Cosmological Inference using Gravitational Wave Standard Sirens: A Mock Data Challenge

Rachel Gray,[1,*] Ignacio Magaña Hernandez,[2,†] Hong Qi,[3,‡] Ankan Sur,[4,5,§] Patrick R. Brady,[2]
Hsin-Yu Chen,[6] Will M. Farr,[7,8] Maya Fishbach,[9] Jonathan R. Gair,[10,11] Archisman Ghosh,[4,12,13,14]
Daniel E. Holz,[9] Simone Mastrogiovanni,[15] Christopher Messenger,[1] Danièle A. Steer,[15] and John Veitch[1]

[1]*SUPA, University of Glasgow, Glasgow G12 8QQ, United Kingdom*
[2]*University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA*
[3]*Cardiff University, Cardiff CF24 3AA, United Kingdom*
[4]*Nikhef, Science Park 105, 1098 XG Amsterdam, The Netherlands*
[5]*Nicolaus Copernicus Astronomical Center, Polish Academy of Sciences, 00-716, Warsaw, Poland*
[6]*Black Hole Initiative, Harvard University, Cambridge, Massachusetts 02138, USA*
[7]*Department of Physics and Astronomy, Stony Brook University, Stony Brook NY 11794, USA*
[8]*Center for Computational Astronomy, Flatiron Institute, New York NY 10010, USA*
[9]*University of Chicago, Chicago, IL 60637, USA*
[10]*School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom*
[11]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Potsdam-Golm, 14476, Germany*
[12]*Delta Institute for Theoretical Physics, Science Park 904, 1090 GL Amsterdam, The Netherlands*
[13]*Lorentz Institute, Leiden University, PO Box 9506, Leiden 2300 RA, The Netherlands*
[14]*GRAPPA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*
[15]*AstroParticule et Cosmologie, Université Paris Diderot, F-75205 Paris, France*

(Dated: October 7, 2019)

The observation of binary neutron star merger GW170817, along with its optical counterpart, provided the first constraint on the Hubble constant H_0 using gravitational wave standard sirens. When no counterpart is identified, a galaxy catalog can be used to provide complementary redshift information. However, the true host might not be contained in a catalog which is not complete out to the limit of gravitational-wave detectability. These electromagnetic and gravitational-wave selection effects must be accounted for. We describe and implement a method to estimate H_0 using both the counterpart and the galaxy catalog standard siren methods. We perform a series of mock data challenges using binary neutron star mergers to confirm our ability to recover an unbiased estimate of H_0. Our simulations used a simplified universe with no redshift uncertainties or galaxy clustering, but with different magnitude-limited catalogs and assumed host galaxy properties, to test our treatment of both selection effects. We explore how the incompleteness of catalogs affects the final measurement of H_0, as well as the effect of weighting each galaxy's likelihood of being a host by its luminosity. In our most realistic simulation, where the simulated catalog is about three times denser than the density of galaxies in the local universe, we find that a 4.4% measurement precision can be reached using galaxy catalogs with 50% completeness and 249 binary neutron star detections with sensitivity similar to that of Advanced LIGO's second observing run.

arXiv:1908.06050v3 [gr-qc] 4 Oct 2019

I. INTRODUCTION

The idea that gravitational waves (GW) detections can be used for the inference of cosmological parameters, such as the Hubble constant (H_0), was first proposed over three decades ago by Bernard Schutz [1]. The key to this process is that GW signals from compact binary coalescences (CBCs) act as standard sirens, in the sense that they provide a self-calibrated luminosity distance to the source. This can be obtained directly from the GW signal, and is therefore entirely independent of the cosmic distance ladder [2–9]. With the addition of redshift information for each source we then have the required input for cosmological inference.

At the time of writing, the current percent level state-of-the-art electromagnetic (EM) measurements of H_0 are in tension with each other. The Planck experiment uses measurements of cosmic microwave background (CMB) anisotropies and provides a value of $H_0 = 67.4 \pm 0.5\ \mathrm{km\ s^{-1}\,Mpc^{-1}}$ [10]. The Supernovae, H_0, for the Equation of State of Dark energy (SH0ES) experiment measures distances to Type Ia supernovae standard candles making use of the cosmic distance ladder, and gives $H_0 = 74.03 \pm 1.42\ \mathrm{km\ s^{-1}\,Mpc^{-1}}$ [11]. These two independent measurements of H_0 are in tension at the level of $\sim 4.4\text{-}\sigma$ [11]. While the early-universe Planck measurements are also favored by measurements using supernovae calibrated with Baryon Acoustic Oscillations [12], and the SH0ES results agree with local gravitational lensing measurements by the H0LiCOW Collaboration [13], calibration of supernovae using the Tip of the Red Giant Branch yields H_0 midway between the two [14].

This indicates the possibility that at least one of these measurements is subject to unknown systematics, or it could be an indication of new physics causing the discrepancy between the local measurements and the non-local (early universe) CMB based measurement. This makes a GW standard siren measurement of H_0 particularly interesting, as this will provide a local constraint on H_0 which is entirely independent. In this manner, the use of GWs as standard sirens may allow us to

* rachel.gray@ligo.org
† ignacio.magana@ligo.org
‡ hong.qi@ligo.org
§ ankan.sur@ligo.org

arbitrate the current situation, indicating either a bias in the current measurements, or pointing towards new physics.

The detection of the binary neutron star (BNS) event GW170817 [15], together with its optical counterpart [16, 17] led to the first standard siren measurement of H_0 [18]. The counterpart associated with GW170817 allowed for the identification of its host galaxy, NGC4993, and hence a direct measurement of its redshift, which in turn resulted in the inferred value H_0=70^{+12}_{-8} km s^{-1} Mpc^{-1}. Future counterpart standard siren measurements are expected to constrain H_0 to the percent level [3–7].

Central to the aims of this paper, is the case where an EM counterpart is not observed, and how H_0 inference can still be performed. In particular, the method proposed by Schutz in 1986 [1, 19] allows the use of galaxy catalogs to provide redshift information for potential host galaxies within the event's GW sky-localization. The idea is that, by marginalizing over the possible discrete values of redshift for each GW detection we account for uncertainty as to which galaxy is the true host. By combining the information from many GW events, the contributions from the true host galaxies will grow since they will all share the same true H_0. Contributions from the others will statistically average out, leading to a constraint on H_0 and possibly other cosmological parameters.

Over the course of the first observing run (O1) and the second observing run (O2) a total of 11 GW events were detected by the advanced LIGO and Virgo detectors: 10 are binary black hole (BBH) events and one is a BNS event [20]. The "galaxy catalog" method has been independently applied to both the BNS event GW170817 (without assuming NGC4993 is the host) [21], and the BBH event GW170814 [22] resulting in posterior probability distributions on H_0 where the posterior from GW170814 was broader than (but consistent with) that obtained from GW170817. The difference in the widths of the H_0 constraints is an expected result due to the larger localization volume associated with GW170814, and the high number of galaxies it contained. Using the detections from O1 and O2, multiple GW events have been combined to give the latest standard siren measurement of H_0 [23] using the methodology presented in this paper.

Predictions suggest that it will be possible to constrain H_0 to less than 2% within 5 years of the start of the third observing run (O3) and to 1% within a decade, though this is dependent on the number of events observed with EM counterparts [6], and this may change as our understanding of astrophysical rates improves. Simulations in [6] and [21], which assume complete catalogs based on realistic large-scale structure simulations, find that for BNSs without counterparts, the convergence is $40\%/\sqrt{N}$. The convergence found there for BBHs is much slower, as BBHs are typically detected at greater distances with larger localization volumes.

The prospects of identifying a transient EM counterpart will certainly increase, and correspondingly, the number of candidate host galaxies in a catalog will decrease, with improved event sky-localizations as future GW observatories join the detector network [24]. With the Japanese detector KAGRA set to join late in 2019 [25] and LIGO-India approved for construction [26], the next decade of standard siren cosmology is set to be very exciting.

O3 began on April 1$^{\text{st}}$ 2019 and will last for one full year. The sensitivities of the LIGO and Virgo detectors have improved since O2, leading to an increased detection rate of GW candidates[1] [27]. This is the first observing run for which there will be 3 detectors operating for the entirety of the run. Having more detectors improves the duty-cycle of the network, *i.e.* the fraction of run time for which one or more detectors in the network is online, and also increases the rate of three-detector detections, which will likely be better localized on the sky than the two-detector ones. This is important, both in terms of performing EM follow-up for EM counterparts practically [28], and for reducing the number of possible host galaxies for events in the case where a counterpart is not observed.

This paper presents the Bayesian framework behind the `gwcosmo` code, a product of the LIGO and Virgo Collaboration (LVC) which was used to measure H_0 using detected GW events from O1 and O2 [23]. We present results from a series of mock data challenges (MDCs) which were designed specifically to test this method's robustness against some of the most common pitfalls, in particular, GW selection effects which affect all H_0 measurements, and EM selection effects, which are relevant in the content of galaxy catalogs. This method builds upon the Bayesian framework first presented in [19] which has subsequently been extended, modified and independently derived by multiple authors [5, 6, 21, 22, 29]. With specific care regarding selection effects we outline methods for constraining H_0 using both the "galaxy catalog" and "EM counterpart" approaches.

This paper is structured as follows. Section II presents the Bayesian framework used to estimate the posterior on H_0. Section III discusses the design and preparation of the MDCs. In Section IV we present our results. We conclude in Section V giving a detailed discussion of results and providing guidance for future work. Some of the details of the Bayesian method have been set aside to be discussed in an Appendix.

II. METHODOLOGY

The late-time cosmological expansion in a Friedmann-Lemaître-Robertson-Walker universe is characterized by the Hubble parameter as a function of the redshift z,

$$H(z) = H_0 \sqrt{\Omega_{\rm m}(1+z)^3 + \Omega_{\rm k}(1+z)^2 + \Omega_\Lambda}\,, \qquad (1)$$

where H_0 is the Hubble constant, the rate of expansion in the current epoch, and $\Omega_{\rm m}$ and Ω_Λ are the fractional matter density (including baryonic and cold dark matter) and fractional dark energy density (assumed to be due to a cosmological constant) respectively; $\Omega_{\rm k}$ is the fractional curvature energy

[1] At present in O3 the detectors appear to be averaging the detection of one GW candidate per week. If all of these candidates are ultimately identified as real GW events, then O3 within its first two months will have exceeded the total number of detections of O1 and O2.

density which is identically zero for a "flat" universe consistent with observations. Additionally, we have the constraint $\Omega_{\rm m} + \Omega_{\rm k} + \Omega_\Lambda = 1$ for all the components contributing to the energy density of universe at the present epoch.

The expansion history of the universe maps to a "redshift-distance relation" associating the redshift z of observable sources to their luminosity distance $d_L(z)$ (see *e.g.* [30]) as,

$$d_L(z) = \frac{c\,(1+z)}{H_0} \int_0^z \frac{H_0}{H(z')}\, dz' \,. \tag{2}$$

From the relation between observed z and d_L to sources (EM sources such as variable stars or supernovae, or GW sources), one can measure the cosmological parameters appearing in $H(z)$. With knowledge of the other cosmological parameters $\{\Omega_{\rm m}, \Omega_{\rm k}, \Omega_\Lambda\}$ coming from independent observations, the redshift-distance relation can be used to measure H_0. We would like to note that with prior knowledge on the other cosmological parameters coming from EM observations, the measurement made with GW detections are not strictly independent measurements.

At low redshifts $z \ll 1$, the redshift-distance relation can be approximately described by the linear Hubble relation,

$$d_L(z) \approx c\, z/H_0 \,, \tag{3}$$

which contains H_0 but is independent of the other cosmological parameters. With this approximate linear relation at low redshifts, any measurement of H_0 with GWs is independent of the values of the other cosmological parameters.

A. Standard Sirens

The amplitude of the observed strain is inversely proportional to the luminosity distance to the GW source. For compact binary sources, to the leading order (see, *e.g.* [31]),

$$A \sim \frac{\mathcal{M}_z^{5/3}}{d_L} \left[f(\mathcal{M}_z, t)\right]^{2/3} \,, \tag{4}$$

where A is the observed amplitude, $\mathcal{M}_z \equiv \mathcal{M}(1+z)$ is the detector frame "chirp" mass (redshifted relative to its source frame value $\mathcal{M}$), and $f(\mathcal{M}_z, t)$ is the evolution of the observed frequency. $\mathcal{M}_z$ can be estimated from the observed $f(\mathcal{M}_z, t)$. The luminosity distance, d_L, can then obtained directly from the amplitude of signal. This makes compact binaries self-calibrated luminosity distance indicators or "standard sirens" unlike EM distance indicators which need to undergo calibration via multiple rungs of the cosmic distance ladder. The redshift of the GW source, also required for cosmological inference, remains degenerate with the source's mass, contained within $\mathcal{M}_z$, and needs to be estimated in alternate ways.

