Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **439,** 3986–4010 (2014)
Advance Access publication 2014 February 27

doi:10.1093/mnras/stu254

Targeting supermassive black hole binaries and gravitational wave sources for the pulsar timing array

Pablo A. Rosado[1]★ and Alberto Sesana[2]

[1]*Max Planck Institute for Gravitational Physics, Albert Einstein Institute, Callinstraße 38, D-30167 Hanover, Germany*
[2]*Max Planck Institute for Gravitational Physics, Albert Einstein Institute, Am Mühlenberg 1, D-14476 Golm, Germany*

Accepted 2014 February 5. Received 2014 January 28; in original form 2013 November 25

ABSTRACT

This paper presents a technique to search for supermassive black hole binaries (MBHBs) in the Sloan Digital Sky Survey (SDSS). The search is based on the peculiar properties of merging galaxies that are found in a mock galaxy catalogue from the Millennium Simulation. MBHBs are expected to be the main gravitational wave (GW) sources for pulsar timing arrays (PTAs); however, it is still unclear if the observed GW signal will be produced by a few single MBHBs, or if it will have the properties of a stochastic background. The goal of this work is to produce a map of the sky in which each galaxy is assigned a probability of having suffered a recent merger, and of hosting a MBHB that could be detected by PTAs. This constitutes a step forward in the understanding of the expected PTA signal: the skymap can be used to investigate the clustering properties of PTA sources and the spatial distribution of the observable GW signal power; moreover, galaxies with the highest probabilities could be used as inputs in targeted searches for individual GW sources. We also investigate the distribution of neighbouring galaxies around galaxies hosting MBHBs, finding that the most likely detectable PTA sources are located in dense galaxy environments. Different techniques are used in the search, including Bayesian and machine learning algorithms, with consistent outputs. Our method generates a list of galaxies classified as MBHB hosts, that can be combined with other searches to effectively reduce the number of misclassifications. The spectral coverage of the SDSS reaches less than a fifth of the sky, and the catalogue becomes severely incomplete at large redshifts; however, this technique can be applied in the future to larger catalogues to obtain complete, observationally based information of the expected GW signal detectable by PTAs.

Key words: black hole physics – gravitational waves – methods: data analysis – pulsars: general – galaxies: evolution – galaxies: statistics.

1 INTRODUCTION

Studies of the dynamics of nearby galaxies (Kormendy & Richstone 1995; Richstone et al. 1998) suggest that a supermassive black hole (MBH) must reside at their centres, and there now exists plenty of observational evidence that almost all massive galaxies host a MBH, our Milky Way being the most striking example (Ghez et al. 2008; Gillessen et al. 2009). There is also a variety of investigations that confirm that the mass of the MBH is highly correlated to the mass and velocity dispersion of the bulge of the hosting galaxy (Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000; Marconi & Hunt 2003; Haering & Rix 2004; Graham et al. 2011; McConnell & Ma 2013). In the context of the Λ cold dark matter (ΛCDM) cosmology, large dark matter structures in the Universe build up hierarchically (White & Rees 1978). Galaxies form as gas cools at the centres of dark matter haloes; small dark matter haloes fall on to greater ones, and the galaxies of the former become satellites of the new host. At some later time, the smaller galaxies can merge on to the more massive ones, that lie at the bottom of the potential well. Within this framework, following galaxy mergers, a large number of supermassive black hole binaries (MBHBs) must form along the cosmic history (Begelman, Blandford & Rees 1980; Volonteri, Haardt & Madau 2003).

Depending on the mass ratio (i.e. the mass of the satellite galaxy over the mass of the primary) of the two systems, it is customary to divide galaxy interactions in *minor mergers* and *major mergers*. In a *minor merger*, the satellite is much lighter than the primary and its material can be disrupted before the two centres merge (Guo et al. 2011). Furthermore, dynamical friction (Chandrasekhar 1943), which is the main mechanism that brings the two MBHs towards each other, can become inefficient (causing the merger to

★E-mail: pablo.rosado@aei.mpg.de

take longer than the age of the Universe) if the masses of the two galaxies differ too much. Alternatively, in a *major merger*, the two galaxies have similar masses and their merger can be completed within less than a few Gyr (Kitzbichler & White 2008). Once the separation between the two galaxies is smaller than a few tens of kpc, the two MBHs can efficiently transfer energy and angular momentum to the surrounding stars and gas (Quinlan 1996; Escala et al. 2005; Dotti et al. 2007; Khan, Just & Merritt 2011; Preto et al. 2011), spiraling towards each other. At pc separations, they form a bound Keplerian binary; when the binary is tight enough (of the order of 0.01 pc) gravitational wave (GW) emission takes over, leading to coalescence.

Gravitational radiation emission from binary systems (Misner, Thorne & Wheeler 1973; Maggiore 2008) is predicted by Einstein's theory of General Relativity; however, only indirect proofs of this phenomenon have been achieved so far (Weisberg, Nice & Taylor 2010).[1] When GW emission becomes the main mechanism of energy loss of a binary, the inspiral process is well described by General Relativity in its lower order, quadrupolar approximation (Peters 1964). The period of the orbit decreases with time, while the amplitude of the emitted GWs increases. At the end of the inspiral phase, when the binary approaches the last stable orbit (see e.g. box 25.6 of Misner et al. 1973), the coalescence occurs, in which the amplitude of the GWs reaches its maximum (Ajith et al. 2011); after this, the binary enters the ring-down phase, and the GW emission rapidly decays. Merging MBHBs are the most powerful GW emitters in the Universe (Sesana et al. 2004).

The direct detection of GWs, which is the main goal of several state of the art experiments around the world, including GEO600, LIGO, Virgo and KAGRA [Abbott et al. (LIGO Scientific Collaboration) 2009; Abadie et al. (LIGO Scientific Collaboration) 2011; Pitkin et al. 2011; Accadia et al. 2012; Somiya (for the KAGRA Collaboration) 2012], will mark the beginning of the era of GW astronomy (Schutz 1999; Sathyaprakash & Schutz 2009). Many other experiments have been proposed, like eLISA [Shaddock 2008; Amaro Seoane et al. (eLISA Consortium) 2013] and ET (Punturo et al. 2010) to fully exploit this new window to the Universe, which will unveil valuable information not only about astrophysics, but also about cosmology and fundamental physics (Babak et al. 2011).

One of the most fascinating means of detecting GWs directly involve the timing of an ensemble of millisecond pulsars (MSPs, pulsars with spinning periods of $\sim$1 ms; Lorimer 2008), forming a pulsar timing array (PTA). MSPs are the most regular astronomical objects known, and they play a double role in GW astronomy. On the one hand, they are potential sources of GWs (Prix 2009; Andersson et al. 2011; Rosado 2012); on the other hand, they can also be used as parts of a galactic scale GW detector. GWs perturb the space–time metric between the pulsars and Earth, and these small perturbations affect the times of arrival (TOAs) of the pulses (Foster & Backer 1990; Jenet et al. 2005). The differences between the expected and measured TOAs are the timing residuals. By studying the timing residuals, PTAs aim to detect GWs of frequencies between $\sim 10^{-6}$ and $\sim 10^{-9}$ Hz. There are three independent PTA collaborations around the globe: the EPTA (Ferdman et al. 2010), NANOGrav (Jenet et al. 2009), and the PPTA (Manchester et al. 2013), which work jointly in the IPTA (Hobbs et al. 2010).

[1] A similar result based on the double pulsar PSR J0737−3039 (Burgay et al. 2003; Lyne et al. 2004) by Kramer et al. is in preparation (Kramer 2012).

Two main sources are expected to contribute to the GW spectrum in the frequency band of the PTA: MBHBs (Rajagopal & Romani 1995; Jaffe & Backer 2003; Wyithe & Loeb 2003) and cosmic strings (Regimbau et al. 2012; Sanidas, Battye & Stappers 2012). In particular, the incoherent superposition of the radiation of the numerous MBHBs in the Universe may produce a background of GWs (Sesana, Vecchio & Colacino 2008; Rosado 2011). Combining our current theoretical models with observational constraints, the present PTA sensitivity limit lies within the $\sim$95 per cent confidence level of the amplitude of the GW background from MBHBs (Sesana 2013b); this means that a detection may occur before the end of the decade.

To date, it is still unknown what kind of signal will be detected by the PTA; it can be dominated by the radiation of a handful of individually resolvable MBHBs (Sesana, Vecchio & Volonteri 2009; Mingarelli et al. 2012), or it can be an incoherent superposition of unresolvable sources, i.e. a stochastic background (Maggiore 2000). Very efficient searching algorithms have been developed to detect a Gaussian isotropic GW background (van Haasteren et al. 2011, 2013), but the actual properties of the background (especially its isotropy) are currently under investigation (Ravi et al. 2012; Mingarelli et al. 2013). Alternatively, different data analysis techniques are under development to detect the signature of individually resolvable MBHBs (Babak & Sesana 2012; Ellis, Siemens & Creighton 2012; Petiteau et al. 2013). Within this context, it is therefore meaningful to use available observations to better understand the distribution of MBHBs in the neighbouring Universe.

The goal of this work is to assign to each galaxy of a catalogue a probability of containing a MBHB. By doing that, we complement our theoretical models of galaxy mergers with information about their spatial distribution. This can be useful for on-going investigations regarding the anisotropy of the GW background (Mingarelli et al. 2013; Taylor & Gair 2013). A skymap of potential nearby sources of GWs also provides candidates on which a targeted search for GWs can be applied (Burke-Spolaor 2013), and on which algorithms for single MBHB searches can be tested. From a theoretical perspective, it is interesting to investigate the environments where MBHBs are formed, whether or not they are more likely to be found in galaxy clusters, and the relation of MBHBs with active galactic nuclei (AGN; Lynden-Bell 1969; Kauffmann et al. 2003b).

In order to find a criterion to identify galaxies that may contain a MBHB, we rely on a simulated galaxy catalogue. It is constructed from the Millennium Simulation (MS; Springel et al. 2005), using the galaxy formation models from Guo et al. (2011) and the all-sky light-cone produced by Henriques et al. (2012), with the stellar population from Bruzual & Charlot (2003). In this fake universe we find that galaxies that suffered a major merger in the 'recent' past (meaning in less than a few hundreds of Myr, which is the time lapse between snapshots of the simulation) have a distribution in redshifts and masses that does not follow that of non-merging galaxies. Moreover, they present a distinctive statistical distribution of neighbours in their surroundings (to distances up to a few Mpc). Therefore, we use the mass, the redshift, and the distribution of neighbouring galaxies to characterize the signatures of major mergers.

A galaxy that recently experienced a major merger will be referred to as a *B-galaxy*, since it may contain a MBHB. Conversely, a galaxy that did *not* merge in less than a few hundred Myr, will be called an *N-galaxy*. Only a fraction of B-galaxies can contain a MBHB, because the binary lifetime is generally shorter than the time lapse between the MS snapshots, which is used to define

B-galaxies.[2] B-galaxies containing a MBHB that could be observed by the PTA (when emitting in some frequency interval accessible to the PTA) will be called *PTA-galaxies*. Once we are able to identify B-galaxies in the fake catalogue, we adapt this catalogue to the observational limitations of a real catalogue (including, for example, the fact that redshifts are affected by peculiar velocities, and the incompleteness of observations at the low-mass/luminosity end). We then perform a similar search on the adapted catalogue, and study how the efficiency of the search is affected by these limitations. Finally, we perform the same search on a real catalogue, namely the Sloan Digital Sky Survey (SDSS) seventh data release (DR7; York et al. 2000; Abazajian et al. 2009), and obtain probabilities for real candidates of B- and PTA-galaxies.

The search for B-galaxies is performed using several algorithms. The simplest of them (based on Bayesian statistics) takes into account only redshifts and masses to characterize galaxies. When considering the spatial distribution of neighbours around B-galaxies, the search is performed via a machine learning algorithm (MLA). The method presented here to search for MBHBs in galaxy catalogues provides an alternative to other recent proposals (Tsalmantza et al. 2011; Eracleous et al. 2012; Ju et al. 2013; Shen et al. 2013): its advantage is that it is applicable to all galaxies, independent of their emission properties; its disadvantage is that it is a statistical, indirect method, that can be only used to pick candidates which are *more likely* to host a binary. The outcomes of this exploratory study could be improved in several ways (as discussed in Section 5); among other things, the way we adapt our fake catalogue to the SDSS' observational limitations is not optimal, which affects the selection of MBHB candidates.

The outline of the paper is as follows. In Section 2 we describe the galaxy catalogues (both fake and real) employed in this work; here we also explain the process used to adapt the fake catalogue to the observational limitations of the real one. Section 3 presents the methods applied to assign galaxies a probability of having suffered a major merger in the recent past, and the probability of containing an observable source of GW in the PTA frequency band. In Section 4 we show the results of our study of the clustering of B- and PTA-galaxies, as well as the efficiency of the different searches. We also present a skymap of the SDSS galaxies with the largest probabilities of being a B- or a PTA-galaxy. In Section 5, the main drawbacks of the searches and possible improvements are discussed. The main achievements, conclusions and caveats of the paper are summarized in Section 6. In Appendix A one can find additional material on how the data used in the paper have been obtained (from the MS and SDSS data bases).

2 DESCRIPTION OF THE CATALOGUES

2.1 Real catalogue

The real galaxy catalogue is the MPA/JHU value-added galaxy catalogue[3] (Kauffmann et al. 2003a; Brinchmann et al. 2004; Tremonti et al. 2004), which is based on the SDSS DR7[4] (York et al. 2000; Stoughton et al. 2002; Abazajian et al. 2009). In fact, we use the stellar masses in the MPA/JHU updated to DR8 photometry[5] (Aihara et al. 2011a,b). The SDSS is the largest and most complete redshift survey to date, covering roughly a quarter of the sky. It contains photometry and imaging of galactic and extragalactic objects, and spectroscopy for a fraction of them. The MPA/JHU complements the spectroscopy with the stellar masses of galaxies, which are calculated following the prospects of Kauffmann et al. (2003b) and Salim et al. (2007). The maximum redshift measured in this catalogue is $z_{\rm max} = 0.7$, which is the maximum redshift considered in our search. More precisely, the search will be described and performed (in Section 3.1) with a maximum redshift of 0.1. Then, in Section 3.4, the search will be extended to $z \leq 0.7$. We set a minimum redshift of $z_{\rm min} = 0.01$, because the SDSS imaging is frequently broken up below this redshift (Blanton et al. 2005); moreover, distances cannot be calculated from the redshift (since the assumption that galaxies drift with the Hubble flow does not hold), instead, one needs to complement the SDSS with other catalogues of nearby galaxies. The maximum and minimum stellar masses of galaxies found in the real catalogue are $m_{\rm max} = 10^{13}\,{\rm M}_{\odot}$ and $m_{\rm min} = 10^{6}\,{\rm M}_{\odot}$, respectively.

The SDSS spectroscopic redshift catalogue has the advantage of permitting very precise calculations of the positions of galaxies (unlike photometric redshift catalogues, in which redshifts have much larger uncertainties); however, the surveyed area covers only 19.5 per cent of the sky, and its completeness is affected by several effects (Blanton et al. 2005; Guo, Zehavi & Zheng 2012). Spectroscopic targets of the SDSS are assigned fibres using a tiling algorithm that optimizes completeness (Blanton et al. 2003). But fibres have a finite size, so they cannot be placed close enough to each other in order to aim at targets that have too small angular separations. Thus, in areas of the sky with high galaxy density, the completeness of the spectroscopic catalogue decreases considerably. This issue, called *fibre collision*, affects the measurements on galaxy clustering (Guo et al. 2012), specially at scales $\lesssim$1 Mpc. Besides the fibre collision, the completeness of the spectroscopic catalogue changes from one region of the sky to another. All these effects should be taken into account when adapting the simulated catalogue to the limitations of the SDSS spectroscopic catalogue (see Section 5). However, we apply a simple method (described in Section 2.3) that does not take them into account. In Fig. 1 we show how galaxies in the real catalogue are distributed in stellar mass and redshift. This distribution will be used to adapt our fake catalogue to the observational limitations of the SDSS (see Section 2.3).

2.2 Fake catalogue

The MS (Springel et al. 2005) is an N-body simulation in which 10^{10} particles of dark matter evolve in time, in a cubic region of comoving side $\sim 500 h_{100}^{-1}$ Mpc, where $h_{100} = H_0/[100\,{\rm km\ s^{-1}Mpc^{-1}}]$ and H_0 is the present-day Hubble expansion rate. These particles interact and form structures in a ΛCDM universe. Haloes and subhaloes are identified using the methods described in Springel et al. (2001), and baryonic matter is then assigned to the haloes, following the semi-analytical models of Guo et al. (2011). The distribution of haloes and galaxies is recorded in 64 different snapshots, from redshift $z = 127$ to 0. Since galaxies at $z \gtrsim 0.1$ are distributed in a comoving volume larger than the simulation cube, the latter is repeated periodically. The mock catalogues (in which equatorial coordinates and apparent

[2] We assume that once the dynamical friction process has ended the two black holes coalesce in an interval of time shorter than a few hundreds of Myr. This assumption is justified in Section 5.

[3] Maintained by Jarle Brinchmann at http://www.mpa-garching.mpg.de/SDSS/.

[4] http://www.sdss.org/dr7/

[5] We thank Jarle Brinchmann for providing us with the updated stellar mass catalogue.

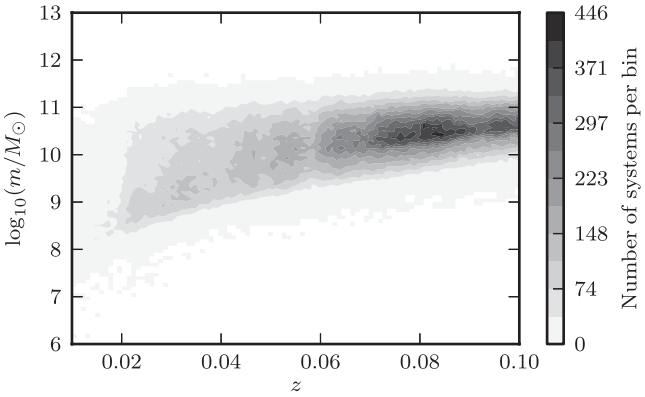

Figure 1. Contour plot of stellar mass versus apparent redshift of the real catalogue. The horizontal axis is divided into 100 redshift intervals (or z bins), and the vertical one into 100 intervals of logarithmic mass (let us call them m bins for simplicity). The grey-scale gives the number of systems contained in each redshift–mass pixel (or z–m bin).

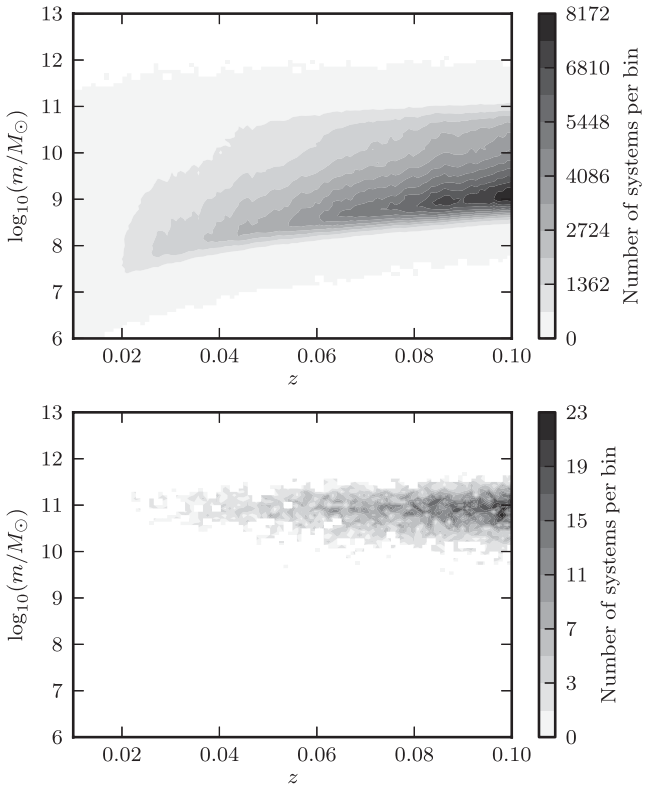

Figure 2. Contour plot of stellar mass versus redshift of the fake galaxy catalogue (for systems with $z < 0.1$). Both axes are divided into 100 equal bins; the grey-scale gives the number of systems contained in each z–m bin. The upper plot considers all galaxies in the fake catalogue, while the lower plot contains only B-galaxies.

redshifts are assigned to galaxies as if we were observers in the fake universe) are constructed as explained in Henriques et al. (2012).

