Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **471,** 1167–1180 (2017)
Advance Access publication 2017 July 4

doi:10.1093/mnras/stx1682

Matched filtering with interferometric 21 cm experiments

Martin White[1,2,3]⋆ and Nikhil Padmanabhan[4]

[1] *Department of Astronomy, University of California, Berkeley, CA 94720, USA*
[2] *Department of Physics, University of California, Berkeley, CA 94720, USA*
[3] *Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[4] *Department of Physics, Yale University, New Haven, CT 06511, USA*

Accepted 2017 June 30. Received 2017 June 23; in original form 2017 May 26

ABSTRACT

A new generation of interferometric instruments is emerging, which aims to use intensity mapping of redshifted 21 cm radiation to measure the large-scale structure of the Universe at $z \simeq 1$–2 over wide areas of the sky. While these instruments typically have limited angular resolution, they cover huge volumes and thus can be used to provide large samples of rare objects. In this paper we study how well such instruments could find spatially extended large-scale structures, such as cosmic voids, using a matched filter formalism. Such a formalism allows us to work in Fourier space, the natural space for interferometers, and to study the impact of finite $u - v$ coverage, noise and foregrounds on our ability to recover voids. We find that in the absence of foregrounds, such instruments would provide enormous catalogs of voids, with high completeness, but that control of foregrounds is key to realizing this goal.

Key words: gravitation – galaxies: haloes – galaxies: statistics – cosmological parameters – large-scale structure of Universe.

1 INTRODUCTION

Intensity mapping with radio interferometers has emerged as a potentially powerful means of efficiently mapping large volumes of the Universe, albeit at low spatial resolution. Several groups are fielding 21 cm intensity mapping experiments using a variety of technical designs (Ansari et al. 2012; Chen 2012; Pober et al. 2013; Ali & Bharadwaj 2014; Vanderlinde & Chime Collaboration 2014; Newburgh et al. 2016). Even though such instruments do not have the angular resolution to see individual galaxies, or even large clusters, they are capable of mapping the larger elements of the cosmic web (e.g. protoclusters and cosmic voids). Protoclusters are the progenitors of the most massive systems in the Universe today (Overzier 2016). Cosmic voids make up most of the Universe, by volume (Rood 1988; van de Weygaert & Platen 2011). As we shall show, in each case the system size is sufficiently large that they can be reliably found with upcoming intensity mapping experiments if foregrounds can be controlled sufficiently well.

Cosmic voids, regions almost devoid of galaxies, are intrinsically interesting as the major constituent of the cosmic web by volume, and as an extreme environment for galaxy evolution (Rood 1988; van de Weygaert & Platen 2011). They may be an excellent laboratory for studying material that clusters weakly like dark energy (Lee & Park 2009; Lavaux & Wandelt 2012) or neutrinos (Banerjee & Dalal 2016)) or for testing modified gravity (Clampitt, Cai & Li 2013; Hamaus, Sutter & Wandelt 2014; Cai, Padilla & Li 2015; Hamaus et al. 2015; Cai et al. 2016; Falck et al. 2017; Hamaus et al. 2017). In this paper we show that 21 cm instruments aimed at measuring the large-scale power spectrum, either proposed or under construction, should enable a search of enormous cosmic volumes at high redshift for these rare objects, which are large enough to be detected at high significance (see also Battye, Davies & Weller 2004).

Voids and protoclusters are inherently ‘configuration space objects’, in the sense of being highly coherent under- or overdensities in the matter field in configuration space. However, most future 21 cm experiments are interferometers that naturally work in Fourier space. We will use a matched filter formalism to allow us to work in the interferometer’s natural space, where the noise and sampling are easy to understand. This formalism also provides a natural way to combine data sets that live in different domains, e.g. optical imaging data with 21 cm interferometry.

We will illustrate our ideas by focusing on cosmic voids, though much of what we say could be applied to protoclusters as well. Our goal in this paper will thus be the detection of voids, and the challenges associated with this. We assume that these candidate voids will be appropriately analysed or followed up for different science applications. It is worth keeping in mind that for certain applications, the intermediate step of constructing an explicit void catalog and characterizing its purity and completeness may not be necessary. One might be able to construct estimators of the quantities of interest directly from the visibilities. We shall not consider such approaches in this paper.

The outline of the paper is as follows. In Section 2 we establish our notation and provide some background on interferometry, foregrounds for 21 cm experiments and matched filters. Section 3 describes the numerical simulations that we use to test our matched

⋆ E-mail: mwhite@berkeley.edu

Published by Oxford University Press on behalf of the Royal Astronomical Society

Table 1. Useful quantities and conversion factors as a function of redshift, z, assuming a flat ΛCDM model with $\Omega_m = 0.3$. These are (a) the observing frequency, ν, in MHz for 21 cm radiation emitted at z, (b) the comoving distance to z, χ, in h^{-1} Mpc, (c) the foreground wedge angle [equation 6] (d) the k mode that which a 10 m baseline maps to at redshift z and (e) the differential conversion from frequency (in MHz) to comoving distance (in h^{-1} Mpc).

z	ν [MHz]	χ [h^{-1} Mpc]	μ	k/D_{10} [hMpc^{-1}]	$\lvert d\chi/d\nu\rvert$ [h^{-1} Mpc MHz^{-1}]
0.50	947	1322	0.359	0.1509	3.630
0.75	811	1854	0.473	0.0922	4.256
1.00	710	2313	0.562	0.0647	4.796
1.25	631	2709	0.632	0.0491	5.267
1.50	568	3055	0.687	0.0392	5.685
1.75	516	3358	0.732	0.0324	6.061
2.00	473	3626	0.767	0.0275	6.405
2.25	437	3865	0.796	0.0238	6.724
2.50	406	4079	0.820	0.0210	7.023

filter and the profiles of voids in those simulations. Our main results are given in Section 4, and we present our conclusions in Section 5. We relegate a number of technical details to a series of appendices. In particular Appendix A discusses instrument noise for an interferometer in the cosmological context, Appendix B describes the formalism of transiting telescopes (including cylinder telescopes) and the flat-sky limit and Appendix C discusses the manner in which neutral hydrogen might be expected to trace the matter field at intermediate redshift.

2 BACKGROUND AND REVIEW

In this section we provide some background information, to set notation and provide an easy reference for our later derivations.

2.1 Visibilities

In an interferometer the fundamental datum is the correlation between two feeds (or antennas), known as a visibility. For an intensity measurement the visibility is (Thompson, Moran & Swenson 2017)

$$V_{ij} \propto \int \mathrm{d}^2n\ A^2(\hat{\boldsymbol{n}})T(\hat{\boldsymbol{n}})\,\mathrm{e}^{2\pi i\hat{\boldsymbol{n}}\cdot\boldsymbol{u}_{ij}}, \tag{1}$$

where $T(\hat{\boldsymbol{n}})$ is the brightness temperature in the sky direction $\hat{\boldsymbol{n}}$, $A(\hat{\boldsymbol{n}})$ is the primary beam (assumed the same for all feeds) and $\boldsymbol{u}_{ij}$ is the difference in position vectors of the i^{th} and j^{th} feeds in units of the observing wavelength. It is common to normalize the visibilities so that they return brightness temperature. We convert from brightness temperature to cosmological overdensity throughout, so we omit the exact normalization here. We will work in visibility space, since this is the natural space for the interferometer and has the simplest noise properties. Some useful conversions between common quantities are given in Table 1.

Visibilities are measured over a range of frequencies, and we shall follow the common procedure in 21 cm studies of Fourier transforming in the frequency direction to obtain a data cube in 3D Fourier space, $\boldsymbol{k}$. The conversion from frequency to distance (and hence Fourier mode) is

$$\left|\frac{\mathrm{d}\chi}{\mathrm{d}\nu}\right| = \frac{c}{H(z)}\,\frac{(1+z)^2}{\nu_0} \tag{2}$$

with $\nu_0 = 1420$ MHz.

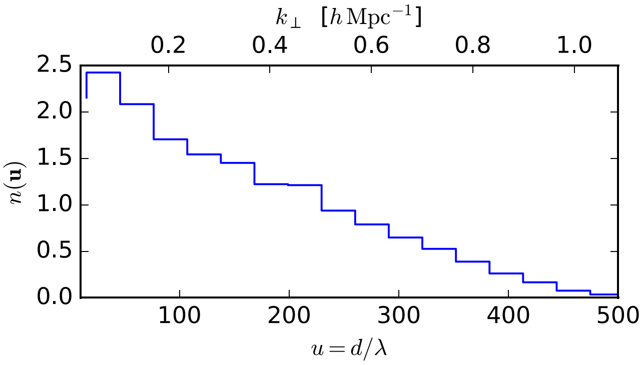

Figure 1. The circularly averaged baseline distribution for an HIRAX-like experiment. The lower axis shows the length baseline separation in units of the wavelength for $\lambda = 42$ cm (i.e. $z = 1$) while the upper axis converts to $k_\perp$ in h Mpc^{-1}. The distribution is normalized to integrate to the number of antenna pairs.

For small sky areas, the visibility thus measures the Fourier transform of the sky signal, apodized by the primary beam. Approximating the sky as flat and assuming the signal of interest, τ, is azimuthally symmetric (see Appendix B)

$$\tau(k) = 2\pi \int \tilde{\omega}\,\mathrm{d}\tilde{\omega}\ J_0(k\,\tilde{\omega})\,\tau(\tilde{\omega}), \tag{3}$$

where $\tilde{\omega}$ is an angular, radial coordinate and

$$V_{ij} \propto [\tau \star B]\,(2\pi u_{ij}) \tag{4}$$

with the ⋆ representing a convolution and τ and B being the Fourier transforms of $\tau(\hat{\boldsymbol{n}})$ and $A^2(\hat{\boldsymbol{n}})$, respectively. The surveys of interest to us here will cover large sky areas. The (very correlated) visibilities from the different pointings can be combined to produce higher resolution in the $u-v$ plane (a process known as mosaicking), entirely analogously to the manner in which the many slits in a diffraction grating sharpen the transmitted lines (e.g. Thompson et al. 2017, or see the discussion in White et al. 1999 for the cosmological context). In such a case, the effective B is determined by the survey area rather than the primary beam (analogous to a survey window in a galaxy survey) and will be very small. Combined with the fact that our signals will be very smooth in $u-v$, this allows us to neglect B to simplify our presentation. Reinstating it does not change any of our conclusions.

2.2 21 cm interferometers

We shall start by considering an interferometer consisting of an array of dishes (the interesting case of transiting, cylinder telescopes presents only technical modifications and is described in Appendix B). As a concrete example, we use the HIRAX experiment (Newburgh et al. 2016). HIRAX will use 1024 6 m parabolic dishes in a compact grid covering the frequency range $400 < \nu < 800$ MHz (i.e. $0.8 < z < 2.5$ for 21 cm radiation). HIRAX is a transit telescope: all dishes will be pointed at the meridian with a given declination, and the sky will rotate overhead in a constant drift-scan. Each declination pointing will give access to a 6° wide stripe of the sky and the complete survey will cover 15 000 square degrees.

Fig. 1 plots the circularly averaged distribution of baselines at $z = 1$, given our assumptions for HIRAX. The x-axis is the baseline separation in units of the wavelength, $|\boldsymbol{u}|$, while the y-axis is the number density of baselines per d^2u, conventionally normalized to integrate to the number of antenna pairs. The upper x-axis shows what k modes these baselines map to. Recalling that the noise scales as the inverse of the baseline density (see Appendix A), we see that an HIRAX-like experiment is sensitive to a broad range of k scales well suited to the detection of voids and protoclusters.

