Methods for reducing false alarms in searches for compact binary coalescences in LIGO data

Methods for reducing false alarms in searches for compact binary coalescences in LIGO data

J Slutsky[1], L Blackburn[2], D A Brown[3], L Cadonati[4], J Cain[5],
M Cavaglià[5], S Chatterji[6], N Christensen[7], M Coughlin[7], S Desai[8],
G González[1], T Isogai[7], E Katsavounidis[2], B Rankins[5], T Reed[9],
K Riles[10], P Shawhan[11], J R Smith[3], N Zotov[9] and J Zweizig[6]

[1] Louisiana State University, Baton Rouge, LA 70803, USA
[2] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3] Syracuse University, Syracuse, NY 13244-1130, USA
[4] University of Massachusets—Amherst, Amherst, MA 01003, USA
[5] University of Mississippi, University, MS 38677, USA
[6] California Institute of Technology, Pasedena, CA 91125, USA
[7] Carleton College, Northfield, MN 55057, USA
[8] NCSA at University of Illinois, Urbana-Champaign, IL 61801, USA
[9] Louisiana Tech University, Ruston, LA 71272, USA
[10] University of Michigan, Ann Arbor, MI 48109, USA
[11] University of Maryland, College Park, MD 20742, USA

E-mail: jsluts@tigers.lsu.edu

Abstract

The LIGO detectors are sensitive to a variety of noise transients of non-astrophysical origin. Instrumental glitches and environmental disturbances increase the false alarm rate in the searches for gravitational waves. Using times already identified when the interferometers produced data of questionable quality, or when the channels that monitor the interferometer indicated non-stationarity, we have developed techniques to safely and effectively veto false triggers from the compact binary coalescences search pipeline.

PACS numbers: 04.30.Tv, 04.80.Nn, 07.60.Ly

(Some figures in this article are in colour only in the electronic version)

1. Introduction

In October 2007, the Laser Interferometer Gravitational-wave Observatory (LIGO) [1] detectors completed a fifth data run over a 2 year long period, denoted as S5, during which 1 year of triple-coincidence data was collected at design sensitivity. During S5, LIGO consisted of two interferometers in Hanford, Washington and one in Livingston, Louisiana. In Hanford, the interferometers had arm lengths of 4 km and 2 km, referred to as H1 and H2, respectively.

In Livingston, there was a single 4 km interferometer, referred to as L1. The LIGO detectors were sensitive to the coalescences of massive compact binaries up to 30 Mpc for neutron stars [2, 3], and even farther for binary black holes. Ongoing 'Enhanced' LIGO upgrades, as well as the future 'Advanced' LIGO upgrades, are planned to increase the sensitivity significantly.

The LIGO Scientific Collaboration (LSC) performs astrophysical searches for gravitational waves from compact binary coalescences (CBCs), including the inspiral, merger and ringdown of compact binary systems of neutron stars and black holes. These searches use matched filtering with template banks that include a variety of durations and frequency ranges for inspirals or ringdowns. When a LIGO interferometer is locked and operational, data are recorded in what is called 'science mode'. In order to confidently make statements on astrophysical upper limits and detections from science mode data, characterization of the LIGO detectors and their data quality is vital. The LIGO detectors are sensitive to a variety of noise transients of non-astrophysical origin, including disturbances within the instrument and environmental noise sources. Triggers generated by these disturbances may occur at different times and with different amplitudes for each template, increasing the false alarm rate in the searches for gravitational waves.

In this paper we discuss techniques for vetoing non-astrophysical transient noises effectively, and thereby reducing their effect on searches for gravitational waves. These methods were developed on searches for low mass CBCs in the first of the 2 years of S5 as described in [2, 3], though they are applicable to future searches as well.

In section 2 we present the broad range of data quality issues that were present in LIGO data. In section 3 we briefly review the search methods for which these techniques were developed. The techniques that we have so far developed and implemented to evaluate vetoes are explained in section 4. The categorization and application of these vetoes is described in section 5. In section 6 we describe proposed methods to extend and automate our vetoes for use in future CBC searches. We present our conclusions in section 7.

2. Data quality studies

There are two broad categories of spurious transients in LIGO data: instrumental and environmental noises. Within these two classes of noise sources, there are dozens of identified phenomena that require vetoing. LIGO records hundreds of channels of data on the state of internal degrees of freedom of the interferometers, as well as the output from environmental sensors located nearby.

When a set of transients affects the stability or sensitivity of the gravitational wave data, members of the Detector Characterization and Glitch groups within the LSC work to determine the source of these transients, using the auxiliary channel data [4]. When a given noise source is identified, it is documented with sets of time intervals referred to as data quality flags, which begin and end on integer GPS seconds. These data quality flags are discussed in sections 2.1 and 2.2.

Alternatively, one can begin with the hundreds of data channels, and search for correlations between the outputs of some algorithm on these channels and the gravitational wave data channel. The 'auxiliary channel' vetoes used by the CBC search are described in section 2.3.

2.1. Data quality flags from instrumental noise transients

Each LIGO interferometer makes an extremely sensitive comparison of the lengths of its arms. This necessitates the ability to sense and control minute changes in displacement and alignment of the suspended optics, as well as intensity and phase in the laser [1]. The control

systems have both digital and analog components, as well as a variety of complex filtering schemes. Instrumental noise transients, or *glitches*, sometimes correspond to fluctuations of a large amplitude and short duration in the control systems. While these may be prompted at times by environmental effects, they can be identified and vetoed by using the control channels alone, as they are well-known failure modes of the control systems. We describe some examples in the following paragraphs.

- Overflows: the feedback control signals used to control the interferometer arm lengths and mirror alignments are processed and recorded in digital channels. When the amplitude of such a signal exceeds the maximum amplitude the channel can accommodate, it 'overflows', and the signal abruptly flattens to read as this maximum value, until the quantity falls back below this threshold. This discontinuity in the control signal usually introduces transients at the time of the overflow.
- Calibration line dropouts: signals of a single frequency are continuously injected into the feedback control system to provide calibration. Occasionally, these signals 'drop out' for short periods of time, usually 1 s. This discontinuous jump in the control signal produces artifacts in the data both when the calibration line drops out and when it resumes.
- Light scattering: the two interferometers at the Hanford site share the same vacuum enclosure. During times when one of the interferometers was locked and in the science mode, and the other was not locked, the swinging of the mirrors of the unlocked interferometer scattered light into the locked interferometer. This produced strong, short-duration transients, though not necessarily overflows.
- Arm cavity light dips: brief mirror misalignments caused drops in the power in the arm cavities, and thus transients in the data.

2.2. Data quality flags from environmental noise transients

Environmental noise transients correspond to the coupling of mechanical vibrations and electromagnetic glitches that enter into the interferometer. Seismic motion, human activity near the LIGO sites and weather are the most common sources of mechanical vibrations. Similar to instrumental transients, we describe some examples that have been identified by the LIGO detector characterization group in the following paragraphs.