The outcomes of the MS have been contrasted with many observations, confirming that the properties of the fake universe match well the current population of galaxies and MBHs (see e.g. Marulli et al. 2008; Bonoli et al. 2009). The cosmological parameters assumed in the MS are a combination of the 2dFGRS (Colless et al. 2001) with the first year of data from *WMAP* (Spergel et al. 2003). To be consistent with the cosmological model of the simulation, when dealing with MS data we assume that the density parameters of matter and dark energy are $\Omega_{\rm m} = 0.25$ and $\Omega_\Lambda = 0.75$, respectively, and $h_{100} = 0.73$. Alternatively, SDSS derived data are treated with the cosmological parameters $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.7$, and $h_{100} = 0.7$, which are the values assumed in the MPA/JHU. Neither of these sets of cosmological parameters agree with the most recent measurements; the possible effect of the 'outdated' cosmological parameters in the results is commented on in Section 5. A new release of the mock galaxy catalogues has been made public during the writing of this paper; the simulation has been rescaled (Guo et al. 2013) to adapt the results to a cosmology based on the data of *WMAP* 7 (Komatsu et al. 2011). Redoing this investigation using the 'updated' fake universe may be considered for a future work.

Our fake galaxy catalogue can be downloaded from the MS website[6] (Lemson & Virgo Consortium 2006), using the SQL query given in Appendix A. Each galaxy in the fake catalogue has a unique identification number (called `galID`). However, as already pointed out, the simulated universe has a cubic finite size; this cube is repeated periodically, to permit galaxies at larger distances. A galaxy in one of the cubes has the same mass (as well as other intrinsic properties) and `galID` of its analogous ones in other cubes, but a different sky position and redshift.

B-galaxies are *descendants* of two (or rarely three) merging *progenitors*. A descendant suffered a *major merger* if the mass of at least two of the progenitors is $\geq$0.2 times the mass of the descendant. Moreover, the merger had to occur between the snapshot corresponding to the redshift of the descendant and the immediately previous snapshot. In Appendix A one can find more details on the selection of B-galaxies, and the query used to download their `galID`. Once we know which systems are B-galaxies, we can perform searches for B-galaxies in the fake catalogue to check how well the methods work. These searches (described in Section 3) are first applied to the local Universe (with a maximum redshift of 0.1); then in Section 3.4 we extend the algorithms to a maximum redshift of 0.7. For $z < 0.1$ we find (using the second query of Appendix A) 8400 B-galaxies, of which ~91 per cent are unique (and the rest are repetitions).

In Fig. 2, redshift–mass histograms are plotted for all galaxies and for only B-galaxies. There we see that B-galaxies are biased towards larger masses. The histograms in Fig. 2 provide a prescription to distinguish B-galaxies from N-galaxies. With only this information, one can already assign probabilities of galaxies in the real catalogue (assuming that we know their redshifts and masses well enough). The description of such a search is given in Section 3.1.

2.3 Adapted catalogue

We now turn to the procedure we have used to adapt our fake catalogue to the observational constraints of the SDSS spectroscopic catalogue. As already mentioned in Section 2.1, this may not be optimal; more sophisticated methods (like the one described in Li et al. 2006b) should be used to properly account for the SDSS incompleteness.

Redshifts in the adapted catalogue are *apparent redshifts*; these are the redshifts that would be measured if we were observers in the simulated universe, taking into account that galaxies have peculiar velocities. These apparent redshifts are included in the MS data base (labelled `z_app` in the queries of Appendix A).

[6] http://www.mpa-garching.mpg.de/millennium/

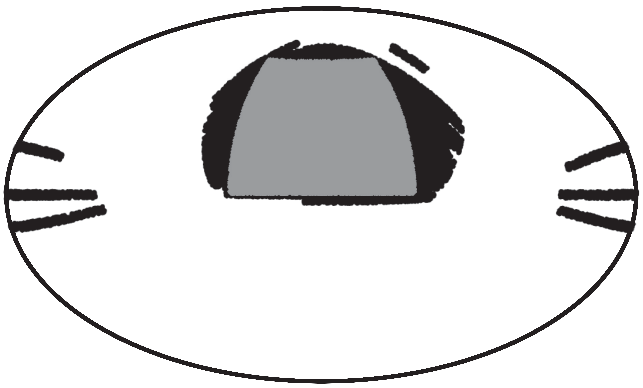

Figure 3. Two-dimensional (Hammer) projection of the part of the sky covered by our real catalogue (black area). The grey area is the border-free central region we have chosen as model for the redshift and mass distributions of real galaxies, in order to construct the adapted catalogue.

We first select a region of the sky that is 'almost completely included' in the SDSS spectroscopic catalogue, i.e. far from the borders of the SDSS footprint. In Fig. 3, a projected skymap of galaxies with spectroscopy is plotted. The selected border-free region is highlighted in grey; this area is limited by two fixed values of right ascension (RA) and declination (Dec.), selected in such a way that the area is exactly 1/9th of the entire sky. The fake catalogue is then divided into nine regions of equal area. A z–m histogram is calculated for the SDSS central region and for the nine different patches of the simulated sky. In all histograms, the axes are divided into 100 bins; where redshifts and masses fulfil $z \in [0.01, 0.11]$, and $\log_{10}(m/\mathrm{M}_{\odot}) \in [6, 13]$, respectively. Each histogram is thus a 100×100 matrix in which each element gives the number of systems with redshifts and masses in a particular z–m bin. We compare each of the nine histograms of the fake catalogue with the SDSS histogram. On the one hand, if one z–m bin from the fake catalogue contains n more galaxies than the same pixel in the SDSS, n galaxies in that bin are randomly chosen and deleted from the fake catalogue. On the other hand, if one pixel in the fake catalogue contains fewer galaxies than the analogous SDSS pixel, nothing is done. This implies that the number of systems in the adapted catalogue is slightly smaller than in the SDSS for some regions of the z–m histogram; in fact, at the high-mass end, the fake catalogue presents a small shortage of systems with respect to the real one (we will comment on this in Section 5). Finally, the adapted catalogue is the result of combining the nine sky regions of the fake catalogue from which systems have been subtracted.

In Fig. 4, a map of the real local Universe (from the real catalogue) is compared to a map of the simulated local universe (from the adapted catalogue). The z–m histogram of the adapted catalogue is shown in the upper plot of Fig. 5. This is, as expected, very similar to that of Fig. 1, except for the fact that the real catalogue contains fewer systems at each pixel, due to the smaller sky region that it covers. The lower plot of Fig. 5 shows the z–m histogram of B-galaxies in the adapted catalogue.

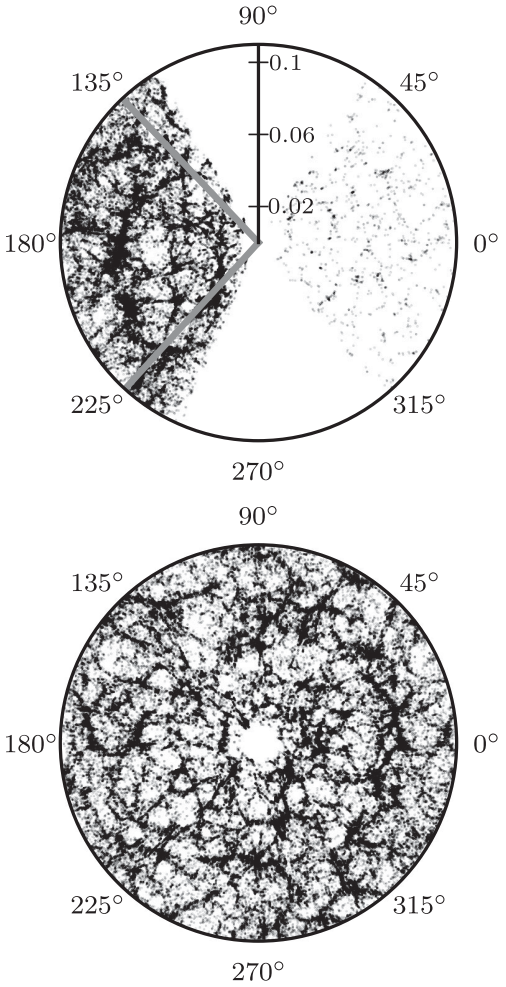

Figure 4. Sky distribution of galaxies (projected over the equatorial plane) with RA, Dec., and redshift in the ranges RA$\in [0^{\circ}, 360^{\circ})$, Dec.$\in [1^{\circ}, 8^{\circ}]$, and $z \in [0.01, 0.11]$, respectively. The upper plot corresponds to the real catalogue, and the lower one to the adapted catalogue. The grey lines in the first plot delimit the region that has been chosen as reference to construct the adapted catalogue (the grey region in Fig. 3).

3 DESCRIPTION OF THE SEARCHES

3.1 Probabilities of B-galaxies

In our searches, each system is characterized by a vector of parameters $\boldsymbol{\theta}$. We start here with the identification of B-galaxies through their peculiar mass and redshift distribution, therefore, for the time being, $\boldsymbol{\theta} = \{z, m\}$. For practical purposes, we divide each parameter range in 100 bins, so that the $z-m$ parameter space forms a matrix of 10^4 elements. We name a generic element of this matrix θ_i, where $i = 1, \ldots, 10^4$. We now define two functions: $n_f^{\mathrm{G}}(\theta_i)$ is the number of galaxies (B- or N-galaxies) in the fake catalogue with parameters within θ_i; $n_f^{\mathrm{B}}(\theta_i)$ is the number of B-galaxies in the fake catalogue with parameters within θ_i. The total number of galaxies in the fake catalogue is thus

$$\mathcal{N}_f^{\mathrm{G}} = \sum_i n_f^{\mathrm{G}}(\theta_i). \tag{1}$$

Similarly,

$$\mathcal{N}_f^{\mathrm{B}} = \sum_i n_f^{\mathrm{B}}(\theta_i) \tag{2}$$

is the total number of B-galaxies in the fake catalogue.

The probability of a system in the fake catalogue being a B-galaxy, in the case of total ignorance about the parameters $\boldsymbol{\theta}$, is

$$p_f(\mathrm{B}|I) = \frac{\mathcal{N}_f^{\mathrm{B}}}{\mathcal{N}_f^{\mathrm{G}}}, \tag{3}$$

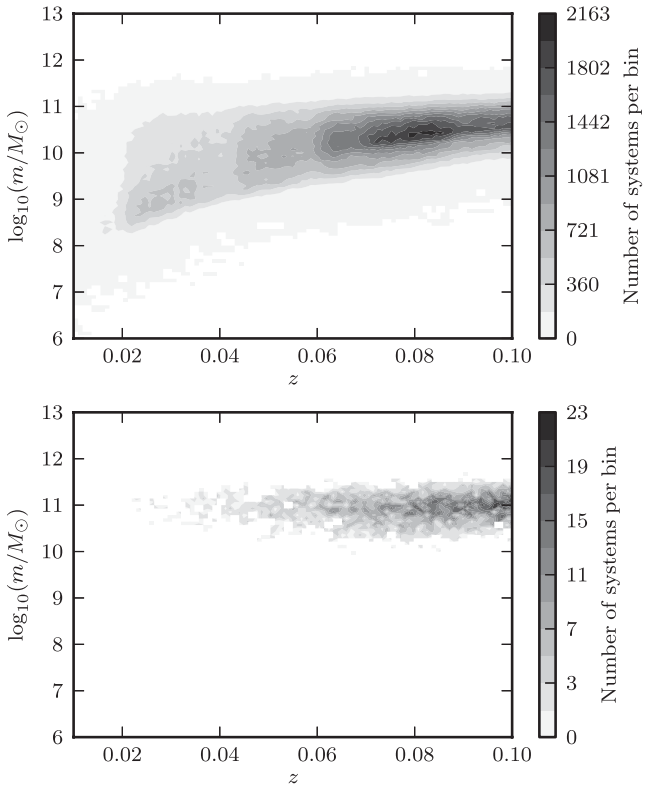

Figure 5. Contour plots of stellar mass versus apparent redshift of the adapted catalogue. Both axes are divided into 100 bins; the grey-scale gives the number of systems contained in each pixel. The upper plot considers all galaxies in the adapted catalogue. This plot is to be compared with that of Fig. 1 (but note that the numbers of systems per z–m bin are larger here than in Fig. 1, because the area of the sky covered by the real catalogue is smaller than that of the adapted catalogue). The lower plot is obtained for the subset of B-galaxies in the adapted catalogue.

which is our prior, using the typical notation and nomenclature of Bayesian statistics. The probability of a system in the fake catalogue having z and m within θ_i, given that it is a B-galaxy, is

$$p_f(\theta_i|\mathrm{B}) = \frac{n_f^{\mathrm{B}}(\theta_i)}{\mathcal{N}_f^{\mathrm{B}}}, \tag{4}$$

which is the likelihood. Equations (3) and (4) can be adapted to the case of N-galaxies; then,

$$p_f(\mathrm{N}|I) = \frac{\mathcal{N}_f^{\mathrm{G}} - \mathcal{N}_f^{\mathrm{B}}}{\mathcal{N}_f^{\mathrm{G}}} \tag{5}$$

is the probability of a system being an N-galaxy, in the absence of other information, and

$$p_f(\theta_i|\mathrm{N}) = \frac{n_f^{\mathrm{G}}(\theta_i) - n_f^{\mathrm{B}}(\theta_i)}{\mathcal{N}_f^{\mathrm{G}} - \mathcal{N}_f^{\mathrm{B}}} \tag{6}$$

is the probability of an N-galaxy having parameters within θ_i. Using Bayes' theorem, the probability of a system in the fake catalogue being a B-galaxy, given that it has z and m within θ_i, is

$$p_f(\mathrm{B}|\theta_i) = p_f(\theta_i|\mathrm{B})\frac{p_f(\mathrm{B}|I)}{p_f(\theta_i|I)}. \tag{7}$$

Here, the term in the denominator is the normalization, given by

$$p_f(\theta_i|I) = p_f(\theta_i|\mathrm{B})\,p_f(\mathrm{B}|I) + p_f(\theta_i|\mathrm{N})\,p_f(\mathrm{N}|I) = \frac{n_f^{\mathrm{G}}(\theta_i)}{\mathcal{N}_f^{\mathrm{G}}}. \tag{8}$$

Introducing equations (3), (4), and (8) in (7), we get

$$p_f(\mathrm{B}|\theta_i) = \frac{n_f^{\mathrm{B}}(\theta_i)}{n_f^{\mathrm{G}}(\theta_i)}, \tag{9}$$

which is an expected result: the probability of a system within a z–m bin being a B-galaxy is just the ratio of the number of B-galaxies over the total number of galaxies in that pixel.

The same statistics can be applied to systems in the adapted catalogue. The probability of a system in the adapted catalogue to be a B-galaxy, given that it has z and m within θ_i, is

$$p_a(\mathrm{B}|\theta_i) = \frac{n_a^{\mathrm{B}}(\theta_i)}{n_a^{\mathrm{G}}(\theta_i)}, \tag{10}$$

where we have introduced $n_a^{\mathrm{G}}(\theta_i)$, the number of galaxies in the adapted catalogue with parameters within θ_i, and $n_a^{\mathrm{B}}(\theta_i)$, the number of B-galaxies in the adapted catalogue with parameters within θ_i. From now on, we will call $p_x(\mathrm{B}|\theta_i)$ the *B-galaxy probability* of a system of catalogue x (where x can be 'f', 'a', or 'r', corresponding to the fake, adapted, or real catalogue, respectively).

The number of B-galaxies in the real catalogue, $n_r^{\mathrm{B}}(\theta_i)$, is (of course) unknown, but we do know $n_r^{\mathrm{G}}(\theta_i)$, the number of galaxies in the real catalogue with parameters within θ_i. The function $n_r^{\mathrm{G}}(\theta_i)$ should be almost identical to $n_a^{\mathrm{G}}(\theta_i)$ (by construction of the adapted catalogue), except for an overall normalization factor (given that the real catalogue does not cover the entire sky). Then,

$$p_r(\mathrm{B}|\theta_i) = p_a(\mathrm{B}|\theta_i) \tag{11}$$

is assumed to be the probability of a system in the real catalogue being a B-galaxy, given that it has z and m within θ_i.

Since we want to test the efficiency of the searches, the probability matrices $p_x(\mathrm{B}|\theta_i)$ (for $x = f$ or a) are calculated using systems of only one half of the sky (with $0^\circ \leq \mathrm{RA} < 180^\circ$). These systems form the *training set*. Afterwards, the efficiency of the searches are tested (as will be explained in Section 4.2) using systems from the other half (with $180^\circ \leq \mathrm{RA} < 360^\circ$), that form the *testing set*. Furthermore, the probability matrices $p_x(\mathrm{B}|\theta_i)$ are calculated as the average over 1000 realizations; in each realization, B- and N-galaxies are randomly chosen until covering 19.5 per cent of the sky (the area of the SDSS spectroscopic footprint). In the case of $p_a(\mathrm{B}|\theta_i)$, in each realization we use a different adapted catalogue (from a list of 100 different adapted catalogues, each one built as described in Section 2.3). This process reduces the amount of systems that are contained both in the training and in the testing sets (because of the repetitions of the simulated cube, mentioned in Section 2.2).

An additional remark about the probabilities $p_x(\mathrm{B}|\theta_i)$ needs to be made. A certain galaxy may have z and m within a bin θ_i that contains zero B-galaxies, even when all bins around that particular one do contain B-galaxies. The B-galaxy probability would thus be zero for that system. But this would not be fair: the probabilities would strongly depend on the sizes of the z and m bins (since, for a different choice of the sizes, that bin θ_i would not be empty of B-galaxies). Moreover, our results would also depend too much on the particular realization of the universe that the MS provides. To avoid these biases, $p_x(\mathrm{B}|\theta_i)$ is smoothed with a two-dimensional Gaussian filter. The results nevertheless do not change significantly when using different types of filter or using no filter at all.

Finally, each system (from the fake, adapted, and real catalogues) is assigned a value of B-galaxy probability from the smoothed probability matrix $p_x(\mathrm{B}|\theta_i)$, depending on the z–m bin θ_i it falls

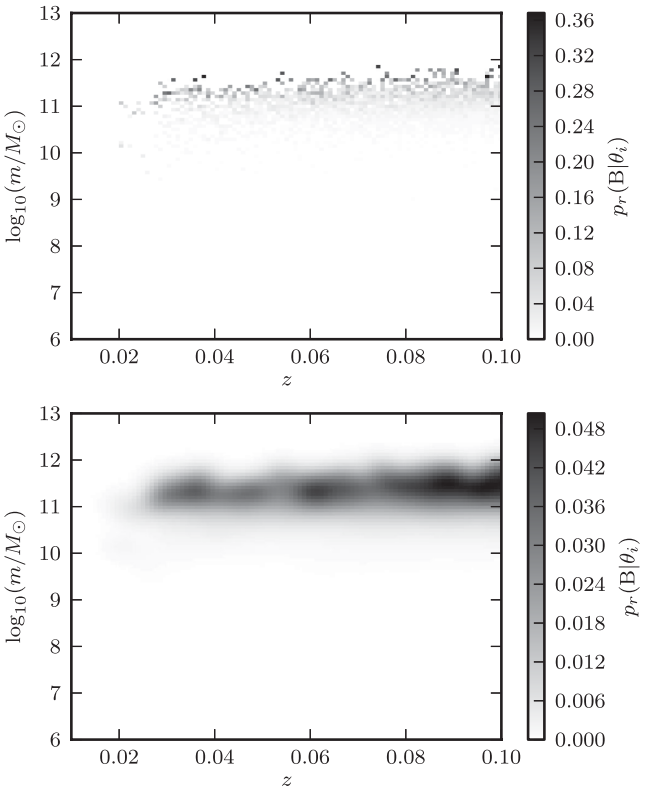

Figure 6. B-galaxy probability map as a function of redshift and mass, for systems in the adapted and real catalogues, i.e. $p_r(\mathrm{B}|\theta_i)$. The upper plot is the (unfiltered) average probability matrix obtained with the galaxies of the training set, as explained in the text. The lower plot shows the same matrix smoothed with a Gaussian filter; this is the probability matrix that will be used to assign B-galaxy probabilities to systems in the adapted and real catalogues.

into. The unfiltered and filtered probability matrices are plotted in Fig. 6.

3.2 Probabilities of PTA-galaxies

A galaxy is said to be a B-galaxy if, among other conditions mentioned in Section 2.2 or in Appendix A, it is the descendant of two (or more, in the case of a multimerger) progenitors that existed only until the previous snapshot in the simulation. In other words, between a snapshot and the following one, two (or more) different galaxies became one. The resulting galaxy may or may not have a MBHB, depending on when the merger actually started and on the lifetime of the binary. We define the function $T^{\mathrm{snap}}(z)$ as the interval of time elapsed between the current snapshot of a galaxy with redshift z and the previous one (in the fake catalogue or in the adapted catalogue).

