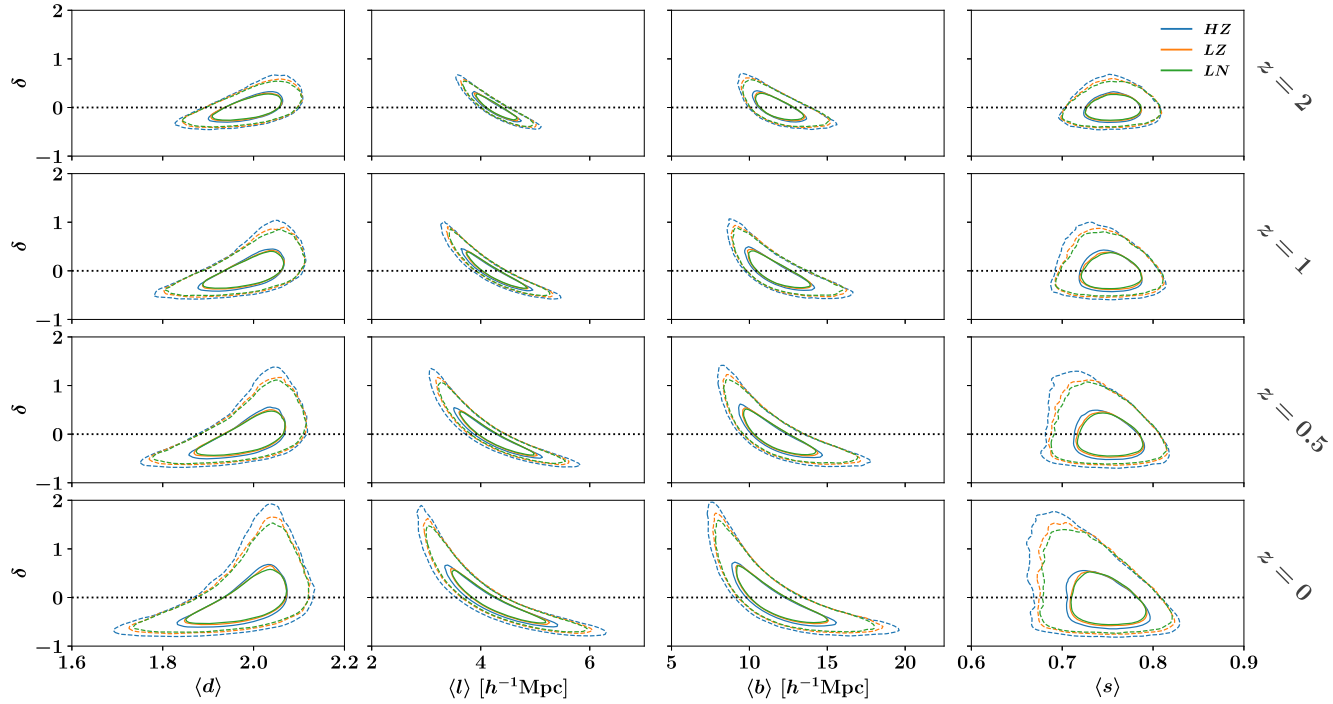

Figure 6. Contour plots of the average density contrast (δ) is plotted against the MST statistics [from left to right: the average degree ($\langle d \rangle$), edge length ($\langle l \rangle$), branch length ($\langle b \rangle$), and branch shape ($\langle s \rangle$)] in 25 h^{-1} Mpc cubes. The 1σ and 2σ contours are indicated by solid and dashed lines, respectively. The relation for HZ is in blue, LZ in orange, and LN in green. See Section 5.2.2 for a detailed explanation of the relation and their evolution.

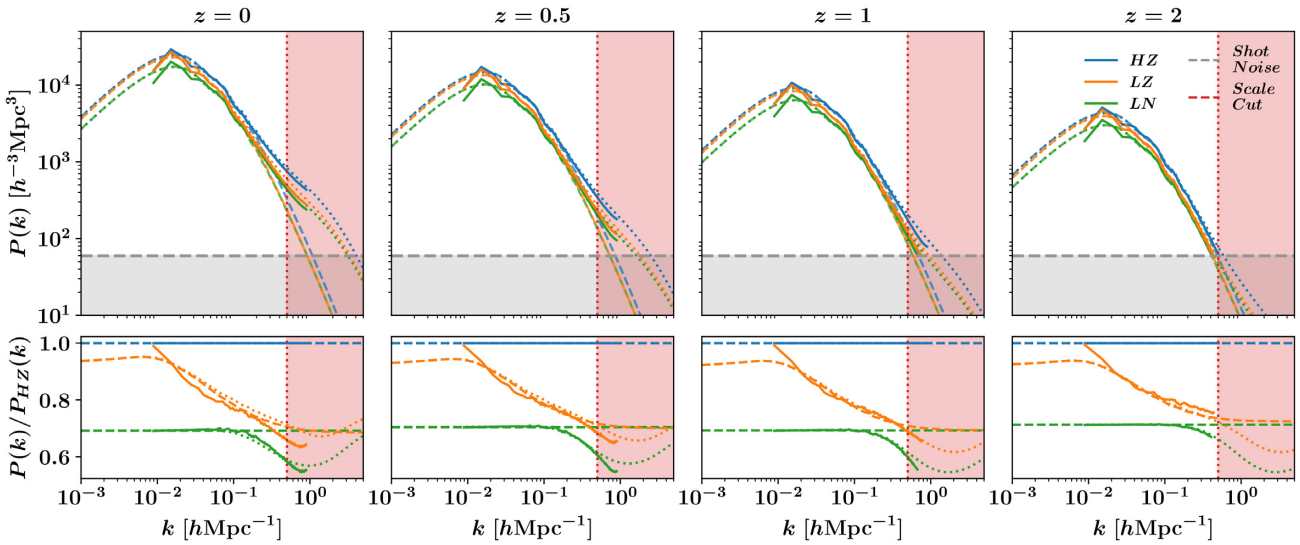

Figure 7. In the top panels, the matter power spectra, $P(k)$, are plotted for redshift (from left to right) 2, 1, 0.5, and 0 for simulations HZ (blue), LZ (orange), and LN (green). In the bottom subplots we plot the ratio with respect to the HZ power spectra. Solid lines correspond to the measured $P(k)$ from the respective simulations, while dashed and dotted lines correspond to the theoretical linear and non-linear $P(k)$, respectively. The dashed grey lines shows the level at which the measured $P(k)$ will be affected by the shot noise of the simulation and the regions in red show the scales for which we apply a scale cut in the construction of the MST. Here, we see (that at all redshifts) the power at all k is highest for HZ, and then LZ and lastly LN. Note that LZ is close to HZ at high k and close to LN at low k.

k_3), and MST are therefore estimated from a grid of simulations in parameter space. To limit the noise in the estimates of the theory we take the mean of five simulations rather than just one at each point in parameter space. Additionally, since our simulation grid is rather sparse we use Gaussian process (GP) regression to interpolate the data vector. Finally, we use a corrected likelihood function (see Sellentin & Heavens 2016; Jeffrey & Abdalla 2018) which accounts for the use of an estimated covariance matrix.