B. Galaxy Information

There are multiple ways in which EM observations can provide complementary redshift[2] information. A BNS event may be detected in coincidence with an EM counterpart, which can be associated with the host galaxy to provide a direct measurement of the redshift of the source. More generically, a GW event may not have a detected EM counterpart, in which case one needs to fall back on the method outlined by Schutz [1] and use potential host galaxies within the event's sky localization region for the redshift information for the source. Two possibilities come up: (i) to use available galaxy catalogs, or (ii) to conduct dedicated EM follow-up on the event's sky region, mapping the galaxies within that area to as great a depth as possible to maximize the redshift information available.

When using galaxy catalogs to provide the prior redshift information, the possibility that the host galaxy lies beyond the reach of the catalog must be taken into account. EM telescopes are flux limited, and can be reasonably modeled as having an apparent magnitude limit. This means that galaxy catalogs are inherently biased towards containing objects which are brighter and/or nearer-by, although there may be other selection effects due to galaxy color or size, depending on the catalog. These EM selection effects must be compensated for. Carrying out dedicated EM follow-up will, to some degree, mitigate this issue, as it will allow for far deeper coverage over a small section of the sky. For nearby events, the possibility that the host galaxy lies beyond the telescope's sight may be negligible. However, the time and resources required for dedicated EM follow-up means that the default approach for GW events observed without counterparts will be to use pre-existing catalogs.

In either case, the uncertainty associated with each galaxy's redshift must be taken into account, including the redshift error due to the galaxy's peculiar velocity, v_p, and, in cases where the redshift is estimated photometrically, a much larger uncertainty due to the photometric algorithm. Peculiar velocities are significant for nearby galaxies. The effect of v_p on the measurement of H_0 may be small if there are a large number of potential host galaxies in the GW event's sky-localization, but for a small number of galaxies, and for the counterpart case, this effect is particularly noticeable (see the treatment in [18]). Photometric redshift uncertainties on the other hand become important even slightly farther away due to lack of spectroscopic data in galaxy catalogs (see, *e.g.*, discussion in [23]).

[2] There are ways of obtaining the redshift independent of EM observations, by using known population properties such as the mass distribution [32, 33], or the neutron star equation-of-state [34].

C. Bayesian Framework

The posterior probability on H_0 from $N_{\rm det}$ GW events is computed as follows:

$$p(H_0|\{x_{\rm GW}\},\{D_{\rm GW}\}) \propto p(H_0)p(N_{\rm det}|H_0)\prod_i^{N_{\rm det}} p(x_{{\rm GW}i}|D_{{\rm GW}i},H_0) \tag{5}$$

where $\{x_{\rm GW}\}$ is the set of GW data, $D_{\rm GW}$ indicates that the event was detected as a GW and $p(H_0)$ is the prior on H_0. For a given H_0, the term $p(N_{\rm det}|H_0)$ is the likelihood of detecting $N_{\rm det}$ events. It depends on the intrinsic astrophysical rate of events in the source frame, $R = \frac{\partial N}{\partial V \partial T}$. The total number of expected events is given by $N_{\rm det} = R\langle VT\rangle$, where $\langle VT\rangle$ is the average of the comoving volume multiplied by the observation time. By choosing a "non-informative" prior on rate, $p(R) \propto 1/R$, the dependence on H_0 drops out [35]. For simplicity this approximation is made throughout the analysis.

The remaining term factorizes into likelihoods for each detected event. Using Bayes' theorem we can write it as,

$$\begin{aligned} p(x_{\rm GW}|D_{\rm GW},H_0) &= \frac{p(D_{\rm GW}|x_{\rm GW},H_0)p(x_{\rm GW}|H_0)}{p(D_{\rm GW}|H_0)} \\ &= \frac{p(x_{\rm GW}|H_0)}{p(D_{\rm GW}|H_0)}, \end{aligned} \tag{6}$$

where we set $p(D_{\rm GW}|x_{\rm GW},H_0) = 1$, since the analysis is only carried out when the signal-to-noise ratio (SNR), ρ, associated with $x_{\rm GW}$ passes some detection statistic threshold $\rho_{\rm th}$ – it is a prerequisite that the event has been detected. Calculating $p(D_{\rm GW}|H_0)$ requires integrating over all possible realizations of GW events, with a lower integration limit of $\rho_{\rm th}$:

$$p(D_{\rm GW}|H_0) = \int_{\rho>\rho_{\rm th}}^{\infty} p(x_{\rm GW}|H_0)dx_{\rm GW}. \tag{7}$$

For explicit details on the calculation of $p(D_{\rm GW}|H_0)$ see Appendix 5. The term $p(D_{\rm GW}|H_0)$ depends on properties of the GW source population (*e.g.* the mass distribution), but in this work, for simplicity, it is assumed that the population properties are known exactly.

1. The galaxy catalog method

In the galaxy catalog case, the EM information enters the analysis as a prior, made up of a series of possibly smoothened delta functions[3] at the redshift, right ascension (RA) and declination (dec) of the possible source locations. As we are in the regime where (especially for BBHs) galaxy catalogs cannot be considered complete out to the distances to which GW events are detectable, we have to consider the possibility that the host galaxy is not contained within the galaxy catalog, but lies somewhere beyond it. In order to do so, we marginalize the likelihood over the case where the host galaxy is, and is not, in the catalog (denoted by G and $\bar{G}$ respectively):

$$p(x_{\rm GW}|D_{\rm GW},H_0) = \sum_{g=G,\bar{G}} p(x_{\rm GW}|g,D_{\rm GW},H_0)p(g|D_{\rm GW},H_0)\,. \tag{8}$$

We defer detailed derivations for each of the components of Eq. (8) to Appendix 2.

[3] While uncertainties on the galaxy sky-coordinates can be safely ignored, the error on the redshift can be modeled with a Gaussian or a more complicated distribution.

2. The counterpart method

The method outlined above is for the galaxy catalog case, in which no EM counterpart is observed, or expected. We also consider the case where we observe an EM counterpart. The main difference is the inclusion of a likelihood term for the EM counterpart data, mirroring that of the GW data.

The likelihood in this case, which is the term within the product in Eq. (5), is given by:

$$\begin{aligned} &p(x_{\rm GW},x_{\rm EM}|D_{\rm GW},D_{\rm EM},H_0) \\ &\quad= \frac{p(x_{\rm GW},x_{\rm EM}|H_0)p(D_{\rm GW},D_{\rm EM}|x_{\rm GW},x_{\rm EM},H_0)}{p(D_{\rm GW},D_{\rm EM}|H_0)}, \\ &\quad= \frac{p(x_{\rm GW}|H_0)p(x_{\rm EM}|H_0)}{p(D_{\rm EM}|D_{\rm GW},H_0)p(D_{\rm GW}|H_0)}, \\ &\quad\propto \frac{p(x_{\rm GW}|H_0)p(x_{\rm EM}|H_0)}{p(D_{\rm GW}|H_0)}. \end{aligned} \tag{9}$$

where $x_{\rm EM}$ refers to the EM counterpart data and $D_{\rm EM}$ denotes that the counterpart was detected. In the numerator we have assumed that the GW and EM data are independent of each other and so the joint GW-EM likelihood factors out. $p(D_{\rm GW},D_{\rm EM}|x_{\rm GW},x_{\rm EM},H_0) = 1$ whenever we have GW and EM data. We make the assumption that the detection of an EM counterpart is also flux-limited, and hence also dependent on the luminosity distance (as opposed to the redshift). Since both $D_{\rm GW}$ and $D_{\rm EM}$ are functions of only the luminosity distance, the term $p(D_{\rm EM}|D_{\rm GW},H_0)$ is then some constant independent of H_0.

In an ideal scenario, the observation of an EM counterpart will allow for the identification of one of the galaxies in the neighboring region as the host of the GW event. In the case where the EM counterpart cannot be unambiguously linked to a host galaxy, this uncertainty can also be taken into account. See Appendix 4 for more details.

III. THE MOCK DATA CHALLENGE

In this section we describe a series of mock data challenges (MDCs) that we use to test our implementation of the Bayesian formalism described in Section II and its ability to infer the posterior on H_0 under different conditions. For each case, the MDC consists of (i) simulated GW data, and (ii) a corresponding mock galaxy catalog. In all cases, we make several idealized assumptions regarding both the GW

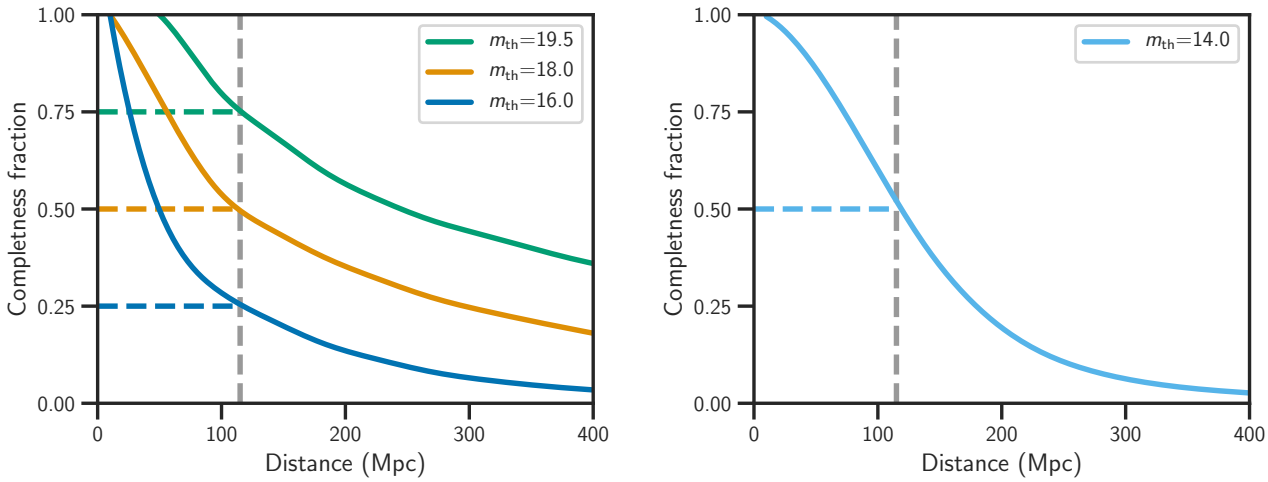

FIG. 1. Galaxy catalog completeness fractions for MDC2 and MDC3. *Left panel:* Galaxy number completeness fraction defined in Eq. (11) as a function of luminosity distance for the three MDC2 sub-catalogs. The lines in green, orange and blue correspond to the catalogs with $m_{\rm th}$ = 19.5, 18, and 16 respectively; these correspond to completeness fractions of 75%, 50% and 25% out to a fiducial reference distance of 115 Mpc (shown as a vertical grey line). *Right panel:* The galaxy luminosity completeness fraction defined in Eq. (15) as a function of luminosity distance for the MDC3 catalog, with $m_{\rm th}$ = 14. At the reference distance of 115 Mpc (vertical grey line), this is corresponds to a completeness fraction of ∼ 50%.

and galaxy data. On the GW side, the detection efficiency and the source population properties are assumed to be known exactly. On the galaxy side, the luminosity function and magnitude limit are also assumed to be known exactly in each case, so that the incompleteness correction can be calculated exactly. Further, we neglect the effect of large-scale structure in the mock catalogs.

For each of the MDCs we use an identical set of simulated BNS events from The First Two Years of Electromagnetic Follow-Up with Advanced LIGO and Virgo dataset [36, 37][4]. The set of BNS events comes from an end-to-end simulation of approximately 50,000 "injected" events in detector noise corresponding to a sensitivity similar to what was achieved during O2. Only a subset (approximately 500 events) were "detected" by a network of two or three detectors with the `GstLAL` matched filter based detection pipeline [38]. From the above detections, 249 events were randomly selected (in a way that no selection bias was introduced), and these events underwent full Bayesian parameter estimation using the `LALInference` software library [39] to obtain gravitational wave posterior samples and skymaps. Consistency with the First Two Years parameter estimation results in terms of sky localization areas and 3D volumes was demonstrated in [40]. It is these 249 events of the First Two Years dataset and the associated GW data which we use for our analysis.

[4] The set of simulations in [37] are more realistic with the same injections in (recolored) detector data as opposed to Gaussian noise used in [36]. Correspondingly, the detection criterion is in terms of a false alarm rate (FAR) rather than a threshold on the SNR. This is an important distinction, particularly affecting events marginally close to the detection threshold. We use the simplified set of simulations in [36] noting potential caveats.