2.3 Noise for 21 cm experiments

The major difficulty facing upcoming 21 cm experiments is astrophysical foregrounds (e.g. Furlanetto, Oh & Briggs 2006; Shaw et al. 2014, 2015; Pober 2015; Seo & Hirata 2016). Foregrounds have been extensively studied in the context of (high z) epoch of re-ionization studies, i.e. at lower frequencies than of direct interest here. However the amplitudes of the signal and foreground scale in a roughly similar manner with frequency, so many of the lessons hold in our case (see Pober 2015, for a recent discussion). Since the main (Galactic) foregrounds are relatively smooth in frequency, their removal impacts primarily the slowly varying modes along the line of sight, i.e. the low $k_\parallel$ modes. However since the foregrounds are very bright (compared to the signal) and no instrument can be characterized perfectly, there is also some leaking of foreground power into other parts of the $k_\perp - k_\parallel$ plane.

The precise range of scales accessible to 21 cm experiments after foreground removal is currently a source of debate. We do not attempt to model foreground subtraction explicitly, but take into account its effects by restricting the range of the $\boldsymbol{k}_\perp - k_\parallel$ plane we use.

There are two regions of this plane we could lose to foreground removal. The first is low $k_\parallel$ modes, i.e. modes close to transverse to the line of sight. This boundary is slightly fuzzy and not well known. For instance, Shaw et al. (2014, 2015) claim that foreground removal leaves modes with $k_\parallel > 0.02\,h\mathrm{Mpc}^{-1}$ available for cosmological use, while Pober (2015) claims $k_\parallel < 0.1\,h\mathrm{Mpc}^{-1}$ modes are unusable. We shall consider the impact of a $k_\parallel$ cut within this range and we will see that our ability to find voids is quite sensitive to this cut.

In addition to low $k_\parallel$, non-idealities in the instrument lead to leakage of foreground information into higher $k_\parallel - k_\perp$ modes. This is usually phrased in terms of a foreground 'wedge' (for recent discussions, see Pober 2015; Shaw et al. 2015; Cohn et al. 2016; Seo & Hirata 2016, and references therein). The wedge does not form a hard boundary, but delineates a region where modes far from the line-of-sight direction can become increasingly contaminated. For a spatially flat Universe we can define the wedge geometrically as (Cohn et al. 2016; Seo & Hirata 2016)

$$\mathcal{R} = \frac{\chi H}{c(1+z)} = \frac{E(z)}{1+z}\int_0^z \frac{\mathrm{d}z'}{E(z')}, \tag{5}$$

where $E(z) = H(z)/H_0$ is the evolution parameter, and we assume we cannot access the signal in modes with $|k_\parallel|/k_\perp < \mathcal{R}$ or

$$\frac{|k_\parallel|}{k} < \mu_{\min} = \frac{\mathcal{R}}{\sqrt{1+\mathcal{R}^2}} \approx 0.6 \tag{6}$$

with the last step being for $z = 1$. It is worth emphasizing that this foreground 'wedge' does not represent a fundamental loss of information, and may be mitigated with an improved model of the instrument (ideally the wedge can be reduced by $\sin\Theta$, where Θ is the field of view; Liu, Zhang & Parsons 2016). We bracket these cases by considering cases with and without the foreground wedge, and discuss the impact on our void finder.

Finally we must contend with shot noise and receiver noise in the instrument. Castorina & Villaescusa-Navarro (2016) argue that shot noise is sub-dominant to receiver noise for upcoming surveys, so we shall neglect it in what follows (see also Cohn et al. 2016). To simplify our presentation, we shall treat the receiver noise as uncorrelated between visibilities and constant for all pairs of receivers. The noise thus scales with the number of baselines that probe a particular scale, and only an overall scaling is required. If the noise is uncorrelated from frequency channel to frequency channel, and only slowly varying with frequency, then the noise level is independent of $k_\parallel$. It is convenient to quote the thermal noise power in terms of the linear theory power spectrum, P_L, in much the same way as galaxy surveys specify their shot noise by giving $\bar{n}P$ at some fiducial scale. Since one of the design goals of all of these surveys is a measurement of the baryon acoustic oscillation (BAO) scale, we follow the standard practice and specify the receiver noise as a fraction of P_L at $k_{\perp,\,\mathrm{fid}} = 0.2\,h\,\mathrm{Mpc}^{-1}$. The surveys should achieve $P_L/P_{\mathrm{noise}} > 1$ at $k_{\perp,\,\mathrm{fid}}$ and we shall explore a range of values (see Appendix A). Once $P_L/P_{\mathrm{noise}} > 3$, the results become very insensitive to the precise value.

2.4 Matched filters

A matched filter is a convenient means of finding a signal of known shape in a noisy data set. If we write the the data as an amplitude times a template plus Gaussian noise ($d = A\tau + n$), the maximum-likelihood estimate of A and its scatter is given by

$$\hat{A} = \frac{\tau N^{-1} d}{\tau N^{-1}\tau}, \qquad \sigma^{-2} = \tau N^{-1}\tau. \tag{7}$$

We take the 'noise' covariance to include both instrument noise and non-template cosmological signal and shall assume throughout that this noise is diagonal in k-space. The main feature of this expression is that areas of the k-plane that are not sampled or are lost to foregrounds receive zero weight ($N^{-1} = 0$).

We find that our ability to isolate voids is very insensitive to the exact profile chosen for τ. In fact, even a top-hat profile produces a highly pure and complete void catalog for low noise and good k-space sampling. Similarly the performance is not particularly sensitive to the particular choice for $N(k)$, but rather to the larger questions of whether there are significant regions of k-space where $N^{-1} = 0$ or very uneven sensitivity of the instrument due to the spacing of the feeds.

We shall use N-body simulations for our signal, and work with a periodic, cubic box. In such situations, given a 3D density field, $\delta(\boldsymbol{x})$, and a template, $\tau(\boldsymbol{x})$, we can implement the flat-sky version of the matched filter very efficiently using FFTs if the noise is diagonal in k-space. Recalling that a shift in configuration space amounts to multiplication by a phase in Fourier space, the matched filter for a void centred at $\boldsymbol{a}$ is

$$\tau N^{-1} d \to \sum_{\boldsymbol{k}} \mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{a}}\, \frac{\tau_0(\boldsymbol{k})\delta^\star(\boldsymbol{k})}{N(\boldsymbol{k})}, \tag{8}$$

where τ_0 is the template for a void centred at the origin. The sum is simply an (inverse) Fourier transform, so we can test for all $\boldsymbol{a}$ at once. A similar set of steps can be used for the denominator $\tau N^{-1}\tau$, allowing a fast computation of S/N for any position, $\boldsymbol{a}$. Thus with forward Fourier transforms of the template and data and one inverse transform, we can compute the matched filter amplitude, A, everywhere in space and hence its (volume weighted) distribution at random locations and at the positions of voids.

There is, in principle, no reason why the matched filter cannot be modified to remove spectrally smooth foregrounds, at the same time as searching for voids or protoclusters. We choose not to implement

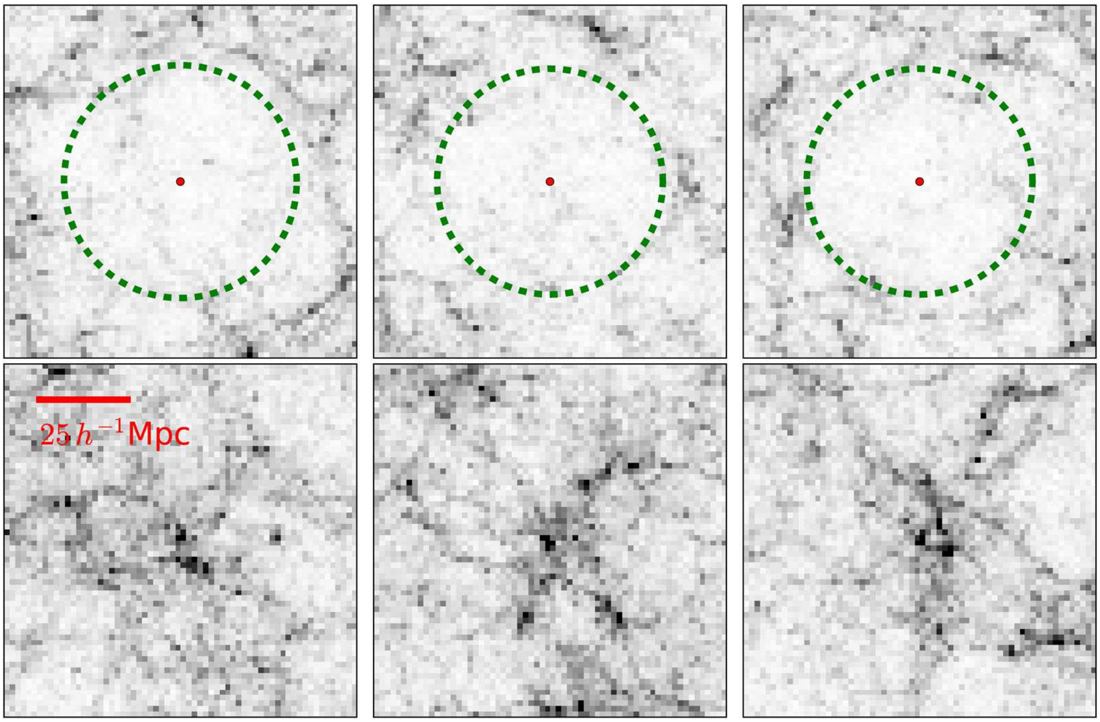

Figure 2. Slices through one of our N-body simulations at $z = 1$. Each panel shows the projected, redshift-space density field, on an arcsinh scale saturating at $10\,\bar{\rho}$, with a (line-of-sight) depth of $\pm 20\,h^{-1}$ Mpc and transverse dimensions $\pm 50\,h^{-1}$ Mpc. The panels in the top row are centred on voids, with the void centre marked with a dot and the radius with a dashed circle. Those in the bottom row are centred on (randomly selected) massive protoclusters.

this approach, preferring instead to set $N^{-1} = 0$ for modes that we deem unusable due to foregrounds.

3 SIMULATIONS

3.1 *N*-body

To illustrate our ideas, we make use of several N-body simulations, each of the ΛCDM family. Specifically we use the $z = 1$ outputs of 10 simulations run with the *TreePM* code described in White (2002). This code has been extensively compared to other N-body codes in Heitmann et al. (2008), and these simulations have been previously used (and described) in Reid et al. (2014) and White et al. (2015). Each run utilized 2048^3 particles in a periodic box of side $1380\,h^{-1}$ Mpc to model a cosmology with $\Omega_m = 0.292$, $h = 0.69$ and $\sigma_8 = 0.82$. This gives a particle mass of $m_p = 2.5 \times 10^{10}\,h^{-1}\,\mathrm{M}_{\odot}$. Slices of the redshift-space density field at $z = 1$ from one of the simulations are shown in Fig. 2, to illustrate the types of structures we are searching for. There are spatially coherent regions of over- or underdensity with scales of $\mathcal{O}(10\,\mathrm{Mpc})$ clearly visible in the figure.

Properly modelling the distribution of HI at z =1–2 is beyond the scope of this paper. Our simulations would need much higher resolution, to resolve the haloes likely to host neutral hydrogen at z = 1–2, and the halo occupancy is anyway highly uncertain (see discussion in e.g. Castorina & Villaescusa-Navarro 2016; Seehars et al. 2016). Instead we assume that the HI is an unbiased tracer of the matter field, and simply use the dark matter density. In Appendix C we use a halo model of HI in a higher resolution (but smaller volume) simulation to show that this is a conservative approximation for the purposes of establishing how well 21 cm experiments can find voids.

3.2 Voids in the simulations

We define voids through a spherical underdensity algorithm (for a comparison with other void finders, see Stark et al. 2015b, and for a general comparison of void finders, see Colberg et al. 2008). The dark matter particles are binned on to a regular, Cartesian grid of 1380^3 points. Around each density minimum with $1 + \delta < 0.2$, we grow a sphere until the mean enclosed density is $1 + \bar{\delta} < 0.4$. Visually such an underdensity gives voids that match expectations (see Fig. 2). The voids are then ordered by their radius R_V and overlapping voids with smaller radii are removed from the list. As is the case for the large overdensities (protoclusters), these large underdensities (voids) are very rare, necessitating surveys of large volumes. The number density of redshift-space voids at $z = 1$ is $10^{-5}\ h^{-3}\,\mathrm{Mpc}^3$ for $10 < R_v < 15\,h^{-1}$ Mpc and 6×10^{-7} for $20 < R_v < 25\,h^{-1}$ Mpc and falls quickly with redshift.

