SUBMITTED TO APJ
Preprint typeset using LaTeX style emulateapj v. 12/16/11

PARAMETER ESTIMATION FOR BINARY NEUTRON-STAR COALESCENCES WITH REALISTIC NOISE DURING THE ADVANCED LIGO ERA

CHRISTOPHER P. L. BERRY[1], ILYA MANDEL[1], HANNAH MIDDLETON[1], LEO P. SINGER[2], ALEX L. URBAN[3], ALBERTO VECCHIO[1], SALVATORE VITALE[4], KIPP CANNON[5], BEN FARR[6,1], WILL M. FARR[1], PHILIP B. GRAFF[7,8], CHAD HANNA[9,10], CARL-JOHAN HASTER[1], SATYA MOHAPATRA[11,4], CHRIS PANKOW[3], LARRY R. PRICE[2], TREVOR SIDERY[1], AND JOHN VEITCH[1]

Submitted to ApJ

ABSTRACT

Advanced ground-based gravitational-wave (GW) detectors begin operation imminently. Their intended goal is not only to make the first direct detection of GWs, but also to make inferences about the source systems. Binary neutron-star mergers are among the most promising sources. We investigate the performance of the parameter-estimation pipeline that will be used during the first observing run of the Advanced Laser Interferometer Gravitational-wave Observatory (aLIGO) in 2015: we concentrate on the ability to reconstruct the source location on the sky, but also consider the ability to measure masses and the distance. Accurate, rapid sky-localization is necessary to alert electromagnetic (EM) observatories so that they can perform follow-up searches for counterpart transient events. We consider parameter-estimation accuracy in the presence of realistic, non-Gaussian noise. We find that the character of the noise makes negligible difference to the parameter-estimation performance. The source luminosity distance can only be poorly constrained, the median 90% (50%) credible interval scaled with respect to the true distance is 0.85 (0.38). However, the chirp mass is well measured. Our chirp-mass estimates are subject to systematic error because we used gravitational-waveform templates without component spin to carry out inference on signals with moderate spins, but the total error is typically less than $10^{-3}M_{\odot}$. The median 90% (50%) credible region for sky localization is ~ 600 deg^2 (~ 150 deg^2), with 3% (30%) of detected events localized within 100 deg^2. Early aLIGO, with only two detectors, will have a sky-localization accuracy for binary neutron stars of hundreds of square degrees; this makes EM follow-up challenging, but not impossible.

Keywords: gravitational waves — methods: data analysis — stars: neutron — surveys

1. INTRODUCTION

The goal of gravitational-wave (GW) astronomy is to learn about the Universe through observations of gravitational radiation. This requires not only the ability to detect GWs, but also to infer the properties of their source systems. In this work, we investigate the ability to perform parameter estimation (PE) on signals detected by the upcoming Advanced LIGO (aLIGO) instruments (Harry 2010) in the initial phase of their operation (Aasi et al. 2013b).

Compact binary coalescences (CBCs), the GW-driven inspiral and merger of stellar-mass compact objects, are a prime source for aLIGO and Advanced Virgo (AdV; Acernese et al. 2009, 2014). Binary neutron-star (BNS) systems may be the most abundant detectable CBCs (Abadie et al. 2010). We focus on BNS mergers in this study.

Following the identification of a detection candidate, we wish to extract the maximum amount of information from the signal. It is possible to make some inferences using selected components of the data. However, full information regarding the source system, including the component objects' masses and spins, is encoded within the gravitational waveform, and can be obtained by comparing the data to theoretical waveform models (Cutler & Flanagan 1994; Jaranowski & Krolak 2005). Doing so can be computationally expensive.

PE is performed within a Bayesian framework. We use algorithms available as part of the LALINFERENCE toolkit for the analysis of CBC signals. The most expedient code is BAYESTAR (Singer et al. 2014; Singer 2014), which infers sky location from data returned from the detection pipeline. Exploring the posterior probability densities for the parameters takes longer for models where the parameter space is larger or the likelihood is more complicated. Calculating estimates for parameters beyond sky location is done using the stochastic-sampling algorithms of LALINFERENCE (Veitch et al. 2014). There are three interchangeable sampling algorithms: LALINFERENCE_NEST (Veitch & Vecchio 2010), LALINFERENCE_MCMC (van der Sluys et al. 2008a; Raymond et al. 2009), and LALINFERENCE_BAMBI (Graff et al. 2012), which we refer to as LALINFERENCE for short. These compute waveform templates for use in the likelihood. Using the least computationally expensive waveforms allows for posteriors to be estimated on timescales of hours to days; potentially more accurate estimates can be calculated with more expensive waveforms. In this paper, we discuss what can be

cplb@star.sr.bham.ac.uk
[1] School of Physics & Astronomy, University of Birmingham, Birmingham, B15 2TT, UK
[2] LIGO Laboratory, California Institute of Technology, Pasadena, CA 91125, USA
[3] Leonard E. Parker Center for Gravitation, Cosmology, and Astrophysics, University of Wisconsin–Milwaukee, Milwaukee, WI 53201, USA
[4] Massachusetts Institute of Technology, 185 Albany St, Cambridge, Massachusetts 02139, USA
[5] Canadian Institute for Theoretical Astrophysics, 60 St. George Street, University of Toronto, Toronto, Ontario, M5S 3H8, Canada
[6] Department of Physics and Astronomy & Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA), Northwestern University, Evanston, IL 60208, USA
[7] Department of Physics, University of Maryland–College Park, College Park, MD 20742, USA
[8] Gravitational Astrophysics Lab, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[9] Perimeter Institute for Theoretical Physics, Ontario, N2L 2Y5, Canada
[10] The Pennsylvania State University, University Park, PA 16802, USA
[11] Syracuse University, Syracuse, NY 13244, USA

achieved using low-latency (BAYESTAR) and medium-latency (LALINFERENCE with inexpensive waveforms) PE; a subsequent paper will evaluate what can be achieved on longer timescales using more expensive waveform templates.

With the detection of GWs, it is also possible to perform multi-messenger astronomy, connecting different types of observations of the same event. BNS mergers could be accompanied by an electromagnetic (EM) counterpart (Metzger & Berger 2012). To associate an EM event with a GW signal, it is beneficial to have an accurate sky location: timing information can also be used for EM signals that are independently detected, such as gamma-ray bursts (Aasi et al. 2014b). To provide triggers for telescopes to follow up a GW detection, it is necessary to provide rapid sky localization.

Several large-scale studies investigated the accuracy with which sky position can be reconstructed from observations with ground-based detector networks. The first only used timing information from a multi-detector network to triangulate the source position on the sky (e.g., Fairhurst 2009, 2011). Subsequently, further information about the phase of the gravitational waveform was folded into the timing triangulation (TT) analysis (Grover et al. 2014). The most sophisticated techniques perform a coherent Bayesian analysis to reconstruct probability distributions for the sky location (e.g., Veitch et al. 2012; Kasliwal & Nissanke 2014; Grover et al. 2014; Sidery et al. 2014). Singer et al. (2014) used both BAYESTAR and LALINFERENCE to analyse the potential performance of aLIGO and AdV in the first two years of their operation. They assumed the detector noise was stationary and Gaussian. Here, we further their studies (although we use the same analysis pipeline) by using a set of injections into observed noise from initial LIGO detectors recoloured (see section 2.1) to the expected spectral density of early aLIGO.[12] This provides results closer to those expected in practice, as real interferometer noise includes features such as non-stationary glitches (Aasi et al. 2013b, 2014a). Our results are just for the first observing run (O1) of aLIGO, expected in the latter half of 2015, assuming that this occurs before the introduction of AdV. As the sensitivity of the detectors will increase with time, and because the introduction of further detectors increases the accuracy of sky localization (Schutz 2011), these set a lower bound for the advanced-detector era. Estimates for sky-localization accuracy in later observing periods can be calibrated using our results.

PE beyond sky localization, considering the source system's mass, spin, distance and orientation, has been subject to similar studies. The initial investigations estimated PE using the Fisher information matrix (e.g., Cutler & Flanagan 1994; Poisson & Will 1995; Arun et al. 2005). The Fisher information matrix only gives an approximation to true PE potential (Vallisneri 2008). More reliable (but computationally expensive) results are found by simulating a GW event and analysing it using our PE codes, mapping the posterior probability distributions (e.g., Rover et al. 2006; van der Sluys et al. 2008b; Veitch & Vecchio 2010; Rodriguez et al. 2014). This has even been done for a blind injection during the run of initial LIGO (Aasi et al. 2013a). As with sky-localization, more general PE can improve with the introduction of more detectors to the network (Veitch et al. 2012).

To be as faithful as possible, our analysis is performed using one of the pipelines intended for use during O1. We make use of various components of the LIGO Scientific Collaboration Algorithm Library (LAL).[13] In particular, we shall make use of GSTLAL,[14] one of the detection pipelines, to search for signals and LALINFERENCE for PE on detection candidates.

We begin by describing the source catalogue and detector sensitivity curve used for this study in section 2. In section 3 we explain how the data is analysed to produce sky areas and other parameter estimates. Many of the details of these two sections are shared with the preceding work of Singer et al. (2014), so the interested reader is recommended to consult that for further details. In section 4 we present the results of our work. We first discuss the set of events that are selected by the detection pipeline in section 4.1 (with supplementary information in appendix A); then we examine PE, considering sky-localization accuracy in section 4.2, and mass and distance measurements in section 4.3. We conclude with a discussion of these results in section 5; this includes in section 5.1.2 an analysis of estimates for sky localization in later observing periods with reference to our findings. Estimates of the computational costs associated with running BAYESTAR and full LALINFERENCE PE are given in appendix B.

Our main findings are:

- The detection pipeline returns a population of sources that is not significantly different from the input astrophysical population, despite a selection bias based upon the chirp mass.
- Both BAYESTAR and LALINFERENCE return comparable sky-localization accuracies (for a two-detector network). The latter takes more computational time (a total CPU time of $\sim 10^6$ s per event compared with $\sim 10^3$ s), but returns estimates for more parameters than just location.
- At a given signal-to-noise ratio (SNR), the character of the noise does not affect sky localization or other PE.
- Switching from a detection threshold based upon SNR to one based upon the false alarm rate (FAR) changes the SNR-distribution of detected events. A selection based upon FAR includes more low-SNR events (the distribution at high SNRs is unaffected).
- TT provides a poor predictor of sky localization for a two-detector network; it does better (on average) for a three-detector network when phase coherence is included, but remains imperfect.
- Systematic errors from uncertainty in the waveform template are significant for chirp-mass estimation. Neglecting the mass-spin degeneracy by using non-spinning waveforms artificially narrows the posterior distribution.

For O1, we find that the luminosity distance is not well-measured, the median 50% credible interval (interquartile range) divided by the true distance is 0.38 and the median 90% credible interval divided by the true distance is 0.85. Despite being subject to systematic error, the chirp mass is still accurately measured, with the posterior mean being less than $10^{-3}M_{\odot}$ from the true value is almost all (96%) cases. We find

[12] We refer to the noise as recoloured as it is first whitened (removing its colour), to eliminate initial LIGO's frequency dependence, and then passed through a linear response filter (reintroducing colour) so that, on average, it has the aLIGO spectral density.

[13] http://www.lsc-group.phys.uwm.edu/lal

[14] https://www.lsc-group.phys.uwm.edu/daswg/projects/gstlal.html

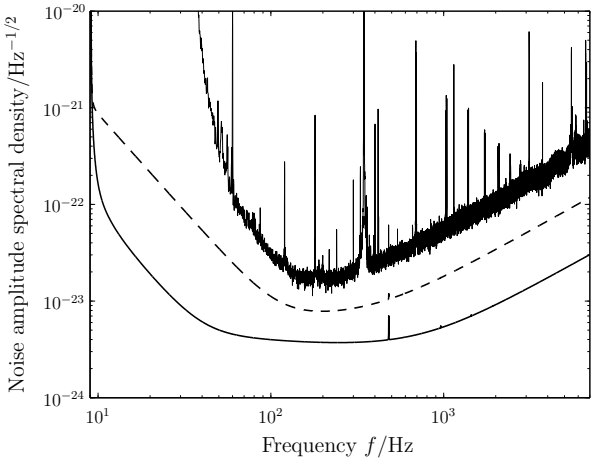


Figure 1. Initial and Advanced LIGO noise amplitude spectral densities. The upper line is the measured sensitivity of the initial LIGO Hanford detector during S6 (Aasi et al. 2014a). The dashed line shows the early aLIGO sensitivity and the lower solid line the final sensitivity (Barsotti & Fritschel 2012). The early sensitivity is used as a base here.

that the median area of 50% sky localization credible region is 154 deg^2 and the median area of the 90% credible region is 632 deg^2; the median searched area (area of the smallest credible region that encompasses the source location) is 132 deg^2. EM follow-up to BNS mergers in 2015 will be challenging and require careful planning.

2. SOURCES AND SENSITIVITIES

Our input data consists of two components: simulated detector noise and simulated BNS signals. We describe the details of these in the following subsections, before continuing with the analysis of the data in section 3.

2.1. *Recoloured 2015 noise*

We consider the initial operation of the advanced detectors at LIGO Hanford and LIGO Livingston. The sensitivity is assumed to be given by the early curve of Barsotti & Fritschel (2012), which has a BNS detection range of ~ 55 Mpc (assuming Gaussian noise). This configuration corresponds to the 2015 observing scenario in Aasi et al. (2013b). Figure 1 plots the noise spectral density, the square root of the power spectral density (Moore et al. 2014), as measured during the sixth science (S6) run of initial LIGO,[15] the early aLIGO sensitivity curve, and final aLIGO curve (Shoemaker 2010).

The noise is constructed from data from the sixth science (S6) run of initial LIGO (Christensen 2010; Aasi et al. 2014a), recoloured to the early aLIGO noise spectral density. Two calendar months (21 August 2010–20 October 2010) of S6 data were used. The recoloured data are constructed using GSTLAL_FAKE_FRAMES.[16] The recolouring process can be thought of as applying a finite-impulse response filter to whitened noise. The result is a noise stream that, on average, has the same power spectral density as expected for early aLIGO, but contains transients that are similar to those found in S6. We use real noise, instead of idealised Gaussian noise, to try to capture a realistic detector response including transients; however, the S6 noise can only serve as a proxy for the actual noise in aLIGO since the detectors are different.

[15] http://www.ligo.caltech.edu/~jzweizig/distribution/LSC_Data/

[16] https://ldas-jobs.ligo.caltech.edu/~gstlalcbc/doc/gstlal-0.7.1/html/gstlal__fake__frames.html

2.2. *Binary neutron-star events*

BNS systems constitute the most probable and best understood source of signals for advanced ground-based GW detectors. There is a wide range in predicted event rates as a consequence of uncertainty in our knowledge of the astrophysics. Abadie et al. (2010) gives a BNS merger rate for the full-sensitivity aLIGO–AdV network of 0.01–10 Mpc^{-3} Myr^{-1}, with 1 Mpc^{-3} Myr^{-1} as the most realistic estimate (Kalogera et al. 2004).

We use exactly the same list of simulated sources as in Singer et al. (2014). The neutron-star masses are taken to be uniformly distributed from $m_{\rm min} = 1.2M_\odot$ to $m_{\rm max} = 1.6M_\odot$, which safely encompasses the observed mass range of BNS systems (Kiziltan et al. 2013). Their (dimensionless) spin magnitudes are uniformly distributed between $a_{\rm min} = 0$ and $a_{\rm max} = 0.05$. The most rapidly rotating neutron star observed in a binary with another neutron star that is close enough to merge in a Hubble time is PSR J0737−3039A (Burgay et al. 2003; Kramer & Wex 2009). This has been estimated to have a spin within this range (Mandel & O'Shaughnessy 2010; Brown et al. 2012): since we do not know precisely the neutron-star equation of state (Lattimer 2012), it is not possible to exactly convert from a spin period to a spin magnitude. The spin orientations are distributed isotropically. The binaries are uniformly scattered in volume, and all their various orientation angles are drawn from isotropic distributions. This set of parameters is motivated by our understanding of the astrophysical population of BNSs.