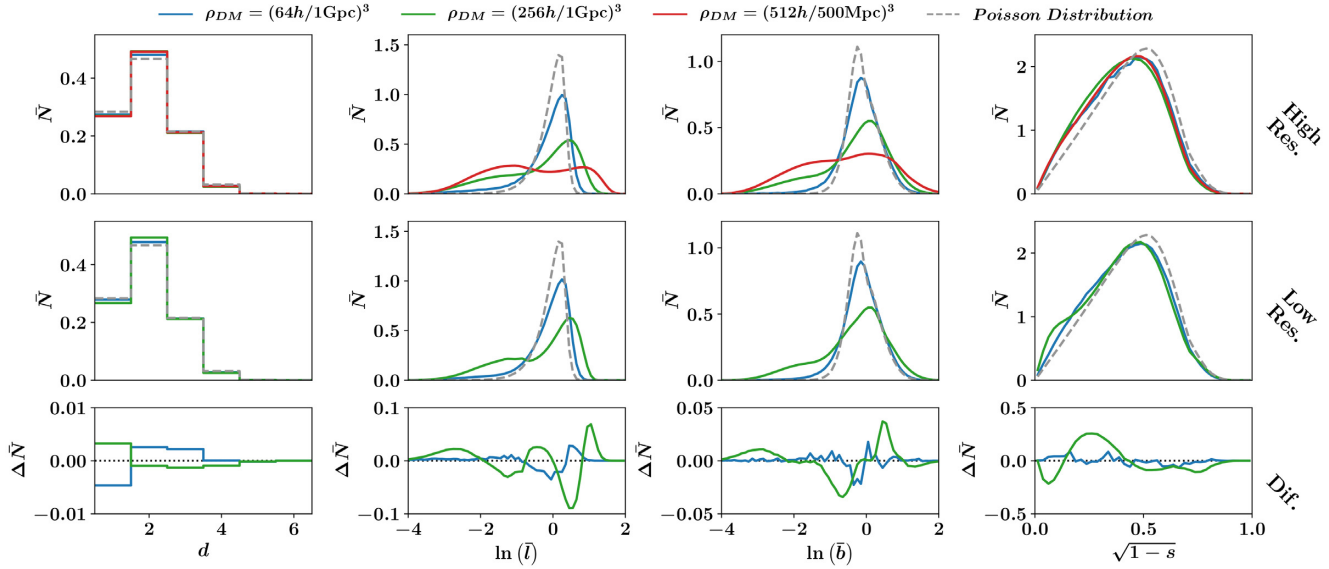

Figure 8. The distribution of degree (d), normalized edge and branch lengths ($\ln(\bar{l})$ and $\ln(\bar{b})$), and branch shape (s) are displayed from left to right. Top panel: the distributions of the high-resolution versions of HZ (dashed line) and LN (dotted line) (i.e. HZHR and LNHR) simulation are shown in red and subsequently subsampled versions are shown in green and blue with dark matter particle densities (ρ_{DM}) 256^3, 128^3 and 64^3 per $(h^{-1}\,\mathrm{Gpc})^3$ respectively. Middle panel: the distribution for the HZ (dashed line) and LN (dotted line) simulation is shown. Bottom panel: the differences between the high-resolution (top panel) and low-resolution (middle panel) simulations are shown. We additionally illustrate the distribution for random points (dashed grey).

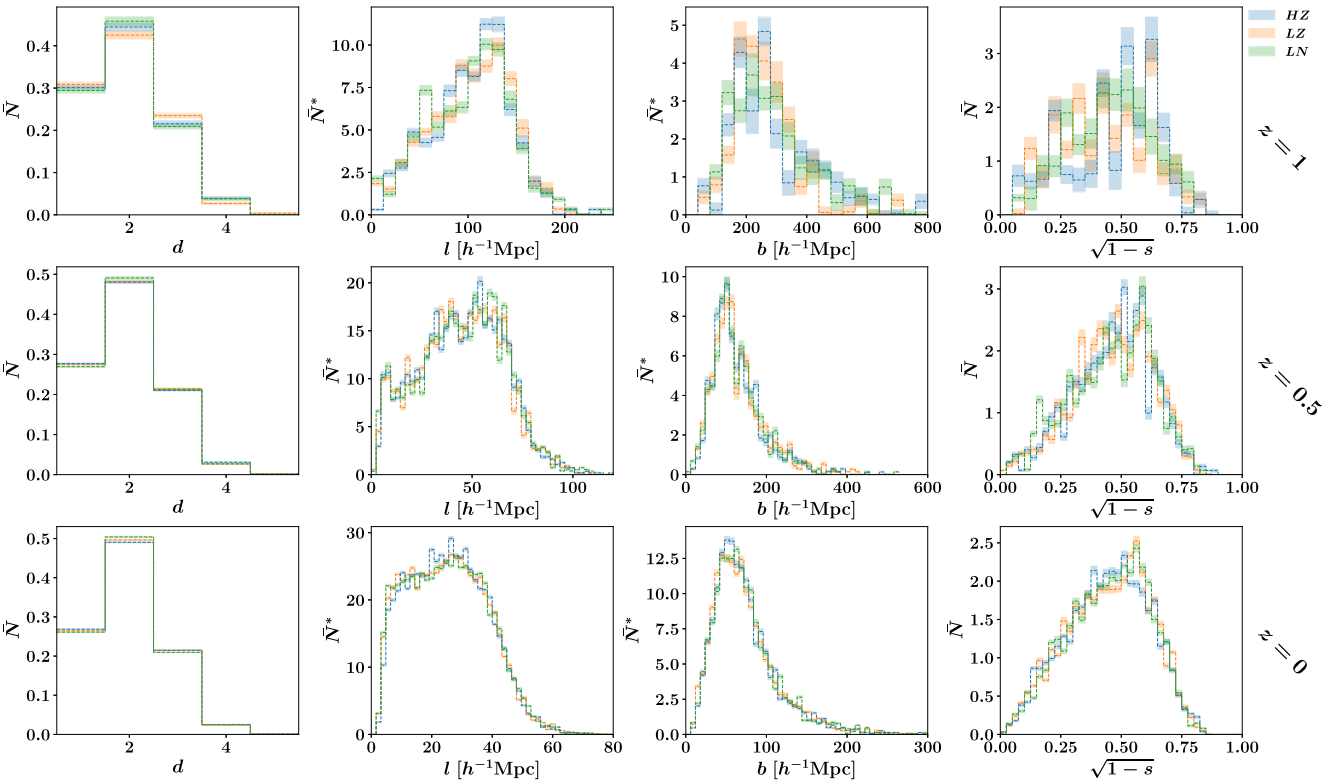

Figure 9. The MST constructed on halo catalogues derived from the HZ (blue), LZ (orange), and LN (green) N-body simulations. From left to right are the MST statistics: degree (d), edge length (l), branch length (b), and branch shape (s). They are plotted from top to bottom according to snapshots at redshift 1, 0.5, and 0. Corresponding shaded areas show the jackknife uncertainties in the measurements. The distribution of the MST statistics are indistinguishable from each other at all redshifts, demonstrating that we should expect to see similar lines of degeneracy as power spectrum. Note $\bar{N}^* = 10^3\bar{N}$.

Table 3. The number of haloes found in each simulation (HZ, LN, and LZ) for each redshift (z) snapshot. The number of haloes at $z = 2$ was far too little for a meaningful MST and presumably would be uninformative.

Redshift	HZ	LN	LZ
0	17911	11168	9892
0.5	6717	3017	2392
1	1585	458	262
2	16	2	1

6.1 COLA simulation suites

A suite of COLA (Tassev, Zaldarriaga & Eisenstein 2013) simulations were constructed using the MG-PICOLA software (Winther et al. 2017, an extension to L-PICOLA by Howlett, Manera & Percival 2015) which, among other things, can model the effects of massive neutrinos (Wright, Winther & Koyama 2017). This allowed us to generate N-body-like simulations relatively cheaply (in terms of computation time), albeit by sacrificing accuracy at small scales. All simulations are run in boxes of lengths 250 h^{-1} Mpc, with 256^3 dark matter particles and a discrete Fourier transform (DFT) density grid of $(3 \times 256)^3$. The latter is set to satisfy a requirement to produce accurate haloes from COLA simulations (Izard, Crocce & Fosalba 2016). The dependence on A_s, Ω_m, and $\sum m_\nu$ are explored, while $h = 0.6711$, $\Omega_b = 0.049$, and $n_s = 0.9624$ are constant in all simulations. Haloes and particles are outputted at redshift $z = 0.5$, using 20 steps from an initial redshift $z = 10$. Further details on the simulation suites are summarized in Table 4.

