Enhancing the capabilities of LIGO time-frequency plane searches through clustering

R Khan[1], S Chatterji[2]

[1] Columbia Astrophysics Laboratory, Columbia University, Pupin Labs Rm 1027, MC 5247, New York, NY 10027 USA

[2] LIGO Laboratory, California Institute of Technology, MS 18-34, Pasadena, CA 91125 USA

E-mail: rmk2109@columbia.edu, shourov@ligo.caltech.edu

Abstract.

One class of gravitational wave signals LIGO is searching for consists of short duration bursts of unknown waveforms. We present a density-based clustering algorithm to improve the performance of time-frequency searches for gravitational wave bursts of unknown waveforms that are extended in time and/or frequency. Potential sources include core collapse supernovae, GRBs, and the merger of binary black holes or neutron stars. We have implemented this algorithm as an extension to the Q Pipeline search for bursts, which determines the statistical significance of events based solely on the peak significance observed in the minimum uncertainty regions of the time-frequency plane. Density based clustering improves the performance of such a search by considering the aggregate significance of arbitrarily shaped regions in the time-frequency plane and rejecting isolated noise triggers. In this paper, we present test results and show that density based clustering improves the performance of Q pipeline for extended signals.

PACS numbers: 04.80.Nn, 07.05.Kf, 95.55.Ym, 95.75.Pq

Submitted to: *Class. Quantum Grav.*

1. Introduction

General Relativity predicts that as concentrations of mass-energy rapidly change quadrupole moment, such as in the case of asymmetric supernovae explosion, the merger of binary compact objects, neutron star instabilities, etc., they create gravitational waves — space-time fluctuations that propagate through the universe at the speed of light [1, 2, 3, 4]. The current generation of gravitational-wave observatories such as the Laser Interferometer Gravitational Wave Observatory (LIGO; Washington and Louisiana, USA) [5], Virgo (Pisa, Italy) [6], GEO600 (Hanover, Germany) [7], TAMA300 (Tokyo, Japan) [8], and proposed ACIGA (Perth, Australia) [9] observatiories apply optical interferometery to detect the 1 part in 10^{21} strain expected due to passing gravitational waves. The sources of GWs detectable to ground based interferometers are customarily classified into four major groups. The inspiral phase of coalescing binary compact objects such as neutron stars or black holes manifest as "chirp" like signals. Stochastic signals can come from either relic GWs from the very early universe or from the cumulative effect of many unresolved sources. Spinning compact objects such pulsars can produce long duration periodic signals if their mass is asymmetrically distributed [3, 4]. Finally, GW signals lasting from a few millisecond to a few seconds, and for which we do not have sufficient theoretical understanding of the source to predict a waveform, are classified as GW burst signals. This last category is the primary focus of this work and includes the merger phase of binary coalescence, core collapse supernovae, gamma ray bursts, and possibly unexpected sources.

For GW waveforms based on theoretical predictions, matched filtering (projection of data onto the expected waveforms) is used. For bursts of unmodeled waveforms, the data is typically projected onto a convenient basis of abstract waveforms that are chosen to cover a targeted region of the time-frequency plane. One of these search algorithms, the Q Pipeline [10, 11], projects the data onto the basis of Gaussian enveloped sinusoids and determines the statistical significance of events based on the most significant single projection in the time-frequency plane. We investigated extensions to this approach that also consider the combined statistical significance of arbitrarily shaped clusters of projections in the time-frequency plane while rejecting noise. Density based clustering algorithms have proven to be the best for our purpose. We present detailed test results and show that density based clustering improves the performance of Q Pipeline for signals that are extended in time and/or frequency.

This paper is structured as follows. Section 2) briefly describes the Q Pipeline burst search algorithm. Section 3) explains the motivations for exploring the advantages of clustering as an extension to the Q Pipeline algorithm. Section 4) describes the concept of density based clustering, and presents a flow chart of the algorithm that we implemented. Section 5) discusses the tests performed for simulated GW burst injections of different waveforms and their results. The conclusions are presented in Section 6).