The GW signals were constructed using a post-Newtonian (PN) inspiral template, the SpinTaylorT4 approximant (Buonanno et al. 2003, 2009) which is a time-domain approximant accurate to 3.5PN order in phase and 1.5PN order in amplitude. There exist more accurate but more expensive waveforms. This template only contains the inspiral part of the waveform and not the subsequent merger: this should happen outside of the sensitive band of the detector for the masses considered and so should not influence PE (Mandel et al. 2014). We do not use SpinTalyorT4 templates either for detection or PE in this study, instead we use a less expensive approximant. In a future study, we shall investigate the effects of using SpinTaylorT4 templates for PE, such that the injection and recovery templates perfectly match.

3. ANALYSIS PIPELINE

To accurately forecast sky localization prospects in O1, we run our simulated events through the same data-analysis pipeline as is intended for real data. The results of this pipeline are analysed in the next section (section 4). A GW search is performed using GSTLAL_INSPIRAL (Cannon et al. 2010, 2011, 2012, 2013); this is designed to provide GW triggers in real time with $\sim$ 10–100 s latency during LIGO–Virgo observing runs. A trigger is followed up for sky localization if its calculated FAR is less than 10^{-2} yr^{-1}, which is roughly equivalent to a network SNR threshold of $\varrho \gtrsim 12$ (Aasi et al. 2013b).

In using the FAR to select triggers, our method differs from that used in Singer et al. (2014). Since they considered Gaussian noise, which is free of glitches, their FAR would not be representative of those computed using real noise; the FAR calculated with Gaussian noise corresponds to a SNR-threshold that is too low for detection in realistic noise. Therefore, they also imposed a network-SNR cut of $\varrho \geq 12$, in addition to the FAR selection. This joint SNR and FAR threshold was found to differ negligibly from an SNR-only threshold: in effect, they

selected by SNR alone. While this is a small difference in selection criteria, we shall see in section 4.2 that this has an impact on our sky-localization results.

To recover the GW signal, another PN inspiral approximant, TaylorF2 (Damour et al. 2001, 2002; Buonanno et al. 2009), was used as a template. This is a frequency-domain stationary-phase approximation waveform accurate to 3.5PN order in phase and Newtonian order in amplitude. It does not include the effects of spin, although it can be modified to incorporate these (Mikoczi et al. 2005; Arun et al. 2009; Bohé et al. 2013). We neglect spin as this should not lead to a significant reduction in detection efficiency for systems with low spins (Brown et al. 2012), which we confirm in section 4.1.2. TaylorF2 does not incorporate as many physical effects as SpinTaylorT4, notably it does not include precession, but is less computationally expensive, permitting more rapid follow-up.

Rapid sky localization is computed using BAYESTAR (Singer et al. 2014). This reconstructs sky position using a combination of information associated with the triggers: the times, phases and amplitudes of the signals at arrival at each detector. It coherently combines this information to reconstruct posteriors for the sky position. BAYESTAR makes no attempt to infer intrinsic parameters such as the BNS masses and, hence, can avoid computationally expensive waveform calculations. The sky-position distributions can be formulated in under a minute (see appendix B).

Full PE, computing posterior distributions for sky localization parameters as well as the other parameters for the source system like masses, orientation and inclination, is performed using LALINFERENCE (Veitch et al. 2014). LALINFERENCE maps the posterior probability distribution by stochastically sampling the parameter space (e.g., MacKay 2003, chapter 29). There are three codes within LALINFERENCE to sample these posterior distributions: LALINFERENCE_NEST (Veitch & Vecchio 2010), a nested sampling algorithm (Skilling 2006); LALINFERENCE_MCMC (van der Sluys et al. 2008a; Raymond et al. 2009), a Markov-chain Monte Carlo algorithm (Gregory 2005, chapter 12), and LALINFERENCE_BAMBI (Graff et al. 2012), another nested sampling algorithm (Feroz et al. 2009) which incorporates a means of speeding up likelihood evaluation using machine learning (Graff et al. 2014). All three codes use the same likelihood and so should recover the same posteriors; consistency of the codes has been repeatedly checked. While the codes produce the same results, they may not do so in the same times, depending upon the particular problem. All the results here were computed with LALINFERENCE_NEST.

TaylorF2 waveforms were used again in constructing the LALINFERENCE posterior. Since these do not exactly match the waveforms used for injection, there may be a small bias in the recovered parameters (Buonanno et al. 2009). Using TaylorF2 is much less computationally expensive than using SpinTaylorT4, in this case a LALINFERENCE run takes $\sim 10^6$ s of CPU time (see appendix B).

4. RESULTS

4.1. *Detection catalogue*

We ran sky-localization codes on a set of 333 events recovered from the detection pipeline. We shall compare these to the results of Singer et al. (2014) who used Gaussian noise for the same sensitivity curve. They ran BAYESTAR on a sample of 630 events, but only ran LALINFERENCE on a sub-sample of 250 events. We first consider the set of detected events before moving on to examine sky-localization accuracies in section 4.2, and mass and distance measurement in section 4.3.

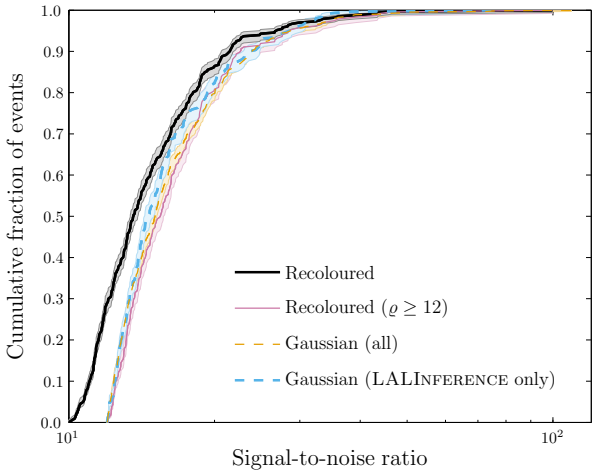


Figure 2. Cumulative fractions of events with network signal-to-noise ratios smaller than the abscissa value. The SNR distribution assuming recoloured noise is denoted by the thick solid line; we also show the distribution subject to a lower cutoff of $\varrho \geq 12$, denoted by the thin solid line. The SNR distribution for the complete set of 630 events with Gaussian noise analysed with BAYESTAR is denoted by the thinner dashed line and the distribution for the subset of 250 events analysed with both BAYESTAR and LALINFERENCE is denoted by the thicker dashed line (Singer et al. 2014). The 68% confidence intervals (1σ for a normal distribution) are denoted by the shaded areas, these are estimated from a beta distribution (Cameron 2011).

4.1.1. *Signal-to-noise ratio distribution*

Unsurprisingly, the distribution of SNRs differs between the recoloured and Gaussian data sets. This is shown in figure 2. The recoloured SNR distribution includes a tail at low SNR ($\varrho \simeq 10$–12). If we impose a lower threshold $\varrho \geq 12$ for the recoloured data set, as was done for the Gaussian data set, we find that the SNR distributions are similar. With the shared SNR cut, the distributions agree within the expected sampling error; performing a Kolmogorov–Smirnov (KS) test (DeGroot 1975, section 9.5) comparing the recoloured SNR distribution to the complete (LALINFERENCE only) Gaussian SNR distribution returns a p-value of 0.311 (0.110).

Comparing injections between the recoloured and Gaussian data sets, there are 255 events that have been detected in both sets. There are 108 events shared between the recoloured data set and the sub-sample of the Gaussian data set analysed with LALINFERENCE. Considering individual events, we may contrast the SNR for recoloured noise ϱ_{R} and Gaussian noise ϱ_{G}. The ratio of the two SNRs are shown in figure 3. Considering the entire population of shared detections, the mean value of the ratio of SNRs is $\varrho_{\mathrm{R}}/\varrho_{\mathrm{G}} = 0.938 \pm 0.006$, showing a small downwards bias as an effect of the differing cutoffs used for the two samples. To limit selection effects that could skew the distribution of the ratio of SNRs, we can impose an SNR cut of $\varrho_{\mathrm{R}} \geq 12$. This reduces the number of events detected in both noise sets to 214 using the full Gaussian set and 88 for the LALINFERENCE Gaussian sub-sample. There is a small difference between the SNR as calculated with Gaussian noise and with recoloured noise. This does not appear to be a strong function of the SNR. However, the scatter in the ratio decreases as SNR increases, approximately decreasing as ϱ^{-1}. This is as expected as the inclusion of random noise realisations in the signal should produce fluctuations in the SNR of order ± 1; these fluctuations become less significant for louder events. After imposing the cut $\varrho \geq 12$ on both sets, the mean value of

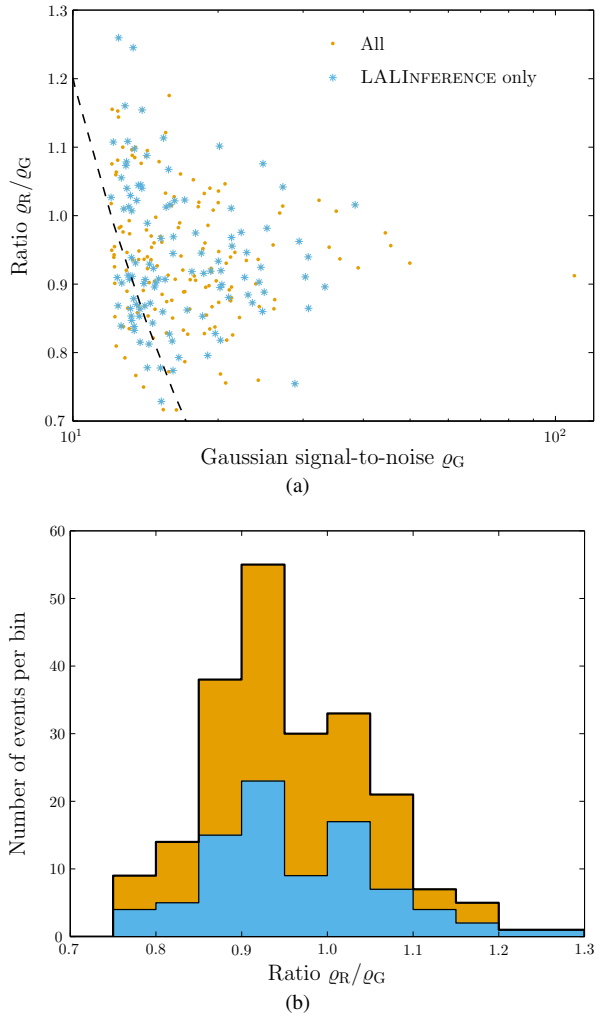


Figure 3. Comparison of signal-to-noise ratios from injections with Gaussian noise ϱ_G and from injections with recoloured noise ϱ_R. (a) The ratio ϱ_R/ϱ_G as a function of ϱ_G. The dashed line shows the locus of $\varrho_R = 12$. (b) Distribution of ϱ_R/ϱ_G with both $\varrho_G \geq 12$ and $\varrho_R \geq 12$, using a bin width of 0.5. Events that fall within the sub-sample of Gaussian events analysed with LALInference are highlighted with blue (star-shaped points) and the complete set of events detected in both the Gaussian and recoloured data sets is indicated by orange (round points).

the ratio of SNRs is $\varrho_R/\varrho_G = 0.955 \pm 0.006$. Although there is a small difference in SNRs, we shall see that this does not impact our PE results.

4.1.2. *Selection effects*

The population of detected events should not match exactly the injected distribution; depending upon their parameters, some systems are louder and hence easier to detect. Here, we look at the selection effects of the most astrophysically interesting parameters: mass and spin. We expect there to be a selection based upon mass, as the component masses set the amplitude of the waveform. We do not expect there to be a dependence upon the spin because the spin magnitude is small, but since we injected with a spinning waveform and recovered with a non-spinning waveform, there could potentially be a selection effect due to waveform mismatch. Checking these distributions confirms the effectiveness of the detection pipeline for this study.

To leading order, the GW amplitude is determined by the

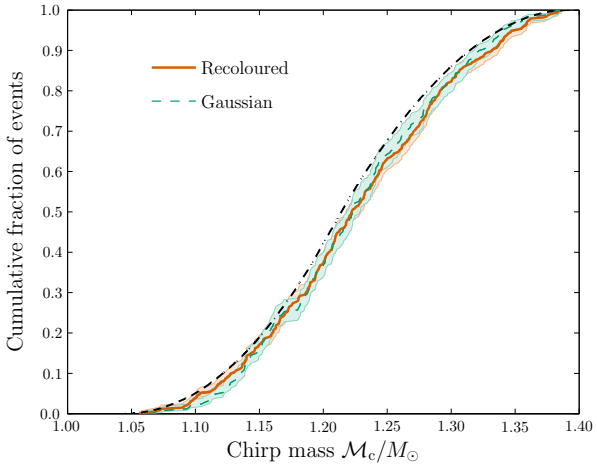


Figure 4. Cumulative fractions of detected events with chirp masses smaller than the abscissa value. Results using recoloured noise are denoted by the solid line, and results from the subset of 250 events with Gaussian noise analysed with LALInference are denoted by the dashed line (Singer et al. 2014). The 68% confidence intervals are denoted by the shaded areas. The injection distribution, based upon a uniform distribution of component masses, is indicated by the dot–dashed line.

(5/6 power of the) chirp mass (Sathyaprakash & Schutz 2009)

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}, \tag{1}$$

where m_1 and m_2 are the individual component masses. We therefore expect to preferentially select systems with larger chirp masses.

Figure 4 shows the recovered distribution of (injected) chirp masses and the injection distribution (which is calculated numerically). We do detect fewer systems with smaller chirp masses (and more with larger chirp masses), as indicated by the curve for the recovered distribution lying below the curve for the injection distribution. However, this selection effect does not alter the overall character of population. The difference is only marginally statistically significant with this number of events (a KS test with the injection distribution yields p-values of 0.315 and 0.068 for the Gaussian and recoloured noise respectively). This is consistent with expectations for this narrow chirp-mass distribution; in appendix A we use a simple theoretical model to predict that we would need $\sim 10^3$ detections (or a broader distribution of chirp masses in the injection set) to see a significant difference between the injected and recovered populations. The character of the noise does not influence the chirp mass distribution (a KS test gives a p-value of 0.999).

For completeness, in appendix A we present the distributions for the individual component masses, the asymmetric mass ratio and the total mass. The selection effects on these depend upon their correlation with the chirp mass; the total mass, which is most strongly correlated with the chirp mass, shows the most noticeable difference between injection and detected distributions.

Since we injected with a spinning waveform and recovered with a non-spinning waveform, there could also be a selection bias depending upon the spin magnitude. Figure 5 shows the recovered distribution of (injected) spins. The detected events are consistent with having the uniform distribution of spins used for the injections. This is to be expected, since the spin magnitudes are always small and, thus, should not significantly influence the waveform. We conclude that the presence of spin

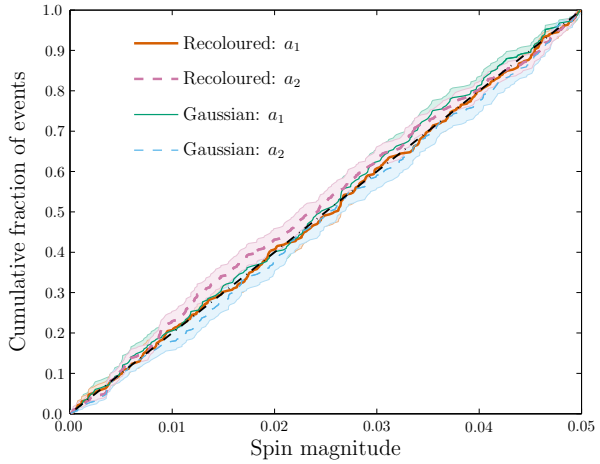


Figure 5. Cumulative fractions of detected events with spin magnitudes smaller than the abscissa value. The spin distribution for the first neutron star a_1 is denoted by the solid line, and the distribution for the second neutron star a_2 is denoted by the dashed line. Results using recoloured noise are denoted by the thicker red–purple lines, and results from the subset of 250 events with Gaussian noise analysed with LALINFERENCE are denoted by the thinner blue–green lines (Singer et al. 2014). The 68% confidence intervals are denoted by the shaded areas. The expected distribution for spins uniform from $a_{\min} = 0$ to $a_{\max} = 0.05$ is indicated by the black dot–dashed line.

does not affect the detection efficiency for BNS systems, in agreement with Brown et al. (2012).