The matched filter essentially performs a ‘weighted convolution’ of the density field with a profile, and thus requires some knowledge of the shape of the object it is trying to ‘match’. While the performance of the filter is relatively insensitive to the precise profile we use, we describe the choices we have made based on the N-body simulations described above.

To begin, we note that a void has an extent $\mathcal{O}(10\,\mathrm{Mpc})$ and thus covers only a small region of sky (<1 arcminute) and a small portion of the frequency coverage of the telescope. We are thus justified in treating the sky as locally flat and the $\boldsymbol{k}_\perp$ coverage as approximately wavelength independent.[1] We expect the profile to have a significant power at $k \sim 0.1\,h\,\mathrm{Mpc}^{-1}$, well within the band of

[1] Recall that the conversion from $\boldsymbol{u}$ to $\boldsymbol{k}_\perp$ depends on the frequency of the observation.

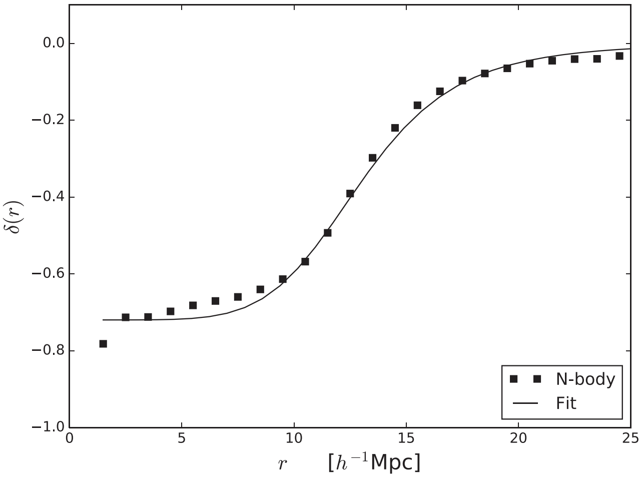

Figure 3. Real-space profile of voids in our simulation. Squares show the average profile measured from the *N*-body simulations described in the text for voids with radii $10 < R_{\rm v} < 15\,h^{-1}$ Mpc at $z = 1$. The solid line is the analytic fit of equation 9 with $r_{\rm s} = 12.5\,h^{-1}$ Mpc.

sensitivity of 21 cm interferometers aimed at large-scale structure observations (see Fig. 1 before).

Fig. 3 shows the averaged real space profile of voids with $10 < R_V < 15\;h^{-1}$ Mpc in our *N*-body simulations at $z = 1$. There are numerous analytic void profiles in the literature (e.g. Hamaus et al. 2014; Hawken et al. 2016, for recent examples). Most of these do not fit our *N*-body results particularly well, which is most likely due to the different choices of void finder employed. In particular, our void profile approaches 0 smoothly from below at large radius, i.e. we do not find a prominent ‘compensation wall’ at the void edge. This echoes the findings of Cai et al. (2016), who also found no compensation wall for voids that are not part of a larger overdensity.

A simple, 2-parameter form that does provide a good fit to our *N*-body data is

$$\delta = \frac{\delta_0}{1 + (r/r_{\rm s})^6}, \tag{9}$$

where δ_0 and $r_{\rm s}$ are the interior underdensity and void scale radius, respectively. This is shown in Fig. 3 as the solid line. In Fourier space, this profile becomes

$$\tau(k) = 4\pi \int r^2\,{\rm d}r\;\delta(r)\;j_0(kr) \tag{10}$$

$$= \frac{2\pi^2}{3\,\kappa}\delta_0\,r_s^3\,{\rm e}^{-\kappa/2}\left[{\rm e}^{-\kappa/2} + \sqrt{3}\sin y - \cos y\right], \tag{11}$$

where $\kappa = k r_{\rm s}$ and $y = \sqrt{3}\kappa/2$. For large k the profile is exponentially suppressed. Since the profile is not compensated, $\tau(k \ll 1) \simeq (2\pi^2/3)(1 - \kappa^2/3 + \cdots)\delta_0 r_{\rm s}^3$ does not go to zero as $k \to 0$. This is clearly only an approximation, since on sufficiently large scales the profile must go to zero due to mass conservation, but it does not seem to adversely affect our filter. We remind the reader that it is this Fourier space form that is input into the matched filter.

The above was all in real space. An analytic model for a void in redshift space could simply use the linear theory analysis of Kaiser (1987). A better alternative would be to make use of the Gaussian streaming model (Reid & White 2011). Hamaus et al. (2015) have shown that this model works well if linear theory expressions for the mean pairwise velocity and dispersion are computed from the assumed profile. We have taken a simpler approach, using the simulations to measure the anisotropy. Fig. 4 shows the Fourier transform of the same voids shown in Fig. 3, except now in redshift space. Material that is outflowing causes the void to appear deeper and wider in the line-of-sight direction (Kaiser 1987), enhancing the profile along $k_\parallel$. In principle, this makes the redshift-space profile less sensitive to loss of modes in the ‘wedge’ than would be anticipated from the real-space profile (though the filter will tend to downweight the line-of-sight modes more due to the enhanced cosmic clustering close to the line of sight).

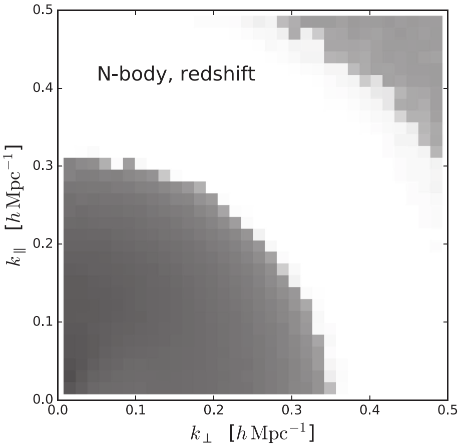

Figure 4. The (stacked) profile in Fourier space for voids with radii $10 < R_{\rm v} < 15\,h^{-1}$ Mpc from our *N*-body simulations in redshift space at $z = 1$.

However, for intermediate scales $k \sim 0.2h{\rm Mpc}^{-1}$, the void profiles are remarkably close to spherical, with only a very mild quadrupole. Given this small anisotropy, we shall continue to use a spherically symmetric void profile even in redshift space. This choice was motivated purely for the simplicity of the presentation and does not represent a limitation of the method, and we expect these choices to be revisited in future work.

4 RESULTS

4.1 Filter amplitude distributions

We now turn to the performance of the matched filter. Recall that we can evaluate the matched filter at an arbitrary point – ideally positions centred on voids would have significantly larger values of $\hat{A}$ than a randomly chosen point.[2] The left-hand panel of Fig. 5 plots the distribution of $\hat{A}$ in the ideal case of an effectively noiseless $\bar{n}P = 10$ survey. The distribution is close to Gaussian with a width of 0.86; this compares with the analytically predicted value (equation 7) of 0.90. The Gaussianity of this distribution is easily understood by observing that the matched filter simply smoothes the (configuration space) density field with a kernel that is $\mathcal{O}(10)$ Mpc wide; on these scales, the density field is very close to Gaussian. We do see evidence of non-Gaussianity from collapsed objects in a slight skew towards negative values of $\hat{A}$. Although the matched filter has the void radius as an input parameter, we find that the shapes of the distributions (after scaling out the variance) are very

[2] Since our input void profile has a negative central underdensity, we expect voids to have positive values of $\hat{A}$.

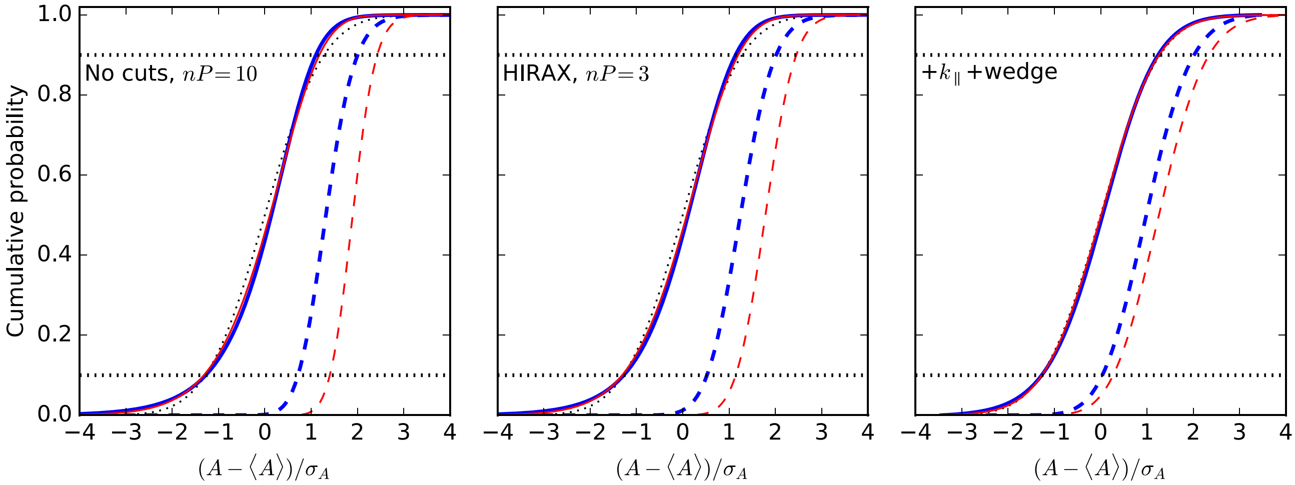

Figure 5. The (normalized) distribution of matched filter amplitude, A, at the locations of voids (dashed) compared to the full (volume weighted) distribution (solid) for voids of radius 10–15 h^{-1} Mpc (thick, blue) and 20–25 h^{-1} Mpc (thin, red). In each panel the dotted (black) line shows a unit-variance Gaussian for reference. The left panel shows void recovery for perfect sampling of the $k_{\perp}$ plane with minimal noise (we have quoted the noise as $\bar{n}P = 10$ at $k_{\rm fid} = 0.2\,h\,{\rm Mpc}^{-1}$, in analogy with galaxy surveys but recall $\bar{n}$ is in reality thermal noise as a function of $k_{\perp}$). The middle panel shows a baseline distribution for a HIRAX-like telescope with $\bar{n}P = 3$ and the right-hand panel shows the additional effects of removing modes with $k_{\parallel} < 0.05\,h\,{\rm Mpc}^{-1}$ and $\mu_k < 0.56$. The horizontal dotted lines in each panel mark the 10th and 90th percentiles for reference.

similar. We therefore simply standardize all our distributions by the appropriate variance.

We can now compare the above distribution with the matched filter evaluated at void centres. We consider two sets of voids: $10 < R_V < 15\,h^{-1}$ Mpc and $20 < R_V < 25\,h^{-1}$ Mpc and we set the filter radius to $12.5\,h^{-1}$ Mpc and $22.5\,h^{-1}$ Mpc, respectively. We find that the distribution of $\hat{A}$ evaluated at the void centres is clearly separated from the full distribution of the matched filter. Approximately 80 per cent of the smaller voids are detected at $>1\sigma$ from zero while ∼90 per cent of the larger voids are detected at $>1.5\sigma$. It is also worth noting that our reference distribution includes points that are in voids. Indeed ∼8 per cent of the simulation volume is contained in voids larger than $10\,h^{-1}$ Mpc, which would correspond to a threshold choice of $\sim 1\sigma$. Note that this is somewhat different from the Gaussian expectation of $\sim 1.5\sigma$; this difference can be traced to non-Gaussianity in the tails of the distribution of $\hat{A}$.