2. Q Pipeline

The Q pipeline is a comprehensive analysis pipeline for the detection of gravitational-wave bursts using data from single interferometric detector [11]. It is an unmodeled burst search algorithm analogous to matched filtering for waves having sine-Gaussian waveform. It consists of whitening by zero-phase linear prediction [12], application of the discrete Q transform [13], thresholding on the white noise significance of Q transform coefficients [10], and identification of the most significant set of non-overlapping time-frequency tiles. A final stage excludes all but the most significant time-frequency tile within a specified time window in order to prevent the redundant reporting of candidate events. The analysis tool of Q pipeline is the Q-transform [14] which is a modification of the standard short time Fourier Transform [15] in which the analysis window duration varies inversely with frequency. As a result the time-frequency plane is covered by 'tiles' of a constant 'Q' which can be interpreted as a dimensionless quality factor for bursts which is the ratio of the center frequency to the characteristic bandwidth (in terms of second central moments in frequency) of the burst.

The Q pipeline analyzes the time-frequency signal plane looking for non-overlapping tiles that have highest energy among tiles overlapping each other. It finds the most significant event above a threshold in a given signal space. As it projects data into regions of the time-frequency plane, an uncertainty relation applies. The minimum uncertainty signal is a sine-Gaussian, ie. a sinusoid with a Gaussian envelope. The algorithm is therefore based on searching for signals that have this waveform:

$$h(t) = h_0 \sin(\omega_\circ t) e^{-\frac{(t-t_0)^2}{\tau^2}} \tag{1}$$

where $h_\circ$ is the amplitude, $t_\circ$ is the center-time, τ is the length in time, and central angular $\omega_\circ = 2\pi f_\circ$. It is efficient at successfully detecting signals identified by a single tile that are not extended in time and/or frequency scale. However, for extended signals that are so less well localized in the time-frequency plane that its energy is distributed across multiple tiles, their detectability is currently determined by their maximum projection onto the space of sine Gaussian. For extended signals, only identifying the highest energy tile will underestimate the total signal energy most of the time, and hence the associated Signal to Noise Ratio (SNR). The signal can even be missed altogether if the lower value does not pass the threshold of the search. As a result, the Q pipeline's efficiency is not optimal for bursts that are poorly localized in the time-frequency plane. In particular, it only considers the statistical significance of the single most significant tile with minimum time-frequency uncertainty.

Fig. 1 (top) shows the time-frequency map of a hardware injection (a simulated signal physically injected in the detector for test purposes) of the inspiral phase of a binary neutron star coalescence at 5 Mpc away as seen by the Q pipeline, and the non-overlapping tiles produced by the Q pipeline after removing overlapping tiles of lesser energy. In this case, the performance of the Q -pipeline is determined by the highest energy tile (the dark tile at the center of the plot). However, since the Q