The galaxy catalogs for each iteration of the MDC described below are designed to test a new part of the `gwcosmo` methodology in a cumulative fashion, starting with GW selection effects, adding in EM selection effects, and finally testing the ability to utilize the information available in the observed brightness of host galaxies, by weighting the galaxies with a function of their intrinsic luminosities.

The starting point for the galaxy catalogs is to take all 50,000 injected events from the First Two Years dataset and simulate a mock universe, which contain a galaxy corresponding to each injected event's sky location and luminosity distance, where the latter is converted to a redshift using a fiducial "simulated" H_0 value of 70 km s^{-1} Mpc^{-1}. The First Two Years data was originally simulated in a universe where GW events followed a d_L^2 distribution, and there was no distinction between the source frame and the (redshifted) detector frame masses. Though not ideal, this data reasonably mimics a low redshift universe ($z \ll 1$) in which the linear Hubble relation of Eq. (3) holds, and galaxies follow a z^2 distribution. We use the same linear relation for the generation of the MDC universe (*i.e.* a set of simulated galaxy catalog parameters) for each of the MDCs. It should be emphasized that the Bayesian method for estimation of H_0 outlined in Section II above is general, and can be extended to realistic scenarios with a non-linear cosmology with $\{\Omega_{\rm m}, \Omega_{\rm k}, \Omega_\Lambda\}$ held fixed. So, in particular, the method is applicable for events which are detected at higher distances, where the low redshift approximation breaks down. The restriction to a linear cosmology in this paper comes only due to the use of the MDC dataset. We would like to note that by using a linear cosmology, we are not testing possible effects introduced by the presence of other cosmological parameters. The analysis at large redshifts may, for example, be sensitive to the values (or the assumed

prior ranges) of the parameters like Ω_m and Ω_Λ.

The first four columns of Table I summarize the characteristics of each of the galaxy catalogs created and how they correspond to each MDC. We give a brief description for each of the cases below.

A. MDC0: Known Associated Host Galaxies

MDC0 is the simplest version of the MDCs, in which we identify with certainty the host galaxy for each GW event, and is equivalent to the direct counterpart case. As the galaxies are generated with no redshift uncertainties or peculiar velocities, the results will be (very) optimistic. This MDC provides the "best possible" constraint on H_0 using the 249 events, which then allows for comparison with the other MDCs.

B. MDC1: Complete Galaxy Catalog

The MDC1 universe consists of the full set of 50,000 galaxies out to $z \approx 0.1$ ($d_L \approx 428\,\mathrm{Mpc}$) in the original First Two Years dataset. This gives a galaxy number density of ~ 1 per $7000\,\mathrm{Mpc}^3$, which is ~ 35 times sparse compared to the actual density of galaxies in the local universe [41]. Additional galaxies are generated beyond the edge of the dataset universe, uniformly across the sky and uniformly in comoving volume, thereby extending the universe out to a radius of 2000 Mpc ($z = 0.467$ for $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$). This means that, even allowing H_0 to be as large as $200\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, the edge of the MDC universe is more than twice the highest redshift associated with the farthest detection (which is at ~ 270 Mpc)[5]. Each of the 249 detected BNS have a unique associated host galaxy contained within the MDC1 catalog. This catalog is thus *complete* in the sense that it contains every galaxy in the simulated universe. We refer to the MDC universe as MDC1 throughout the rest of the paper, and similarly for the subsequent MDCs.

MDC1 is designed to test our treatment of GW selection effects, by ensuring that given a set of sources and access to a complete catalog, our methodology and analysis produces a result consistent with the simulated value of H_0.

C. MDC2: Incomplete Galaxy Catalog

MDC2 is designed to test our treatment of EM selection effects, by applying an apparent magnitude threshold to the MDC universe, such that a certain fraction of the host galaxies is not contained in it. This is a necessary consideration, given that we are in the regime where GW signals are being detected beyond the distance to which the current galaxy catalogs can be considered to be complete. This has been true for BBHs detections since O1, and is true of BNSs as well now we are in O3.

In order to create the catalog for MDC2, we start with the initial MDC1 universe and assign luminosities to each of the galaxies within it. We assume that the luminosity distribution of the galaxy catalog is known to the observer throughout and follows a Schechter function of the form (see *e.g.* [42])

$$\phi(L)\,dL = n^* \left(\frac{L}{L^*}\right)^{\alpha} e^{-L/L^*}\,\frac{dL}{L^*}\,, \tag{10}$$

where L denotes a given galaxy luminosity and $\phi(L)\,dL$ is the number of galaxies within the luminosity interval $[L, L + dL]$. The characteristic galaxy luminosity is given by $L^* = 1.2 \times 10^{10}\,h^{-2}\,L_\odot$ with solar luminosity $L_\odot = 3.828 \times 10^{26}\,\mathrm{W}$, and $h \equiv H_0/(100\mathrm{km\,s^{-1}\,Mpc^{-1}})$[6], $\alpha = -1.07$ characterizes the exponential drop off of the luminosity function, and n^* denotes the number density of objects in the MDC universe (in practice, this only acts as a normalization constant). The integral of the Schechter function diverges at $L \to 0$, requiring a lower luminosity cut off for the the dimmest galaxies in the universe which we set to $L_{\mathrm{lower}} = 0.001L^*$. This choice is arbitrary for our purpose here, but small enough to include almost all objects classified as galaxies in real catalogs like GLADE [41].

These luminosities are then converted to apparent magnitudes using $m \equiv 25 - 2.5\log_{10}(L/L^*) + 5\log_{10}(d_L/\mathrm{Mpc})$, and an apparent magnitude threshold m_{th} is applied as a very crude characterization of the selection function of an optical telescope observing only objects with $m < m_{\mathrm{th}}$. MDC2 is broken into three sub-MDCs, in order to test our ability to handle different levels of galaxy catalog completeness dictated by different telescope sensitivity thresholds. In each case, the catalog completeness is defined as the ratio of the number of galaxies inside the catalog relative to the number of galaxies inside the MDC universe, out to a reference fiducial distance d_L,

$$f_{\mathrm{completeness}}(d_L) = \frac{\sum_j^{\mathrm{MDC2}}(d_{L_j} < d_L)}{\sum_k^{\mathrm{MDC1}}(d_{L_k} < d_L)}\,, \tag{11}$$

where the numerator is a sum over the galaxies contained within the MDC2 catalog out to some reference distance d_L, and the denominator is a sum over the galaxies in the MDC1 catalog.

Apparent magnitude thresholds of $m_{\mathrm{th}} = 19.5$, 18, and 16 are chosen for the three sub-MDCs, which correspond to cumulative number completeness fractions of 75%, 50% and 25% respectively, evaluated at a distance of $d_L = 115\,\mathrm{Mpc}$, chosen such that given the luminosity distance distribution of detected BNSs, the completeness fraction for the sub-MDC to this distance is roughly indicative of the percentage of host galaxies which remain inside the galaxy catalog. The left panel of Fig. 1 shows how the completeness of each of the MDC2 catalogs drop off as a function of distance.

[5] For MDC1 and for all subsequent MDCs, it has been tested that the artificial "edge of the universe" has no bearing on the results.

[6] We note that the parameter L^* of the Schechter luminosity function itself depends on H_0, which we allow to vary and hence take into account within our formalism.

D. MDC3: Luminosity Weighting

MDC3 is designed to test the effect of weighting the likelihood of any galaxy being host to a GW event as a function of their luminosity. It is likely that the more luminous galaxies are also more likely hosts for compact binary mergers; the luminosity in blue (B-band) is indicative of a galaxy's star formation rate, for example, while the luminosity in high infrared (K-band) is a tracer of the stellar mass. The bulk of the host probability is expected to be contained within a smaller number of brighter galaxies, effectively reducing the number of galaxies which need to be considered. Additional information from luminosity is thus expected to improve the constraint on H_0 by narrowing its posterior probability density.

For MDC3, the probability of a galaxy emitting a GW signal is taken to be proportional to the galaxy's luminosity. As with MDC2, the luminosity distribution of the galaxies in the universe is assumed to follow the Schechter luminosity function as in Eq. (10) (referred to from now on as $p(L)$). However, the joint probability of a single galaxy having luminosity L and containing an emitting source, s, is

$$\begin{aligned} p(L, s) &= p(s|L)\, p(L) \\ &\propto L\, p(L)\,, \end{aligned} \tag{12}$$

where we assume that the probability of a galaxy of luminosity L hosting a source is proportional to the luminosity itself[7]. All host galaxies thus have luminosities sampled from $L\,p(L)$. In this context, we must consider all galaxies which emitted signals, not just those which were detected. With this in mind, the overall luminosity distribution has the following form:

$$p(L) = \beta \frac{L}{\langle L \rangle} p(L) + (1 - \beta)\, x(L) \tag{13}$$

where β is the fraction of emitting galaxies to total galaxies over the observed time period ($1 \geq \beta \geq 0$), $L/\langle L \rangle$ is the normalized luminosity, and $x(L)$ is the unknown luminosity distribution of galaxies which did not emit gravitational waves which we can sample for a given value of β.

Rearranging to obtain the only unknown, $x(L)$, gives

$$x(L) = \frac{p(L)}{1 - \beta} \left[1 - \beta \frac{L}{\langle L \rangle} \right], \tag{14}$$

and from this we see there is an additional constraint on β, because the term inside the brackets must be > 0. The maximum value that β can take is given by $\beta_{\rm max} = \langle L \rangle / L_{\rm max}$, where $L_{\rm max}$ is the maximum luminosity from the Schechter function, and $\langle L \rangle$ is the mean. From the Schechter function parameters detailed in section III C, $\beta_{\rm max} \approx 0.015$.

[7] Luminosity weighting is motivated by the expectation the brighter galaxies are more likely hosts. Here we assume a linear relation between luminosity of a galaxy and its probabiliy of hosting a source. Luminosities in the B- or K-bands, respectively indicative of galaxies' star formation rate or stellar mass, for example, may be expected to be approximately proportional to their probability of hosting compact binary sources.

We recall that the original First Two Years data was generated by simulating $\sim 50,000$ BNS events, of which ~ 500 were detected, of which 249 randomly selected detections underwent parameter estimation. The number of "emitting" and "non-emitting" galaxies have to be rescaled to represent this. Approximately half of the original events are chosen to be "emitters", including the 249 detected ones. With this in mind, luminosities can not be assigned to their host galaxies without adding a greater density of non-emitting galaxies in order to satisfy the requirements for β. Thus for MDC3, the density of galaxies is increased by a factor of 100, with the acknowledgement that this would lead to a broadening of the final posterior. MDC3 has a galaxy density of ~ 1 galaxy per $70\,{\rm Mpc}^3$, which is about 3 times denser than the actual density of galaxies in the local universe [41]. This also means that MDC3 is not directly comparable with the previous MDC versions, save MDC0. The galaxies which are considered "emitters" are assigned luminosities from $Lp(L)$, and "non-emitters" from $x(L)$ above.

In order to include EM selection effects, an apparent magnitude cut $m_{\rm th}$ of 14 is applied, such that the completeness of the galaxy catalog is $\sim 50\%$ out to the same fiducial distance of 115 Mpc as in MDC2. In this case, completeness is however defined in terms of the fractional luminosity contained in the catalog, rather than in terms of numbers of objects:

$$f_{\rm completeness}(d_L) = \frac{\sum_j^{\rm MDC3} L_j(d_{L_j} < d_L)}{\sum_k^{\rm complete} L_k(d_{L_k} < d_L)}\,, \tag{15}$$

where the numerator is summed over the galaxies inside the MDC3 apparent magnitude-limited catalog, and the denominator is summed over the galaxies in the whole MDC3 universe. This is shown in the right panel of Fig. 1. As the emitting galaxies are luminosity weighted, the cumulative luminosity completeness is representative of the percentage of BNS event hosts inside the catalog.

IV. RESULTS

In this section we summarize the results for the mock data challenges described in Section III. We show the combined posteriors on H_0 for each MDC, discuss the convergence to the simulated value of $H_0 = 70$ km s^{-1} Mpc^{-1}and calculate the precision of the combined measurement under each set of conditions. In Table I we list the measured values of the Hubble constant for the combined 249 event posterior (maximum a-posteriori and 68.3% highest density posterior intervals) all computed with a uniform prior on H_0 in the range of $[20, 200]$ km s^{-1} Mpc^{-1}, as well as the corresponding fractional uncertainties for each of the MDCs.