We now consider how survey non-idealities impact the efficiency of the matched filter. There are two aspects relevant to the 21 cm interferometer case. The first is that the instrument only samples particular k-modes and that this sampling is modulated by the number of baselines in the interferometer. The second is that, as discussed in Section 2.3, astrophysical foregrounds and instrumental imperfections can contaminate both low $k_{\parallel}$ modes and the so-called 'wedge', further restricting the accessible k-space. The impact of these is summarized in the middle and right-hand panels of Fig. 5. The relatively wide and dense coverage in k-space of our HIRAX-like survey implies that the filter's performance does not degrade significantly compared to the ideal case. Removing modes contaminated by foregrounds has a more significant effect. While we still see a separation between voids and randomly chosen points, only 50 per cent of the voids are now above the thresholds discussed above.

While the detailed performance of the void finder will depend on the details of the interferometer, the principal conclusion of the above discussion is that for the designs that are being considered voids are relatively easily detected in the absence of foregrounds but the loss of low $k_{\parallel}$ modes is a serious matter and some foreground mitigation strategy is necessary. Fig. 6 shows similar performance plots for an idealization of the CHIME experiment (see Appendix B for details). As with our HIRAX example, we find a clear separation between the distribution of voids and random points with similar recovered fractions of voids for the cases without any foregrounds, and a loss of separation when foregrounds become important.

As with all matched filter applications, there are a number of input choices. The choice of the void profile is the most notable example in this case. We experimented with different choices of void shapes and sizes and find that the results above are quite robust. A different complication arises from the fact that our void profiles are estimated from the dark matter. Appendix C explores the shapes of voids with a more realistic modelling of the 21 cm density field. We find that shapes of the voids here are very similar (and possibly more pronounced) to those in the dark matter. We therefore expect our results to be qualitatively unchanged with more realistic modelling of the 21 cm field.

Another choice in our matched filter is the power spectrum used in the noise covariance matrix to account for large-scale structure noise. While different choices here change the exact width of the distribution of $\hat{A}$, it does not change our basic result that voids are detected with very high significance in the absence of foregrounds.

4.2 An example application: a void catalog

As an example application, we discuss how to use such a matched filter to construct a void catalog. Our intention here is not to attempt to quantify (or optimize) the purity and completeness of such an algorithm, since this will be data and instrument specific and so much depends upon the manner in which foregrounds are subtracted. Instead, we outline the steps of a possible algorithm and perform some simple calculations with it, and defer detailed discussions to future work.

For this demonstration, we choose a single simulation box from our suite of 10 simulations. We run the matched filter on this box with the void radius R_V varying from $33.3\,h^{-1}$ Mpc to $20\,h^{-1}$ Mpc in 10 per cent steps. We keep a list of all points where the matched filter amplitude, A, exceeds $2\,\sigma$. Starting from the largest void(s)

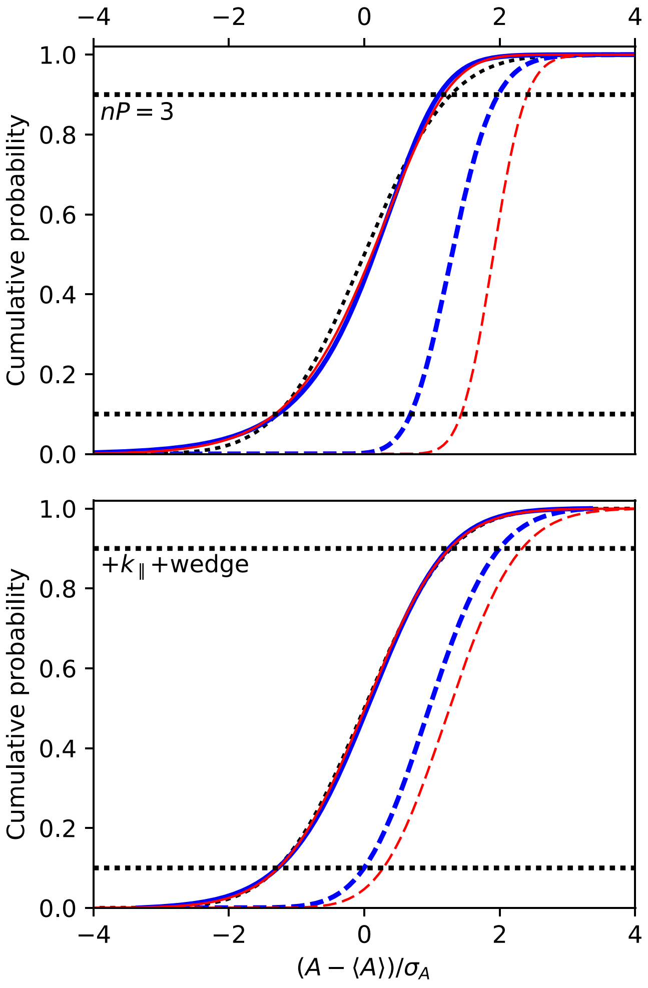

Figure 6. As in Fig. 5 but for an idealization of the CHIME telescope. The top panel shows the impact of the CHIME baseline distribution with a noise level appropriate for a BAO detection, while the lower panel shows the additional impact of removing modes with $k_{\parallel} < 0.05\,h\,\mathrm{Mpc}^{-1}$ and $\mu_k < 0.56$.

and working down in radius, we eliminate any voids that overlap. If two overlapping voids have the same radius, the one with the smallest A is removed. The result of this procedure is our 'void catalog'.

In a single 1380 h^{-1} Mpc box the largest 1000 voids have radius above about 20 h^{-1} Mpc. With full $u - v$ coverage and low noise we find that all but 1 of the 32 largest 'true' voids contain a match in our catalog within 0.75 R_V and these matches are all in the upper 5$^{\mathrm{th}}$ percentile of the A distribution. Just under half of them (14 of the 32 voids) show significant ($>R_V/3$) mis-centring, i.e. the detected void centre is $>R_V/3$ away from the centre of the closest true void. For the $u - v$ coverage of our HIRAX-like experiment, and $\bar{n}P = 3$, four of the 32 largest 'true' voids do not have a match within 0.75 R_V and again all are highly significant. The situation changes dramatically as we include a $k_{\parallel,\,\mathrm{min}}$ and μ cut. For $k_{\parallel,\,\mathrm{min}} = 0.05\,h\,\mathrm{Mpc}^{-1}$ and $\mu > 0.56$, we find only 5 of the top 32 voids in our catalog, though these voids are in the extreme tails of the A distribution. Most of this effect is driven by the $k_{\parallel}$ cut. If we relax the cut to $0.02\,h\,\mathrm{Mpc}^{-1}$, then we recover 10 of the 32 largest voids and for a cut of $0.01\,h\,\mathrm{Mpc}^{-1}$ we recover 20 of them.

We can recast the results of this and the previous section into the more traditional forms of the completeness and purity of the sample.

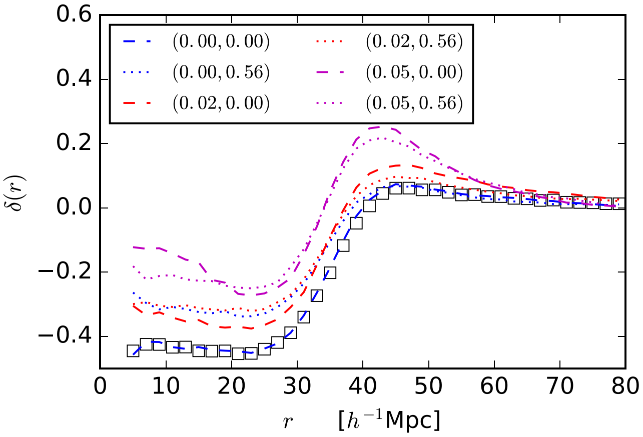

Figure 7. Redshift-space density profile stacked around the 1000 voids in our 'candidate catalog', as described in the text. Black squares show the stacked profile assuming perfect $u - v$ coverage and no noise. Clearly, in the absence of foregrounds, our candidates correspond to large, coherent underdensities. Compared to Fig. 3, the shallower profile at small radius is due to the miscentering described in the text. The lines are all for the HIRAX $u - v$ coverage with $\bar{n}P = 3$, and show the impact of losing modes at low $k_{\parallel}$ and in the wedge. The legend gives the cuts as $k_{\parallel,\,\mathrm{min}}$ and μ pairs.

In the absence of foregrounds, our detected void catalog is both pure (only ∼10 per cent of detected voids do not correspond to true voids) and complete (>90 per cent of true voids are detected at better than 1.5σ) for large (∼20 h^{-1} Mpc) voids. However, both of these numbers are sensitive to foregrounds. For our most conservative case of foregrounds contaminating all modes with $k_{\parallel} < 0.05\,h\mathrm{Mpc}^{-1}$ and $\mu < 0.56$, the majority of the most prominent detections do not correspond to true underlying voids and only ∼50 per cent of true voids are detected at high significance.

It is possible that some of the low $k_{\parallel}$ information lost to the interferometer by foregrounds could be replaced by another experiment. As an example, modern photometric surveys can achieve high photometric redshift precision for certain types of galaxies, and thus can map the low $k_{\parallel}$ modes of the 3D density field. In fact, such surveys have been used to search for voids (Sánchez et al. 2017). Including the photometric survey in our matched filter presents no problem in principle – one simply augments the data vector and includes a model for the void in configuration space – but could be difficult in practice. Assuming the combination recovers all of the $k_{\parallel}$ range, we recover our no-foreground forecasts. If there is a gap in coverage, the results are adversely affected. To take a pessimistic example: if we lose modes $0.02 < k_{\parallel} < 0.05\,h\,\mathrm{Mpc}^{-1}$, we are able to recover 12 of our top 32 voids. For $0.03 < k_{\parallel} < 0.05\,h\,\mathrm{Mpc}^{-1}$ it is half of our top 32 voids.

These lost $k_{\parallel}$ modes potentially could be reconstructed from higher-point information in the 21cm field itself (Zhu et al. 2016). There is considerable interest in developing these reconstruction schemes for 21 cm surveys to enable cross-correlations with photometric surveys or CMB lensing maps. Initial results (Zhu et al. 2016) suggest that modes $k_{\parallel} < 0.01\,h\mathrm{Mpc}^{-1}$ and $k_{\perp} < 0.05\,h\mathrm{Mpc}^{-1}$ could be recovered. As with the example above, the efficiency of the void finder will depend on the details of the performance of these reconstructions.

We can visualize this information in another way. Fig. 7 shows the stacked matter profile around our top 1000 void candidates for various choices of $k_{\parallel,\,\mathrm{min}}$ and μ_{min}. With full $u - v$ coverage, there is a clear, coherent underdensity at the locations of the void candidates. The shallower inner profile in Fig. 7, when compared to

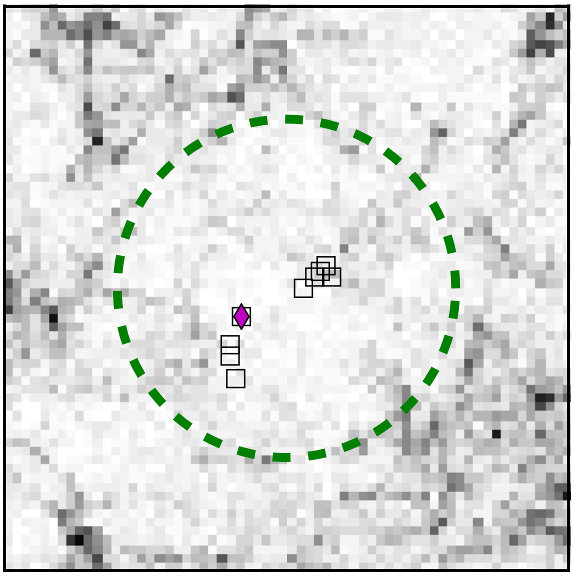

Figure 8. An example of a detected void in our catalog, for the case of no foregrounds. The image shows the matter distribution centred around a $R_V \simeq 30\,h^{-1}$ Mpc void (bounded by the dashed line). The slice is $100\,h^{-1}$ Mpc wide and $20\,h^{-1}$ Mpc thick. The squares show significant matched filter detections, with the filled, magenta, diamond being the most significant detection (as determined by our pruning algorithm). We see that, while the void centre is detected by the matched filter, it happens not to be the most significant detection, resulting in a mis-centred void.