The reliability of these simulations is evaluated by comparing the power spectrum, calculated on the dark matter particles from the fiducial suite, to the non-linear power spectrum calculated from CAMB. We plot the 1σ difference variation in the power spectrum in Fig. 10. Although this test shows the simulations can be trusted up to $k < 0.7\,h\mathrm{Mpc}^{-1}$, we apply a conservative scale cut of $k_{\max} < 0.5\,h\mathrm{Mpc}^{-1}$ in Fourier space and $l_{\min} > 4\pi\,h^{-1}$ Mpc in real space.

6.2 Measurements

We use haloes from MG-PICOLA as a proxy for galaxies. These are found using the friends-of-friends halo finder `MatchMaker`[14] which was found to be consistent (for the heaviest haloes) to the phase space halo finder ROCKSTAR (Behroozi, Wechsler & Wu 2013). Unlike $P(k)$ and $B(k_1, k_2, k_3)$ which are unaffected by the density of tracers, the MST will exhibit different profiles purely based on the different halo counts. Since different number of haloes are produced from simulations with different cosmologies we mitigate this issue by performing our measurements on only the heaviest 5000 haloes. In practice such a restriction would not be imposed on $P(k)$ or $B(k_1, k_2, k_3)$ measurements, but here we wish to simply establish whether the MST improves on the constraints of $P(k) + B(k_1, k_2, k_3)$.

We will explore replicating realistic survey properties in later work but in practice if we were simulating a galaxy catalogue, we would have to use a halo occupation distribution (HOD) model where we would tune the parameters of the HOD to have the same galaxy density as the actual survey. What we do here is a simplified version of that. The simulations constructed used haloes with masses between 10^{12} and 10^{15} $\mathrm{M}_\odot$. The number density ($\sim 3.2 \times 10^{-4}\ h^{-3}\,\mathrm{Mpc}^3$) is similar to the BOSS LOWZ sample between redshift 0.3–0.4 and to the CMASS sample between redshift 0.5–0.6 (see fig. 1 of Tojeiro et al. 2014). Assuming a linear bias of $b^2 = P_{\mathrm{haloes}}(k)/P(k)$ we found the fiducial simulations to have a bias of $b \sim 1.3$; this is more similar to the bias observed in eBOSS for emission line galaxies ($b \sim 1.4$) than in BOSS for luminous red galaxies ($b \sim 2$).

6.2.1 *Power spectrum and bispectrum*

Power spectrum and bispectrum measurements are performed through DFT algorithms as implemented by FFTW3.[15] We use the cloud-in-cell (CIC) mass assignment scheme using 64^3 Cartesian grid cells to define a discrete overdensity field in configuration space, later transformed into Fourier space. The size of the simulation box is $L_{\mathrm{box}} = 250\,h^{-1}$ Mpc and therefore, the mass resolution of the discrete over-density field is $\sim 3.9\,\mathrm{Mpc}h^{-1}$. We compute the power spectrum between the fundamental frequency, $k_f = 2\pi/L_{\mathrm{box}}$, and a maximum frequency, $k_{\max} = 0.5\,h\mathrm{Mpc}^{-1}$, in bins of k_f.

The power spectrum and bispectrum measurements are performed using the code and estimator described in Gil-Marín et al. (2017). For the bispectrum, we initially perform the measurements in bin sizes of k_f. In this case, we ensure that the three k-vectors of the bispectrum form closed triangles, and without loss of generality we define $k_1 \leq k_2 \leq k_3$. We include all the closed triangles with $k_3 < k_{\max}$. The bispectrum data vector, $B(k_1, k_2, k_3)$, contains around 700 elements. In Fig. 11, the bispectra measured on dark matter particles from the fiducial simulations are compared to theoretical values, showing good agreement until we reach non-linear regimes where the theory can no longer be trusted.

Using measurements of the power spectrum and bispectrum on the haloes of the fiducial suite, we were able to determine the skewness and kurtosis of the individual elements of the data vector. We found that elements with $k < 0.125\,h\mathrm{Mpc}^{-1}$ contained much higher than expected skewness and kurtosis (i.e. exceeded the expected skewness and excess kurtosis of a Gaussian data set by 2σ) and as such we limit the power spectrum and bispectrum measurements to $k > 0.125\,h\mathrm{Mpc}^{-1}$. This reduced the bispectrum data vector from ∼700 to ∼500. We then use a maximal compression technique (based on the work of Tegmark, Taylor & Heavens 1997; Heavens et al. 2017) to compress the bispectrum data vector to three elements (following Gualdi et al. 2018, 2019). Such a compression allows us to estimate the covariance matrix for a number of triangle configurations much larger than the number of available simulations.

6.2.2 *Minimum spanning tree*

The MST measurements are made with a scale cut of $l_{\min} > 4\pi\,h^{-1}$ Mpc, which corresponds to the wavelength ($\lambda = 2\pi/k$) of the largest k modes ($k_{\max}$) probed by $P(k)$ and $B(k_1, k_2, k_3)$. The MST statistics are then binned, which presents a problems as counts are discrete. For large counts, the distribution can be approximated by a Gaussian and as such we only select bins which we found the mean of our fiducial data vectors to have counts of greater than 50.

[14] https://github.com/damonge/MatchMaker

[15] Fastest Fourier Transform in the West, http://www.fftw.org

Table 4. Properties of the simulations suites are shown above; including the reference names, cosmological parameters, realizations, and information on their eventual uses.

Name	$10^9 A_s$	Ω_m	$\sum m_\nu$ (eV)	Realizations	Notes
Grid	[1, 3.5]	[0.2, 0.5]	[0, 0.6]	5	Simulations carried out at 216 points defined across a 6 × 6 × 6 grid in parameter space
Fiducial	2	0.3	0	500	Used to calculate covariance matrices
Mock	2.13	0.3175	0.06	10	Treated as real data

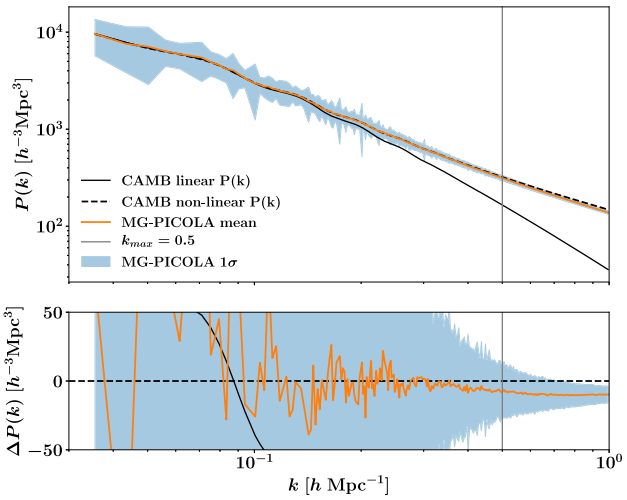

Figure 10. In the top panel, we compare the mean (blue) and 1σ distributions (blue envelopes) of the power spectra calculated on dark matter particles from our fiducial suite of simulations to the linear and non-linear CAMB power spectra. In the bottom panels, we show the difference between the measured and non-linear CAMB power spectra. The power spectra from MG-PICOLA appears to be accurately reproduced up to about $k = 0.7$, but we conservatively apply a scale cut of $k < k_{\rm max}$ where $k_{\rm max} = 0.5$.