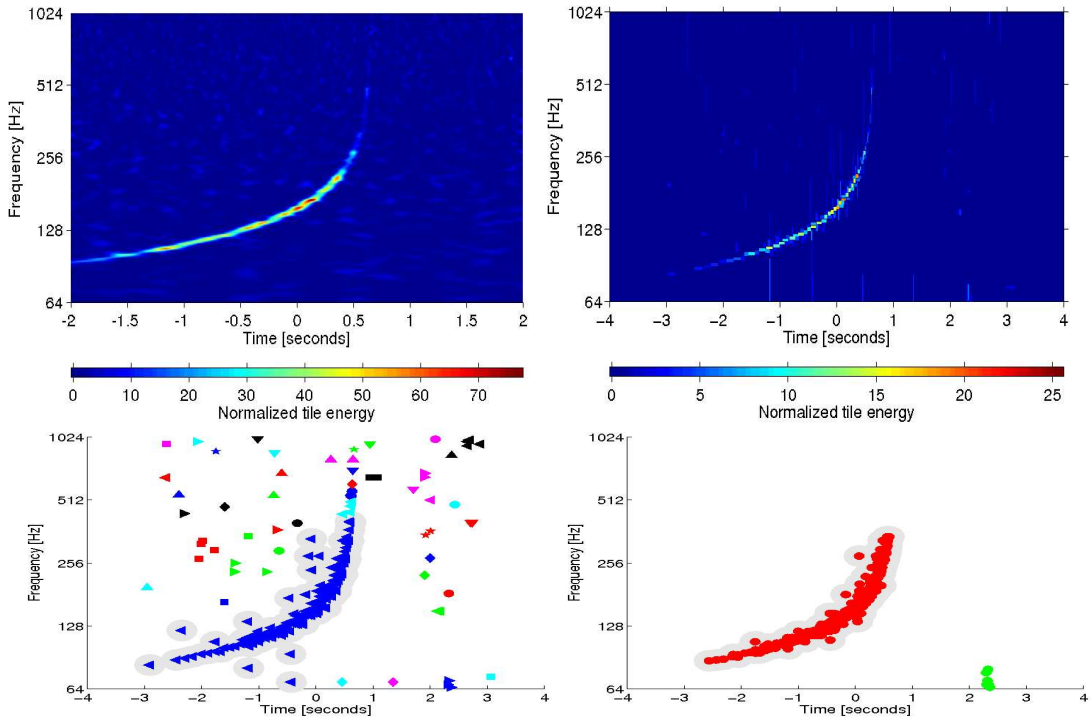

Figure 1. A hardware injection for the inspiral phase of an optimally oriented 1.4-1.4 solar mass binary neutron star merger at 5 Mpc as seen by the Q pipeline (top-left). The Q pipeline keeps only the most significant non-overlapping tiles (top-right). Hierarchical clustering [16] clusters together most of the injection tiles, but also includes some noise tiles. Many individual noise clusters are produced as well. Here each color and shape combination represents one of 68 clusters (bottom-left). Density based clustering [17] clusters together most of the signal-energy while removing most of the noise. The large cluster is related to the injection, the small one is a low frequency detector glitch (bottom-right).

pipeline considers each tile as an individual event, it cannot identify the other tiles of the injection as part of the same signal, and consider them as less significant individual events instead. Also, had there been a nearby glitch with higher energy than the center tile of the injection, the Q pipeline would simply identify that as the most significant event on the signal space even if the total energy of the injection were higher than that of the glitch, and miss the injection.

3. Motivations for Using Clustering

Clustering is the method of grouping elements of data into classes based on their specific properties. In our case, it was expected that clustering to collect energy associated with neighboring tiles would help to more precisely estimate the significance of extended signals, and thus increase the detection efficiency of Q pipeline for signals extended in time and/or frequency. Potential clustering methods include partitioning, hierarchical clustering, and density based clustering [18, 19].

Partitioning approaches, such as the kmeans [20] algorithm, repeatedly divides data points into smaller groups until a certain threshold is met. It was determined not to be useful to find clusters of arbitrary number and shape. Hierarchical algorithms, which build many smaller clusters and keep merging them until a certain threshold is reached, were tested using preexisting MATLAB Statistics Toolbox functions *linkage* [21] and *cluster* [22] in conjunction with a customized measure of distance (Section 4.2) between tiles. It was shown that much of the injection could be clustered together (Fig. 1, bottom-left). Though much of the injection energy is included into one cluster, a lot of noise related clusters are also produced. This makes identifying the most significant cluster statistically difficult. Moreover, noise identified as part of the injection distorts information about the shape of the injected waveform. While hierarchical clustering shows the potential advantage of clustering, density based clustering (Section 4) has been found to be most advantageous for the purpose of this project because of its efficiency in finding arbitrarily shaped regions in the time-frequency space. While most other clustering algorithms classify noise tiles as single member clusters or as part of larger clusters, density based algorithms keep noise, or data points that could not be grouped together with other data points, out of all clusters and identify them as noise.

4. Density Based Clustering

4.1. Concept [17]

Density based clustering facilitates searches for signals of unknown shape, while picking up only significant clusters over a large data set. It does not clutter the output with a list of numerous noise clusters that contain one or just a few data points. The algorithm looks for neighbors of those points that have at least a given number of neighbors within a given distance on the time-frequency plane, and forms clusters of data-points that can be related through their common neighbors (Fig. 2, left). Our implementation of density based clustering algorithm takes two parameter: minimum neighbor number and neighborhood radius.