During a galaxy merger several physical mechanisms contribute to shrink the distance between the two MBHs, the final one being the emission of GWs. At this stage, the time a MBHB spends emitting between two observed GW frequencies f_1 and f_2 can be calculated using the quadrupolar approximation, and gives

$$\tau(f_1, f_2) = \frac{5c^5[1+z]^{-8/3}}{256\pi^{8/3}[G\mathcal{M}]^{5/3}}\left[f_1^{-8/3} - f_2^{-8/3}\right]. \quad (12)$$

Here, c and G are the speed of light and the gravitational constant, respectively, and

$$\mathcal{M} = \frac{[m^1_{\mathrm{BH}} m^2_{\mathrm{BH}}]^{3/5}}{[m^1_{\mathrm{BH}} + m^2_{\mathrm{BH}}]^{1/5}} \quad (13)$$

is the chirp mass of a binary composed of two MBHs of masses m^1_{BH} and m^2_{BH}. The lighter the mass of the binary, the longer the time interval it spends emitting in a certain frequency interval. We disregard the other mechanisms of energy loss that may play an important role when the binary orbits at distances larger than $\sim$0.1 pc. These mechanisms would enhance the loss of angular momentum at low frequencies, reducing the amount of time the MBHB is emitting GWs (Kocsis & Sesana 2011; Sesana 2013a). The inclusion of these effects in the calculations could be subject of future work.

The GWs produced by a MBHB in a quasi-circular orbit, at observed GW frequency f, would produce a strain amplitude of

$$h_0 = \frac{2[G\mathcal{M}]^{5/3}\,[\pi f[1+z]]^{2/3}}{c^4 r(z)}. \quad (14)$$

Here, the function $r(z)$ is the comoving distance between Earth and a galaxy of redshift z, given in a ΛCDM universe by

$$r(z) = \frac{c}{H_0}\int_0^z \left[\Omega_{\mathrm{m}}[1+z']^3 + \Omega_\Lambda\right]^{-1/2} \mathrm{d}z'. \quad (15)$$

The values assumed for the cosmological parameters H_0, Ω_{m}, and Ω_Λ are the ones given in Section 2.2 (we use different parameters for systems from the MS and from the MPA/JHU). The maximum frequency at which a system can be observed is

$$f_{\mathrm{lso}} = \frac{c^3}{6\sqrt{6}\pi G[m^1_{\mathrm{BH}} + m^2_{\mathrm{BH}}][1+z]}, \quad (16)$$

which is the frequency of the last stable orbit. The minimum frequency that can be observed is chosen in such a way that the interval of time until the coalescence is not longer than 0.1 Gyr; we assume that, at lower frequencies, other mechanisms of energy loss would dominate over the GW emission.

The PTA is sensitive to GWs within a certain interval of observed frequencies $[f_{\mathrm{min}}, f_{\mathrm{max}}]$. We choose this *frequency window* to be

$$[f_{\mathrm{min}}, f_{\mathrm{max}}] = \left[[10\,\mathrm{yr}]^{-1}, [1\,\mathrm{week}]^{-1}\right]. \quad (17)$$

The lower limit is given by the duration of the PTA campaign; we take 10 years as a default value. The upper limit is set by the cadence of individual pulsar observations, typically of one per week. The exact choice of this upper limit does not make any difference in the results, since we do not expect observable sources at such high frequencies anyway. The PTA frequency window $[f_{\mathrm{min}}, f_{\mathrm{max}}]$ is divided into 100 frequency bins, equally separated in logarithmic scale. Galaxies are assigned a probability of being PTA-galaxies at each frequency bin, i.e. they are assigned the probability of producing a strain amplitude larger than a certain threshold within a certain observed interval of frequencies.

Let us first calculate the probability of a system to be a PTA-galaxy at a certain frequency bin, assuming that it is a B-galaxy that contains a binary. To calculate this probability, $p_x(\mathrm{P}|\mathrm{B}; \mathcal{M}, f)$, we follow an iterative process. In each realization (of a total of 100), a total black hole mass m_{BH} is drawn from a log-normal distribution, with mean given by the fitting formula from McConnell & Ma (2013),

$$\log_{10}\left(\frac{m_{\mathrm{BH}}}{\mathrm{M}_\odot}\right) = 8.46 + 1.05\log_{10}\left(\frac{m_{\mathrm{bulge}}}{10^{11}\mathrm{M}_\odot}\right), \quad (18)$$

and standard deviation $\sigma = 0.34$. In equation (18), $m_{\rm bulge}$ is the mass of the galactic bulge. Bulge masses are obtained directly from the MS data base for systems in the fake and adapted catalogues (called `bulgemass` in the first query of Appendix A); in Section 4.6 we explain how to calculate $m_{\rm bulge}$ for galaxies in the real catalogue. To avoid considering very light MBH masses,[7] we discard all MBHs with masses smaller than $10^6\,{\rm M}_\odot$. The mass ratio of the binary, $q = m^1_{\rm BH}/m^2_{\rm BH}$ (with $m^1_{\rm BH} \leq m^2_{\rm BH}$), is drawn at each realization from the distribution of the mass ratios of progenitors' black hole masses in the fake catalogue. Hence, at each realization we have $m^1_{\rm BH} = q\, m_{\rm BH}$, and $m^2_{\rm BH} = m_{\rm BH} - m^1_{\rm BH}$. Then, the probability that a galaxy, assuming it contains a MBHB, of catalogue x with redshift z is detectable by the PTA for a given minimum strain amplitude $h_0^{\rm thres}$ (the strain amplitude threshold) at a certain observed frequency bin $[f_1, f_2]$, is

$$p_x({\rm P}|{\rm B};\mathcal{M},f) = \frac{1}{100}\sum_j^{100} p_x^j({\rm P}|{\rm B};\mathcal{M},f). \tag{19}$$

Here, $f = [f_1 + f_2]/2$ is the central frequency of the bin, j denotes the number of the realization, and

$$p_x^j({\rm P}|{\rm B};\mathcal{M},f) = \begin{cases} \min\left(1, \frac{\tau(f_1,f_2)}{T^{\rm snap}(z)}\right) & \text{if} \quad h_0 \geq h_0^{\rm thres} \\ 0 & \text{if} \quad h_0 < h_0^{\rm thres} \end{cases}. \tag{20}$$

In the previous equation we have introduced the function min ( ), to avoid probabilities larger than unity.[8]

Now, applying the product rule, the probability that a galaxy with z and m within θ_i, and MBHB chirp mass $\mathcal{M}$ (if the galaxy hosts a MBHB), is a B-galaxy and a PTA-galaxy in a frequency bin centred at f is

$$p_x({\rm B},{\rm P}|\theta_i,\mathcal{M},f) = p_x({\rm B}|\theta_i)\, p_x({\rm P}|{\rm B};\mathcal{M},f). \tag{21}$$

Finally, the *PTA-galaxy probability* is

$$p_x({\rm B},{\rm P}|\theta_i,\mathcal{M}) = \sum_k p_x({\rm B}|\theta_i)\, p_x({\rm P}|{\rm B};\mathcal{M},f_k), \tag{22}$$

where f_k is the centre of a frequency bin; the summation sweeps all frequency bins within the PTA window (as long as the frequency of the last stable orbit, $f_{\rm lso}$, is not exceeded, and the interval of time until the coalescence is not longer than 0.1 Gyr). The PTA-galaxy probability of a system (with m and z within θ_i) in catalogue x is, therefore, its probability to be a B-galaxy that contains a MBHB (of chirp mass $\mathcal{M}$) producing a strain amplitude larger than $h_0^{\rm thres}$ within the PTA frequency band.

Note that later on we want to apply this machinery to galaxies in the real catalogue, where we do not know with certainty which systems are B-galaxies. For this reason we calculate $p_x({\rm B},{\rm P}|\theta_i,\mathcal{M})$ for all galaxies in the different catalogues, even if we know that they are N-galaxies. Also note that it is meaningless to talk about snapshots of the real catalogue (we observe only one snapshot of the Universe); however, we expect the probabilities $p_a^j({\rm P}|{\rm B};\mathcal{M},f)$ to have a similar behaviour in the simulated universe and in the real one. Hence we keep the definition of equation (20) to calculate $p_r^j({\rm P}|{\rm B};\mathcal{M},f)$, where $T^{\rm snap}(z)$ is the time between snapshots in the simulation.

[7] We point out that the minimum black hole mass found among B-galaxies in the fake catalogue is of $10^{6.2}\,{\rm M}_\odot$.

[8] This, in practice, does not affect the results, since massive binaries never spend a time exceeding $T^{\rm snap}(z)$ at an observable PTA frequency interval.

3.3 Including clustering in the search

We find that B-galaxies tend to cluster differently than N-galaxies (as will be shown in Section 4.1); this fact motivates us to refine the search by adding information about the clustering of galaxies. Characterizing galaxies by means of their clustering properties is a common technique in observational astrophysics (see e.g. Li et al. 2006a; Wang & White 2012). These investigations are usually carried out by using the two-point correlation function (TPCF; Peebles 1980; Hamilton 1993), which is defined by the joint probability of finding an object simultaneously in two volume elements separated by a certain distance. The comoving distance between two galaxies is simply calculated as

$$D_{1,2} = \sqrt{[X_2 - X_1]^2 + [Y_2 - Y_1]^2 + [Z_2 - Z_1]^2}, \tag{23}$$

where (X_j, Y_j, Z_j) are the Cartesian coordinates of galaxy j (for $j \in \{1, 2\}$), related to the equatorial coordinates by

$$\begin{cases} X_j = r(z)\cos({\rm Dec.})\cos({\rm RA}) \\ Y_j = r(z)\cos({\rm Dec.})\sin({\rm RA}) \\ Z_j = r(z)\sin({\rm Dec.}) \end{cases}. \tag{24}$$

These same equations are applied to obtain the positions of galaxies in the fake, the adapted, and the real catalogue. In the adapted and real catalogues z is affected by the peculiar movement of galaxies, but we neglect this effect when calculating distances. The comoving distance $r(z)$ is defined in equation (15), and the cosmological parameters are given in Section 2.2. One can find several definitions for the TPCF in the literature (Davis & Peebles 1983; Hamilton 1993; Landy & Szalay 1993), which account for possible biases and selection effects of the catalogue. The TPCF is thus a statistical tool that can be used to characterize the clustering of an ensemble of point particles; it is meaningless to talk about the TPCF of an individual galaxy.

Instead, we introduce the *number of neighbours at different shells* (NNDS): this is a set of numbers that measures how many galaxies are contained in spherical comoving shells around a selected object. For simplicity, we will use the term NNDS both for an ensemble of systems (in which case it denotes the average NNDS over all systems of the ensemble) and for individual systems. The systems for which the NNDS is calculated (i.e. systems that are at the centres of the shells when counting neighbours) are called *foreground galaxies*. The rest of the systems (that may be counted as neighbours of some other foreground galaxies) are *background galaxies*. For the calculation of the NNDS, systems in the *adapted* catalogue will always be the foreground galaxies; we then investigate two different cases.

(i) All galaxies of the fake catalogue are considered as background galaxies. Redshifts are not affected by peculiar velocities (they are cosmological redshifts).

(ii) The set of background galaxies is the same as that of foreground galaxies (i.e. galaxies of the adapted catalogue). Redshifts are affected by peculiar velocities (they are apparent redshifts).

The first case corresponds to an idealized case in which we have perfect knowledge of the positions of all galaxies in the universe. The second is a more realistic approach: the positions of galaxies cannot be precisely calculated (because we do not know the velocity and direction of the peculiar movement of the galaxies), and not all galaxies can be observed. For both cases, the NNDS is calculated for different *shell sets*, corresponding to different choices of the shells' sizes. The borders of the shells of some of the sets are separated

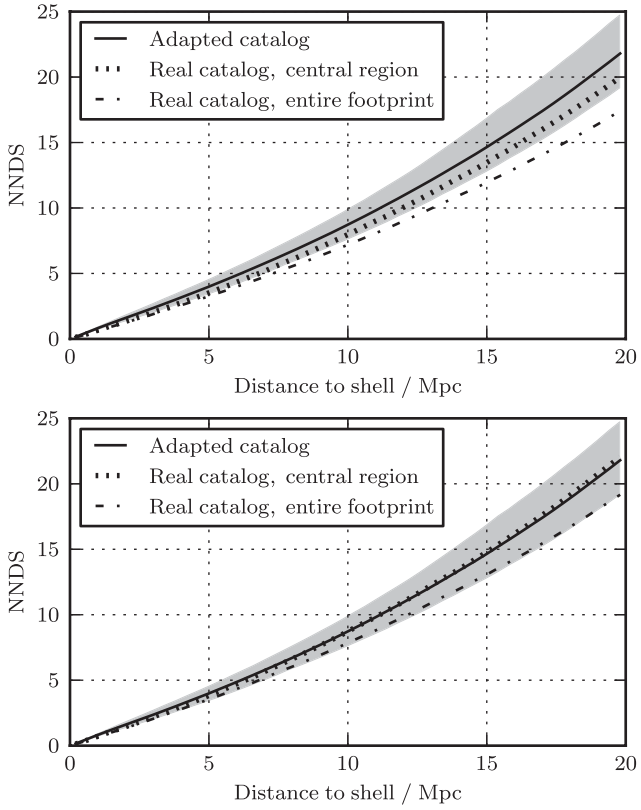

Figure 7. Average NNDS of systems in the adapted catalogue (solid line), in the real catalogue (dot–dashed line), and in a volume of the real catalogue that is not affected by the borders of the observed area of the sky (dotted line). The filled area contains the NNDS of each of the nine patches in which the adapted catalogue is divided. The number of neighbours is counted for each galaxy at 50 shells with borders equally separated by 400 kpc. For the upper plot, distances between systems of the real catalogue are calculated assuming the cosmological parameters used by the MPA/JHU, whereas for the lower plot the parameters are those of the MS.

linearly, and some other logarithmically. The number of shells in all sets equals 50.

Let us consider case (ii) and a shell set made of 50 shells with borders linearly separated by 400 kpc. For each galaxy, we count the number of neighbouring galaxies at a distance of less than 400 kpc; then we count the neighbours that are between 400 and 800 kpc away, then between 800 and 1200 kpc, etc. We keep counting neighbours at different shells until reaching a maximum distance of 20 Mpc (which corresponds to the 50th shell) from the initial galaxy. Then we average (over all galaxies) the number of neighbours at each shell. This is shown in the upper plot of Fig. 7. The dotted curve shows the NNDS of systems inside the selected central region of the SDSS sky (the grey area in Fig. 3). The solid line is the NNDS of all systems in the adapted catalogue (the filled area contains the NNDS of each of the nine patches of the sky into which the fake catalogue is divided). The NNDS is also calculated for systems (in the redshift range [0.01, 0.1]) from the entire real catalogue, including the region outside the selected central one (dot–dashed curve). The latter NNDS is affected by border effects: galaxies close to the border of the observed area of the sky (those galaxies on the black area of Fig. 3) have fewer neighbours. This effect is increasingly important (and the dot–dashed line differs more from the other two) as more distant shells are considered. The dotted and solid lines agree quite well, even if there are other sources of incompleteness in the SDSS catalogue (besides the effect of the border of the observed sky area) that are not taken into account.

As we mentioned in Section 2.2, the MPA/JHU uses different cosmological parameters than the MS. In the upper plot of Fig. 7, distances of systems from the real catalogue are calculated with the cosmological parameters used by the MPA/JHU; in the lower plot, these distances are calculated with the same parameters used by the MS, and the agreement is much better. The difference between these plots is the effect of the different sets of cosmological parameters. In order to properly compare the NNDS of the adapted and real catalogues, it would be convenient to use the updated simulated galaxy catalogue of the MS (Guo et al. 2013) (this issue is commented on in Section 5).

In Section 3.1 we described a search in which each system is characterized by two parameters, z and m; we now describe a search, carried out with a MLA, in which galaxies are characterized by a vector of 52 parameters, $\boldsymbol{\theta} = \{z, m, \mathrm{NNDS}^1, \ldots, \mathrm{NNDS}^{50}\}$, where NNDS^k gives the number of neighbours at the shell k (which goes from 1 to 50). Machine learning is a growing subject of artificial intelligence, and include a vast variety of techniques, in which a program is trained on a number of samples of data (which are characterized by one or more parameters $\boldsymbol{\theta}$ called *features*), and tries to predict the characteristics of different sets of data. We use supervised learning methods of the Scikit-learn[9] (Pedregosa et al. 2011) library of PYTHON[10] (Drake & van Rossum 2011); in particular, a method that seems to be particularly fast and effective is the Stochastic Gradient Descent. The algorithm uses a training set as a playground to set up the engine that afterwards assigns a probability of being a B-galaxy, $p_x(\mathrm{B}|\boldsymbol{\theta})$, to each of the elements in a testing set.[11]

The sky of the simulated universe is divided into nine patches; the systems in one of them make up our testing set. We subtract from the rest of the sky all systems which `galID` is contained in the testing set. This subtraction is performed to avoid systems in the training and in the testing set to be equal (these repetitions would artificially enhance the efficiency of the search). Then, from the part of the sky outside the testing set, we randomly pick 50 per cent of B-galaxies and 5 per cent of N-galaxies, to construct the training set.[12] The training set is used as input for the MLA. Then, the MLA calculates the probabilities $p_x(\mathrm{B}|\boldsymbol{\theta})$ of systems in the testing set. In Section 4.4 we show the efficiency of this search.

3.4 Extending the search to larger redshifts

The search in which the clustering is taken into account (described in Section 3.3) is performed at redshifts below 0.1. We do not attempt to extend this search to larger redshifts for several reasons. First, the completeness of a spectroscopic catalogue decreases with distance; the distinct features that are found in the NNDS of B-galaxies will vanish as more neighbours become unobservable. Secondly, we expect that the method we use to build the adapted catalogue will be less trustworthy (regarding the clustering) at larger redshifts. Finally, as will be shown in Section 4.4, the inclusion of the NNDS

[9] http://scikit-learn.org
[10] http://www.python.org/
[11] The mathematical machinery that is used by the algorithm to obtain these probabilities can be consulted at http://scikit-learn.org/stable/modules/sgd.html
[12] Note that the exact amount of B- and N-galaxies in the training set is not relevant for the results; one just needs to be sure that one has a large enough number of systems of each class to train the MLA properly.

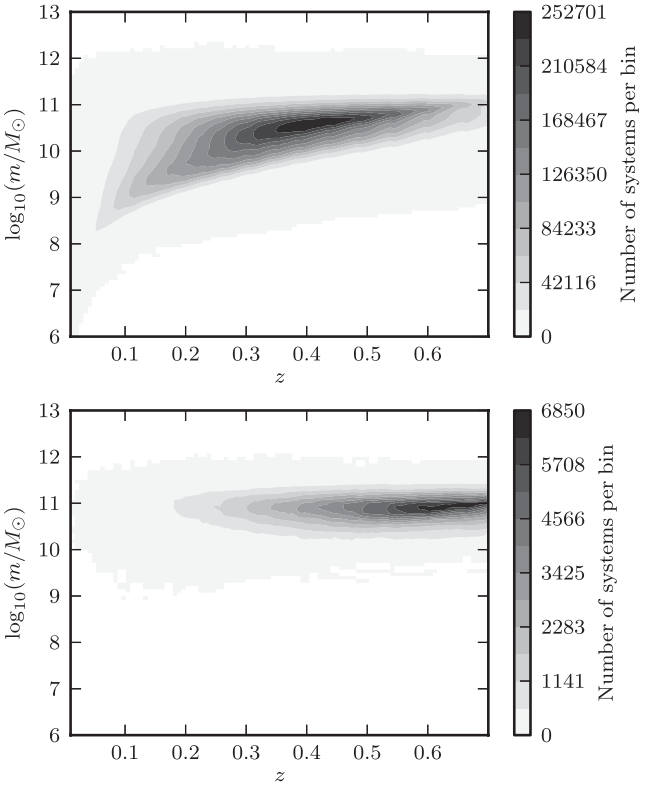

Figure 8. Contour plots of stellar mass versus apparent redshift of the entire fake catalogue (with $z < 0.7$). Both axes are divided into 100 bins; the grey-scale gives the number of systems contained in each pixel. Upper and lower plots consider all galaxies and B-galaxies, respectively.