- Electromagnetic disturbances: the electronic systems of the interferometers are susceptible to electromagnetic interference, both due to glitches in the power lines and electronics noise at both sites. Magnetometers arrayed around the detector are used to diagnose these signals.
- Weather-related transients: the Hanford site is arid, with little to block the wind from pushing against the buildings housing the interferometers. Times at Hanford with wind speeds above 30 mph are problematic. Weather and ocean waves also contribute to ground motion in the frequency range 0.1–0.35 Hz at both sites, particularly Livingston.
- Seismic disturbances: the seismic activity from different noise sources have different characteristic frequencies. Earthquakes around the globe introduce transient noise in the frequency range 0.03–0.1 Hz. Nearby human activities, such as trucks, logging, and trains, produce disturbances with frequencies greater than 1 Hz, and can be so extreme that the interferometers often cannot stay locked. Even when they remain locked, significant noise transients frequently occur.

The ground motion described in the latter two points above occurs at frequencies below a few Hz, while the frequency range in which the interferometers are most sensitive is from 100–1000 Hz. There is a nonlinear coupling from the ground motion into the interferometers,

which results in increased glitching at higher frequencies during times of high ground motion, and for this reason, these environmental effects are especially important.

2.3. KleineWelle triggers from auxiliary channels

Data quality flags identify times affected by specific issues and use auxiliary channels as appropriate to locate the problems. Rather than starting with a known or plausible noise coupling, we search over hundreds of the auxiliary channels for transients that are coincident between any of these channels and the gravitational wave channel. This information can be used both for producing veto intervals and for finding clues about the problems left unflagged.

A wavelet-based algorithm, KleineWelle (KW) [5], is used to analyze interferometer control and environmental data. It is a valuable source of triggers for detector characterization because of its low computational cost allowing it to be applied to many data channels. The algorithm that produces trigger lists contains the peak time and significance of the trigger. During S5, KW analyzed the gravitational wave channel and a variety of important auxiliary channels for the three LIGO detectors, including interferometer channels used in the feedback systems of the detectors and channels containing data from the environmental monitors.

3. Searches for gravitational waves from CBCs

Matched filtering is the optimal method of finding known signals in data with stationary Gaussian noise [6]. The searches for CBCs in S5 data used a matched filter method to compare theoretically predicted waveforms with the LIGO gravitational wave channel data [7]. Because the masses of the components of the binary determine both the duration and the frequency profile of the gravitational radiation, template banks consisting of many different waveforms are used [8].

When the signal to noise ratio ρ of a template rises above a pre-set threshold ρ_*, a trigger corresponding to the time of coalescence for the binary system is recorded [7]. Single interferometer triggers are compared for coincidence in time and component mass parameters between the different interferometers, to produce coincident triggers. These coincident triggers are further subjected to signal consistency checks, for example a χ^2 test in time/frequency bins between the template and the data [9]. In order to estimate the accidental rate of the coincident triggers, the searches perform 'time-slide' analysis, in which the data from the interferometers are offset by large time shifts compared to the light travel time between the detectors, resulting in the production of coincident triggers that can only be of an instrumental origin. Knowing the accidental rate, we can estimate the significance of unshifted coincidences. Those with the highest significance are examined as candidate events and, in the absence of a detection, used to calculate astrophysical upper limits [2, 3, 10].

Transients of non-astrophysical origin, as described in section 2, often produce triggers of large ρ, as there is significant power in these transients [11]. Even with the signal consistency checks, disturbances of non-astrophysical origin increase the false alarm rate in the time-slides, as well as producing accidental coincidences from non-astrophysical events in the unshifted data. This has the effect of reducing the significance of the loudest events which are not caused by these transients, as the rate of coincidences in the time-slides is increased, and thus the measured false alarm rate of the events in unshifted time is elevated. It therefore has the effect of 'burying' good gravitational wave candidates, as coincidences due to transient detector noises can produce significant outliers. In order to reduce these effects, we have learned to define time intervals within which triggers should not be trusted. These are called *vetoes*.
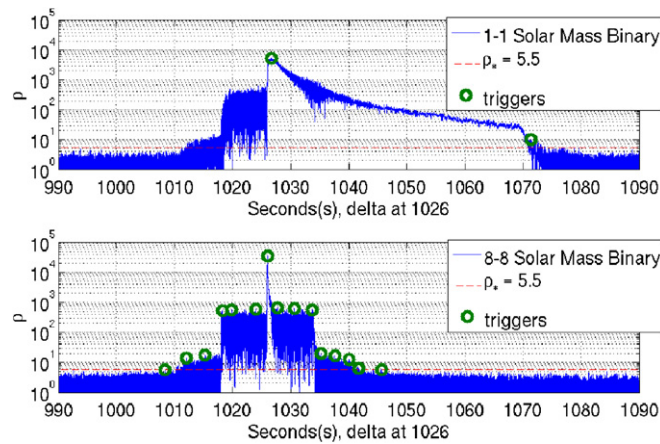
Figure 1. Matched filtering an impulse in Gaussian noise, colored based on the LIGO design sensitivity, using the CBC search with a single template. The result for two templates, top consisting of a binary of 1 solar mass objects, and bottom a binary of 8 solar mass objects. The red dashed line is the search threshold ρ_* of 5.5, and the green circles indicate the positions of the resulting triggers.

The CBC searches use banks of templates, each starting from a low frequency cutoff of 40 Hz (defined by the detector sensitivity) and increasing in amplitude and frequency until the coalescence time of the represented system. These templates have durations of up to 44 s for the lowest mass templates (binary of 1 solar mass objects).

Figure 1 depicts the filtering of an impulse in Gaussian noise, colored to match the LIGO noise spectrum, using the CBC search for a single template. We show the result for two different templates, the top panel consisting of a binary of 1 solar mass objects and the bottom panel, a binary of 8 solar mass objects. The dashed line is the search threshold ρ_* of 5.5, and the circles indicate the positions of the resulting triggers. The response of the matched filter search to loud impulsive transients in the data is complicated, with multiple simultaneous effects of the data and search code. While the ρ time series of both templates have a clear peak near the time of the impulse, the peaks do not perfectly overlap. The lower mass template has a long tail following the peak, and for both templates there is a plateau of high ρ extending 8 s before and after the time of the impulse. Additionally, while filtering with both templates results in a trigger with a large ρ value, the higher mass template results in many more triggers. While many of the triggers shown would fail the χ^2 test mentioned above, any trigger that even marginally pass this test will increase the rate of accidental coincidences in the time-slides, adversely affecting the significance calculation of unshifted coincidences.