4.2. Distance Metric

Any clustering algorithm requires measurement of the pairwise distances between all data points and in our case, the pairwise distance between all tiles produced by Q Pipeline. However, the tiles have varied shapes which make measurement of distance between any pair of data points rather difficult. We implemented a distance metric that takes into account the problem of varied tile shapes by utilizing the fact that each tile covers a time-frequency area of 1, that is, for a tile with length in time d and length in frquency f, $df = 1$. It also inflates the distance on frequency scale relative to the distance in time scale. This step has to be taken in order to compensate for the fact that most nonlocalized signals are quite limited on the time scale (seconds) while being comparatively more extended on the frequency scale (Hz). For a pair of tiles with center

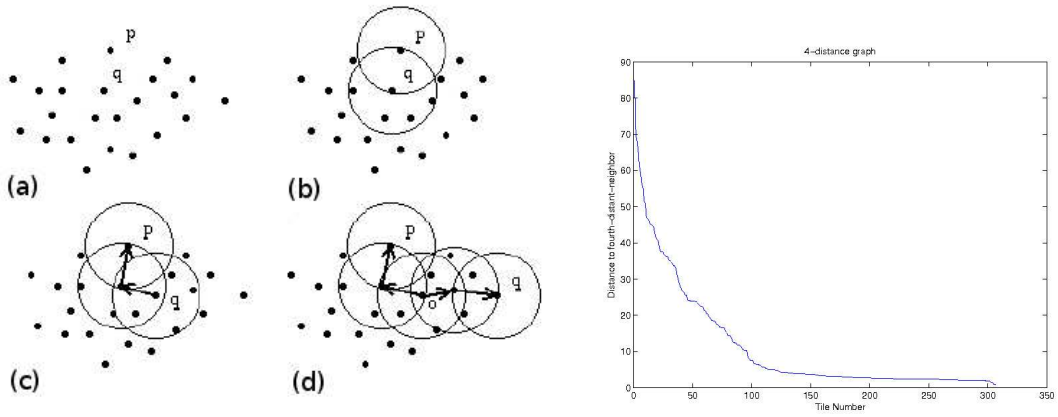

Figure 2. Density based clustering first finds a tile's nearest neighbors, then the neighbors' neighbors, and so on (Figure from [17]). (a) Data points before clustering. (b) If the density of data points within a given distance around a point is above a given threshold to form a cluster, that point becomes a cluster seed. (c) Neighboring data points having a sufficient number of neighbors are then included in the cluster. (d) This process repeats as long as data points with sufficient number of neighbors are found (left). The four-distance graph has the distance of the fourth closest neighbor of every point along y-axis for every corresponding point on x-axis. The sharp turn close to distance 8 provides us with the numerical value of neighborhood radius (right).

times t_1 and t_2, center frequencies f_1 and f_2, Q of q_1 and q_2, and normalized energy of z_1 and z_2, the distance on the time-frequency plane D is measured from the following ralations:

$$D = \sqrt{{D_t}^2 + 30{D_f}^2} \tag{2}$$

$$D_t = \frac{|t_2 - t_1|}{S_t}, D_f = \frac{|f_2 - f_1|}{S_f}, S_t = \frac{d_1 z_1 + d_2 z_2}{z_1 + z_2}, S_t = \frac{b_1 z_1 + b_2 z_2}{z_1 + z_2} \tag{3}$$

$$d_1 = \frac{1}{b_1}, d_2 = \frac{1}{b_2}, b_1 = 2\sqrt{\pi}\frac{f_1}{q_1}, b_2 = 2\sqrt{\pi}\frac{f_2}{q_2} \tag{4}$$

where D_t is the distance on the time scale, D_f is the distance on the frequency scale, S_t is the scale factor on the time scale, S_f is the scale factor on the frequency scale, d_1 and d_2 are durations, and b_1 and b_2 are bandwidths. The distance metric can be compacted as:

$$D = (z_1 + z_2)\sqrt{\left(\frac{2\sqrt{\pi} f_1 f_2 (t_2 - t_1)}{z_1 q_1 f_2 + z_2 q_2 f_1}\right)^2 + 30\left(\frac{q_1 q_2 (f_2 - f_1)}{2\sqrt{\pi}(z_1 q_2 f_1 + z_2 q_1 f_2)}\right)^2}. \tag{5}$$

4.3. Neighborhood Radius

The exact numerical value of the neighborhood radius is determined using a 4-distance graph that has the distance of the fourth closest neighbor of every point along y-axis for every corresponding point on x-axis. The points are sorted according to descending