Fig. 3, arises due to mis-centring. While some voids were well centred, a significant fraction had offsets. After visually inspecting these voids, we find that, in most cases, the true void centre was a significant detection in the matched filter, but happened not to be the most significant detection and was removed by our relatively simple pruning algorithm. Fig. 8 shows an example of such a case. Both the fraction of found voids and the degree of mis-centring get worse when modes are lost to foregrounds, as the other curves in Fig. 7 show. For our pessimistic scenario of $k_\parallel > 0.05\,h\,\mathrm{Mpc}^{-1}$ and $\mu > 0.56$ there is barely any underdensity detected at all. This suggests that doing science with voids selected via 21 cm experiments will be difficult unless the foregrounds can be brought under control. We however note that the algorithm used to construct our void catalog is relatively simple; e.g. a more robust algorithm might use multi-scale information to get more robust measurements and there is significant potential for complementarity between optical imaging surveys and 21 cm measurements.

5 CONCLUSIONS

Recent advances in technology have made it feasible to study the 21 cm emission from objects at cosmological distances. A new generation of telescopes is being designed and built, which aims to survey enormous volumes of the Universe with modest resolution at redshifts $z \simeq 1$–2. A primary focus of these facilities is the measurement of the power spectrum of large-scale structure, as traced by neutral hydrogen, which will hopefully improve constraints on our cosmological model. While these instruments do not have sufficient angular resolution to resolve the emission from individual objects, we point out that they should be able to make catalogs of the largest members of the cosmic web – protoclusters and voids – if they are able to control foregrounds sufficiently.

We have considered instruments that measure the sky interferometrically, which means they naturally operate in Fourier space. The finite sampling of the Fourier plane and the loss of sensitivity in some modes due to foregrounds make it difficult to generate a real-space, 3D map from the data and hence to search for exotica whose properties are not known in advance. However, our understanding of the cosmic web allows us to specify in advance what sort of objects we are interested in finding and searches for objects of known shape do not need to go through the map-making step: a matched filter provides a natural method for finding such objects. The matched filter formalism also allows us to mix multiple data sets, each of which is provided in its own domain.

The cosmic web contains voids on a variety of scales, and voids that touch or merge. We have only studied the simplest matched filter. The algorithm can be modified to iteratively add voids to an existing catalog, always adding the void, which leads to the largest increase in the likelihood given the already-found voids (see e.g. Kochanek et al. 2003; Dong et al. 2008). This involves a scan over void (or protocluster) sizes, and increases the complexity of the algorithm. A multi-probe approach could use deep, optical imaging data in conjunction with 21 cm data in much the same way as multi-frequency information is sometimes used for cluster finding (e.g. Melin, Bartlett & Delabrouille 2006; Rykoff et al. 2014). As our main aim was to assess the feasibility of void detection with 21 cm surveys, we defer further consideration of such a process to future work.

Throughout we have focused our discussion on voids as exemplars of large structures in the cosmic web. Of course, the matched filter algorithm is more general and the huge volume and sensitivity of upcoming experiments can be used to search for a number of exotic objects. At the other end of the density distribution from voids are the large, coherent overdensities associated with protoclusters.

Despite keen interest in the community in how clusters form and evolve, and years of observational and numerical efforts, the study of early cluster formation (at high z) remains observationally limited. Protoclusters are rare, present only modest overdensities and lack many of the features used to discover clusters (e.g. a hot ICM or a red sequence). Observations of protoclusters at high z would provide important clues into cluster assembly and the processes of galaxy formation (Overzier 2016). Given the diversity of protoclusters, having large samples with well-understood selection is important. Like voids, protoclusters form large coherent structures amenable to discovery in upcoming 21 cm experiments. Assuming a mean interior density of 200 times the background, the linear size of the mean-density region from which material accretes into a present-day cluster is several (comoving) Mpc. The progenitors of large clusters should thus be identifiable in relatively low-resolution maps that can cover large volumes (see e.g. Overzier 2016, and Fig. 2).

Slices through the density field in one of our simulations are shown in Fig. 2, where the large extended mass profile of the protoclusters is evident. In fact, the most massive clusters in the mature Universe form not from the most overdense regions at high z but from large, possibly only moderately overdense regions such as shown in Fig. 2 (Overzier 2016). While we do not show it here, the typical protocluster covers a larger volume at $z \simeq 2$, rendering it potentially easier to see while still being well within the redshift reach of HIRAX or CHIME.

The abundance of such protoclusters is identical to the abundance of the clusters at $z = 0$: for a mass threshold of $3 \times 10^{14}\,h^{-1}\,\mathrm{M}_\odot$, it is $4 \times 10^{-6}\,h^3\,\mathrm{Mpc}^{-3}$. This emphasizes the need for a survey to cover a large volume in order to properly sample the heterogeneous population of protoclusters. As an example, if it covered 15 000 deg^2

between $z = 1$ and 2 HIRAX would survey $50(h^{-1}\,\mathrm{Gpc})^3$ encompassing ∼200 000 protoclusters.[3] CHIME is anticipated to cover a similar volume in the Northern hemisphere. In some models the star formation associated with haloes in protoclusters makes up a significant fraction of the ionizing photon budget for re-ionization (Chiang et al. 2017) at $z \simeq 6$–7. If foregrounds could be controlled, using interferometers designed for studying re-ionization to search for protoclusters could provide an interesting synergy.

ACKNOWLEDGEMENTS

We thank Emanuele Castorina and Richard Shaw for useful discussions and comments on an early version of this paper. N.P. thanks Laura Newburgh for useful conversations. We thank the referee for detailed comments on the paper. M.W. is supported by DOE. N.P. is supported in part by DOE DE-SC0008080. This work made extensive use of the NASA Astrophysics Data System and of the astro-ph preprint archive at arXiv.org. The analysis made use of the computing resources of the National Energy Research Scientific Computing Center.

REFERENCES

Ali S. S., Bharadwaj S., 2014, J. Astrophys. Astron. 35, 157
Alonso D., Ferreira P. G., Santos M. G., 2014, MNRAS, 444, 3183
Alonso D., Ferreira P. G., Jarvis M. J., Moodley K., 2017, preprint (arXiv:1704.01941)
Ansari R. et al., 2012, A&A, 540, A129
Bagla J. S., White M., 2003, in Ikeuchi S., Hearnshaw J., Hanawa T., eds, ASP Conf. Ser. Vol. 289, The Proceedings of the IAU 8th Asian-Pacific Regional Meeting, Vol. 1, pp. 251–254 (astro-ph/0212228)
Bagla J. S., Khandai N., Datta K. K., 2010, MNRAS, 407, 567
Banerjee A., Dalal N., 2016, J. Cosmology Astropart. Phys., 11, 015
Barnes L. A., Haehnelt M. G., 2010, MNRAS, 403, 870
Barnes L. A., Haehnelt M. G., 2014, MNRAS, 440, 2313
Battye R. A., Davies R. D., Weller J., 2004, MNRAS, 355, 1339
Bond J. R., Efstathiou G., 1987, MNRAS, 226, 655
Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21
Bunn E. F., White M., 2007, ApJ, 655, 21
Cai Y.-C., Padilla N., Li B., 2015, MNRAS, 451, 1036
Cai Y.-C., Taylor A., Peacock J. A., Padilla N., 2016, MNRAS, 462, 2465
Castorina E., Villaescusa-Navarro F., 2016, preprint (arXiv:1609.05157)
Chang T.-C., Pen U.-L., Bandura K., Peterson J. B., 2010, Nature, 466, 463
Chen X., 2012, in Int. J. Mod. Phys. Conf. Ser. pp. 256–263, preprint (arXiv:1212.6278)
Chiang Y.-K., Overzier R. A., Gebhardt K., Henriques B., 2017, preprint (arXiv:1705.01634)
Clampitt J., Cai Y.-C., Li B., 2013, MNRAS, 431, 749
Cohn J. D., White M., Chang T.-C., Holder G., Padmanabhan N., Doré O., 2016, MNRAS, 457, 2068
Colberg J. M. et al., 2008, MNRAS, 387, 933
Crighton N. H. M. et al., 2015, MNRAS, 452, 217
Datta K. K., Choudhury T. R., Bharadwaj S., 2007, MNRAS, 378, 119
Davé R., Katz N., Oppenheimer B. D., Kollmeier J. A., Weinberg D. H., 2013, MNRAS, 434, 2645
Dong F., Pierpaoli E., Gunn J. E., Wechsler R. H., 2008, ApJ, 676, 868
Falck B., Koyama K., Zhao G., Cautun M., 2017, preprint (arXiv:1704.08942)
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gong Y., Chen X., Silva M., Cooray A., Santos M. G., 2011, ApJ, 740, L20
Hamaus N., Sutter P. M., Wandelt B. D., 2014, Phys. Rev. Lett., 112, 251302
Hamaus N., Sutter P. M., Lavaux G., Wandelt B. D., 2015, JCAP, 11, 036
Hamaus N., Cousinou M.-C., Pisani A., Aubert M., Escoffier S., Weller J., 2017, JCAP, 07, 014
Hawken A. J. et al., 2016, preprint (arXiv:1611.07046)
Heitmann K. et al., 2008, Computational Science and Discovery, 1, 015003
Kaiser N., 1987, MNRAS, 227, 1
Kochanek C. S., White M., Huchra J., Macri L., Jarrett T. H., Schneider S. E., Mader J., 2003, ApJ, 585, 161
Lavaux G., Wandelt B. D., 2012, ApJ, 754, 109
Lee J., Park D., 2009, ApJ, 696, L10
Liu A., Zhang Y., Parsons A. R., 2016, ApJ, 833, 242
Marín F. A., Gnedin N. Y., Seo H.-J., Vallinotto A., 2010, ApJ, 718, 972
McQuinn M., Zahn O., Zaldarriaga M., Hernquist L., Furlanetto S. R., 2006, ApJ, 653, 815
Melin J.-B., Bartlett J. G., Delabrouille J., 2006, A&A, 459, 341
Newburgh L. B. et al., 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser., p. 99065X, preprint (arXiv:1607.02059)
Overzier R. A., 2016, A&A Rev., 24, 14
Padmanabhan H., Refregier A., 2017, MNRAS, 464, 4008
Padmanabhan H., Choudhury T. R., Refregier A., 2015, MNRAS, 447, 3745
Pober J. C., 2015, MNRAS, 447, 1705
Pober J. C. et al., 2013, AJ, 145, 65
Rao S. M., Turnshek D. A., Nestor D. B., 2006, ApJ, 636, 610
Reid B. A., White M., 2011, MNRAS, 417, 1913
Reid B. A., Seo H.-J., Leauthaud A., Tinker J. L., White M., 2014, MNRAS, 444, 476
Rood H. J., 1988, ARA&A, 26, 245
Rykoff E. S. et al., 2014, ApJ, 785, 104
Sánchez C. et al., 2017, MNRAS, 465, 746
Seehars S., Paranjape A., Witzemann A., Refregier A., Amara A., Akeret J., 2016, J. Cosmology Astropart. Phys., 3, 001
Seo H.-J., Hirata C. M., 2016, MNRAS, 456, 3142
Seo H.-J., Dodelson S., Marriner J., Mcginnis D., Stebbins A., Stoughton C., Vallinotto A., 2010, ApJ, 721, 164
Shaw J. R., Sigurdson K., Pen U.-L., Stebbins A., Sitwell M., 2014, ApJ, 781, 57
Shaw J. R., Sigurdson K., Sitwell M., Stebbins A., Pen U.-L., 2015, Phys. Rev. D, 91, 083514
Solanes J. M., Manrique A., García-Gómez C., González-Casado G., Giovanelli R., Haynes M. P., 2001, ApJ, 548, 97
Stark C. W., White M., Lee K.-G., Hennawi J. F., 2015a, MNRAS, 453, 311
Stark C. W., Font-Ribera A., White M., Lee K.-G., 2015b, MNRAS, 453, 4311
Switzer E. R. et al., 2013, MNRAS, 434, L46
Thompson A. R., Moran J. M., Swenson G. W., Jr, 2017, Interferometry and Synthesis in Radio Astronomy, 3rd Edition. Springer-Verlag, Berlin
Tinker J. L., Conroy C., 2009, ApJ, 691, 633
van de Weygaert R., Platen E., 2011, Int. J. Mod. Phys. Conf. Ser., Vol. 1, p. 41
Vanderlinde K., Chime Collaboration 2014, in Exascale Radio Astronomy BAAS, 46, 101.02
Villaescusa-Navarro F. et al., 2016, MNRAS, 456, 3553
White M., 2002, ApJS, 143, 241
White M., Carlstrom J. E., Dragovan M., Holzapfel W. L., 1999, ApJ, 514, 12
White M., Reid B., Chuang C.-H., Tinker J. L., McBride C. K., Prada F., Samushia L., 2015, MNRAS, 447, 234
Wolz L., Blake C., Wyithe J. S. B., 2017, preprint (arXiv:1703.08268)
Zaldarriaga M., Furlanetto S. R., Hernquist L., 2004, ApJ, 608, 622
Zhu H.-M., Pen U.-L., Yu Y., Chen X., 2016, preprint (arXiv:1610.07062)

APPENDIX A: SIGNAL-TO-NOISE RATIO

We present a self-contained derivation of the instrument noise power spectrum, converted to cosmological units. Our derivation is similar to that in Alonso et al. (2017), but related expressions have also appeared in White et al. (1999), Zaldarriaga, Furlanetto & Hernquist

[3] Almost by definition, the number density of protoclusters is redshift independent.

