Density-clustering of continuous gravitational wave candidates from large surveys

B. Steltner,[1, 2, a] T. Menne,[1, 2] M. A. Papa,[1, 2, 3] and H.-B. Eggenstein[1, 2]

[1] *Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Callinstrasse 38, 30167 Hannover, Germany*
[2] *Leibniz Universität Hannover, D-30167 Hannover, Germany*
[3] *University of Wisconsin Milwaukee, 3135 N Maryland Ave, Milwaukee, WI 53211, USA*

Searches for continuous gravitational waves target nearly monochromatic gravitational wave emission from e.g. non-axysmmetric fast-spinning neutron stars. Broad surveys often require to explicitly search for a very large number of different waveforms, easily exceeding $\sim 10^{17}$ templates. In such cases, for practical reasons, only the top, say $\sim 10^{10}$, results are saved and followed-up through a hierarchy of stages. Most of these candidates are not completely independent of neighbouring ones, but arise due to some common cause: a fluctuation, a signal or a disturbance. By judiciously clustering together candidates stemming from the same root cause, the subsequent follow-ups become more effective. A number of clustering algorithms have been employed in past searches based on iteratively finding symmetric and compact over-densities around candidates with high detection statistic values. The new clustering method presented in this paper is a significant improvement over previous methods: it is agnostic about the shape of the over-densities, is very efficient and it is effective: at a very high detection efficiency, it has a noise rejection of 99.99% , is capable of clustering two orders of magnitude more candidates than attainable before and, at fixed sensitivity it enables more than a factor of 30 faster follow-ups. We also demonstrate how to optimally choose the clustering parameters.

I. INTRODUCTION

Continuous gravitational waves are long-lasting signals that may come from fast-spinning non-axisymmetric neutron stars, unstable r-modes [1, 2], the fast inspiral of dark-matter objects [3, 4] or emission from clouds of axion-like particles around black holes [5, 6].

Independently of the emission mechanism, the expected signal is a nearly monochromatic wave that due to the relative motion between the source and the detector, is frequency- and amplitude- modulated. The signal shape is described by a frequency and its derivatives, the source position in the sky, the signal amplitude, the polarization angle and the source inclination with respect to the line of sight.

Searches for continuous gravitational wave signals typically use template banks for frequency, frequency derivatives and source position only, because the remaining parameters can be analytically maximized over, and do not need to be searched for explicitly. The parameter space of broad surveys grows quickly with the observation span, and for observations lasting months $\sim 10^{17}$ template waveforms need to be considered.

For template banks that are this big, typically only the top results are saved – say the $\sim 10^{10}$ results with the highest detection statistic values. These are then considered for further follow-up investigations [7–9]. Even though at this stage most of the results are not statistically significant, because they will be subject to more inspections, they are referred to as "candidates".

To reduce the loss in signal-to-noise ratio due to template-signal waveform mismatch, our templates have a high overlap and thus are not independent. Therefore a loud signal or disturbance triggers multiple templates, generating a high number of candidates close to each other in parameter space. Following-up each one independently would result in a waste of resources.

Clustering is an important step in the post-processing of the results that organizes and reduces the $\sim 10^{10}$ candidates to a more useful set of $\approx$ independent $\sim 10^{6}$ candidates.

While the clustering algorithm details vary, the core is to find candidates likely due to the same root cause, bundle (*cluster*) them and consider them as a single entity in follow-up studies.

Each cluster is represented by the parameters of the so-called *seed* candidate and by a *containment region*. The latter measures how far from the seed associated with a signal, the true signal parameters are. In follow-up studies the entire containment region around each seed is surveyed. The containment region is the same for all seeds and it is determined statistically, such that it holds for a very large fraction ($> 99\%$) of signals, across the parameter space.

It has also been observed that a threshold on the minimum number of candidates in a cluster is effective at discarding noise-clusters. With a fixed computing budget for follow-ups, fewer candidates means that freed-up computational capacity can be used on additional, lower significance candidates which translates in deeper and more sensitive searches.

Clustering is a crucial step in the analysis of the results of broad continuous wave surveys with very large template banks, such as those carried out on the volunteer computing project Einstein@Home[1] [10–12]. For

[a] benjamin.steltner@aei.mpg.de

[1] www.einsteinathome.org/

this reason clustering procedures have been in use for a long time: One of the first non-trivial clustering procedures is box-clustering [13, 14], which dates back to nearly a decade ago. More recently a more flexible adaptive clustering technique has been used [15] which however does not converge fast enough when used on many data points. This is a significant drawback, as we want to set lower thresholds, which means considering more candidates in the follow-ups. Attempts to use machine-learning for clustering have been successful for directed searches, but not for all-sky searches [16, 17].

