Enhancing the Capabilities of LIGO Time-Frequency Plane Searches Through Clustering

Authors: Rubab Khan (GECo, Columbia), Shourov Chatterji (LIGO, Caltech)

Abstract: One class of signals LIGO is searching for consists of short duration gravitational wave bursts of unknown waveforms. Potential sources include core collapse supernovae and the coalescence of binary black holes. To detect such events, existing search algorithms project the LIGO data stream onto various time-frequency bases and then search for regions of excess signal energy. One of these search algorithms, the Q Pipeline, determines the statistical significance of events based solely on the loudest element observed in the time-frequency plane. We investigated extensions to this approach that also considers the statistical significance of arbitrarily shaped clusters 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 extended in time and/or frequency.

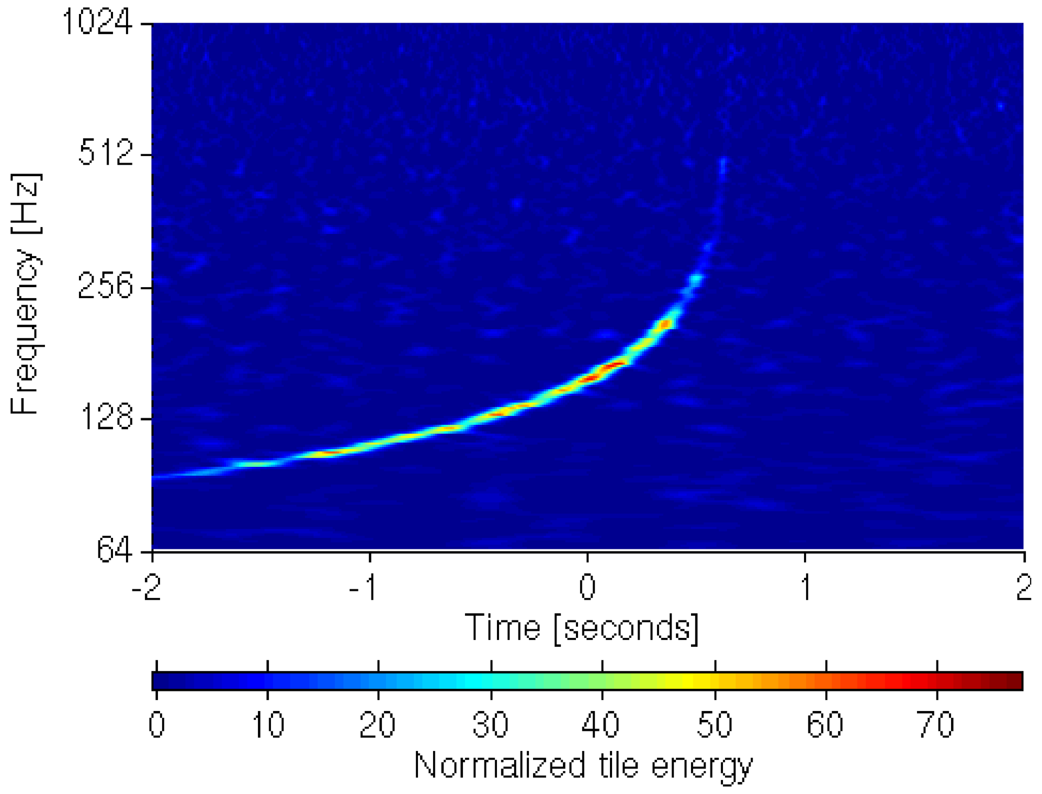


Figure-1: A hardware injection for the inspiral phase of an optimally oriented 1.4, 1.4 solar mass binary neutron star merger at 5 mega-parsecs as seen by the Q pipeline.