A. MDC0: Known Associated Host Galaxies

We first consider the simple case where we identify the true host galaxy for every event and determine the resulting 249-event combined H_0 posterior. Fig. 2 presents the

MDC	Host galaxy preference	Completeness[a]	$m_{\rm th}$	Analysis assumption	H_0 (km s^{-1} Mpc^{-1})	Fractional uncertainty
0	Known host	-	-	direct counterpart	$69.08^{+0.79}_{-0.80}$	1.13%
1	equal weights	100%	-	unweighted catalog	$68.91^{+1.36}_{-1.22}$	1.84%
2a	equal weights	75%	19.5	unweighted catalog	$69.69^{+1.66}_{-1.44}$	2.21%
2b	equal weights	50%	18	unweighted catalog	$69.76^{+1.79}_{-1.65}$	2.46%
2c	equal weights	25%	16	unweighted catalog	$69.64^{+2.44}_{-2.10}$	3.24%
3a	luminosity weighted	50%	14	weighted catalog	$70.38^{+3.49}_{-2.64}$	4.38%
3b	luminosity weighted	50%	14	unweighted catalog	$68.95^{+4.41}_{-3.54}$	5.68%

[a] The completeness is calculated as a number completeness using Eq. (11) for MDCs 1 and 2, and as a luminosity completeness using Eq. (15) for MDC 3, out to a fiducial distance of 115 Mpc, such that it is indicative of the fraction of host galaxies which are inside the galaxy catalog in both cases.

TABLE I. A summary of the main results. We quote the peak value and the 68.3% highest density error region for the posterior probability on H_0 for each of the MDCs combining all 249 events. The fractional uncertainty is defined as the half-width of the 68.3% highest density probability interval divided by the simulated value of $H_0 = 70$ km s^{-1} Mpc^{-1}.

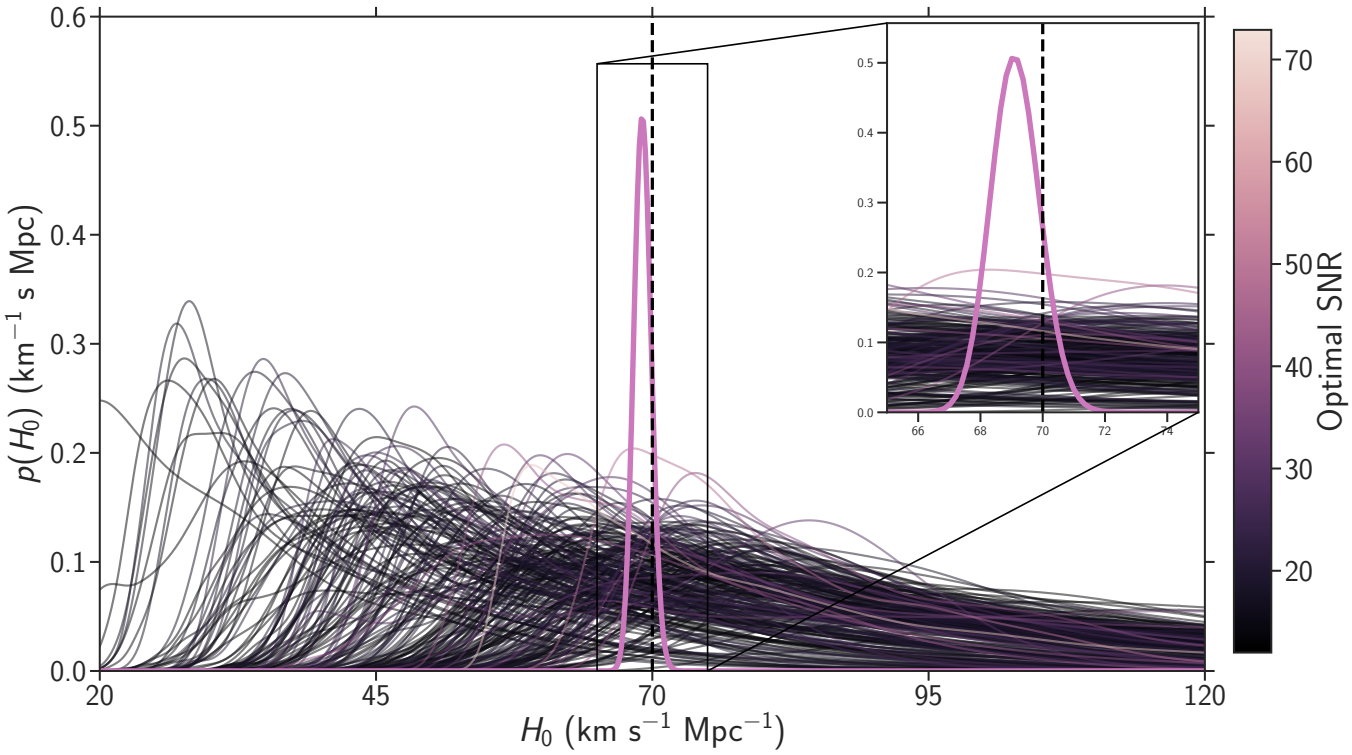

FIG. 2. Individual and combined results for MDC0 (known host galaxy or direct counterpart case). The thick purple line shows the combined posterior probability density on H_0. Individual likelihoods (scaled by an arbitrary value), for each of the 249 events, are shown as thin lines with shades corresponding to their optimal SNR. The simulated value of H_0 is shown as a vertical dashed line.

results of this analysis. The likelihoods for each individual GW event are shown (normalized relative to each other but scaled w.r.t. the combined posterior for clarity) shaded by the event's optimal SNR. In this case, the likelihood is informative, having a clearly-defined peak corresponding to finding the likely values of H_0 for the known galaxy redshift. Each curve traces the information in the corresponding d_L distribution, which is usually unimodal, but in some cases may have two or more peaks [36, 37]. We see that the peaks of the individual likelihoods do not necessarily correspond to the true value $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, but there is always support for it. Therefore it is the only consistent value for all events, driving the combined posterior, which is overlaid in thick purple. This gives us a statistical estimate for the maximum a-posteriori value and 68.3% maximum-density credible interval for H_0 as $69.08^{+0.79}_{-0.80}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. The final result

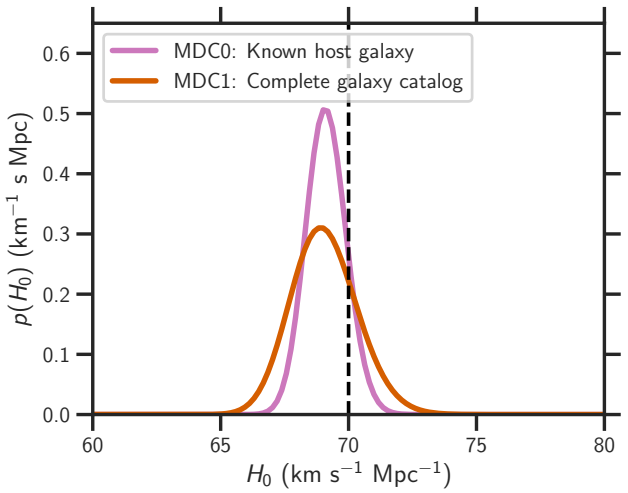

FIG. 3. Comparison of the galaxy catalog method with the known host galaxy case. Joint posterior probability density on H_0 using all 249 events for MDC0 (known host galaxy) and MDC1 (complete galaxy catalog) are shown respectively in purple and red. For this set of simulations, uncertainty with the galaxy catalog is only about 1.63 times larger than with known host galaxies.

combining all the 249 events have converged to a relatively symmetric "Gaussian" distribution.

The result of MDC0 provides us with the best possible H_0 estimate given the set of GW detections, since this case corresponds to perfect knowledge of the host galaxies. This gives us a good benchmark against which other versions of the MDC can be compared. Since this is a best-case scenario, we have the least statistical uncertainty in the final result, making any systematic bias more apparent than for the subsequent MDCs. For the combined result with 249 events, the simulated value is contained within the support of the posterior distribution of H_0. Assuming that the GW selection function is known with sufficient precision, this demonstrates that the H_0 measurement will be unbiased with the employed formalism [6, 9, 18, 43].

B. MDC1: Complete Galaxy Catalog

The next more complex case is MDC1, where we assume no counterpart was observed, and resort to using a galaxy catalog. MDC1 uses a *complete* galaxy catalog containing all potential hosts, and is an optimistic scenario. Moreover EM selection effects are not present yet. The results with MDC1 already show a wider posterior distribution on H_0 ($68.91^{+1.36}_{-1.22}$ km s^{-1} Mpc^{-1}) because of lack of certainty of the host galaxy (Fig. 3). The introduction of this uncertainty means that the contributions from each event will be smoothed out, depending on the size of the event's sky localization and the number of galaxies within it. As can be seen in Fig. 4, there is a far higher proportion of events for which the likelihood is relatively broad and less informative, in comparison to Fig. 2. However, many events clearly have a small number of galaxies in their sky-area, and hence still show clear peaks.

MDC0 and MDC1 demonstrate the importance of correctly accounting for GW selection effects. We are biased towards detecting sources which are nearer-by, and which are optimally orientated (closer to face-on). If an analysis is performed without taking into consideration the denominator $p(D_{\rm GW}|H_0)$ of Eq. (6), that corrects for the above GW selection effects, the posterior density on H_0 converges to a value different from its simulated value of 70 km s^{-1} Mpc^{-1}.

C. MDC2: Incomplete Galaxy Catalog

The next most complex scenario is the case where we have incomplete galaxy catalogs, limited by an apparent magnitude threshold. This gives us the first case where accounting for EM selection effects is important. To investigate this, we consider three galaxy catalogs, with apparent magnitude thresholds of $m_{\rm th} = 19.5$, 18 and 16, with respective completeness fractions of 75%, 50% and 25% in addition to the complete catalog for MDC1 (see III C for details). The combined 249-event posterior distributions on H_0 are shown in Fig. 5.

As the catalogs become less complete, the combined H_0 posterior becomes wider. This is because the probability that the host galaxy is inside the catalog decreases. The contribution from the galaxies within the catalog is reduced, and the uninformative contribution from the out-of-catalog term in Eq. (8) increases. This is visible in the individual likelihoods shown in Fig. 6, where instead of decreasing toward zero at high values of H_0, the individual likelihoods tend toward a constant. This is because, in the absence of EM data, and with the linear Hubble relation assumed in this work, the number of unobserved galaxies increases without limit as d_L^2.

We estimate $H_0 = 69.69^{+1.66}_{-1.44}$, $69.76^{+1.79}_{-1.65}$, and $69.64^{+2.44}_{-2.10}$ km s^{-1} Mpc^{-1} respectively for galaxy catalogs of 75%, 50%, and 25% completeness. See section IV E for a more in depth comparison of how galaxy catalog completeness affects posterior width.

Our exercise demonstrates that we need to know (or assess) the completeness of galaxy catalogs, and put in an appropriate out-of-catalog term in the analysis. For any of the MDC2 catalogs, if we assume that the galaxy catalog is complete, when in reality it is not, we get a posterior distribution on H_0 which is inconsistent with a value of 70 km s^{-1} Mpc^{-1}. This is because the well-localized events for which the host is not inside the catalog do not have support for the correct value of H_0. In real catalogs, galaxy clustering might ensure that there are nearby bright galaxies in the catalog, partially mitigating this bias.

D. MDC3: Luminosity Weighting

Until now we have considered all galaxies in our catalog to be equally likely to host a gravitational-wave source. In MDC3 we analyze the case where this is no longer true by constructing a galaxy catalog such that the probability of any