The difference in time and ρ of the peak occurs because ρ is recorded at the coalescence time of the matching template, and frequencies in each template are weighted by the frequency-dependent noise spectrum of the detector. The time of the top panel trigger is 0.66 s after that of the bottom panel, whereas their ρ values are 5000 and 34000, respectively. For a broadband transient, the time between when the transient occurs and when the coalescence time is recorded is determined by the time remaining in the waveform after its frequency content matches the sensitive frequency band of the detector (40 Hz to 1 kHz). No transient is a perfect delta function, and there are many transient types that have different timescales and frequency content. The ρ tail visible in the top panel of figure 1 occurs due to the aforementioned 44 s template duration. When any part of the template is matched against the impulse, the ρ is significantly above threshold. Rather than recording a trigger for all times
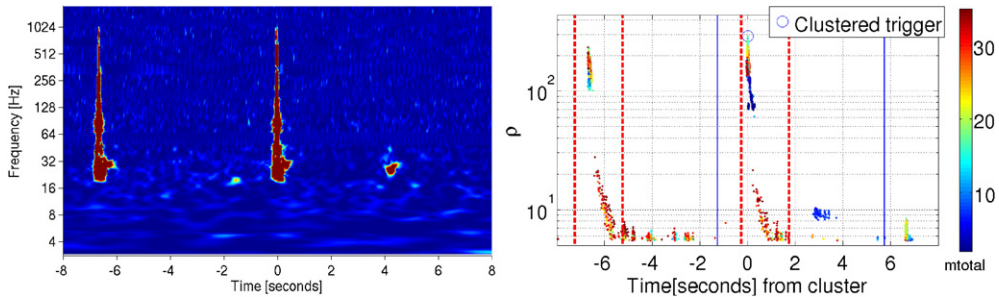
Figure 2. A pair of overflows in the length sensing and control system of H2. On the left, a time-frequency representation [5]. On the right, the effect of the transient on the production of unclustered triggers for the CBC search with the total mass between 2 and 35 solar masses (dots), the 10 s clusters of these raw triggers (circle), the original data quality flags (dashed lines) and the expanded data quality veto after duration paddings are applied (solid lines).

with $\rho \geqslant \rho_*$, a trigger is only recorded if there is no larger value of ρ within one template duration of its coalescence time. The higher mass template therefore results in more triggers because its duration is of order 3 s, while the plateau is 16 s long. The plateau is not caused by any attribute of the waveform, nor by the data adjacent the impulse, but rather is a phenomenon intrinsic to the method used to estimate the power spectral density of the data. Specifically, the data is analyzed in 256 s chunks, and a wrap-around effect of the fast-Fourier transform would cause an impulsive transient in the data to contaminate the power spectral density calculation for this entire duration; an algorithm used to truncate this effect to 16 s is detailed in sections 4.6 and 4.7 of [12]. Vetoes of impulsive transients for the CBC searches must include this time.

In cases of extraordinarily powerful transients, we observe a second trigger slightly more than one template duration from the trigger at the peak of the ρ time series, as is the case in the top panel of figure 1. Because the actual search is performed using a bank of thousands of templates of different durations, a significant number of triggers multiple seconds away from the impulse are recorded, as can be observed for the triggers with $\rho \sim 100$ to the right of the peak ρ triggers in figure 2.

For all these reasons, intervals containing transients of non-astrophysical origin often must be padded with extra duration to make them into effective vetoes for the CBC searches. This is done by examining the falloff of the triggers in ρ before and after the transients in vetoed times, in order to include those triggers associated with the transient while working to minimize the deadtime by not padding more than necessary. For the S5 searches, this was determined by examining the plots of the maximum and median amplitude transients, as measured by the peak trigger ρ. The paddings were not always symmetrical, and added from 0 to 16 s before and from 0 to 25 s after a given data quality flag. Transients of higher energy generally required more window padding. Efforts to automate this decision process are ongoing, as discussed in section 5.2.

Prior to S5, all CBC searches used a single set of veto definitions [5, 13]. During S5, the CBC searches extended over a broad range of component masses. The wide distribution in the template durations of the waveforms caused the triggers associated with transients in the data to appear at different times, and be sensitive to different frequency ranges. Thus, veto window paddings must be defined based on template waveforms included in each search, which are determined by the possible component masses of the binaries.

4. Established veto techniques

In this section we discuss techniques for using the aforementioned detector characterization work to create vetoes for the matched filter CBC searches. We illustrate our methods with examples from the CBC searches in the first year of LIGO's fifth science run, on which they were developed and implemented [2, 3]. This work was done simultaneously and in consultation with the veto efforts for the searches for unmodeled gravitational wave bursts [14]. The veto techniques discussed in this section are implemented by creating lists of times during which triggers from a search are suspected of originating not from gravitational radiation, but from instrumental or environmental disturbances.

When deciding to use sets of time intervals as vetoes, we need to include all of the bad times of the interferometers, and as little of the surrounding science mode as possible. Since not all vetoes are well understood, we need to create figures of merit, or metrics, to evaluate the effectiveness of the vetoes. These veto intervals are derived both from data quality flags (section 4.3) and disturbances in auxiliary channels (section 4.4). We then classify these vetoes into categories (section 5).

The *safety* of the veto intervals must also be ensured. A veto is unsafe if it could be triggered by a true gravitational wave. In order to ensure that our instrumental vetoes are safe, we investigate their correlation with *hardware injections*, intentional transients introduced into the interferometer in order to properly tune the various searches in LIGO for gravitational wave signals. The signals are injected directly into the gravitational wave channel itself, the differential arm length servo, to simulate the effect that a gravitational wave would have on the detector. Since these hardware injections are intentional and controlled, they exist entirely within known time intervals.

4.1. Veto metrics

Veto metrics were developed on single-interferometer triggers, from the CBC searches in the first year of LIGO's S5 science run, as well as previous runs. For this purpose, the triggers were clustered by keeping a trigger only when there are no triggers with a larger value of ρ within 10 s. With this clustering, all the triggers from a single loud transient occurred in one, or at most two, clusters. This had the effect of making the figures of merit independent of the number of waveforms in the template bank of the search. Different searches may still use different clustering times, based on the different waveform durations. A minimum ρ of 8 for the clustered triggers was chosen in order to be sensitive to the glitches that produce loud triggers from the template bank. This threshold applies to the clusters used to measure the metrics, but the resulting vetoes are applied to all triggers falling in vetoed times.

The percentage of triggers vetoed defines the *efficiency* of the veto $E = \frac{N_{vt}}{N_t} \cdot 100\%$, where N_{vt} is the number of clustered triggers vetoed and N_t is the total number of clustered triggers. If all outliers came from a single source, then the ideal veto would have 100% efficiency, especially at large ρ. In reality, there are many different sources of transient noise, as detailed in section 2, and each may be responsible for only a few percent of the clustered triggers, and only in some specific range of ρ values. The efficiency quantifies the effectiveness of the veto for removing clusters, but more information is required to determine whether this removal is warranted. Because the clustering is by loudest ρ, veto efficiency as a function of ρ can be used to learn more about what population of clusters correspond to the transient event being vetoed.