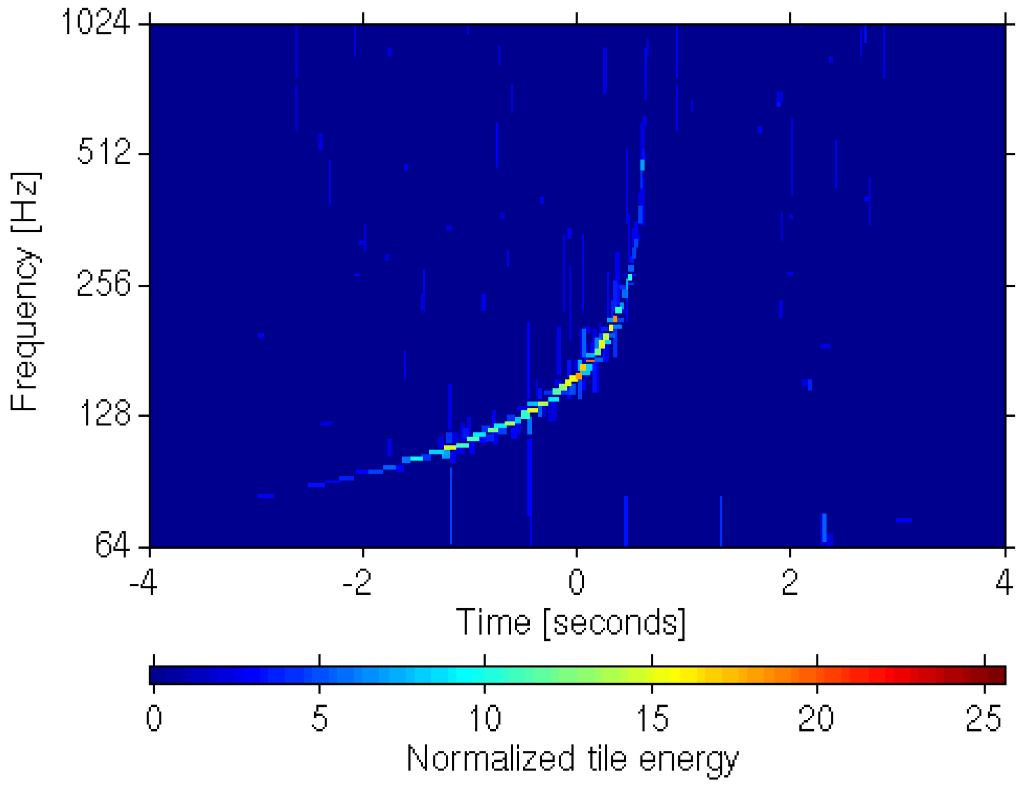


Figure-2: The Q pipeline keeps only the most significant non-overlapping tiles.

