Geophysical Journal International

Geophys. J. Int. (2018) **213,** 1984–1998
Advance Access publication 2018 March 16
GJI Seismology

doi: 10.1093/gji/ggy100

Detecting earthquakes over a seismic network using single-station similarity measures

Karianne J. Bergen[1] and Gregory C. Beroza[2]

[1]*Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA. E-mail: kbergen@stanford.edu*
[2]*Department of Geophysics, Stanford University, Stanford, CA 94305, USA*

Accepted 2018 March 15. Received 2018 January 29; in original form 2017 October 27

SUMMARY

New blind waveform-similarity-based detection methods, such as Fingerprint and Similarity Thresholding (FAST), have shown promise for detecting weak signals in long-duration, continuous waveform data. While blind detectors are capable of identifying similar or repeating waveforms without templates, they can also be susceptible to false detections due to local correlated noise. In this work, we present a set of three new methods that allow us to extend single-station similarity-based detection over a seismic network; event-pair extraction, pairwise pseudo-association, and event resolution complete a post-processing pipeline that combines single-station similarity measures (e.g. FAST sparse similarity matrix) from each station in a network into a list of candidate events. The core technique, pairwise pseudo-association, leverages the pairwise structure of event detections in its network detection model, which allows it to identify events observed at multiple stations in the network without modeling the expected moveout. Though our approach is general, we apply it to extend FAST over a sparse seismic network. We demonstrate that our network-based extension of FAST is both sensitive and maintains a low false detection rate. As a test case, we apply our approach to 2 weeks of continuous waveform data from five stations during the foreshock sequence prior to the 2014 M_w 8.2 Iquique earthquake. Our method identifies nearly five times as many events as the local seismicity catalogue (including 95 per cent of the catalogue events), and less than 1 per cent of these candidate events are false detections.

Key words: Time-series analysis; Self-organization; Computational seismology; Earthquake monitoring and test-ban treaty verification.

1 INTRODUCTION

Waveform similarity search has emerged as a powerful technique for detecting weak seismic events in continuous waveform data. Most similarity-based search methods in seismology are informed, in that they use known waveform signatures (Gibbons & Ringdal 2006; Peng & Zhao 2009) or waveform characteristics (Harris 2006; Barrett & Beroza 2014) from previously identified earthquakes to detect new events. A subset of similarity-based search methods are uninformed, or blind, such that they use similarity alone to detect new events without prior knowledge of specific waveform signatures or characteristics (Brown *et al.* 2008; Yoon *et al.* 2015; Skoumal *et al.* 2016). The capability for uninformed search is important in the common situation where expected signal characteristics are incompletely known. Fingerprint and Similarity Thresholding (FAST) (Yoon *et al.* 2015) detection has been applied to a single channel of continuous data; however, because similarity should extend over a network, we can improve detection sensitivity and reduce false alarms by developing a multistation version of FAST.

In this paper, we present an approach that combines the output of single-channel similarity-based detection over a network. We apply our new three-step network detection pipeline to the output from FAST, but the approach is general. Because our method identifies events observed over a network, which is referred to as 'association' when applied to phase arrivals from typical energy-based earthquake detectors (Mykkeltveit & Bungum 1984; Ringdal & Kværna 1989; Johnson *et al.* 1997), we refer to the core network detection step as pairwise pseudo-association. The distinguishing feature of our approach is that it does not exploit information on the moveout related to the expected arrival times of seismic waves, but instead leverages pairwise detections to associate the detected events.

Energy-based methods such as the short-term average/long-term average (STA/LTA) algorithm (Allen 1982; Withers *et al.* 1998) are efficient, widely used real-time detectors, but suffer from limitations including difficulty detecting low signal-to-noise waveforms or non-impulsive arrivals. As a result, sensitive detectors based

on waveform similarity have become popular tools for identifying events that are missing in earthquake catalogues.

Waveform cross-correlation, also known as template matching or matched filtering, is a highly sensitive detector based on waveform similarity that has been successfully applied to a range of detection problems (Shelly *et al.* 2007; Peng & Zhao 2009; Skoumal *et al.* 2014). Template matching requires precise information about the signals we wish to detect, in the form of a library of template waveforms; this makes it highly sensitive but also limited to detecting events with known sources and in areas with an existing seismicity catalogue. Subspace detectors (Harris 2006; Barrett & Beroza 2014), a related class of generalized correlation-detectors, can identify non-repeating sources, but these still require prior knowledge of waveform signatures.

Not all waveform-similarity-based detectors require prior knowledge of waveform signatures. Autocorrelation (Brown *et al.* 2008), a template-free variant on waveform cross-correlation, identifies candidate earthquake signals using a blind, brute-force search for similar waveforms in continuous data. While autocorrelation can detect events with previously unknown waveform signatures, it has poor scaling properties that limit its use to short-duration data sets (Aguiar *et al.* 2017).

Recent methods such as FAST (Yoon *et al.* 2015) and Repeating Signal Detector (RSD) (Skoumal *et al.* 2016) address some of the limitations of autocorrelation for blind, similarity-based detection. In particular, FAST leverages locality-sensitive hashing (Andoni & Indyk 2006) to perform a computationally efficient similarity search to detect similar earthquake waveforms. In contrast with template-matching, FAST does not require template waveforms for detection. FAST is more computationally efficient than autocorrelation and uses time-frequency features for improved detection performance. Previously published studies (Yoon *et al.* 2015; Bergen *et al.* 2016; Yoon *et al.* 2017) have applied FAST to detect similar earthquake waveforms in single or three-component continuous seismic data at a single station. One challenge of blind, similarity-based detectors is their possible susceptibility to false detections, especially at the single-station level, as persistent local noise sources can also produce repeating or similar waveforms. Thus, to maintain high precision (i.e. a low false detection rate), we would like to extend FAST for multistation detection.

In the case of template matching, simultaneous detection over multiple stations enables the identification of very weak earthquake signals (Peng & Zhao 2009); summing cross-correlation values over the network makes it easier to separate the signal from noise without incurring significant false detections. It is relatively straightforward to extend template matching to multiple channels and stations. Since multistation templates are created using data from known events, both the waveform signatures and the relative arrival times at each station in the network are already known, and no phase association is necessary. Even recent variations on template matching (Zhang & Wen 2015) that allow for some uncertainty in relative arrival times still require prior information about the approximate source location and arrival times. In contrast, our task of extending blind similarity search for multiple station detection is more challenging because the moveout of wave arrivals across the network is unknown and determining the relative arrival times at each station becomes part of the search problem.

In developing a blind, waveform-similarity-based detector that can operate over a seismic network, we consider two possible directions. The first direction is network-based detection on the input side, using multistation features as inputs to the detector. The alternative, which we pursue in this work, is to apply single-station detection separately for each station and perform network-based post-processing on the output side.

Network detection via multistation features is an attractive approach that would mirror the use of multistation template waveforms in template matching. This approach is expected to produce few false positives (prior to any post-processing steps aimed at further reduction), and is attractive from a computational standpoint as it would not require running FAST separately for each channel and station. However, this approach is not robust to station dropout or changes in network architecture, which are likely to occur in any real seismic network over an extended duration. Furthermore, in a blind search without prior knowledge of the expected moveout over the network, it is difficult to create effective multistation features, especially in sparse networks where the moveout across stations is large. It is also challenging to select the optimal subset of stations for multistation features, and poor selection is likely to hurt performance.

