21 cm intensity mapping has arisen as a powerful probe of the high-redshift universe, but its potential is limited by extremely bright foregrounds and high source confusion. In this Letter, we propose a new analysis which can help solve both problems. From the combination of an intensity map with an overlapping galaxy survey we construct a new one-point statistic which is unbiased by foregrounds and contains information left out of conventional analyses. We show that our method can measure the HI mass function with unprecedented precision using observations similar to recent 21 cm detections.

PACS numbers:

Many experiments are studying the evolution of the universe with the redshifted 21 cm line [1–5]. These experiments seek to map vast swaths of the sky from the local universe to cosmic dawn using line intensity mapping [6–11]. Intensity maps do not resolve individual emitters, but instead map fluctuations in the density of neutral hydrogen. This gives them sensitivity to the aggregate emission from all galaxies, as well as the neutral intergalactic medium. By targeting a narrow emission line, maps can be made in three dimensions by observing at many closely-spaced frequencies. Intensity mapping surveys can quickly access large volumes of space, allowing unprecedented constraints on cosmology and fundamental physics [12–16]. In addition, they can provide statistical information about faint objects below detection thresholds of conventional surveys. Theory predicts how it is distributed with the HI mass function (HIMF) \( \phi_{\text{HI}}(M_{\text{HI}}) \), which gives the number density of halos with HI masses between \( M_{\text{HI}} \) and \( M_{\text{HI}} + dM_{\text{HI}} \). Let us consider an HIMF with a slightly-modified Schechter form:

\[
\phi(M_{\text{HI}}) = \ln(10) \phi_\star \left( \frac{M_{\text{HI}}}{M_\star} \right)^{1+\alpha} e^{-M_{\text{HI}}/M_\star - M_{\text{min}}/M_\star} .
\]

This basic form has been used many times in the literature [21–27]. We have added an additional low-mass cutoff at \( M_{\text{min}} \), necessary as intensity maps lack hard detection thresholds. A power spectrum, sensitive to only the Gaussian part of a field, can only access the first two moments of the HIMF [28], while it would take at least four numbers, \( (\phi_\star, M_\star, \alpha, M_{\text{min}}) \), to fully determine Eq. (1). Furthermore, there are substantial degeneracies between \( \phi \) and fundamental cosmological parameters that intensity maps might otherwise be ideally situated to measure, such as halo bias, the growth rate of density fluctuations, or the amplitude of primordial non-Gaussianity [22–31].

It was suggested in Refs. [32, 33] that this non-Gaussianity could be accessed using one-point statistics, as opposed to two-point statistics like power spectra. A technique called P(D) analysis, which has seen use for decades across the electromagnetic spectrum [34–39], allows mapping between the HIMF and the probability distribution function (PDF) \( P(T) \) of a voxel (or three-dimensional pixel) having intensity between \( T \) and \( T + dT \). This statistic, termed the Voxel Intensity Distribution (VID), has been shown to add significantly to the astrophysical information which can be gained from an intensity map [40].

Unfortunately, as with the auto-spectrum, the VID of a 21 cm map would be dangerously contaminated with foregrounds. Even after foreground cleaning, there can be substantial residual contamination. Rather than expose ourselves to this large potential source of bias, we in-
roduce here a one-point analogue to the cross-spectrum, extending the VID formalism to remove foreground bias. We will then forecast how this technique can be applied to the Parkes/2dF and GBT/WiggleZ data.

The approach we propose here relies on a simple fact: for independent random variables $T_1$ and $T_2$, the PDF $P_{1+2}(T)$ of their sum $T = T_1 + T_2$ is the convolution of their individual PDFs $P_1(T)$ and $P_2(T)$. This is straightforward to prove. We can write

$$P_{1+2}(T) = \int P_1(T')P_2(T-T')dT',$$

(2)

where $\delta_D$ is a Dirac delta function. Evaluating one integral leaves a convolution,

$$P_{1+2}(T) = \int P_1(T')P_2(T-T')dT' = (P_1 \ast P_2)(T),$$

(3)