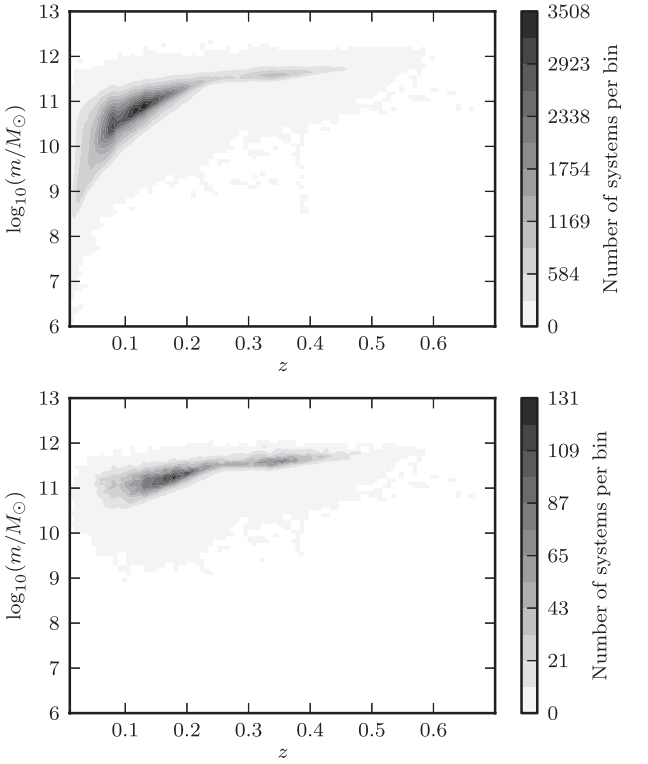

Figure 9. Contour plots of stellar mass versus apparent redshift of the extended adapted catalogue (that includes systems with $z < 0.7$). Both axes are divided into 100 bins; the grey-scale gives the number of systems contained in each pixel. Upper and lower plots consider all galaxies and B-galaxies, respectively.

in the search for B-galaxies is already quite ineffective in the local universe; there is no reason why it should improve for $z > 0.1$. Therefore, we spare ourselves the computationally intricate task of calculating the NNDS at larger redshifts.

Conversely, the simple Bayesian search described in Section 3.1 can be easily extended to $z_{\max}$. As equations (9)–(11) reveal, the only information we need to assign probabilities is the number of galaxies and B-galaxies within different z–m bins. We do not need to download all systems with $z < 0.7$ from the MS data base, but just a z–m histogram of components $n_f^{\rm G}(\theta_i)$, which means 10^4 integer numbers (for the choice of a 100×100 z–m grid) with the numbers of galaxies within each pixel, and a histogram of components $n_f^{\rm B}(\theta_i)$, with the number of B-galaxies. In Fig. 8, z–m histograms of galaxies and B-galaxies from the fake catalogue up to $z_{\max}$ are displayed as contour plots. For this extended fake catalogue, apparent redshifts are used.

There is a simple way to construct an adapted catalogue using only the functions $n_f^{\rm G}(\theta_i)$ and $n_f^{\rm B}(\theta_i)$, and the function $n_r^{\rm G}(\theta_i)$. The histogram components $n_a^{\rm G}(\theta_i)$ and $n_a^{\rm B}(\theta_i)$ can be calculated as

$$n_a^{\rm G}(\theta_i) = \min\left(1, \frac{n_r^{\rm G}(\theta_i)}{n_f^{\rm G}(\theta_i)}\right) n_f^{\rm G}(\theta_i), \tag{25}$$

and

$$n_a^{\rm B}(\theta_i) = \min\left(1, \frac{n_r^{\rm G}(\theta_i)}{n_f^{\rm G}(\theta_i)}\right) n_f^{\rm B}(\theta_i). \tag{26}$$

What we are imposing here is that the number of systems in a certain z–m bin of the adapted catalogue cannot be larger than the same bin of the real catalogue. Fig. 9 is analogous to 8, but for galaxies and B-galaxies of the adapted catalogue.

Once we have the functions $n_f^{\rm B}(\theta_i)$ and $n_f^{\rm G}(\theta_i)$ we can construct the B-galaxy probabilities in the extended fake catalogue, $p_f({\rm B}|\theta_i)$, using equation (9). Analogously, with the functions $n_a^{\rm B}(\theta_i)$ and $n_a^{\rm G}(\theta_i)$, the probabilities for the extended adapted and real catalogues can be calculated using equations (10) and (11), respectively. Finally, the probabilities $p_x({\rm B}|\theta_i)$ are smoothed with a Gaussian filter, as performed at $z < 0.1$.

4 RESULTS

4.1 Clustering of B-galaxies

In Section 3.3 we mentioned that B-galaxies present a characteristic clustering that could be used to distinguish them from N-galaxies; we now present this distinct shape of the NNDS of B-galaxies, and check how significantly it differs from the NNDS of N-galaxies, when using the fake and the adapted catalogues. First, the NNDS of B-galaxies, $\rm NNDS_B$, needs to be calculated. Then, for each B-galaxy, we find an N-galaxy that has similar mass and redshift of that B-galaxy [the matching tolerances are $\Delta\log_{10}(m/{\rm M}_\odot) = 0.1$ and $\Delta z = 0.001$]. The NNDS of these selected N-galaxies is calculated, $\rm NNDS_N$. The mean NNDS at each shell is obtained both for B- and for N-galaxies. In Fig. 10 the ratio $\rm NNDS_B/NNDS_N$ is plotted; the upper plots correspond to case (i) (as described in Section 3.3), in which all systems from the fake catalogue can be counted as neighbours; the lower plots correspond to case (ii), and only systems from the adapted catalogue can be background galaxies. The plots on the left are obtained for a set of shells with

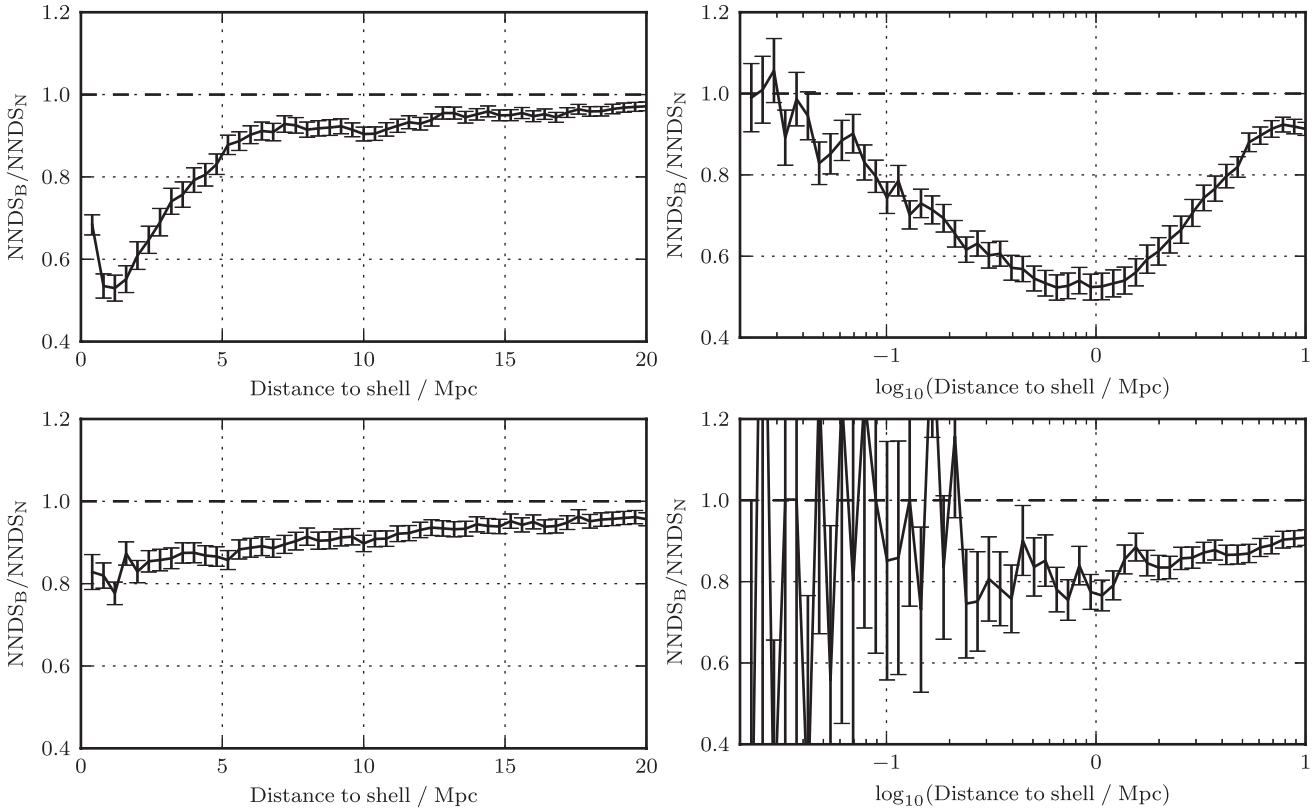

Figure 10. Ratio of the NNDS of B-galaxies over the one of N-galaxies, i.e. $\mathrm{NNDS_B/NNDS_N}$, for systems in the fake catalogue (upper plots) and in the adapted catalogue (lower plots). N-galaxies are chosen to have the same masses and redshifts as B-galaxies (with matching tolerances of $\Delta\log_{10}(m/\mathrm{M}_\odot) = 0.1$ and $\Delta z = 0.001$). Left- and right-hand plots correspond to different shell sets: shells with borders linearly separated from 0 to 20 Mpc (left) and logarithmically separated from 1 kpc to 10 Mpc (right). Similar patterns are found when using other shell sets and for other choices of matching N-galaxies. See the text for an explanation on the sizes of the error bars.

borders linearly separated by 400 kpc. The borders of the shells used for the right plots are logarithmically separated from 1 kpc to 10 Mpc. The error of $\mathrm{NNDS_B}$ and $\mathrm{NNDS_N}$ is assumed to be the square root of the variance of the mean; the error bars in Fig. 10 are the propagated error of the ratio of both quantities. We point out that the error bars in the lower-right plot are not meaningful at small distances; within those small shells, galaxies in the adapted catalogue usually count zero or at most a few neighbours.

There are two reasons why the ratio $\mathrm{NNDS_B/NNDS_N}$ is closer to one for systems in the adapted catalogue. First, distances in the adapted catalogue are calculated using apparent redshifts (affected by peculiar velocities), which introduce some level of randomness in the positions of the neighbours. Secondly, many of the neighbours have been deleted in the process of building the adapted catalogue, so the amount of information contained in the NNDS is smaller than when observing all galaxies.

Fig. 10 shows the interesting fact that B-galaxies present an underdensity of galaxies at ∼1 Mpc scale, when compared to N-galaxies of same redshift and mass. A similar pattern has already been found in the TPCF of narrow-line AGN in SDSS data (see e.g. fig. 3 of Li et al. 2006b, but note that, in that paper, projected proper distances are used, instead of spatial comoving distances). They find that this pattern is also typical of galaxies located at the centre of their dark matter haloes, where AGN preferentially reside. It is thus interesting to check whether or not the underdensity of neighbours is due to the relative position of B-galaxies within their haloes. This information can also be extracted from the MS data base, by means of the parameter `type` (see section 3.6 of Guo et al. 2011). Galaxies of `type` 0 are the principal galaxies of their haloes, whereas `type` 1 and 2 are satellite galaxies. In the fake catalogue we find that 80 per cent of B-galaxies are of `type` 0, and 20 per cent of `type` 1. No B-galaxy has `type` 2 (which corresponds to the so-called *orphan galaxies*). We construct a sample of N-galaxies that match mass, redshift, and `type` of the B-galaxies (this is the reason why the parameter `type` is included in the query of Appendix A), and recalculate $\mathrm{NNDS_B}$, $\mathrm{NNDS_N}$, and the ratio of both. In the upper plot of Fig. 11 we show the NNDS of B- and N-galaxies of `type` 0 and 1; the ratio $\mathrm{NNDS_B/NNDS_N}$ is shown in the lower plot. There we see that, in the case of `type` 0 galaxies, the underdensity is due to ∼ 1 neighbour at distances of a few Mpc, hence the clustering of `type` 0 B- and N-galaxies of similar masses and redshifts are almost indistinguishable. B-galaxies of `type` 1, however, still present a significant underdensity of neighbours with respect to matching N-galaxies at around 1 Mpc.

In Fig. 12 we show the NNDS of B- and N- galaxies at different mass intervals. Galaxies within the same mass interval have NNDS displayed with the same colour; the NNDS of B-galaxies is marked with diamonds. Here one can clearly see that B-galaxies have on average fewer neighbours (especially at distances of a few Mpc) than N-galaxies of similar masses. An important feature to note is that the larger the mass is, the larger the number of neighbours becomes, both for B- and for N-galaxies.

An interesting question is whether or not PTA-galaxies are more likely to be found in galaxy clusters. In a previous section we

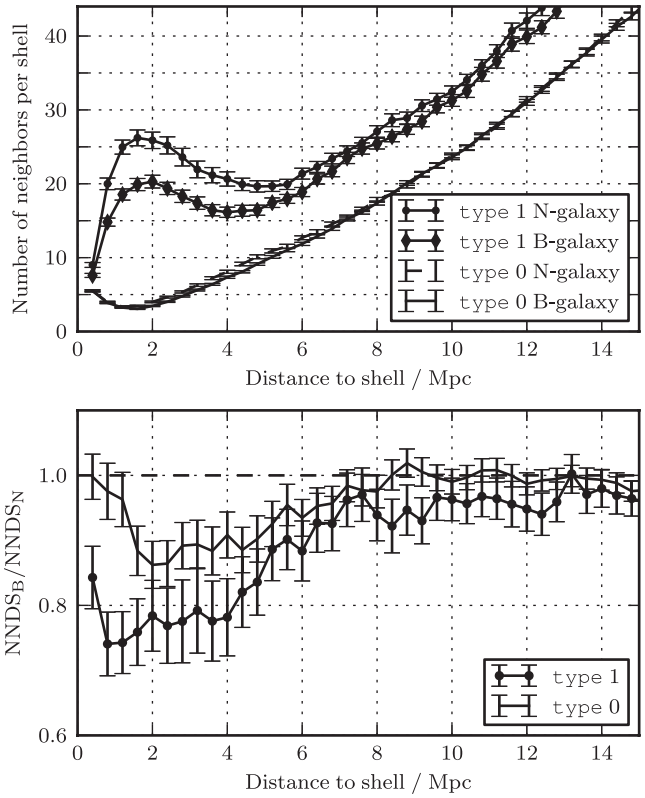

Figure 11. The upper plot shows the NNDS of N- and B-galaxies from the fake catalogue, of `type` 0 and 1 (which determines the relative position of the galaxies in their dark matter haloes). Galaxies of `type` 0 are central galaxies of their halo, whereas galaxies of `type` 1 are associated with a non-dominant subhalo. The ratios $\mathrm{NNDS_B/NNDS_N}$ of galaxies of same `type` are plotted below. N-galaxies are chosen in such a way that they match redshift, mass, and `type` of B-galaxies.

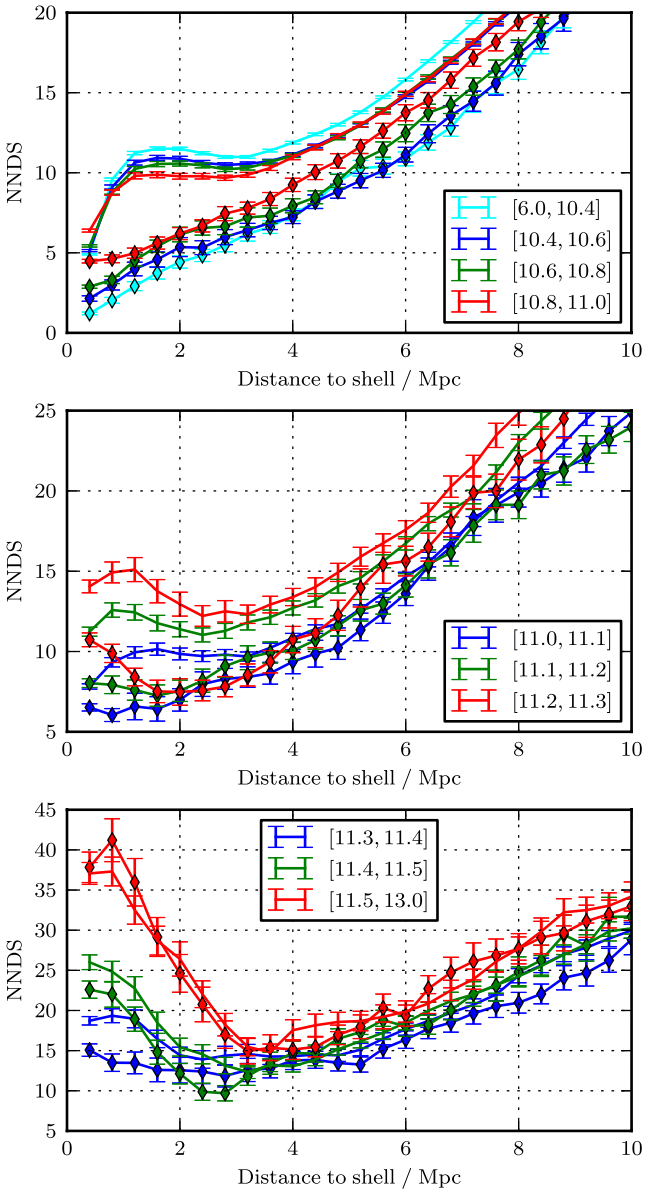

Figure 12. NNDS of B- and N-galaxies within different ranges of masses. Different intervals of masses m [depicted in the legends in units of $\log_{10}(m/\mathrm{M}_\odot)$] correspond to different colours. The NNDS of N- and B-galaxies are plotted using points and diamonds, respectively.

defined PTA-galaxies as B-galaxies that can contain a MBHB emitting GWs (of frequencies within the PTA window) that produces strain amplitudes larger than a certain threshold h_0^{thres}. According to equation (14), the strain amplitude is proportional to $\mathcal{M}^{5/3}$, so MBHBs need to be very massive to produce large strains. However, the time a binary spends in the PTA frequency band is proportional to $\mathcal{M}^{-5/3}$, meaning that more massive binaries are less likely to be found in the GW emission phase. There exists a trade-off between the two arguments, that happens to favour larger masses. In Section 4.6, we will make a list of real PTA-galaxy candidates; their masses lie between $\sim 10^{11.1}$ and $10^{12.1}\,\mathrm{M}_\odot$, with a mean of $10^{11.7}\,\mathrm{M}_\odot$. As Fig. 12 shows, galaxies with such masses tend to have significantly more neighbours than average (lower mass) galaxies. This argument is not enough to conclude that PTA-galaxies are usually in big galaxy clusters, but we can nonetheless claim that they are more likely located in dense neighbourhoods. A more precise answer to the question of PTA-galaxies being found preferentially or not in clusters could be achieved by performing an exhaustive study of the list of PTA-galaxy candidates of Section 4.6.

We remark that it is more customary to use projected distances in studies regarding galaxy clustering and neighbours, rather than three-dimensional distances. Indeed, the actual magnitude that we are able to observe from two distant objects is their two-dimensional distance (projected on the plane perpendicular to the line of sight), and it is often not possible to measure the redshift of all neighbours of a particular galaxy. Nevertheless, since the clustering study of this paper is restricted to the simulated universe (in which all spatial information is provided), the use of three-dimensional distances is sufficient. A more detailed investigation of the clustering properties of B-galaxies, using projected distances, may be subject of a forthcoming work.

4.2 Efficiency of the search for B-galaxies

In Section 3.1 we described how a training set of galaxies (chosen from one half of the simulated sky) is used to create a matrix of probabilities $p_x(\mathrm{B}|\theta_i)$; we now address the question of how well we can identify B-galaxies from the testing set (constructed with systems of the other half of the sky). Galaxies are randomly chosen from the testing set, until covering 19.5 per cent of the sky. We define a vector of threshold probabilities p_T, with components in the range [0, 1]. For each value of p_T, we count the number of N-galaxies

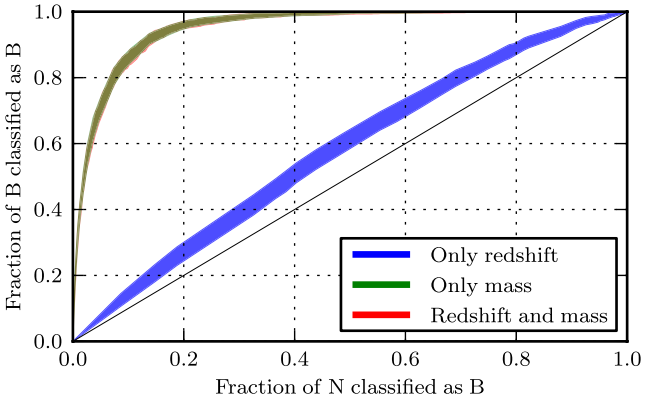

Figure 13. Results of 1000 realizations of the search for B-galaxies in the fake catalogue using the simple Bayesian approach described in Section 3.1. The coloured areas contain the ROC curves produced in each of the 1000 realizations. The red area contains the curves obtained when both mass and redshift are considered as parameters in the search. The green (or blue) areas contain the ROC curves obtained when characterizing galaxies only by their mass (or redshift). The brown area is the overlap of the red and green ones, that are almost identical.

classified as B-galaxies (i.e. with $p_f(\mathrm{N}|\theta_i) > p_T$) and the number of B-galaxies classified as B-galaxies (with $p_f(\mathrm{B}|\theta_i) > p_T$). This process is repeated 1000 times.