Instead, we prefer a multistation detection approach that combines single-station detection results in post-processing. Although it requires more computational resources to run a detector separately for each station, we chose to implement network detection as a post-processing routine to ensure the method will be robust and flexible enough for practical use with real-world data for a variety of detection tasks. Our approach was originally developed for multistation detection with FAST with non-trivial moveout, but it can be used with any single-station detector that returns pairwise detections.

In this work, we propose a method for network detection that uses pairwise similarity measures for events detected in single channel or station data, and combines these measures over a seismic network. The effectiveness of our approach comes from the pairwise treatment of detections, which enables identification of a given pair of events (where both events in the pair share a similar source) over the network without having to solve for consistency with a source. Our approach has three key strengths: (1) it automatically handles the unknown moveout across the network, (2) it is robust to missing or low-quality data at one or more stations, and (3) it is flexible enough to be adapted for a variety of detection tasks and network geometries.

Our proposed method involves three steps. The initial processing step, *event-pair extraction*, extracts the essential data from the raw FAST output, and is applied separately to the output from each station. This is followed by the core network detection step, referred to as *pairwise pseudo-association*, which identifies subsets of arrivals across the network that can be attributed to the same source event. Both event-pair extraction and pairwise pseudo-association operate on pairwise detections, so we require a final step, *event resolution*, to produce the final list of individual events. We present the method in the context of extending detection with FAST over a seismic network, and demonstrate the success of this approach in identifying uncatalogued events over a sparse network using the 2014 Iquique foreshock sequence as a test case.

2 FAST EARTHQUAKE DETECTOR

2.1 Single-channel detection with FAST

In this work, we present our method for network detection with single-station blind, waveform-similarity-based detectors in the context of FAST, first introduced in Yoon *et al.* (2015). FAST detects earthquakes by performing a blind search for similar waveforms in

a single-channel of continuous data. FAST does not require any templates or labelled examples of earthquake waveforms as inputs, instead it treats earthquake detection as a pattern-mining problem. FAST extracts a set of features, called waveform fingerprints, for each short-duration interval in the continuous data, then searches the collection of waveform fingerprints to identify similar or repeating patterns. Rather than using a brute-force search, as in the autocorrelation method (Brown *et al.* 2008), FAST searches for similar waveform fingerprints using an efficient indexing and search procedure. The core of FAST involves two key steps: fingerprint extraction and computationally efficient similarity search.

Fingerprint extraction is the process by which each short-duration interval of waveform data is converted into a set of binary features called a *waveform fingerprint* (Bergen *et al.* 2016). Waveform fingerprints are used as proxies for earthquake waveforms in the similarity search step, so these fingerprints must be discriminative; similar waveforms should produce similar fingerprints, and fingerprints corresponding to noise should have low similarity to each other. FAST uses a data-adaptive variant of the feature extraction used in the Waveprint (Baluja & Covell 2008) audio search and retrieval algorithm.

FAST similarity search step involves two stages: the first is indexing a set of database fingerprints, and the second is querying the index to identify similar waveform fingerprints. Because FAST assumes no prior information about waveform signatures, we set up the similarity search to identify any repeating or similar signals among the full set of waveform fingerprints. Specifically, we store the full set of fingerprints in the search index and we use the full set of fingerprints as queries against the index, in a true blind (all-to-all) search. The index is designed such that for each query fingerprint, FAST can identify similar fingerprints without having to scan the entire index. Rather than an exhaustive search, FAST performs an approximate similarity search (Andoni & Indyk 2006), designed to identify similar waveforms with high probability, which enables improved scalability and reduced runtime.

The output of the FAST similarity search step is a sparse similarity matrix, $\mathcal{M}$, with non-zero values representing pairs of fingerprints identified as having relatively high similarity. Each row and column of the matrix corresponds to a particular waveform fingerprint, and the matrix values are a similarity measure between fingerprints, such that $\mathcal{M}[i, j] = m_{ij}$ gives the similarity of fingerprints i and j. The output is returned in the form of a list of triplets corresponding to the non-zero values in the matrix. Each triplet contains the integer index of each fingerprint in the pair and their similarity value: (i, j, m_{ij}), with $m_{ij} > 0$ and $i < j$. In this work, we consider this sparse similarity matrix to be the final output of single-channel FAST.

A detailed review of single-channel FAST detection (Yoon *et al.* 2015) can be found in Supporting Information Section S1, including a new variation called all-to-some similarity search (Supporting Information Sections S1.4 and S4). A schematic diagram illustrating the key steps of single-channel FAST is shown in Fig. 1.

2.2 Extending FAST for multistation detection

Our goal in this work is to develop a method to combine the detection results from single-station, blind similarity-based detectors, like FAST, across a network of stations for improved detection of weak earthquake signals with limited false detections. Though the strategy for combining single-station detections over a network will depend in part on the station density and expected moveout, we focus on the detection of regional events in a sparse network. The sparse network case forces us to address the problem of unknown moveout across the network, and the methods for this case can be adapted to a range of network geometries. Approaches for handling the simpler case of limited moveout over the network, which has been partially addressed in previous work (Yoon *et al.* 2014), and the limitations of extending the same approaches to the sparse network case are discussed in Supporting Information Sections S2 and S6, respectively.

In many seismic networks, station spacing is tens or even hundreds of kilometres. In these cases, the expected moveout over the network will be too large to assume significant temporal overlap of signals recorded across multiple stations. For blind waveform-similarity-based detectors like FAST, where neither waveform signatures nor the relative arrival times are known in advance, we need a more general approach for network detection that can account for substantial, unknown moveout. In order to identify FAST detections across multiple stations, we need to perform a task similar to association: we need to determine which detections observed at different stations can be attributed to the same event. This will allow us to eliminate false detections due to local, persistent noise sources.

In the next section, we introduce a new three-step post-processing pipeline that converts the sparse similarity matrices from single-station FAST to a final list of events observed across multiple stations: (1) event-pair extraction, (2) pairwise pseudo-association, and (3) event resolution.

First, event-pair extraction identifies pairs of similar events in the sparse similarity matrix from each individual station. The output of event-pair extraction, a list of pairwise event detections at each station, is then fed into the core network detection technique, *pairwise pseudo-association*. In the context of energy detectors, *phase association* is the process of taking a collection of phase arrivals from a network of seismic stations and identifying subsets of those arrivals that are consistent with a seismic event (Mykkeltveit & Bungum 1984; Ringdal & Kværna 1989; Johnson *et al.* 1997). Pairwise pseudo-association performs a task similar to phase association, identifying sets of FAST detections from multiple stations that can be attributed to the same event, but pairwise pseudo-association operates on pairs of detections, rather than individual phases. In contrast with phase association, the network detection model used in pairwise pseudo-association does not require that arrivals are consistent with a source, but rather that the moveout for a pair of events observed across the network is consistent with both events sharing a similar source. Through this stage in the pipeline, all detections (fingerprints or events) are treated as pairs of similar detections; event resolution, the final step of the network detection pipeline, converts the results into a list of individual (unpaired) candidate events. A single event may belong to multiple pairs of similar events, so event resolution must identify and merge multiple detections corresponding to the same event. The entire multistation detection pipeline is illustrated in Fig. 2.