6.3 Parameter estimation

Using the noisy estimates of the theory $\boldsymbol{d}_{\rm Grid}$ (the mean of five grid simulations at each point in parameter space) we can interpolate using GPs (see Appendix A) from a 6 × 6 × 6 to a 20 × 20 × 20 grid with theoretical data vectors $\boldsymbol{\mu}_{GP}$ and uncertainty $\boldsymbol{\sigma}_{GP}$ which is used instead of an MCMC due to the low dimensionality of the parameters. The sample covariance matrix, $\mathbf{S}$, is estimated from 400 fiducial simulations (the other 100 fiducial simulations are used to apply a coverage correction, Sellentin & Starck 2019). The posterior for each of our ten mocks, denoted by the data vector $\boldsymbol{d}$, is evaluated using the likelihood function (which accounts for an estimated sample covariance, see Sellentin & Heavens 2016; Jeffrey & Abdalla 2018)

$$\mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}) \propto \det(\mathbf{C})^{-1/2} \left[1 + \frac{(\boldsymbol{d} - \boldsymbol{\mu}_{\rm GP})^\top \cdot \mathbf{C}^{-1} \cdot (\boldsymbol{d} - \boldsymbol{\mu}_{\rm GP})}{N - 1}\right]^{-\frac{N}{2}}, \tag{7}$$

where the uncertainty in the GPs regression is added to the sample covariance, i.e. $\mathbf{C} = \mathbf{S} + \mathbf{S}_{\rm GP}$, where elements of $(\mathbf{S}_{\rm GP})_{\rm ij} = \boldsymbol{\sigma}_{\rm GP,i}\boldsymbol{\sigma}_{\rm GP,j}\delta_k(v_{\rm i}, v_j)$ where $\delta_{\rm k}$ is the Kronecker delta function and $v_{\rm i}$ and $v_{\rm j}$ are only equal if the same GPs hyperparameters were used to construct these elements of the data vector (following Bird et al. 2019; Rogers et al. 2019, which assume maximal dependency between elements of the data vector constructed from the same GPs hyperparameters).

Finally, we apply a coverage correction (Sellentin & Starck 2019) using 100 fiducial simulations not included in the calculation of the covariance matrix. This accounts for unrecognized sources of biases. We found that all methods exhibited overconfident confidence contours. For $P(k)$ and $B(k_1, k_2, k_3)$, this is believed to have arisen due to non-Gaussian features in the data set. Although we attempted to limit this by selecting regions of the data vector that had fairly low skewness and kurtosis, we found that the skewness for $P(k)$ tended to be consistently positive, whilst the excess kurtosis for the maximally compressed $B(k_1, k_2, k_3)$ was always $>1\sigma$ than expected if the data were Gaussian. For the MST, this effect is larger which we suspect occurs due to two reasons: (1) similar to $P(k)$ and $B(k_1, k_2, k_3)$ the data vector is non-Gaussian and (2) the scale cut adds an extra stochasticity to the data vector that is not fully captured by the covariance matrix.

6.4 Comparison

The posterior distributions are measured for the three statistics and their combinations. Correlations between each statistic are accounted for by using a covariance matrix that is not block diagonal. In Figs 12–14, we show the posterior distributions measured on the mean of the data vectors from 10 mocks allowing for better visual comparison of the errors whilst improvement in parameter constraints are stated according to the average improvement when measured on the mocks independently.

6.4.1 Components of the minimum spanning tree

We compare the constraints from the four individual components of the MST. The elements of the MST statistics are counts, and as such they follow a Poisson distribution. We apply a cut on the data vector based on where the mean of the fiducial MST statistics had counts >50, where expect the Poisson distribution to be approximately characterized by a Gaussian. In Fig. 12, we display the constraints from the individual components of the MST. Of the four statistics *s* is the least constraining and provides very little information; this is followed by *d* which, although it has very broad posteriors, appears at least to rule out parts of the parameter space (low A_s, Ω_m, and high $\sum m_\nu$). The MST statistics *l* and *b* provide constraints having similar degeneracies with *l* providing somewhat tighter constraints.

6.4.2 P(k)*,* B(k_1*,* k_2*,* k_3*), and MST*

In Fig. 13, we compare the constraints from $P(k)$, $B(k_1, k_2, k_3)$ and MST. All three appear to have similar degeneracies and as such are unable to establish meaningful constraints on A_s and $\sum m_\nu$. The constraints on Ω_m are more conclusive but are fairly similar. The constraints on $\sum m_\nu$ tend to show a broad peak towards the centre of the prior range. Since the constraints on neutrino mass are poor the kernel length-scale for $\sum m_\nu$ of the GPs is quite broad and as

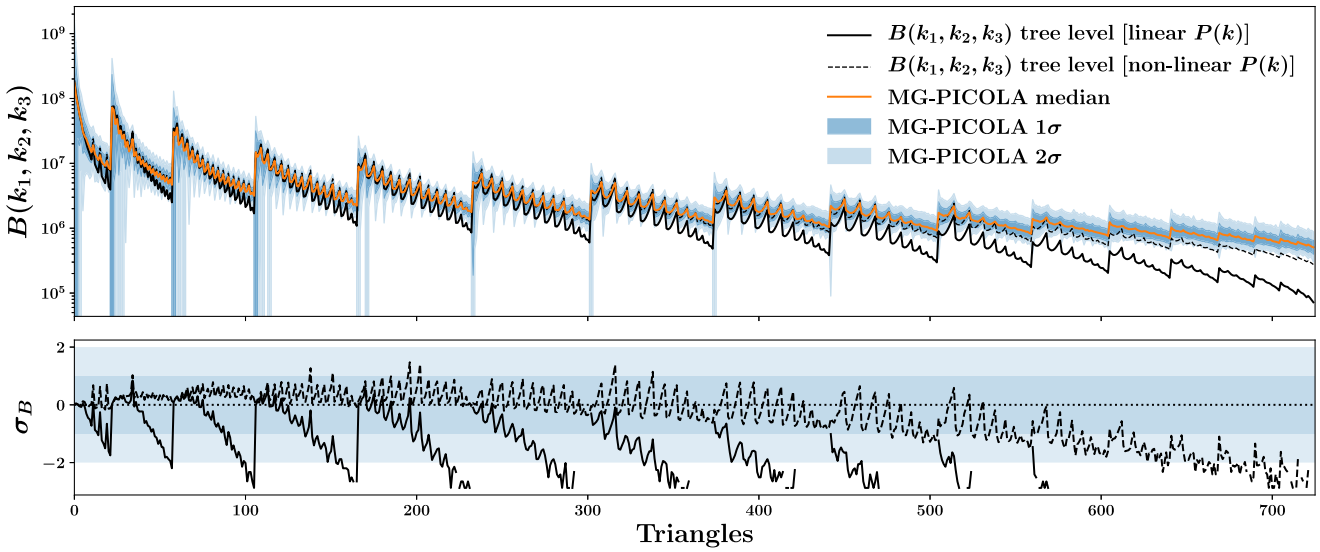

Figure 11. In the top panel, we compare the mean (blue) and 1σ and 2σ distributions (blue envelopes) of the bispectrum calculated on dark matter particles (from our fiducial suite of simulations) against theoretical bispectra calculated using the linear and non-linear CAMB power spectra. The x-axis displays triangle index (generated by listing triangles in lexographic order based on sides k_1, k_2, and k_3 where all elements are below $k_{\rm max}$). In the bottom panel, we show the significance between the measured and theoretical values. The theoretical bispectrum measurements are made using Gualdi et al. (2018) and will only be accurate up to the quasi-linear regime; since we are pushing to more non-linear scales the discrepency for smaller triangles is expected. Using the non-linear $P(k)$ for the bispectrum is an approximation that only helps in partially reducing the discrepancy between the tree-level model and the measurements by using loop corrections for the power spectrum. A better model would be given by using one-loop corrections to the bispectrum.