To determine the statistical significance of the efficiency, we need to compare it with the *deadtime*. The percentage of the science mode contained in a set of veto intervals defines the

deadtime $D = \frac{T_v}{T} \cdot 100\%$, where T_v is the time vetoed and T is the total science mode time, including the vetoed time. If the veto only includes truly bad times, then by vetoing triggers within these times we are not reducing our chance of detection, as the noise transients had already polluted this data. In practice, the integer second duration of data quality flags, as well as the need for padded veto windows due to the nature of the CBC search (described in section 4.2), limits how small the deadtime can be for veto intervals that remove common transient noises. It also adds to the probability that a true gravitational wave event, occurring when the detectors are in the science mode, will be missed if these times are vetoed. Most vetoes have deadtimes of at least several tenths of a percent of the science mode time, although less understood or longer duration disturbances may lead to the vetoes with larger deadtime percentages.

The ratio of the efficiency over the deadtime is a useful metric to determine which vetoes are effective. Effective vetoes have a deadtime small compared to the efficiency, indicating that many more clusters are vetoed than one would expect by random chance. This ratio is unity for ineffective vetoes, and large for effective vetoes, and can be expressed as

$$R_{ED} = \frac{E}{D} = \frac{N_{vt} \cdot T}{N_t \cdot T_v}. \tag{1}$$

The percentage of veto intervals that contain at least one clustered trigger defines the *used percentage* such that $U = \frac{N_{wt}}{N_w} \cdot 100\%$, where N_{wt} is the number of veto windows that contain at least one cluster and N_w is the total number of windows. For an ideal veto, every vetoed interval should contain at least one cluster, corresponding to one or more loud transient noises. For vetoes with short time spans compared to the clustering time, it is more common to obtain values of the used percentage of less than 100%, even for effective vetoes. The statistical significance of this metric is made by comparing the used percentage for a veto to that expected if its intervals are uncorrelated with the triggers. This expected used percentage is obtained by multiplying the length of a given veto interval T_w by the average trigger rate, given by dividing the number of triggers by the available science mode time. The ratio behaves similarly to R_{ED}, and can be expressed as

$$R_U = \frac{U}{T_w \frac{N_t}{T} \cdot 100\%} = \frac{N_{wt} \cdot T}{N_w \cdot T_w \cdot N_t} = \frac{N_{wt} \cdot T}{N_t \cdot T_v}. \tag{2}$$

In the S5 run, there was a clustered trigger on average every 6–17 min, depending on the interferometer. Effective vetoes have an expected used percentage small compared to the actual used percentage, indicating that more intervals contain clusters than one would expect by a random chance.

To evaluate the safety of each veto, the percentage of hardware injections that are vetoed is compared to the veto deadtime. If veto intervals are correlated with the hardware injections, as indicated by R_{ED} for injections significantly greater than 1, the veto could be generated by an actual gravitational wave signal. Such a veto is therefore 'unsafe' and is not used. In S5, zero data quality flags, and only one auxiliary channel, that were considered were found to be unsafe.

4.2. Vetoes from data quality flags

We apply our veto metrics and window paddings to data quality flags created by the Detector Characterization and Glitch groups within the LSC, as mentioned in section 2. A concrete example of a veto based on a data quality flag marking instrumental transients can be seen by examining the case of identified intervals containing an overflow in the length sensing and control loops for the H2 interferometer. These overflows cause severe glitches, and are
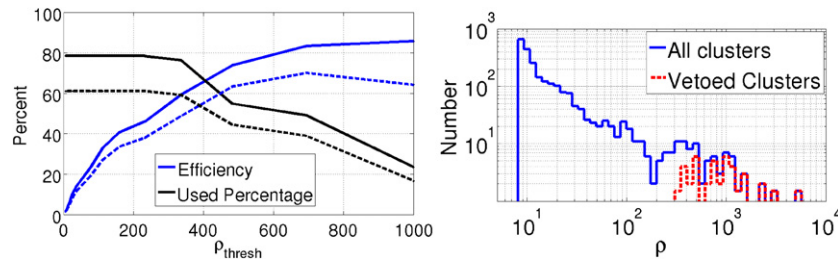
Figure 3. The left plot shows the efficiency and used percentage as a function of a minimum threshold single-interferometer cluster ρ for the H2 length sensing and control overflow veto. The dashed lines show the values for the data quality flag before window paddings are added. The solid lines show the values after windows are added to veto the associated triggers. The right plot shows a log–log histogram for the same veto. All clusters found in the science mode are shown in solid lines and vetoed clusters are shown in dashed lines.

identified within a second of their occurrence. The overflows themselves are caused by other disturbances to the control systems such as seismic motion, but irrespective of the physical origin, the overflow itself produces glitches in the gravitational wave channel. Figure 2 shows a time–frequency representation [5] of the gravitational wave channel, as well as a plot of the unclustered triggers as a function of time, for the CBC search with a total mass between 2 and 35 solar masses, around two typical transients caused by a type of overflow. The data quality flag intervals start and stop on GPS seconds, and have a minimum duration of 2 s, centered around the times of the overflows, to ensure that the glitches are not too close to the edges of the intervals. The loudest raw triggers occur near the transients, corresponding to the clustered trigger. Raw triggers subsequently fall off in ρ over the next several seconds after the data quality interval.

As shown in figure 3, this flag has a used percentage of 62% for a typical month during the first year of S5, indicating that these veto segments are well suited to vetoing triggers from transients. These data quality veto segments have an efficiency on all triggers of 1.4% and a deadtime of 0.0037%. The ratio of the efficiency to the deadtime is more than 300. This indicates that a veto with a statistically significant correlation to the triggers, as the expectation for random chance, would be a ratio of 1. The efficiency is strongly dependent on ρ of the clustered triggers; for the clusters with $\rho \geqslant 50$ it is 14%, while for the clusters with $\rho \geqslant 1000$, the efficiency is 64%. Two-thirds of the loudest triggers found from the H2 interferometer in this month were due to overflows.

To attempt to account for as many triggers as possible associated with this transient, the veto interval is given 1 s of padding prior to the data quality interval, and 4 s after. This alters the metrics, leading to a deadtime of 0.013%, an efficiency for all clusters of 1.7%, a used percentage of 78% and an efficiency to the deadtime ratio of 130. The expected used percentage from the trigger rate was only 0.58%, giving R_U of slightly over 130. For the veto intervals with duration equal or less than the clustering time, only one cluster can be vetoed per interval; thus, the number of veto windows used N_{wt} approaches the number of triggers vetoed N_{vt}. Comparing equations (1) and (2), this means we expect $R_{ED} \approx R_U$, as we see in this example with a value of 130.