3 MULTISTATION DETECTION PIPELINE

Multistation detection with FAST involves three post-processing steps: event-pair extraction, pairwise pseudo-association, and event resolution. In the description of these methods that follows, we will use the term *fingerprint-pair* to refer to a set of two binary waveform fingerprints, each corresponding to a waveform of fixed time duration, that FAST identifies as similar to each other. We will use *event-pair* to refer to a set of two similar events, each

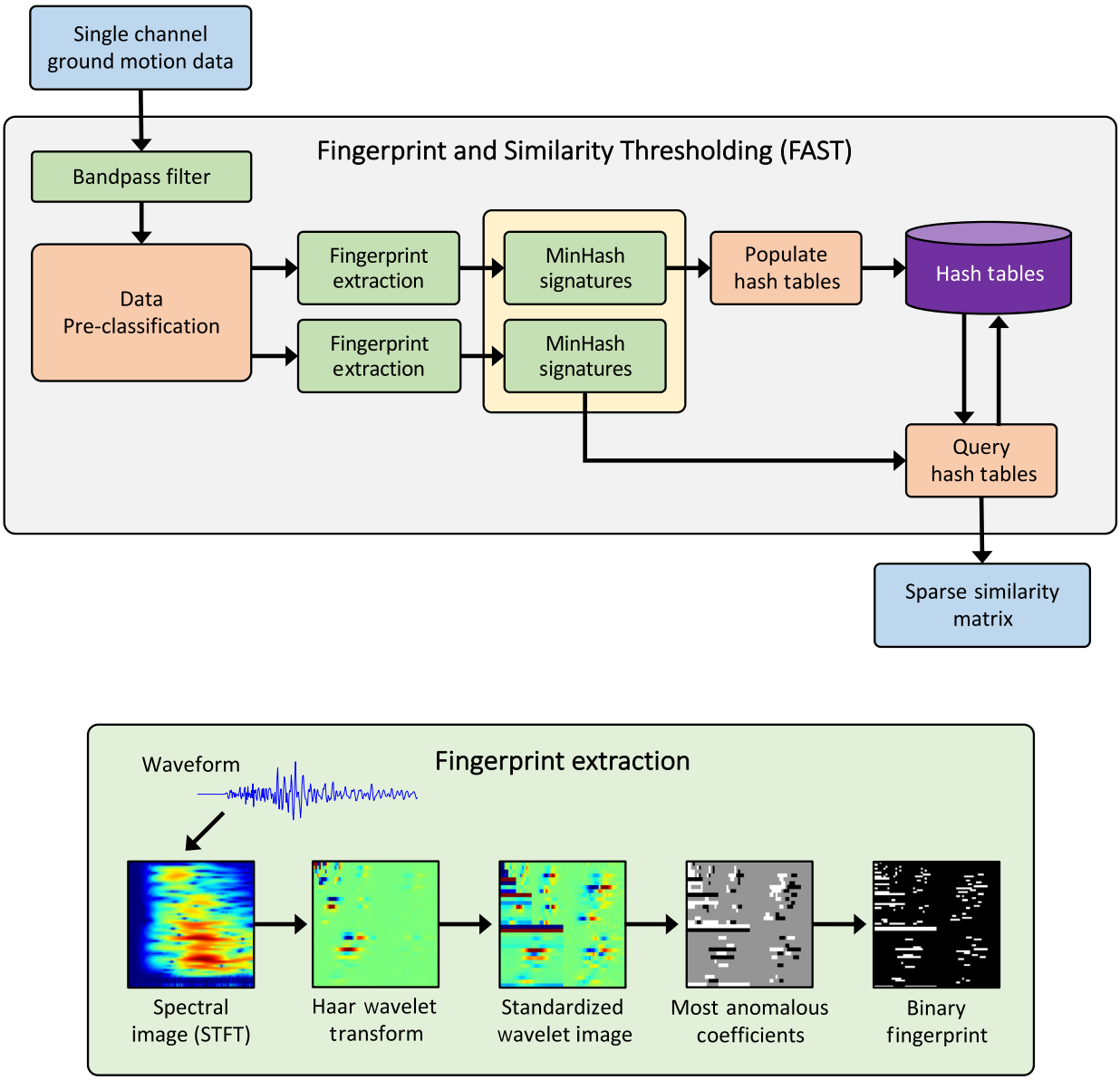

Figure 1. Overview of single-station FAST (Yoon *et al.* 2015). Diagram illustrates key steps in the single-channel implementation of FAST (top), and the key steps in the FAST fingerprint extraction algorithm (bottom). Detailed explanations of the fingerprint extraction process can be found in Yoon *et al.* (2015) and Bergen *et al.* (2016).

corresponding to waveforms of arbitrary but equal time duration, both observed at the same station; event-pairs are composed of one or more overlapping or sequential fingerprint-pairs. A *network event-pair* (or *source-pair*) refers to an event-pair that is observed across multiple stations in the network.

3.1 Event-pair extraction

The first step in our post-processing pipeline for network detection is to apply *event-pair extraction* to each single-station sparse similarity matrix. Event-pair extraction organizes a sparse similarity matrix into a list of event-pairs, the input format required for pairwise pseudo-association. Event-pair extraction acts as a deduplication step, condensing the output from FAST by representing pairwise detections in terms of events rather than fingerprints, and enables improved thresholding to eliminate excess single-station detections.

3.1.1 Event-pair model

FAST processes continuous data using a sliding window to map each short, overlapping waveform segment to a waveform fingerprint. The efficient similarity search step identifies pairs of similar fingerprints (and corresponding waveforms), returned in the form of a sparse similarity matrix. A single pair of similar events is expected to produce multiple detections (fingerprint-pairs) because adjacent time windows overlap and an event waveform may be longer in duration than a single fingerprint window. We expect that if fingerprints corresponding to times t_1 and t_2 have high similarity, then the fingerprints corresponding to times $t_1 + \ell_f$ and $t_2 + \ell_f$ will also have high similarity (see Fig. 3), especially when there is significant overlap between adjacent fingerprints (i.e. the lag between fingerprints, ℓ_f, is small compared to the length of the fingerprint window, w_f). As a result, multiple (sequential) detections will appear tightly clustered along a diagonal in the sparse similarity matrix. Each diagonal in