We can further make use of the Fourier convolution theorem to turn this into a product in Fourier space:

$$\hat{P}_{1+2}(T) = \hat{P}_1(T)\hat{P}_2(T),$$

(4)

where $T = 2\pi/T$ is the Fourier conjugate of the intensity $T$ and $\hat{P}(T)$ is the Fourier transform of $P(T)$ (also known as the characteristic function). In an intensity map, the observed $T$ in a given voxel is the sum of a signal component $T_S$ with PDF $P_S$ and a contribution from noise and foregrounds which we will abbreviate as $T_{FG}$ with PDF $P_{FG}$. The full VID of a map is then the convolution of $P_S$ and $P_{FG}$.

As stated above, $P_{FG}$ is hard to model to sufficient precision, so it will be difficult to apply the VID statistic directly to 21 cm data. We will therefore use a separate data set with different systematics to isolate our signal. Assume that our volume contains both a 21 cm map and an optical galaxy survey. In each voxel we know the total radio intensity $T_S+T_{FG}$ and the number $N_{\text{det}}$ of detected optical galaxies. $T_S$ and $N_{\text{det}}$ will be correlated, due both to large-scale structure and the HI content of the optical galaxies. We can therefore construct conditional PDFs $P(T|N_{\text{det}})$. We refer to these PDFs as conditional Voxel Intensity Distributions, or CVIDs.

Crucially, each CVID will be a convolution of a signal part, which depends on $N_{\text{det}}$, and a noise/foreground part, which does not. If we compare voxels with different $N_{\text{det}}$, we can write,

$$\hat{P}(T|N_{\text{det}}) = \hat{P}_S(T|N_{\text{det}})\hat{P}_{FG}(T) = \hat{P}_S(T|N_{\text{det}})\hat{P}_S(T|N_{\text{det}})\hat{P}_{FG}(T),$$

(5)

From the second equality, it is clear that the above ratio is unbiased by foregrounds, as the deconvolution cancels out the component which is common to both CVIDs.

In practice, we do not compute continuous PDFs directly from maps. Instead, we estimate PDFs using histograms $B_i \approx P(T_i)\Delta T N_{\text{vox}}$, where $B_i$ is the number of voxels in a bin of width $\Delta T$ centered at $T_i$, and $N_{\text{vox}}$ is the total number of voxels. If we separate our map by $N_{\text{det}}$ values, compute histograms $B_{i|N_{\text{det}}}$ from each part, then compute their Fourier transforms $\tilde{B}_{i|N_{\text{det}}} = \Delta T \sum_j B_{j|N_{\text{det}}} \exp(i\kappa_i T_j)$, we can write down the CVID Ratio (CVR),

$$\tilde{R}_{i|N_{\text{det}}} = \tilde{B}_{i|N_{\text{det}}} / \tilde{B}_{i|N_{\text{det}}}^2$$

(6)

It is easy to show that the expectation value of $\tilde{R}_{i|N_{\text{det}}} N_{\text{det}}^2$ is proportional to the ratio from Eq. (6), and thus is unbiased by foregrounds.

Figure 11 shows an example of how this works. The left panel shows predicted histograms for a Parkes/2dF data set with a Gaussian $P_{FG}$, with the signal contribution computed as described below. The right panel shows the CVR computed only from the signal. Both panels include, for illustration, a toy data realization. If we assume that $T_S$ and $T_{FG}$ in each voxel are independent draws from $P_S$ and $P_{FG}$, then each $B_i$ is an independent draw from a binomial distribution with mean $B_i$ and variance

$$\text{var}(B_i) = \langle B_i \rangle (1 - \langle B_i \rangle /N_{\text{vox}}) \approx \langle B_i \rangle.$$ 

(7)

Based on the nearly-diagonal correlation matrices shown in Figure 2 of Ref. [40], this approximation seems reasonable, at least for low signal-to-noise data. We can therefore create a sample histogram by drawing randomly from binomial distributions, with errors following Eq. (7). Detailed testing of this binomial approximation will require actual mock maps, which we leave for future work.