4.2. *Sky-localization accuracy*

The recovered sky positions from BAYESTAR and LALINFERENCE appear in good agreement. A typical example of the recovered posterior probability density is shown in figure 6. This is a bimodal distribution, reflecting the symmetry in the sensitivity of the detectors, which is common (Singer et al. 2014). We use geographic coordinates to emphasise the connection to the position of the detectors. A catalogue of results will be made available online in a format similar to Singer et al. (2014).

To quantify the accuracy of sky localization, we use credible regions: areas of the sky that include a given total posterior probability. We denote the credible region for a total posterior probability p as CR_p: it is defined as

$$\mathrm{CR}_p \equiv \min A \tag{2}$$

such that the sky area A satisfies

$$p = \int_A \mathrm{d}\mathbf{\Omega} P_\Omega(\mathbf{\Omega}), \tag{3}$$

where $P_\Omega(\mathbf{\Omega})$ is the posterior probability density over sky position $\mathbf{\Omega}$ (Sidery et al. 2014). A smaller CR_p at a given p indicates more precise sky localization.

We also consider the searched area: the area of the smallest credible region that includes the true location, and, hence, the area of the sky that we expect would have to be observed before the true source was found.

The self-consistency of our sky areas can be checked by calculating the fraction of events that fall within the credible region at the given probability. We expect that a fraction p of true sky positions are found within CR_p, that is the frequentist confidence region agrees with our Bayesian credible region (Sidery et al. 2014). Figure 7 shows the fraction of events found within a given CR_p as a function of p. The distributions are consistent with expectations: performing a KS test with the predicted distribution yields p-values of 0.455 and 0.546 for LALINFERENCE and BAYESTAR respectively. Both LALINFERENCE and BAYESTAR produce self-consistent and unbiased sky areas in the presence of recoloured noise.

The recovered sky areas are plotted in figure 8. This shows the cumulative distribution of areas for $\mathrm{CR}_{0.5}$, $\mathrm{CR}_{0.9}$ and searched areas A_* as recovered from LALINFERENCE and BAYESTAR. For comparison, we plot both the results using recoloured noise and the results using Gaussian noise from Singer et al. (2014). All the results are similar. LALINFERENCE produces (marginally) more accurate sky localizations than BAYESTAR, but the rapid code does a successful job of reconstructing the sky position in a much shorter time (see appendix B for estimates of computation time). The recovered areas are (generally) marginally smaller for LALINFERENCE as this makes use of more information and so is expected to perform better (a KS test returns p-values of 0.740 when comparing $\mathrm{CR}_{0.9}$ for Gaussian noise and 0.181 for recoloured noise).

The difference between the Gaussian and recoloured results can be understood as a consequence of the SNR distribution (see figure 2). The SNR is the dominant factor affecting sky localization. For example, there is no strong correlation between the time delay between detection at the two LIGO sites and the sky-localization accuracy. The inclusion of more low-SNR events means that, on average, the results using recoloured noise are worse.

The sky-localization accuracy is expected to scale with ϱ^2: doubling the SNR reduces the sky area by approximately a factor of four. The uncertainty in each direction on the sky scales inversely with the SNR, hence the area scales inversely with the square of the SNR (cf. Fairhurst 2009, 2011). This SNR scaling can be verified by plotting recovered sky areas as a function of ϱ as shown in figure 9. The recovered areas do show the expected correlation, although there is considerable scatter resulting from the variation in intrinsic parameters.

We have plotted fiducial best-fit lines with the expected scaling. The fitting was done simply using a naive least-squares method, fitting a straight line to $\log\varrho$ and $\log A$ for each sky area A. Allowing the slope of the line to vary from -2 yields negligible change to the fit. There is little difference between the trends for the recoloured and Gaussian results, indicating that the variation in the sky-localization accuracies is primarily an effect of the different distribution of SNRs. There is a small discrepancy between LALINFERENCE and BAYESTAR in both cases, but the difference is not significant and is within the uncertainty expected from the scatter of results. The general trend for the sky-localization areas can be approximated as

$$\log_{10}\left(\frac{\mathrm{CR}_{0.5}}{\mathrm{deg}^2}\right) \approx -2\log_{10}\varrho + 4.46, \tag{4a}$$

$$\log_{10}\left(\frac{\mathrm{CR}_{0.9}}{\mathrm{deg}^2}\right) \approx -2\log_{10}\varrho + 5.06. \tag{4b}$$

Sky-localization accuracy (at a given SNR) does not appear to be sensitive to the Gaussianity of the noise.

From our fits (4), we can immediately see that the ratio $\mathrm{CR}_{0.9}/\mathrm{CR}_{0.5}$ is about $10^{0.6} \simeq 4$. Considering this ratio for each posterior, the mean value of $\log_{10}(\mathrm{CR}_{0.9}/\mathrm{CR}_{0.5})$ is 0.60 and the standard deviation is 0.07. For comparison, if the posterior were a 1-d Gaussian, we would expect the ratio to be $\mathrm{erf}^{-1}(0.9)/\mathrm{erf}^{-1}(0.5) \simeq 2.4 \simeq 10^{0.39}$, and if it were a 2-d Gaussian, the ratio would be $\ln(1-0.9)/\ln(1-0.5) \simeq 3.3 \simeq 10^{0.52}$

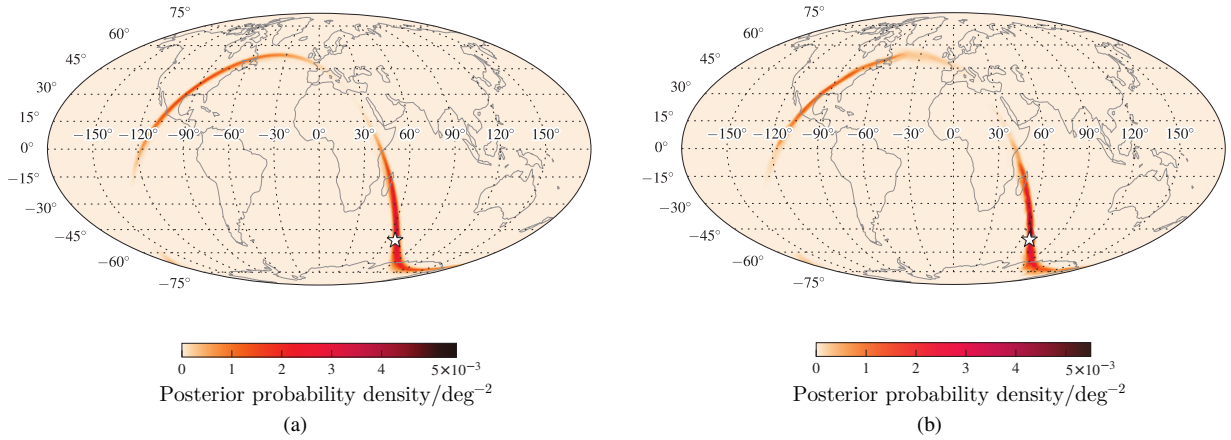


Figure 6. Posterior probability density for sky location, plotted in a Mollweide projection in geographic coordinates. The star indicates the true source location. (a) Computed by BAYESTAR. (b) Computed by LALINFERENCE. The event has simulation ID 1243 and a network SNR of $\varrho = 13.2$.

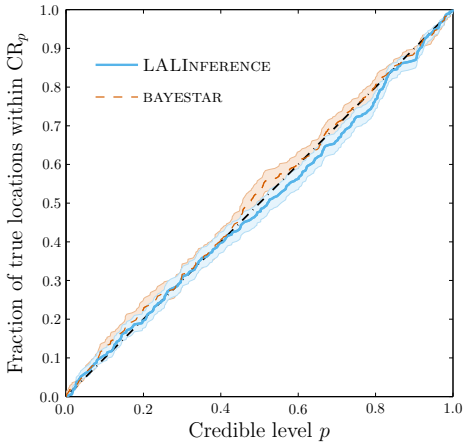


Figure 7. Fraction of true locations found within a credible region as a function of encompassed posterior probability. Results from LALINFERENCE are indicated by the solid line, results from BAYESTAR are indicated by the dashed line and the expected distribution is indicated by the dot–dashed diagonal line. The 68% confidencee interval is enclosed by the shaded regions, this accounts for sampling errors and is estimated from a beta distribution (Cameron 2011).

(Fairhurst 2009, 2011). Neither of these agree well. The sky-location posteriors can have complicated shapes, and cannot be accurately modelled by a simple Gaussian description.

To verify that SNR distribution is the dominant cause of difference between the Gaussian and recoloured results, we impose a cut on the recoloured data set of $\varrho \geq 12$ to match the Gaussian set. This reduces the number of events from 333 to 236. The cumulative distribution of sky-localization areas for results with $\varrho \geq 12$ are shown in figure 10. The distributions do overlap as expected: the Gaussian and recoloured results are in agreement (a KS test on $\mathrm{CR}_{0.9}$ gives a p-value of 0.550 when comparing LALINFERENCE results between noise realizations and 0.673 for BAYESTAR).

The key numbers describing the distributions are given in tables 1 and 2; the former gives the fraction of events with sky-localization areas smaller than fiducial values, and the latter gives median sky-localization areas. Our results are discussed further in section 5.1.

4.3. *Mass and distance estimation*

Independent of any EM counterpart, GW astronomy is still informative. GW observations allow for measurement of various properties of the source system. Here, we examine the ability to measure luminosity distance and mass (principally the chirp mass of the system).

Accurate mass and distance measurements have many physical applications. Measurement of the chirp mass distribution can constrain binary evolution models (Bulik & Belczynski 2003). Determining the maximum mass of a neutron star would shed light on its equation of state (e.g., Read et al. 2009), and, potentially, on the existence of a mass gap between neutron stars and black holes (Özel et al. 2010; Farr et al. 2011; Kreidberg et al. 2012). Combining mass and distance measurement, it may be possible to construct a new (independent) measure of the Hubble constant (Taylor et al. 2012). GW observations shall give us unique insight into the properties of BNS systems.

In addition to component masses and the distance to the source, the component spins are of astrophysical importance (e.g., Mandel & O'Shaughnessy 2010). Unfortunately, we cannot estimate the component spins as we are using non-spinning waveform templates. Measurement of the spins will be examined in a future study investigating PE using SpinTaylorT4 waveforms.

4.3.1. *Luminosity distance*

Quantifying the precision of distance estimation is simpler than for sky localization as we are now working in a single dimension. The equivalent of a credible region is a credible interval. We denote the distance credible interval for a total posterior probability p as CI_p^D. It is defined to exclude equal posterior probabilities in each of the tails; it is given by

$$\mathrm{CI}_p^D = C_D^{-1}\left[\frac{1+p}{2}\right] - C_D^{-1}\left[\frac{1-p}{2}\right], \tag{5}$$

where $C_D^{-1}(p)$ is the inverse of the cumulative distribution function

$$C_D(D) = \int_0^D \mathrm{d}D' P_D(D') \tag{6}$$

for distance posterior $P_D(D)$. The same symmetric definition for the credible interval was used by Aasi et al. (2013a). A smaller CI_p^D for a given p indicates more precise distance estimation.

The self-consistency of our distance estimates can be verified by calculating the fraction of true values that fall within

Table 1
Fractions of events with sky-localization areas smaller than a given size from this study using recoloured noise and Singer et al. (2014), which uses Gaussian noise. Results are quoted for the full catalogue of results with recoloured noise and imposing a SNR cut of $\varrho \geq 12$ to match the Gaussian catalogue. Figures for the 50% credible region $\mathrm{CR}_{0.5}$, the 90% credible region $\mathrm{CR}_{0.9}$ and the searched area A_* are included. A dash (—) is used for fractions less than 0.01.

		Gaussian noise		Recoloured noise		Recoloured noise $\varrho \geq 12$	
		BAYESTAR	LALINFERENCE	BAYESTAR	LALINFERENCE	BAYESTAR	LALINFERENCE
$\mathrm{CR}_{0.5} \leq$	5 deg^2	—	—	—	—	—	—
	20 deg^2	0.02	0.03	0.01	0.02	0.02	0.03
	100 deg^2	0.30	0.37	0.21	0.30	0.30	0.41
	200 deg^2	0.74	0.80	0.58	0.64	0.76	0.80
	500 deg^2	1.00	1.00	1.00	0.99	1.00	1.00
	1000 deg^2	1.00	1.00	1.00	1.00	1.00	1.00
$\mathrm{CR}_{0.9} \leq$	5 deg^2	—	—	—	—	—	—
	20 deg^2	—	—	—	—	—	—
	100 deg^2	0.03	0.04	0.02	0.03	0.03	0.04
	200 deg^2	0.10	0.13	0.06	0.08	0.09	0.12
	500 deg^2	0.44	0.48	0.31	0.38	0.44	0.52
	1000 deg^2	0.98	0.93	0.78	0.80	0.96	0.94
$A_* \leq$	5 deg^2	0.03	0.04	0.03	0.04	0.03	0.06
	20 deg^2	0.14	0.19	0.12	0.14	0.15	0.16
	100 deg^2	0.45	0.54	0.40	0.45	0.47	0.52
	200 deg^2	0.64	0.70	0.60	0.60	0.66	0.68
	500 deg^2	0.87	0.89	0.82	0.83	0.87	0.89
	1000 deg^2	0.97	0.99	0.96	0.95	0.98	0.97

Table 2
Median sky-localization areas from this study using recoloured noise, and Singer et al. (2014), which uses Gaussian noise. Results are quoted for the full catalogue of results with recoloured noise and imposing a SNR cut of $\varrho \geq 12$ to match the Gaussian catalogue. Figures for the 50% credible region $\mathrm{CR}_{0.5}$, the 90% credible region $\mathrm{CR}_{0.9}$ and the searched area A_* are included.

		Gaussian noise		Recoloured noise		Recoloured noise $\varrho \geq 12$	
		BAYESTAR	LALINFERENCE	BAYESTAR	LALINFERENCE	BAYESTAR	LALINFERENCE
Median	$\mathrm{CR}_{0.5}$	138 deg^2	124 deg^2	175 deg^2	154 deg^2	145 deg^2	118 deg^2
	$\mathrm{CR}_{0.9}$	545 deg^2	529 deg^2	692 deg^2	632 deg^2	524 deg^2	481 deg^2
	A_*	123 deg^2	88 deg^2	145 deg^2	132 deg^2	118 deg^2	88 deg^2

the credible interval at a given p. This is shown in figure 11 for results from both the Gaussian and recoloured noise results. Both distributions are consistent with expectations (performing a KS test with the predicted distribution yields p-values of 0.168 and 0.057 for the recoloured and Gaussian noise respectively). LALINFERENCE does return self-consistent distance estimates.

The cumulative distributions of credible intervals are plotted in figure 12. We divide the credible interval by the true (injected) distance $D_\star$; this gives an approximate analogue of twice the fractional uncertainty. The fractional uncertainty $\mathrm{CI}^D_p/D_\star$ appears insensitive to the detection cut-off (a KS test between $\mathrm{CI}^D_{0.9}/D_\star$ for the recoloured and Gaussian results gives a p-value of 0.077). This appears in contrast to the case for sky areas, but the differing SNR distributions are accounted for by scaling with respect to the distance (which is inversely proportional to the SNR). The estimation of the distance, like that for sky areas, does not depend upon the character of the noise.

Distance estimation is imprecise: the posterior widths are frequently comparable the the magnitude of the distance itself. This is a consequence of a degeneracy between the distance and the inclination (Cutler & Flanagan 1994; Aasi et al. 2013a). The key numbers summarising distance estimation are given in tables 3 and 4; the former gives the fraction of events with $\mathrm{CI}^D_p/D_\star$ smaller than fiducial values, and the latter gives median values.

Table 3
Fractions of events with fractional distance estimate uncertainties smaller than a given size. Results using recoloured noise and Gaussian noise are included (Singer et al. 2014). Figures for the 50% credible interval $\mathrm{CI}^D_{0.5}$ and the 90% credible interval $\mathrm{CI}^D_{0.9}$ are included, both are scaled with respect to the true distance $D_\star$. A dash (—) is used for fractions less than 0.01.