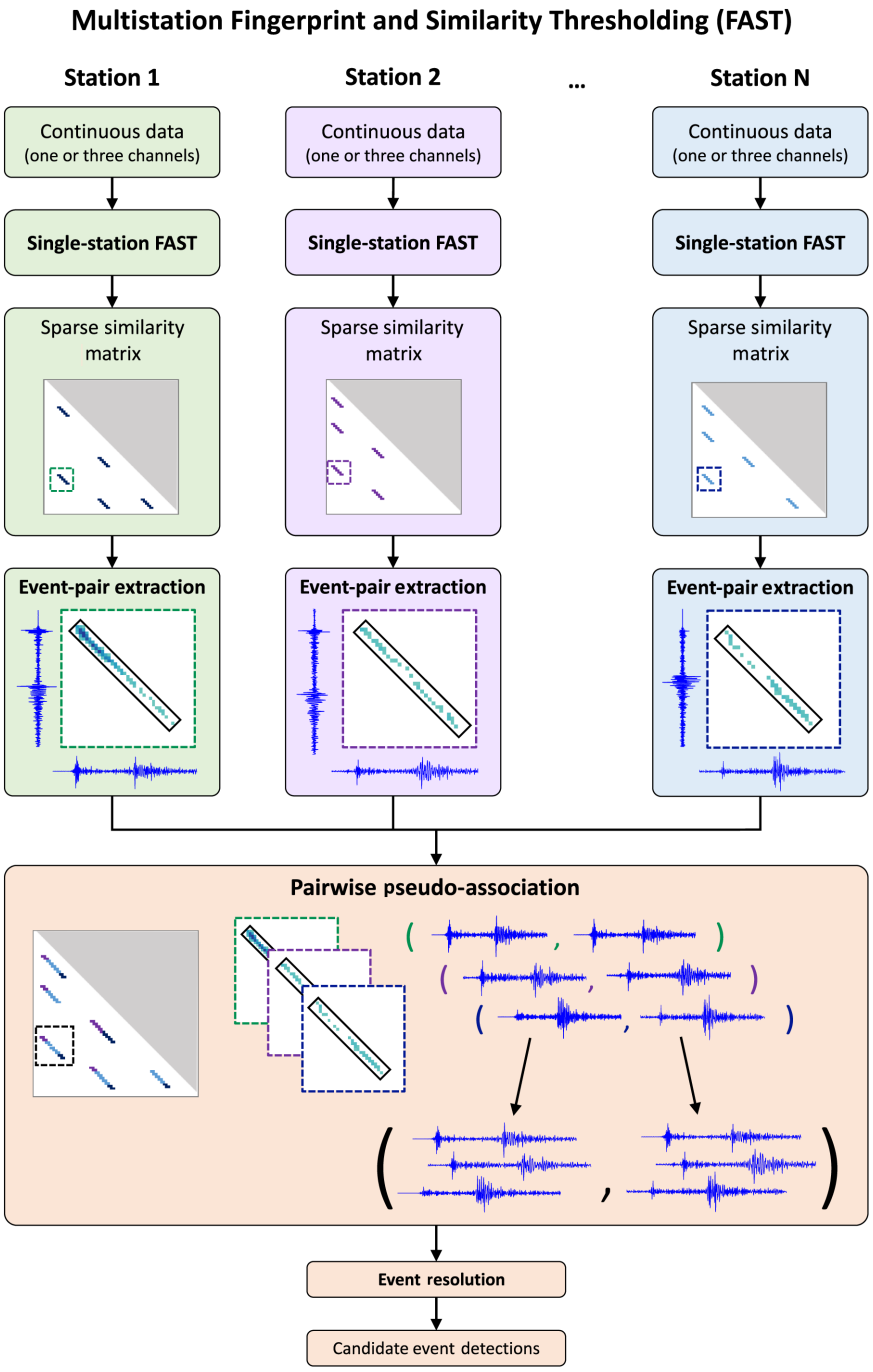

Figure 2. Diagram of multistation detection with FAST. Single-station FAST produces a sparse similarity matrix for each station, which is the input to the three-step processing pipeline for multistation detection: (1) Event-pair extraction converts sparse similarity matrix into list of pairwise event detections for each station and (2) Pairwise pseudo-association identifies event-pairs that are observed at multiple stations in the seismic network; and (3) Event resolution converts pairwise detections into list of candidate events.

the similarity matrix corresponds to a fixed value of the interevent time, $t_j - t_i$. By identifying diagonal structures in the sparse similarity matrix, we can convert the list of similar fingerprint-pairs into a list of similar event-pairs.

Examples of diagonal clusters in the sparse similarity matrix that correspond to event-pairs are shown in Supporting Information Figs S1–S3. It should be noted that not all event-pairs correspond to pairs of similar earthquake waveforms; Supporting Information Fig. S3 shows event-pairs detected by (single-station) FAST that correspond to noise.

3.1.2 Identifying event-pairs in the similarity matrix

We developed a technique called event-pair extraction to link all of the detections (fingerprint-pairs) associated with a single pair of events with similar waveforms. Event-pair extraction takes the sparse similarity matrix as input and assigns each fingerprint-pair

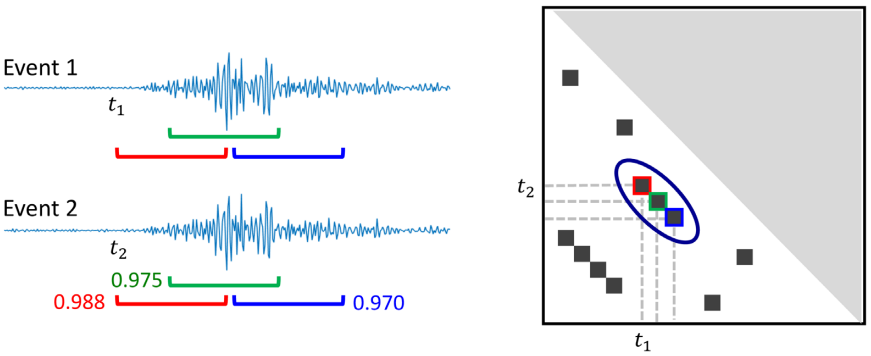

Figure 3. Diagonal structures in similarity matrix correspond to multiple detections of a single event-pair. For events 1 and 2, with arrival times t_1 and t_2, respectively, there is correspondence between fingerprints $(t_1,\ t_2)$, $(t_1 + \ell_f,\ t_2 + \ell_f)$ and $(t_1 + 2\ell_f,\ t_2 + 2\ell_f)$, where ℓ_f is the time lag between adjacent sliding windows. Note that for all three pairs of similar fingerprints the time elapsed between the first and second event in the pair is $t_2 - t_1$. The normalized cross-correlation for corresponding waveforms is given.

$(i,\ j)$ to a cluster, $\mathcal{C}$, containing all the individual pairs of similar fingerprints that correspond to a pair of similar events. Specifically, event-pair extraction identifies narrow, diagonally oriented clusters in the sparse similarity matrix; in following the event-pair model, each cluster corresponds to an event-pair, though some clusters may contain only a single, isolated fingerprint-pair.

Extraction of diagonal features in a similarity matrix has been used in natural language processing applications to identify repeating speech patterns (syllables) in unlabelled audio recordings (Jansen & Van Durme 2011). In that study, the authors treat the similarity matrix as a dense matrix and apply image processing techniques to extract the diagonal features. Because we expect to apply event-pair extraction to very large sparse matrices, with nearly one hundred thousand rows and columns for each day of continuous data, we do not wish to form a dense matrix. Therefore, we treat event-pair extraction as a clustering problem on 2-D data points, and use an iterative algorithm to identify diagonal clusters of points.

The input for event-pair extraction is a sparse similarity matrix, in coordinate list form: $(i,\ j,\ m_{ij})$, or alternatively $(t_i,\ t_j,\ m_{ij})$, using the timestamp t_i corresponding to fingerprint i. We require the index for each fingerprint be assigned sequentially and account for any time gaps in the data such that there is a direct correspondence between the index and timestamp for each fingerprint: $t_i = i \times \ell_f$, where ℓ_f is the parameter controlling the time lag between adjacent fingerprints. This is required so the kth matrix diagonal $\mathcal{D}_k$ contains all pairs of fingerprints with fixed interevent time dt_k: $\mathcal{D}_k = \{(i,\ j) : t_j - t_i = dt_k\}$, where $dt_k = k \times \ell_f$; below we use timestamps for readability.

Event-pair extraction uses an iterative clustering method that is specifically tailored for narrow, diagonally oriented clusters. In the first step, we convert the representation of the fingerprint-pairs to diagonal coordinates: $(t_i,\ t_j,\ m_{ij}) \rightarrow (dt_k = t_j - t_i,\ t_i,\ m_{ij})$. We then group the fingerprint-pairs by their diagonal coordinate dt_k and for each diagonal we sort the fingerprint-pairs by t_i, the timestamp of the earlier fingerprint within the pair. After grouping and sorting, we can identify sequences of detections along each diagonal, $\mathcal{D}_k$, using a parameter g_L to control the largest allowable gap between detections within a sequence. Each of the identified sequences defines an initial cluster, containing fingerprint-pairs with a single diagonal coordinate value (i.e. clusters of width 1), and is assigned a unique cluster ID. We then allow the initial clusters to merge; we make several passes through the data, merging adjacent clusters in diagonals dt_k and dt_{k+1} for all values of k by reassigning them to shared cluster IDs.