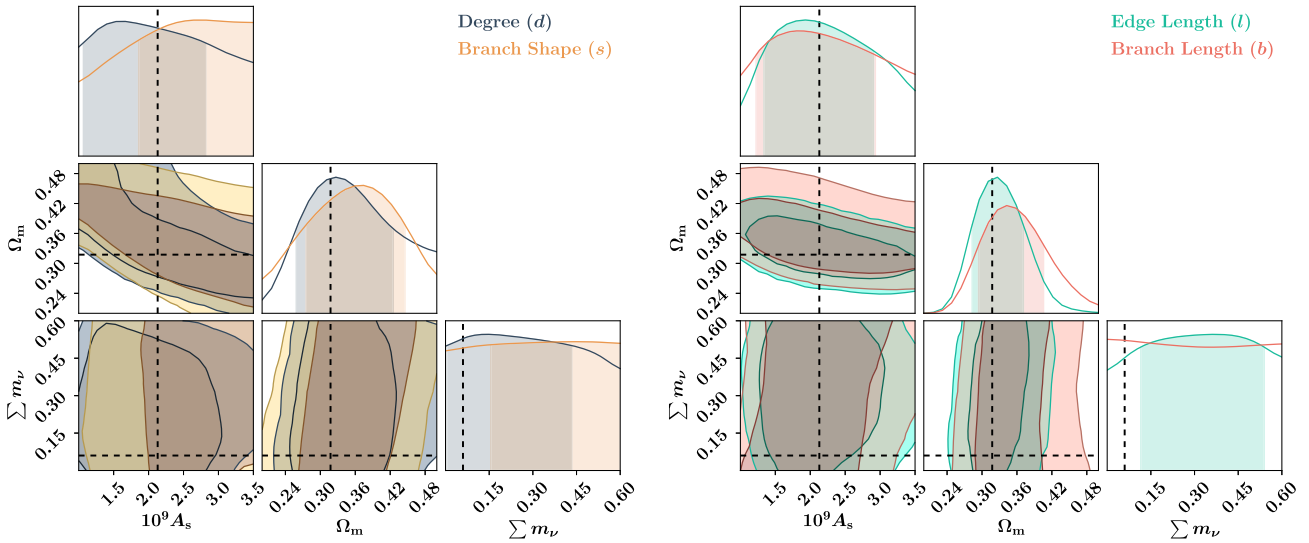

Figure 12. Posterior distributions on cosmological parameters as constrained by the individual components of the MST. On the left, we show those from the degree and branch shape and on the right from edge and branch lengths. Branch shapes are the least sensitive, whilst the degree gives broad constraints but rules out parts of the parameter space. Edge and branch length show similar posterior distributions with tighter constraints coming from edges.

such the estimates of the theory vector are smoother in the centre. This creates a slight bias towards the centre of the parameter space. This effect is also seen in Fig. 14.

6.4.3 *Combining* P(k)*,* B(k_1*,* k_2*,* k_3*), and MST*

In Fig. 14, we combine the statistics and compare their relative constraints which is more clearly shown in Fig. 15. In combining $P(k)$ and $B(k_1, k_2, k_3)$, we find an improvement of ~ 6 per cent in the constraints of $\Omega_{\rm m}$ and ~ 3 per cent for $A_{\rm s}$. When combined with the MST the constraints on $\Omega_{\rm m}$ improve by ~ 17 per cent and on $A_{\rm s}$ improve by ~ 12 per cent with respect to (with respect to) $P(k)$ (~ 12 per cent for $\Omega_{\rm m}$ and ~ 10 per cent for $A_{\rm s}$ with respect to $P(k) + B(k_1, k_2, k_3)$). Since we have ensured the same scale cuts, i.e. $k_{\rm max} = 0.5\,h{\rm Mpc}^{-1}$ for $P(k)$ and $B(k_1, k_2, k_3)$ and $l_{\rm min} = 4\pi\,h^{-1}$ Mpc, we can be fairly certain that the additional information is not coming

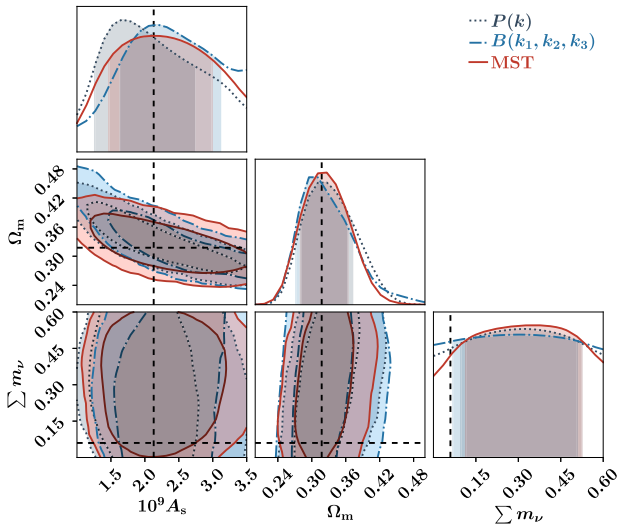

Figure 13. The posterior distributions are shown for power spectrum ($P(k)$, shown in grey), bispectrum ($B(k_1, k_2, k_3)$, shown in blue) and MST (shown in red). The tightest constraints on A_s and Ω_m are given by the MST, whilst $B(k_1, k_2, k_3)$ provides better constraints on $\sum m_\nu$.

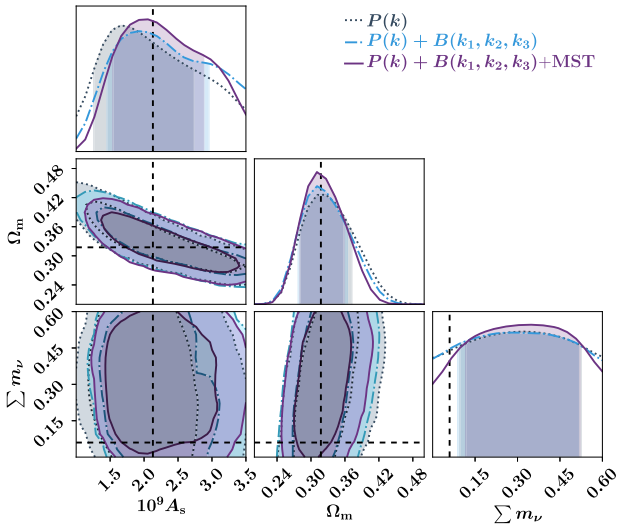

Figure 14. The posterior distributions for cosmological parameters as constrained by (a) power spectrum ($P(k)$, shown in dark grey), (b) power spectrum and bispectrum ($P(k) + B(k_1, k_2, k_3)$, shown in blue) and (c) power spectrum, bispectrum, and MST ($P(k) + B(k_1, k_2, k_3)$ + MST, shown in purple).

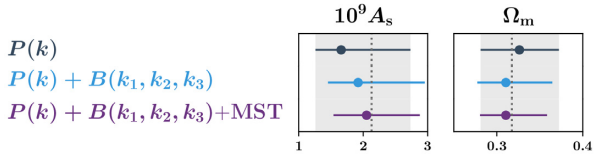

Figure 15. The 1σ constraints on A_s and Ω_m are shown for $P(k)$ (dark grey), $P(k) + B(k_1, k_2, k_3)$ (blue), and $P(k) + B(k_1, k_2, k_3)$ + MST (purple). This plot shows how including the MST improves constraints on A_s by $\sim$ 12 per cent ($\sim$ 10 per cent) and on Ω_m by $\sim$ 17 per cent ($\sim$ 12 per cent) with respect to $P(k)$ ($P(k) + B(k_1, k_2, k_3)$).

from the MST having access to smaller scales. Furthermore, the maximally compressed $B(k_1, k_2, k_3)$ has been shown by Gualdi et al. (2018) to improve parameter constraints by allowing the inclusion of many more triangle configurations than standard bispectrum analysis. Therefore, we can be fairly certain that the additional information is coming from the MST's detection of patterns in the cosmic web, information which would be present in higher order functions such as the trispectrum, thus confirming the heuristic arguments made in Section 3.1.

7 DISCUSSION

In this paper, we have sought to understand whether the MST can be used for parameter inference in cosmology. Until now, the MST has been predominantly used to search for large-scale features. This type of information has largely been overlooked as traditionally two-point statistics are completely insensitive to phase information. In constructing the MST we hope to pick up patterns in the cosmic web and use this to improve parameter constraints.

In Section 3, we argue heuristically why the MST should be sensitive to higher order statistics (i.e. three point and beyond). This is demonstrated using simulated galaxies (from the Illustris N-body simulation) and a random walk simulation (produced using an adjusted Lévy Flight algorithm) with virtually identical 2PCF by design but different higher order statistics.