We present here the new *Density Clustering* algorithm, able to process orders of magnitude more candidates than previous clustering strategies at comparable, if not lower, computing cost. We show how to choose the clustering parameters, and demonstrate its performance on real data. We concentrate on clustering results from very large template banks – with over 10^{16} points – and hence refer to the Einstein@Home results. This method can also be employed in less challenging environments.

The paper is organized as follows: In Section II we describe the input data; in Section III the method itself; in Section IV the choice of the clustering parameters; in Section V the implementation; in Section VI the method is compared with Adaptive Clustering under realistic conditions, i.e. by applying it to the data of the Stage 0 results of the Einstein@Home all-sky search for continuous gravitational waves in Advanced LIGO data of the second observation run (O2) [9, 18].

II. INPUT DATA TO CLUSTERING

Clustering works on a set of candidates, i.e. selected results from a search. A candidate is described by the values of the template that produced the detection statistic result, and the detection statistic result. For an all-sky search including up to second-order spin-down parameters, a generic candidate i is of the form

$$\left(f_i, \dot{f}_i, \ddot{f}_i, \alpha_i, \delta_i, \chi_i = \hat{\beta}_{\mathrm{S/GLtL}i}\right), \tag{1}$$

where f indicates the signal-template frequency, α, δ the source sky position and χ the detection statistic. Consistently with the Einstein@Home searches we have indicated $\hat{\beta}_{\mathrm{S/GLtL}}$ the line-and-transient-line robust statistic [19] as the detection statistic of choice. We will illustrate clustering for these 5 dimensions; fewer or more dimensions are treated analogously.

Since continuous waves are modulated by the Earth's rotation and orbit around the Sun, the sky grids are set-up in sky coordinates projected on the ecliptic plane,$x_{\mathrm{ecl}}, y_{\mathrm{ecl}}$. Therefore for clustering we convert for the candidates $(\alpha_i, \delta_i) \to (x_{\mathrm{ecl}i}, y_{\mathrm{ecl}i})$ – see Eq.s (14-15) in [15] for the conversion between $(\alpha, \delta) \to (x_{\mathrm{ecl}}, y_{\mathrm{ecl}})$.

The sky grids are approximately uniform hexagonal grids on the ecliptic plane and are defined by the hexagon edge length d:

$$d(m_{\mathrm{sky}}) = \frac{1}{f}\frac{\sqrt{m_{\mathrm{sky}}}}{\pi\tau_E}, \tag{2}$$

with $\tau_E \simeq 0.021$ s being half of the light travel-time across the Earth and m_{sky} a constant which controls the resolution of the sky grid [9]. From Eq. 2 it is clear that the sky-grid density increases with frequency f.

III. DENSITY CLUSTERING

We bin the parameter space in equally-spaced cells of size

$$\delta b = (\delta f, \delta\dot{f}, \delta\ddot{f}, \delta x_{\mathrm{ecl}}, \delta y_{\mathrm{ecl}}) \tag{3}$$

in each dimension. The $\delta f, \delta\dot{f}, \delta\ddot{f}$ are each an integer multiple of the search grid spacing. The sky grid has a hexagonal tilling, so the square tiling of the bins above does not match it. The bins are usually chosen to be large enough that this does not matter and the square covering greatly simplifies the binning and the identification of neighbouring bins. The bin size is always a multiple of the hexagon side, so the bins shrink with increasing frequency as the sky-grid pixels, keeping the average number of candidates per bin the same.

We only consider candidates with detection statistic values above a threshold Γ_L. In each bin j we count the number of candidates $N_{occ,j}$ with parameters in that bin. Bins with $N_{occ,j} \leq N_{occ,\mathrm{min}}$ are discarded. $N_{occ,\mathrm{min}}$ is one of the clustering parameters and its optimal value depends on the search set-up and on the bin size.