All fingerprint-pairs assigned to the same cluster at the end of this process define an event-pair. The properties of the extracted event-pairs will depend on the clustering parameters; for instance, permitting modest gaps in the sequences of detections along the diagonal helps ensure P and S arrivals are assigned to the same event-pair. Complete algorithmic details are included in Supporting Information Section S5.1.

3.1.3 *Event-pair summaries*

In order to reduce the size of the output and pass only essential information as an input to pairwise pseudo-association for network detection, we produce a condensed summary for each event-pair. For each cluster $\mathcal{C}$, the event-pair summary includes the coordinates of the bounding box containing the fingerprint-pairs assigned to the cluster and summary statistics that capture the degree of similarity between the two events. For fingerprint-pairs $(i,\ j) \in \mathcal{C}$, we can define a bounding box:

$$\left(dt^{(\mathcal{C})}_{\min} = \min_{(i,j)\in\mathcal{C}}(t_j - t_i),\ dt^{(\mathcal{C})}_{\max} = \max_{(i,j)\in\mathcal{C}}(t_j - t_i),\right.$$
$$\left.t^{(\mathcal{C})}_{\min} = \min_{(i,j)\in\mathcal{C}} t_i,\ t^{(\mathcal{C})}_{\max} = \max_{(i,j)\in\mathcal{C}} t_i\right), \quad (1)$$

where the first two coordinates, $dt^{(\mathcal{C})}_{\min}$ and $dt^{(\mathcal{C})}_{\max}$, give the range of the diagonal coordinate values for fingerprint-pairs assigned to the cluster, and the third and fourth coordinates, $t^{(\mathcal{C})}_{\min}$ and $t^{(\mathcal{C})}_{\max}$, give the range of the time coordinate values.

We also compute summary statistics for each event-pair cluster $\mathcal{C}$, including the number of fingerprint-pairs in the cluster, $\mathrm{ndet}^{(\mathcal{C})}$, the strongest similarity value of any fingerprint-pair in the cluster, $\mathrm{pk}^{(\mathcal{C})}$, and sum (or 'volume') of all similarity values of fingerprint-pairs included in the cluster, $\mathrm{v}^{(\mathcal{C})}$:

$$\left(\mathrm{ndet}^{(\mathcal{C})} = |\mathcal{C}|,\ \mathrm{pk}^{(\mathcal{C})} = \max_{(i,j)\in\mathcal{C}}(m_{ij}),\ \mathrm{v}^{(\mathcal{C})} = \sum_{(i,j)\in\mathcal{C}}(m_{ij})\right). \quad (2)$$

Event-pair extraction converts the output of FAST into a useful format for further analysis or verification. FAST may output dozens of fingerprint-pairs associated with the same pair of events, but in our analysis we are interested in events rather than fingerprints. Event-pair extraction collects the fingerprint-pairs corresponding to a single event-pair to form a more manageable aggregate output for analysis. Event-pair extraction provides the first step toward compiling the final detection list, grouping duplicate detections while

retaining pairwise detections. Pairwise detections are required for the network detection model used by pairwise pseudo-association for multistation detection. Even for single-station detection, the event-pair extraction retains useful information about the pairwise similarity of event detections. And event-pair extraction, combined with event-pair pruning described below, is a useful means of reducing the size of the output of FAST.

3.1.4 Updated thresholding: event-pair pruning

Event-pair extraction enables improved thresholding and pruning of events in the single-station FAST output. Typically, in the similarity search step of FAST we choose a low initial similarity threshold (τ_0) on the output, since it is more straightforward to reduce excess detections than to rerun FAST with an adjusted threshold. In Yoon *et al.* (2015) a second threshold $\tau_1 \geq \tau_0$ is selected and any fingerprint-pairs (i, j) with similarity below the new threshold, $m_{i,j} < \tau_1$, are eliminated. Event-pair extraction allows us to eliminate or 'prune' event-pairs, rather than individual fingerprint-pairs, to reduce the number of excess detections. Criteria for event-pair pruning may include a threshold on the summary statistics (e.g. a minimum cluster size, ndet $\geq |\mathcal{C}|_{\min}$), in eq. (2), or event duration, from eq. (1); the pruning criteria will depend on the data set and detection target.

The advantage of this approach is that low-similarity fingerprint-pairs are not all automatically eliminated based on a single secondary threshold on all fingerprint-pairs. Thus, we can avoid throwing away the low-similarity fingerprint-pairs that belong to an event-pair, but eliminate those that represent isolated fingerprint-pairs.

3.2 Pairwise pseudo-association

The core step in our post-processing pipeline is the identification of event-pairs observed at multiple stations across the network with *pairwise pseudo-association*. FAST and event-pair extraction together perform pairwise event detection, identifying pairs of events with similar waveforms independently for each station in the network. Pairwise pseudo-association uses these pairwise event detections, and a model for the arrival times of events with similar sources, to identify the same *pair* of detections across a network of stations.

In the following discussion, an event-pair refers to two events that are identified by FAST as having similar waveforms, and the *interevent time* refers to the amount of time that lapses between the two events belonging to the same event-pair.

3.2.1 Network detection model

A typical phase associator groups individual phases observed across the network into subsets of picks with arrival times that are consistent with a source model. Pairwise pseudo-association uses an alternate model constraint to identify event-pairs observed at multiple stations in the network. Unlike a typical association model, our *pseudo*-association model does not require phase arrivals to be consistent with a source (i.e. it does not attempt to constrain, determine or verify the source). Instead, our detection model requires an event-pair to have consistent relative arrival times over the network (by requiring fixed interevent times). As a result, pairwise pseudo-association can only be applied when detections are returned as pairs of similar events and not to single, unpaired event detections.

Specifically, our model attributes two or more event-pairs observed at different stations to the same source-pair if the interevent time is identical across the network and the arrivals are close together in time. Consider two events, Event 1 and Event 2, with similar sources and origin times t_1 and t_2, respectively. Under the assumption of a similar source, the arrival times of the P phases at station q will be $t_1 + dp^{(q)}$ and $t_2 + dp^{(q)}$, and $t_1 + ds^{(q)}$ and $t_2 + ds^{(q)}$ for S phases, where $dp^{(q)}$ and $ds^{(q)}$ are the P and S travel times from the source to station q. Therefore, a pair of similar events will have a fixed interevent time, dt, for all stations in the network independent of the particular phase detected (see Fig. 4):

$$\begin{aligned} dt &= t_2 - t_1 = (t_2 + \cancel{dp^{(q)}}) - (t_1 + \cancel{dp^{(q)}}) \\ &= (t_2 + \cancel{ds^{(q)}}) - (t_1 + \cancel{ds^{(q)}}) \quad \text{for any station } q. \end{aligned} \tag{3}$$

This allows us to directly apply the model to event detections (which may include the P arrival, S arrival or both) without explicitly identifying phases. The implicit assumption in this model, that the same phase is detected for both events at each station, holds automatically for similarity-based detectors like FAST.

3.2.2 Pairwise pseudo-association

The simplicity of our network model for pairs of detections with similar sources makes pairwise pseudo-association relatively straightforward to implement; our algorithm identifies sets of event-pairs at multiple stations that occur close together in time and share a common interevent time across the network. In terms of the sparse similarity matrices, our task is to identify sets of near- or partially overlapping diagonal clusters (event-pairs) in the matrices for two or more stations. In particular, the diagonal clusters should be very closely aligned along the diagonal coordinate, but may be shifted in time to allow for moveout across the network (an example is shown in Fig. 5).