In Section 4, we look at the effects of boundaries and masks, RSD, and scale cuts. Boundaries and masks[16] tended to produce longer edge lengths, whilst the degree and branch shape appeared to be unaffected. RSD is shown to have a significant impact on the MST statistics and thus should be incorporated in any future study. Lastly, we develop a strategy to impose a scale cut on the MST. This is done by removing edges below a set length in the kNN graph and then constructing the MST from this. Unfortunately this creates some artefacts in the degree and branch shape distributions. It is also believed that this method distorts some of the information we are trying to learn. As such alternatives or improvements to this method should be explored.

In Section 5, we look to determine what the MST actually measures, finding the MST to be highly sensitive to its local density. This is demonstrated by the fact that nodes in overdensities tended to have a degree of 2.

Lastly in Section 6, we determine whether the MST provides information not present in power spectrum and bispectrum. We do this by obtaining parameter constraints on A_s, Ω_m, and $\sum m_\nu$ for 10 halo mock catalogues. To keep the density of haloes the same in all our simulations we use only the most massive 5000 haloes and measure the power spectrum $P(k)$, bispectrum $B(k_1, k_2, k_3)$, and MST statistics. The individual methods provided similar constraints although due to the degeneracies with Ω_m we were unable to obtain meaningful constraints on $\sum m_\nu$. We found that combining the three methods narrows the 1σ constraints on Ω_m by $\sim$ 17 per cent and on A_s by $\sim$ 12 per cent with respect to $P(k)$ and $\sim$ 12 per cent on Ω_m and $\sim$ 10 per cent on A_s with respect to $P(k) + B(k_1, k_2, k_3)$, thus showing that the MST is providing information not present in the power spectrum or bispectrum. We expect this to improve with improved implementation of scale cuts and greater statistical power from larger samples.

The MST provides several advantages over existing methods but has some important limitations. The main advantages are: (1) it

[16]Boundaries can be thought of as a survey's footprint, whilst the mask would also include holes and varying completeness levels.

is sensitive to patterns in the cosmic web and (2) the algorithm is computationally inexpensive. The naive brute force implementation of N-point statistics for n points is an $\mathcal{O}(n^N)$ process. While there exist faster implementations of the 2PCF and 3PCF (see Scoccimarro 2015; Slepian & Eisenstein 2016), there are no such methods for higher order statistics. On the other hand, the MST is sensitive to higher order statistics and the Kruskal algorithm used here is approximately an $\mathcal{O}(n \log n)$ process. In the MST, we have a window into these higher order statistics but at a fraction of the computational cost. The main limitations of the MST: (1) we need simulations to estimate the statistics and (2) the statistic is dependent on the density of the tracer. This means we will need to create simulations that both match the survey properties as well as the density of the tracers used.

In future work, we look to apply the MST to current and future galaxy redshift surveys. In doing so we hope to better understand how to implement scale cuts and mitigate any of the resulting effects that occur as a result. One thing we have not studied in this paper is the effect of galaxy bias which should be explored in future. This could be achieved by varying HOD parameters. Lastly, ML algorithms and AI are powerful new tools to cosmology (see Ravanbakhsh et al. 2017; Fluri et al. 2018), however it is difficult to gain an intuition into what these algorithms are learning. Since the MST is relatively simple this could be used to gain insight into this work, providing a bridge between the traditional two-point and a full ML/AI approach.

Finally, the MST statistics presented in this paper have been produced by the PYTHON module `MiSTree` (Naidoo 2019), which implements the procedures detailed in Section 2. The module is made publicly available (see https://github.com/knaidoo29/mistree for documentation) and can handle data sets provided in 2D and 3D Cartesian coordinates, spherical polar coordinates, and coordinates on a sphere (either celestial RA, Dec., or simply longitude and latitude).

ACKNOWLEDGEMENTS

We thank Donnacha Kirk for his contributions to the early stages of this project and Niall Jeffrey for useful discussions.

Many of the figures in this paper were made using MATPLOTLIB[17] (Hunter 2007), whilst the corner plots were made using `ChainConsumer`[18] (Hinton 2016).

KN acknowledges support from the Science and Technology Facilities Council grant ST/N50449X. DG acknowledges support from European Union's Horizon 2020 research and innovation programme ERC (BePreSySe, grant agreement 725327), Spanish MINECO under projects AYA2014-58747-P AEI/FEDER, UE, and MDM-2014-0369 of ICCUB (Unidad de Excelencia María de Maeztu). OL acknowledges support from a European Research Council Advanced Grant FP7/291329 and from an STFC Consolidated Grant ST/R000476/1. MV is supported by INFN PD51 INDARK grant. AFR was supported by an STFC Ernest Rutherford Fellowship, grant reference ST/N003853/1.

REFERENCES

Abbott T. M. C. et al., 2018, Phys. Rev. D, 98, 043526
Abbott B. P. et al., 2017, Nature, 551, 85
Adami C., Mazure A., 1999, A&AS, 134, 393
Adami C. et al., 2010, A&A, 509, A81
Ahmad Q. R. et al., 2001, Phys. Rev. Lett., 87, 071301
Alam S. et al., 2017, MNRAS, 470, 2617
Allison R. J., Goodwin S. P., Parker R. J., Portegies Zwart S. F., de Grijs R., Kouwenhoven M. B. N., 2009, MNRAS, 395, 1449
Alpaslan M. et al., 2014, MNRAS, 438, 177
Alvarez M. A., Rosasco L., Lawrence N. D., 2012, Now Publishers, Inc., Foundations and Trends in Machine Learning, 4, 3, 195
Balázs L. G., Horváth I., Vavrek R., Bagoly Z., Mészáros A., 2008, in Galassi M., Palmer D., Fenimore E., eds, AIP Conf. Ser. Vol. 1000, New Statistical Results on the Angular Distribution of Gamma-Ray Bursts. p. 52 preprint (arXiv:0902.4812)
Barrow J. D., Bhavsar S. P., Sonoda D. H., 1985, MNRAS, 216, 17
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Beuret M., Billot N., Cambrésy L., Eden D. J., Elia D., Molinari S., Pezzuto S., Schisano E., 2017, A&A, 597, A114
Bhavsar S. P., Ling E. N., 1988, PASP, 100, 1314
Bhavsar S. P., Splinter R. J., 1996, MNRAS, 282, 1461
Bird S., Rogers K. K., Peiris H. V., Verde L., Font-Ribera A., Pontzen A., 2019, J. Cosmol. Astropart. Phys., 2019, 050
Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603
Campana R., Massaro E., Bernieri E., 2018a, Ap&SS, 363, 144
Campana R., Massaro E., Bernieri E., 2018b, A&A, 619, A23
Clarke S. D., Williams G. M., Ibáñez-Mejía J. C., Walch S., 2019, MNRAS, 484, 4024
Colberg J. M., 2007, MNRAS, 375, 337
Coles P., Pearson R. C., Borgani S., Plionis M., Moscardini L., 1998, MNRAS, 294, 245
Cormen T. H., Leiserson C. E., Rivest R. L., Stein C., 2009, Introduction to Algorithms. MIT Press, Cambridge, MA
Crocce M., Castander F. J., Gaztañaga E., Fosalba P., Carretero J., 2015, MNRAS, 453, 1513
Cybulski R., Yun M. S., Fazio G. G., Gutermuth R. A., 2014, MNRAS, 439, 3564
de Sainte Agathe V. et al., 2019, A&A, 629, A85
Demiański M., Doroshkevich A., Pilipenko S., Gottlöber S., 2011, MNRAS, 414, 1813
DESI Collaboration et al., 2016, preprint (arXiv:1611.00036)
Doroshkevich A. G., Müller V., Retzlaff J., Turchaninov V., 1999, MNRAS, 306, 575
Doroshkevich A. G., Tucker D. L., Fong R., Turchaninov V., Lin H., 2001, MNRAS, 322, 369
Durret F. et al., 2011, A&A, 535, A65
Fluri J., Kacprzak T., Refregier A., Amara A., Lucchi A., Hofmann T., 2018, Phys. Rev. D, 98, 123518
Font-Ribera A., McDonald P., Mostek N., Reid B. A., Seo H.-J., Slosar A., 2014, J. Cosmol. Astropart. Phys., 5, 023
Fukuda Y. et al., 1998, Phys. Rev. Lett., 81, 1562
Gil-Marín H., Percival W. J., Verde L., Brownstein J. R., Chuang C.-H., Kitaura F.-S., Rodríguez-Torres S. A., Olmstead M. D., 2017, MNRAS, 465, 1757
Gualdi D., Manera M., Joachimi B., Lahav O., 2018, MNRAS, 476, 4045
Gualdi D., Gil-Marín H., Schuhmann R. L., Manera M., Joachimi B., Lahav O., 2019, MNRAS, 484, 3713
Heavens A. F., Sellentin E., de Mijolla D., Vianello A., 2017, MNRAS, 472, 4244
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Hinton S. R., 2016, J. Open Source Softw., 1, 00045
Hong S., Coutinho B. C., Dey A., Barabási A.-L., Vogelsberger M., Hernquist L., Gebhardt K., 2016, MNRAS, 459, 2690
Howlett C., Manera M., Percival W. J., 2015, Astron. Comput., 12, 109
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Izard A., Crocce M., Fosalba P., 2016, MNRAS, 459, 2327
Jeffrey N., Abdalla F. B., 2019, MNRAS, 490, 5749
Kaiser N., 1987, MNRAS, 227, 1
Katriel I., Sanders P., Träff J. L., 2003, Algorithms - ESA 2003, Eur. Symp. Algorithms, Springer, Berlin, Heidelberg.. p. 679