Among the surviving bins, we cluster together nearby ones, to create a cluster. The basic notion of vicinity is controlled by two parameters: N^j and N_c. A bin b_a is a neighbour of bin b_c if the distances k^j in integer bin spacings

$$b_a - b_c = \left(k^1\delta f, k^2\delta\dot{f}, k^3\delta\ddot{f}, k^4\delta x_{\mathrm{ecl}}, k^5\delta y_{\mathrm{ecl}}\right) \tag{4}$$

satisfy the following conditions:

$$\begin{cases} k^j \leq N^j \quad \text{with } j = 1, .., M \\ \sum_{j=1}^{M} k^j \leq N_c, \end{cases} \tag{5}$$

where M is the number of dimensions. The first condition sets the maximum distance in every dimension, whereas the second condition sets an overall maximum distance. With $M = 3$, $N_c = 1$ means that the two nearby bins have to share a face, $N_c = 2$ that they have to share an edge and $N_c = 3$ that they have to share a vertex. Default values are $N^j = 1$, equal for all j, and $N_c = M$.

Among the clusters from the previous step, we remove the ones with too few bins: $N_{bins} \leq N_{\mathrm{bins,min}}$.

For each remaining cluster a representative candidate becomes the seed. The seed is by default the candidate

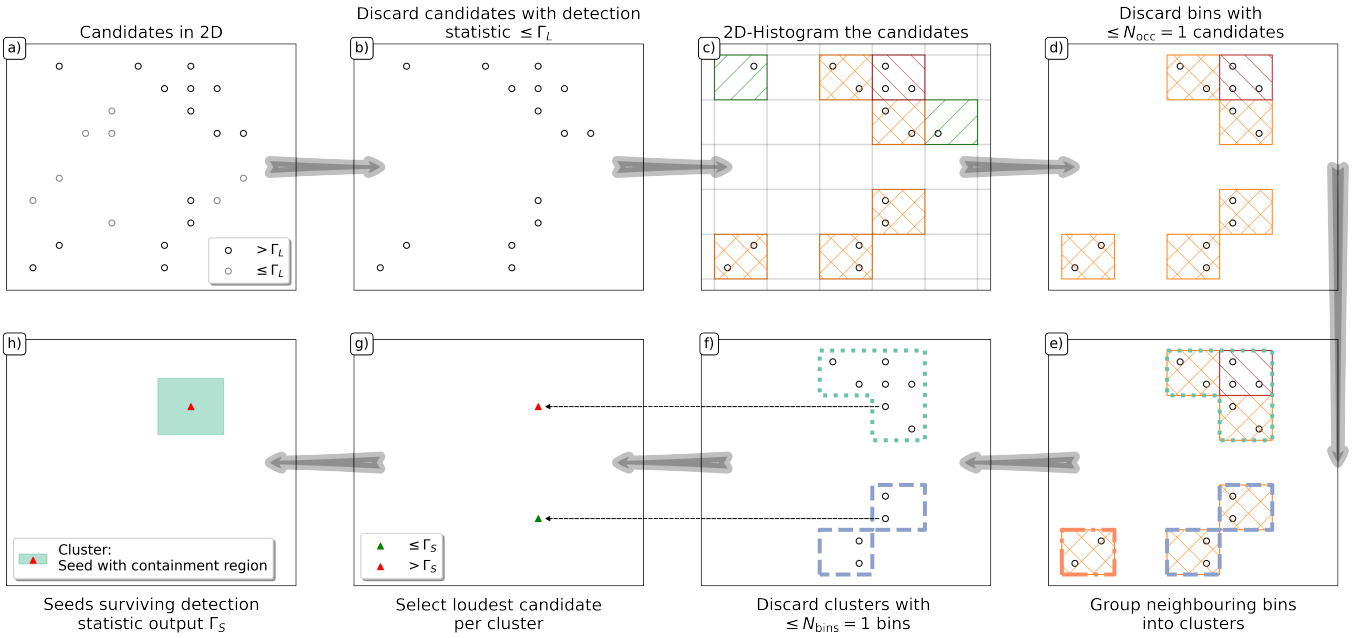

FIG. 1. Schematic illustration of the main steps of Density Clustering

with the highest detection statistic value (the loudest) of all candidates in the cluster. In noisier data it may make sense to look at the loudest candidate in the bin with the most candidates (densest bin) or the loudest candidate in the bin with the highest sum over all detection statistic values of the candidates within that bin (loudest bin).

Finally all clusters with a seed with detection statistic value smaller than Γ_S are discarded.

An additional parameter can be used to mitigate binning effects: an overdensity of candidates may not be perfectly contained within one bin, but may extend across bin boundaries. For faint signals with just enough candidates to surpass the occupancy threshold $N_{occ,\min}$, this effect can make the difference between recovering a signal or not. Boundary effects can be partly mitigated by smoothing over bins, e.g. adding bin counts over neighbouring bins or adding bin counts weighted with a Gaussian kernel. The overall impact of using smoothing procedures should be evaluated within the general framework of choosing the optimal clustering parameters, as described in the next Section, but we will not explicitly consider it here.