		Gaussian noise	Recoloured noise
$\frac{\mathrm{CI}^D_{0.5}}{D_\star} \leq$	0.25	0.04	0.03
	0.50	0.77	0.74
	0.75	0.95	0.93
	1.00	0.98	0.98
	2.00	1.00	1.00
$\frac{\mathrm{CI}^D_{0.9}}{D_\star} \leq$	0.25	—	—
	0.50	—	—
	0.75	0.40	0.35
	1.00	0.70	0.66
	2.00	0.96	0.97

4.3.2. *Chirp mass*

The chirp mass should be precisely measured as it determines the GW phase evolution. We again use the credible

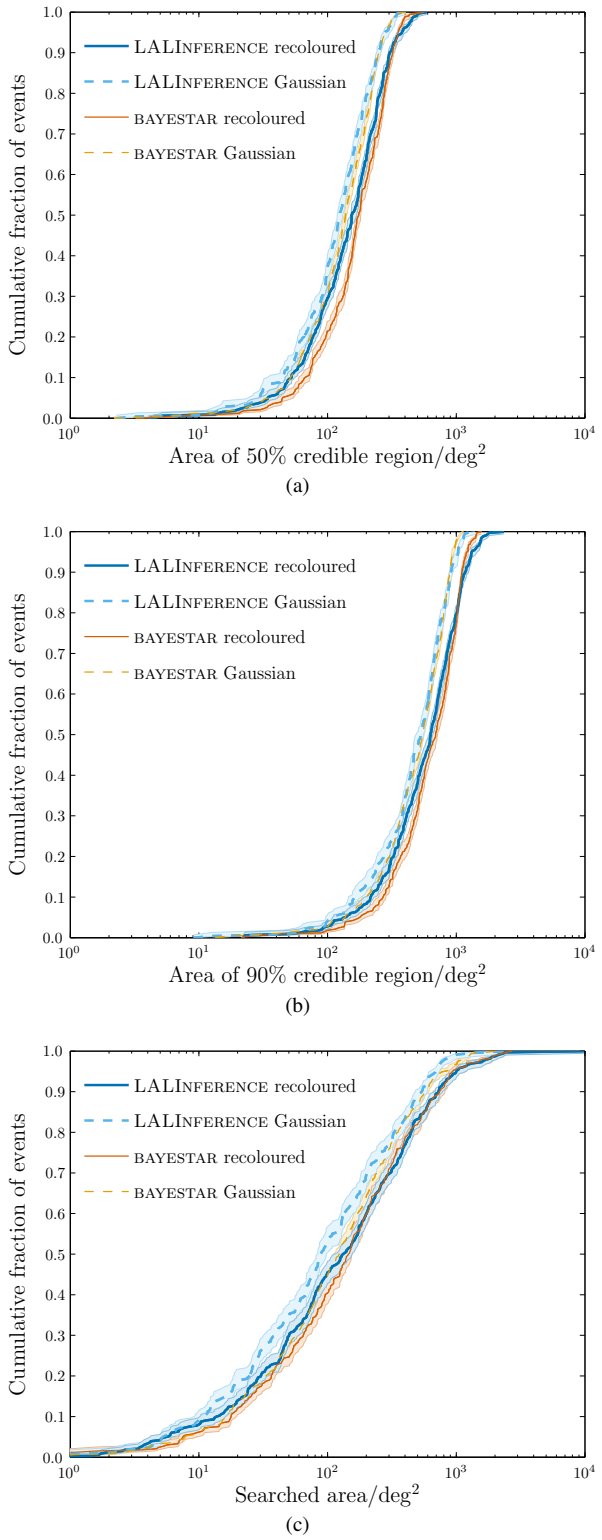


Figure 8. Cumulative fractions of events with sky-localization areas smaller than the abscissa value. (a) Sky area of 50% credible region $\mathrm{CR}_{0.5}$, the (smallest) area enclosing 50% of the total posterior probability. (b) Sky area of $\mathrm{CR}_{0.9}$. (c) Searched area $A_\star$, the area of the smallest credible region containing the true position. LALInference and BAYESTAR results are denoted by thicker blue and thinner red–orange lines respectively. The results of this study are indicated by a solid line, while the results of Singer et al. (2014), which uses Gaussian noise, are indicated by a dashed line. The 68% confidence intervals are denoted by the shaded areas.

Table 4
Median distance credible intervals (divided by the true distance) using recoloured noise and Gaussian noise (Singer et al. 2014). Figures for the 50% credible interval $\mathrm{CI}^D_{0.5}$ and the 90% credible interval $\mathrm{CI}^D_{0.9}$ are included.

		Gaussian noise	Recoloured noise
Median	$\mathrm{CI}^D_{0.5}/D_\star$	0.36	0.38
	$\mathrm{CI}^D_{0.9}/D_\star$	0.82	0.85

interval to quantify measurement precision; the chirp-mass credible interval $\mathrm{CI}^{\mathcal{M}_c}_p$ is defined equivalently to its distance counterpart in (5).

The fraction of true chirp masses that fall within $\mathrm{CI}^{\mathcal{M}_c}_p$ at a given p is plotted in figure 13. Neither the results calculated using Gaussian noise nor those using recoloured noise fit our expectations: the posteriors are not well calibrated. However, the two sets of results are entirely consistent with each other (a KS test between the two gives a p-value of 0.524), indicating that the PE is not affected by the noise. There appears to be a systematic error in our posterior distributions of the chirp mass.

The discrepancies between our posterior estimates for the chirp masses and their true values are a consequence of our use of non-spinning TaylorF2 waveform templates. This has two consequences. First, by using a non-spinning waveform, we do not explore the degeneracy between mass and spin (Cutler & Flanagan 1994; van der Sluys et al. 2008b; Baird et al. 2013). This results in an artificially narrow marginalized posterior for mass parameters such as the chirp mass. In effect, we are pinning the spin to be zero, which is information we should not have a priori. Second, we have used a template that does not exactly match the injected waveform (SpinTaylorT4). The small difference in approximants results in a mismatch in estimated parameters (Buonanno et al. 2009; Aasi et al. 2013a). Since the posterior on the chirp mass is narrow, because it is intrinsically well-measured and because we have not included degeneracy with spin, even a small difference in templates is sufficient to offset the posterior from the true chirp mass by a statistically significant amount.

To examine the offset between the estimated and true chirp masses, we plot in figure 14 the difference between the posterior mean $\bar{\mathcal{M}}_\mathrm{c}$ and the true value $\mathcal{M}_\star$ divided by the standard deviation of the posterior $\sigma_{\mathcal{M}_\mathrm{c}}$. Using the median in place of the mean, or $\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.68}/2$ in place of $\sigma_{\mathcal{M}_\mathrm{c}}$, gives only a small quantitative difference. Over this narrow mass range, the offset is not a strong function of the chirp mass. The offset is a combination of both error introduced by the presence of noise and theoretical error from the mismatch between the injected waveform and template waveforms (Cutler & Vallisneri 2007). If only the former were significant, we would expect the mean offset to be zero, and the typical scatter of offsets to be of order of the posterior's standard deviation. Neither of these is the case. The average scaled offset $(\bar{\mathcal{M}}_\mathrm{c}-\mathcal{M}_\star)/\sigma_{\mathcal{M}_\mathrm{c}}$ across the recoloured (Gaussian) data set is -1.3 ± 0.1 (-0.9 ± 0.1). This shows that there is a systematic error. However, it is not as simple as just systematically underestimating the chirp mass; there is a large scatter in the offsets, the standard deviation of the scaled offset for the recoloured (Gaussian) data set is 2.07 ± 0.08 (2.09 ± 0.09). This is consistent with our expectation that the mass–spin degeneracy should broaden the posterior; these results imply that the posterior should be a

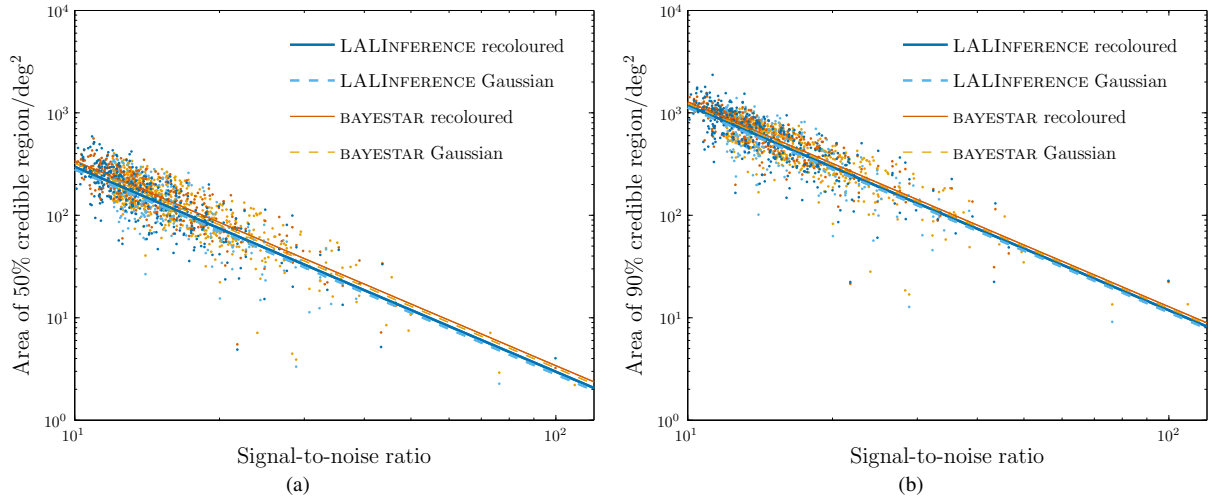


Figure 9. Sky-localization areas as a function of signal-to-noise ratio ϱ. (a) Sky area of 50% credible region $\mathrm{CR}_{0.5}$. (b) Sky area of $\mathrm{CR}_{0.9}$. Individual results are indicated by points. We include simple best-fit lines assuming that the area $A \propto \varrho^{-2}$. LALINFERENCE and BAYESTAR results are denoted by thicker blue and thinner red–orange lines respectively. The results of this study are indicated by a solid line, while the results of Singer et al. (2014), which uses Gaussian noise, are indicated by a dashed line.

factor of ~ 2 wider (cf. Poisson & Will 1995).

While the theoretical error is important in determining the accuracy to which we can infer the chirp mass, it does not completely dominate the noise error. To illustrate the scale of the errors, we plot distribution of the 50% and 90% credible intervals in figures 15(a) and 15(b), and the absolute magnitudes of the offsets in figure 15(c). For a well calibrated posterior, we would expect the offset to be smaller than $\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.9}/2$ ($\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.5}/2$) in approximately 90% (50%) of events. Figure 13 shows that this is not the case, that we do have systematic error. Figure 14 confirms this and shows that the theoretical error is of a comparable size to the noise error. In figure 15, we see that the presence of theoretical error does not radically affect the distribution of offsets. The median value of the offsets are $(2.6 \times 10^{-4})M_\odot$ and $(2.4 \times 10^{-4})M_\odot$, and the median values of $\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.5}/2$ are $(1.2 \times 10^{-4})M_\odot$ and $(1.3 \times 10^{-4})M_\odot$ for the recoloured and Gaussian data sets respectively; the theoretical error approximately doubles the total uncertainty on the chirp mass. The key numbers summarising the distributions are given in tables 5 and 6, which give the fraction of events with uncertainties smaller than fiducial values and the median uncertainties respectively.

Furthermore, figure 15 shows that the (in)ability to measure the chirp mass is not significantly influenced by the character of the noise or the detection threshold used (a KS test comparing the $\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.9}$ and $|\bar{\mathcal{M}}_\mathrm{c} - \mathcal{M}_\star|$ distributions between the Gaussian and recoloured data sets gives p-values of 0.805 and 0.507 respectively). The latter is a consequence of both thresholds recovering equivalent chirp-mass distributions (figure 4).

It should be possible to incorporate knowledge of theoretical waveform error into PE by marginalizing out the uncertainty. This can be done using parametric models for the uncertainty if a specific form of the waveform error is suspected, or non-parametrically if we wish to be agnostic. The effect of folding in this additional uncertainty is to broaden the posteriors and possibly shift their means; doing so should make posterior estimates consistent with the true values.

While we cannot correctly reconstruct the posterior distribution for the chirp mass, the error in the estimate is still small. We can measure the chirp mass accurately, even though we are affected by systematic error.

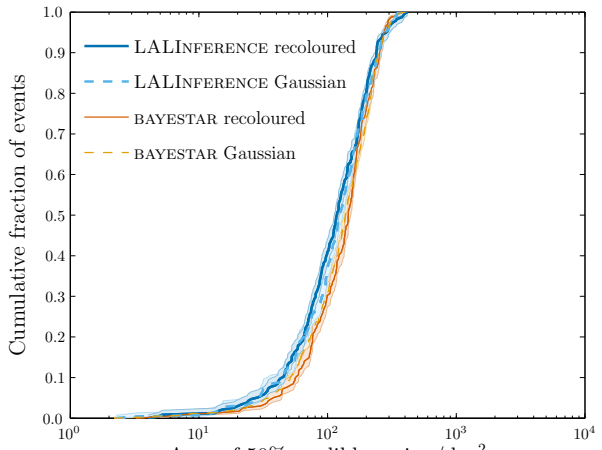
Table 5
Fractions of events with chirp-mass estimate errors smaller than a given value. Results using recoloured noise and Gaussian noise are included (Singer et al. 2014). Included are figures for the 50% credible interval $\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.5}$ and the 90% credible interval $\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.9}$, which only include statistical error from the noise, and for the posterior mean offset relative to the true chirp mass $|\bar{\mathcal{M}}_\mathrm{c} - \mathcal{M}_\star|$, which includes both noise error and theoretical error. A dash (—) is used for fractions less than 0.01.

		Gaussian noise	Recoloured noise
$\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.5} \leq$	$(5 \times 10^{-5})M_\odot$	—	—
	$(1 \times 10^{-4})M_\odot$	0.05	0.03
	$(2 \times 10^{-4})M_\odot$	0.34	0.33
	$(5 \times 10^{-4})M_\odot$	0.89	0.88
	$(1 \times 10^{-3})M_\odot$	1.00	0.99
	$(2 \times 10^{-3})M_\odot$	1.00	1.00
$\mathrm{CI}^{\mathcal{M}_\mathrm{c}}_{0.9} \leq$	$(5 \times 10^{-5})M_\odot$	—	—
	$(1 \times 10^{-4})M_\odot$	—	—
	$(2 \times 10^{-4})M_\odot$	0.01	0.02
	$(5 \times 10^{-4})M_\odot$	0.29	0.29
	$(1 \times 10^{-3})M_\odot$	0.77	0.79
	$(2 \times 10^{-3})M_\odot$	1.00	0.97
$\lvert\bar{\mathcal{M}}_\mathrm{c} - \mathcal{M}_\star\rvert \leq$	$(5 \times 10^{-5})M_\odot$	0.09	0.11
	$(1 \times 10^{-4})M_\odot$	0.20	0.21
	$(2 \times 10^{-4})M_\odot$	0.42	0.41
	$(5 \times 10^{-4})M_\odot$	0.83	0.79
	$(1 \times 10^{-3})M_\odot$	0.98	0.96
	$(2 \times 10^{-3})M_\odot$	1.00	1.00

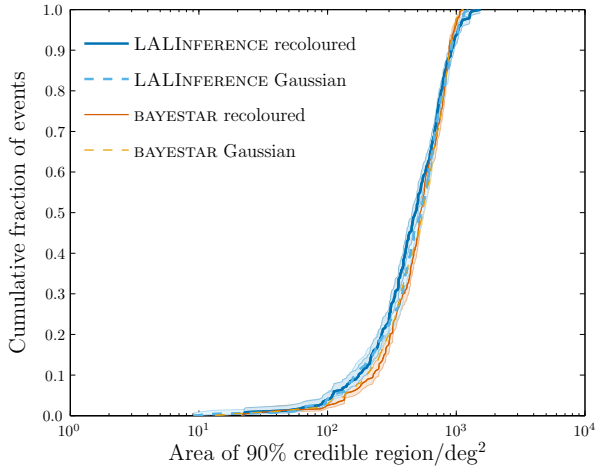
4.3.3. *Component masses*

The chirp mass is a combination of the component masses; in some cases it can be used to infer whether the source is a BNS or a binary black-hole system (Hannam et al. 2013; Vitale & Del Pozzo 2014), but the component masses are of greater interest. The mass–spin degeneracy affects our ability to construct accurate estimates for the individual masses. Since we have already seen a systematic error in the chirp mass, we expect an analogous (larger) phenomenon here.