For such loud transients, we expect all veto intervals to be used, but even for the loudest intervals we have tens of percent of the intervals that contain no cluster, as is the case with the aforementioned overflow flags. Out of 22% of the overflow veto intervals that are unused, 20% are within a clustering time of a clustered trigger. These intervals, therefore, may well have many raw triggers, as the overflow on the left-hand side of figure 2 did, but only the

overflow with the loudest raw trigger was marked by a clustered trigger. Future efforts to automatically classify the effectiveness of vetoes (section 5.1) would likely be sensitive to the anomalously low used percentages mentioned above, but for future searches the problem is significantly mitigated by employing a clustering window of 4 s rather than 10 s.

The efficiency versus ρ of the veto interval is shown in figure 3. This rises rapidly with the minimum ρ of the clustered triggers, and the efficiency to the deadtime ratio reaches 1300 for the clusters with ρ above 50 and over 5000 for the clusters with ρ above 500. Efficiency and the used percentage would be independent of ρ if the times vetoed were random and uncorrelated with transient noises.

An example of a data quality flag for an environmental transient that we used to make a veto was the flag marking the times of elevated seismic motion due to trains passing through Livingston. Early in S5, investigations of loud noise transients in L1 indicated that many such transients occurred in the minutes preceding loss of interferometer lock due to the passage of trains near the detector. At each LIGO site, seismometers are located in each major building. Since the trains pass closest to the end of the 'Y' arm, the seismometer located there is most sensitive to the trains. Specifically, the train-induced seismic motion was most pronounced along the direction of the arm in the 1–3 Hz frequency band. Seismic disturbances due to trains were visible upon examining the 1–3 Hz band limited root mean square (BLRMS) value for the minutes of the aforementioned seismometer channel. Setting a minimum seismic threshold in that channel of $0.75\,\mu\mathrm{m\,s^{-1}}$ in the 1–3 Hz BLRMS to identify times of passing trains, the two to three trains per day that passed the interferometer were identified. Studies of the seismic motion induced by these trains compared with single interferometer online glitch monitoring codes [4] showed correlation for up to 1 min before and after 1 min of the peak seismic amplitude for each train. Data quality flag vetoes were defined to mark these times.

For this example, our metrics then yielded a deadtime of 0.69%, an efficiency of 2% for all clusters and 20% for clusters with $\rho_{\mathrm{thresh}} \geqslant 100$ and used percentage of 60% and 32% for each threshold respectively. R_{ED} therefore increases from 3 to 30 with increasing ρ_{thresh}. R_U for the same values of ρ_{thresh} increases from 0.78 to 16. This veto is effective at eliminating a population of significant glitches, though not as loud or common as those from the overflows mentioned earlier. Techniques for combining together the information from these data quality vetoes is discussed in section 4.5.

4.3. Auxiliary channel used percentage veto

Even after taking all DQ flags into account, the number of triggers left unflagged is still very much in excess of those expected from random noise. Another approach is also possible, using the KW triggers that are generated on hundreds of important auxiliary channels.

The KW-based veto method developed in S5 is similar to other KW-based vetoes implemented in LIGO's S1, S2, S3 and S4 CBC searches [13, 15]. Comparisons were made between KW triggers in interferometer and environmental channels and the clustered CBC triggers. The KW significance can be used in the same way the ρ is used for the CBC searches.

For each channel, the threshold on KW significance was initially chosen to be above the background in the exponentially distributed triggers from Gaussian noise but low enough to catch noise transients. This threshold was then incremented until a used percentage for the veto of at least 50% was achieved. This was chosen to ensure that the times identified were likely to contain transients. Veto intervals were then generated by taking intervals ± 1 s from the KW trigger times, and rounding away from the trigger to create intervals of 3 s total duration. Only channels that achieved a used percentage of 50% were considered for veto use.

In the tuning of the veto, already diagnosed problematic times, as described in sections 5.1 through 5.3, were excluded from consideration. Once each veto was defined, a list of time intervals to be excluded was created for all S5 science mode data with the tuned parameters. The safety of each veto channel was determined using hardware injections, similar to the method used for data quality flags; those channels that produced a statistically significant overlap with hardware injection times were not implemented as vetoes. While the vetoes were defined via comparison with the clustered inspiral triggers, the vetoes required padding to ensure that the unclustered triggers associated with each transient were also vetoed. For the channels that triggered coincident with large amplitude transients in the gravitational wave channel, we added 7 s veto window paddings to the beginning due to the plateau effect (section 3) and end of the initial 3 s intervals. For each CBC search these KW-based used percentage vetoes were tuned and defined with respect to the single interferometer triggers from the specific analysis.

Only one channel considered was found to be unsafe. The LIGO interferometers readout differential arm motion at the anti-symmetric port using a heterodyne scheme [1]. In addition to the differential arm length signal, there is an orthogonal modulation which should in theory be insensitive to gravitational wave yet still be sensitive to alignment fluctuations. In practice, this channel was found to have a small coupling to differential arm motion, and thus for large amplitude transients in the gravitational wave channel, this channel was unsafe. This channel therefore was not used to make CBC vetoes in S5.

In the first year of S5 there were a number of critical veto channels found in this manner. An important veto associated with an interferometer control channel was the feedback loop that keeps the H1 recycling cavity resonant. A veto based on environmental monitors came from the magnetometers located at the end of the Y-arm for L1.

5. Veto categorization

The goal of using vetoes is to reduce the false alarm rate, in order to more accurately assess whether gravitational wave candidates are true detections. Upon evaluating the available vetoes, we found that they do not all perform similarly, and divided them into categories. Well-understood vetoes have a low probability of accidentally vetoing gravitational waves, and significantly reduce the background. More poorly understood vetoes can also reduce the background significantly, but with an increased chance of falsely dismissing actual gravitational waves. We classified the vetoes into categories in order to allow searches to choose between using only the well-established vetoes or aggressively using more poorly understood vetoes.

Those vetoes classified as well-understood almost always had higher R_{ED} and R_U, as well as a lower overall deadtime than the less-understood transients that correspondingly had less effective, longer intervals with poorer ratios. This was because when the mechanism behind a transient was well understood, such as an overflow in the digital control channels of the interferometer, it was easier to identify the specific times at which these transients occurred. In some cases, however, well-understood vetoes may include little enough time that statistics are difficult to perform due to the small number involved, and these can still be categorized as well understood, providing sufficient evidence that coupling is present. Conversely, when only the general cause was known, as in the case of transients related to passing trains at the Livingston site, time intervals when this elevated ground motion was present needed to be vetoed in order to capture the related transients. Despite the short duration of each particular transient, these intervals ranged from several to tens of minutes. In this latter case, a statistical argument based on the veto metrics was required to prove the utility of a set of veto intervals.