IV. CHOOSING THE PARAMETERS OF THE CLUSTERING PROCEDURE

A number of parameters define the Density Clustering algorithm, and they are summarized in Table I. We choose the parameters such that at fixed computational cost for the follow-up of the resulting seeds, the sensitivity of the clustering procedure is maximized.

The sensitivity of the clustering procedure is measured by the gravitational wave signal amplitude $h_0^{90\%}$ at which the detection efficiency ϵ of the clustering procedure is 90%, for signals with parameters in the search range. $h_0^{90\%}$ scales with the amplitude spectral density of the noise $\sqrt{S_h(f)}$, so we will consider the quantity $\mathcal{D}^{90\%} = \sqrt{S_h(f)}/h_0$, instead, that does not depend on frequency. $\mathcal{D}$ is also known as the sensitivity depth [14].

Since there is no way to predict the detection efficiency of the clustering procedure, we measure it with a Monte Carlo, where we add fake signals to the noise with amplitudes corresponding to a given value of $\mathcal{D}$. For each signal we perform a search like the one whose results we wish to cluster, cluster the results and produce seeds. If one of the seeds comes from the added signal, we consider the signal detected. The fraction of detected signals to total signals gives the detection efficiency at that sensitivity depth: $\epsilon(\mathcal{D})$. $\mathcal{D}^{90\%}$ is then

$$\epsilon(\mathcal{D}^{90\%}) = 90\%, \tag{6}$$

and $\mathcal{D}^{90\%}$ measures the sensitivity of the clustering procedure: the higher is $\mathcal{D}^{90\%}$, the higher is the sensitivity.

We consider different clustering set-ups corresponding to different choices of clustering parameters. For each we estimate

- $\mathcal{D}^{90\%}$
- the containment region (see Section I).
- the false alarm rate. This is done by running the clustering on a sub-set on the search results, at different frequencies. Since we operate in the regime of very rare signals, we take this as a measure of the false alarm.

Parameter	Function
Input-Threshold $\Gamma_{\rm L}$	Discards candidates with detection statistic $\leq \Gamma_{\rm L}$. Filters input-candidates
Bin sizes δb	Binning
Smoothing	Smooth histogram or not
Occupancy-threshold $N_{occ,\rm min}$	Discard bins with $N_{occ} \leq N_{occ,\rm min}$ candidates
Neighbour-criterion, N^J and N_c	Defines what a neighbour is
Cluster-size-threshold N_{bins}	Discard clusters with $N_{bins} \leq N_{\rm bins,min}$ bins
Seed criterion	Loudest candidate in cluster, loudest in most-populated bin or in bin with highest average detection statistic
Output-Threshold $\Gamma_{\rm S}$	Discards cluster whose seed has detection statistic $\leq \Gamma_{\rm S}$. Reduces false alarms

TABLE I. Parameters of Density Clustering in the order that they are employed

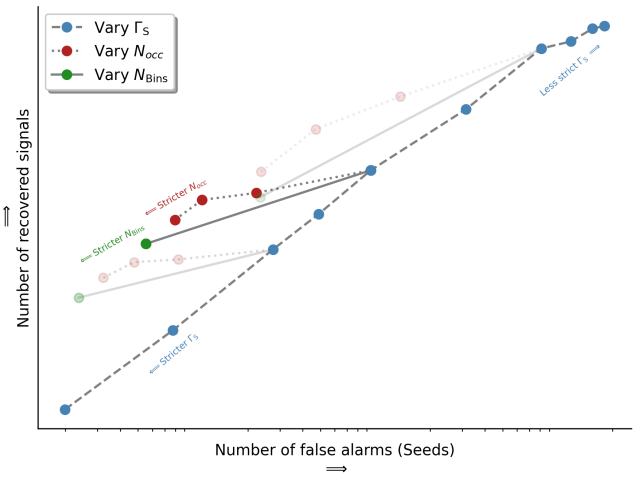

FIG. 2. We show how the number of false alarms (in log-scale) and the number of detected signals changes while varying just a single clustering parameter. Along the dashed line the output threshold $\Gamma_{\rm S} \in [-4, 10]$ varies. The stricter (higher) $\Gamma_{\rm S}$ is, the fewer signals are recovered and the fewer are the false alarms. Similarly stricter $N_{occ,\rm min}$ (dotted line) or $N_{\rm bins,min}$ (solid line) reduce the number of false alarms. The number of recovered signals is also reduced, but less than by using a higher $\Gamma_{\rm S}$. It is often beneficial to allow lower significance candidates (lower $\Gamma_{\rm S}$) and veto more aggressively based on $N_{occ,\rm min}$ and $N_{\rm bins,min}$, instead.