In Fig. 13 we plot *receiver operating characteristic* (ROC) curves of the search in the fake catalogue. A ROC graph represents the false alarm rate (number of N-galaxies classified as B-galaxies divided by the number of N-galaxies in the testing set) on the horizontal axis and the detection rate (number of B-galaxies classified as B-galaxies divided by the number of B-galaxies in the testing set) on the vertical one. The thin black line that crosses the plot diagonally would be the result of a totally random search (for each probability value, one gets the same fraction of good and bad classifications). The red area contains 1000 ROC curves (each one corresponding to a different testing set) for which galaxies are characterized by the parameters $\boldsymbol{\theta} = \{z, m\}$. The green area contains also 1000 ROC curves; in this case galaxies are characterized only by their mass, so $\boldsymbol{\theta} = \{m\}$. Finally, the blue area is filled with 1000 ROC curves for which galaxies have been characterized by their redshift, $\boldsymbol{\theta} = \{z\}$. This plot makes clear that the most useful piece of information to distinguish B-galaxies is their mass. The same procedure to test the efficiency of the search is applied to systems in the adapted catalogue. Fig. 14 is analogous to Fig. 13, but now the training and testing sets are constructed with systems from the adapted catalogue.

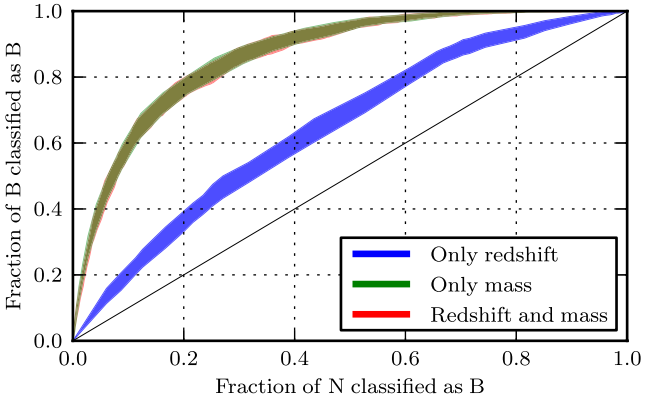

Figure 14. Analogous to Fig. 13; now only systems from the adapted catalogue are considered.

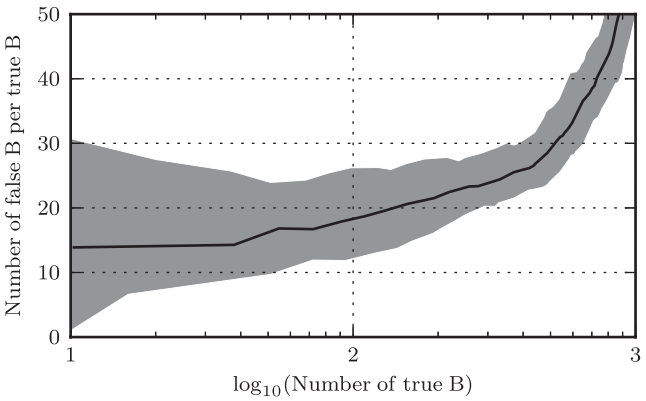

Figure 15. Number of wrong classifications per good one versus well-classified B-galaxies from the adapted catalogue, after using the search described in Section 3.1. This figure is a different representation of the data contained in Fig. 14. The grey area contains the curves corresponding to the 1000 realizations of the search (each one for a different choice of the galaxies in the testing set). The black curve is the average over the 1000 curves; each point corresponds to a value of B-galaxy probability.

ROC curves are commonly used to test the efficiency of classification algorithms, however, for our purposes, we find more convenient to present the data in a slightly different manner. In Fig. 15 we present the number of well-classified B-galaxies from the adapted catalogue (on the horizontal axis) and the number of incorrectly classified N-galaxies per well-classified B-galaxy (on the vertical axis). The grey area contains the curves obtained for the 1000 realizations; the black line is the average over all of them. Each point of a curve corresponds to a certain value of probability $p_a(\mathrm{B}|\theta_i)$. The points on the right part of the plot (where more B-galaxies are well classified) correspond to lower probabilities.

A probability threshold value p_T^{B} needs to be chosen, so that all galaxies with $p_a(\mathrm{B}|\theta_i) \geq p_T^{\mathrm{B}}$ are considered *candidates* for B-galaxies. We choose a threshold such that an average of ∼100 B-galaxies are counted as candidates. More precisely, for $p_T^{\mathrm{B}} = 4.16 \times 10^{-2}$, in the worst realizations we find 82 B-galaxies among 2166 candidates, whereas in the best ones, 143 B-galaxies out of 2106 galaxies are found. The average result produces 110 B-galaxies among 2168 candidates. Further on we apply this same search to galaxies in the real catalogue. B-galaxy candidates are chosen using the same probability threshold p_T^{B}, so we expect to have a similar number of galaxies with $p_r(\mathrm{B}|\theta_i) \geq p_T^{\mathrm{B}}$, and thus a similar number of real B-galaxies among them too. We point out that our choice of p_T^{B} is arbitrary. One could have chosen a smaller one, and more B-galaxies would be counted among the candidates; nonetheless, as Fig. 15 shows, for a smaller threshold the ratio of N-galaxies per B-galaxy considered as candidates would also be larger.

4.3 Efficiency of the search for PTA-galaxies

We now test the efficiency of the search for PTA-galaxies in the adapted catalogue, following a similar procedure to the one explained in Section 4.2. The strain amplitude threshold is set to $h_0^{\mathrm{thres}} = 10^{-15}$; with this we calculate, for each system, its PTA-galaxy probability, $p_x(\mathrm{B}, \mathrm{P}|\theta_i, \mathcal{M})$. Because of the iterative nature of the procedure used to assign PTA-galaxy probabilities, systems cannot be repeated: even if two galaxies have the same `galID` and m, the masses of their MBHs will be different, leading to different

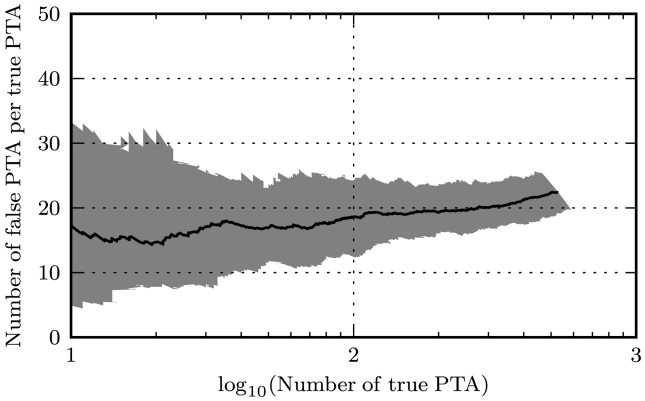

Figure 16. Efficiency of the search for PTA-galaxies in the adapted catalogue. This plot is analogous to that in Fig. 15, but now using the probabilities $p_a(\mathrm{B}, \mathrm{P}|\theta_i, \mathcal{M})$. Hence, galaxies in this plot are selected such that their hosted MBH produces (if it is a binary) a strain amplitude larger than the threshold $h_0^{\mathrm{thres}} = 10^{-15}$.

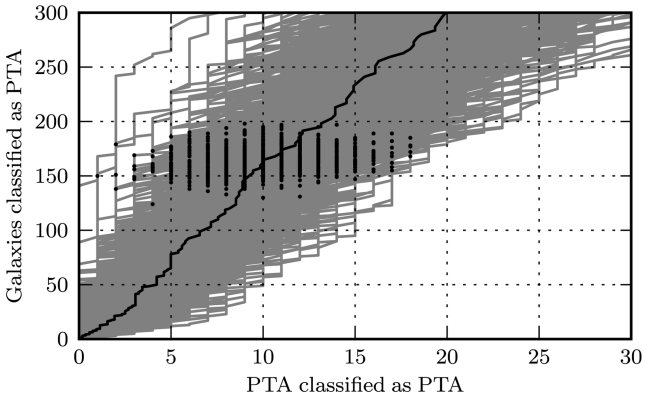

Figure 17. Total number of galaxies classified as PTA-galaxies versus the number of good classifications. Each grey curve corresponds to a different realization (a different choice of galaxies from the adapted catalogue covering 19.5 per cent of the sky). The black line is the average over all 1000 realizations. For each realization, we plot a black point corresponding to the PTA-galaxy probability threshold p_T^{PTA}. Galaxies considered here are such that their hosted MBH produces (if it is a binary) a strain amplitude larger than the threshold $h_0^{\mathrm{thres}} = 10^{-15}$.

probabilities. For this reason, the adapted catalogue is not divided into a training and a testing set. Galaxies (that could contain a MBHB producing a strain amplitude larger than h_0^{thres}) are chosen randomly from the whole simulated sky until covering an area of the sky of 19.5 per cent; we then count the number of B- and N-galaxies passing each value of the probability threshold. This procedure is repeated 1000 times.

Fig. 16 is analogous to 15, but now the PTA-galaxy probabilities are used, instead of the B-galaxy probabilities, to test the efficiency of the search in the adapted catalogue. This plot is used to decide upon a probability threshold p_T^{PTA} to select our real PTA-galaxy candidates later on. Let us first consider the case in which the chosen probability threshold corresponds to the rightmost point on the plot. The worst results for such a threshold find 462 PTA-galaxies among 11 764 candidates, whereas the best results have 588 PTA-galaxies of 11 610 candidates. The average is a number of 525 PTA-galaxies of a total of 11 770 candidates. An important remark is that these $\sim$500 PTA-galaxies are all B-galaxies with $z < 0.1$ that produce a strain amplitude larger than 10^{-15} in the adapted catalogue. In other words, *all* possible PTA-galaxies contained in the adapted catalogue are counted as candidates, by choosing such a probability threshold.

A detection rate of 100 per cent is a great feature, however, the false alarm rate is large. The list of candidates can be reduced, by choosing a larger probability threshold, so that the number of bad candidates per good one decreases (by the cost of losing PTA-galaxies). We now want to select a threshold such that only the $\sim$10 most likely PTA-galaxies are counted as candidates. For that, we choose a probability threshold corresponding to the leftmost point in Fig. 16. This threshold is $p_T^{\mathrm{PTA}} = 8.01 \times 10^{-06}$. Now, in the worst realizations, 1 PTA-galaxy is found among 190 candidates; whereas in the best realizations 18 PTA-galaxies are counted among 137 candidates. On average, we count 10 PTA-galaxies in a list of 164 candidates.

In Fig. 17 we have plotted the same data as in Fig. 16 in a different manner (and at a different range of probabilities). Each grey line corresponds to one realization; the black line contains the average values over all (1000) realizations. Each one of the 1000 black points gives the number of galaxies (on the vertical axis) and PTA-galaxies (on the horizontal axis), for a particular realization, that have PTA-galaxy probabilities larger than p_T^{PTA}. Here we see that, for such a threshold, $\sim$10 galaxies among the $\sim$150 selected as candidates are well classified. We expect similar numbers when applying the search to the real catalogue.

4.4 Efficiency of the search when including clustering information

We now turn to the results of the search including clustering information, as described in Section 3.3. We applied the MLA 1000 times; for each realization, the systems of the training set were different, but the testing set was always composed of the galaxies of one particular patch of the sky (of the nine into which the adapted catalogue is divided). After each realization we construct a ROC curve with the probabilities assigned to the B- and N-galaxies of the testing set. In order to determine the amount of information provided by the parameters (z, m, and NNDS), the MLA is used in three different circumstances. First, we characterize galaxies only using the NNDS (blue curves in the plots of Fig. 18). Secondly, only z and m are used as features (green curves). Thirdly, all pieces of information (z, m, and NNDS) are considered (red curves). The NNDS is always calculated for 50 shells with borders separated from 0 to 4 Mpc in linear steps.

Fig. 18 shows the ROC curves resulting from this search. Upper and lower plots correspond to cases (i) and (ii) (as described in Section 3.3), respectively. The coloured areas contain all ROC curves after the 1000 realizations. In the lower plot, the NNDS is affected by the inclusion of peculiar velocities and by the subtraction of neighbours; the efficiency of the search using only the NNDS is not significantly different from a totally random search. A discouraging result is that the inclusion of the NNDS introduces some level of randomness in the search. In the upper plot of Fig. 18, we see that the red and green areas are approximately equal; but in the lower plot, the red area is shifted to the right (to higher values of false alarm rate) with respect to the green one.

Summarizing, the inclusion of the clustering does not ameliorate [in case (i)] or even deteriorates [in case (ii)] the efficiency of the search. However, the fact that the blue curve differs from the black thin line, in the upper plot, means that the clustering of B-galaxies

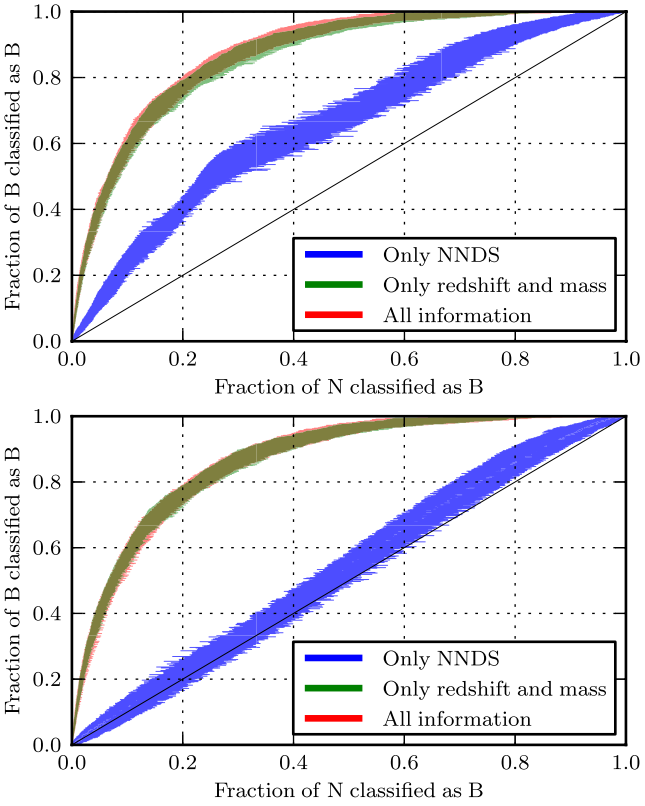

Figure 18. ROC curves obtained for the search for B-galaxies using a MLA. The upper and lower plots correspond to case (i) and (ii) (described in Section 3.3), respectively. Filled areas contain all the curves produced after the 1000 realizations. Blue and green curves correspond to searches in which only the NNDS, and only z–m, respectively, are taken as features. Red curves are the results of searches in which all parameters (z, m, and NNDS) are considered.

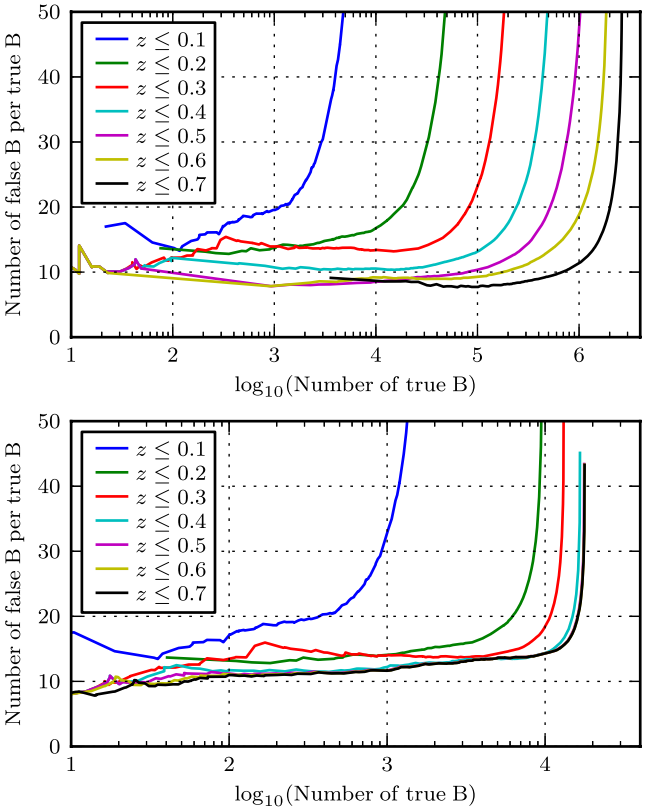

Figure 19. Estimate of the efficiency of a search for B-galaxies extended to larger redshifts (up to 0.7). The plot shows the ratio of bad classifications per good one versus good classifications of B-galaxies, within different redshift intervals. The upper (or lower) plot is obtained when applying the extended search to the fake (or adapted) catalogue.

presents in fact features that distinguish them from N-galaxies. The plots in Fig. 18 change, although not drastically, by using other choices of the shells, reducing the number of shells (from 50 to 10 or 20), or assigning the testing set to another patch of the sky. Nevertheless, the conclusions already drawn about the efficiency of the searches using the NNDS remain unaffected. By comparing the ROC curves obtained in this and in Section 4.2, one can also conclude that the efficiency of our Bayesian search is consistent to that of the MLA (when characterizing galaxies only by their redshift and mass).

4.5 Efficiency of the search at larger redshifts

Here we show how well the search for B-galaxies performs at redshifts as large as 0.7. The probabilities $p_x(\mathrm{B}|\theta_i)$ are calculated following the method described in Section 3.4. We build a z–m histogram of 100×100 pixels; redshifts and masses are in the ranges $[z_{\min}, z_{\max}]$ and $[m_{\min}, m_{\max}]$, respectively. Each pixel is assigned a value of $p_x(\mathrm{B}|\theta_i)$. We then count the number of B- and N-galaxies contained in pixels with probabilities larger than a certain value that goes from 0 to 1. We do this for different maximum redshifts: 0.1, 0.2, ..., 0.7. The results are depicted in Fig. 19, that contains two plots similar to those in Figs 15 and 16. Upper and lower plots correspond to the probabilities $p_f(\mathrm{B}|\theta_i)$ (using the fake catalogue) and $p_a(\mathrm{B}|\theta_i)$ (using the adapted catalogue), respectively.

One notices with Fig. 19 that the number of bad classifications per good one does not grow (and even decreases) when considering galaxies at larger redshifts. For example, considering systems in the adapted catalogue all the way up to $z_{\max}$, there exists a certain probability threshold such that ∼1000 B-galaxies (and a factor of ∼11 more N-galaxies) have larger probabilities than that threshold. Applying this search to the real catalogue we could then make a list with ∼11 000 candidates containing ∼1000 true B-galaxies. If the trend found on the search for PTA-galaxies at $z < 0.1$ (in Section 4.3) also holds at larger redshifts, we could then presumably make a list with the, say, ∼1000 most likely single PTA sources contained in the SDSS footprint (for $z < 0.7$).

Although extending the searches to larger redshifts looks potentially very interesting, such a task is not performed for the moment. We believe that, to create a trustworthy list of real PTA-galaxy candidates at such large redshifts, the method used to adapt our fake catalogue to the limitations of the real one should be more accurate than that described in Section 3.4. A proper extension of the search towards larger redshift is thus left for a possible follow-up work.