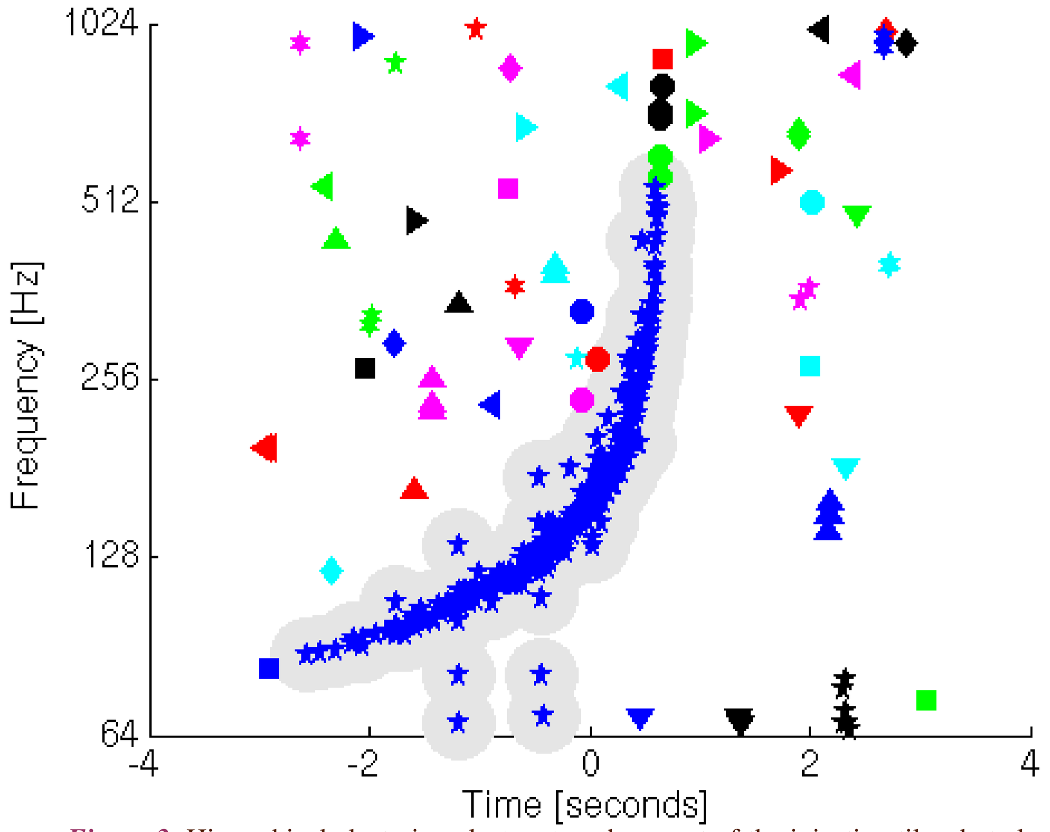


Figure-3: Hierarchical clustering clusters together most of the injection tiles, but also includes some noise tiles. A lot of individual noise clusters are produced as well. Here each color and shape combinations represent a cluster, totaling ~68.

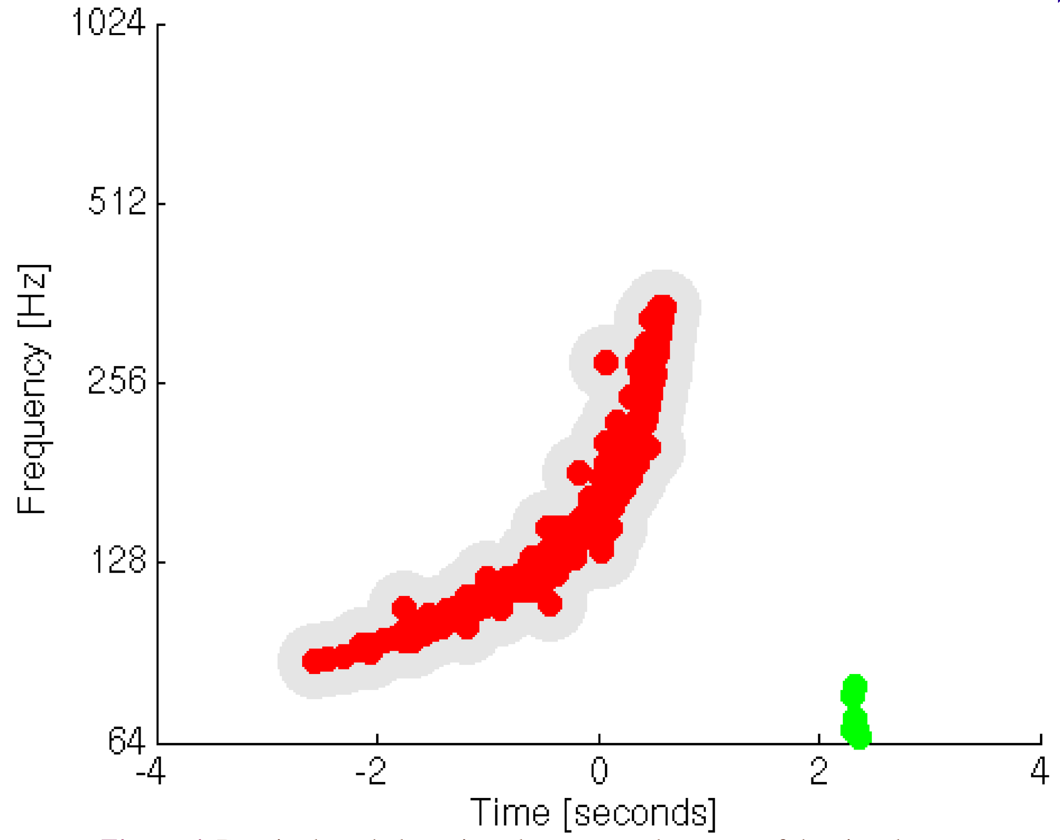


Figure-4: Density based clustering clusters together most of the signal-energy while removing most of the noise. The large red cluster is related to the injection, the small green one is a low frequency detector "glitch".