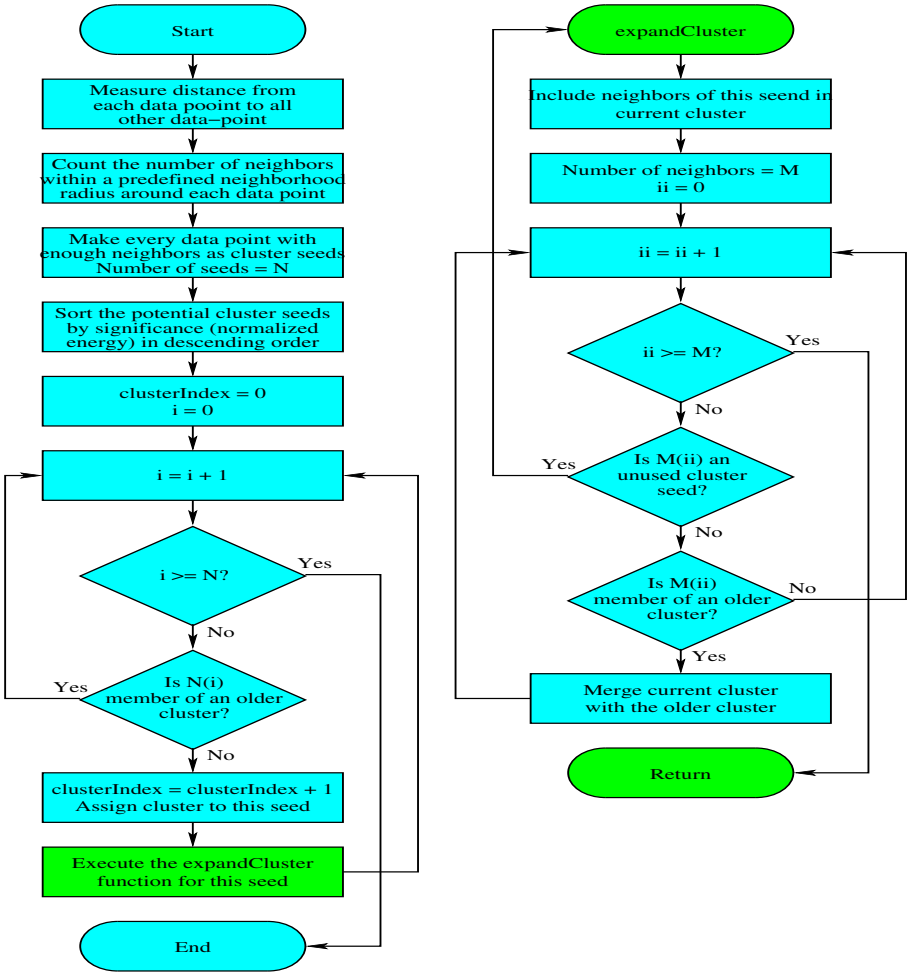

Figure 3. Flowchart of density based clustering algorithm.

order of their 4-distance value. Close observation of the 4-distance graph provides a cut-off distance whose numerical value depends on the distance metric used. For the specific distance metric we used, the numeric value of 8 has been chosen as the neighborhood radius from observing the 4-distance plot that we produced (Fig. 2, right) that evidently shows a sharp turn in the plot near that point.

4.4. Algorithm

The main clustering function first uses the distance function to measure pairwise distance between all tiles, and calls the expandCluster function which recursively calls itself to induct more data points into the cluster (Fig. 3). Clustering starts at the highest energy data-point and then proceeds to the next significant data point that is not in a cluster, considering only such points as cluster seeds that have enough (4 in our case) neighbors to ensure that the least number of loops are executed. If any qualifying member of the current cluster is found to be already in a cluster, the two clusters merge. Thus, regardless of which data-point the algorithm starts clustering from, it will always find

the cluster. For speed optimization, though, our density based clustering function picks the more significant data-points first. Fig. 1 (bottom-right) shows a cluster built using density based clustering algorithm. It shows that density based clustering has clustered together the most significant part of the previously discussed injection successfully, and almost all the noise is removed. While it loses the high-frequency end of the injection, that part contains very little energy which does not significantly contribute to the duration or significance estimation of the detected trigger. The only noise cluster on the signal space is a low frequency detector glitch.