The idea of this categorization scheme was to allow the followup of candidate triggers after applying sequentially each category of vetoes with the consequently lowered background false alarm rate, in order to search for detections [2, 3]. In the CBC searches in LIGO's S5 science run, we decided on four categories in descending order of understanding of the problems involved.

5.1. Category 1

The first category includes vetoed times when the detectors were not taking data in the design configuration. A fundamental list of science mode times is compiled for each interferometer, and only the data in these times is analyzed. These times are logged automatically by the detectors with high reliability, though on rare occasions DQ flags marking non-science mode data mistakenly marked as science quality need to be generated after the fact. These are the same for all searches, and do not need to be padded with extra windows, as the data are not analyzed.
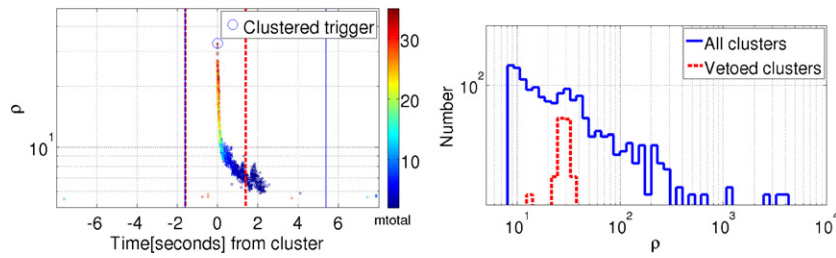
5.2. Category 2

The second category contains well-understood vetoes with well-tuned time intervals, low deadtimes and a firm model for the coupling into the gravitational wave channel. For many transients, this results in a high efficiency, particularly at high ρ, though this is not necessarily the determining factor in categorization. A well-understood noise coupling into the gravitational wave channel may consistently produce triggers of moderate amplitude, or at a lower rate than more common transients. These are still considered to be of category 2 if several conditions are met. The ratios R_{ED} and R_U should statistically be significant, of order 10 or higher, for all clusters above some ρ threshold characteristic of the transients.

One example of a category 2 veto is the overflow veto mentioned in section 4.3. As is clear from the right-hand plot in figure 3, these veto intervals include most of the loudest clusters. Not all category 2 vetoes need to have this level of efficiency, or any efficiency at all at the most extreme ρ. For instance, we also used vetoes based on data quality flags for glitches in the lasers for the thermal compensation system (TCS). The TCS heats the mirrors in order to offset changes in curvature due to heating by the main laser. These flags, with 4 s added to the end of the original 2 s intervals, had a R_{ED} ratio of nearly 500 for triggers above ρ of 20, but zero efficiency above ρ of 40. R_U for this veto was over 100. As is clear in figure 4, there is a population of clusters with ρ from 20 to 40 that correspond to the transients from TCS glitches in this particular search.

5.3. Category 3

The third category contains vetoes which were significantly correlated with transients, but with less understanding of the exact coupling mechanism, and thus often poorer performance in the metrics of deadtime and used percentage than category 2 vetoes. There are many sources of transient noises whose coupling is only partly understood. Site-wide events of significant duration, such as heavy winds or elevated seismic motion, intermittently lead to loud transient noises. The auxiliary channel vetoes discussed in section 4.3, because of their statistical nature, also fall into this category. Category 3 vetoes also include the minutes immediately preceding the loss of lock of the interferometer, when the triggers were likely due to the same instabilities that contributed to the lock loss. These vetoes, based more on the probability of transients than a direct measurement, tend to have lower used percentages, higher deadtimes and therefore smaller ratios between the efficiency and deadtime.

Figure 4. A category 2 data quality veto for a month containing glitches in the TCS lasers of H2. On the right, a log–log histogram of single interferometer clusters for the CBC search with a total mass between 2 and 35 solar masses, with all clusters in blue and vetoed clusters in red. On the left, the effect of the transient on the production of unclustered triggers (dots), the 10 s clusters of these raw triggers (circle), the original data quality flag (dashed lines) and the expanded data quality veto after the window paddings are applied (solid lines).
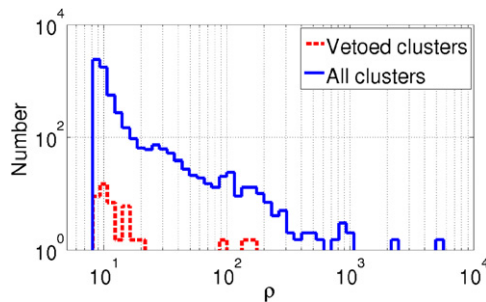
Figure 5. A category 3 data quality veto for a month of high winds at Hanford. A log–log histogram of single interferometer clusters, total in solid lines and vetoed in dashed lines.

The train data quality flag veto mentioned in section 4.2 was in this category: while the trains themselves were well understood, the nonlinear coupling to create sporadic high frequency glitches was not. This caused large windows defined by the presence of heightened ground motion alone to be created, rather than the targeted vetoes of the individual noise transients.

Another example was the elevated winds above 30 mph at the Hanford site. This data quality veto had R_{ED} of 17 and R_U of 31, and as illustrated by the histogram in figure 5, vetoed a small population of triggers with ρ of order 100. While this is significant, it is less than the typical value of category 2 vetoes.

By definition, the auxiliary channel vetoes have a large used percentage, always greater than 50%. However, they are classified as category 3 because the coupling is not well understood. An example of such a veto made using the above technique based on an interferometer control channel is the veto made from the H1 feedback loop that keeps the recycling cavity resonant. This veto had R_{ED} of 20. Another veto, this time based on an environmental channel, used the magnetometers located at the end of the L1 Y-arm and had R_{ED} of 10.

5.4. Category 4

The fourth category contains vetoes with low statistical significance, often with high deadtimes. The used percentages are often near 100%, but this is a representative of the long intervals of science mode time flagged, and thus the high probability that at least one cluster will be within the time defined (as mentioned earlier, the average clustered trigger rate was of order 1 per 10 min). Seismic flags with lower thresholds, aircraft passing within miles of the detectors and problems recorded in the electronic logbooks at the detectors all fall within this category. These long intervals are not used as vetoes for searches, but rather are identified for the purposes of providing input to the follow up of gravitational wave candidates, when all possible factors that prompted the creation of these flags at the time of the detection candidate are scrutinized.