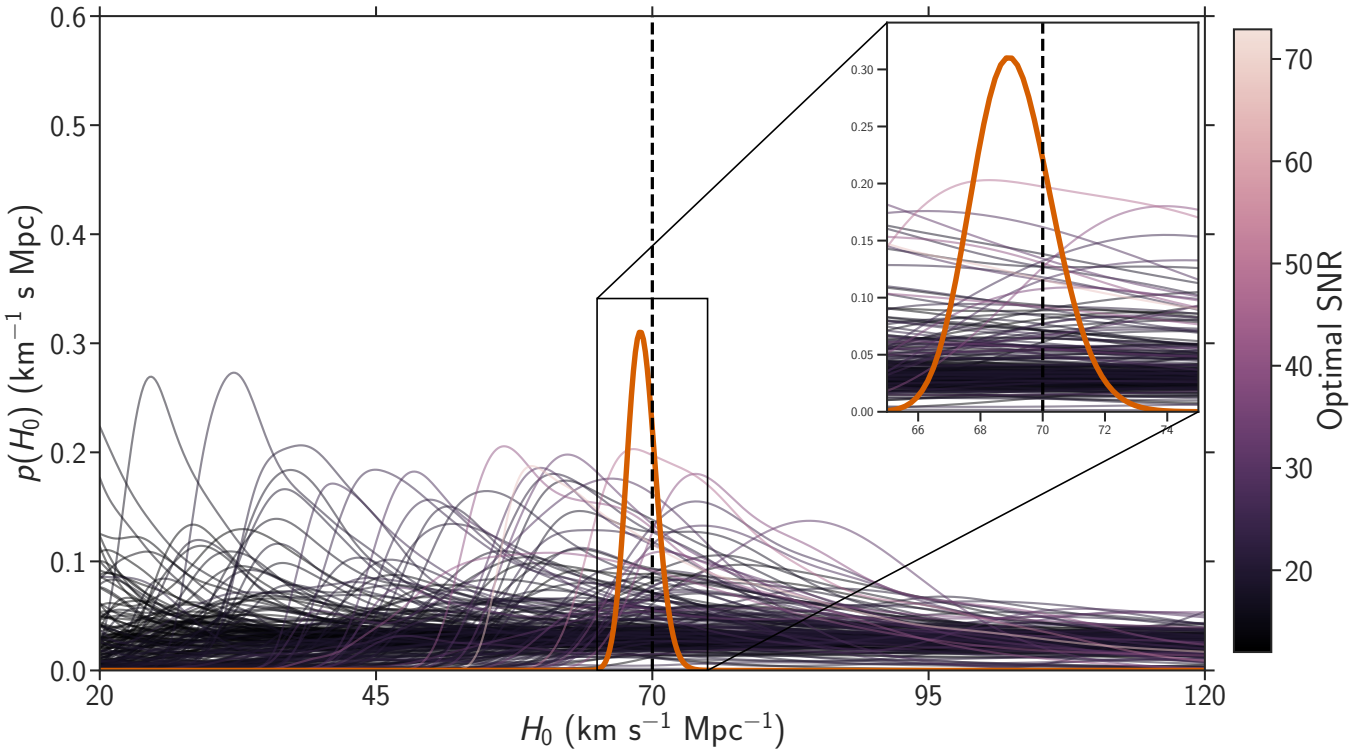

FIG. 4. Individual and combined results for MDC1 (complete galaxy catalog). The thick red line shows the combined posterior probability density on H_0. Individual likelihoods (scaled by an arbitrary value), for each of the 249 events, are shown as thin lines with shades corresponding to their optimal SNR. The simulated value of H_0 is shown as a vertical dashed line. Many of the individual likelihoods do not have sharp features, however the final result converges to the simulated value with redshift information present in the galaxy catalogs. This demonstrates the applicability of our methodology.

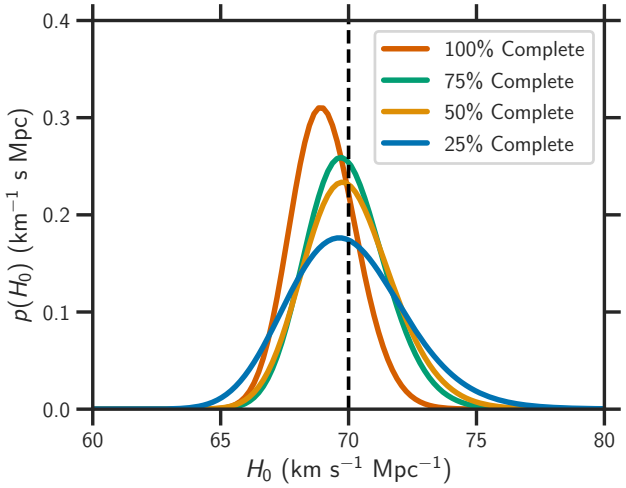

FIG. 5. Comparison of results with varying galaxy catalog completeness. In MDC2, the simulated apparent magnitude threshold is varied to obtain galaxy catalogs of 100%, 75%, 50%, and 25% completeness. The corresponding posterior probability densities on H_0 using all 249 events are shown in red, green, yellow, and blue respectively.

single galaxy hosting a GW source is directly proportional to its luminosity. This models the case in which GW events trace the luminosity of galaxies, commonly expected to be true in the real universe, as luminosity can be a tracer for star formation rate (B-band) or galaxy mass (K-band), both of which would be correlated with higher rates of GW mergers. MDC3 includes the same EM selection effects as MDC2, in the sense that the catalog is magnitude limited. The completeness of this catalog, defined in terms of luminosity rather than numbers of galaxies, as defined in Eq. (15), is 50% out to 115 Mpc. This is indicative that approximately 50% of the detected GW events have host galaxies inside the catalog.

To investigate the importance of luminosity weighting, MDC3 was analyzed twice under different assumptions, given in Eq. (A.3). In the first, the analysis was matched to the known properties of the galaxy catalog, such that the probability of any galaxy hosting a GW event was proportional to its luminosity. In the second, we feigned ignorance and ran the analysis with the assumption that each galaxy was equally likely to be host to a GW event (as was true in MDCs 1 and 2). This allows us to determine the effect of ignoring galaxy weighting with this dataset. The combined H_0 posteriors for both cases are shown in Fig. 7. The estimated values of the Hubble constant are $70.38^{+3.49}_{-2.64}$ km s^{-1} Mpc^{-1}(assuming hosts are luminos-

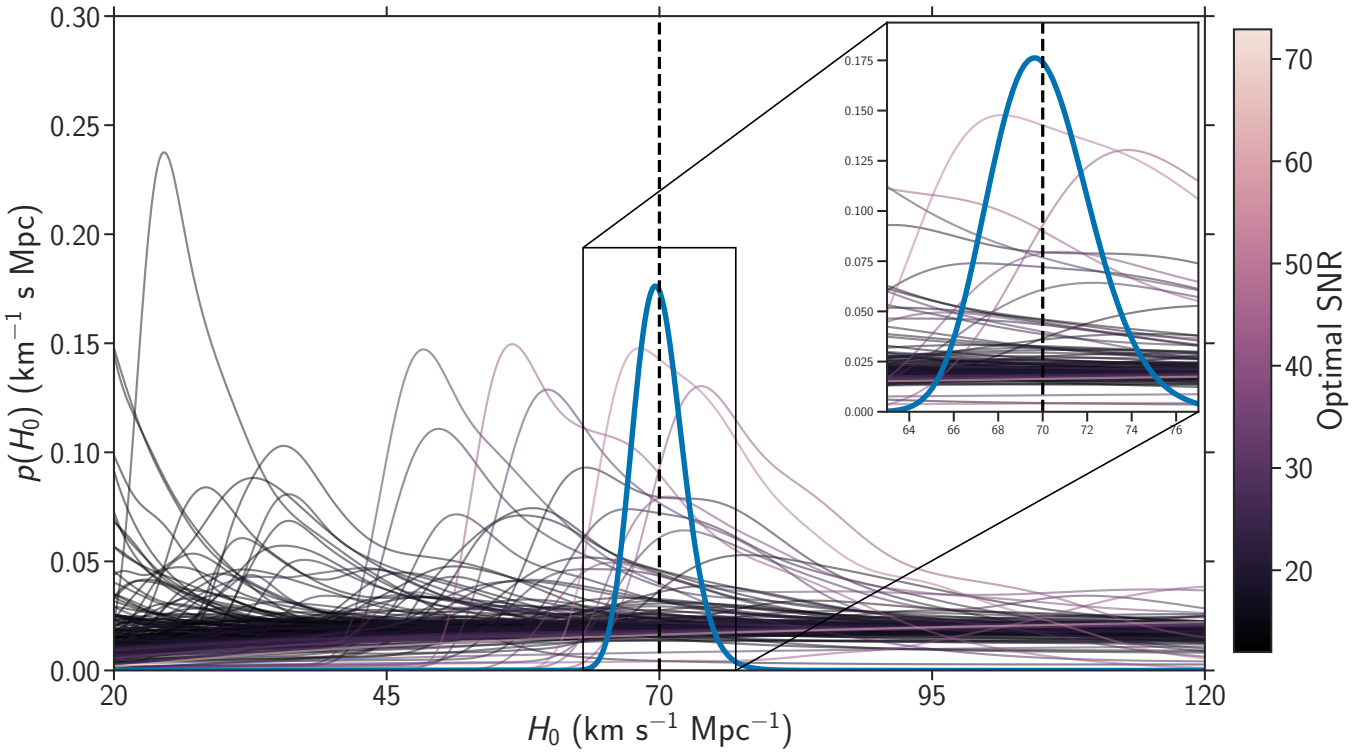

FIG. 6. Individual and combined results for MDC2 with a 25% complete galaxy catalog. The thick blue line shows the combined posterior probability density on H_0. Individual likelihoods (scaled by an arbitrary value), for each of the 249 events, are shown as thin lines with shades corresponding to their optimal SNR. The simulated value of H_0 is shown as a vertical dashed line. Compared to MDC0 (Fig. 2) and MDC1 (Fig. 4), fewer individual likelihoods are peaked here. Although the final H_0 estimate is less precise, the results converge to the simulated value, demonstrating the applicability of our methodology to threshold-limited galaxy catalogs of about 25% completeness.

ity weighted) and $68.95^{+4.41}_{-3.54}$ km s^{-1} Mpc^{-1}(assuming equal weights). By weighting the host galaxies with the correct function of their luminosies, which happens to be known in this case, the constraint on H_0 improves — the uncertainty narrows by a factor of 1.3, compared to the case in which equal weights are assumed. Both results are consistent with the fiducial H_0 value of 70 km s^{-1} Mpc^{-1}. In the limit of a far greater number of events, one might expect to see a bias emerge in the case in which the assumptions in the analysis do not match those with which the catalog was simulated. The luminosity weighting of host galaxies, by its very nature, increases the probability that the host galaxy is inside the galaxy catalog; assuming equal weighting gives disproportionate weight to the contribution that comes from beyond the galaxy catalog. However, for the 249 BNS events considered here, the final posteriors are too broad to be able to detect any kind of bias.

E. Comparison between the MDCs

So far we have focused on individual event likelihoods and combined results for all 249 events. Our data-set also allows us to to assess the convergence for the combined Hubble posterior as we add events. We calculate the intermediate combined posteriors as a function of the number of events, and show the resulting convergence in Fig. 8. We plot the fractional H_0 uncertainty (defined here as the half-width of the 68.3% credible interval divided by H_0, $\Delta^{68.3\%}_{H0}/2H_0$), against the number of events we include in a randomly-selected group. The scatter between realizations of the group is indicated by the error bars, which encompass 68.3% of their range. There is a considerable variation between different realizations, for the incomplete catalogs. For example, of the 100 realizations we used, for 25% completeness and 40 events, there are groups that give $\sim$ 10% precision, but others that give $\sim$ 70% precision.

With a sufficiently large number of events, we expect a $1/\sqrt{N}$ scaling of the uncertainty with the number of events [5, 6]. To check whether this behavior is indeed true, we fit the results for each MDC to the expected scaling, obtaining the coefficient of $1/\sqrt{N}$ by maximizing its likelihood given the fractional uncertainties and their variances from the different realizations. The coefficient of the scaling is automaticlly dominated by the fractional uncertainties at large N where the variances are small. We show this scaling for each MDC as a set of dashed lines in Fig. 8.

It can be seen that for each MDC, the results converge to the expected $1/\sqrt{N}$ scaling. The number of events required before this behavior is reached is dependent on the amount of EM information available on average for each event, in agree-

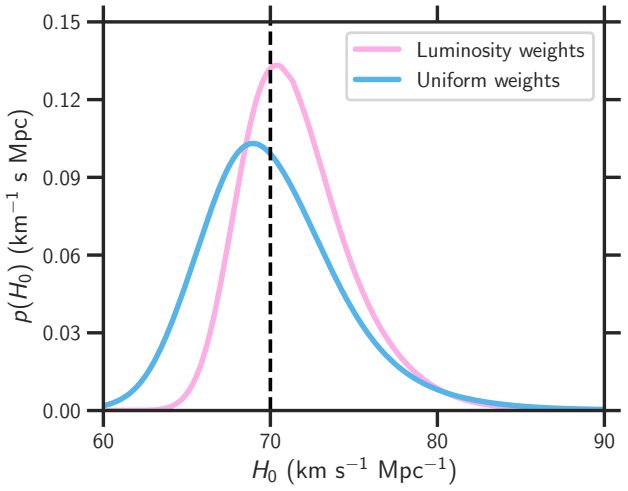

FIG. 7. Comparison of results with and without luminosity weighting. In MDC3, by construction, the probability of any galaxy hosting a GW event is proportional to its luminosity. The pink curve shows the posterior probability density on H_0 for the case where we take this into account in our analysis as a weighting by the galaxy's luminosity. The blue curve shows the posterior density for the case where we ignore this extra information, and treat every galaxy as equally likely to be hosts. Luminosity weighting improves the precision in the results by a factor of 1.3 for this set of simulations.

ment with the results of [6]. The direct counterpart case is always on the trend after $O(10)$ events, and shows a $\sim 18\%/\sqrt{N}$ convergence, comparable to and consistent with the results in [6, 7]. With the most complete galaxy catalogs, if the host galaxy is not directly identified it will take tens of events before this behavior is reached. However, even the least complete catalog (25%) appears to have reached this behavior by the time all 249 events are combined. It should be noted that as the catalogs for MDCs 1 and 2 were not simulated realistically, their low density relative to the density of the universe means that these numbers should not be taken as predictions of how fast $1/\sqrt{N}$ may be reached (except, perhaps, in the counterpart case, although on should bear in mind that even for that case, peculiar velocities and redshift uncertainties have been neglected). Even with a galaxy catalog which is 25% complete, MDC2 gives a result which is only about a factor of 3 times worse than the counterpart case.