(2004), McQuinn et al. (2006), Seo et al. (2010), Bull et al. (2015), Seo & Hirata (2016) and Wolz, Blake & Wyithe (2017).

The brightness temperature, T_b, is defined in terms of the intensity at frequency ν as $I_\nu = 2k_B T_b(\nu/c)^2 = 2k_B T_b/\lambda^2$. We begin by noting that if we normalize our visibilities in terms of temperature (rather than intensity) the power can be written in terms of the brightness temperature power spectrum and window function as

$$\left\langle |V_i|^2 \right\rangle = \int \mathrm{d}^2 u\ P_T(\boldsymbol{u}) W(\boldsymbol{u}) \approx P_T(\boldsymbol{u}) \int \mathrm{d}^2 u\ W(\boldsymbol{u}) \tag{A1}$$

with the last approximation holding if the window function is compact and P_T is smooth. Conventionally the beam is normalized to unity at peak, so its area in the $u - v$ plane integrates to unity and thus the window function integrates to the inverse area

$$\int \mathrm{d}^2 u\ W(\boldsymbol{u}) \sim \frac{1}{\mathrm{d}^2 u}, \tag{A2}$$

which gives

$$\left\langle |V_i|^2 \right\rangle \approx \frac{P_T(\boldsymbol{u})}{\mathrm{d}^2 u}. \tag{A3}$$

It may be helpful to derive equation (A3) differently. If we treat the visibility as measuring a single Fourier mode of the 2D brightness temperature field, we can relate this to the 2D power spectrum of this field

$$\langle V(\boldsymbol{\ell}) V^\star(\boldsymbol{\ell}') \rangle = (2\pi)^2 \delta^D(\boldsymbol{\ell} - \boldsymbol{\ell}') P_T(\boldsymbol{\ell} = 2\pi \boldsymbol{u}) \approx \delta^K_{\boldsymbol{\ell},\boldsymbol{\ell}'} \frac{P_T(\boldsymbol{u})}{\mathrm{d}^2 u}, \tag{A4}$$

where $\delta^{D,K}$ are the Dirac and Kronecker δ functions. The above equation explicitly relates $\boldsymbol{u}$ to the 2D wavevector $\boldsymbol{\ell}$, and the last approximation comes from assuming a discretized set of wavevectors.[4]

In the same units the visibility noise is diagonal (Thompson et al. 2017)

$$\left\langle |N_i|^2 \right\rangle = \left[\frac{2k_B}{\lambda^2} \right]^{-2} \left[\frac{2k_B T_{\rm sys}}{A_{\rm e}} \right]^2 \frac{1}{\Delta\nu\, t_{\rm p}} = \left[\frac{\lambda^2 T_{\rm sys}}{A_{\rm e}} \right]^2 \frac{1}{\Delta\nu\, t_{\rm p}} \tag{A5}$$

per baseline. Here $T_{\rm sys}$ is the system temperature, $A_{\rm e}$ the effective area of the telescope (equal to the aperture efficiency times the physical area), $\Delta\nu$ is the bandwidth, $t_{\rm p}$ is the observing time per pointing and we have assumed a single polarization.

The above is all that is needed to implement a matched filter on the data, where we can work at the level of the visibilities. It is however useful to translate this into the cosmological units used in the paper. We start by defining the number of baselines per unit area in the $u - v$ plane, $n(\boldsymbol{u})$, normalized such that

$$\int \mathrm{d}^2 u\, n(\boldsymbol{u}) = N_{\rm pairs} = \frac{N_a(N_a - 1)}{2}, \tag{A6}$$

where N_a is the number of antennas and $N_{\rm pairs}$ is the number of pairs (i.e. instantaneous baselines). Averaging over the number of baselines, the noise becomes $\langle N_i^2 \rangle / (n(\boldsymbol{u}) \mathrm{d}^2 u)$. Using equation (A3), we obtain

$$P_N(\boldsymbol{u}) = \left[\frac{\lambda^2 T_{\rm sys}}{A_{\rm e}} \right]^2 \frac{1}{n(\boldsymbol{u})} \frac{1}{\Delta\nu\, t_{\rm p}} = \left[\frac{\lambda^2 T_{\rm sys}^2}{A_{\rm e}} \right] \frac{1}{n(\boldsymbol{u})} \frac{4\pi f_{\rm sky}}{\Delta\nu\, t_{\rm obs}}. \tag{A7}$$

The last equality follows from $N_p \Omega_p = 4\pi f_{\rm sky}$, where $N_{\rm p} = t_{\rm obs}/t_{\rm p}$ is the number of pointings and $t_{\rm obs}$ is the total observing time. The area covered by each pointing $\Omega_{\rm p}$ is approximately given by $\lambda^2/A_{\rm e}$. Physically, the above equations assume that each pointing yield a disjoint set of modes.

To convert this visibility noise into a cosmological power spectrum, we divide by the mean cosmological brightness temperature (Seo et al. 2010)

$$\bar{T} = 188 \frac{x_{\rm HI}(z) \Omega_{H,0} h (1+z)^2}{H(z)/H_0}\ \mathrm{mK}, \tag{A8}$$

with $x_{\rm HI}$ the neutral hydrogen fraction, and convert from $\boldsymbol{u}$ to $\boldsymbol{k}_\perp$ in comoving coordinates and similarly for frequency to $k_\parallel$ to obtain

$$P_N = \left(\frac{T_{\rm sys}}{\bar{T}} \right)^2 \left(\frac{\lambda^2}{A_{\rm e}} \right) \frac{4\pi f_{\rm sky}}{t_{\rm obs} n(\boldsymbol{u})} \frac{\mathrm{d}^2 V}{\mathrm{d}\Omega\, \mathrm{d}\nu}, \tag{A9}$$

where in a spatially flat model

$$\frac{\mathrm{d}^2 V}{\mathrm{d}\Omega\, \mathrm{d}\nu} = \chi^2 \frac{\mathrm{d}\chi}{\mathrm{d}z} \frac{\mathrm{d}z}{\mathrm{d}\nu} = \chi^2 \frac{c\,(1+z)^2}{H(z)\, \nu_0} \tag{A10}$$

with $\nu_0 = 1420$ MHz. Unfortunately the value of $\Omega_{H,0} h$ is quite uncertain and it enters quadratically in the noise power spectrum. Rao, Turnshek & Nestor (2006) measure $10^3 \Omega_{H,0} \simeq 0.9 \pm 0.3$ at $z \approx 1$ through the abundance of damped Lyman-α systems (see also the compilations of data in Crighton et al. 2015; Padmanabhan, Choudhury & Refregier 2015). The measurement of $\Omega_{H,0} b$ through 21 cm auto-correlations by Switzer et al. (2013) has a similar value and fractional error. We will consider the range (0.6–1.2)$\times 10^{-3}$ or $\Omega_{H,0}\, h$=(4–9)$\times 10^{-4}$. For $\Omega_{H,0}\, h = 4 \times 10^{-4}$ and the HIRAX-like interferometer described in the text operating for 3 yr we obtain $P_N \approx 600\, h^{-3}$ Mpc^3 at $z = 1$ and $k_\perp = 0.2\, h^{-1}$ Mpc. Comparing to the linear matter spectrum, and assuming $b = 1$, we have $(P_L/P_N)(k_\perp) \approx 1$. If $\Omega_{H,0}\, h = 9 \times 10^{-4}$, we obtain $P_N \approx 150\, h^{-3}$ Mpc^3 and have $(P_L/P_N)(k_\perp) \approx 4$.

While this is similar in spirit to nP in galaxy surveys, it is worth emphasizing that this quantity is intrinsically 2D, while nP is spherically symmetric. In particular, at fixed k, the average value of $k_\perp$ is $(\pi/4)k$.

APPENDIX B: TRANSIT TELESCOPES AND THE *m*-MODE FORMALISM

The interferometers for 21 cm intensity mapping experiments are designed to be transit telescopes, using the Earth's rotation to map large areas of the sky. This mapping process simultaneously performs two operations that are traditionally treated separately – filling in the $u - v$ plane,[5] and improving the resolution in the $u - v$ plane[6] by 'mosaicking'. Furthermore, some upcoming experiments, notably CHIME[7] and Tianlai,[8] use a close-packed array of cylinders rather than traditional dishes. In the CHIME configuration, four cylinders (each 20 m in diameter and ∼100 m long, oriented north–south) are placed adjacent in the east–west direction (Vanderlinde &

[4] For instance, this is exactly what happens on an FFT grid in a simulation.

[5] As we discuss later in this section, a more appropriate basis for discussing these telescopes are spherical harmonics. We use the $u - v$ plane here to mean an appropriate 'Fourier' transform of the sky.

[6] Recall that a single visibility measurement is smeared in the '$u - v$' plane by the Fourier transform of the primary beam and 'mosaicking' combines observations of different areas of the sky to make this window function more compact.

[7] http://chime.phas.ubc.ca/

[8] http://tianlai.bao.ac.cn

Chime Collaboration 2014). The primary beam from such a configuration is highly extended in the north–south direction, while being focused by the cylinders in the east–west direction.

Both of these features naturally cover large angles on the sky. The natural basis for describing these telescopes is not the usual Fourier basis, but rather spherical harmonics. However, most astrophysical signals (including the voids discussed here) cover small areas in the sky and are easily described in a flat-sky limit. The goal of this Appendix is to make the connection between the wide-angle and flat sky formalism explicit.

We start with a review of the m-mode formalism, following Shaw et al. (2014, 2015), who state the fundamental visibility measurements in a spherical harmonic basis. We then take the flat-sky limit of this result and show that we recover the traditional $u-v$ plane interpretation. Making this connection also allows us to explicitly see how the Earth's rotation fills in the $u-v$ plane. We then develop the matched filter formalism in this basis. We conclude with a worked example of the m-mode formalism, to help build intuition.