5.5. Examining candidate events after vetoing times

As mentioned earlier, candidate coincident events that occur during the times of categories 2, 3 and 4 vetoes are not automatically discarded. The total deadtime of category 2 vetoes for the low mass CBC search, for example, was of order 1%. Category 3 had a deadtime of order 5%, and category 4 many times that. As the search reported no detections, it therefore included a calculation of the upper limit on the number of CBCs. The first three categories, including both data quality flag vetoes and auxiliary channel used percentage vetoes, were applied in this example search before calculating the upper limit, as they reduced the false alarm rate from these transients. Because the total veto deadtime of the applied categories was between 5% and 10% per interferometer, the probability that a true gravitational wave could be in a vetoed interval is significant.

   The decisions on which vetoes to use are made prior to the examination of candidates. While the veto choices were tuned on single interferometer triggers, the end product of the CBC searches are detection candidates found in coincidence in the data from multiple interferometers. In the rest of this section, we will discuss the effect that the veto categories have on the coincident CBC searches.

   While we are unwilling to precipitously remove all candidates in vetoed times from consideration, it is imperative to reduce the rate of accidental coincidences from the noise transients that have been identified. This can be done by examining all significant candidates with respect to the background present after each category is consecutively applied. If a candidate is not vetoed by successive veto categories, it becomes more significant, as more of the background false alarm candidates against which it is compared are removed. If a candidate is in vetoed time for a given category, it is not completely ruled out as a detection, but further investigation would be required to show such a candidate is not an artifact due to the disturbance that triggered the veto. Candidates that are vetoed by lower numbered categories are more suspect, given the firmer understanding of the vetoes that populate the first two categories. A true gravitational wave is not impossible, as a sufficiently nearby (within the Milky Way) binary system coalescence could theoretically overflow the feedback control systems. It would be apparent, however, in follow up investigations as the spectrogram for the data would show a large amplitude chirp signature leading up to the overflow, coherently between multiple detectors.

   The diagram in figure 6 illustrates the effect of vetoes that we anticipate on the significance of detection candidates. Superimposed are four hypothetical candidates with the labels A, B, C and D. For illustration only, let us assume that there is a gravitational wave candidate at one of these points.

14
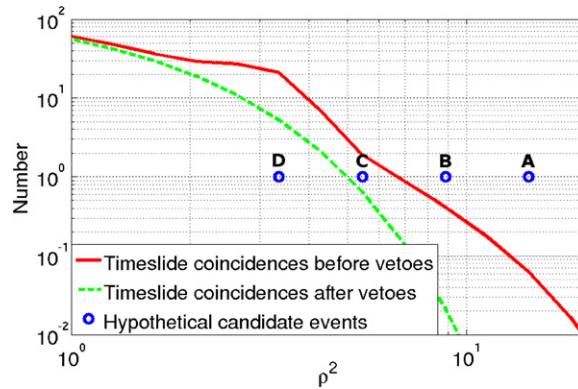
Figure 6. This diagram schematically represents the anticipated effect of vetoes on the significance of candidate events. The lines represent the estimated background number of coincident events expected from time-slides before (solid) and after (dashed) vetoes. The circles denote the number of foreground coincidences with ρ^2 equal to or greater than the *x* axis value. For the purpose of the discussion in section 5.5, the points A, B, C and D denote the hypothetical detection candidates.

If the candidate is at point A, it is visible above the background and significant before any vetoes are applied. If A is not vetoed after subsequent veto categories are applied, it will be the loudest candidate and significantly above the rest of the distribution, and thus, a strong gravitational wave candidate. If A is vetoed, it would be plausible to believe it could still be significant with strong evidence that it did not originate from the same problems used to define the veto, as in the hypothetical example of the galactic coalescence mentioned above. Even if A is recovered, it would be compared to the background estimation before vetoes are applied, and thus have a lower significance than had it survived the vetoes originally.

If the candidate is at point B, it is visible above the background, though not as significant as if it were at point A. If it is not vetoed, then B is a good candidate, and having cleared away the understood accidental triggers from transient noises, it can be followed up in depth. The reason for defining vetoes is precisely to uncover these candidates, which would be buried in the background otherwise. If it is vetoed, it is again necessary to confirm that the data artifacts prompting the veto intervals are not responsible for the candidate.

If the candidate is at point C, it is likely among the triggers with the largest ρ^2 before the vetoes are applied. Surviving the vetoes improves its ranking, reducing the background of triggers with equal or lower ρ^2. If it is vetoed, it is a problematic candidate, given the population of spurious triggers that it sits in. If it survives the vetoes, it is still only a marginal candidate. Follow up analysis of the highest ρ^2 triggers will likely uncover reasons to distrust the surrounding loud candidates, but that is not enough to make C a strong candidate. Additional veto definitions and revisions would make a candidate at C somewhat more significant, providing it is not vetoed, plausibly making it significant.

A candidate at point D is within the accidental population of triggers after vetoes are applied, with tens of triggers surrounding it with similar ρ^2. Such candidates are not detectable without additional reduction of the background. Additional veto definitions and revisions might make a candidate at D marginally more significant, but it is not at all likely to become detectable through veto efforts.

5.6. Results of veto efforts in S5

In S5, hundreds of data quality flags and auxiliary channels were evaluated as mentioned above in sections 4 and 5. The resulting veto metrics and categorization for each of the vetoes used in the CBC searches were archived in a technical document [16]. The data quality flags used by the unmodeled burst search group were selected via the same metrics, using KW and other burst search techniques instead of single interferometer CBC triggers, and almost always resulted in the same flags in the same categories as the CBC vetoes, though no paddings were added to the flags.

6. Proposed veto techniques

The techniques mentioned in section 4 were by and large refined months after the data were recorded. Additionally, the decisions on categorization, veto window padding and utility of auxiliary channels for used percentage vetoes were determined largely by individual human examination of the behavior of each data quality flag and auxiliary channel. While this was necessary for the development and early implementation, the low latency and rigor associated with automating as much of this decision making process as possible is desirable. Below are the discussions of our current and near future efforts to realize the goal of automated evaluation, categorization and extension of data quality and auxiliary channel vetoes.

6.1. Automated categorization using a χ^2 test

One promising method that has been developed, and that has the potential to help automate recommendations for veto categories, is a figure of merit based on a χ^2 test. For each DQ flag this χ^2 statistic is given by

$$\chi^2(\rho) = \sum_{k=1}^{R} \frac{(n_k(\rho) - T_k \langle n_t(\rho) \rangle)^2}{T_k \langle n_t(\rho) \rangle}, \tag{3}$$

where $\langle n_t(\rho) \rangle$ is the average number of triggers per unit time in the science run above a certain threshold ρ, T_k is the duration of the flagged window k, R is the number of flag windows and n_k is the actual number of triggers above the threshold ρ in the same window.

The null hypothesis is that the triggers are Poisson distributed, i.e. there is no correlation between the presence of triggers and the DQ flags. In our analysis we compute the figure of merit and test the null hypothesis at a confidence level of 95%. The higher the figure of merit, the higher the correlation between triggers and DQs and thus the lower the category.