4.6 Assigning probabilities to real galaxies

A B-galaxy probability is now assigned to each system in the real catalogue. This probability, determined by the matrix $p_r(\mathrm{B}|\theta_i)$ (whose construction is explained in Section 3.1), depends only on the bin θ_i in which the values of z and m of the galaxy are contained. In Fig. 20 we plot a projected skymap of the systems from the real catalogue that have probabilities larger than the threshold p_T^{B}

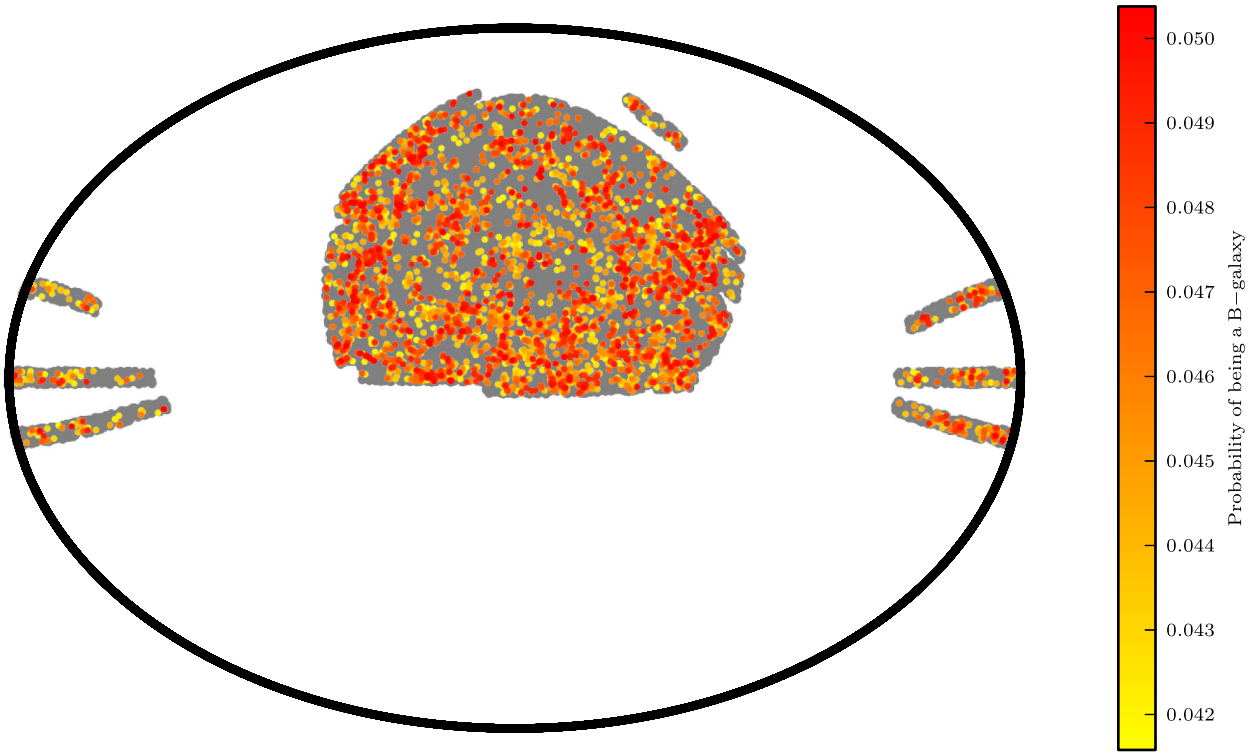

Figure 20. Projected skymap (using a Hammer projection) of galaxies from the real catalogue. Coloured points are B-galaxies candidates. The colour bar gives the B-galaxy probability $p_r(\mathrm{B}|\theta_i)$; redder points are galaxies with larger probabilities. Grey points are galaxies with probabilities below the chosen threshold p_T^{B}. From the 3870 candidates we expect ∼196 of them to be real B-galaxies.

introduced in Section 4.2. All galaxies that are not candidates are grey points; B-galaxy candidates have colours corresponding to different values of probabilities $p_r(\mathrm{B}|\theta_i)$. Redder points are candidates with larger B-galaxy probabilities. In this skymap there are 3870 candidates, a factor of ∼1.79 more systems than in the adapted catalogue for the same probability threshold p_T^{B}. The reason for this is the overabundance of high-mass galaxies in the SDSS with respect to the MS, discussed in Section 5. We thus expect that, among these candidates, $\sim 1.79 \times 110 \approx 196$ are true B-galaxies (since 110 was the number of B-galaxies found in the adapted catalogue for the same probability threshold).

In order to assign to each real galaxy its probability of hosting a MBHB observable by the PTA, $p_r(\mathrm{B}, \mathrm{P}|\theta_i, \mathcal{M})$, we follow the prescriptions described in Section 3.2; for that we first need the bulge mass m_{bulge} of each galaxy,

$$m_{\mathrm{bulge}} = f_{\mathrm{b}} m, \tag{27}$$

where f_{b} is the bulge mass fraction. Elliptical galaxies are expected to have f_{b} close to 1, while for spiral galaxies, reasonable values of f_{b} lie between ∼0.1 and ∼0.3. We now explain how these bulge mass fractions have been obtained. The Galaxy Zoo (Lintott et al. 2008, 2011) is a project in which volunteers assign SDSS galaxies a morphological classification by visual criterion. The data are public[13] and contain a final morphological-type classification, constructed after processing the votes of the volunteers and reducing possible visual biases. The three possible types are 'elliptical', 'spiral', and 'unknown' (in those cases when the voting for elliptical or spiral was not significant enough). Unfortunately, not all galaxies in our real catalogue have a final classification in the Galaxy Zoo. We use those galaxies with a type classification different than 'unknown' that are contained in both catalogues, to adopt a criterion on the morphologies of all galaxies in our real catalogue.

In Fig. 21 we show contour plots of galaxies classified in the Galaxy Zoo as ellipticals (on top) and as spirals (bottom). The horizontal axis shows the *surface mass density* (SMD), defined as

$$\mathrm{SMD} = \log_{10}\left(\frac{m}{2\pi R^2}\frac{\mathrm{kpc}^2}{\mathrm{M}_\odot}\right). \tag{28}$$

The radius R is the half-light proper radius in the z-band, calculated using

$$R = [1+z]^{-1} r(z) \mathcal{R}_{50,z}, \tag{29}$$

where $r(z)$ is the comoving distance to the galaxy (equation 15), and $\mathcal{R}_{50,z}$ is the angular radius in which 50 per cent of the Petrosian flux in the z-band is contained (called `petroR50_z` in the SDSS server). The vertical axis shows the concentration parameter C, for which we use the definition

$$C = \frac{\mathcal{R}_{50,\mathrm{r}}}{\mathcal{R}_{90,\mathrm{r}}}. \tag{30}$$

This means, C is the ratio of the 50 and 90 per cent Petrosian radii in the r-band (called `petroR50_r` and `petroR90_r` in the SDSS server, respectively). The query used to obtain the parameters `petroR50_z`, `petroR50_r`, and `petroR90_r` is included in Appendix A. From Fig. 21 we see that elliptical and spiral galaxies cannot be clearly distinguished in a certain region of the SMD–C plane; there is an important overlap. The darkest regions in the upper and lower plots show the maximum accumulation of elliptical and spiral galaxies, respectively. The intermediate value of C between those two maxima is $\overline{C} = 0.335$ (which is plotted as a black horizontal line). The vertical line is at the value of SMD above which

[13] http://data.galaxyzoo.org

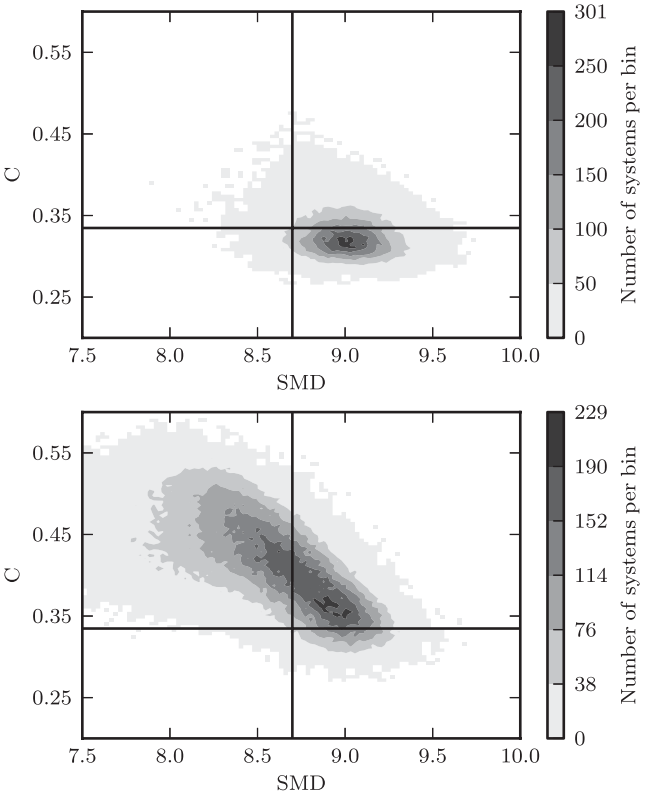

Figure 21. Contour plots of the distribution of galaxies classified as elliptical (above) and spiral (below) in the Galaxy Zoo. The horizontal axis shows the SMD, and the vertical one is the concentration parameter. Both axes are divided into 100 linearly spaced bins, and the colour bar gives the number of systems per bin. We use these distributions to find a criterion to calculate the bulge mass fractions of systems in our real catalogue.

90 per cent of elliptical galaxies are contained, $\overline{\mathrm{SMD}} = 8.70$. We take these two values and construct the functions

$$f_{\mathrm{b}}^{\mathrm{SMD}}(c_s, \overline{\mathrm{SMD}}, \mathrm{SMD}) = \min\left(1, \exp\left(c_s[\mathrm{SMD} - \overline{\mathrm{SMD}}]\right)\right) \quad (31)$$

and

$$f_{\mathrm{b}}^{C}(c_r, \overline{C}, C) = \min\left(1, \exp\left(c_r[C - \overline{C}]\right)\right). \quad (32)$$

Then, the bulge mass fraction is calculated as the product of the two previous functions,

$$f_{\mathrm{b}}(\mathrm{SMD}, C) = f_{\mathrm{b}}^{\mathrm{SMD}}(7, 8.70, \mathrm{SMD}) f_{\mathrm{b}}^{C}(-13, 0.335, C), \quad (33)$$

where the parameters c_s and c_r have been chosen in such a way that the average f_{b} is $\sim$0.9 for elliptical galaxies and $\sim$0.3 for spiral galaxies. We point out that this particular choice of functions and parameters is arbitrary; the aim of this calculation is to construct a simple procedure to assign bulge mass fractions based on observational data. In Fig. 22, a contour plot of $f_{\mathrm{b}}(\mathrm{SMD}, C)$ is shown, and on top we have superimposed the distributions of ellipticals (on green) and spirals (on red) previously shown (in Fig. 21).

Once the bulge masses are known, the rest of the calculation of the PTA-galaxy probabilities is as described in Section 3.2. The list of real PTA-galaxy candidates is constructed as explained in Section 4.3: we select galaxies from the real catalogue that have PTA-galaxy probabilities larger than p_T^{PTA}. In Fig. 23 we show a projected skymap with these candidates. They are 232, which is a factor of $\sim$1.41 larger than the average number of candidates produced in the adapted catalogue. Therefore, we expect that $\sim 1.41 \times 10 \approx 14$

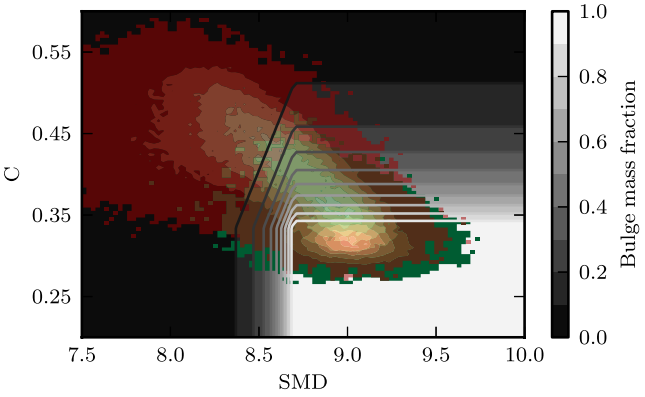

Figure 22. Contour plot of the function $f_b(\mathrm{SMD}, C)$, that gives, at each point of the SMD-C plane, the bulge mass fraction that we assign to galaxies in our real catalogue. Galaxies falling within the lower-right square (with a white background) will have a bulge mass fraction equal to 1, whereas the bulge mass fraction of galaxies far from these region will decay exponentially. On top of it, the contour plots of Fig. 21 have been superimposed: elliptical galaxies from the Galaxy Zoo on green and spiral ones on red.

true PTA-galaxies are counted among them. These candidates are the most likely single PTA sources in the local Universe (contained in the SDSS window at $z < 0.1$). Note that we do not expect to observe $\sim$14 MBHBs emitting at $h_0 \geq 10^{-15}$ in the SDSS window; in fact, the PTA-galaxy probability of each of the candidates is fairly small (see the numbers in the colour bar of Fig. 23), since they spend a relatively short interval of time emitting at frequencies at which they are observable. However, if we do observe a PTA source from this part of the sky with $z < 0.1$, it will most likely be one of these candidates. If one wanted to perform a targeted search, one could use this list of galaxies; the list could also be reduced by combining ours with other searching criteria, for example, by looking for signs of recent galaxy interaction in the SDSS images (that can be accessed from the SDSS server).

As we pointed out in Section 4.3, we could also construct a list of PTA-galaxy candidates such that *all* possible PTA-galaxies observed by the SDSS are counted. Taking into account this factor of $\sim$1.41 difference with respect to the adapted catalogue, the resulting list would contain $\sim 1.41 \times 11770 \approx 1.66 \times 10^4$ galaxies, of which $\sim 1.41 \times 525 \approx 740$ would be PTA-galaxies. They would be the only galaxies in the local Universe that could possibly be observed emitting GWs of a strain amplitude $h_0 \geq 10^{-15}$ within the spectroscopic SDSS catalogue.

In Fig. 24 the strain amplitude of systems in the real catalogue is plotted as a function of the observed GW frequency. Each point of the upper plot gives the average number of galaxies from the real catalogue with $z < 0.1$ that can be found within a certain frequency bin (whose central frequency is given by the horizontal axis), producing a strain amplitude larger than a certain value (given by the vertical axis). The numbers written on top of the graph are the sum over all frequency bins of the window, for a particular strain amplitude threshold (10^{-17}, 10^{-16}, 10^{-15}, and 10^{-14}). For example, we expect on average 1.1 systems in the real catalogue producing a strain amplitude larger than 10^{-16} within the PTA frequency window. The number of systems is calculated as the sum of the probabilities $p_r(\mathrm{B}, \mathrm{P}|\theta_i, \mathcal{M})$ of all galaxies. The lower plot in Fig. 24 represents the same as the upper one, but for systems in the adapted catalogue. We see that the numbers in both plots agree well; although the numbers are slightly smaller in the adapted catalogue case, again due to the shortage of systems at the high-mass

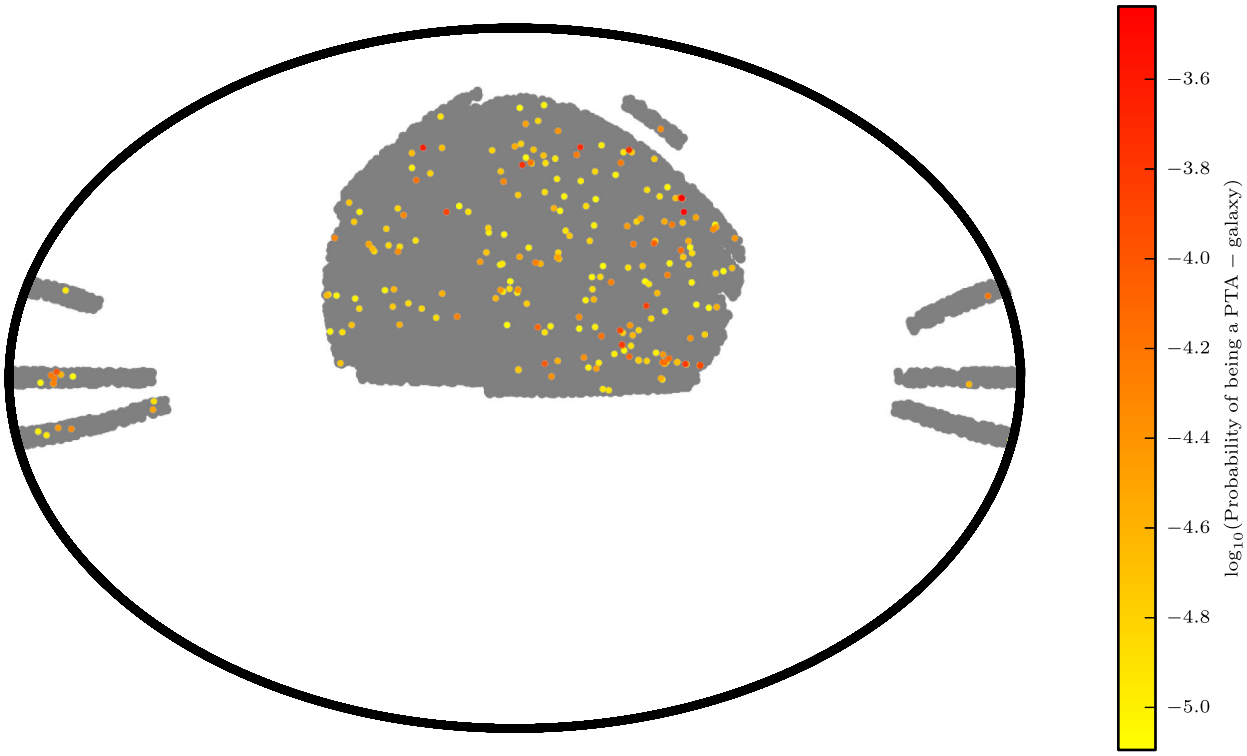

Figure 23. Projected skymap of galaxies from the real catalogue (analogous to Fig. 20). Now the colour bar gives the PTA-galaxy probabilities. Coloured points in this plot correspond to real PTA-galaxy candidates, i.e. galaxies that may host a MBHB emitting GWs that produce a maximum strain amplitude larger than $h_0^{\rm thres} = 10^{-15}$. We expect ∼14 of these 232 candidates to actually be B-galaxies (so they may contain an observable MBHB). The probability of observing one of them is small (as the numbers in the colour bar reveal); however, if we do observe a single source from this region of the sky (at $z < 0.1$) with a sensitivity of $h_0^{\rm thres}$, it will most likely be one of these candidates.

end, with respect to the real catalogue (see Section 5). In Fig. 25 the total number of PTA-galaxies expected to be observed with $z < 0.1$ is plotted against the strain amplitude threshold. The grey area contains the curves obtained for each of the 100 realizations described in Section 3.2. The black line is the average over all realizations. Upper and lower plots count systems from the real and adapted catalogues, respectively.

The upper plot in Fig. 26 is analogous to the upper plot in Fig. 24, but now considering systems with $z < 0.7$. Still, in order to count an average number of PTA-galaxies larger than 1 in the real catalogue, a strain amplitude threshold smaller than $\sim 10^{-16}$ is necessary. The lower plot in Fig. 26 shows the same as the upper one, but considering only B-galaxies from the whole fake catalogue. In this case, the probabilities $p_f(\mathrm{B}, \mathrm{P}|\theta_i, \mathcal{M}, f)$ are also defined by equation (21), but the B-galaxy probabilities $p_f(\mathrm{B}|\theta_i)$ are identically 1 for all galaxies (since we know they are B-galaxies). Hence, this graph reveals the average number of PTA-galaxies that would be contained in an ideal all-sky galaxy catalogue (more complete than a spectroscopic catalogue like the one we use). One could plot, on top of the graphs in Fig. 26, the exact sensitivity of the PTA to single MBHBs for a given array of MSPs (Ellis et al. 2012); the sum of the pixels swept by the sensitivity curve would give the average number of systems that should be observable for such an array. The total number of PTA-galaxies with $z < 0.7$ as a function of the strain amplitude threshold is shown in Fig. 27, for galaxies of the real catalogue (upper plot) and for B-galaxies of the entire fake catalogue (lower plot). The upper plot is analogous to the upper plot in Fig. 25, but now considering all systems in the real catalogue, and not only those with $z < 0.1$. The lower graph informs on the average number of PTA-galaxies that could be observed with an ideally complete all-sky galaxy catalogue.

5 CAVEATS AND ASSUMPTIONS

Our selection of B-galaxies is based on the assumption that MBHBs form and coalesce in less than ∼300 Myr, which is the approximate time between MS snapshots (at least for $z < 0.7$). In practice, in the MS a galaxy merger starts when the dark matter subhalo of the satellite galaxy is disrupted, and ends after a dynamical friction time computed according to equation 31 of Guo et al. (2011) (from Binney & Tremaine 2008). It is at this moment that the system is recorded as 'merger'; however, the two MBHs are likely still orbiting each other at a separation of few parsecs, because dynamical friction becomes inefficient on them once they have paired in a Keplerian binary.