Figure 2 shows how the detection efficiency and the false alarm rate change as different clustering parameters vary.

The number of seeds and the containment region are used to estimate the computing cost of the follow-up. In principle one could optimize the follow-up search setup for each clustering setup. This would however be extremely expensive and experience has shown that a set-up choice guided by the sensitivity gain with respect to the previous stage, at accessible computing cost, lands a choice not significantly far from optimum. So we assume here that the follow-up set-up is fixed.

We can now identify the clustering set-up that yields the highest $\mathcal{D}^{90\%}$, within the computing budget. This is illustrated in Figure 3 for the results of the Stage-0 Einstein@Home search [9].

V. IMPLEMENTATION

In the previous Section we have described how the optimal combination of clustering parameters is identified. As we have seen, this requires a Monte Carlo in order to measure the false alarm and 90% detection-efficiency signal-amplitude $\mathcal{D}^{90\%}$, for every clustering set-up.

For each setup we cluster $\gtrsim$ 2000 result-files corresponding to data with different fake signals – this is to determine $\mathcal{D}^{90\%}$. We cluster $\gtrsim$ 500 search result-files with no fake signals, in order to estimate the false alarm. These operations can be quite time-consuming, so we describe here how to reduce the computing cost of this step.

Einstein@Home search results typically come in files that cover a 50 mHz range of template frequencies, with size varying between a few MB to few GB, due to the different sky resolutions in the range $50 - 600$ Hz. Each clustering instance uses as input one of these 50 mHz results-files. Since the time to cluster is $\lll$ the time it takes to load such a file, it is faster to load a results-file, keep it in memory, and test different clustering set-ups.

Further savings are obtained by re-using intermediate results:

- we compute a histogram for a choice of $\Gamma_{\rm L}$ and δb, and re-use it to produce the bin-counts for different values of $N_{occ,\rm min}$
- similarly, for a choice of $\Gamma_{\rm L}$, δb and $N_{occ,\rm min}$ from the bin-counts we produce different clusters for different values of $N_{\rm bins,min}$
- for each cluster different seeds are produced, based on different seed-selection criteria, e.g. the loudest cluster candidate, the loudest in the densest bin, the loudest in the bin with the highest average detection statistic (we call this the "loudest bin"). By default this step is not performed and the loudest cluster candidate is directly considered as the ultimate cluster seed.
- finally, each seed, and with it the whole cluster, may be discarded depending on the value of $\Gamma_{\rm S}$

With this scheme, testing a single clustering set-up costs (on average, over many set-ups) just under a second, with

more than half the time spent on the initial histogram and clustering. In order to reduce memory usage, the candidates are internally addressed only by an id. For thresholding on Γ_S and for computing the containment region, the actual seed parameters must be retrieved. This operation accounts for another 20% of the computing time. The remaining time is due to fluctuations in these estimates due to varying number of seeds and the initial overhead.

Given the computing-load profile described above, we parallelise the work among different independent processors, with each processor working only with a single results file and several $(\Gamma_L, \delta b)$-combinations. Say we have 2500 result-files, 1000 $(\Gamma_L, \delta b)$ combinations and 1000 combinations of the remaining parameters, each processors analyses 100 $(\Gamma_L, \delta b)$-combinations, exhausting all 1000 combinations of the remaining parameters. Hence, with 10 processes per result file, 25000 processes are spawned in total.

Using the large-capacity and fast-loading hdf5 and FITS file formats, and a HDD-raid configuration results-file server, testing a single $(\Gamma_L, \delta b)$-combination and all 1000 combinations of the remaining parameters, takes $\approx 0.26\,\mathrm{h}$. Thus one processor exhausting 100 $(\Gamma_L, \delta b)$-combinations takes $\approx$ a day. On the ATLAS cluster[2] using 25000 parallel processes the full testing of 1000×1000 set-ups is carried out in a day.