5. Testing for Different Waveforms

We evaluated our density based clustering code implementation on single detector searches for injections of simulated bursts of different waveforms at constant SNR. Our test program loads segments of LIGO S5 detector data (that is data collected by LIGO during its ongoing fifth science run) and runs clustering over the noise-only signal space. Every detection on the noise-only space is considered a false detection. Then it repeatedly injects constant SNR signals of specific waveforms at random times with random signal parameters (bandwidth, duration, strength, mass of component stars etc.), and clusters noise-injection data over the same signal space separately. Every injection that is successfully detected after clustering is considered a successful correct detection. We produced receiver operating characteristic (ROC) curves for each injected signal population. ROC curves plot the false-rate of a search algorithm on the x-axis against its detection-efficiency on the y-axis as the detection threshold is varied, which in this case is a threshold on the energy of a tile or total energy of all tiles in a cluster. Three other figures plotting detection-efficiency against energy, false-rate against energy, and energy against number-of-tiles are also produced for analyzing the effects of clustering on injections of different waveforms (Fig. 4).

A total of five waveform families have been tested for, including two non-localized waveforms: inspiral and noise-burst, and three localized waveforms: ringdown, sine-Gaussian, and Gaussian. Fig. 4 shows ROC curves produced for the inspiral and sinegaussian waveform injections. These ROC curves are produced for injections at constant SNR with one injection per 32 second LIGO data collected during the ongoing fifth science run (S5). The red curves represent the performance of the Q pipeline without clustering and the blue curves represent the performance of the Q pipeline with clustering. Since density based clustering identifies and removes tiles that do not have enough neighboring tiles within the given neighborhood radius, localized injections that have most of their energy contained in a single tile are identified as noise. Thus, use of density based clustering can cause Q pipeline to miss extremely localized signals that contain most of its energy in a single tile on the time-frequency signal space. To maintain the original Q pipeline's sensitivity for localized signals while expanding it to find non-localized signals through clustering, results from Q pipeline with and without clustering can be combined carefully avoiding double-counting. In ROC's presented