From our toy data, it is clear that the simulation, which includes foregrounds, gives the same CVR as the foreground-free theory. Though we used a Gaussian $P_{FG}$, the same would hold for any general PDF. If we can estimate the error on $B_{i|N_{\text{det}}}$, either with Eq. (7) or with simulations, then we can propagate this error through to get the error on the CVR directly from the data. This has two important implications. First, it means that we do not need a model of the foregrounds to estimate errors on $\tilde{R}_{i|N_{\text{det}}}$. Second, it accounts for instabilities which appear when the denominator of Eq. (9) approaches zero, as seen at high-$T$ in Figure 11. Though the measured CVR deviates significantly from the expectation, the large error bar means that these points get correspondingly little weight.

Now we need to connect the CVR to the HIMF. We can relate the two using a modified $P(D)$ analysis [33, 37, 41]. In each voxel, there are $N_{\text{det}}$ detected optical galaxies and an unknown number $N_{\text{un}}$ of unresolved HI emitters, so:

$$\tilde{P}(T|N_{\text{det}}) = \tilde{P}_{\text{det}}(T|N_{\text{det}})\tilde{P}_{\text{un}}(T|N_{\text{det}}).$$

(8)

To compute $\tilde{P}_{\text{det}}$ and $\tilde{P}_{\text{un}}$, we need separate HIMFs for the detected and undetected galaxy populations. For now, we assume an exponential separation,

$$\phi_{\text{un}}(M_{\text{HI}}) = \phi(M_{\text{HI}})e^{-M_{\text{HI}}/M_{\text{cut}}},$$

(9)
galaxies and unresolved HI emitters. If there is no detection, we do not know the value of \( N_{\text{un}} \), so we have

\[ \phi_{\text{det}}(M_{\text{HI}}) = \phi(M_{\text{HI}}) \left( 1 - e^{-M_{\text{HI}}/M_{\text{cut}}} \right), \tag{10} \]

with free parameter \( M_{\text{cut}} \).

If \( N_{\text{det}} = 0 \), there is no contribution from detected galaxies and \( \mathcal{P}_{\text{det}}(T|0) = \delta_D(T) \). If \( N_{\text{det}} = 1 \), then \( \mathcal{P}_{\text{det}}(T|1) \) is proportional to \( \phi_{\text{det}} \), with appropriate normalization (see [33] for details). For higher values of \( N_{\text{det}} \), we can recursively apply Eq. (4) to get

\[ \tilde{\mathcal{P}}_{\text{det}}(T|N_{\text{det}}) = \left[ \tilde{\mathcal{P}}_{\text{det}}(T|N_{\text{det}} = 1) \right]^{N_{\text{det}}}. \tag{11} \]

Note that in Eq. (11) we have implicitly assumed that the HIMF in a voxel is independent of how many galaxies it contains. This is known to be inaccurate, as more massive objects will be more strongly biased, shifting the HIMF to larger masses in dense voxels. Accurately modeling this effect will likely require simulations, so we neglect it for now.

For undetected galaxies, we do not know the value of \( N_{\text{un}} \), so we have

\[ \mathcal{P}_{\text{un}}(T|N_{\text{det}}) = \sum_{N_{\text{un}}} \left[ \tilde{\mathcal{P}}_{\text{un}}(T|N_{\text{un}} = 1) \right]^{N_{\text{un}}} \mathcal{P}(N_{\text{un}}|N_{\text{det}}). \tag{12} \]

where \( \mathcal{P}_{\text{un}}(T|N_{\text{un}} = 1) \) is proportional to \( \phi_{\text{un}} \), \( \mathcal{P}(N_{\text{un}}|N_{\text{det}}) \) describes the correlation between optical galaxies and unresolved HI emitters. If there is no clustering, \( \mathcal{P}_{\text{un}} \) has no dependence on \( N_{\text{det}} \), and will cancel out of the CVR. However, we know that HI and optical galaxies should be at least somewhat correlated, though that correlation may be color- and scale-dependent [21, 42, 43]. Here we will employ a simple model to capture the leading order effects.