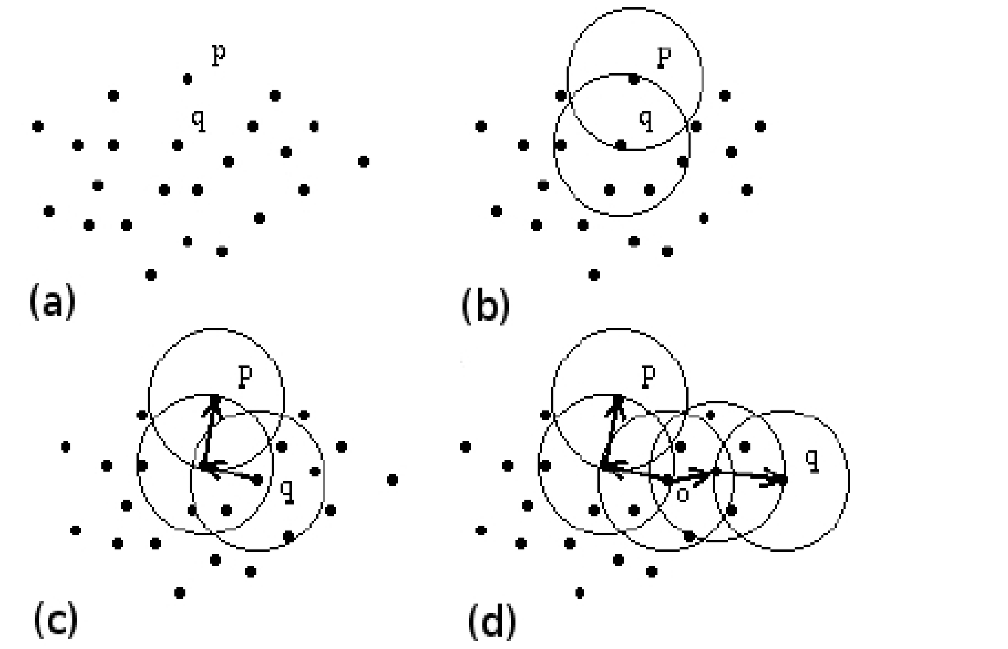


Figure-5: Density based clustering first finds a tile's nearest neighbors, then the neighbors' neighbors, and so on. (a) Data points before clustering. (b) If the density of data points "near" a point is sufficient, that point becomes a cluster seed. (c) Neighboring data points having a sufficient number of neighbors are then included in the cluster. (d) This process repeats as long as data points with sufficient number of neighbors are found. (Figure from M. Ester et. al.)

Introduction:

1. Gravitational Wave signals lasting from a few millisecond to a few seconds long events, and for which we do not have sufficient understanding to predict the waveform, are classified as GW bursts. That includes merger phase of binary coalescence, core collapse supernovae, gamma ray bursts, and other possible unexpected sources.
2. For known GW burst waveforms, matched filtering based on the projection of data onto the expected waveforms, is used.
3. For bursts of unmodeled waveform, the data is typically projected onto a convenient basis of abstract waveforms that are chosen to cover a targeted region of the time-frequency signal space. We investigated extensions to this approach by exploiting the advantages of data clustering.

Q-Pipeline:

1. Unmodeled burst search algorithm analogous to matched filtering for waves having sinusoidal-Gaussian waveform.
2. Analyzes the time-frequency signal plane looking for non-overlapping tiles (approximately: pixels) that have higher energy than nearby tiles. Finds the most significant "event" above a certain threshold in a given signal space (Fig-2).
3. Works very well for signals identified by a single tile that are not extended in time and/or frequency scale.
4. For signals with energy distributed across multiple tiles, only identifying the highest energy tile will underestimate the total signal energy (and therefore SNR). The signal can be missed altogether if the lower value does not pass the threshold of the search.