Let us suppose that a merger occurred between two consecutive snapshots, $n - 1$ and n. This merger is first registered by the simulation at snapshot n. At this moment, the descendant galaxy has three possibilities: (i) it contains a single MBH (because the MBHB has already coalesced between $n - 1$ and n), (ii) it contains a wide MBHB (that is not yet emitting strong enough GWs to be observed), (iii) it contains a MBHB that is emitting observable GWs (and will coalesce well before $n + 1$). Ideally many B-galaxies would be in (iii), however, this is in reality the least likely case; the time a MBHB emits GWs that produce a strain on Earth larger than, for instance, 10^{-15}, is usually much smaller than the time between

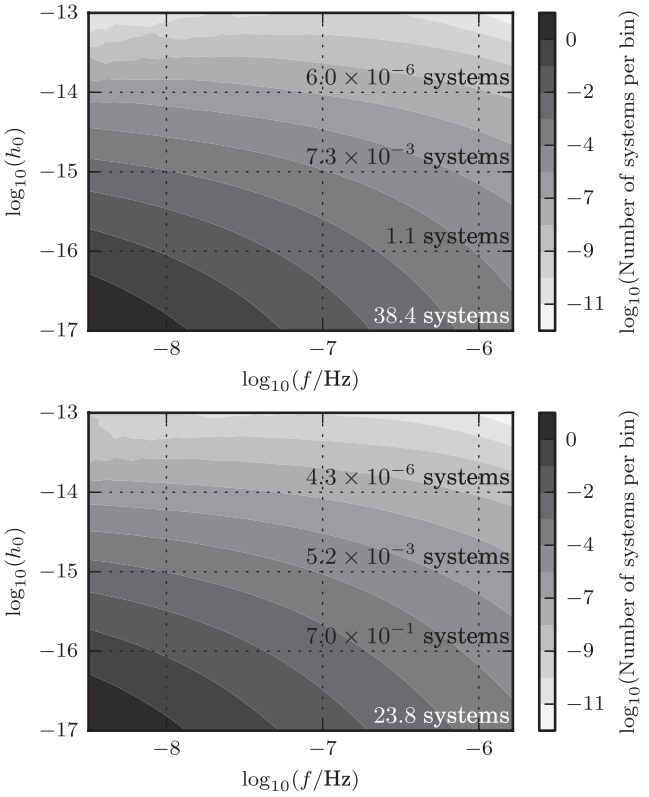

Figure 24. Strain amplitude versus observed GW frequency (averaged over 100 realizations). The horizontal axis is divided into 50 frequency bins, and the vertical axis into 50 strain amplitude bins. The colour of each pixel in the upper plot gives the sum of probabilities of all galaxies from the real catalogue (with $z < 0.1$) that produce a strain amplitude larger than a certain value (given in the vertical axis) within a certain frequency bin (given in the horizontal axis). The numbers written on top give the sum of probabilities over all frequency bins at a fixed strain amplitude, i.e. these are the average numbers of PTA-galaxies (for different strain amplitude thresholds) that are contained in the real catalogue at $z < 0.1$. The lower plot is analogous to the upper one, but for systems in the adapted catalogue.

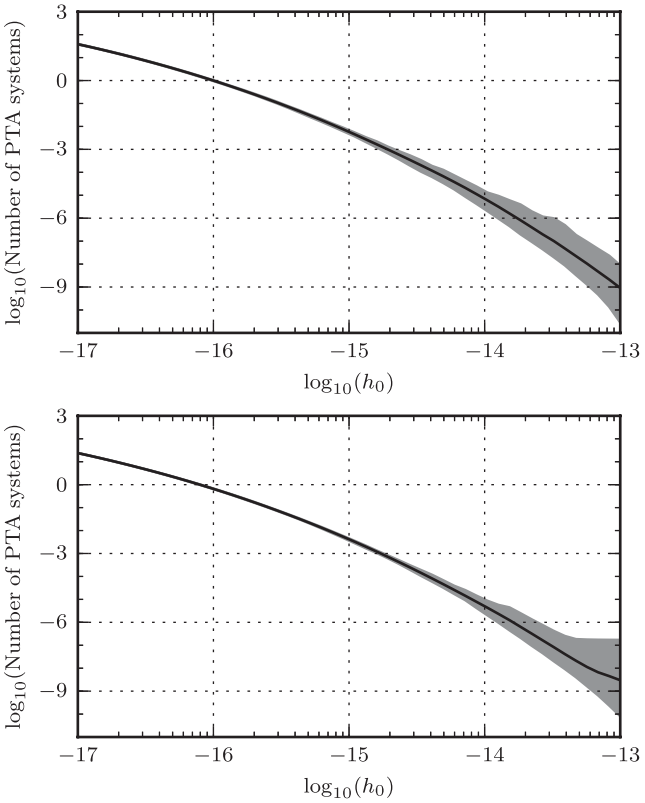

Figure 25. Average number of PTA-galaxies that are contained in the real (above) and adapted (below) catalogue for $z < 0.1$, as a function of the strain amplitude threshold. For example, to have an average of 1 galaxy in the real catalogue, the PTA needs to be able to detect strain amplitudes smaller than 10^{-16}. These plots are obtained by integrating the plots in Fig. 24 over all frequencies in the PTA frequency window.

snapshots.[14] Hence, we do not expect to have many B-galaxies in (iii), but we do assume the most of the coalescences to occur closer to n than either $n - 1$ or $n + 1$, i.e. that the binary lifetime is less than ∼300 Myr. This assumption is supported by recent simulations of MBHB mergers in stellar bulges (Preto et al. 2011; Khan et al. 2012), which found that the 'hardening phase' (i.e. the lifetime) of most MBHBs would last at most a few hundred Myr. While the role and impact of cold gas on such massive systems is less clear, any additional source of friction would help making this time-scale shorter.

Suppose now that the mechanisms taking part after the dynamical friction evolve over time-scales larger than ∼300 Myr. Then, the systems selected in n as B-galaxies should be those galaxies that merged between $n - 2$ and $n - 1$, or even at earlier snapshots, in order to account for the delay. This alternative selection would introduce a displacement in the redshift distribution of B-galaxies. However, as Figs 13 and 14 show, redshifts do not ameliorate much the search for B-galaxies, so the displacement would not change our results considerably. The discriminating parameter in the search is the mass, which is not expected to vary significantly between two snapshots after the merger. As a conclusion, as long as the coalescence is reached in a lapse that is not significantly larger than a few hundreds of Myr, the search is not affected. Nonetheless, for completeness, an improvement of this work could take into account possible longer intervals of time, both for the selection of B-galaxies and for the assignation of B- and PTA-probabilities.

The searches described in Section 3 can be improved in several ways. One of the main shortcomings is in the method used to construct the adapted catalogue. We are assuming that all galaxies with the same redshift and mass are affected by the same observational limitations. Moreover, we assume that the same limitations hold for galaxies that recently suffered a major merger and galaxies that did not. If we consider that the latter assumption is not too crude, then our adaptation method should be good enough when calculating the efficiency of the simple Bayesian search explained in Section 3.1. This is so because the efficiency of the search depends only on the distribution of galaxies in z and m, which we know for our real catalogue, regardless of the observational processes that caused the distribution to be so. When considering the galaxy clustering as a piece of information for the search, the adaptation method becomes crucial. Nevertheless, the NNDS does not ameliorate the search performed using only z and m as parameters. Therefore, even if the clustering information is affected by our adaptation method, the inclusion of the NNDS in a search on a properly adapted catalogue is not expected to be efficient. There may be yet other ways to improve the searches by including the clustering; as shown in Sections 3.3

[14] This is the reason why the PTA-probabilities calculated in Section 3.2 are so small.

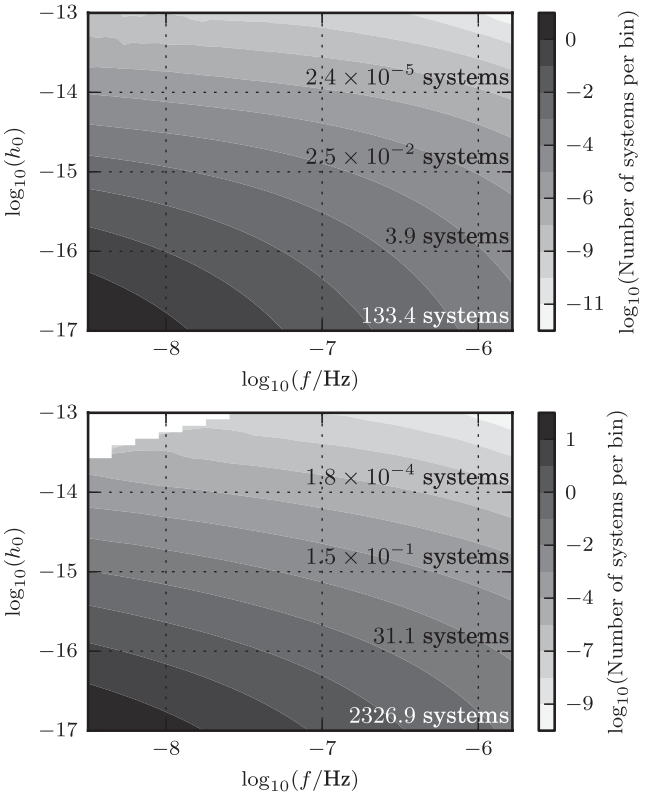

Figure 26. The upper plot is analogous to the upper plot in Fig. 24; now, all systems in the real catalogue (with $z < 0.7$) are considered. The lower plot is calculated considering all B-galaxies from the entire fake catalogue; on top are written the average numbers of PTA-galaxies (for different strain amplitude thresholds) that could be observed in a very complete, all-sky galaxy catalogue (with $z < 0.7$). The upper-left corner is empty simply because, among the 100 realizations, there were no B-galaxies observable at that region of frequencies producing such a large strain.

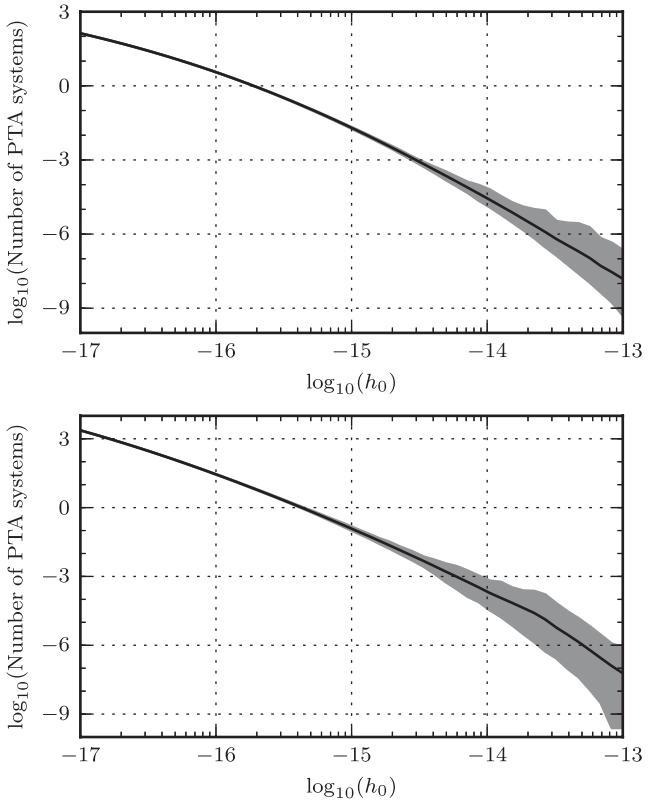

Figure 27. The upper plot is analogous to the upper plot in Fig. 25, but now all systems in the real catalogue are considered (with $z < 0.7$). The lower plot gives the average number of PTA-galaxies that would be contained in an very complete, all-sky galaxy catalogue (the fake catalogue) as a function of their maximum strain amplitude.

and 4.4 there are indeed some features contained in the number of neighbours that distinguish B- from N-galaxies.

Another drawback of this work is the incompleteness of our real catalogue. The spectroscopic SDSS catalogue covers a small fraction of the sky. It would be interesting to apply the algorithms of this paper to a full sky survey. With such a catalogue we could do a more complete study of the distribution of PTA-galaxy candidates. Data from Pan-STARRS (Schlafly et al. 2012; Tonry et al. 2012) will soon be publicly released. Their coverage is of roughly three quarters of the sky, although the redshifts will be photometric (instead of spectroscopic), inferred from the four different wavebands measured, which would be a source of uncertainty in the calculations of redshifts and masses. Additionally, the incompleteness of our real catalogue at very low redshifts ($z_{\min} < 0.01$) could be easily solved by combining this with other complete catalogues of nearby galaxies. Nonetheless, at such low redshifts we do not expect to find more than $\sim$1 PTA-galaxy candidates.

The cosmological parameters assumed by the MS are based on *WMAP* 1 data, which significantly differ from more modern measurements [Komatsu et al. 2011; Ade et al. (Planck Collaboration) 2013]. According to Guo et al. (2013), updating the simulation to the cosmological parameters of *WMAP* 7 does not affect galaxy clustering significantly, since the changes in the values of Ω_{m} and σ_8 (the amplitude of mass fluctuations at $8h_0^{-1}$ Mpc) effectively compensate each other, at least below $z \lesssim 3$. Overall, they report a small difference in the outcomes of the simulation for the two sets of cosmological parameters. Nevertheless, an update of this work to the most recent models would be a sensible follow up. Also, to avoid our method to be too dependent on the particular realization of the universe provided by the MS, we could redo our calculations using other simulations, like DEUS[15] (Alimi et al. 2012).

In different sections of the paper, we have mentioned that the adapted catalogue contains fewer high-mass systems than the real catalogue. This difference in masses is the reason why the number of B- and PTA-galaxy candidates found in the real and adapted catalogues disagree (by less than a factor of 2). For the calculations of stellar masses of real galaxies, it is necessary to calculate the distance to the galaxy, which depends on the cosmological parameters; therefore, the reason for the discrepancy at the high-mass end could be related to the different parameters assumed by the two catalogues. In Fig. 28 one can clearly see that discrepancy in masses. Again, using a simulated universe updated to the most recent cosmological parameters could be the solution for this issue.

In Fig. 29 we show the distributions of B-galaxy and PTA-galaxy probabilities (upper and lower plots, respectively) from the adapted and real catalogues. Each point in a curve tells the number of galaxies (on the vertical axis) that have probabilities larger than a certain value (on the horizontal axis). The vertical lines mark the probability thresholds p_T^{B} and p_T^{PTA} chosen to select B- and PTA-galaxy

[15] http://www.deus-consortium.org/

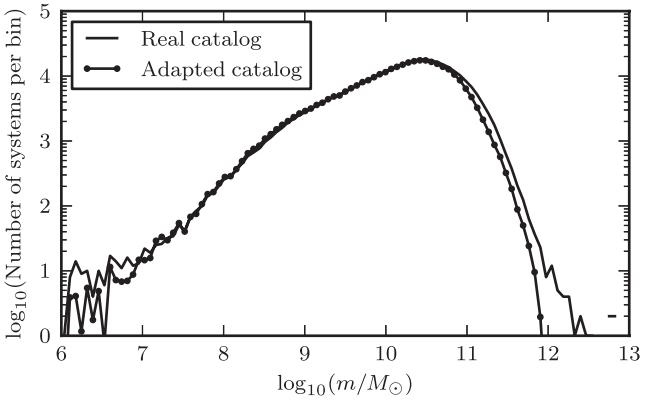

Figure 28. Mass histograms of systems in the real and adapted catalogues (taking a fraction of the sky in the adapted catalogue that is equal to the area covered by the real one). Here we see that the MS has an underdensity of high-mass galaxies with respect to the SDSS.

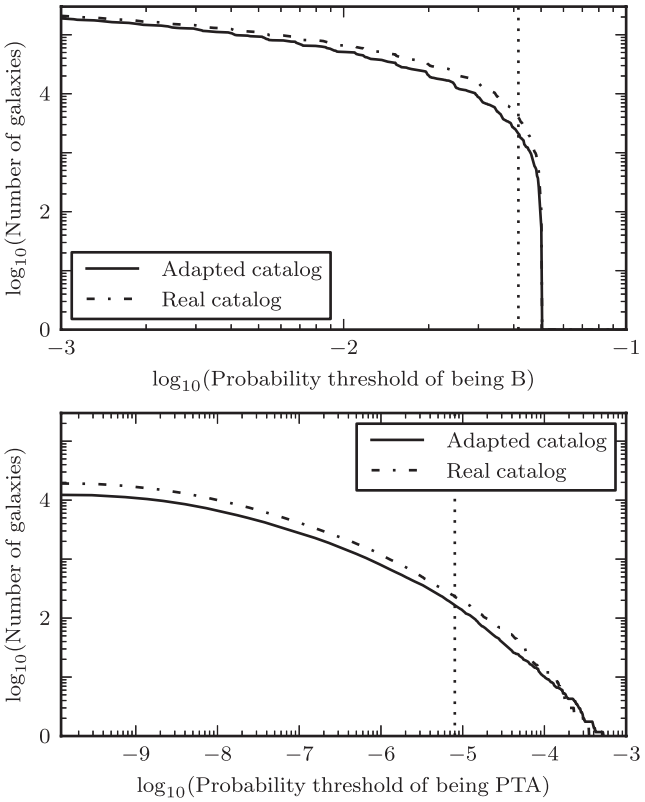

Figure 29. Distribution of B-galaxy probabilities (upper plot) and PTA-galaxy probabilities (below) for the real (dot–dashed lines) and adapted (solid lines) catalogues with $z < 0.1$. The vertical axis gives the number of galaxies that have probabilities larger than a certain threshold, given on the horizontal axis. The numbers for the adapted catalogue have been multiplied by the fraction of the sky covered by the real catalogue (19.5 per cent), so that solid and dot–dashed lines are comparable. The probability thresholds $p_T^{\rm B}$ and $p_T^{\rm PTA}$ chosen to select B-galaxy candidates and PTA-galaxy candidates are marked with vertical dotted lines. Systems on the right of the vertical lines are B-galaxy (PTA-galaxy) candidates, in the upper (lower) plot.

candidates. This plot demonstrates that probabilities are assigned in the real catalogue in a similar way than in the adapted catalogue, except for a factor of $\lesssim 2$ due to the mass discrepancy that we just mentioned.

6 CONCLUSIONS

MBHBs are expected to form in the centre of massive galaxies following merger events, and the detection of their emitted GWs is the main goal of ongoing PTAs. Whether or not MBHBs will be observed as an unresolvable background or as the sum of only a few bright signals is so far unknown, and depends on the spatial distribution of very massive systems in the low-redshift Universe. As such, an effective method for predicting the properties of the PTA signal is to study the distribution of putative MBHB hosts in large galaxy surveys.

In this paper we have first investigated possible criteria to assign to each system a probability of hosting a MBHB (technically, of being a B-galaxy, i.e. a galaxy that suffered a major merger less than ∼300 Myr before emitting the radiation we observe now). We have used a fake (simulated) galaxy catalogue (result of the MS with the semi-analytical galaxy formation models of Guo et al. 2011) and used the peculiar two-dimensional mass–redshift distribution of merging galaxies as a selection criterion. The fake catalogue was then adapted to the observational constraints of the real catalogue, the SDSS spectroscopic catalogue, and the same search for B-galaxies was performed. We caution that this method of adaptation may not be optimal (as commented on in Sections 2.3 and 5).

For each galaxy, we also calculated the probability of being a PTA-galaxy, i.e. a B-galaxy that contains a MBHB emitting GWs that produce a strain amplitude $h_0 \geq h_0^{\rm thres} = 10^{-15}$ (in some frequency interval within the PTA frequency band). To do this, we needed to populate galaxies with MBHs, which requires a knowledge of their bulge mass. To infer the latter, we have constructed a simple model based on the morphological classification of the Galaxy Zoo (explained in Section 4.6).

Our fiducial search is based only on the mass–redshift distribution of galaxies, and extends in the redshift interval $0.01 < z < 0.1$. The search has been extended up to $z = 0.7$, even though the severe incompleteness of the SDSS spectroscopic catalogue at those redshifts and our simple fake catalogue adaptation technique make the results of this search less robust. We also included clustering information in our search (using a MLA), which did not improve the efficiency of the searches for B- or PTA- galaxies.

The main results are summarized below.

(i) In the fake universe, B-galaxies show a distinct distribution in redshift and mass (as shown in Fig. 2): they tend to have larger masses than average (N-) galaxies, which is a reasonable consequence of the conditions in which major mergers take place.

(ii) By using only this information we were able to construct a list of 3870 candidates for B-galaxies with $z < 0.1$ in the SDSS footprint, of which ∼196 are expected to be actual B-galaxies (a skymap with these real candidates is shown in Fig. 20). All of these have stellar masses larger than 10^{11} M$_\odot$.

(iii) Applying our PTA-galaxy search to the real catalogue, we created a list of 232 real PTA-galaxy candidates; this list is expected to include the ∼14 most likely PTA sources in the local Universe that are observed by the SDSS (a skymap with these candidates is given in Fig. 23). PTA-galaxies also have masses $\geq 10^{11}$ M$_\odot$.

(iv) According to Fig. 12, galaxies with such large masses ($\geq 10^{11}$ M$_\odot$) are expected to have more neighbours than average galaxies, which suggests that B- and PTA-galaxies are more likely to be found in large groups or clusters.