As the density of galaxies in MDC3 was increased by 2 orders of magnitude over MDCs 1 and 2, the final posteriors cannot be directly compared between MDCs. However, by plotting the equivalent convergence figure for MDC3 (including the "known host" case as a reference, see Fig. 8), the impact of increasing the density of galaxies in the universe on the rate at which the posterior converges on the $1/\sqrt{N}$ behavior becomes clear. When there are more host galaxies, the results are overall less precise, and take longer to reach the $1/\sqrt{N}$ trend. As expected, using luminosity-weighting of potential host galaxies as an assumption in the analysis concentrates the probability to a smaller number of galaxies, leading to a more precise result.

F. Limited Robustness Studies

Our results are expected to be sensitive to the luminosity distribution parameters — if one uses values of the Schechter function parameters α and L^* in the analysis which are different from the ones used to simulate the galaxy catalogs, one would expect to end up with a bias in the results. With variations of these parameters within their current measurement uncertainties, we have however demonstrated that the resulting variation in the final result is small compared to the statistical uncertainties reached with the current set of MDCs. Furthermore we have also demonstated that our results are robust against a small $O(1)$ variation in the value of the telescope sensitivity threshold $m_{\rm th}$.

V. CONCLUSIONS AND OUTLOOK

The H_0 measurement using GW standard sirens has been demonstrated with recent events, including both the counterpart method for GW170817 [18], and the galaxy catalog method [21, 22]. These approaches are combined in the analysis of both BNS and BBH events from the first two observing runs of the advanced detector network [23], using the method described in this paper. Future measurements will rely on a combination of counterpart and catalog methods, as appropriate for each new detected event, with catalog incompleteness playing an important role for the more distant, yet more common, BBHs. This paper outlines a coherent approach that tackles both of these scenarios, including treatment of selection effects in both GWs (due to the limited sensitivity of GW detectors) and EM (due to the flux-limitations of EM observing channels). We performed a series of MDCs to validate our method using up to 249 observed events. For each of the MDCs analyzed, the final posterior on H_0 is found to be consistent with the value of $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ used to simulate the MDC galaxy catalogs, demonstrating that our method can produce sufficiently unbiased results for treating these numbers of events, in our simulations.

GW selection effects are inherent in every version of the MDC and were corrected for by the term $p(D_{\rm GW}|H_0)$ in the denominator of Eq. (6). EM selection effects are addressed in MDCs 2 and 3 by the out-of-catalog terms containing $\bar{G}$ in Eq. (8). In both these MDCs, in spite of having an apparent magnitude-limited galaxy catalog, we are able to accurately infer H_0 without any bias. MDC2 further demonstrates our ability to account for missing host galaxies down to a level where only 25% of events have hosts inside the catalog. Even in this case, we converge to the correct H_0 value, to the level of precision which could be reached by 249 events.

MDC3 demonstrates a clear tightening of the posterior distribution when we can assume that GW events trace the galaxy luminosities, compared to the case in which we treat all galaxies as equally likely hosts. The "uniform weights" analysis of MDC3 remains consistent with the simulated H_0 value. Hence

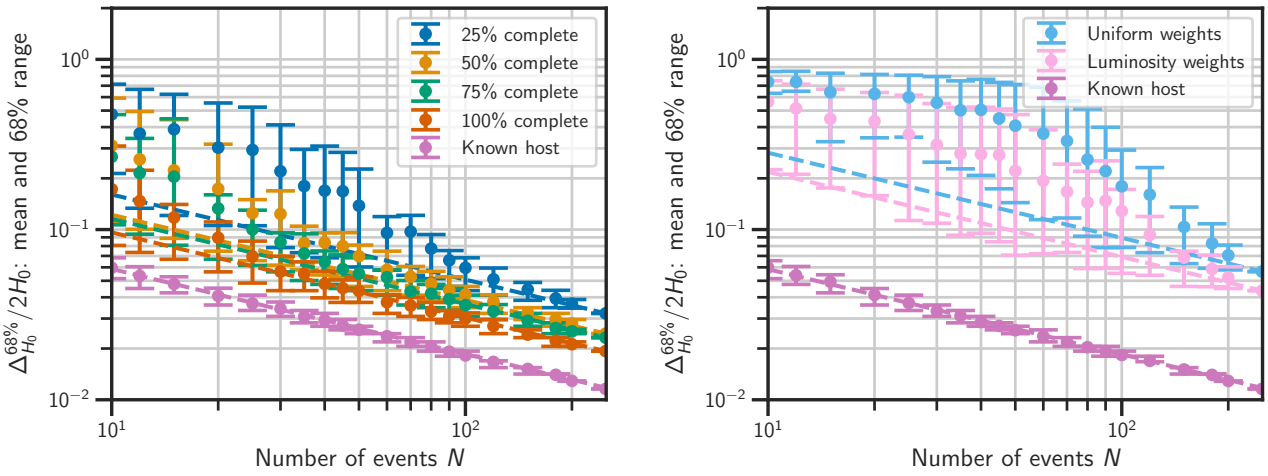

FIG. 8. Fractional uncertainty in H_0 as a function of the number N of the events for the combined H_0 posteriors. The fractional uncertainty in H_0 is defined as the half-width of the 68.3% highest probability interval divided by 70 km s^{-1} Mpc^{-1}, and is shown as the plotted dots for all cases. The error bars contain 68% of the scatter arising from different realisations of the events. (left) In purple, red, green, yellow and blue we show the associated host galaxy case (MDC0), complete galaxy catalog (MDC1) case, and the 75%, 50% and 25% completeness cases; we find a fractional H_0 uncertainty of 1.13%, 1.84%, 2.21%, 2.46% and 3.24% respectively for the combined H_0 posterior from 249 events. (right) convergence for MDC3 (event probability proportional to galaxy luminosity), analyzed with luminosity-weighted likelihood (pink) or equally-weighted likelihood (light blue). We find fractional H_0 uncertainties of 4.38% and 5.68% respectively. MDC0 (purple) is included for reference. We plot the expected $1/\sqrt{N}$ scaling behavior for large values of N for all cases with the dashed lines. This scaling behavior is met by all MDCs as the number of events reaches 249, but for the less informative, lower completeness MDCs the trend is slower to emerge. This is even more evident in MDC3, where the density of galaxies is 100 times greater, producing more potential hosts for each event. This is mitigated somewhat by the effect of luminosity-weighting the potential hosts (pink).

we are unable to conclude whether an incorrect assumption would lead to a biased result, as one might expect. We used only 249 events for our MDCs. With enough events of comparable nature the bias would be detected. Future work will expand these studies to include a larger numbers of simulated GW events, and will be able to discern smaller sources of systematic effects.

Although the galaxy-catalog standard-siren measurement of H_0 is less precise than the counterpart measurement, it is still able to constrain H_0, but requires at least an order of magnitude more events in order to reach a comparable accuracy (in the most realistic case of MDC3). These MDCs have validated our method and implementation in simplified scenarios. However future work will be needed to improve on this in several directions, to test its applicability to BBHs (which are detectable out to much farther distances), realistic cosmology, and real galaxy catalogs [6, 23].

In both the counterpart and galaxy catalog cases, the lack of redshift uncertainties and peculiar velocities implies that the contributions from individual galaxies are a lot more precise than they would be in reality. Moreover, the simulated catalogs in MDCs 1 and 2 have a low density of galaxies compared to the universe, making them more informative than real catalogs. MDC3, with a galaxy density of 1 galaxy per 70 Mpc^3, comes closest to the actual density of galaxies in the local universe of $\sim$ 1 galaxy per 200 Mpc^3 [41]. In this scenario there is still a clear convergence towards the simulated H_0 value. In comparison to actual catalogs such as GLADE [41], the apparent magnitude threshold of 14 is very low, so we expect a real catalog-only analysis to fall somewhere between MDCs 2 and 3. We caution the reader that with tens of events, the precision of results can vary by almost an order of magnitude depending on the particular realization of the detected population, before eventually converging to the expected $1/\sqrt{N}$ behaviour [5, 6]. Analyzing more realistic catalogs will also require a sky-varying EM selection function, as the magnitude threshold varies significantly on the sky according to the design of particular surveys.

The galaxy distribution in these simulated catalogs is uniform in comoving volume. Although it has not been studied here, clustering of galaxies is expected to improve the constraint on H_0 (see, *e.g.* [6, 44]), since even when the host is not in the catalog, it is likely that there will be observed galaxies nearby.

The Advanced LIGO - Virgo second observing run [20] has confirmed that BBH systems are detected at higher rates than BNSs. Since their greater mass allows them to be observed at much greater distances, where galaxy catalogs are incomplete, the catalog method including EM selection effects is particularly important. With the catalog of GW events expected to expand at an increasing rate in future observing runs, our analysis will evolve to meet the challenges that come with it, and give us the fullest picture of cosmology as revealed by gravitational waves.

ACKNOWLEDGMENTS

We thank members of the LIGO-Virgo Collaboration for valuable discussions pertaining to the writing of this paper, and in particular Nicola Tamanini for a careful reading of the manuscript. AG additionally thanks P. Ajith, Walter Del Pozzo, Anuradha Samajdar, and Chris Van Den Broeck for discussion at various stages of the work. RG, CM and JV are supported by the Science and Technology Research Council (grant No.ST/L000946/1). IMH is supported by the NSF Graduate Research Fellowship Program under grant DGE-17247915. IMH also acknowledges support from NSF Award PHY-1607585. HQ is supported by Science and Technology Facilities Council (grant No.ST/T000147/1). AS thanks Nikhef for its hospitality and support from the Amsterdam Excellence Scholarship (2016-2018). HYC was supported by the Black Hole Initiative at Harvard University, through a grant from the John Templeton Foundation. MF and DEH were supported by NSF grant PHY-1708081. They were also supported by the Kavli Institute for Cosmological Physics at the University of Chicago through an endowment from the Kavli Foundation. AG is supported by the research programme of the Netherlands Organisation for Scientific Research (NWO). DEH gratefully acknowledges support from the Marion and Stuart Rice Award. We are grateful for computational resources provided by the Leonard E Parker Center for Gravitation, Cosmology and Astrophysics at the University of Wisconsin-Milwaukee, and those provided by Cardiff University, and funded by an STFC grant supporting UK Involvement in the Operation of Advanced LIGO. This article has been assigned LIGO document number LIGO-P1900017.

[1] B. F. Schutz, Nature (London) **323**, 310 (1986).

[2] D. E. Holz and S. A. Hughes, Astrophys. J. **629**, 15 (2005), arXiv:astro-ph/0504616 [astro-ph].

[3] N. Dalal, D. E. Holz, S. A. Hughes, and B. Jain, Phys. Rev. **D74**, 063006 (2006), arXiv:astro-ph/0601275 [astro-ph].

[4] S. Nissanke, D. E. Holz, S. A. Hughes, N. Dalal, and J. L. Sievers, Astrophys. J. **725**, 496 (2010), arXiv:0904.1017 [astro-ph.CO].

[5] S. Nissanke, D. E. Holz, N. Dalal, S. A. Hughes, J. L. Sievers, and C. M. Hirata, (2013), arXiv:1307.2638 [astro-ph.CO].

[6] H.-Y. Chen, M. Fishbach, and D. E. Holz, Nature **562**, 545 (2018), arXiv:1712.06531 [astro-ph.CO].

[7] S. M. Feeney, H. V. Peiris, A. R. Williamson, S. M. Nissanke, D. J. Mortlock, J. Alsing, and D. Scolnic, Phys. Rev. Lett. **122**, 061105 (2019), arXiv:1802.03404 [astro-ph.CO].

[8] E. Di Valentino, D. E. Holz, A. Melchiorri, and F. Renzi, Phys. Rev. **D98**, 083523 (2018), arXiv:1806.07463 [astro-ph.CO].

[9] D. J. Mortlock, S. M. Feeney, H. V. Peiris, A. R. Williamson, and S. M. Nissanke, (2018), arXiv:1811.11723 [astro-ph.CO].

[10] N. Aghanim *et al.* (Planck), (2018), arXiv:1807.06209 [astro-ph.CO].

[11] A. G. Riess, S. Casertano, W. Yuan, L. M. Macri, and D. Scolnic, Astrophys. J. **876**, 85 (2019), arXiv:1903.07603 [astro-ph.CO].

[12] E. Macaulay *et al.* (DES), Mon. Not. Roy. Astron. Soc. **486**, 2184 (2019), arXiv:1811.02376 [astro-ph.CO].