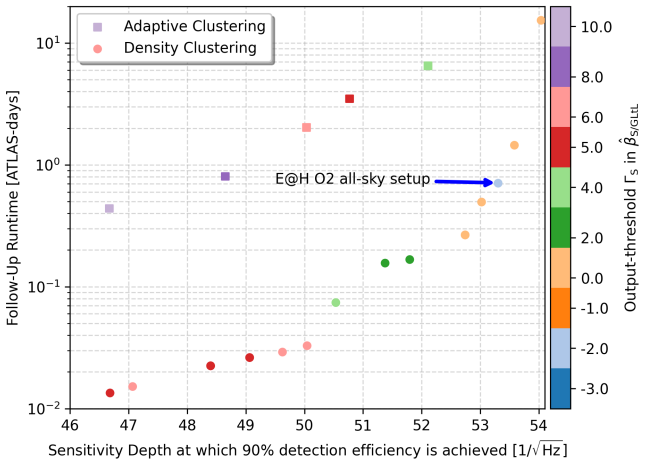

FIG. 3. Performance comparison between the previous clustering method, Adaptive Clustering, and Density Clustering. Each point represents a different clustering set-up, used on the results of the Einstein@Home search [9]. To avoid excessive clutter we do not show all considered set-ups, but rather only those with runtime close to the smallest runtime at each $\mathcal{D}^{90\%}$. The color encodes the Γ_S threshold parameter value. The arrow indicates the density clustering set-up chosen for the follow-up analysis reported in [9].

[2] ATLAS is the super-computer cluster at the MPI for Gravitational Physics in Hannover: `https://www.atlas.aei.uni-hannover.de/`

VI. PERFORMANCE ON E@H O2 ALL-SKY

We compare our Density Clustering with the Adaptive Clustering [15] on the results of the Stage-0 Einstein@Home O2 all-sky search [9].

We characterize the detection efficiency on a set of $\sim$ 2900 fake signals from the target source population of the search: signals with spin-frequencies uniformly distributed; spin-downs log-uniform distributed and all other parameters distributed uniformly: orientation $\cos\iota \in [-1, 1)$, polarization angle $\psi \leq |\pi/4|$, sky position $0 \leq \alpha \leq 2\pi$ and $-1 \leq \sin\delta \leq 1$. The signal amplitude h_0 ranges from loud to faint signals with ~ 1000 signals too faint to be detectable by either method.

The results of the procedure described in the previous Section in order to identify the optimal Density Clustering parameters, are shown in Fig. 3. We compare with the results for the optimal parameter choice for Adaptive Clustering.

The Density Clustering set-up chosen in [9] with a first-stage follow-up runtime-cost of ≤ 1 ATLAS-day is $\approx 10\%$ more sensitive than the Adaptive Clustering set-up at the same computing cost. In continuous gravitational wave searches a 10% improvement, solely due to a better search method, is a big gain.

Perhaps more immediately impressive is the fact that at fixed sensitivity, Density Clustering enables follow-ups that are a factor of $\gtrsim 30$ faster than previous methods.

This gain can be re-invested in deeper follow-ups by using a lower Γ_S, albeit the gain in practice is limited by the steep increase in computing cost for $\Gamma_S \lesssim 4$. With a threshold $\Gamma_S = -3.7$ Density Clustering is able to process two orders of magnitude more candidates than with a threshold $\Gamma_S = 4$, whereas Adaptive Clustering could not be used at all.

The performance of the Adaptive Clustering was characterised in [15] by the detection efficiency and the noise rejection NR defined as

$$\mathrm{NR} := 1 - \frac{N_{out}}{N_{in}}, \tag{7}$$

where N_{in} is the number of candidates above the threshold Γ_S and N_{out} is the number of seeds produced by the clustering procedure.

With a threshold of $\Gamma_S \geq 4$ Adaptive Clustering and Density Clustering achieve similar performance with NR $\geq 99\%$ and detection efficiencies above 98%. At lower thresholds, Adaptive Clustering does not converge in weeks of runtime, indicating that the method struggles to identify over densities due to faint signals. Density Clustering, instead, can probe threshold values as low as -3.7, still achieving NR $\geq 99.99\%$ and attaining a very respectable detection efficiency (now at the 85% level) on a set that includes very faint signals with detection statistic values $\in [-3.7, 4]$, which are much harder to find.

VII. CONCLUSION AND OUTLOOK

We have presented a new, fast and efficient clustering method - Density Clustering - for continuous gravitational wave search post-processing.

Density Clustering works by identifying over-densities of candidates in parameter space: clusters are purely build on candidates' closeness to each other and the detection statistic value is nearly irrelevant. This result may be somewhat surprising because the detection statistic ranks results based on the likeliness of they originating from a signal. However some of our faintest - but still recoverable - signals show detection statistic values at which there are thousands to millions of louder candidates purely from noise. Our results show that in this regime over-densities are a better detection criterion than the significance given by the detection statistic value alone, even in Gaussian noise. This is probably due to the fact that the search is a semi-coherent search.