(v) The probability of actually observing these PTA-galaxy candidates is small (since they spend a relatively short interval of time producing a strain amplitude larger than $h_0^{\rm thres}$), ranging from 10^{-3} to 10^{-5}; nevertheless, if the PTA manages to detect single sources from the part of the sky covered by the SDSS with $z < 0.1$, those sources are expected to be among the list of candidates shown in Fig. 23.

(vi) Sensitivities to strain amplitudes smaller than $\sim 10^{-16}$ are required to have a sizable number of detectable sources. This result supports the idea that the first PTA detection will most likely involve a low frequency stochastic background signal (Siemens et al. 2013), as opposed to a loud individual source.

(vii) Among the $\approx 3.4 \times 10^5$ galaxies in the real catalogue with $z < 0.1$ we are able to select $\approx 1.7 \times 10^4$ candidates[16] that should include *all* PTA-galaxies ($\approx 7.4 \times 10^2$), if the adapted catalogue resembles well the real one. In other words, we are able to correctly classify $\approx$95 per cent of the whole galaxy population as non-PTA-galaxies, and 100 per cent of the PTA-galaxies are counted among the candidates. Despite this great feature of the search, the number of systems misclassified as PTA-galaxies is still large (we have $\sim$20 misclassifications per true PTA-galaxy). This is an important caveat that can be overcome (and we plan to do it in a future work) by combining our search criteria and methodology with others from the literature.

(viii) The search seems to keep (and even ameliorate) its efficiency when applied to redshifts larger than $\sim$0.1, as shown in Fig. 19.

As shown in Fig. 25, the expected number of observable PTA-galaxies in the local Universe contained in our real catalogue is, for a threshold of $h_0^{\rm thres} = 10^{-15}$, smaller than 0.01. This number does not contradict the expected number of PTA-galaxy candidates contained in the skymap of Fig. 23, which is $\sim$14. Among the systems in the skymap, we do expect $\sim$14 of them to be producing a strain amplitude $\geq h_0^{\rm thres}$, but only during a relatively short interval of time; the sum of the PTA-galaxy probabilities of the candidates in the skymap is therefore less than 0.01. More encouraging numbers can be achieved by setting a smaller strain amplitude threshold (to which future PTA campaigns will be sensitive), or by considering a more complete (or deeper) set of galaxies. The upper plot in Fig. 27 is analogous to the upper plot in Fig. 25, but for redshifts up to 0.7. The lower plot in this figure gives insight on the average number of systems that could be observed simultaneously by the PTA and an ideal telescope, able to produce a complete all-sky galaxy catalogue up to 0.7.

Our work has important practical implications for MBHB and PTA-source searches. If our understanding of the galaxy formation process is correct, there *must* be hundreds of binaries at $z < 0.1$, and our method provides a useful way to narrow down the number of selected targets for deep imaging and spectroscopy to unveil MBHBs in the local Universe. Note, moreover, that these are massive, very low redshift systems, where the search for kinematic signatures of massive binaries at several parsec separations might be already possible with the *Hubble Space Telescope*,[17] and will definitely be within the capabilities of the *European Extremely Large Telescope* (Gilmozzi & Spyromilio 2008). Several tens of such binaries are PTA sources, but unfortunately only a few of them will produce a strain amplitude $\geq 10^{-16}$ which might be detectable with the SKA (Lazio 2009). However, even if very small, there is a chance that a nearby galaxy hosts a loud source of GWs detectable in the PTA band, and our method proved effective in selecting the most likely candidates. Our list of PTA-galaxy candidates can be used to perform targeted searches, also in combination with other searching criteria (Ellison et al. 2013), for example, by looking for signs of recent galaxy interaction in the SDSS images. Finally, being able to assign a probability to each galaxy in the Universe is a powerful tool for constructing PTA signals with the 'right' spatial properties.

The SDSS spectroscopic catalogue covers $\sim$20 per cent of the sky, but surveys like Pan-STARRS and LSST [Ivezic et al. (for the LSST Collaboration) 2011] will cover almost the entire sky. Our technique applied to all-sky, deep galaxy catalogues will allow a complete study of the expected properties (in terms of number and location of putative resolvable sources and level of anisotropy) of the low frequency GW signal in the PTA band. This has a double value for the PTA community: on the one hand, it will provide useful guidance to the development of data analysis algorithms to search for GWs in PTA data; on the other hand, in the presence of a detection, it will provide a useful tool to interpret the results from an astrophysical standpoint.

[16] This number corresponds to the rightmost point in the black curve of Fig. 16, multiplied by the factor 1.41 (explained in Section 4.6) that accounts for the overabundance of SDSS massive galaxies with respect to the MS.
[17] www.hubblesite.org

ACKNOWLEDGEMENTS

We thank Roberto Decarli, Iraklis Konstantopoulos and Jarle Brinchmann for the comments and suggestions; we also thank the latter for his guidance on the use of the real catalogue. PAR wants to thank Bruce Allen, Colin Clark, Rutger van Haasteren, David Keitel, Drew Keppel, Reinhard Prix, Francesco Salemi, Miroslav Shaltev, and especially Tito Dal Canton for the fruitful discussions and help while the preparation of this work. PAR also thanks Guinevere Kauffmann for the valuable advices and the bibliography suggested.

The MS data bases used in this paper and the web application providing online access to them were constructed as part of the activities of the German Astrophysical Virtual Observatory (GAVO).

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the US Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS website is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

This work was supported by the IMPRS on Gravitational Wave Astronomy.

REFERENCES

Abadie J. et al. (LIGO Scientific Collaboration), 2011, Nat. Phys. Lett., 7, 962
Abazajian K. et al., 2009, ApJS, 182, 543
Abbott B. P. et al. (LIGO Scientific Collaboration), 2009, Rep. Prog. Phys., 72, 076901
Accadia T. et al. (Virgo Collaboration), 2012, J. Instrum., 7, 3012
Ade P. A. R. et al. (Planck Collaboration), 2013, preprint (arXiv:1303.5076)
Aihara H. et al., 2011a, ApJS, 193, 29
Aihara H. et al., 2011b, ApJS, 195, 26
Ajith P. et al., 2011, Phys. Rev. Lett., 106, 241101
Alimi J.-M. et al., 2012, preprint (arXiv:1206.2838)
Amaro Seoane P. et al. (eLISA Consortium), 2013, preprint (arXiv:1305.5720)
Andersson N., Ferrari V., Jones D. I., Kokkotas K. D., Krishnan B., Read J., Rezzolla L., Zink B., 2011, Gen. Relativ. Grav., 43, 409
Babak S., Sesana A., 2012, Phys. Rev. D, 85, 044034
Babak S., Gair J. R., Petiteau A., Sesana A., 2011, Class. Quantum Grav., 28, 114001
Begelman M., Blandford R., Rees M., 1980, Nature, 287, 307
Binney J., Tremaine S., eds, 2008, Galactic Dynamics, 2nd edn. Princeton Univ. Press, Princeton, NJ
Blanton M. R., Lin H., Lupton R. H., Miller Maley F., Young N., Zehavi I., Loveday J., 2003, AJ, 125, 2276
Blanton M. R. et al., 2005, AJ, 129, 2562
Bonoli S., Marulli F., Springel V., White S. D., Branchini E., Moscardini L., 2009, MNRAS, 396, 423
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Burgay M. et al., 2003, Nature, 426, 531
Burke-Spolaor S., 2013, Class. Quantum Grav., 30, 224013
Chandrasekhar S., 1943, ApJ, 97, 255
Colless M. et al., 2001, MNRAS, 328, 1039
Davis M., Peebles P. J. E., 1983, ApJ, 267, 465
Dotti M., Colpi M., Haardt F., Mayer L., 2007, MNRAS, 379, 956
Drake F. L., van Rossum G., 2011, The Python Language Reference Manual: for Python version 3.2. Network Theory Ltd.
Ellis J., Siemens X., Creighton J., 2012, ApJ, 756, 175
Ellison S. L., Mendel J. T., Patton D. R., Scudder J. M., 2013, MNRAS, 435, 3627
Eracleous M., Boroson T. A., Halpern J. P., Liu J., 2012, ApJS, 201, 23
Escala A., Larson R. B., Coppi P. S., Mardones D., 2005, ApJ, 630, 152
Ferdman R. D. et al., 2010, Class. Quantum Grav., 27, 084014
Ferrarese L., Merritt D., 2000, ApJ, 539, L9
Foster R., Backer D., 1990, ApJ, 361, 300
Gebhardt K. et al., 2000, ApJ, 539, L13
Ghez A. et al., 2008, ApJ, 689, 1044
Gillessen S., Eisenhauer F., Trippe S., Alexander T., Genzel R., Martins F., Ott T., 2009, ApJ, 692, 1075
Gilmozzi R., Spyromilio J., 2008, Proc. SPIE, 7012, 19
Graham A. W., Onken C. A., Athanassoula E., Combes F., 2011, MNRAS, 412, 2211
Guo Q. et al., 2011, MNRAS, 413, 101
Guo H., Zehavi I., Zheng Z., 2012, ApJ, 756, 127
Guo Q., White S., Angulo R. E., Henriques B., Lemson G., Boylan-Kolchin M., Thomas P., Short C., 2013, MNRAS, 428, 1351
Haering N., Rix H.-W., 2004, ApJ, 604, L89
Hamilton A., 1993, ApJ, 417, 19
Henriques B., White S., Lemson G., Thomas P., Guo Q., Marleau G.-D., Overzier R., 2012, MNRAS, 421, 2904
Hobbs G. et al., 2010, Class. Quantum Grav., 27, 084013
Ivezic Z. et al. (for the LSST Collaboration), 2011, preprint (arXiv:0805.2366)
Jaffe A. H., Backer D. C., 2003, ApJ, 583, 16
Jenet F. A., Hobbs G. B., Lee K. J., Manchester R. N., 2005, ApJ, 625, L123
Jenet F. et al., 2009, preprint (arXiv:0909.1058v1)
Ju W., Greene J. E., Rafikov R. R., Bickerton S. J., Badenes C., 2013, ApJ, 777, 44
Kauffmann G. et al., 2003a, MNRAS, 341, 33
Kauffmann G. et al., 2003b, MNRAS, 346, 1055
Khan F. M., Just A., Merritt D., 2011, ApJ, 732, 89
Khan F. M, Berentzen I., Berczik P., Just A., Mayer L., Nitadori K., Callegari S., 2012, ApJ, 756, 30
Kitzbichler M. G., White S. D. M., 2008, MNRAS, 391, 1489
Kocsis B., Sesana A., 2011, MNRAS, 411, 1467
Komatsu E. et al., 2011, ApJS, 192, 18
Kormendy J., Richstone D., 1995, ARA&A, 33, 581
Kramer M., 2012, in van Leeuwen J., ed., Proc. IAU Symp., 291, Probing gravitation with pulsars. Cambridge Univ. Press, Cambridge, p. 19
Landy S., Szalay A., 1993, ApJ, 412, 64
Lazio J., 2009, preprint (arXiv:0910.0632)
Lemson G., the Virgo Consortium, 2006, preprint (arXiv:astro-ph/0608019)
Li C., Kauffmann G., Jing Y. P., White S. D. M., Börner G., Cheng F. Z., 2006a, MNRAS, 368, 21
Li C., Kauffmann G., Wang L., White S. D. M., Heckman T. M., Jing Y. P., 2006b, MNRAS, 373, 457
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Lintott C. et al., 2011, MNRAS, 410, 166
Lorimer D., 2008, Living Rev. Relativ., 11, 8
Lynden-Bell D., 1969, Nature, 223, 690
Lyne A. et al., 2004, Science, 303, 1153
McConnell N. J., Ma C.-P., 2013, ApJ, 764, 184
Maggiore M., 2000, preprint (arXiv:gr-qc/0008027v1)
Maggiore M., 2008, Gravitational Waves Volume 1: Theory and Experiments. Oxford Univ. Press, Oxford
Magorrian J. et al., 1998, AJ, 115, 2285
Manchester R. et al., 2013, Publ. Astron. Soc. Aust., 30, 17
Marconi A., Hunt L. K., 2003, ApJ, 589, L21
Marulli F., Bonoli S., Branchini E., Moscardini L., Springel V., 2008, MNRAS, 385, 1846
Mingarelli C. M. F., Grover K., Sidery T., Smith R. J. E., Vecchio A., 2012, Phys. Rev. Lett., 109, 081104
Mingarelli C. M. F., Sidery T., Mandel I., Vecchio A., 2013, Phys. Rev. D, 88, 062005
Misner C., Thorne K., Wheeler J., 1973, Gravitation. Freeman & Co., San Francisco
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peebles P., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Peters P. C., 1964, Phys. Rev., 136, 4B
Petiteau A., Babak S., Sesana A., de Araújo M., 2013, Phys. Rev. D, 87, 064036
Pitkin M., Reid S., Rowan S., Hough J., 2011, Living Rev. Relativ., 14, 5
Preto M., Berentzen I., Berczik P., Spurzem R., 2011, ApJ, 732, L26
Prix R., 2009, in Becker W., ed., Astrophysics and Space Science Library, Vol. 357, Neutron Stars and Pulsars. Springer-Verlag, Berlin, p. 651
Punturo M. et al., 2010, Class. Quantum Grav., 27, 194002
Quinlan G. D., 1996, New Astron., 1, 35
Rajagopal M., Romani R. W., 1995, ApJ, 446, 543
Ravi V., Wyithe J. S. B., Hobbs G., Shannon R. M., Manchester R. N., Yardley D. R. B., Keith M. J., 2012, ApJ, 761, 84
Regimbau T., Giampanis S., Siemens X., Mandic V., 2012, Phys. Rev. D, 85, 066001
Richstone D. et al., 1998, Nature, 395, A14
Rosado P. A., 2011, Phys. Rev. D, 84, 084004
Rosado P. A., 2012, Phys. Rev. D, 86, 104007
Salim S. et al., 2007, ApJS, 173, 267
Sanidas S. A., Battye R. A., Stappers B. W., 2012, Phys. Rev. D, 85, 122003
Sathyaprakash B., Schutz B. F., 2009, Living Rev. Relativ., 12, 2
Schlafly E. F. et al., 2012, ApJ, 756, 158
Schutz B. F., 1999, Class. Quantum Grav., 16, A131
Sesana A., 2013a, Class. Quantum Grav., 30, 224014
Sesana A., 2013b, MNRAS, 433, L1
Sesana A., Haardt F., Madau P., Volonteri M., 2004, ApJ, 611, 623

Sesana A., Vecchio A., Colacino C. N., 2008, MNRAS, 390, 192
Sesana A., Vecchio A., Volonteri M., 2009, MNRAS, 394, 11
Shaddock D. A., 2008, Class. Quantum Grav., 25, 114012
Shen Y., Liu X., Loeb A., Tremaine S., 2013, ApJ, 775, 49
Siemens X., Ellis J., Jenet F., Romano J. D., 2013, Clas. Quantum Grav., 30, 224015
Somiya K., (for the KAGRA Collaboration) 2012, Class. Quantum Grav., 29, 124007
Spergel D. N. et al., 2003, ApJS, 148, 175
Springel V., White S., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Springel V. et al., 2005, Nature, 435, 629
Stoughton C. et al., 2002, AJ, 123, 485
Taylor S. R., Gair J. R., 2013, Phys. Rev. D, 88, 084001
Tonry J. et al., 2012, ApJ, 750, 99
Tremonti C. A. et al., 2004, ApJ, 613, 898
Tsalmantza P., Decarli R., Dotti M., Hogg D. W., 2011, ApJ, 738, 20
van Haasteren R. et al., 2011, MNRAS, 414, 3117
van Haasteren R., Mingarelli C. M. F., Vecchio A., Lassus A., 2013, preprint (arXiv:1301.6673)
Volonteri M., Haardt F., Madau P., 2003, ApJ, 582, 15
Wang W., White S. D. M., 2012, MNRAS, 424, 2574
Weisberg J. M., Nice D. J., Taylor J. H., 2010, ApJ, 722, 1030
White S., Rees M., 1978, MNRAS, 183, 341
Wyithe J. S. B., Loeb A., 2003, ApJ, 590, 16
York D. G. et al., 2000, AJ, 120, 1579

APPENDIX A: QUERIES FOR THE DATA

The SQL query sent to the MS site to download the fake catalogue (for systems with redshifts $z < 0.11$) is the following:

```
select h.galID,
       h.ra,
       h.dec,
       h.z_geo,
       h.z_app,
       g.stellarmass,
       g.blackholemass,
       g.bulgemass,
       g.type

from Henriques2012a.wmap1.bc03_AllSky_001 h,
    Guo2010a..MR g

where h.galID = g.galaxyid
 and g.stellarmass > 7e-5
 and h.z_geo <= 0.11
```

The stellar mass in the data base is given in units of $10^{10}\,M_{\odot}/h_{100}$, where $h_{100} = 0.73$ (for the set of cosmological parameters assumed by the MS), so the minimum mass imposed in this query is of $\approx 9.6 \times 10^5\,M_{\odot}$. Once the catalogue is downloaded, we select only galaxies with masses $\geq 10^6\,M_{\odot}$ (minimum mass of the real catalogue).

This query is limited to redshifts smaller than 0.11, and outputs $\approx 1.0 \times 10^7$ galaxies. A query with a maximum redshift of 0.7 would produce $\approx 1.8 \times 10^8$ galaxies. As we explain in Section 3.4, we do not need to download all these data to extend the search to all redshifts considered in the real catalogue. Instead, we just need a histogram with the number of galaxies contained in each z–m bin.

The searches described in Sections 3.1 and 3.2 are restricted to $z < 0.1$. However, we download systems with $z < 0.11$ to avoid border effects: when calculating the NNDS of systems close to $z = 0.1$, one needs to consider background galaxies that have larger redshifts; for our choices of shells, all background galaxies are safely contained below $z = 0.11$.

The query used to construct the list of B-galaxies in the fake catalogue is

```
select h.galID,
       h.z_geo,
       h.z_app,
       p1.blackholemass,
       p2.blackholemass

from Henriques2012a.wmap1.bc03_AllSky_001 h,
    Guo2010a..MR p1,
    Guo2010a..MR p2,
    Guo2010a..MR d

where h.galID = d.galaxyid
 and h.z_geo <= 0.7
 and p1.descendantid = d.galaxyid
 and p2.descendantid = d.galaxyid
 and p1.galaxyid < p2.galaxyid
 and p1.snapnum = p2.snapnum
 and p1.stellarmass >= 0.2*d.stellarmass
 and p2.stellarmass >= 0.2*d.stellarmass
 and p1.blackholemass > 1e-6
 and p2.blackholemass > 1e-6
 and d.blackholemass > 1e-6
 and d.stellarmass > 7e-5
 and p1.disruptionon = 0
 and p2.disruptionon = 0
 and p1.snapnum = d.snapnum-1
```

This query outputs the `galID` of galaxies that suffered a major merger between their snapshot and the previous one. Since the `galID` can be repeated (as mentioned in Section 2.2), cosmological and apparent redshifts are also downloaded, to be able to identify systems without any ambiguity. The progenitors' black hole masses are downloaded to construct the black hole mass ratio, necessary for the calculations presented in Section 3.2.

The condition of major merger is that at least two progenitors of a galaxy must have mass $\geq$0.2 times the mass of the galaxy. Furthermore, we discard progenitors that were disrupted before the merger. We also impose that the progenitor must have a black hole mass larger than $\approx 1.4 \times 10^4\,M_{\odot}$. The same condition is imposed to the descendant, although this condition turns out to be unnecessary, since the minimum black hole mass found among B-galaxies is of $\approx 10^{6.2}\,M_{\odot}$.

Several major multimergers are found with this query. These are mergers of three galaxies[18] that have a mass larger than 0.2 times the mass of the descendant. When such a multimerger occurs, the query outputs the descendant galaxy three times, because three possible pairs of progenitors are considered (first and second, first and third, and second and third). If that descendant galaxy appears twice (or three times, or four times, etc.) in the galaxy catalogue (due to a repetition of the cube of the simulation), the query will output 6 (or 9, or 12, etc.) times the same galaxy. One has to properly correct for these repetitions to avoid ambiguities when identifying galaxies.

[18] The unlikely cases of major multimergers involving more than three progenitors were not found with this query.

The query used in the SDSS DR7 server to obtain the morphological parameters introduced in Section 4.6 is the following:

```
select g.objid,
       s.specobjid,
       g.petror50_r,
       g.petror90_r,
       g.petror50_z

from galaxy g,
    specobj s
where g.objid=s.bestobjid
 and s.z<=0.7
```

The two first items, `objid` and `specobjid` were used to identify the galaxies of this query with the ones of our real catalogue.

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