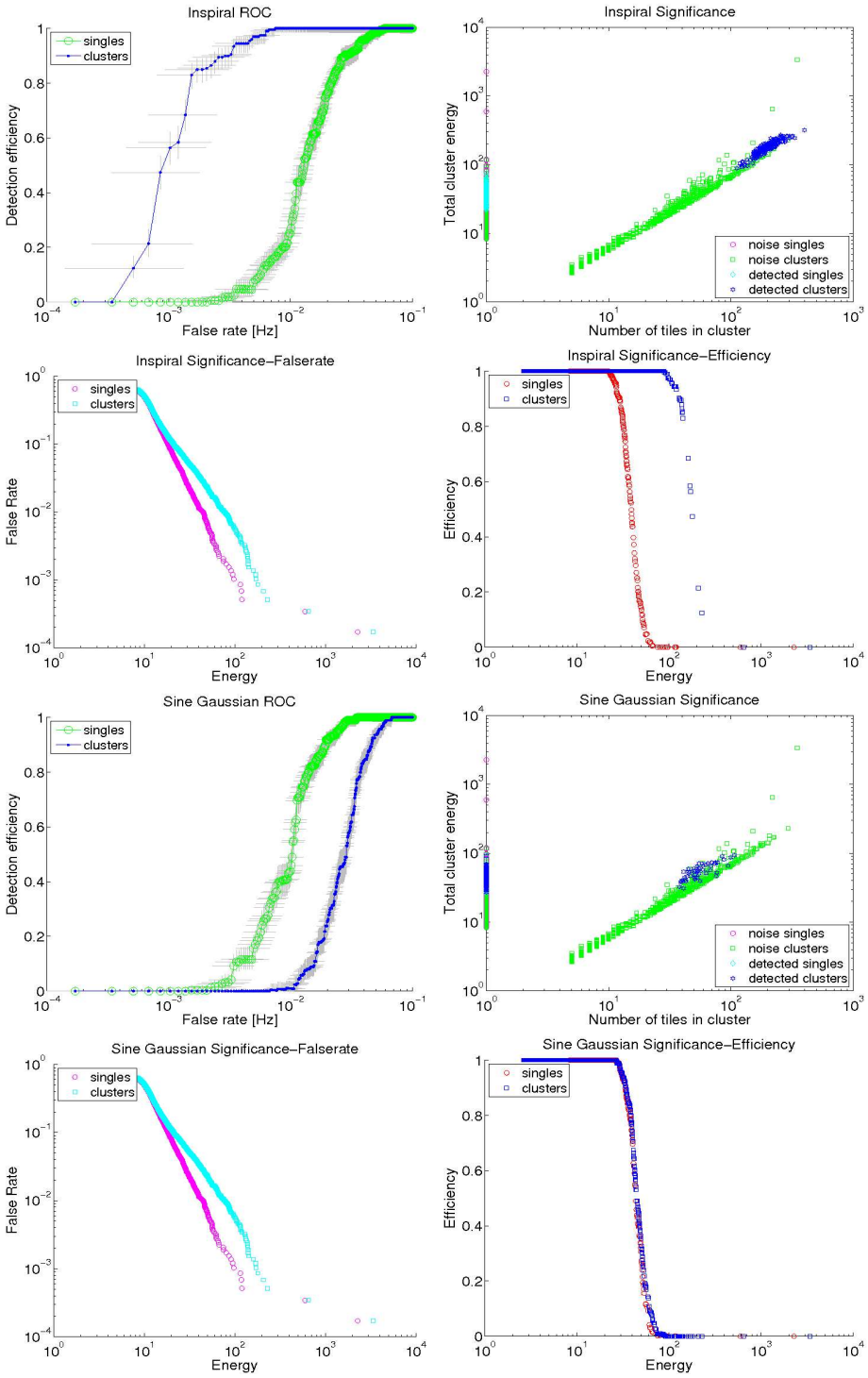

Figure 4. ROC curve, number of tiles vs. energy plot, false-rate vs. energy plot, and efficiency vs. energy plot for 200 Inspiral injections (top) and sine Gaussians (bottom) at constant signal to noise ratio injected into LIGO data collected during the ongoing fifth science run (S5).

here, the blue curves represent such merged results.

The ROC curves indicate that for the extended waveforms (top), clustering significantly improves the ROC curves as efficiency increases. For the localized waveforms (bottom), the ROC curves slightly worsens though efficiency remains unchanged. Higher false-rate is observed (false-rate vs. energy plots in Fig. 4) in all the cases which was expected due to merger of results. However, merging the results can be a necessary step when searching for signals that are not extended on the time-frequency plane to ensure that Q pipeline with clustering does not disregard significant localized triggers that it would otherwise find without clustering. This step is not applicable when searching for only extended signals. However, it also demonstrates that to search only for signals localized both in time and frequency, density based clustering would not offer any advantage. It is possible to recover any performance in between the two ROC curves by setting separate thresholds for clustering and non-clustering triggers. However, to do that we need to specify whether to look for extended or loaclized waveforms.

6. Conclusion

Methods of clustering the measurements from neighboring or overlapping basis functions have been employed to more efficiently detect signals that are not well represented by this particular choice of basis. Adding density based clustering algorithm to Q Pipeline for statistically significant events led to an improvement in the detectability GW burst signals extended in time and/or frequency scales. Our implementation of density based clustering facilitates Q Pipeline to find clusters of unknown shapes and rejects noise without slowing down the search.

Since all the testing so far has been done using single detector data, the logical next step is to incorporate clustering in to Q pipeline, and implement coherent and co-incident search capabilities. An improvement of the false rate is expected for coherent and co-incident searches. Clustering can help us to extract information about signal shapes by identifying GW signal energy distributions across time and frequency, and investigate other signal characteristics that potential search methods can not recognize without clustering. We recommend further research to explore these promising aspects.

Acknowledgments

The authors are grateful for the support of the United States National Science Foundation under cooperative agreement PHY-04-57528, California Institute of Technology, and Columbia University in the City of New York. We are grateful to the LIGO collaboration for their support. We are indebted to many of our colleagues for frequent and fruitful discussion. In particular, wed like to thank Albert Lazzarini for his valuable suggestions regarding this project, and Luca Matone, Zsuzsa Márka, Sharmila Kamat, Jameson Rollins, Peter Kalmus, John Dwyer, and Szabolcs Márka for their thoughtful comments on the manuscript.