The over-densities are uncovered by binning the parameter space and this is performed in one pass instead of the previously employed slower iterative procedures. The clustering step is thus largely independent of the number of input candidates, and this allows to process orders of magnitude more candidates with comparable computing resources, probing deeper into the noise.

Until now Einstein@Home searches have returned about $\mathcal{O}\left(10^4\right)$ candidates per work-unit (e.g. [9]), which was more than adequate for what previous clustering algorithms could process. Density Clustering can cluster orders of magnitude more candidates, which means that more results can be inspected, allowing to recover fainter signals in upcoming searches.

The previous clustering method, Adaptive Clustering, assumes compact over-densities, whereas signals typically present X- or Y-shaped over-densities which are hard to capture (and practically impossible to predict). Density Clustering is agnostic about the shape of the over-densities and for this reason it is significantly more effective at identifying even very weak signals.

A different approach of using machine learning for clustering was developed and applied to the Einstein@Home O2 all-sky dataset in [16, 17]. They cluster in $f, \dot{f}$ and achieve better sensitivity depths at fixed false alarms, but lack in sky localization to the point of clustering together candidates from "seemingly unrelated sky positions" [17]. This means that a follow-up would entail searching over the whole sky, whereas Density Clustering restricts the sky position to a patch of $\sim 9\%$ to 0.01% of the full sky, depending on the frequency, between $\sim 20\,\mathrm{Hz}$ to $600\,\mathrm{Hz}$, respectively. Even with the smaller uncertainties in $f, \dot{f}$ and only half the false alarms [17], the computational cost of their approach is higher by one order of magnitude compared to Density Clustering. They propose to generalize to include sky, and the results will be interesting to see.

Clustering is not a problem unique to gravitational wave astronomy, and a number of generic clustering methods exist. For example k-means [20], is a clustering method widely used in a variety of applications including signal-, image- and text-processing, health, cyber security, machine learning and big data [21]. It works based on minimising the cluster-occupants' distance to the cluster center. Limitations of k-means are that the number of clusters must be known a-priori and clusters are assumed to be roughly spherical and similar size. Density-based clustering applications exist: for example DBSCAN [22, 23] and its many generalizations, like e.g. OPTICS [24] or HDBSCAN [25], identify over-densities generated by a minimum number of points within a given volume. They are however not suitable for the large number of points in our results, and they are not as efficient as Density Clustering on our data.

A major advantage of our approach is the versatility of the method. Density Clustering can cluster in any combination of dimensions, so it is easily extendable to e.g. third / higher order spindowns $\ddot{f}, ...$ or to the 5 additional orbital parameters for searches for neutron stars in binary systems. In these searches signal-template offsets in orbital parameters can be to some extent compensated by offsets in frequency- and derivative(s). This translates into correlations between different templates and results in more candidates due to the same root cause [26], making clustering all the more important. All-sky binary searches are computationally extremely expensive and so are the follow-ups. A first test of Density Clustering on the results-data from [27] showed promising results within a few hours of clustering in 6 dimensions $(f, \alpha, \delta, \tau_{\mathrm{asc}}, P_b, a)$, showcasing the flexibility and ease of use of the method presented here.

ACKNOWLEDGMENTS

This work has utilised the ATLAS cluster computing at MPI for Gravitational Physics Hannover. We thank Carsten Aulbert and Henning Fehrmann for their support. We use results from the Einstein@Home search [9] on LIGO data obtained for that search from the Gravitational Wave Open Science Center (gw-openscience.org). We thank again the LIGO-Virgo-KAGRA Collaboration for this service, and LIGO for producing that data.

[1] P. D. Lasky, Gravitational Waves from Neutron Stars: A Review, Publ. Astron. Soc. Austral. **32**, e034 (2015).

[2] B. J. Owen, L. Lindblom, C. Cutler, B. F. Schutz, A. Vecchio, and N. Andersson, Gravitational waves from hot young rapidly rotating neutron stars, Phys. Rev. D **58**, 084020 (1998).

[3] C. J. Horowitz and S. Reddy, Gravitational Waves from Compact Dark Objects in Neutron Stars, Phys. Rev. Lett. **122**, 071102 (2019).

[4] C. Horowitz, M. Papa, and S. Reddy, Gravitational waves from compact dark matter objects in the solar system, Phys. Lett. B **800**, 135072 (2020).