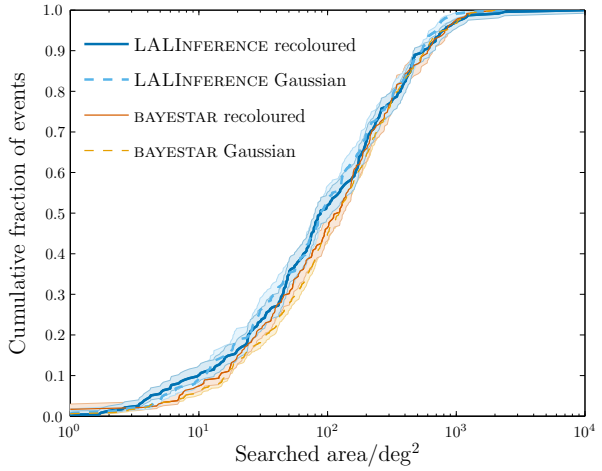
We are again working in two dimensions, so we use credible regions to quantify PE precision. The mass-space credible



(a)



(b)



(c)

Figure 10. Cumulative fractions of events with sky-localization areas smaller than the abscissa value as in figure 8 but imposing an SNR cut of $\varrho_R \geq 12$. (a) Sky area of $CR_{0.5}$. (b) Sky area of $CR_{0.9}$. (c) Searched area A_*. LALInference and BAYESTAR results are denoted by thicker blue and thinner red–orange lines respectively. The results of this study are indicated by a solid line, while the results of Singer et al. (2014), which uses Gaussian noise, are indicated by a dashed line. The 68% confidence intervals are denoted by the shaded areas.

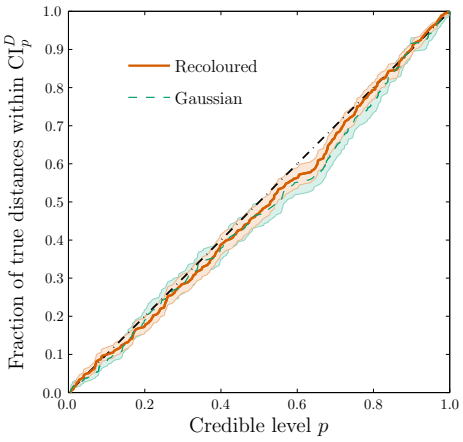


Figure 11. Fraction of true luminosity distances found within a credible interval as a function of encompassed posterior probability. Results using recoloured noise are indicated by a solid line, while the results using Gaussian noise (Singer et al. 2014) are indicated by a dashed line. The expected distribution is indicated by the dot–dashed diagonal line. The shaded regions enclose the 68% confidence intervals accounting for sampling errors.

Table 6
Median chirp mass credible intervals and posterior estimate offset using recoloured noise and Gaussian noise (Singer et al. 2014). Included are figures for the 50% credible interval $CI_{0.5}^{\mathcal{M}_c}$ and the 90% credible interval $CI_{0.9}^{\mathcal{M}_c}$, and the posterior mean offset relative to the true value $|\bar{\mathcal{M}}_c - \mathcal{M}_\star|$.

		Gaussian noise	Recoloured noise
Median	$CI_{0.5}^{\mathcal{M}_c}$	$(2.6 \times 10^{-4})M_\odot$	$(2.5 \times 10^{-4})M_\odot$
	$CI_{0.9}^{\mathcal{M}_c}$	$(6.4 \times 10^{-4})M_\odot$	$(6.4 \times 10^{-4})M_\odot$
	$\lvert\bar{\mathcal{M}}_c - \mathcal{M}_\star\rvert$	$(2.4 \times 10^{-4})M_\odot$	$(2.6 \times 10^{-4})M_\odot$

region CR_p^{m1-m2} is defined analogously to its sky-area counterpart in (2); it is easier to compute as we do not have to contend with the spherical geometry of the sky or with as intricate posterior distributions. We plot in figure 16 the fraction of injected masses that fall within $CR_p^{m_1-m_2}$ at a given p. As for the chirp mass, the posterior is not well calibrated, approximately 40% (38% for results with recoloured noise and 42% for Gaussian) of the true component masses lie altogether outside the range of the estimated posterior, but the two sets of results are consistent with each other (performing a KS test gives a p-value of 0.969). We cannot accurately reconstruct the component masses using our non-spinning waveforms.

To give an indication of the scale of the uncertainty in m_1–m_2 space, we plot the 90% credible region in figure 17. Since our estimates for the component masses are inaccurate, with many true values lying outside the posterior, CR_p^{m1-m2} is a lower bound on the typical scale for measurement accuracy. This does not reflect how well we can actually measure the component masses, to produce accurate estimates, we must include the mass–spin degeneracy which broadens the posterior.

It is apparent that a statement regarding measurement of component masses must wait until an analysis is done using waveforms that include spin. We will return this question in a future publication.

5. DISCUSSION AND CONCLUSIONS

5.1. *Observing scenarios*

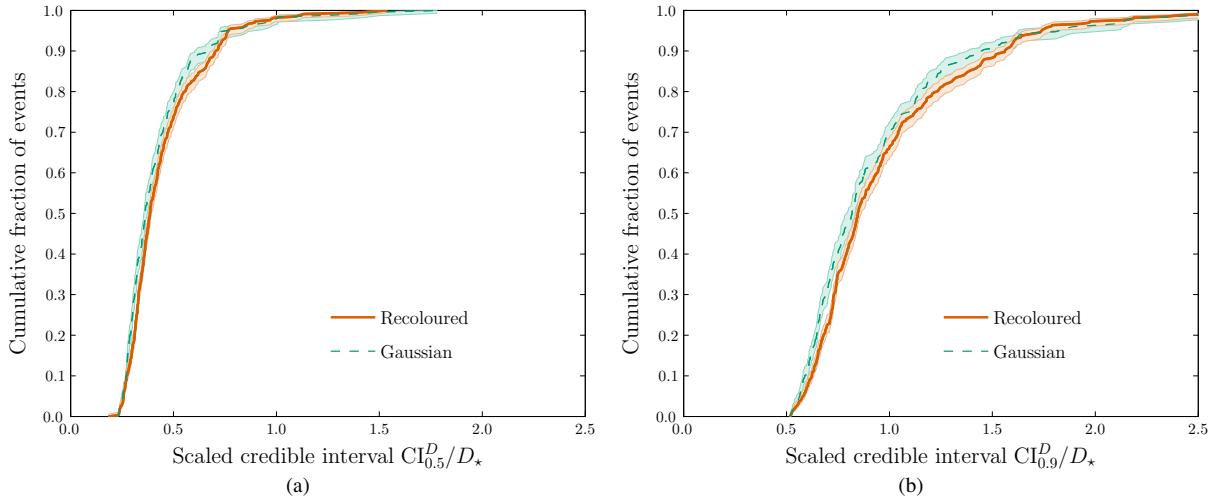


Figure 12. Cumulative fractions of events with luminosity-distance credible intervals (divided by the true distance) smaller than the abscissa value. (a) Scaled 50% credible interval $\mathrm{CI}^D_{0.5}/D_\star$. (b) Scaled 90% interval $\mathrm{CI}^D_{0.9}/D_\star$. Results using recoloured noise are indicated by a solid line and the results using Gaussian noise (Singer et al. 2014) are indicated by a dashed line. The 68% confidence intervals are denoted by the shaded areas.

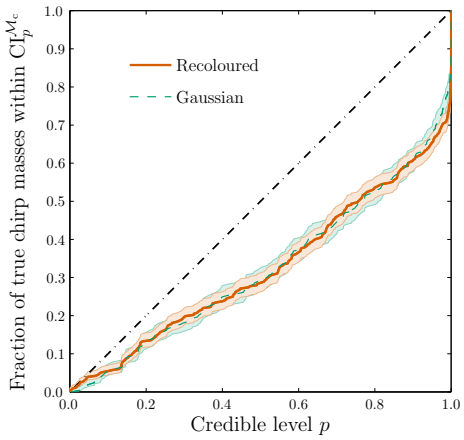


Figure 13. Fraction of true source chirp masses found within a credible interval as a function of encompassed posterior probability. Results using recoloured noise are indicated by a solid line, while the results using Gaussian noise (Singer et al. 2014) are indicated by a dashed line. The expected distribution is indicated by the dot–dashed diagonal line. The shaded regions enclose the 68% confidence intervals accounting for sampling errors.

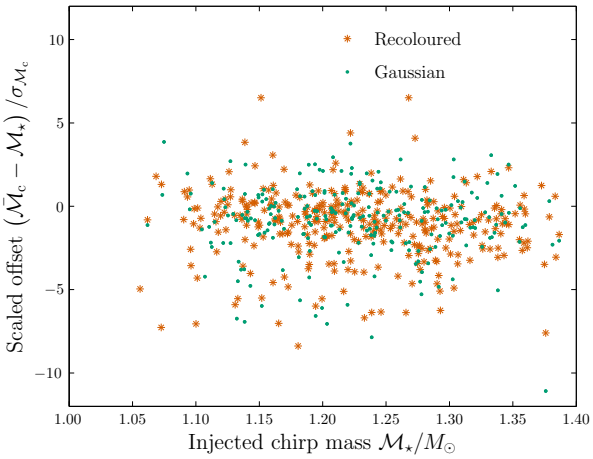


Figure 14. Offset between the posterior mean estimate for the chirp mass $\bar{\mathcal{M}}_c$ and the true (injected) value $\mathcal{M}_\star$ divided by the standard deviation of the posterior distribution $\sigma_{\mathcal{M}_c}$. The round (green) points are for the results using Gaussian noise (Singer et al. 2014) and the star-shaped (red) points are for results using recoloured noise.

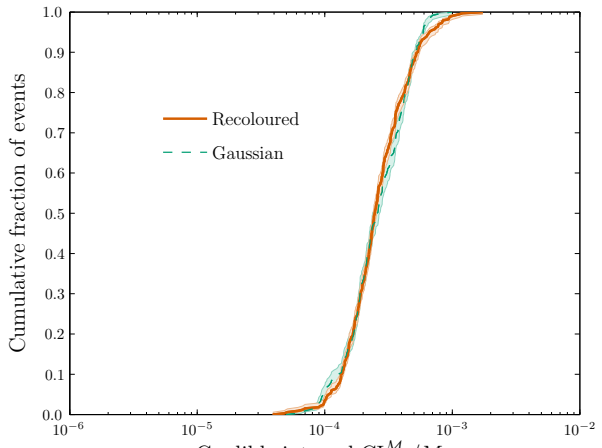
Having determined the sky-localization accuracy expected for O1, we now use our results to compare with current predictions for observing scenarios in the advanced-detector era. In section 5.1.1 we consider the two-detector network of O1. In section 5.1.2 we extend our discussion to consider predictions for sky-localization in subsequent observing runs using a three-detector network.

5.1.1. *Two-detector sky-localization accuracy*

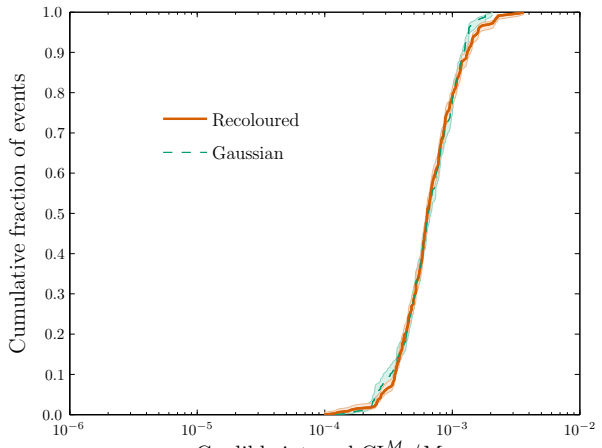
Prospects for sky localization in the advanced-detector era are specified by Aasi et al. (2013b). This states that any events detected in 2015 would not be well localized. This has been shown to not be the case (e.g., Nissanke et al. 2011; Kasliwal & Nissanke 2014; Singer et al. 2014). We see that while only a small fraction of events have well-localized sources, this fraction is non-zero. The 90% credible region is almost always smaller than 10^3 deg^2. The 2015 observing scenario of Aasi et al. (2013b) does not give any figures for potential sky-localization accuracy, but we can now be specific using the results of this work.

The sky-localization figures currently included in Aasi et al. (2013b) are calculated using TT (Fairhurst 2009, 2011). This is a convenient means of predicting sky-localization accuracy; it is not a method used to reconstruct the sky-position posterior of detected signals. For a two-detector network, triangulation predicts an unbroken annulus on the sky. The area of this ring linearly scales with the uncertainty on the timing measurement, which is inversely proportional to the SNR. Our results show that, when using a coherent Bayesian approach, the recovered sky area is not (always) a ring, see figure 6, and the area scales inversely with the square of the SNR (Raymond et al. 2009); see Singer et al. (2014) for illustrations. Hence, TT is a poor fit in this case.

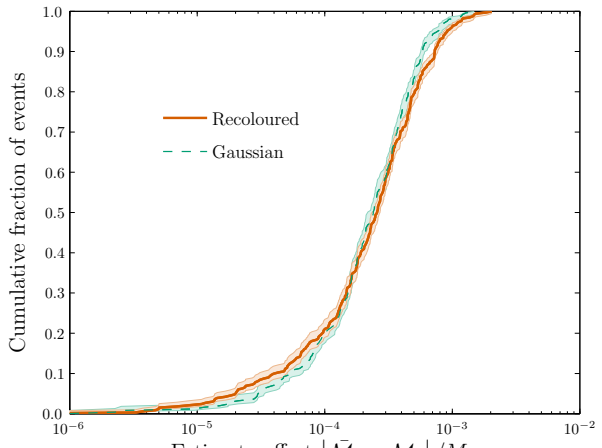
In figure 18 we plot the ratio of the predicted credible region calculated using TT, to the actual credible region calculated using LALINFERENCE PE. We include predictions from both standard TT and also TT including phase coherence (Grover et al. 2014). The former method estimates timing accuracy (and hence the width of the sky annulus) as a function of



(a)



(b)



(c)

Figure 15. Cumulative fractions of events with (a) 50% chirp-mass credible interval, (b) 90% credible interval, and (c) offsets between the posterior mean and true chirp mass smaller than the abscissa value. Results using recoloured noise are indicated by a solid line and the results using Gaussian noise (Singer et al. 2014) are indicated by a dashed line. The 68% confidence intervals are denoted by the shaded areas.

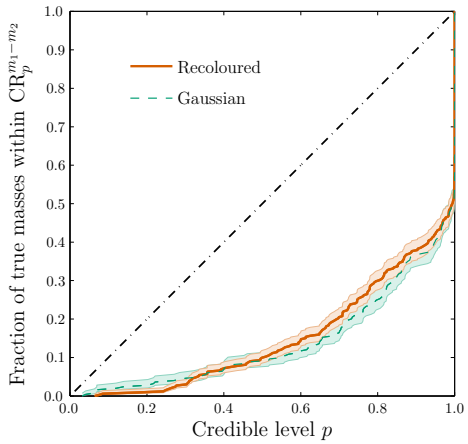


Figure 16. Fraction of true source component masses (m_1, m_2) found within a credible region as a function of encompassed posterior probability. Results using recoloured noise are indicated by a solid line, while the results using Gaussian noise (Singer et al. 2014) are indicated by a dashed line. The expected distribution is indicated by the dot–dashed diagonal line. The shaded regions enclose the 68% confidence intervals accounting for sampling errors.