As an example of the χ^2 categorization scheme, the χ^2 value for the H2 overflow mentioned previously is $\approx$100 times higher than the typical ranges of category 3 vetoes and $\approx$1000 times higher than the ranges typical of category 4 vetoes. For vetoes that have χ^2 values that are near or on the boundaries between the categories we turn to figure of merits mentioned previously such as deadtime, used percentage and efficiency to determine which category a particular veto belongs in. For example, the veto for glitches in the TCS lasers of H2 has a chi-squared value that falls near the lower range for category 2 and the higher range for category 3. However, the high used percentage and the low deadtime distinguish this veto from category 3 vetoes with similar χ^2 values.

The χ^2 method is a step towards automated categorization. An automated monitor that incorporates previously mentioned figures of merit and χ^2 values to organize vetoes into categories is currently being developed.

6.2. Automated veto window padding determination

The determination of window paddings (see section 4.2) is another step of the veto selection pipeline that currently requires human input. A method to automate this step would be to look for quiet time intervals of pre-determined duration around the clusters. If the unclustered triggers remain below the minimum ρ threshold for the chosen duration before and after the glitch's loudest trigger, the earliest and latest triggers above the threshold would determine the left and right padding of the DQ window, respectively. Assuming the duration of the required padding for the DQ flag to be normally distributed, a final recommendation for the padding of that flag would be obtained by taking the average of the values for each window.

6.3. Multiple auxiliary channel veto algorithm

Another approach that has been explored is that of defining vetoes when multiple auxiliary channels glitch coincidently, specifically by examining the output of the 'QScan' time-frequency algorithm [5] over multiple auxiliary channels, at the time of detection candidates. When a number of auxiliary channels glitch simultaneously and at the same time as the gravity wave channel, there is a strong possibility that the glitching has a non-astrophysical cause, particularly when the channels are physically related. For example, a number of the length sensing and control channels may glitch together in different parts of the interferometer, such as the beam splitter and reflected and dark ports. Sometimes the glitches in a set of length sensing and control channels will be associated with glitches in the alignment sensing and control channels. Many of these measure pitch and yaw of mirrors, including the test masses, and when mirror alignment and length disturbances occur simultaneously, it is unlikely that a transient also in the gravity wave channel will be astrophysical. Nevertheless, safety studies have successfully been conducted, using all of the hardware injections from the first year of S5, to verify that these combinations of glitches could not be caused by the arrival of gravitational waves.

Tests of the efficacy of proposed vetoes were carried out on data from the S5 search with a total mass between 25 and 100 solar masses. This method is a promising veto for the coincident triggers with the largest ρ values. Of the coincident triggers remaining after veto categories 1 through 4, all with component single interferometer $\rho \geqslant 200$ would be vetoed. Since these vetoes are run-on small time intervals around the times of the detection candidate triggers, the deadtime is not calculated: a candidate survives or is vetoed.

Future work plan includes running this algorithm over the times of candidates using a larger set of the available channels, rather than only channels corresponding the length and alignment sensing and control, as restricted in the above studies to reduce the computation time. It is expected that additional sets of channels will be useful, particularly the environmental channels.

7. Summary

In this paper, we showed how we developed techniques for vetoing non-astrophysical transient noises safely and effectively, in order to reduce the effect of noise transients on the astrophysical

searches for low mass CBCs in the first year of LIGO's S5 run. We based our vetoes on data quality flags created by detector characterization work, as well as the KW triggers from auxiliary channels with high used percentages. Though we approached each flag and each channel individually, and though different flags and different channels reflected a variety of specific causes, we found that the effects on the gravitational wave channel fell into a few common groupings. Flags and channels that responded to similar phenomena generally required similar windows, had similar deadtimes, were effective for similar populations of triggers, and therefore were placed in the same categories. The LSC used these categories for sequentially studying the significance of gravitational wave candidates rising above the background in the CBC searches.

Going forward, we intend to use the experience gained to finish ongoing automation work both to select veto window paddings and to provide recommendations for veto categorization. There will be data quality flag and auxiliary channel based vetoes developed for the use in LIGO's S6 science run. The goal will be to analyze the auxiliary channel KW triggers in near real time, and to have vetoes defined on a week-by-week basis for both types of vetoes. It is probable that there will be marginal cases for categorization that require further human review, but automation will allow us to focus our time on these cases.

Acknowledgments

References

[1]   Abbott B *et al* (LIGO Scientific Collaboration) 2009 *Rep. Prog. Phys.* **72** 076901 (arXiv:0711.3041)
[2]   Abbott B *et al* (LIGO Scientific Collaboration) 2009 *Phys. Rev.* D **79** 122001 (arXiv:0901.0302)
[3]   Abbott B *et al* (LIGO Scientific Collaboration) 2009 *Phys. Rev.* D **80** 047101 (arXiv:0905.3710)
[4]   Blackburn L *et al* 2008 *Class. Quantum Grav.* **25** 184004 (arXiv:0804.0800)
[5]   Chatterji S, Blackburn L, Martin G and Katsavounidis E 2004 *Class. Quantum Grav.* **21** S1809–18 (arXiv:gr-qc/0412119)
[6]   Wainstein L A and Zubakov V D 1962 *Extraction of Signals From Noise* (Englewood Cliffs, NJ: Prentice-Hall)
[7]   Allen B, Anderson W G, Brady P R, Brown D A and Creighton J D E 2005 FINDCHIRP: an algorithm for detection of gravitational waves from inspiraling compact binaries arXiv:gr-qc/0509116
[8]   Owen B J and Sathyaprakash B S 1999 *Phys. Rev.* D **60** 022002
[9]   Allen B 2005 *Phys. Rev.* D **71** 062001
[10] Abbott B *et al* (LIGO Scientific Collaboration) 2008 *Phys. Rev.* D **77** 062002 (arXiv:0704.3368)
[11] Abbott B *et al* (LIGO Scientific Collaboration) 2007 Tuning matched filter searches for compact binary coalescence *Technical Report LIGO-T070109-01* http://www.ligo.caltech.edu/docs/T/T070109-01.pdf
[12] Brown D A 2005 Search for gravitational radiation from black hole MACHOs in the Galactic halo arXiv:0705.1514
[13] Christensen N *et al* (LIGO Scientific Collaboration) 2005 *Class. Quantum Grav.* **22** S1059–68
[14] Abbott B P *et al* (LIGO Scientific) 2009 *Phys. Rev.* D **80** 102001 (arXiv:0905.0020)
[15] Christensen N, Shawhan P and González G 2004 *Class. Quantum Grav.* **21** S1747–55
[16] Abbott B *et al* (LIGO Scientific Collaboration) 2010 Data quality and veto choices of S5 lowmass CBC searches *Technical Report LIGO-T1000056-v2* https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=8982