Pairwise pseudo-association requires event-pairs as the input, so event-pair extraction must first be applied to the output from the FAST similarity search. The algorithm for pairwise pseudo-association does not require information about all the individual fingerprint-pairs within each event-pair, but instead only requires a representative diagonal coordinate (interevent time) and the time interval for the first event in the event-pair: $(dt^{(c,q)},\ t_{\min}^{(c,q)},\ t_{\max}^{(c,q)})$, for event-pair c observed at station q. Similar to fingerprint-pairs in the event-pair extraction algorithm, we group the event-pairs (from all stations) by interevent time, dt, in the pairwise pseudo-association algorithm. We then sequentially loop through all of the event-pairs and group those event-pairs that share the same or adjacent diagonal coordinates and occur close in time, using a parameter g_N to define the largest allowable time gap between event-pairs observed at different stations. By strictly enforcing the constraint that the interevent times (diagonal coordinates) must be nearly identical for all event-pairs across the network, we can use a more relaxed criterion for the relative arrival times across the network. This allows us to identify multiple event-pairs over the network that can all be attributed to a single source-pair, without having to explicitly use phase information. Implementation details for pairwise pseudo-association are described in Supporting Information Section S5.2.

3.2.3 Strengths of pairwise pseudo-association

Pairwise pseudo-association has several advantages that make it an attractive method for combining single-station similarity-based

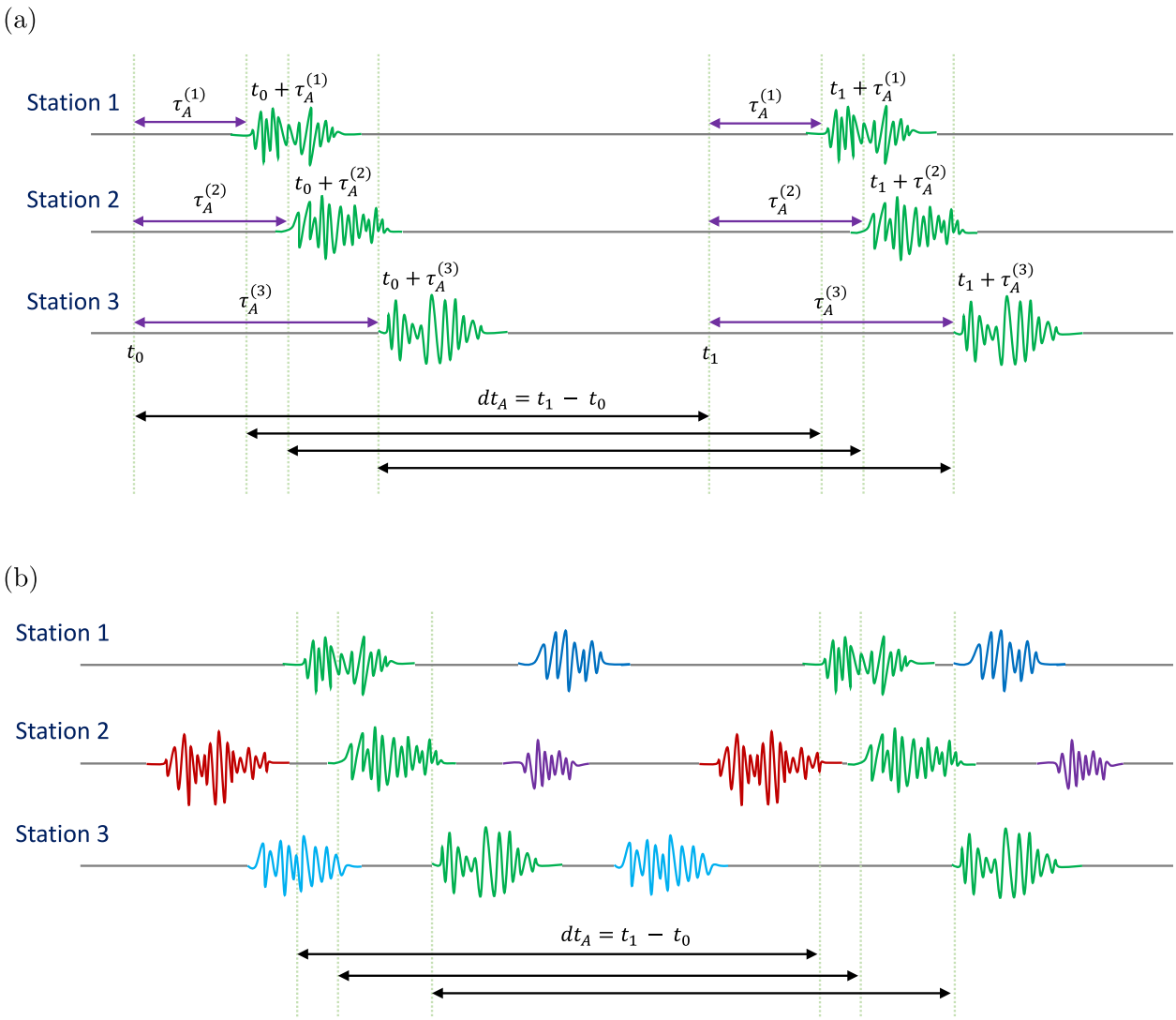

Figure 4. Illustration of fixed interevent times for pairwise pseudo-association. (a) Consider event-pair A which consists of two events with origin times t_0 and t_1. If we let $\tau_A^{(q)}$ be the P wave travel time to station q, then the arrival times at station q are $t_0 + \tau_A^{(q)}$ and $t_1 + \tau_A^{(q)}$, and the interevent time is $dt_A = t_1 - t_0$ for any station q in the network. We can apply the same reasoning to any phase arrivals, not just the P wave, so we expect any similar subinterval of the two events in pair A at any station to be separated by time dt_A. (b) The fixed interevent time requirement provides an additional constraint that allows us to associate events across the network as pairwise detections without requiring us to verify that they are consistent with some source location. In the diagram below, only one pair of events (shown in green) has arrival times over the network (fixed interevent time dt_A) that are consistent with both events sharing a similar source. *Pairwise pseudo-association* tentatively associates event-pairs if the arrival times are relatively close together (e.g. $<$60 s for regional earthquakes in a sparse network) and the interevent times are nearly identical across stations (to within $<$1–2 s, depending on size of lag used in FAST fingerprints).

detections over a sparse network to substantially reduce false detections.

One advantage is the simplicity of the model. Once event-pairs have been identified at each station, the constraints used to group event-pairs across the network are very simple: we require fixed interevent times across the network but use a relaxed constraint on the relative arrival times. Pairwise pseudo-association can be applied directly to the results from (pairwise) event detection without requiring the identification of phase arrivals. The fixed interevent times requirement from our network detection model provides a strong enough constraint on pairwise associations that it is not necessary to identify phases or include a strict constraint on the timing of phase arrivals. Thus for FAST, which as an uninformed similarity-based detector initially identifies candidate events rather than phase arrivals, pairwise pseudo-association provides a means of eliminating the bulk of the false detections with network detection before applying a phase picker to the candidate detections.