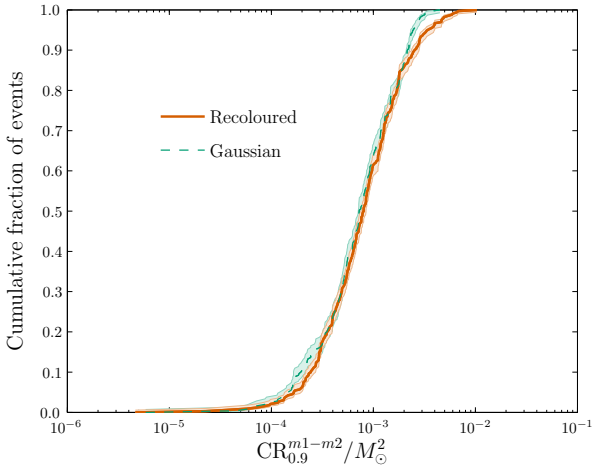


Figure 17. Cumulative fractions of events with m_1–m_2 90% credible regions smaller than the abscissa value. Results using recoloured noise are indicated by a solid line and the results using Gaussian noise (Singer et al. 2014) are indicated by a dashed line. The 68% confidence intervals are denoted by the shaded areas. These results show the typical posterior width using non-spinning waveforms, the failure to include the mass–spin degeneracy means that these posteriors are too narrow.

the SNR and detector bandwidth.[17] The latter method introduces the requirement of phase consistency between detectors, which can significantly aid source localization. These effects are modelled via a correction factor, whose value depends on how marginalization over polarization is taken into account. Here, we use the larger of the two correction factors proposed in Grover et al. (2014), their equation (16), although the degeneracy between phase and polarization means that the correction factor is probably too large for the 2-detector network. The time and phase method does better, but neither technique does a good job at matching the true localization: both are too pessimistic. Agreement worsens at higher SNR as a consequence of the different SNR scalings. We cannot naively use TT to predict sky-localization accuracy for a two-detector network.

[17] In calculating these values we have corrected typos in both equation (28) of Fairhurst (2009), where the prefactor should be $\sqrt{2}\,\mathrm{erf}^{-1}(0.9) \approx 1.65$ rather than 3.3, and equation (15) of Fairhurst (2011), which has an unnecessary factor of D.

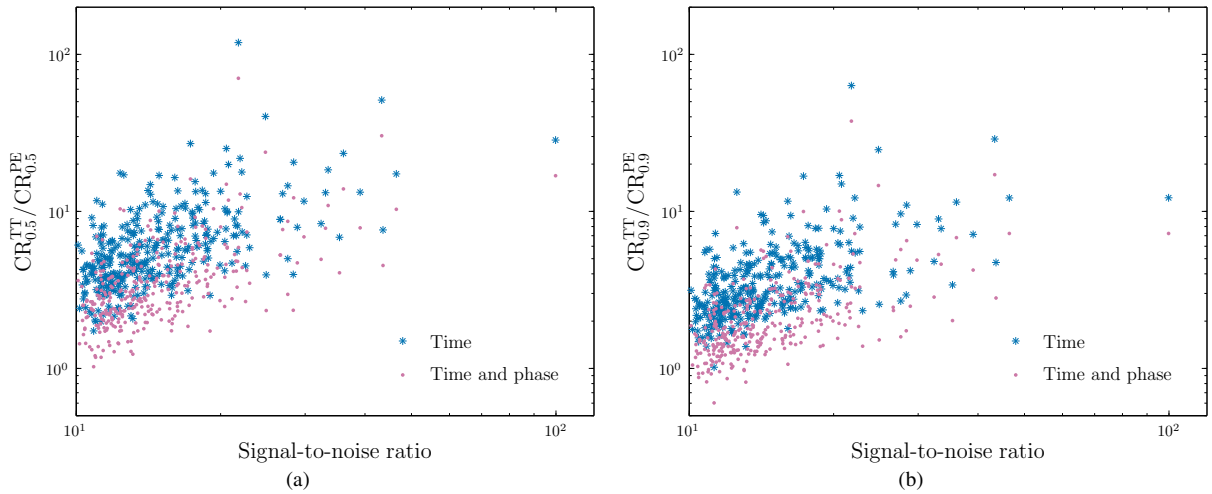


Figure 18. Ratio of the area of credible regions calculated using TT and PE as a function of the SNR. (a) Ratio of 50% credible regions $\mathrm{CR}_{0.5}$. (b) Ratio of $\mathrm{CR}_{0.9}$. TT results are calculated using just time of arrivals (Fairhurst 2009, 2011), indicated by the star-shaped (blue) points, and by also including phase coherence (Grover et al. 2014), indicated by the round (purple) points. PE results are calculated from the posteriors returned by LALINFERENCE.

We have found that sky areas recovered during O1 are likely to be hundreds of square degrees. Covering such a large area to sufficient depth to detect the most plausible EM counterparts ($r \gtrsim 22$–26 mag; Metzger & Berger 2012; Barnes & Kasen 2013; Metzger et al. 2014) is challenging for current EM observatories (Kasliwal & Nissanke 2014); furthermore, posterior distributions for the sky location are commonly multimodal or feature long, narrow arcs making them awkward to cover. It will be necessary to carefully consider how to most efficiently point telescopes to maximise the probability of observing a counterpart; using galaxy catalogues could be one means of increasing the chance of imaging an EM counterpart (Nuttall & Sutton 2010; Hanna et al. 2014; Bartos et al. 2014; Fan et al. 2014).

5.1.2. *Three-detector sky-localization accuracy*

For 2016 onwards, we expect that AdV would also be in operation. The addition of a third detector should significantly improve sky-localization accuracy (Singer et al. 2014).

Aasi et al. (2013b) give figures for sky-localization accuracies in the three-detector era. In 2016, Aasi et al. (2013b) predicts that 2% (5–12%) of BNS detections shall be localized within 5 deg^2 (20 deg^2) at 90% confidence. These values are calculated from TT. Ideally, we would like to compare these to results using Bayesian PE using recoloured noise, but performing three-detector PE runs for later observing periods is outside the range of this study. However, we have demonstrated that the properties of the noise do not impact sky-localization accuracies, provided that the chosen detection threshold yields similar SNR distributions in all cases. Consequently, we can use the three-detector, Gaussian-noise LALINFERENCE results of Singer et al. (2014) as a reference. For comparison, they find that 2% (14%) of events have $\mathrm{CR}_{0.9}$ smaller than 5 deg^2 (20 deg^2). PE with LALINFERENCE provides more optimistic sky-localization accuracies than TT.

In figure 19 we compare the three-detector results of Singer et al. (2014) to the equivalent results calculated using TT. These results are for 2016, assuming the mid noise curve of Barsotti & Fritschel (2012) for the aLIGO detectors, and the geometric mean of the high and low bounds of the early curve of Aasi et al. (2013b) for the Virgo interferometer. Both triangulation and PE produce sky areas that scale with ρ^{-2}, such that their ratio shows no significant trend with SNR, although the scatter seems to decrease as SNR increases.

Comparing the entire population of points, we can calculate average values, which are given in table 7. We consider the logarithm of the ratio, which should be $\log_{10}(1) = 0$ for perfect agreement. The median $\log_{10}(\mathrm{CR}^{\mathrm{TT}}_{0.5}/\mathrm{CR}^{\mathrm{PE}}_{0.5})$ using only time of arrival is 0.61, in complete agreement with the findings of Grover et al. (2014); using time and phase, the median value is 0.13. The TT and PE results have different ratios $\mathrm{CR}_{0.9}/\mathrm{CR}_{0.5}$. The mean value of $\log_{10}(\mathrm{CR}^{\mathrm{PE}}_{0.9}/\mathrm{CR}^{\mathrm{PE}}_{0.5})$ is approximately 0.64 and the standard deviation is 0.13; again (see section 4.2), this does not fit well with a Gaussian model. The 90% credible regions for triangulation and PE are in better agreement with each other, with the time-and-phase triangulation average areas consistent with those from LALINFERENCE. The time-and-phase method produces a reasonable estimate when averaged over the entire population. However, for individual events there is large scatter because TT models are purely predictive and do not take into account the actual data realization.

Despite the good average agreement, there is a large tail of events at low SNRs where credible regions are too small, and the results suggest that at high SNRs the credible regions may be too large; this may introduce errors when considering the sub-populations of the best localized or worst localized events (or if the distribution of events is significantly different from that considered here). Given all these findings, we can be confident that the TT results of Aasi et al. (2013b) are overly pessimistic.

There remains one further caveat before we can state that the sky-localization accuracies of Aasi et al. (2013b) should be revised to give better results. We have seen that using a realistic FAR cut allows us to detect signals with $\varrho < 12$. These low-SNR results shift the distribution of sky-localization accuracies, such that the performance appears worse. Thus, while we can be confident that the events currently included should have a better accuracy than assumed for Aasi et al. (2013b), the total population of detectable events is potentially larger than previously estimated, and may include some low-SNR events with poorer localization.

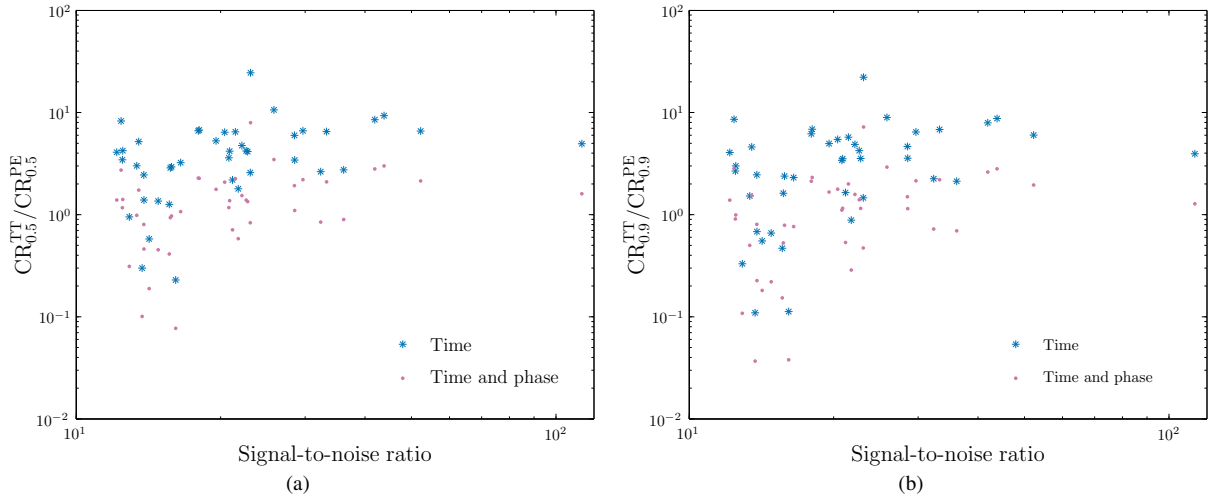


Figure 19. Ratio of the area of credible regions calculated as a function of the SNR as in figure 18, but for a three-detector network as expected in 2016. (a) Ratio of $\mathrm{CR}_{0.5}$. (b) Ratio of $\mathrm{CR}_{0.9}$. TT results are calculated using just time of arrivals (Fairhurst 2009, 2011), indicated by the star-shaped (blue) points, and by also including phase coherence (Grover et al. 2014), indicated by the round (purple) points. PE results with Gaussian noise are calculated from the posteriors returned by LALINFERENCE (Singer et al. 2014).

Table 7
Average values of the logarithm of the ratio of credible regions calculated using TT to those calculated from PE $\log_{10}(\mathrm{CR}^{\mathrm{TT}}_p/\mathrm{CR}^{\mathrm{Full}}_p)$. TT results are calculated using just time of arrivals (Fairhurst 2009, 2011) and by also including phase coherence (Grover et al. 2014). PE results with Gaussian noise are calculated from the posteriors returned by LALINFERENCE (Singer et al. 2014).

Triangulation method	p	Mean	Median	Standard deviation
Time only	0.5	0.53	0.61	0.39
	0.9	0.42	0.55	0.49
Time and phase	0.5	0.05	0.13	0.39
	0.9	–0.07	0.07	0.49

5.2. *Summary*

We provide realistic prospects for sky localization and EM follow-up of CBC sources in the O1 era by simulating a search for BNS sources with a 2-detector aLIGO network at anticipated 2015 sensitivity. Our analysis is designed to be as similar as possible to recent work investigating sky-localization capability in the first two years of the advanced-detector era (Singer et al. 2014). That study assumed Gaussian noise whereas our analysis incorporates more realistic noise, using real data from the S6 observing period recoloured to the anticipated 2015 noise spectrum.

We use the same list of simulated BNS sources as previously used in Singer et al. (2014). The simulated events are passed through the GSTLAL_INSPIRAL data-analysis pipeline which will be used online in O1. Detection triggers from this search with a FAR of $\leq 10^{-2}$ yr^{-1} are then followed up with sky-localization and PE codes.

The pipeline should not significantly distort the population of signals detected compared with the astrophysical population. There appears to be no selection based upon BNS spin. There is a selection effect determined by the chirp mass (systems with smaller chirp masses are harder to detect), but this translates to only a small difference for a small number ($\lesssim 10^2$) of detections.

Comparison of sky-localization areas from BAYESTAR and LALINFERENCE demonstrates that while the former only uses a selection of the information available and employs a number of approximations, it does successfully reconstruct sky position. Furthermore, BAYESTAR does this with sufficiently low latency to be of use for rapid EM follow-up.

Rapid sky-localization with BAYESTAR takes on average 900 s of CPU time per event (appendix B). If it is parallelized in a 32-way configuration (the baseline for online analysis), this correspond to a wall time of 30 s. None of our runs would take longer than 60 s to complete.

PE using LALINFERENCE_NEST with (non-spinning) TaylorF2 waveforms requires a total CPU time of $\sim 2 \times 10^6$ s per event (appendix B). Five CPUs were used for each LALINFERENCE_NEST run, hence the wall time, as a first approximation, can be estimated as ~ 100 hr. These PE results can be produced within a few days, although with more expensive waveforms, the time taken is longer. Ongoing technical improvements should reduce the computational cost in the near future (Veitch et al. 2014).

Considering sky-localization, the median area of $\mathrm{CR}_{0.9}$ ($\mathrm{CR}_{0.5}$) as estimated by LALINFERENCE is 632 deg^2 (154 deg^2), and the median searched area is 132 deg^2. LALINFERENCE finds that 2% of events have $\mathrm{CR}_{0.5}$ smaller than 20 deg^2; fewer than 1% of events have $\mathrm{CR}_{0.5}$ smaller than 5 deg^2 or $\mathrm{CR}_{0.9}$ smaller than 20 deg^2, but 14% of events have searched areas smaller than 20 deg^2 and 4% have searched areas smaller than 5 deg^2. These are worse than predicted using Gaussian noise because of the inclusion of more low-SNR events, but if these additional events are excluded, the results calculated using both types of noise are in agreement. The character of the noise does not noticeably affect sky-localization accuracy, and sky areas are consistent if the same SNR threshold is applied.

The 2015 observing scenario of Aasi et al. (2013b) currently states that any events detected would not be well localized. This is not the case, although recovered areas are still large.

While Aasi et al. (2013b) does not have sky-localization figures for 2015, it does have them for later years. These are calculated using a time triangulation method (Fairhurst 2009, 2011). The Gaussian results of Singer et al. (2014) show that we can achieve better sky localization than expected from TT

alone; this improvement can principally be explained by the incorporation of phase consistency (Grover et al. 2014). Hence, the figures in (Aasi et al. 2013b) may be pessimistic. However, from this study we also know that results using Gaussian noise are liable to be optimistic because they exclude events by using of a detection threshold of $\varrho \geq 12$; in practise, when using a FAR threshold, there is a tail of lower SNR events that skew the distribution. This must be accounted for when quoting the fraction of events located to within a given area. Therefore, updating the numbers in the observing scenarios for later years is not straightforward.

The LALINFERENCE runs also return posteriors for other parameters. We looked at the source luminosity distance, the chirp mass and the component masses. The distance is not well measured; the median $\mathrm{CI}^{D}_{0.9}/D_{*}$ ($\mathrm{CI}^{D}_{0.5}/D_{*}$) is 0.85 (0.38). As a consequence of our use of non-spinning waveform templates that do not exactly match the injected waveforms, the chirp-mass estimates are subject to theoretical error of a size roughly equal to the uncertainty introduced by the noise. This means our posteriors are not well calibrated: they are both (on average) offset from the true position and too narrow (by a factor of $\sim 1/2$). Using spinning waveforms, such that the mass–spin degeneracy can be explored, will broaden the posteriors and resolve this problem, but we will always face a potential systematic bias unless we exactly know the true waveforms of Nature. Despite the systematic effects, the posterior mean of the chirp-mass distribution is within $10^{-3}M_{\odot}$ of the true chirp mass in 96% of events, and the median absolute difference between the two is $(2.6 \times 10^{-4})M_{\odot}$. A larger difference could occur if there is a larger discrepancy between the waveform template and the true waveform, but we expect it to be of a similar order of magnitude. While we can still accurately measure the chirp mass using non-spinning waveforms, the same does not apply for component masses. Estimates for these must be performed using spinning waveforms; we shall examine this in a future study. The character of the noise does not influence PE accuracy.