[17] https://matplotlib.org/index.html
[18] https://samreay.github.io/ChainConsumer/index.html

Kruskal J. B., 1956, Proc. Am. Math. Soc., 7, 48
Krzewina L. G., Saslaw W. C., 1996, MNRAS, 278, 869
Libeskind N. I. et al., 2018, MNRAS, 473, 1195
Loureiro A. et al., 2019, MNRAS, 485, 326
Mandelbrot B. B., 1982, The Fractal Geometry of Nature.W. H. Freeman and Company, New York
Martinez V. J., Jones B. J. T., 1990, MNRAS, 242, 517
Massara E., Villaescusa-Navarro F., Viel M., Sutter P. M., 2015, J. Cosmol. Astropart. Phys., 11, 018
Naidoo K., 2019, J. Open Source Softw., 4, 1721
Nelson D. et al., 2015, Astron. Comput., 13, 12
Palanque-Delabrouille N. et al., 2015, J. Cosmol. Astropart. Phys., 11, 011
Park D., Lee J., 2009, MNRAS, 397, 2163
Pearson R. C., Coles P., 1995, MNRAS, 272, 231
Planck Collaboration et al.2016, A&A, 594, A13
Planck Collaboration et al., 2018, preprint (arXiv:1807.06209)
Rainbolt J. L., Schmitt M., 2017, JINST, 12, P02009
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning, MIT Press, Cambridge, Massachusetts
Ravanbakhsh S., Oliva J., Fromenteau S., Price L. C., Ho S., Schneider J., Poczos B., 2017, ICML, 2407 (arXiv:1711.02033)
Riess A. G. et al., 2016, ApJ, 826, 56
Rodríguez-Torres S. A. et al., 2016, MNRAS, 460, 1173
Rogers K. K., Peiris H. V., Pontzen A., Bird S., Verde L., Font-Ribera A., 2019, J. Cosmol. Astropart. Phys., 2019, 031
Scoccimarro R., 2015, Phys. Rev. D, 92, 083532
Sellentin E., Heavens A. F., 2016, MNRAS, 456, L132
Sellentin E., Starck J.-L., 2019, J. Cosmol. Astropart. Phys., 08, 021
Shim J., Lee J., 2013, ApJ, 777, 74
Shim J., Lee J., Li B., 2014, ApJ, 784, 84
Shim J., Lee J., Hoyle F., 2015, ApJ, 815, 107
Slepian Z., Eisenstein D. J., 2016, MNRAS, 455, L31
Slepian Z. et al., 2017, MNRAS, 468, 1070
Springel V., 2005, MNRAS, 364, 1105
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, J. Cosmol. Astropart. Phys., 6, 036
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
Tojeiro R. et al., 2014, MNRAS, 440, 2222
Ueda H., Itoh M., 1997, PASJ, 49, 131
van de Weygaert R., Jones B. J., Martínez V. J., 1992, Phys. Lett. A, 169, 145
Viel M., Haehnelt M. G., Springel V., 2010, J. Cosmol. Astropart. Phys., 6, 015
Vogelsberger M. et al., 2014, Nature, 509, 177
Winther H. A., Koyama K., Manera M., Wright B. S., Zhao G.-B., 2017, J. Cosmol. Astropart. Phys., 8, 006
Wright B. S., Winther H. A., Koyama K., 2017, J. Cosmol. Astropart. Phys., 10, 054

APPENDIX A: GAUSSIAN PROCESS INTERPOLATION

We will be modelling data vectors following a method similar to that of Rogers et al. (2019) and Bird et al. (2019) in which they emulated the 1D flux power spectrum of the Lyman-α forest using GPs. In this section, we provide a brief introduction to GPs and outline their usage in this paper. A comprehensive overview of GPs and their applications can be found in Rasmussen & Williams (2006), while an overview of their implementations for vectors can be found in Alvarez, Rosasco & Lawrence (2011).

A1 Introduction

GPs are a non-parametric kernel-based regression and interpolation method. In GPs we model the desired function $f(x)$ as a stochastic process with a prior probability over all parametric functions. For a finite input data set $\boldsymbol{X} = \{x_1, ..., x_n\}$, this can be modelled as a multivariate Gaussian,

$$\mathcal{GP} = \mathcal{N}\left(\boldsymbol{m}(\boldsymbol{X}), \mathsf{K}\left(\boldsymbol{X}, \boldsymbol{X}'\right)\right), \tag{A1}$$

with mean $\boldsymbol{m}(\boldsymbol{X})$ and covariance $\mathsf{K}(\boldsymbol{X}, \boldsymbol{X}')$. Given training data $\boldsymbol{Y}_1$ at $\boldsymbol{X}_1$, we model the posterior of the function $f(x)$ at new positions $\boldsymbol{X}_2$ as a multivariate Gaussian,

$$P\left(\boldsymbol{Y}_2|\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{X}_2\right) = \mathcal{N}\left(\boldsymbol{\mu}_{2|1}, \mathsf{S}_{2|1}\right), \tag{A2}$$

with mean $\boldsymbol{\mu}_{2|1}$ and covariance $\mathsf{S}_{2|1}$. Assuming that both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are drawn from the same multivariate Gaussian, as our prior on the function indicates (see equation A1), we can write the relation

$$\begin{bmatrix}\boldsymbol{Y}_1\\ \boldsymbol{Y}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\mathsf{K}_{11} + \mathsf{I}\sigma_{\rm n}^2 & \mathsf{K}_{12}\\ \mathsf{K}_{21} & \mathsf{K}_{22}\end{bmatrix}\right), \tag{A3}$$

where I is the identity matrix and $\sigma_{\rm n}$ is the standard deviation of the training data $\boldsymbol{Y}_1$ (which is either known or fitted later). Thus, assuming the mean function is zero we arrive at the predicted mean and covariance,

$$\boldsymbol{\mu}_{2|1} = \left[\left(\mathsf{K}_{11} + \mathsf{I}\sigma_{\rm n}^2\right)^{-1}\mathsf{K}_{12}\right]^{\top}\boldsymbol{Y}_1, \tag{A4}$$

$$\mathsf{S}_{2|1} = \mathsf{K}_{22} - \left[\left(\mathsf{K}_{11} + \mathsf{I}\sigma_{\rm n}^2\right)^{-1}\mathsf{K}_{12}\right]^{\top}\mathsf{K}_{12}, \tag{A5}$$

where the dependence on K_{21} has been removed due to the symmetry $\mathsf{K}_{12} = \mathsf{K}_{21}^{\top}$. Note that in practice we determine the GPs mean and standard deviation at a single new position and thus the standard deviation is simply a scalar – this means that K_{12} and K_{21} reduce to vectors and K_{22} to a scalar.