[13] S. Birrer *et al.*, Mon. Not. Roy. Astron. Soc. **484**, 4726 (2019), arXiv:1809.01274 [astro-ph.CO].

[14] W. L. Freedman *et al.*, (2019), arXiv:1907.05922 [astro-ph.CO].

[15] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, and et al. (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **119**, 161101 (2017).

[16] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, and et al., ApJ **848**, L12 (2017).

[17] M. Soares-Santos, D. E. Holz, J. Annis, R. Chornock, K. Herner, Berger, and et al., ApJL **848**, L16 (2017).

[18] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, and et al. (LIGO Scientific Collaboration and Virgo Collaboration), Nature (London) **551**, 85 (2017), arXiv:1710.05835.

[19] W. Del Pozzo, Phys. Rev. D **86**, 043011 (2012), arXiv:1108.1317.

[20] B. P. Abbott *et al.* (LIGO Scientific, Virgo), (2018), arXiv:1811.12907 [astro-ph.HE].

[21] M. Fishbach *et al.* (LIGO Scientific, Virgo), Astrophys. J. **871**, L13 (2019), arXiv:1807.05667 [astro-ph.CO].

[22] M. Soares-Santos *et al.* (DES, LIGO Scientific, Virgo), Astrophys. J. **876**, L7 (2019), arXiv:1901.01540 [astro-ph.CO].

[23] B. P. Abbott *et al.* (LIGO Scientific, Virgo), (2019), arXiv:1908.06060 [astro-ph.CO].

[24] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, and R. X. Adhikari, Living Reviews in Relativity **21**, 3 (2018), arXiv:1304.0670 [gr-qc].

[25] D. Castelvecchi, Nature (London) **565**, 9 (2019).

[26] T. Padma, Nature (London) (2019), 10.1038/d41586-019-00184-z.

[27] B. P. Abbott *et al.* (LIGO Scientific, Virgo), (2018), arXiv:1811.12940 [astro-ph.HE].

[28] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. X. Adhikari, V. B. Adya, and C. Affeldt, Astrophys. J. **875**, 161 (2019), arXiv:1901.03310 [astro-ph.HE].

[29] R. Nair, S. Bose, and T. D. Saini, Phys. Rev. **D98**, 023502 (2018), arXiv:1804.06085 [astro-ph.CO].

[30] D. W. Hogg, (1999), arXiv:astro-ph/9905116 [astro-ph].

[31] B. S. Sathyaprakash and B. F. Schutz, Living Rev. Rel. **12**, 2 (2009), arXiv:0903.0338 [gr-qc].

[32] S. R. Taylor, J. R. Gair, and I. Mandel, Phys. Rev. **D85**, 023535 (2012), arXiv:1108.5161 [gr-qc].

[33] W. M. Farr, M. Fishbach, J. Ye, and D. Holz, Astrophys. J. Lett. **883**, L42 (2019), arXiv:1908.09084 [astro-ph.CO].

[34] C. Messenger and J. Read, Phys. Rev. Lett. **108**, 091101 (2012), arXiv:1107.5725 [gr-qc].

[35] M. Fishbach, D. E. Holz, and W. M. Farr, Astrophys. J. **863**, L41 (2018), [Astrophys. J. Lett.863,L41(2018)], arXiv:1805.10270 [astro-ph.HE].

[36] L. P. Singer, L. R. Price, B. Farr, A. L. Urban, C. Pankow, S. Vitale, J. Veitch, W. M. Farr, C. Hanna, K. Cannon, T. Downes, P. Graff, C.-J. Haster, I. Mandel, T. Sidery, and A. Vecchio, Astrophys. J. **795**, 105 (2014), arXiv:1404.5623 [astro-ph.HE].

[37] C. P. L. Berry *et al.*, Astrophys. J. **804**, 114 (2015), arXiv:1411.6934 [astro-ph.HE].
[38] C. Messick, K. Blackburn, P. Brady, P. Brockill, K. Cannon, R. Cariou, S. Caudill, S. J. Chamberlin, J. D. E. Creighton, R. Everett, C. Hanna, D. Keppel, R. N. Lang, T. G. F. Li, D. Meacher, A. Nielsen, C. Pankow, S. Privitera, H. Qi, S. Sachdev, L. Sadeghian, L. Singer, E. G. Thomas, L. Wade, M. Wade, A. Weinstein, and K. Wiesner, Phys. Rev. D **95**, 042001 (2017), arXiv:1604.04324 [astro-ph.IM].
[39] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin, W. Del Pozzo, F. Feroz, J. Gair, C.-J. Haster, V. Kalogera, T. Littenberg, I. Mandel, R. O'Shaughnessy, M. Pitkin, C. Rodriguez, C. Röver, T. Sidery, R. Smith, M. Van Der Sluys, A. Vecchio, W. Vousden, and L. Wade, Phys. Rev. D **91**, 042003 (2015), arXiv:1409.7215 [gr-qc].
[40] W. Del Pozzo, C. P. Berry, A. Ghosh, T. S. F. Haines, L. P. Singer, and A. Vecchio, Mon. Not. Roy. Astron. Soc. **479**, 601 (2018), arXiv:1801.08009 [astro-ph.IM].
[41] G. Dálya, G. Galgóczi, L. Dobos, Z. Frei, I. S. Heng, R. Macas, C. Messenger, P. Raffai, and R. S. de Souza, Mon. Not. Roy. Astron. Soc. **479**, 2374 (2018), arXiv:1804.05709 [astro-ph.HE].
[42] J. Binney and S. Tremaine, "Galactic dynamics," (Princeton University Press, 1987) Chap. 1, pp. 21–22.
[43] I. Mandel, W. M. Farr, and J. R. Gair, Mon. Not. Roy. Astron. Soc. **486**, 1086 (2019), arXiv:1809.02063 [physics.data-an].
[44] C. L. MacLeod and C. J. Hogan, Phys. Rev. **D77**, 043512 (2008), arXiv:0712.0618 [astro-ph].

Appendix: Detailed methodology

1. A note on luminosity weighting and redshift evolution

The probability for a galaxy to host a GW event is not uniform over all the galaxies present in the catalog. Indeed, brighter galaxies are supposed to have an higher star-formation rate and hence have an higher probability to host a GW event. Also galaxies at higher redshifts may be more likely to be hosts, as mergers are expected to be more frequent. Our prior belief for a galaxy at redshift z, sky position Ω and absolute and relative magnitudes M, m, to host a GW source s can be expressed as

$$p(z, \Omega, M, m|s, H_0) = p(m|z, \Omega, M, s, H_0)p(z, \Omega, M|s, H_0), \tag{A.1}$$

where if we assume that z, Ω and M are conditionally independent given s, H_0,

$$\begin{aligned} p(z, \Omega, M, m|s, H_0) &= \delta(m - m(z, M, H_0))p(z|s)p(\Omega)p(M|s, H_0), \\ &= \frac{1}{p(s)p(s|H_0)}\delta(m - m(z, M, H_0))p(s|z)p(z)p(\Omega)p(s|M, H_0)p(M|H_0). \end{aligned} \tag{A.2}$$

In the last equation we used the explicit relation between apparent magnitude and z, M and H_0. The probability $p(z)$ is the prior distribution of galaxies in the universe, taken to be uniform in comoving volume-time, $p(\Omega)$ is the prior on galaxy sky location, assumed uniform over the celestial sphere, and $p(M|H_0)$ is the distribution of absolute magnitudes represented by the Schechter function. In the sections below we will show that the terms $p(s)$ and $p(s|H_0)$ cancel out with other terms, and so their exact form does not need to be considered. $p(s|M, H_0)$ can take the form

$$p(s|M, H_0), \propto \begin{cases} L(M(H_0)), & \text{if GW hosting probability is proportional to luminosity} \\ \text{constant} & \text{if GW hosting probability is independent of luminosity.} \end{cases} \tag{A.3}$$

We refer to the above equation as luminosity weighting. The term $p(s|z)$ represents the probability for the merger rate to depend on the redshift,

$$p(s|z), \propto \begin{cases} \text{function}(z) & \text{if rate evolves with redshift} \\ \text{constant} & \text{if rate is does not evolve with redshift.} \end{cases} \tag{A.4}$$

For the MDCs in this paper with $z \ll 1$, we assume a constant rate model but a more generic model with $p(s|z) \propto (1 + z)^{\lambda}$ can be used with detections at higher redshifts. This was the case of [23], for example, in which a $p(s|z) \propto (1 + z)^3$ was assumed.

2. Individual components of the galaxy catalog case

In this section we provide the individual components of Eq. 8. Note that in the cases where the integration boundaries are not specified, they can be assumed to cover the full parameter space.

a. Likelihood when host is in catalog: $p(x_{GW}|G, D_{GW}, H_0)$

The likelihood in the case where the host galaxy is inside the galaxy catalog, $p(x_{\rm GW}|G, D_{\rm GW}, H_0)$, can be obtained from the marginalization over redshift, sky location, absolute magnitude and apparent magnitude. If we assume that the GW data, $x_{\rm GW}$, is independent of the galaxy catalog G, m and M we can write

$$p(x_{\rm GW}|G, D_{\rm GW}, s, H_0) = \frac{1}{p(D_{\rm GW}|G, s, H_0)} \iiiint p(x_{\rm GW}|z, \Omega, s, H_0) p(z, \Omega, M, m|G, s, H_0) dz d\Omega dM dm. \tag{A.5}$$

The probability density function $p(z, \Omega, M, m|G, s, H_0)$ is taken as a sum of delta functions with specific z, Ω and m corresponding to the location of each galaxy in the catalog. This can be further factorized as

$$p(z, \Omega, M, m|G, s, H_0) = \frac{p(s|z, \Omega, M, m, G, H_0)\delta(M - M(z, m, H_0))p(z, \Omega, m|G)}{p(s|G, H_0)}, \tag{A.6}$$

where we have assumed again a relation between the apparent magnitude, redshift, H_0 and absolute magnitude. This allows us to integrate over the absolute magnitude in Eq. A.5 and obtain

$$p(x_{\rm GW}|G, D_{\rm GW}, s, H_0) = \frac{1}{p(D_{\rm GW}|G, s, H_0)p(s|G, H_0)} \iiint p(x_{\rm GW}|z, \Omega, s, H_0) p(s|z, \Omega, M(z, m, H_0), m, G, H_0) p(z, \Omega, m|G) dz d\Omega dm. \tag{A.7}$$

Remembering that $p(z, \Omega, m|G)$ represents the distribution of the galaxies in the catalog, we can replace the integral above with a sum over the galaxies.

$$p(x_{\rm GW}|G, D_{\rm GW}, s, H_0) = \frac{1}{p(D_{\rm GW}|G, s, H_0)p(s|G, H_0)} \frac{1}{N} \sum_{i=1}^{N} p(x_{\rm GW}|z_i, \Omega_i, s, H_0) p(s|z_i) p(s|M(z_i, m_i, H_0)), \tag{A.8}$$

where we have factorized $p(z_i|s)$ and $p(M(z_i, m_i, H_0)|s)$, together with the term $p(s|z, \Omega, M(z, m, H_0), m, G, H_0)$. Finally expanding the denominator $p(D_{\rm GW}|G, s, H_0)$ in the same way, we can recover the likelihood for the "in catalog" part of the galaxy catalog method.

$$p(x_{\rm GW}|G, D_{\rm GW}, s, H_0) = \frac{\sum_{i=1}^{N} p(x_{\rm GW}|z_i, \Omega_i, s, H_0) p(s|z_i) p(s|M(z_i, m_i, H_0))}{\sum_{i=1}^{N} p(D_{\rm GW}|z_i, \Omega_i, s, H_0) p(s|z_i) p(s|M(z_i, m_i, H_0))}. \tag{A.9}$$

Notably, in the case the galaxies in the catalogs are provided along with their redshift uncertainties $p(z_i)$, these can be implemented in the above equations as:

$$p(x_{\rm GW}|G, D_{\rm GW}, s, H_0) = \frac{\sum_{i=1}^{N_{\rm gal}} \int p(x_{\rm GW}|z_i, \Omega_i, s, H_0) p(s|z_i) p(s|M(z_i, m_i, H_0)) p(z_i) dz_i}{\sum_{i=1}^{N_{\rm gal}} \int p(D_{\rm GW}|z_i, \Omega_i, s, H_0) p(s|z_i) p(s|M(z_i, m_i, H_0)) p(z_i) dz_i}. \tag{A.10}$$

b. Probability the host galaxy is in the galaxy catalog: $p(G|D_{GW}, H_0)$ and $p(\bar{G}|D_{GW}, H_0)$