Following [33, 44], assume each voxel has expected counts \( \mu_{\text{un}} \) and \( \mu_{\text{det}} \) of detected and undetected galaxies, and that the actual counts are Poisson draws with these means. We take \( \mu_{\text{un}} = f_{\text{un}} \mu_{\text{det}} \), where \( f_{\text{un}} \) is the ratio \( N_{\text{un}}/N_{\text{det}} \) of the mean number counts. Galaxy count distributions are known to be reasonably-approximated by lognormal distributions [45, 46],

\[ \mathcal{P}_{\text{LN}}(\mu) = \frac{1}{\mu \sqrt{2\pi \sigma_{\text{G}}^2}} \exp \left\{ -\frac{1}{2\sigma_{\text{G}}^2} \left[ \ln \left( \frac{\mu}{\bar{N}} \right) + \frac{\sigma_{\text{G}}^2}{2} \right]^2 \right\}, \tag{13} \]

where the \( \sigma_{\text{G}} \) parameter acts as the “bias” in this model. The uncorrelated case corresponds here to \( \sigma_{\text{G}} = 0 \). We then can write

\[ \mathcal{P}(N_{\text{un}}|N_{\text{det}}) = \int \mathcal{P}(N_{\text{un}}|f_{\text{un}} \mu_{\text{det}}) \mathcal{P}(\mu_{\text{det}}|N_{\text{det}}) d\mu_{\text{det}}. \tag{14} \]

We can use Bayes’ Theorem to state that,

\[ \mathcal{P}(\mu_{\text{det}}|N_{\text{det}}) \propto \mathcal{P}(N_{\text{det}}|\mu_{\text{det}}) \mathcal{P}_{\text{LN}}(\mu_{\text{det}}), \tag{15} \]

where \( \mathcal{P}_{\text{LN}} \) acts as our “prior”. With our assumption of
Poisson statistics, we then have
\[
\mathcal{P}(N_{\text{un}}, N_{\text{det}}) \propto \int \mathcal{P}_{\text{Pois}}(N_{\text{un}} | \mu_{\text{det}}) \mathcal{P}_{\text{LX}}(\mu_{\text{det}}) \times \mathcal{P}_{\text{Pois}}(N_{\text{det}} | \mu_{\text{det}}) d\mu_{\text{det}},
\]
where \(\mathcal{P}_{\text{Pois}}(N | \mu)\) is the Poisson distribution with mean \(\mu\).

We can now finally predict a CVR from our HIMF model. We will now examine what information could be gained from such an analysis. Consider a model of the Parkes/2dF maps as described in Ref. [21], with free parameters \((\phi_*, M_*, \alpha, M_{\text{min}}, M_{\text{cut}}, \sigma_G)\). Assume that HI evolves negligibly from \(z = 0 - 0.05\), and use the best-fit HIMF from the ALFALFA survey [27] as a model, with \(\phi_* = (4.5 \pm 0.8) \times 10^{-3} \text{ Mpc}^{-3} \text{ dex}^{-1}\), \(\log(M_*/M_\odot) = 9.94 \pm 0.05\), and \(\alpha = -1.25 \pm 0.1\). We arbitrarily choose \(M_{\text{min}} = 10^5 M_\odot\) so that it falls below the ALFALFA detection threshold, and \(M_{\text{cut}} = 3 \times 10^9 M_\odot\) to get the correct number of optically detected galaxies [22]. We compute \(\sigma_G\) following Ref. [32]. We assume a Gaussian \(\mathcal{P}_{FG}\), but note again that this procedure would work regardless of the assumed form. As the strength of the contamination depends on how much prior foreground cleaning is assumed, we consider a pessimistic case with \(\sigma_{FG} = 35 \text{ mK}\) and an optimistic case with \(\sigma_N = 10 \text{ mK}\), arbitrarily chosen to very roughly match the Parkes maps from Figure 1 of Ref. [21]. Note that we do not allow \(\phi_*\) and \(\alpha\) to vary between the detected and undetected populations, but assume a single value for each.