The authors are grateful for useful suggestions from the CBC group of the LIGO–Virgo Science Collaboration and in particular Yiming Hu.

This work was supported by the Science and Technology Facilities Council. PBG acknowledges NASA grant NNX12AN10G. SV acknowledges the support of the National Science Foundation and the LIGO Laboratory. JV was supported by STFC grant ST/K005014/1. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0757058.

Results were produced using the computing facilities of the LIGO DataGrid including: the Nemo computing cluster at the Center for Gravitation and Cosmology at the University of Wisconsin–Milwauke under NSF Grants PHY-0923409 and PHY-0600953; the Atlas computing cluster at the Albert Einstein Institute, Hannover; the LIGO computing clusters at Caltech, and the facilities of the Advanced Research Computing @ Cardiff (ARCCA) Cluster at Cardiff University. We are especially grateful to Paul Hopkins of ARCCA for assistance.

Some results were produced using the post-processing tools of the `plotutils` library at `http://github.com/farr/plotutils`, and some were derived using HEALPix (Gorski et al. 2005).

This paper is has been assigned LIGO document reference LIGO-P1400232. It contains some results originally included in LIGO technical report LIGO-T1400480.

APPENDIX

A. DETECTION AND COMPONENT MASSES

In section 4.1.2, we examined selection effects of the detection pipeline. In particular, we looked at the detected distribution of chirp masses as this sets the GW amplitude. The magnitude of the selection effect depends on the details of the chirp-mass distribution, but can be estimated using a simple model. For low-mass signals whose inspiral spans the sensitive band of the detector, the amplitude of the waveform is proportional to $\mathcal{M}_{\mathrm{c}}^{5/6}$ (Sathyaprakash & Schutz 2009). The sensitive volume is proportional to the cube of this, or $\mathcal{M}_{\mathrm{c}}^{5/2}$. Suppose that half of the injections are made at a chirp mass of $\bar{\mathcal{M}}_{\mathrm{c}} - \delta\mathcal{M}_{\mathrm{c}}$ and the other half at a chirp mass value of $\bar{\mathcal{M}}_{\mathrm{c}} + \delta\mathcal{M}_{\mathrm{c}}$, with $\delta\mathcal{M}_{\mathrm{c}} \ll \bar{\mathcal{M}}_{\mathrm{c}}$. Then the expected fraction of higher-mass systems among all detected systems is

$$\begin{aligned}\mathcal{F}_{\mathrm{high}} &= \frac{\left(\bar{\mathcal{M}}_{\mathrm{c}} + \delta\mathcal{M}_{\mathrm{c}}\right)^{5/2}}{\left(\bar{\mathcal{M}}_{\mathrm{c}} + \delta\mathcal{M}_{\mathrm{c}}\right)^{5/2} + \left(\bar{\mathcal{M}}_{\mathrm{c}} - \delta\mathcal{M}_{\mathrm{c}}\right)^{5/2}} \\ &\simeq \frac{1}{2} + \frac{5}{4}\frac{\delta\mathcal{M}_{\mathrm{c}}}{\bar{\mathcal{M}}_{\mathrm{c}}}. \end{aligned} \tag{A1}$$

If N detections are made in total and the selection effects played no role, the expected number of detections from the higher-mass set would be $N/2$ with a standard deviation of $\sqrt{N}/2$. However, in our model, there is a predicted excess of $5N\delta\mathcal{M}_{\mathrm{c}}/(4\bar{\mathcal{M}}_{\mathrm{c}})$ high-mass detections because of selection effects. Consequently, we expect to have x-σ confidence in observing a selection effect on chirp mass, where

$$x = \frac{5\sqrt{N}}{2}\frac{\delta\mathcal{M}_{\mathrm{c}}}{\bar{\mathcal{M}}_{\mathrm{c}}}. \tag{A2}$$

We can estimate $\bar{\mathcal{M}}_{\mathrm{c}}$ from the mean of the chirp mass distribution, and $\delta\mathcal{M}_{\mathrm{c}}$ from the standard deviation; for our injections set, $\delta\mathcal{M}_{\mathrm{c}}/\bar{\mathcal{M}}_{\mathrm{c}} \approx 0.06$. For the Gaussian data set $N = 250$, and so we expect to observe selection effects at only the ~ 2-σ confidence level; the actual measurements are roughly consistent with this. For such a narrow chirp-mass distribution, $\gtrsim 10^3$ detections are needed to confidently observe the selection effects.

While the chirp mass is of prime importance to GW astronomers (it is their most precisely determined mass parameter), other combinations of mass are of interest in other contexts. Parameters which are correlated with the chirp mass, are also subject to selection effects. However, their significance is proportional to the level of correlation of the parameters with chirp mass; given that selection effects on chirp mass are small, we do not expect statistically significant effects for other mass parameters. Here, we present the distributions of the individual component masses, the asymmetric mass ratio and the total mass.

The distribution of recovered (injected) component masses is shown in figure 20. The detected events show a slight overrepresentation of higher-mass objects, which is the effect of selecting systems with larger chirp masses. The deviation from the injection distribution is small (a KS test with the predicted distribution gives p-values of 0.213 and 0.182 for

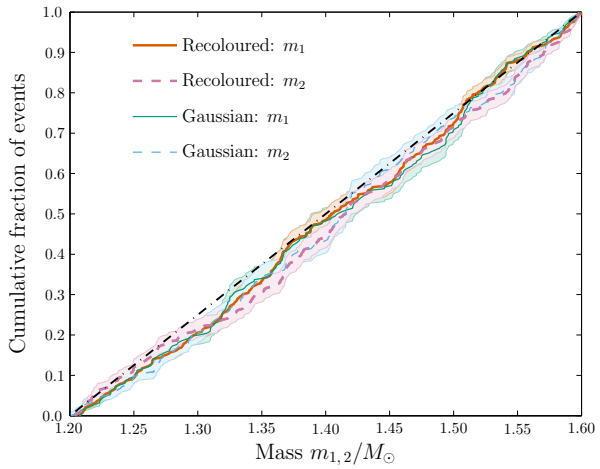


Figure 20. Cumulative fractions of detected events with component masses smaller than the abscissa value. The mass distribution for the first neutron star m_1 is denoted by the solid line, and the distribution for the second neutron star m_2 is denoted by the dashed line. Results with recoloured noise are denoted by the thicker red–purple lines, and results from the subset of 250 events analysed with LALINFERENCE with Gaussian noise are denoted by the thinner blue–green lines (Singer et al. 2014). The 68% confidence intervals are denoted by the shaded areas. The expected distribution for component masses drawn uniformly from $m_{\rm min} = 1.2M_\odot$ to $m_{\rm max} = 1.6M_\odot$ is indicated by the black dot–dashed line.

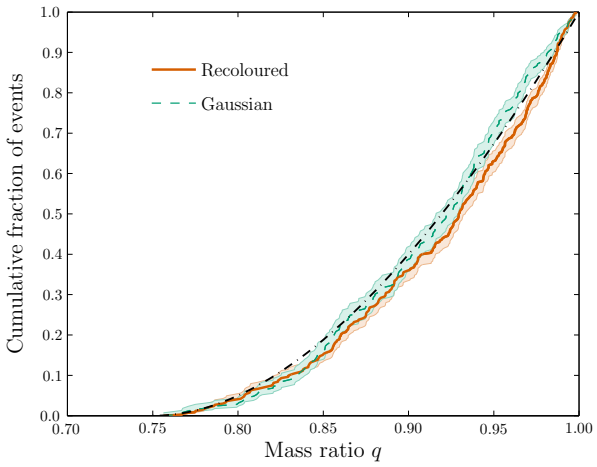


Figure 21. Cumulative fractions of detected events with asymmetric mass ratios smaller than the abscissa value. Results using recoloured noise are denoted by the solid red line, and results from the subset of 250 events with Gaussian noise analysed with LALINFERENCE are denoted by the dashed green line (Singer et al. 2014). The 68% confidence intervals are denoted by the shaded areas. The injection distribution $C_q(q)$ is indicated by the black dot–dashed line.

Gaussian noise, and 0.276 and 0.022 for the recoloured noise), but noticeably more significant than for the spins.

The asymmetric mass ratio is

$$q = \frac{\min\{m_1, m_2\}}{\max\{m_1, m_2\}}. \tag{A3}$$

For uniformly distributed m_1 and m_2 between $m_{\rm min}$ and $m_{\rm max}$, the probability density function for q is

$$P_q(q) = \begin{cases} \dfrac{1}{(m_{\rm max} - m_{\rm min})^2}\left(m_{\rm max}^2 - \dfrac{m_{\rm min}^2}{q^2}\right) & \dfrac{m_{\rm min}}{m_{\rm max}} \leq q \leq 1 \\ 0 & \text{Otherwise} \end{cases}. \tag{A4}$$

Integrating this gives a cumulative distribution function

$$C_q(q) = \begin{cases} 0 & q \leq \dfrac{m_{\rm min}}{m_{\rm max}} \\ \dfrac{1}{(m_{\rm max} - m_{\rm min})^2}\left(m_{\rm max}^2 q - 2m_{\rm min}m_{\rm max} + \dfrac{m_{\rm min}^2}{q}\right) & \dfrac{m_{\rm min}}{m_{\rm max}} \leq q \leq 1 \\ 1 & 1 \leq q \end{cases}. \tag{A5}$$

Figure 21 shows the recovered distribution of mass ratios as well as the injection distribution given by $C_q(q)$. There is a small difference between the injection and recovered distributions (a KS test with the injection distribution returns p-values of 0.536 and 0.050 for the Gaussian and recoloured noise respectively).

The probability density function for the total system mass, $M = m_1 + m_2$, is

$$P_M(M) = \begin{cases} \dfrac{1}{(m_{\rm max} - m_{\rm min})^2}(M - 2m_{\rm min}) & 2m_{\rm min} \leq M \leq m_{\rm min} + m_{\rm max} \\ \dfrac{1}{(m_{\rm max} - m_{\rm min})^2}(2m_{\rm max} - M) & m_{\rm min} + m_{\rm max} \leq M \leq 2m_{\rm max} \\ 0 & \text{Otherwise} \end{cases}. \tag{A6}$$

Consequently, its cumulative distribution function is

$$C_M(M) = \begin{cases} 0 & M \leq 2m_{\rm min} \\ \dfrac{1}{(m_{\rm max}-m_{\rm min})^2}\left(\dfrac{M^2}{2} - 2m_{\rm min}M + 2m_{\rm min}^2\right) & 2m_{\rm min} \leq M \leq m_{\rm min}+m_{\rm max} \\ \dfrac{1}{(m_{\rm max}-m_{\rm min})^2}\left(2m_{\rm max}M - \dfrac{M^2}{2} + m_{\rm min}^2 - 2m_{\rm min}m_{\rm max} - m_{\rm max}^2\right) & m_{\rm min}+m_{\rm max} \leq M \leq 2m_{\rm max} \\ 1 & 2m_{\rm max} \leq M \end{cases} . \quad \text{(A7)}$$

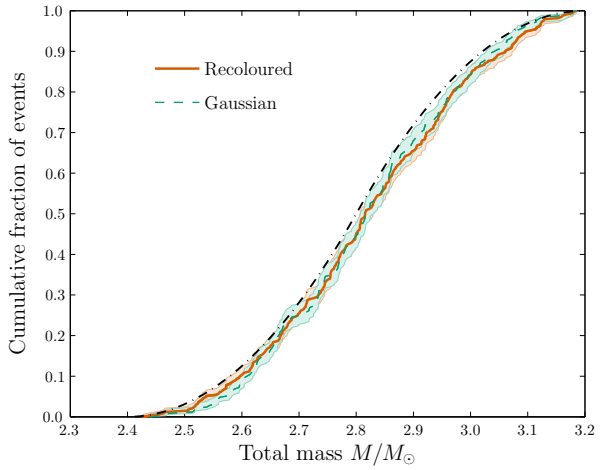


Figure 22. Cumulative fractions of detected events with total masses smaller than the abscissa value. Results using recoloured noise are denoted by the solid red line, and results from the subset of 250 events with Gaussian noise analysed with LALINFERENCE are denoted by the dashed green line (Singer et al. 2014). The 68% confidence intervals are denoted by the shaded areas. The injection distribution $C_M(M)$ is indicated by the black dot–dashed line.

Figure 22 shows the recovered distribution of mass ratios as well as the injection distribution given by $C_M(M)$. The distributions are similar to those seen for the chirp mass in figure 4. This is not surprising, as there is a clear link between the two quantities. We are considering a narrow mass range; individual component masses can be described as $m_{1,2} = m_{\rm min}(1+\varepsilon_{1,2})$, where $\varepsilon_{1,2} \leq (m_{\rm max}-m_{\rm min})/m_{\rm min} \ll 1$. The total mass is $m_{\rm min}(2+\varepsilon_1+\varepsilon_2)$; to first order in $\varepsilon_{1,2}$, the chirp mass can be described as $2^{-6/5}m_{\rm min}(2+\varepsilon_1+\varepsilon_2)$. Hence, the total mass is approximately proportional to the chirp mass across the range of interest. We preferentially select signals with larger total masses as these produce louder signals, although the difference between the injection and recovered distributions is not too large (a KS test with the injection distribution yields p-values of 0.338 and 0.050 for the Gaussian and recoloured noise respectively).

All the mass distributions show a difference between the injection and detected populations. This is as expected. The difference is small, such that for the numbers of events considered in this study, it is only marginally significant. The difference need not always be negligible, it would become more important when considering a larger population of events, or a set of events with a broader chirp mass distribution.

B. COMPUTATIONAL TIME

To perform rapid sky localization, we require that our analysis pipelines are expeditious. Following a detection, BAYESTAR promptly returns a sky localization, and later LALINFERENCE returns estimates of the sky position plus further parameters. Here, we present estimates for the computational time required to perform LALINFERENCE and BAYESTAR runs.

All results are specific to a two-detector network. The LALINFERENCE results are specific to a (non-spinning) TaylorF2 analysis: this is the least expensive waveform family and provides medium-latency results. Computational times can be significantly longer using other waveforms. Efforts are being made to optimise and speed up the methods of LALINFERENCE (e.g., Canizares et al. 2013; Farr et al. 2014; Canizares et al. 2014; Pürrer 2014).

The LALINFERENCE PE is slower than the rapid sky localization. Distributions of estimated CPU times for the runs are shown in figure 23. The LALINFERENCE_NEST times are calculated from log files. This is not entirely reliable as times may not be recorded for a variety of reasons. In this case, the reported time is a lower bound on the true value. In figure 23(a) we show the distribution of run times for both the set of all estimated times and the subset excluding those we suspect are inaccurate due to a reported error message. The distributions are consistent with our expectation that the inaccurate times are lower bounds. In figure 23(c) we show the cumulative distribution of run times using only the more reliable set of estimates. The median (accurately estimated) total CPU time for LALINFERENCE_NEST is 1.96×10^6 s = 545 hr (cf. Veitch et al. 2014) and the median total CPU time for BAYESTAR is 921 s = 15.4 min. Hence, on average LALINFERENCE_NEST takes $\sim$ 2000 times as much CPU time as BAYESTAR to complete.

The actual latency of a technique is given by the wall time, not the CPU time. Five CPU processes were used per LALINFERENCE_NEST run, hence the computational wall time can be estimated as a fifth of the total CPU time. This gives a median approximate wall time of 3.92×10^5 s = 109 hr. Some processes take longer to finish than others, so this is not an exact means of estimating the time taken for a run to finish. These calculations also neglect time spent idle rather than running, which influences the physical wall time required for a job to complete. In online mode, BAYESTAR is generally deployed in a 32–64-way parallel configuration. This gives a median wall time of 28.8 s (14.4 s) for a 32-way (64-way) configuration. BAYESTAR provides sky-localization $\sim 10^4$ times quicker than LALINFERENCE, furthermore, none of our BAYESTAR runs would have taken longer than a minute to complete (Singer 2014, chapter 4).