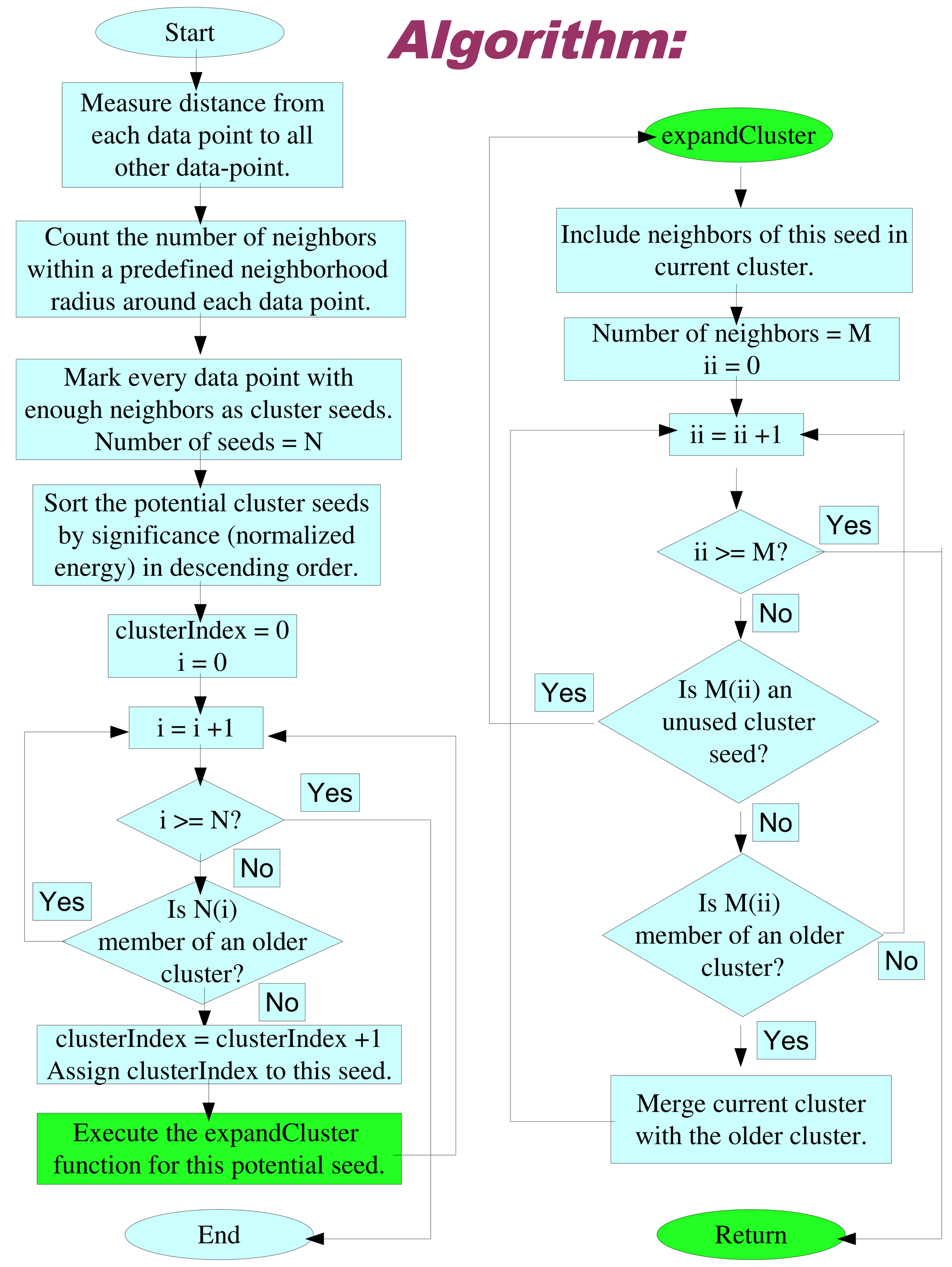
Motivations for Clustering:

1. It was expected that clustering to collect energy would help to more realistically estimate the significance of extended signals.
2. Clustering together multiple tiles from the same signal (or glitch) would thus increase the detection efficiency of Q pipeline for signals extended in time and/or frequency.
3. Hierarchical algorithms were tested using preexisting Matlab functions in conjunction with a customized measure of distance between tiles (Fig-3).
4. 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 included into the injection cluster distorts information about the shape of the injection.

Density Based Clustering:

1. Clusters together the most significant part of an injection successfully, and almost all the noise is removed (Fig-4).
2. Picks a data point that has a given number of neighbors; for all neighbors that also have enough neighbors, include them in the cluster; and so on; until it reaches a point that doesn't have enough neighbors (Fig-5).
3. To maintain 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.
4. Reference: M. Ester et. al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". http://ifsc.ualr.edu/xwxu/publications/kdd-96.pdf

Algorithm:



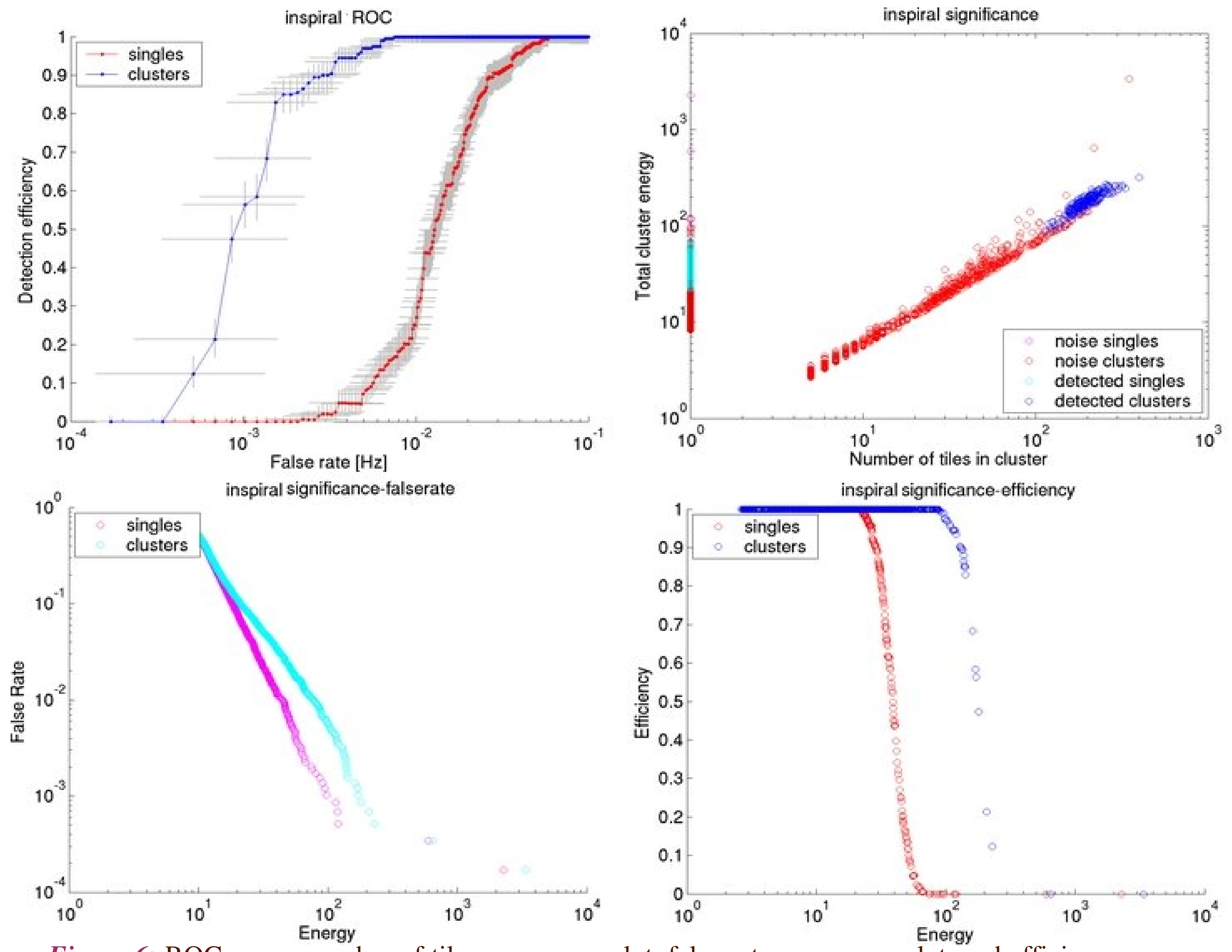


Figure-6: ROC curve, number of tiles vs. energy plot, false-rate vs. energy plot, and efficiency vs. energy plot for 200 Inspiral injections at constant signal to noise ratio (SNR) injected into LIGO data collected during the ongoing fifth science run (S5).

Testing for Different Waveforms:

1. Single detector search for 200 injections; simulated bursts of different waveforms at constant SNR.
2. Five waveforms tested: Signal Waveforms extended in time and/or frequency (Inspiral, and Noise Burst) and those that are localized (Sinusoidal Gaussian, Gaussian, and Ring Down).
3. Extended Waveforms: Significantly improved ROC's as efficiency increases (Fig-6).
4. Localized Waveforms: Slightly worse ROC but efficiency remains unchanged (Fig-7).
5. Higher false-rate in all cases. This was expected due to merger of results.

By setting separate thresholds for clustering and non-clustering triggers it is possible to recover any performance in between the two ROC curves (Fig-6 and Fig-7). However, we need to specify whether to look for extended or non-extended waveforms in that case.

Conclusions and Future Directions:

Conclusions:

1. **Clustering helps finding bursts of unknown waveforms extended in time and/or frequency.**
2. Density based clustering helps finding clusters of arbitrary shapes, and rejects noise.
3. Our implementation of density based clustering adds only trivial processing time to Q pipeline.