We can forecast constraints on our model using the Fisher matrix formalism [47, 48]. We only consider fractional priors of 10 on both \(\phi_*\) and \(M_*\). Though we do not directly forecast the higher-redshift GBT/WiggleZ survey, we can get a rough idea of its performance using a weaker set of priors, which we take as ten times worse than the ALFALFA priors. In both cases, we assume 10% prior knowledge of \(\sigma_G\), which would have to come from simulations, and uninformative fractional priors of 10 on both \(M_{\text{cut}}\) and \(M_{\text{min}}\). We assume a cosmology consistent with the Planck 2015 results [49].

Figure 2 shows the results of our Fisher forecasts. We plot 1σ confidence intervals around our fiducial models for the total, detected, and undetected populations. Even in the most pessimistic case, with strong foregrounds and weak (10x ALFALFA) priors, we get a good measurement...
of the bright end of the HIMF and the detected galaxy HIMF. If we know enough to trust the strong ALFALFA priors, then the CVID adds important constraints on the HIMF of the optical galaxies and a modest measurement of that of the unresolved galaxies, both of which cannot be obtained from ALFALFA alone. With 10 mK foregrounds, the CVID dramatically improves on the priors. The very brightest end of the HIMF is still dominated by ALFALFA, but the CVID has added a wealth of information about faint galaxies which cannot be obtained conventionally. Unfortunately, even the intensity mapping data loses sensitivity at the very brightest end of the HIMF. Even stronger foreground cleaning would be needed to measure $M_{\text{min}}$.

These forecasts clearly demonstrate the utility of this method. However, we have made a number of assumptions here that deserve further study. The lognormal form of the galaxy number count PDF is likely overly simplistic, and could be replaced by a more sophisticated prescription. In Eqs. (9) and (11), we assumed that optical galaxies host the brightest 21 cm emitters. This should hold for blue, gas-rich galaxies, but a number of bright optical galaxies are red and gas-poor, and therefore have weaker HI emission. As mentioned above, we have entirely neglected luminosity-dependent bias in our forecasting. Though the Parkes and GBT surveys are single-dish, many 21 cm experiments are interferometric, which may introduce extra systematics. Finally, we assumed that we can apply some degree of foreground cleaning to our data without affecting our signal. In the Parkes foreground cleaning, the signal was suppressed along with the foregrounds, which if uncorrected would bias our CVR measurements. These caveats motivate additional study of this method using mock data sets, as in Ref. [40].

Though we have focused here on the HIMF, the potential utility of this method extends much further. For example, one could study the HI-color relation mentioned above. With intensity maps of other lines, one could, for example, measure the molecular gas properties of Lyman-alpha emitters, or study AGN feedback. The CVID can also be modified to combine intensity maps of different lines, another common experimental target.

Cross-correlations have long been a powerful tool for cosmologists, and will only become more critical as more intensity mapping surveys come online. With this work we have demonstrated a one-point cross-correlation method that can be used to clean foregrounds and probe astrophysics inaccessible to conventional surveys. With some refinement, this will be a valuable tool for many future experiments.

The authors would like to thank Ue-Li Pen and Hansa Padmanabhan for useful conversations. Philippe Berger was supported by Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Copyright 2019. All rights reserved.