A2 Kernel

GPs use kernels to weight the interdependency of points in parameter space. In our model, we use a Gaussian kernel,

$$\kappa(\theta_{\rm i}, \theta_{\rm j}) = \sigma_{\rm GP}^2 \exp\left(-\frac{r^2}{2}\right). \tag{A6}$$

Here,

$$r = \frac{|\theta_{\rm i,1} - \theta_{\rm j,1}|^2}{2l_{\rm GP,1}^2} + \frac{|\theta_{\rm i,2} - \theta_{\rm j,2}|^2}{2l_{\rm GP,2}^2} + \frac{|\theta_{\rm i,3} - \theta_{\rm j,3}|^2}{2l_{\rm GP,3}^2}; \tag{A7}$$

$\sigma_{\rm GP}$, $l_{\rm GP,\,1}$, $l_{\rm GP,\,2}$, and $l_{\rm GP,\,3}$ are GPs hyperparameters to be fitted with independent scale terms for each axis in the parameter space; and $\boldsymbol{\theta} = \left[10^9 A_{\rm s}, \Omega_{\rm m}, m_{\nu}\right]$. The covariance matrix K is then defined to have elements

$$(\mathsf{K})_{\rm ij} = \kappa(\theta_{\rm i}, \theta_{\rm j}) + \sigma_{\rm n}^2\delta_{\rm k}(\theta_{\rm i}, \theta_{\rm j}), \tag{A8}$$

with an additional noise term $\sigma_{\rm n}$.

A3 Hyperparameter optimization

The hyperparameters $\boldsymbol{\phi} = [\sigma_{\rm GP}, l_{\rm GP,1}, l_{\rm GP,2}, l_{\rm GP,3}]$ are optimized by maximizing the likelihood function

$$\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i}^{n}\mathcal{L}(\boldsymbol{d}_{\rm i}|\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{A9}$$

where $\boldsymbol{D}$ are the ensemble of training data vectors, $\boldsymbol{d}_{\rm i}$ is an element of a specific data vector and

$$\mathcal{L}(\boldsymbol{d}_{\rm i}|\boldsymbol{\theta}, \boldsymbol{\phi}) = -\frac{1}{2}\boldsymbol{d}_{\rm i}^{\top}\mathsf{K}^{-1}\boldsymbol{d}_{\rm i} - \frac{1}{2}\log|\mathsf{K}| - \frac{n}{2}\log 2\pi. \tag{A10}$$

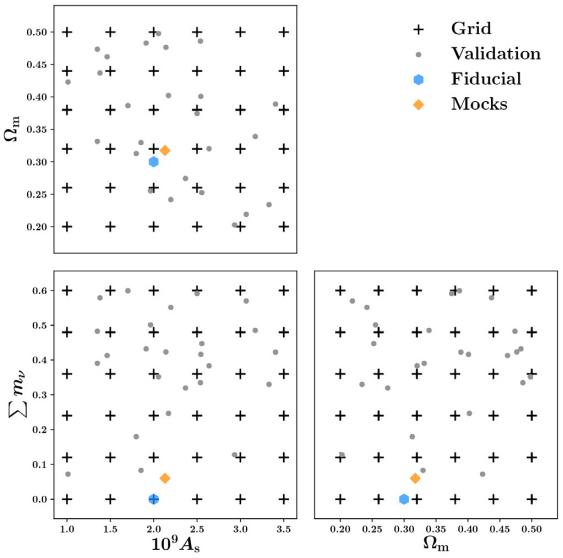

Figure A1. The positions in parameter space of simulations (grid, validation, fiducial, and mocks) used in Section 6. Note that for the grid simulations each cross marks the point of five simulations.

A4 Implementation and validation

The GPs hyperparameters are trained on the measurements of $P(k)$, the maximally compressed $B(k_1, k_2, k_3)$, and the MST statistics d, l, b, and s (see Section 6.2 for further details on these measurements) from the grid simulations separately. In Fig. A1, we show the placement of the grid, fiducial, mock, and validation (used only in this section) simulations in parameter space. To test that our GPs interpolation is emulating the statistics accurately we calculate the residuals between the grid simulations (using the mean of five realizations made at each point in parameter space),

$$\boldsymbol{\sigma}_{\rm Residual} = \frac{\boldsymbol{d} - \boldsymbol{\mu}_{\rm GP}}{\sqrt{\boldsymbol{\sigma}^2_{\rm Fiducial} + \boldsymbol{\sigma}^2_{\rm GP}}}, \tag{A11}$$

where $\boldsymbol{\mu}_{\rm GP}$ and $\boldsymbol{\sigma}_{\rm GP}$ are the GPs mean and standard deviation evaluated at the same points in parameter space as $\boldsymbol{d}$. We plot histograms of the residuals for the grid data vectors in Fig. A2 shown in orange. Notice that since the grid simulations are the mean of five simulations, the distribution follows a Gaussian with mean 0 and standard deviation $1/\sqrt{5}$ (illustrated by the black dotted line). Furthermore to test that our GPs interpolation produces a good fit to simulations not present in the training data, we generate 25 new simulations (called the validation simulations) with randomly drawn cosmological parameters (shown in Fig. A1). We then again compare the residuals to that of our GPs interpolation and find a good agreement (with the exception of $B(k_1, k_2, k_3)$) with a Gaussian with mean 0 and standard deviation 1 illustrated by the black full lines.

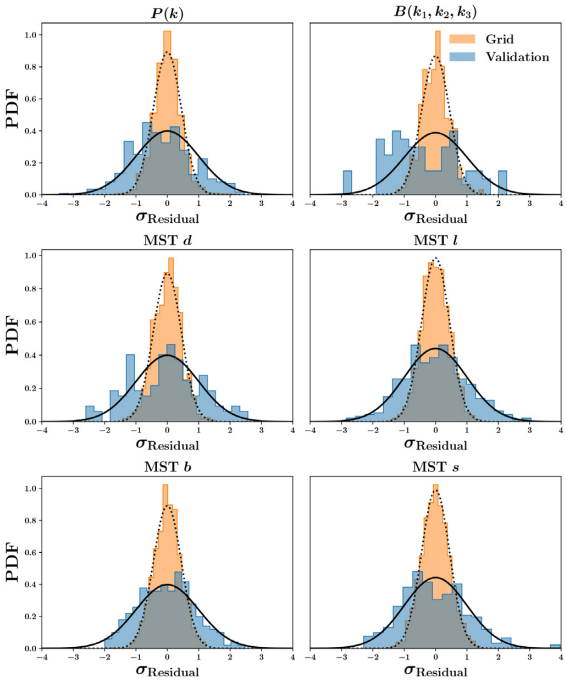

Figure A2. The residuals between the statistics of $P(k)$ (top left), maximally compressed $B(k_1, k_2, k_3)$ (top right), MST degree (middle left), edge length (middle right), branch length (bottom left), and branch shape (bottom right) for the grid (shown by the orange histograms) and validation (shown by the blue histograms) simulations calculated from equation (A11). Since the grid data vectors are the mean of five realizations the residuals are expected to follow a normal distribution of $\mathcal{N}(0, 1/\sqrt{5})$ (shown by the dotted black line), whilst the validation data vector are expected to follow a normal distribution of $\mathcal{N}(0, 1)$. We see that for most of the statistics the agreement is fairly good, with the exception of $B(k_1, k_2, k_3)$ which shows more spread than is expected.

This paper has been typeset from a TEX/LATEX file prepared by the author.