The authors gratefully acknowledge the support of the United States National Science Foundation for the construction and operation of the LIGO Laboratory and the Particle Physics and Astronomy Research Council of the United Kingdom, the Max-Planck-Society and the State of Niedersachsen / Germany for support of the construction and operation of the GEO600 detector. The authors also gratefully acknowledge the support of the research by these agencies and by the Australian Research Council, the Natural Sciences and Engineering Research Council of Canada,the Council of Scientific and Industrial Research of India, the Department of Science and Technology of India, the Spanish Ministerio de Educaciony Ciencia, The National Aeronautics and Space Administration, the John Simon Guggenheim Foundation, the Alexander von Humboldt Foundation,the Leverhulme Trust, the David and Lucile Packard Foundation, the Research Corporation, and the Alfred P. Sloan Foundation. The LIGO Observatories were constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation under cooperative agreement PHY-9210038. The LIGO Laboratory operates under cooperative agreement PHY-0107417. This document has been assigned LIGO document number LIGO-P070041-00-Z.

References

[1] D. Sigg. Gravitational Waves. In H. V. Klapdor, editor, *Neutrinos*, pages 592–+, January 1988.

[2] K. S. Thorne. Gravitational Waves. In E. W. Kolb and R. D. Peccei, editors, *Particle and Nuclear Astrophysics and Cosmology in the Next Millenium*, pages 160–+, 1995.

[3] S. A. Hughes et al. New physics and astronomy with the new gravitational-wave observatories. *ArXiv Astrophysics e-prints*, October 2001.

[4] B. C. Barish and R. Weiss. LIGO and the detection of gravitational waves. *Physics Today*, 52:44–50, 1999.

[5] B. O'Reilly. Status of LIGO. *APS Meeting Abstracts*, November 2006.

[6] F. Acernese et al. The Virgo status. *Classical and Quantum Gravity*, 23:635–+, October 2006.

[7] B. Abbott, R. Abbott, and R. Adhikari. Detector description and performance for the first coincidence observations between LIGO and GEO. *Nuclear Instruments and Methods in Physics Research A*, 517:154–179, January 2004.

[8] S. Fairhurst, the LIGO Scientific Collaboration, H. Takahashi, and the TAMA Collaboration. Status of the joint LIGO TAMA300 inspiral analysis. *Classical and Quantum Gravity*, 22:1109–+, September 2005.

[9] P. J. Barriga et al. Status of ACIGA High Power Test Facility for advanced interferometry. In J. Hough and G. H. Sanders, editors, *Gravitational Wave and Particle Astrophysics Detectors. Edited by Hough, James; Sanders, Gary H. Proceedings of the SPIE, Volume 5500, pp. 70-80 (2004).*, pages 70–80, September 2004.

[10] S. K. Chatterji. *The search for gravitational wave bursts in data from the second LIGO science run*. PhD thesis, Massachusetts Institute of Technology, 2005.

[11] S. Chatterji. The Q Pipeline search for gravitational-wave bursts with LIGO. *APS Meeting Abstracts*, pages 11007–+, April 2006.

[12] J. Makhoul. Linear Prediction: A Tutorial Review. In *Proc. IEEE, Volume 63, p. 561-580*, pages 561–580, 1975.

[13] J. C. Brown. Calculation of a constant Q spectral transform . *Acoustical Society of America Journal*, 89:425–434, January 1991.

[14] S. Chatterji et al. Multiresolution techniques for the detection of gravitational-wave bursts. *Classical and Quantum Gravity*, 21:1809–+, October 2004.

[15] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers*, 93:429–457, November 1946.

[16] S. Krishnamachari and M. Abdel-Mottaleb. Hierarchical clustering algorithm for fast image retrieval. In M. M. Yeung, B.-L. Yeo, and C. A. Bouman, editors, *Proc. SPIE Vol. 3656, p. 427-435, Storage and Retrieval for Image and Video Databases VII, Minerva M. Yeung; Boon-Lock Yeo; Charles A. Bouman; Eds.*, pages 427–435, December 1998.

[17] Martin Ester et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[18] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[19] J. Sylvestre. Time-frequency detection algorithm for gravitational wave bursts. *Phys. Rev. D*, 66(10):102004–+, November 2002.

[20] Tapas Kanungo et al. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):881–892, 2002.

[21] http://www.mathworks.com/access/helpdesk/help/toolbox/stats/linkage.html.

[22] http://www.mathworks.com/access/helpdesk/help/toolbox/stats/cluster.html.