[5] A. Arvanitaki, M. Baryakhtar, and X. Huang, Discovering the QCD Axion with Black Holes and Gravitational Waves, Phys. Rev. D **91**, 084011 (2015).

[6] S. J. Zhu, M. Baryakhtar, M. A. Papa, D. Tsuna, N. Kawanaka, and H.-B. Eggenstein, Characterizing the continuous gravitational-wave signal from boson clouds around Galactic isolated black holes, Phys. Rev. D **102**, 063020 (2020).

[7] M. A. Papa *et al.*, Hierarchical follow-up of subthreshold candidates of an all-sky Einstein@Home search for continuous gravitational waves on LIGO sixth science run data, Phys. Rev. D **94**, 122006 (2016).

[8] B. P. Abbott *et al.* (LIGO Scientific, Virgo), First low-frequency Einstein@Home all-sky search for continuous gravitational waves in Advanced LIGO data, Phys. Rev. **D96**, 122004 (2017).

[9] B. Steltner, M. A. Papa, H. B. Eggenstein, B. Allen, V. Dergachev, R. Prix, B. Machenschalk, S. Walsh, S. J. Zhu, and S. Kwang, Einstein@Home All-sky Search for Continuous Gravitational Waves in LIGO O2 Public Data, Astrophys. J. **909**, 79 (2021).

[10] BOINC, `http://boinc.berkeley.edu/` (2020).

[11] D. P. Anderson, BOINC: A System for Public-Resource Computing and Storage, in *Proceedings of the Fifth IEEE/ACM International Workshop on Grid Computing (GRID04)* (2004) pp. 4–10.

[12] D. P. Anderson, C. Christensen, and B. Allen, Designing a Runtime System for Volunteer Computing, in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing* (2006) pp. 126–136.

[13] J. Aasi *et al.* (LIGO Scientific, VIRGO), Directed search for continuous gravitational waves from the Galactic center, Phys. Rev. D **88**, 102002 (2013).

[14] B. Behnke, M. A. Papa, and R. Prix, Postprocessing methods used in the search for continuous gravitational-wave signals from the Galactic Center, Phys. Rev. D **91**, 064007 (2015).

[15] A. Singh, M. A. Papa, H.-B. Eggenstein, and S. Walsh, Adaptive clustering procedure for continuous gravitational wave searches, Phys. Rev. D **96**, 082003 (2017).

[16] B. Beheshtipour and M. A. Papa, Deep learning for clustering of continuous gravitational wave candidates, Phys. Rev. D **101**, 064009 (2020).

[17] B. Beheshtipour and M. A. Papa, Deep learning for clustering of continuous gravitational wave candidates II: identification of low-SNR candidates, Phys. Rev. D **103**, 064027 (2021).

[18] M. Vallisneri, J. Kanner, R. Williams, A. Weinstein, and B. Stephens, The LIGO Open Science Center, *Proceedings, 10th International LISA Symposium: Gainesville, Florida, USA, May 18-23, 2014*, J. Phys. Conf. Ser. **610**, 012021 (2015).

[19] D. Keitel, Robust semicoherent searches for continuous gravitational waves with noise and signal models including hours to days long transients, Phys. Rev. **D93**, 084024 (2016), 1509.02398.

[20] S. Lloyd, Least squares quantization in pcm, IEEE Transactions on Information Theory **28**, 129 (1982).

[21] M. Ahmed, R. Seraj, and S. M. S. Islam, The k-means algorithm: A comprehensive survey and performance evaluation, Electronics **9**, 10.3390/electronics9081295 (2020).

[22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96 (AAAI Press, 1996) p. 226–231.

[23] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, Dbscan revisited, revisited: Why and how you should (still) use dbscan, ACM Trans. Database Syst. **42**, 10.1145/3068335 (2017).

[24] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, Optics: Ordering points to identify the clustering structure, SIGMOD Rec. **28**, 49–60 (1999).

[25] R. J. G. B. Campello, D. Moulavi, and J. Sander, Density-based clustering based on hierarchical density estimates, in *Advances in Knowledge Discovery and Data Mining*, edited by J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013) pp. 160–172.

[26] A. Singh, M. A. Papa, and V. Dergachev, Characterizing the sensitivity of isolated continuous gravitational wave searches to binary orbits, Phys. Rev. D **100**, 024058 (2019).

[27] P. B. Covas, M. A. Papa, R. Prix, and B. J. Owen, Constraints on r-modes and Mountains on Millisecond Neutron Stars in Binary Systems, Astrophys. J. Lett. **929**, L19 (2022).