B1 Review of the *m*-mode formalism

Following Shaw et al. (2014, 2015), if the beam transfer function pointed at azimuth ϕ is

$$B_{ij}(\hat{\boldsymbol{n}};\phi) \propto A^2(\hat{\boldsymbol{n}};\phi)\,\exp\left[2\pi i\,\hat{\boldsymbol{n}}\cdot\boldsymbol{u}_{ij}(\phi)\right] \tag{B1}$$

then

$$V_{ij}(\phi) = \int \mathrm{d}\hat{\boldsymbol{n}}\;T(\hat{\boldsymbol{n}})B_{ij}(\hat{\boldsymbol{n}};\phi) \tag{B2}$$

(plus noise, of course). We remind the reader to distinguish between the pointing centre of the beam (the azimuth of which is ϕ) and the coordinate that integrates over the beam ($\hat{\boldsymbol{n}}$). Expanding T and $B_{ij}(\phi)$ into spherical harmonics

$$T(\hat{\boldsymbol{n}}) = \sum_{\ell m} a_{\ell m}\,Y_{\ell m}(\hat{\boldsymbol{n}}) \tag{B3}$$

$$B_{ij}(\hat{\boldsymbol{n}};\phi) = \sum_{\ell m} B^{ij}_{\ell m}\,Y^{\star}_{\ell m}(\hat{\boldsymbol{n}}) \tag{B4}$$

we obtain

$$V_{ij}(\phi) = \sum_{\ell m} B^{\ell m}_{ij}(\phi) a_{lm}\,. \tag{B5}$$

The rotation of the Earth in ϕ causes the beam to transform as $B^{\ell m}(\phi) = B^{\ell m}(0)\mathrm{e}^{im\phi}$. Defining

$$V^m_{ij} = \int \frac{d\phi}{2\pi}\mathrm{e}^{-im\phi}V_{ij}(\phi), \tag{B6}$$

we obtain

$$V^m_{ij} = \sum_{\ell} B^{\ell m}_{ij} a_{\ell m}, \tag{B7}$$

where $B^{\ell m}_{ij}$ without an explicit argument is understood to be at $\phi = 0$ (the phase factor cancels out its conjugate in the definition of V^m_{ij}.) These V^m_{ij} (or their Fourier conjugate $V_{ij}(\phi)$ are the fundamental observables of the telescope.

B2 The Flat-Sky Approximation

It is illuminating to show that the above expression recovers the usual flat-sky Fourier representation for small areas of the sky. We will use $\boldsymbol{\ell}$ to represent the 2D Fourier wavevector, with magnitude ℓ and polar angle φ_ℓ (not to be confused with the pointing centre ϕ). The correspondence between $a_{\ell m}$ and $a(\boldsymbol{\ell})$ is (White et al. 1999; Datta, Choudhury & Bharadwaj 2007)

$$a(\boldsymbol{\ell}) = \sqrt{\frac{4\pi}{2\ell+1}}\sum_m i^{-m} a_{\ell m}\mathrm{e}^{im\varphi_\ell} \tag{B8}$$

and

$$a_{\ell m} = \sqrt{\frac{2\ell+1}{4\pi}}\,i^m \int \frac{d\varphi_\ell}{2\pi}\,a(\boldsymbol{\ell})\,\mathrm{e}^{-im\varphi_\ell}\,. \tag{B9}$$

with a similar expansion for $B^{\ell m}_{ij}$. Substituting into the visibility equation, $V_{ij}(\phi) = \sum_{\ell m} B^{\ell m}_{ij}(\phi)a_{\ell m}$, we obtain, for large ℓ,

$$V_{ij}(\phi) \approx \frac{1}{(2\pi)^3}\sum_{\ell m}\int \mathrm{d}\varphi_\ell \mathrm{d}\varphi_{\ell'}\;\ell a(\boldsymbol{\ell})B(\boldsymbol{\ell}',\phi)\mathrm{e}^{im(\varphi_\ell-\varphi_{\ell'})}\,, \tag{B10}$$

where $\boldsymbol{\ell}$ and $\boldsymbol{\ell}'$ have the same magnitude. Doing the sum over m yields a δ-function that collapses one of the azimuthal integrals to yield

$$V_{ij}(\phi) \approx \int \frac{\ell\,\mathrm{d}\ell\,\mathrm{d}\varphi_\ell}{(2\pi)^2}\;a(\boldsymbol{\ell})B(\boldsymbol{\ell},\phi), \tag{B11}$$

where we have approximated the sum over ℓ by an integral. The above shows that the visibilities approximately measure a mode $\boldsymbol{\ell}$, smeared by the Fourier transform of the beam function.

We can use the above results to understand how the rotation of the Earth fills in the $u-v$ plane. In the flat-sky limit, the Fourier transform of the beam is $B(\boldsymbol{\ell}) \sim \sum_m i^{-m}B_{\ell m}\exp[im\varphi_\ell]$. Rotating about the z-axis by α scales the $B_{\ell m}$ by $\exp[im\alpha]$, which is clearly equivalent to rotating $\boldsymbol{\ell}$ by α. The $u-v$ coverage of the telescope traces out circles in the $u-v$ plane as the Earth rotates. We note that this is different from the usual result for interferometers, and reflects the transit nature of these telescopes.

B3 Matched filters

In order to define the matched filter, we need to express the signal in terms of the observable quantities, in this case the visibilities. Since all of the objects of interest in this study are $\mathcal{O}(10\,\mathrm{Mpc})$ in size, at a distance of >1 Gpc, they subtend small angles on the sky, allowing us to express the signal using the same flat-sky Fourier representation used in the main paper.

To begin, consider a single frequency, corresponding to a fixed redshift or (redshift-space) distance. Suppose our template, τ, is centred at $\theta = 0$, is ϕ-independent and non-zero only when $\theta \ll 1$. We expand

$$\tau_{\ell m}(\hat{z}) = \int \mathrm{d}\hat{\boldsymbol{n}}\;Y^{\star}_{\ell m}(\hat{\boldsymbol{n}})\,\tau(\theta) \tag{B12}$$

$$= 2\pi\,\delta^K_{m0}\sqrt{\frac{2\ell+1}{4\pi}}\int \mathrm{d}(\cos\theta)P_\ell(\cos\theta)\,\tau(\theta) \tag{B13}$$

$$\simeq \delta^K_{m0}\sqrt{\frac{2\ell+1}{4\pi}}\left[2\pi\int \tilde{\omega}\,\mathrm{d}\tilde{\omega}\;J_0(\ell\tilde{\omega})\,\tau(\tilde{\omega})\right], \tag{B14}$$

where in the last line we have defined $\tilde{\omega} = 2\sin(\theta/2) \simeq \theta$ and used $P_\ell(\cos\theta) \approx J_0(\ell\theta)$ for $\theta \ll 1$. The $\sqrt{(2\ell+1)/4\pi}$ is just $Y_{\ell 0}(\hat{z})$. If we extend the upper limit of $\tilde{\omega}$-integration to infinity, we recognize in the brackets on the last line the Hankel transform of τ or the 2D Fourier transform of τ with spherical symmetry (e.g. Bond & Efstathiou 1987).

Now we can rotate the template from the north pole ($\hat{z}$) to an arbitrary $\hat{\boldsymbol{n}}$ using Wigner functions, $\mathcal{D}^{\ell}_{m'm}$. However, in our case $\tau_{\ell m} \propto \delta^{K}_{m0}$ and $Y_{\ell m} \propto \mathcal{D}^{\ell}_{0m}$ so that the spherical harmonic coefficients for a template centred on $\hat{\boldsymbol{n}}$ are

$$\tau_{\ell m}(\hat{\boldsymbol{n}}) = \sqrt{\frac{4\pi}{2\ell+1}}\, Y^{\star}_{\ell m}(\hat{\boldsymbol{n}})\, \tau_{\ell 0}(\hat{z}) \tag{B15}$$

$$= \left[2\pi \int \tilde{\omega}\, \mathrm{d}\tilde{\omega}\, J_0(\ell\tilde{\omega})\, \tau(\tilde{\omega})\right] Y^{\star}_{\ell m}(\hat{\boldsymbol{n}}) \tag{B16}$$

(with no implied sum over ℓ). These $\tau_{\ell m}$ can now be inserted into our formula for the m-mode visibility to obtain

$$\mathcal{V}^{m}_{ij}(\hat{\boldsymbol{n}}) = \sum_{\ell} B^{\ell m}_{ij}\, Y^{\star}_{\ell m}(\hat{\boldsymbol{n}}) \left[2\pi \int \tilde{\omega}\, d\tilde{\omega}\, J_0(\ell\tilde{\omega})\, \tau(\tilde{\omega})\right]. \tag{B17}$$

This is the central relation needed for the matched filter, as it expresses a linear relationship between the observable and the template. We recognize the combination $B^{\ell m}_{ij}\, Y^{\star}_{\ell m}$ as the beam transfer function, $B_{ij}(\hat{\boldsymbol{n}};\phi)$, evaluated at the position of the object but now modulated by the Fourier transform of τ.

The above expressions are all for a single frequency. If we now perform the Fourier transform in frequency, the term in square brackets becomes the 3D Fourier transform for an azimuthally symmetric function in cylindrical coordinates: $\tau(k_\perp, k_\parallel)$ with $\ell \simeq |k_\perp|$. For a narrow range of frequencies (corresponding to an astrophysical object such as a void or protocluster, for example) the $k_\perp$ probed by the interferometer are almost constant. For a wide range of frequencies, one must account for the shifting of $\boldsymbol{u}_{ij}$, and ℓ, with wavelength at fixed baseline separation. This represents no difficulty in principle, since we need only evaluate our template where there are data, but it formally breaks the Fourier transform property. It is important to note that this Fourier transform is not necessary for the matched filter, which can be written in visibility-frequency space.

As before, the matched filter is defined by

$$\hat{A}(\hat{\boldsymbol{n}}) = \frac{V^{m}_{ij} N^{-1} \mathcal{V}^{m}_{ij}(\hat{\boldsymbol{n}})}{\mathcal{V}^{m}_{ij}(\hat{\boldsymbol{n}}) N^{-1} \mathcal{V}^{m}_{ij}(\hat{\boldsymbol{n}})}, \tag{B18}$$

where the noise covariance matrix both includes the visibility noise and projects out contaminated modes. There are a few practical differences between this treatment and the flat-sky Fourier version we discuss in the main text. In the simplified flat sky treatment, shifting the matched filter to an arbitrary position $\boldsymbol{x}$ was simply a multiplication of $\hat{A}$ by $\exp(i\boldsymbol{kr})$, which allowed us to efficiently evaluate the matched filter at all possible void positions with inverse FFTs. In particular, the denominator of $\hat{A}$ is translation-invariant. While these simplifications remain true in the azimuthal direction, they no longer hold for the polar or radial directions. Therefore, one must explicitly evaluate the matched filter at all possible void positions. It may be possible to reduce the computational burden by using the Fourier versions of the expressions about more sparsely sampled central void positions. Since the precise implementation will be survey dependent, we do not pursue more detailed implementations here.

B4 A worked example

We conclude with an analytic example to make this formalism more concrete. Our discussion here parallels that in Bunn & White (2007). Consider the interferometer situated at the equator ($\theta_0 = \pi/2$, $\phi_0 = 0$) and looking directly overhead. The baselines, $\boldsymbol{u}_{ij}$, lie in the $y - z$ plane. We will consider two cases, a north–south baseline ($\boldsymbol{u} = u\hat{\boldsymbol{z}}$) and an east–west baseline ($\boldsymbol{u} = u\hat{\boldsymbol{y}}$). For a small field of view, we approximate the sky as flat with Cartesian coordinates ϕ, δ, where $\delta \equiv \pi/2 - \theta$ is the latitude. A Gaussian beam, normalized to unit peak, then has

$$B(\hat{\boldsymbol{n}}) = B(\phi, \delta) = \exp\left[-\frac{\phi^2 + \delta^2}{2\sigma^2}\right] \times \begin{cases} \exp[2\pi i u\phi] \text{ for } \hat{\boldsymbol{y}}(\mathrm{EW}) \\ \exp[2\pi i u\delta] \text{ for } \hat{\boldsymbol{z}}(\mathrm{NS}), \end{cases} \tag{B19}$$

where we have suppressed the ij indices labeling the visibility for convenience.