Pairwise pseudo-association is also designed to be flexible and robust. Pairwise pseudo-association is a post-processing routine that is applied to the output of similarity-based detectors applied independently to each station in the network. As a result this approach is robust to missing data (e.g. due to station dropout) or poor data quality at one or more stations. Pairwise pseudo-association can be applied to any set of pairwise detections from multiple stations in a seismic network identified by similarity-based detection, independent of the specific detection algorithm used. For instance, in the context of FAST, this means we can apply event-pair extraction and pairwise pseudo-association to the sparse similarity matrix outputs generated by FAST for vertical or three-component data, including

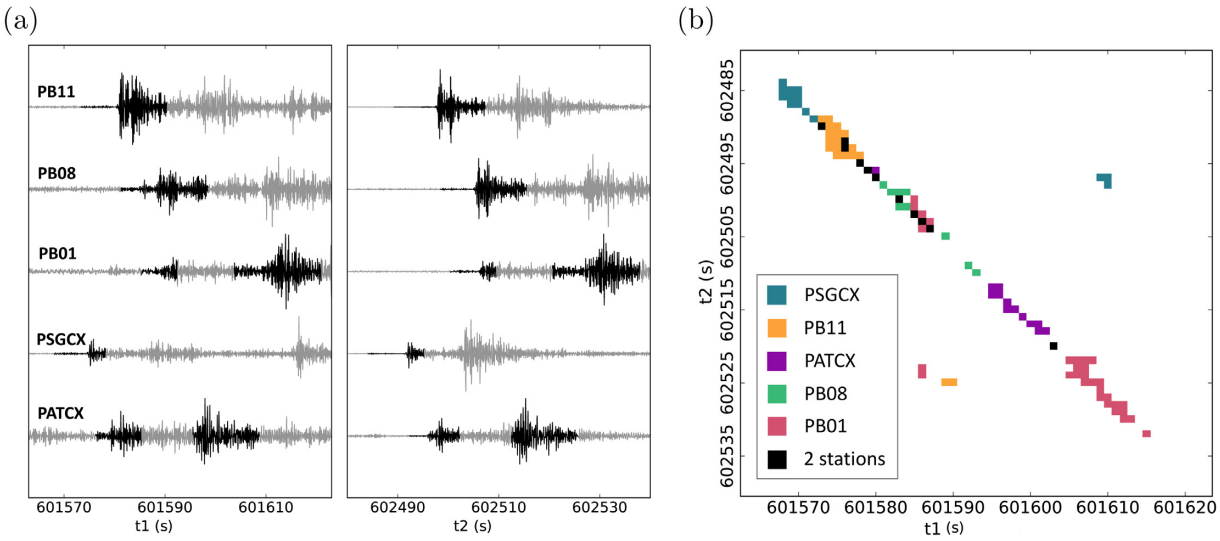

Figure 5. An example of a network event-pair detected across five stations by FAST, but with limited temporal overlap. (a) Waveform data (grey) for two events observed across five stations on 2014-03-21 near Iquique, Chile (this data set is described in Section 4 and station locations are shown in Fig. 6), with similar waveform segments identified by FAST shown in black. (b) Corresponding FAST detections in sparse network similarity matrix, with data for all five stations plotted. Detections for each station are shown in a different shade, with fingerprint-pairs detected on two stations shown in black. There are 11 fingerprint-pairs detected by FAST at exactly two stations, none are identified at 3+ stations.

both variants of three-component detection described in Supporting Information Section S7. There is also flexibility to incorporate additional constraints or criteria into our network detection procedure. For instance, in our implementation of pseudo-association we include the option to restrict the final list of network detections to only those events observed at a minimum number of stations in the network.

The flexibility and robustness we gain using a post-processing routine for network detection also comes with a downside. Since pairwise pseudo-association combines detections from each station as a post-processing step, the ability to detect weak earthquake signals over the network will depend on the detection performance at each station. Pairwise pseudo-association can not identify an event across the network unless it meets the single-station detection criteria independently at two or more stations. Therefore, the single-station detection thresholds should be kept relatively low to capture weak earthquake signals. Although network detection removes most of the additional false detections from the use of a low detection threshold, a threshold value that is set too low will result in a larger output size and slower runtime.

3.3 Event resolution

We require an additional processing step to convert the list of network event-pairs, identified with pairwise pseudo-association, into a list of events (network events). We refer to this step as *event resolution*, as it is an instance of the entity resolution problem (Getoor & Machanavajjhala 2012) in the information retrieval literature. A given event may belong to one or more event-pairs; duplicate occurrences of the event in the list of event-pairs should be linked together to create the final event list. However, in practice resolving events can be challenging due to partial detections, such as detections of only the *P* arrival or *S* arrival, of the same event, or multiple events occurring close together in time during a complex event sequence. An additional challenge arises from resolving events in a manner that is consistent with the observations of the same event across multiple stations, as identified by pairwise pseudo-association. Thus, the approach to event resolution may need to be tailored to the specific data set (see Section 4.2 for event resolution in Iquique foreshock data set, and Supporting Information Section S5.3 for algorithmic details).

4 CASE STUDY: IQUIQUE 2014 FORESHOCK SEQUENCE

We selected the foreshock sequence prior to the Iquique, Chile M_w 8.2 earthquake on 2014 April 1 to test our method for multistation detection with FAST. The foreshocks leading up to the mainshock have been taken as evidence for slow slip that is only marginally detectable in GPS data (Kato *et al.* 2016). The 50–100 km station spacing in the region is small enough to record low-magnitude events on multiple stations, but large enough that the moveout over the network presents a challenge for multistation detection. We demonstrate the ability of FAST to detect weak earthquakes with a low false alarm rate, using our proposed pipeline for combining detection results from five stations.

4.1 Data and catalogues

We apply FAST to data from a 17-day period (2014 March 15–31) during the foreshock sequence prior to the M_w 8.2 mainshock on 2014 April 1. We use three-component continuous broadband data (BHZ, BHN, BHE) from five stations in the IPOC Seismic Network. The data are sampled at 20 Hz, and we apply a bandpass filter to each channel: 2–8 Hz (stations PATCX, PSGCX, PB08 and PB11) and 1–8 Hz (station PB01); note that pairwise pseudo-association is agnostic to the choice of bandpass filter used in single-station FAST, and therefore bandpass parameters can be selected independently for each station in the network. Station locations are shown in Fig. 6.

To confirm the candidate events identified using our FAST network detection pipeline and estimate the false detection rate, we use two reference catalogues: the template matching catalogue of Kato