The length of the LALINFERENCE run depends upon the desired number of posterior samples. We may characterise the computational speed by the average rate at which independent samples are drawn from the posterior: the total number of (independent, as determined by LALINFERENCE) posterior samples divided by the total CPU time. The distribution of sam-

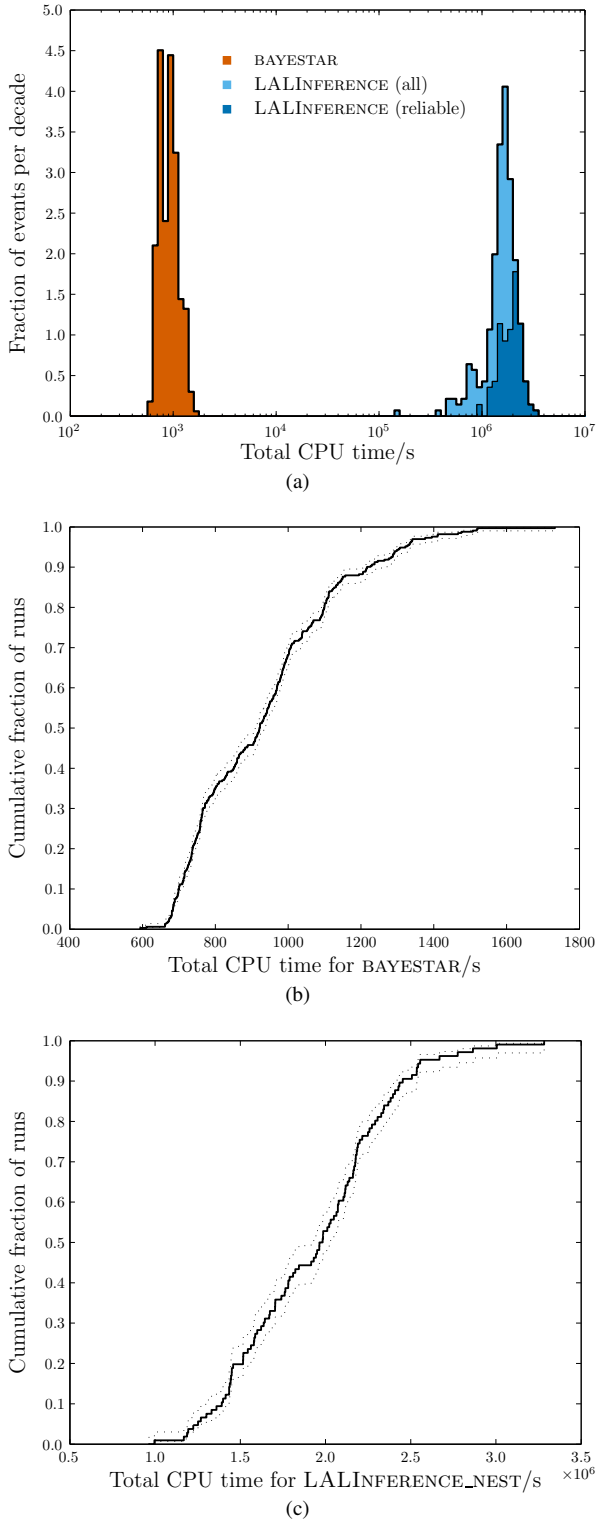


Figure 23. Computation time for a run measured in CPU seconds. (a) Distribution of run times. The left (red) distribution is for BAYESTAR and the right (blue) distribution is for LALINFERENCE_NEST. LALINFERENCE_NEST times which are reliably estimated are shown in dark blue, while the full set of times including potentially inaccurately estimated times are shown in light blue. (b) Cumulative fractions of BAYESTAR runs with computational times smaller than the abscissa value. (c) Cumulative fractions of LALINFERENCE_NEST runs with total CPU times smaller than the abscissa value, only reliable times are used here.The 68% confidence interval is enclosed by the dotted lines, this accounts for sampling errors and is estimated from a beta distribution (Cameron 2011). Each plot has a different scale.

pling speeds is shown in figure 24. We use speeds calculated using both reliably estimated times and those we suspect might be lower bounds (giving upper bounds for sampling speed) in figure 24(a), but only the more reliable values in figure 24(b). The median (accurately estimated) LALINFERENCE_NEST sampling speed is 4.40×10^{-3} s^{-1} = 15.8 hr^{-1} corresponding to one independent posterior sample every 227 s = 6.31×10^{-2} hr of CPU time (cf. Sidery et al. 2014; Veitch et al. 2014).

In contract, BAYESTAR computes the likelihood 24576 times. Its computation speed is thus simply inversely proportional to the total CPU time. The median BAYESTAR computational speed is 26.7 s^{-1} corresponding to one likelihood evaluation every 37.5 ms of CPU time. The difference between the LALINFERENCE and BAYESTAR computational speeds reflects the difference in the complexities of their likelihood functions.

The medium-latency PE runs, using the current code, finish in a few days. This is much longer than is required for BAYESTAR to produce sky-localization estimates. However, LALINFERENCE also provides posterior probability distributions for the other parameters as well as more accurate sky localization than BAYESTAR for three-detector networks (Singer et al. 2014).

REFERENCES

Aasi, J., et al. 2013a, Physical Review D, 88, 062001
—. 2013b, arXiv:1304.0670v1
—. 2014a, arXiv:1410.7764
—. 2014b, Physical Review Letters, 113, 011102
Abadie, J., et al. 2010, Classical and Quantum Gravity, 27, 173001
Acernese, F., et al. 2009, Advanced Virgo Baseline Design, Virgo Technical Report VIR-0027A-09
—. 2014, arXiv:1408.3978
Arun, K., Buonanno, A., Faye, G., & Ochsner, E. 2009, Physical Review D, 79, 104023
Arun, K., Iyer, B. R., Sathyaprakash, B., & Sundararajan, P. A. 2005, Physical Review D, 71, 084008
Baird, E., Fairhurst, S., Hannam, M., & Murphy, P. 2013, Physical Review D, 87, 024035
Barnes, J., & Kasen, D. 2013, Astrophys.J., 775, 18
Barsotti, L., & Fritschel, P. 2012, Early aLIGO Configurations: example scenarios toward design sensitivity, Tech. Rep. LIGO-T1200307-v4
Bartos, I., Crotts, A. P. S., & Marka, S. 2014, arXiv:1410.0677
Bohé, A., Marsat, S., & Blanchet, L. 2013, Classical and Quantum Gravity, 30, 135009
Brown, D. A., Harry, I., Lundgren, A., & Nitz, A. H. 2012, Physical Review D, 86, 084017
Bulik, T., & Belczynski, K. 2003, Astrophysical Journal, 589, L37
Buonanno, A., Chen, Y., & Vallisneri, M. 2003, Physical Review D, 67, 104025
Buonanno, A., Iyer, B., Ochsner, E., Pan, Y., & Sathyaprakash, B. 2009, Physical Review D, 80, 084043
Burgay, M., D'Amico, N., Possenti, A., et al. 2003, Nature, 426, 531
Cameron, E. 2011, Publications of the Astronomical Society of Australia, 28, 128
Canizares, P., Field, S. E., Gair, J., et al. 2014, arXiv:1404.6284
Canizares, P., Field, S. E., Gair, J. R., & Tiglio, M. 2013, Physical Review D, 87, 124005
Cannon, K., Cariou, R., Chapman, A., et al. 2012, Astrophysical Journal, 748, 136
Cannon, K., Chapman, A., Hanna, C., et al. 2010, Physical Review D, 82, 044025
Cannon, K., Hanna, C., & Keppel, D. 2013, Physical Review D, 88, 024025
Cannon, K., Hanna, C., Keppel, D., & Searle, A. C. 2011, Physical Review D, 83, 084053
Christensen, N. 2010, Classical and Quantum Gravity, 27, 194010
Cutler, C., & Flanagan, E. E. 1994, Physical Review D, 49, 2658
Cutler, C., & Vallisneri, M. 2007, Physical Review D, 76, 104018
Damour, T., Iyer, B. R., & Sathyaprakash, B. 2001, Physical Review D, 63, 044023
—. 2002, Physical Review D, 66, 027502
DeGroot, M. H. 1975, Probability and Statistics (Reading, Massachusetts: Addison-Wesley)
Fairhurst, S. 2009, New Journal of Physics, 11, 123006
—. 2011, Classical and Quantum Gravity, 28, 105021
Fan, X., Messenger, C., & Heng, I. S. 2014, Astrophysical Journal, 795, 43
Farr, B., Kalogera, V., & Luijten, E. 2014, Physical Review D, 90, 024014
Farr, W. M., Sravan, N., Cantrell, A., et al. 2011, Astrophysical Journal, 741, 103

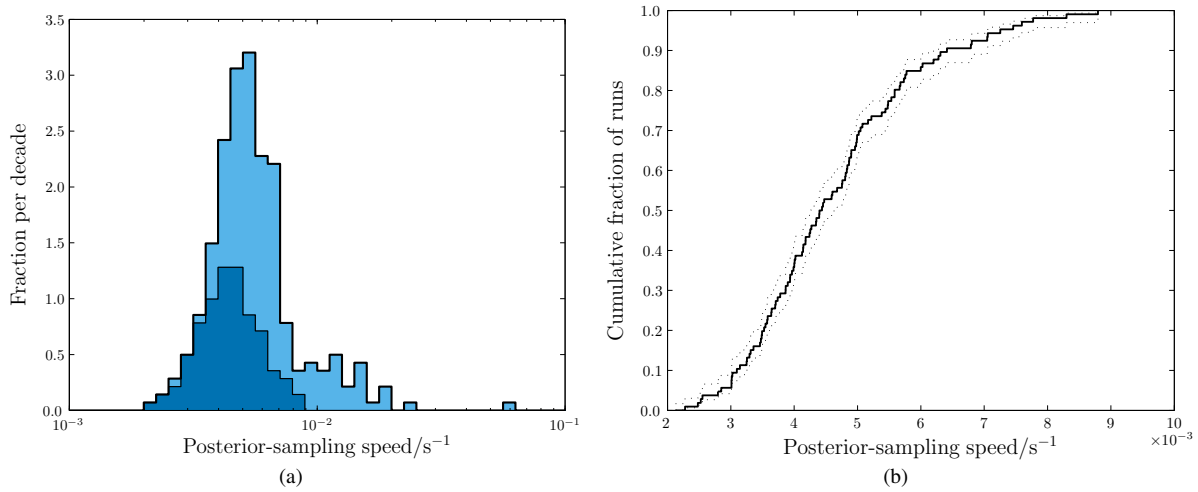


Figure 24. Computation speed of LALINFERENCE_NEST runs measured in independent posterior samples per CPU second. (a) Distribution of sampling speeds. Speeds based on reliably estimated CPU times are shown in dark blue, while the full set of speeds, including those using potentially inaccurately estimated times, are shown in light blue. (b) Cumulative fractions of runs with computational speeds smaller than the abscissa value, only reliable values are used here. The 68% confidence interval is enclosed by the dotted lines. All quantities are calculated based upon total CPU times, not wall times.

Feroz, F., Hobson, M., & Bridges, M. 2009, Monthly Notices of the Royal Astronomical Society, 398, 1601
Gorski, K., Hivon, E., Banday, A., et al. 2005, Astrophys.J., 622, 759
Graff, P., Feroz, F., Hobson, M. P., & Lasenby, A. 2012, Monthly Notices of the Royal Astronical Society, 421, 169
Graff, P., Feroz, F., Hobson, M. P., & Lasenby, A. N. 2014, Monthly Notices of the Royal Astronomical Society, 441, 1741
Gregory, P. C. 2005, Bayesian Logical Data Analysis for the Physical Sciences (Cambridge: Cambridge University Press)
Grover, K., Fairhurst, S., Farr, B., et al. 2014, Physical Review D, 89, 042004
Hanna, C., Mandel, I., & Vousden, W. 2014, Astrophysical Journal, 784, 8
Hannam, M., Brown, D. A., Fairhurst, S., Fryer, C. L., & Harry, I. W. 2013, Astrophysical Journal, 766, L14
Harry, G. M. 2010, Classical and Quantum Gravity, 27, 084006
Jaranowski, P., & Krolak, A. 2005, Living Reviews in Relativity, 8, 3
Kalogera, V., Kim, C.-L., Lorimer, D., et al. 2004, Astrophysical Journal, 601, L179
Kasliwal, M. M., & Nissanke, S. 2014, Astrophysical Journal, 789, L5
Kiziltan, B., Kottas, A., De Yoreo, M., & Thorsett, S. E. 2013, Astrophysical Journal, 778, 66
Kramer, M., & Wex, N. 2009, Classical and Quantum Gravity, 26, 073001
Kreidberg, L., Bailyn, C. D., Farr, W. M., & Kalogera, V. 2012, Astrophysical Journal, 757, 36
Lattimer, J. M. 2012, Annual Review of Nuclear and Particle Science, 62, 485
MacKay, D. J. C. 2003, Information Theory, Inference and Learning Algorithms (Cambridge: Cambridge University Press), 640
Mandel, I., Berry, C. P. L., Ohme, F., Fairhurst, S., & Farr, W. M. 2014, Classical and Quantum Gravity, 31, 155005
Mandel, I., & O'Shaughnessy, R. 2010, Classical and Quantum Gravity, 27, 114007
Metzger, B., & Berger, E. 2012, Astrophysical Journal, 746, 48
Metzger, B. D., Bauswein, A., Goriely, S., & Kasen, D. 2014, arXiv:1409.0544
Mikoczi, B., Vasuth, M., & Gergely, L. A. 2005, Physical Review D, 71, 124043
Moore, C. J., Cole, R. H., & Berry, C. P. L. 2014, arXiv:1408.0740
Nissanke, S., Sievers, J., Dalal, N., & Holz, D. 2011, Astrophysical Journal, 739, 99
Nuttall, L. K., & Sutton, P. J. 2010, Physical Review D, 82, 102002
Özel, F., Psaltis, D., Narayan, R., & McClintock, J. E. 2010, Astrophysical Journal, 725, 1918
Poisson, E., & Will, C. M. 1995, Physical Review D, 52, 848
Pürrer, M. 2014, Classical and Quantum Gravity, 31, 195010
Raymond, V., van der Sluys, M., Mandel, I., et al. 2009, Classical and Quantum Gravity, 26, 114007
Read, J. S., Markakis, C., Shibata, M., et al. 2009, Physical Review D, 79, 124033
Rodriguez, C. L., Farr, B., Raymond, V., et al. 2014, Astrophysical Journal, 784, 119
Rover, C., Meyer, R., & Christensen, N. 2006, Classical and Quantum Gravity, 23, 4895
Sathyaprakash, B., & Schutz, B. 2009, Living Reviews in Relativity, 12, arXiv:0903.0338
Schutz, B. F. 2011, Classical and Quantum Gravity, 28, 125023
Shoemaker, D. 2010, Advanced LIGO anticipated sensitivity curves, Tech. Rep. LIGO-T0900288-v3
Sidery, T., Aylott, B., Christensen, N., et al. 2014, Physical Review D, 89, 084060
Singer, L. P. 2014, PhD thesis, California Institute of Technology
Singer, L. P., Price, L. R., Farr, B., et al. 2014, Astrophysical Journal, 795, 105
Skilling, J. 2006, Bayesian Analysis, 1, 833
Taylor, S. R., Gair, J. R., & Mandel, I. 2012, Physical Review D, 85, 023535
Vallisneri, M. 2008, Physical Review D, 77, 042001
van der Sluys, M., Raymond, V., Mandel, I., et al. 2008a, Classical and Quantum Gravity, 25, 184011
van der Sluys, M., Roever, C., Stroeer, A., et al. 2008b, Astrophysical Journal, 688, L61
Veitch, J., Mandel, I., Aylott, B., et al. 2012, Physical Review D, 85, 104045
Veitch, J., Raymond, V., Farr, B., et al. 2014, arXiv:1409.7215
Veitch, J., & Vecchio, A. 2010, Physical Review D, 81, 062003
Vitale, S., & Del Pozzo, W. 2014, Physical Review D, 89, 022002