The probability that the host galaxy is inside the galaxy catalog, given that a GW signal was detected, can be expressed as

$$\begin{aligned} p(G|D_{\rm GW}, s, H_0) &= \iiiint p(G|z, \Omega, M, m, D_{\rm GW}, s, H_0) p(z, \Omega, M, m|D_{\rm GW}, s, H_0) dz d\Omega dM dm, \\ &= \iiiint \Theta[m_{\rm th} - m] \frac{p(D_{\rm GW}|z, \Omega, M, m, s, H_0) p(z, \Omega, M, m|s, H_0)}{p(D_{\rm GW}|s, H_0)} dz d\Omega dM dm, \\ &= \frac{1}{p(D_{\rm GW}|s, H_0)} \iiiint \Theta[m_{\rm th} - m] p(D_{\rm GW}|z, \Omega, s, H_0) p(z, \Omega, M, m|s, H_0) dz d\Omega dM dm. \end{aligned} \tag{A.11}$$

If we assume that the galaxy catalog is apparent magnitude-limited, such that only galaxies which are observed above some detection threshold are contained within it, we can approximate $p(G|z,\Omega,M,m,D_{\rm GW},s,H_0)$ as a Heaviside step around the detection threshold $m = m_{\rm th}$.

$$p(G|D_{\rm GW},s,H_0) = \frac{1}{p(D_{\rm GW}|s,H_0)}\iiiii \Theta[m_{\rm th}-m]p(D_{\rm GW}|z,\Omega,s,H_0)p(z,\Omega,M,m|s,H_0)dzd\Omega dMdm. \tag{A.12}$$

We now expand $p(z,\Omega,M,m|s,H_0)$ as in Eq A.2 and we obtain

$$p(G|D_{\rm GW},s,H_0) = \frac{1}{p(s)p(s|H_0)}\frac{1}{p(D_{\rm GW}|s,H_0)}\int_0^{z(M,m_{\rm th},H_0)} dz\int d\Omega\int dM p(D_{\rm GW}|z,\Omega,s,H_0)p(s|z)p(z)p(\Omega)p(s|M,H_0)p(M|H_0). \tag{A.13}$$

The term $p(D_{\rm GW}|s,H_0)$ can be expanded in a similar way and finally gives the probability for the host galaxy to be in the catalog.

$$p(G|D_{\rm GW},s,H_0) = \frac{\displaystyle\int_0^{z(M,m_{\rm th},H_0)} dz\int d\Omega\int dM p(D_{\rm GW}|z,\Omega,s,H_0)p(s|z)p(z)p(\Omega)p(s|M,H_0)p(M|H_0)}{\displaystyle\iiint p(D_{\rm GW}|z,\Omega,s,H_0)p(s|z)p(z)p(\Omega)p(s|M,H_0)p(M|H_0)dzd\Omega dM}. \tag{A.14}$$

As the probabilities of being in the catalog and not in the catalog must be complementary, we have,

$$p(\bar{G}|D_{\rm GW},s,H_0) = 1 - p(G|D_{\rm GW},s,H_0). \tag{A.15}$$

c. *Likelihood when host is not in catalog:* $p(x_{GW}|\bar{G},D_{GW},H_0)$

We follow an approach similar to the one presented in App. 2 a. We expand

$$p(x_{\rm GW}|\bar{G},D_{\rm GW},s,H_0) = \frac{1}{p(D_{\rm GW}|\bar{G},s,H_0)}\iiiii p(x_{\rm GW}|z,\Omega,s,H_0)\frac{p(\bar{G}|z,\Omega,M,m,s,H_0)p(z,\Omega,M,m|s,H_0)}{p(\bar{G}|s,H_0)}dzd\Omega dMdm, \tag{A.16}$$

The prior term, $p(z,\Omega,M,m|s,H_0)$ can now be expanded as it was in Eq A.2. Substituting this in, and utilizing a Heaviside step function to represent the galaxy catalog's apparent magnitude threshold for $p(\bar{G}|z,\Omega,M,m,s,H_0)$,

$$p(x_{\rm GW}|\bar{G},s,H_0) = \frac{1}{p(s)p(s|H_0)}\frac{1}{p(\bar{G}|s,H_0)}\int_{z(H_0,m_{\rm th},M)}^{\infty} dz\int d\Omega\int dM p(x_{\rm GW}|z,\Omega,s,H_0)p(s|z)p(z)p(\Omega)p(s|M,H_0)p(M|H_0). \tag{A.17}$$

Expanding the denominator, $p(D_{\rm GW}|\bar{G},s,H_0)$, in the same way gives an equivalent term,

$$p(D_{\rm GW}|\bar{G},s,H_0) = \frac{1}{p(s)p(s|H_0)}\frac{1}{p(\bar{G}|s,H_0)}\int_{z(H_0,m_{\rm th},M)}^{\infty} dz\int d\Omega\int dM p(D_{\rm GW}|z,\Omega,s,H_0)p(s|z)p(z)p(\Omega)p(s|M,H_0)p(M|H_0). \tag{A.18}$$

And substituting this back into Eq A.16 finally gives,

$$p(x_{\rm GW}|\bar{G},D_{\rm GW},s,H_0) = \frac{\displaystyle\int_{z(M,m_{\rm th},H_0)}^{\infty} dz\int d\Omega\int dM p(x_{\rm GW}|z,\Omega,s,H_0)p(s|z)p(z)p(\Omega)p(s|M,H_0)p(M|H_0)}{\displaystyle\int_{z(M,m_{\rm th},H_0)}^{\infty} dz\int d\Omega\int dM p(D_{\rm GW}|z,\Omega,s,H_0)p(s|z)p(z)p(\Omega)p(s|M,H_0)p(M|H_0)}. \tag{A.19}$$

3. The catalog patch case

While in general the galaxy catalog method derived in 2 was for use with a galaxy catalog which covers the entire sky, a small modification allows the use of catalogs which only cover a patch of sky, as long as the patch can be specified using limits in right ascension and declination. If we represent the sky area covered by the catalog as $\Omega_{\rm cat}$, and the area outside the catalog as $\Omega_{\rm rest}$,

such that $\Omega_{\rm cat} + \Omega_{\rm rest}$ covers the whole sky, this can be written as follows:

$$\begin{aligned} p(x_{\rm GW}|D_{\rm GW}, H_0) &= \int p(x_{\rm GW}|\Omega, D_{\rm GW}, H_0) p(\Omega) d\Omega, \\ &= \int^{\Omega_{\rm cat}} p(x_{\rm GW}|\Omega, D_{\rm GW}, H_0) p(\Omega) d\Omega + \int^{\Omega_{\rm rest}} p(x_{\rm GW}|\Omega, D_{\rm GW}, H_0) p(\Omega) d\Omega. \end{aligned} \tag{A.20}$$

The first term is equivalent to the regular galaxy catalog case, but with limits on the integral over Ω, while the second term has no G and $\bar{G}$ terms, and covers the rest of the sky from redshift 0 to ∞.

4. Direct and Pencil Beam Counterpart Cases

The "direct" method assumes that the counterpart has been unambiguously linked to the host galaxy of the GW event, such that the redshift and sky location of that galaxy can be taken to be that of the GW event with certainty, see Eq. 9. Instead the numerator is calculated by evaluating the GW likelihood at the delta-function location of the counterpart in z and Ω, and the term in the denominator is evaluated as:

$$p(D_{\rm GW}|H_0) = \iiint p(D_{\rm GW}|z, \Omega, H_0) p(z) p(\Omega) p(M|H_0) dz d\Omega dM, \tag{A.21}$$

for priors $p(z)$ and $p(\Omega)$ (note that this is independent of galaxy catalog data).

The "pencil-beam" method makes the assumption that while the sky location of the galaxy associated with the counterpart is that of the GW event, we may not make a direct association to a known galaxy but to a set of potential candidate hosts. We can use the EM constrained sky localization and therefore return to the question of whether the host is within or beyond the galaxy catalog. In this case, the likelihood takes the same form as in the galaxy catalog case, but evaluated along the line of sight of the candidate counterparts.

5. GW selection effects

Eq. 7 in section II C can be written as:

$$p(D_{\rm GW}|H_0) = \int p(D_{\rm GW}|x_{\rm GW}, H_0) p(x_{\rm GW}|H_0) dx_{\rm GW}. \tag{A.22}$$

where $p(D_{\rm GW}|x_{\rm GW}, H_0)$ is a binary quantity which is 1 if the SNR of $x_{\rm GW}$ passes ρ_{th}, and 0 otherwise.

As in section 2, $p(D_{\rm GW}|H_0)$ doesn't appear in the expanded versions of any of the equations, but is itself expanded first, such that the actual quantity we require is $p(D_{\rm GW}|z, \Omega, H_0)$. Calculating $p(D_{\rm GW}|z, \Omega, H_0)$ requires integrating over all realizations of GW events (detected and not), for a range of z, Ω and H_0 values, and applying a detection threshold (ρ_{th}) which all events must pass in order to be deemed detected.

Practically, Monte-Carlo integration can be used:

$$p(D_{\rm GW}|z, \Omega, H_0) = \frac{1}{N_{\rm samples}} \sum_{i=1}^{N_{\rm samples}} p(D_{{\rm GW}i}|x_{{\rm GW}i}, z, \Omega, H_0). \tag{A.23}$$

where $x_{{\rm GW}i}$ corresponds to an event, the parameters of which have been randomly drawn from the prior distributions of parameters which affect an event's detectability (mass, inclination, polarization, and sky location) and the event's ρ_i is calculated for specific values of z and H_0.

$$p(D_{{\rm GW}i}|x_{{\rm GW}i}, z, \Omega, H_0) = \begin{cases} 1, & \text{if } \rho > \rho_{th} \\ 0, & \text{otherwise.} \end{cases} \tag{A.24}$$

which gives a smooth function for $p(D_{\rm GW}|z, \Omega, H_0)$, which drops from 1 to 0 over a range of z, Ω and H_0 values.

a. Prior mass distribution

An event's detectability is dependent on its observed (redshifted) detector-frame mass, M_z, but priors on the mass refer to their source-frame mass. When calculating $p(D|H_0)$ the masses are drawn from the priors on source mass, $p(M_1, M_2)$ and then converted to observed masses through the equation:

$$M_z = (1 + z)M. \tag{A.25}$$

However, when we use GW data in the form of posterior samples, the prior used to generate those is uniform on the redshifted mass, M_z [39]. Due to the way the MDC GW data was generated, with masses chosen on the detector-frame, rather than the source-frame, this was not something which had to be considered. With real GW data, as the redshift is linked directly to H_0, it is necessary to take into account the redshifting of the masses explicitly.

In general, when calculating $p(D|H_0)$ for BBHs, the primary mass M_1 is drawn from a power-law with slope α, between limits $[a, b]$ $M_\odot$. The secondary mass, M_2 is drawn from a uniform distribution between $aM_\odot$ and M_1 [27], to give (for $\alpha \neq -1$):

$$p(M_1, M_2) = \frac{(\alpha + 1)M_1^\alpha}{bM_\odot^{(\alpha+1)} - aM_\odot^{(\alpha+1)}} \frac{1}{M_1 - aM_\odot}. \tag{A.26}$$

This is related to the the redshifted mass by the Jacobian:

$$\begin{aligned} p(M_{1,z}, M_{2,z}) &= p(M_1, M_2)\left|\frac{\partial(M_1, M_2)}{\partial(M_{1,z}, M_{2,z})}\right|, \\ &= p(M_1, M_2)\left|\frac{1}{(1 + z)^2}\right|. \end{aligned} \tag{A.27}$$

Substituting in our expression for $p(M_1, M_2)$:

$$\begin{aligned} p(M_{1,z}, M_{1,z}) &= \frac{(\alpha + 1)M_1^\alpha}{bM_\odot^{(\alpha+1)} - aM_\odot^{(\alpha+1)}} \frac{1}{M_1 - aM_\odot} \frac{1}{(1 + z)^2}, \\ &= \frac{(1 + z)^2(\alpha + 1)M_{1,z}^\alpha}{bM_{\odot,z}^{(\alpha+1)} - aM_{\odot,z}^{(\alpha+1)}} \frac{1}{M_{1,z} - aM_{\odot,z}} \frac{1}{(1 + z)^2}, \\ &= \frac{(\alpha + 1)M_{1,z}^\alpha}{bM_{\odot,z}^{(\alpha+1)} - aM_{\odot,z}^{(\alpha+1)}} \frac{1}{M_{1,z} - aM_{\odot,z}}. \end{aligned} \tag{A.28}$$

The factor of $(1 + z)^2$ cancels in the numerator and denominator. As all redshift (and hence H_0) dependence has been removed, no correction is required for the differing priors. For the case in which $\alpha = -1$, it can be shown that all redshift dependence falls out as well, meaning that as long as the prior mass distribution takes the form of a power-law, no prior correction is required. This will not be the case for all mass distributions.