Future Directions:

1. Extract information about signal shapes and energy distributions across time and frequency.
2. Incorporate clustering in to Q pipeline, and implement coherent and co-incident search capabilities.
3. Improvement of the false rate is expected for coherent and co-incident searches.

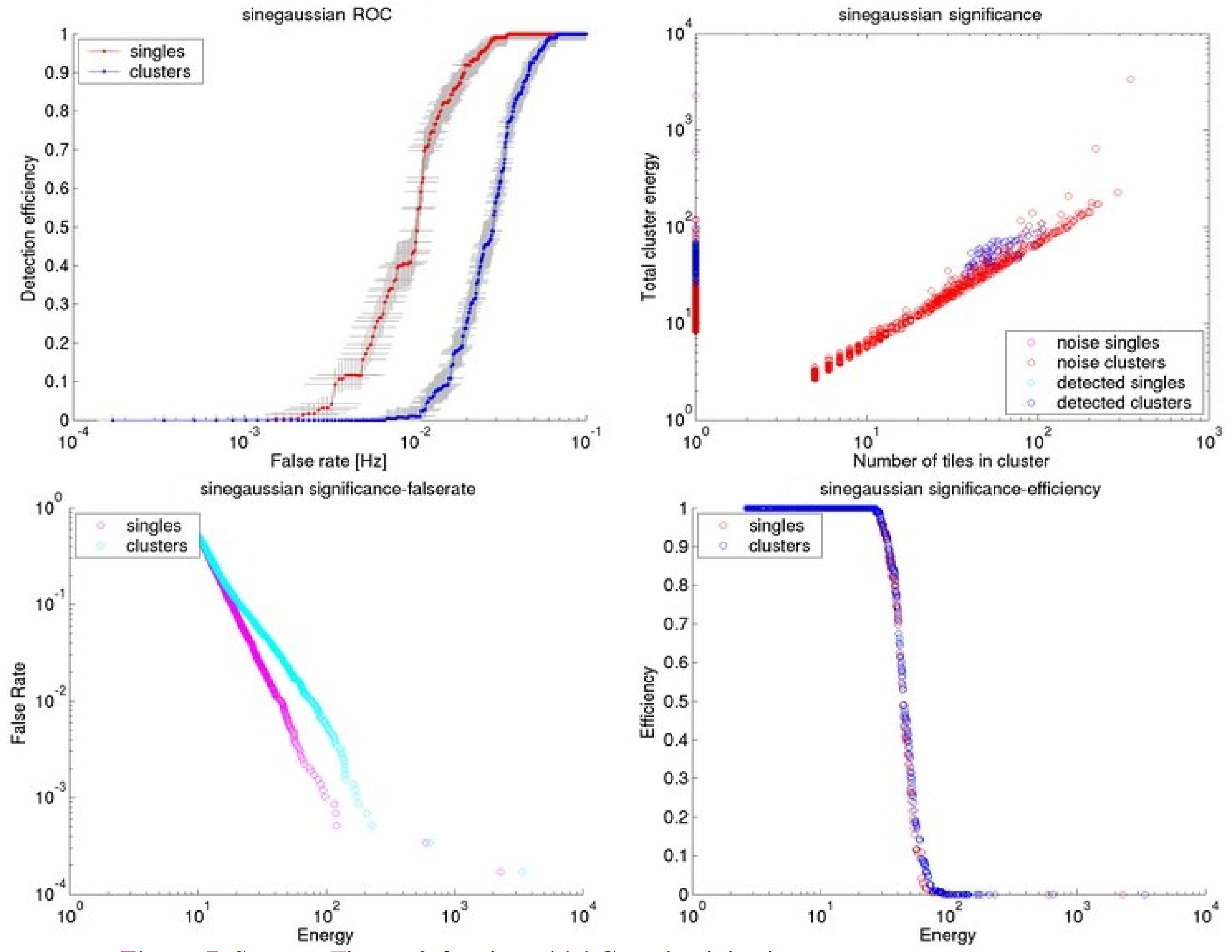


Figure-7: Same as Figure-6, for sinusoidal Gaussian injections at constant.

Acknowledgements: *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, we'd like to thank Albert Lazzarini for his valuable suggestions regarding this project, and Patrick Sutton, Laura Cadonati, Zsuzsa Marka, Luca Matone, and Szabolcs Marka 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.

LIGO-G060607-00-Z

Further Details: http://www.ligo.caltech.edu/~rkhan/burst.html