[1] S. J. Tingay, R. Goeke, J. D. Bowman, D. Emrich, S. M. Ord, D. A. Mitchell, M. F. Morales, T. Booler, B. Crosse, R. B. Wayth, et al., PASA 30, e007 (2013), 1206.6945.
[2] M. P. van Haarlem, M. W. Wise, A. W. Gunst, G. Heald, J. P. McKean, J. W. T. Hessels, A. G. de Bruyn, R. Nijboer, J. Swinbank, R. Fallows, et al., A&A 556, A2 (2013), 1305.3550.
[3] K. Bandura, G. E. Addison, M. Amiri, J. R. Bond, D. Campbell-Wilson, L. Connor, J.-F. Cliche, G. Davis, M. Deng, N. Denman, et al., in Ground-based and Airborne Telescopes V (2014), vol. 9145 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 914522, 1406.2288.
[4] Z. S. Ali, A. R. Parsons, H. Zheng, J. C. Pober, A. Liu, J. E. Aguirre, R. F. Bradley, G. Bernardi, C. L. Carilli, C. Cheng, et al., ApJ 809, 61 (2015), 1502.06016.
[5] Y. Xu, X. Wang, and X. Chen, ApJ 798, 40 (2015), 1410.7794.
[6] L. B. Newburgh, K. Bandura, M. A. Bucher, T.-C. Chang, H. C. Chiang, J. F. Cliche, R. Davé, M. Dobbs, C. Clarkson, K. M. Ganga, et al., in Ground-based and Airborne Telescopes VI (2016), vol. 9906 of Proc. SPIE, p. 99065X, 1607.02059.
[7] D. R. DeBoer, A. R. Parsons, J. E. Aguirre, P. Alexander, Z. S. Ali, A. P. Beardsley, G. Bernardi, J. D. Bowman, R. F. Bradley, C. L. Carilli, et al., PASP 129, 045001 (2017), 1606.07473.
[8] M. Suginoohara, T. Suginoohara, and D. N. Spergel, ApJ 512, 547 (1999), astro-ph/9803236.
[9] M. F. Morales and J. S. B. Wyithe, ARA&A 48, 127 (2010), 0910.3010.
[10] J. R. Pritchard and A. Loeb, Reports on Progress in Physics 75, 086901 (2012), 1109.6012.
[11] E. D. Kovetz, M. P. Viero, A. Lidz, L. Newburgh, M. Rahman, E. Switzer, M. Kamionkowski, J. Aguirre, M. Alvarez, J. Bock, et al., arXiv e-prints arXiv:1709.09066 (2017), 1709.09066.
[12] J. R. Shaw, K. Sigurdson, M. Sitwell, A. Stebbins, and U.-L. Pen, Phys. Rev. D 91, 083514 (2015), 1401.2095.
[13] A. Liu, J. R. Pritchard, R. Allison, A. R. Parsons, U. c. v. Seljak, and B. D. Sherwin, Phys. Rev. D 93, 043013 (2016), URL https://link.aps.org/doi/10.1103/PhysRevD.93.043013
[14] A. Ewall-Wice, J. Hewitt, A. Mesinger, J. S. Dilлон, A. Liu, and J. Pober, Monthly Notices of the Royal Astronomical Society 458, 2710 (2016), ISSN 0035-8711, http://oup.prod.sis.lan/mnras/article-pdf/458/3/2710/8005817/stw452.pdf, URL https://doi.org/10.1093/mnras/stw452
[15] A. Liu and A. R. Parsons, Monthly Notices of the Royal Astronomical Society 457, 1864 (2016), ISSN 0035-8711, http://oup.prod.sis.lan/mnras/article-pdf/457/2/1864/2013449/stw071.pdf, URL https://doi.org/10.1093/mnras/stw071
[16] S. Foreman, P. D. Meerburg, A. van Engelen, and J. Meyers, Journal of Cosmology and Astro-Particle Physics
1903.03144.

[63] S. Dumitru, G. Kulkarni, G. Lagache, and M. G. Haehnelt, MNRAS 485, 3486 (2019), 1802.04804.

[64] D. T. Chung, M. P. Viero, S. E. Church, R. H. Wechsler, M. A. Alvarez, J. R. Bond, P. C. Breysse, K. A. Cleary, H. K. Eriksen, M. K. Foss, et al., ApJ 872, 186 (2019), 1809.04550.

[65] P. C. Breysse and R. M. Alexandroff, arXiv e-prints arXiv:1904.03197 (2019), 1904.03197.

[66] P. C. Breysse and M. Rahman, ArXiv e-prints (2016), 1606.07820.

[67] J. Fonseca, R. Maartens, and M. G. Santos, MNRAS 479, 3490 (2018), 1803.07077.

[68] A. Beane, F. Villaescusa-Navarro, and A. Lidz, ApJ 874, 133 (2019), 1811.10609.