The visibility for this baseline is

$$V(\hat{\boldsymbol{n}}) = \int \mathrm{d}\hat{\boldsymbol{n}}\; B(\hat{\boldsymbol{n}}) T(\hat{\boldsymbol{n}}). \tag{B20}$$

Instead of immediately going to the spherical harmonic expansion, it is algebraicly illuminating and amusing to imagine the sky as a torus. The appropriate orthogonal basis is then the usual Fourier basis

$$V = \int \mathrm{d}\hat{\boldsymbol{n}} \sum B_{nm} \mathrm{e}^{-in\delta} \mathrm{e}^{-im\phi} \sum T_{n'm'} \mathrm{e}^{in'\delta} \mathrm{e}^{im'\phi}, \tag{B21}$$

which collapses to

$$V^{m} = (2\pi)^2 \sum_{nm} B_{nm} T_{nm}, \tag{B22}$$

where we have also implicitly gone to the m-mode basis (to account for the Earth's rotation). This expression is analogous to the spherical harmonic version. The beam multipole moments are then given by

$$B_{nm} = \int \frac{\mathrm{d}\phi\, \mathrm{d}\delta}{(2\pi)^2}\, B(\phi, \delta)\, \mathrm{e}^{in\delta} \mathrm{e}^{im\phi}. \tag{B23}$$

Since we assume the beams are compact in both ϕ and δ, we are free to extend the limits of integration to $\pm\infty$. For the specific case of our Gaussian beam, these integrals are then just Gaussian integrals and can be easily evaluated. For an EW baseline, we get

$$B_{nm} \propto \exp\left[-\frac{\sigma^2 n^2}{2}\right] \exp\left[-\frac{\sigma^2 (m \pm 2\pi u)^2}{2}\right] \tag{B24}$$

while for the NS baseline, we find

$$B_{nm} \propto \exp\left[-\frac{\sigma^2 (n \pm 2\pi u)^2}{2}\right] \exp\left[-\frac{\sigma^2 m^2}{2}\right], \tag{B25}$$

where the ± cases come from the two possible choices for the sign of u. These have a clear physical interpretation – the EW baseline probes modes centred around ($n = 0, m = \pm 2\pi u$) while the NS baseline is centred on ($n = 2\pi u, m = 0$). Note that these expressions indicate that it is the baseline distribution and the primary beam that delineate the range of (ℓm) modes which need to be kept in the sums of the previous section.

Returning to a spherical sky, we will adopt a similar strategy to understand what modes a given baseline probes. Since the beam is compact, we will approximate the spherical harmonics by a Fourier series, in which case the algebra proceeds as in the case of the torus. All that will remain will be to understand the correspondence between mode coefficients n on the torus and (ℓ, m) on the sphere.[9]

[9] Note that in the ϕ direction, both the sphere and the torus have Fourier expansions.

For our specific case, the multipole moments then become

$$B_{\ell m} = \int \mathrm{d}\phi\, \mathrm{d}\sin\delta\; Y_{\ell m}\left(\frac{\pi}{2} - \delta, \phi\right) B(\phi, \delta) \tag{B26}$$

$$\simeq \int_{-\infty}^{\infty} \mathrm{d}\phi\, \mathrm{d}\delta\; Y_{\ell m}\left(\frac{\pi}{2} - \delta, \phi\right) B(\phi, \delta), \tag{B27}$$

where we assume $\delta \ll 1$ in the second line. Near the equator, we have[10]

$$Y_{\ell m} \simeq N_{\ell m} \mathrm{e}^{im\phi} \begin{cases} \cos n_{\ell m}\delta \text{ for } \ell + m \text{ even} \\ -\sin n_{\ell m}\delta \text{ for } \ell + m \text{ odd} \end{cases}, \tag{B28}$$

where $N_{\ell m}$ is a constant and

$$n_{\ell m}^2 = \ell(\ell+1) - m^2 - \begin{cases} 0 \text{ for } \ell + m \text{ even} \\ 1 \text{ for } \ell + m \text{ odd} \end{cases}. \tag{B29}$$

Since we have reduced the problem to the toroidal sky case, we proceed as before and find that EW baselines measure modes centred on ($n_{\ell m} = 0, m = 2\pi u$). In the limit that $\ell \gg 1$, this implies that these baselines measure modes with $m \sim 2\pi u$, $\ell \sim m$. As one might expect, ℓ and m are coupled together by the spherical geometry. For NS baselines, the m-mode visibilities probe ($n_{\ell m} = 2\pi u, m = 0$) or $\ell \sim 2\pi u$, $m \sim 0$. The azimuthal symmetry of the baseline configuration is reflected in the visibilities isolating the $m \sim 0$ modes. These two cases represent the two limiting cases; baselines with components in both the EW and NS directions will probe more general ℓ, m modes.

For this particular case, this also completes the correspondence with the usual flat-sky treatment where a baseline measures a particular $\boldsymbol{\ell}$ Fourier mode. Here, the visibility m modes measure particular ℓ, m modes.

APPENDIX C: MODELING THE 21 CM SIGNAL

In the main text we have assumed that neutral hydrogen traces the mass field in an unbiased manner for the purposes of testing our matched filter on simulations. In this appendix we present a more refined model and argue that this assumption is conservative (for our purposes).

At low z most of the hydrogen in the Universe is ionized, and the 21 cm signal comes only from self-shielded regions such as galaxies.[11] Unfortunately there are not many observational constraints on the manner in which HI traces galaxies and haloes in the high-z Universe. There have been a large number of approaches to modelling this uncertain signal. Some approaches work directly at the level of the density field. For example, Shaw et al. (2014, 2015) use Gaussian density fields. Bull et al. (2015) assume a constant bias times the matter power spectrum (this is implicitly what we do in the main text, with $b = 1$). The CRIME code by Alonso, Ferreira & Santos (2014) uses lognormal realizations. Bagla & White (2003) selected dark matter particles based on a density threshold to mock up self-shielded regions.

An alternative is to use a halo-based approach, specifying the mass of HI to assign to a dark matter halo of a given mass, M_h. A popular model was introduced by Bagla, Khandai & Datta (2010), which populated haloes with circular velocities above 30 km/s with HI such that the HI mass saturates at high halo mass. A similar model was proposed by Barnes & Haehnelt (2010, 2014), who modelled the low-M cut-off as an exponential. Marín et al. (2010) use abundance matching between blue galaxies in the HI mass function at $z \approx 0$. Gong et al. (2011) employ a double power-law model. Seehars et al. (2016) propose a form with an exponential cut-off at both low and high halo masses. Padmanabhan & Refregier (2017) allow a non-unity slope in addition to the high and low mass cut-offs. The model we shall follow is due to Castorina & Villaescusa-Navarro (2016), which assumes

$$M_{\mathrm{HI}} \propto M_h^{\alpha}\, \mathrm{e}^{-M_{\mathrm{cut}}/M_h} \tag{C1}$$

with the constant of proportionality adjusted to match the observed value of Ω_{HI}. Aside from the normalization, this model has two free parameters, α and M_{cut}, which control the behaviour at high and low halo masses. There is evidence from simulations that $\alpha < 1$ (e.g. Davé et al. 2013; Villaescusa-Navarro et al. 2016) with $\alpha \approx 3/4$ a reasonable estimate. We shall use this value. Note that in contrast to some of the other models this assumption puts significant HI mass in higher mass haloes. There is some evidence at $z \simeq 0$ that HI is depleted in galaxies within clusters (e.g. Solanes et al. 2001), but the behaviour at $z \sim 1$ is unknown. In the simulations of Castorina & Villaescusa-Navarro (2016), the trend of M_{HI} with M_h is different at high and low redshifts. The remaining free parameter, M_{cut}, then adjusts the bias[12] of the HI. While a range of values is allowed within the observational constraints, typical values for the low-mass cut-off, M_{cut}, are around $10^{11}\, h^{-1}\, \mathrm{M}_{\odot}$. We shall explore a range around this value ($\lg M_{\mathrm{cut}} = 10.5$, 11 and 11.5 with masses in $h^{-1}\, \mathrm{M}_{\odot}$) to illustrate the effects.

The simulations used in the main body of this paper do not have sufficient resolution to track the haloes expected to host much of the HI at $z \sim 1$. Thus, in this appendix we use a different simulation, run with the same code, which employed 2560^3 particles in a box of side $256\, h^{-1}$ Mpc. This is the same simulation as used in Stark et al. (2015a,b), to which the reader is referred for more details. We generate a mock HI field from the $z \simeq 1$ halo catalog using the mapping of equation (C1).

We find voids in this simulation using the same technique as described in the main text. For completeness we also find protoclusters, in a manner similar to that of Stark et al. (2015a): starting from a friends-of-friends halo catalog (with a linking length of 0.168 times the mean interparticle spacing), we select each $z = 0$ halo more massive than $10^{14}\, h^{-1}\, \mathrm{M}_{\odot}$. We then track the particles within a few hundred kpc of the most bound particle back to $z = 1$. The centre of mass of these is taken to be the protocluster position at $z = 1$.

A comparison of the (real-space) profiles of protoclusters and voids in the dark matter and mock HI at $z \simeq 1$ is shown in Fig. C1 for three values of M_{cut}. The curves are noisier than from the larger volume simulations, due to the poorer statistics, however we see that the protoclusters in the HI have just as much broad, distributed

[10] The approximation agrees to the first two terms in the Taylor series. For completeness, we note that

$$N_{\ell m} = 2^m \sqrt{\pi} \sqrt{\frac{2\ell+1}{4\pi}} \sqrt{\frac{(\ell-m)!}{(\ell+m)!}}\, \frac{1}{\Gamma\left(\frac{1}{2} - \frac{(l+m)}{2}\right) \Gamma\left(1 + \frac{(l-m)}{2}\right)}. \tag{B30}$$

[11] Most likely between the outskirts of discs until where the gas becomes molecular within star-forming regions.

[12] For $\alpha = 3/4$ at $z \approx 1$ the bias ranges from 1.4 to 1.7 as $\lg M_{\mathrm{cut}}$ runs from 10.5 to 11.5 in $h^{-1}\, \mathrm{M}_{\odot}$ units. This is consistent with the amplitude of the measured clustering at $z \sim 1$ by Chang et al. (2010); Switzer et al. (2013) but those measurements are not precise enough to place strong limits on the bias.

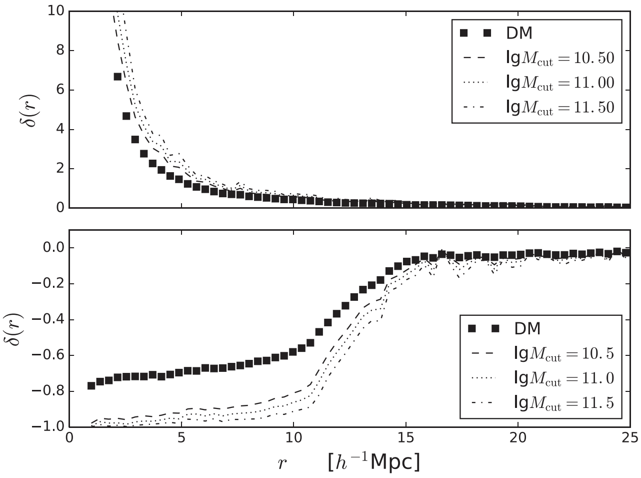

Figure C1. The (real-space) profiles of voids and protoclusters at $z \simeq 1$, as in Fig. 3. The upper panel shows the $M > 10^{14}\,h^{-1}\,\mathrm{M}_{\odot}$ protocluster profile in the mass and in the HI for three values of the cut-off mass in equation (C1), specified as $\log_{10}$ of the mass in $h^{-1}\,\mathrm{M}_{\odot}$. The lower panel shows the same comparison for voids of $10 < r_s < 15\,h^{-1}$ Mpc.

emission as the matter profiles. The voids in the HI have a qualitatively similar 'bucket shaped' profile to the mass density, but are notably more empty. As noted by Tinker & Conroy (2009), the halo mass function shifts dramatically to lower masses in underdense regions. Thus we expect to see voids in the massive halo and HI distributions be 'more empty' than in the mass. Given the greater contrast in HI than in the matter, our approximation in the main text is conservative from the point of view of finding protoclusters and voids with 21 cm experiments.

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