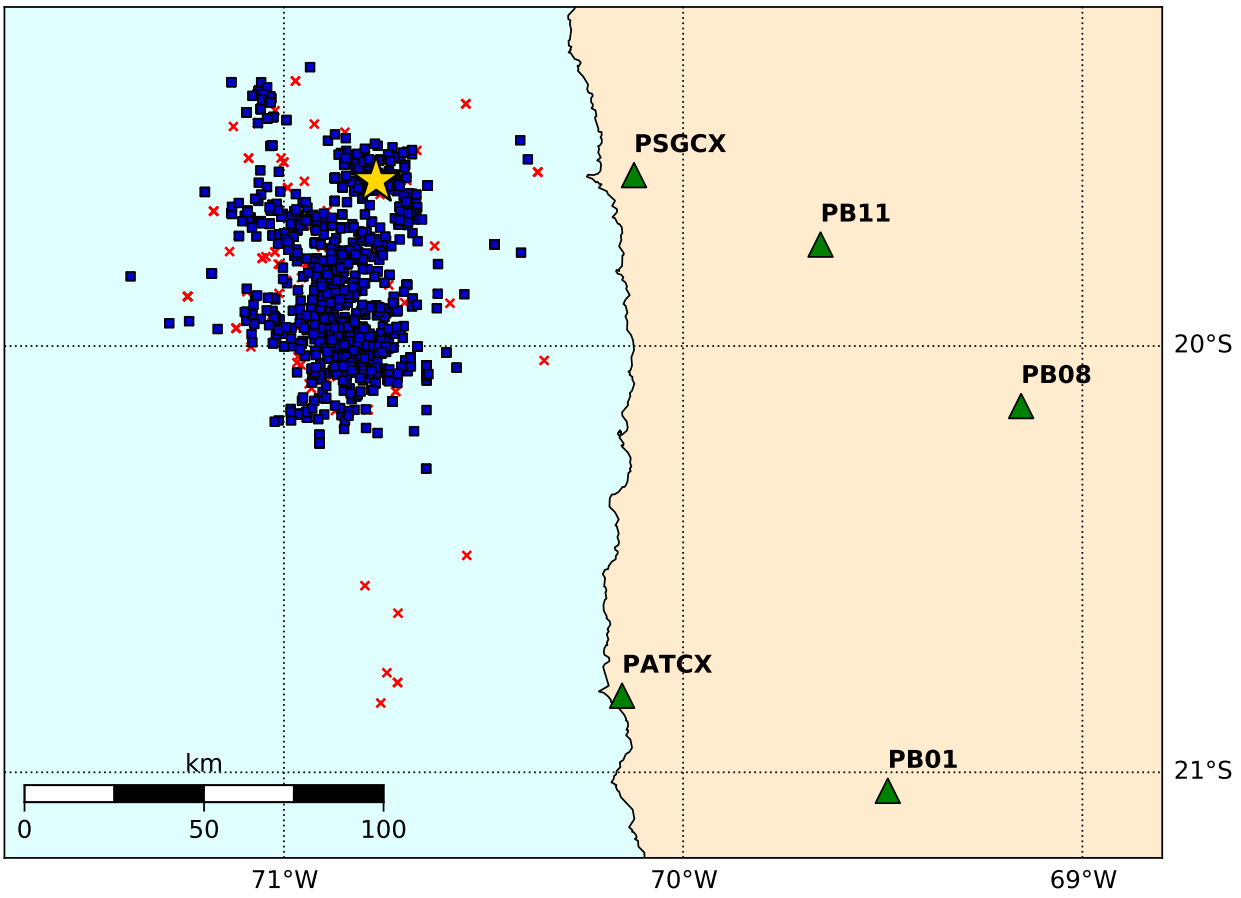

Figure 6. Data and catalogues: Triangles indicate the locations of the five stations in the Plate Boundary Observatory Network Northern Chile (CX) network. Blue squares indicate the locations of 1166 events detected by multistation FAST; only those FAST detections that are also included in the Kato *et al.* template matching and/or local seismicity catalogue are shown, and we use locations given by those catalogues. Red x's indicate locations of events in the template matching and local catalogues that were not detected by multistation FAST. Location of 2014 April 1 mainshock is indicated by yellow star.

et al. (2016), and the local seismicity catalogue from the Chilean National Seismological Center (CSN). The Kato *et al.* template matching catalogue contains 2372 events during this 17-day period. The Kato *et al.* catalogue is based on 684 template events between 2008 January and 2014 May, and detects over three-component data from 10 stations. The CSN local seismicity catalogue contains 780 events, of which 571 are in the Iquique region, bounded by latitudes −19.0° to −21.0° and longitudes −71.5° to −70.0°. Of these events, 277 are included in both catalogues. The FAST detections and the two catalogues are generated using different station data, so we primarily use these catalogues as references, rather than for a direct performance comparison.

4.2 Network detection with FAST in the Iquique region

In the first step of multistation detection with the Iquique data set, we apply fingerprint extraction and similarity search independently for each channel. In the results presented below, we use only the data from the vertical channel at each of the five stations (see Supporting Information Section S7 for detection results using three-component data). The application of FAST follows closely with that described in Yoon *et al.* (2015) (see Supporting Information Section S1 for a review of FAST), with an adjustment in the fingerprint extraction process, and with the addition of data pre-selection for all-to-some similarity search (see Supporting Information Sections S3.1 and S3.2). The parameters used for fingerprint extraction are given in Supporting Information Table S1. Seventeen days of data with 1 s fingerprint lag yields a total of 1468800 fingerprints per channel. We apply event-pair extraction and pruning to the output from each channel, and use pairwise pseudo-association to identify event-pairs that are observed at a minimum of four out of five stations. We use event resolution to generate the final list of candidate detections. This process is illustrated in Fig. 2.

The parameters used for event-pair extraction and pruning and for pairwise pseudo-association are given in Table S2. The network detection pipeline converts 275 million fingerprint-pairs in the FAST output from five stations into a list of 2788 candidate events. The initial output from FAST contains 20–93 million fingerprint-pairs per station. Event-pair extraction and pruning reduce this to 0.4–2.1 million event-pairs per station, and these are used as the inputs to pairwise pseudo-association. Pairwise pseudo-association identifies 27463 of these event-pairs at 4+ stations (see Supporting Information Fig. S4).

While pairwise pseudo-association identifies any event-pair observed at two or more stations in the network, we exclude event-pairs observed at fewer than four (out of five) stations when we generate our final detection list. By requiring detection at four stations, we ensure a low false detection rate. We have not, however, included any constraint that would limit the detections to only events inside the region of interest, so as FAST searches for similar event waveforms it may return detections of real earthquake events that

occurred outside of the Iquique region. We have avoided applying overly strict constraints on arrival times in pseudo-association to account for uncertainty about which phase is identified by FAST, though such a criterion would be straightforward to incorporate into an implementation of pairwise pseudo-association.

We use event resolution to generate the final event list from the pairwise detections produced with pairwise pseudo-association. Using only the set of event-pairs identified at four or more stations with pairwise pseudo-association, we resolve the events separately for each station in the network. Specifically, to avoid duplicate events in the final event list, we group events recorded at the same station based on temporal overlap, and associate all with a shared initial timestamp. Then we use the network event-pairs to produce a final event list that includes the event timestamp at each station.

4.3 Detection results for multistation FAST

FAST identifies 2788 candidate events across the network. We divide these detections into three groups for validation. The first group is events identified by FAST that are also in one of the available catalogues: the Kato *et al.* template-matching catalogue or the local CSN catalogue. The second group is new events identified by FAST that are similar to one or more events in the Kato or local catalogues. The third group is new candidate events identified by FAST that are not similar to any catalogue events. The detection results are summarized in Table 1. We estimate a false detection rate of less than 1 per cent among the candidate events identified by FAST with our proposed network detection algorithm.

To determine whether an event is in the catalogue, we compare the initial arrival time observed by FAST with the estimated *P* arrival times at each station. Because FAST is designed to search for similar waveform intervals rather than to pick phase arrivals, the beginning of the event waveform detected by FAST does not necessarily correspond to the *P* wave. Therefore, to determine if the FAST event is in one of the catalogues, we require that the FAST arrival time matches the estimated *P* arrival at one or more station to within 12 s (additional details in Supporting Information Section S3.4).

Eight hundred and ninety-four of the multistation FAST detections are verified by the template-matching catalogue. Five hundred and forty-five FAST detections are verified by the local CSN catalogue (only includes events in Iquique region); events located outside of the Iquique region are treated separately in Table 1. A total of 1166 FAST detections are included in one or both catalogues. The detections in the Kato *et al.* catalogue during this period belong to 296 families (based on location assignment), of which 225 contain at least two events. FAST detects events in 200 of these 225 families (89 per cent), and FAST detects 545 of the 571 events (95 per cent) in the CSN catalogue. Thus, network detection with FAST achieves good coverage of known waveform signatures for this period. The locations and magnitudes of the FAST detections that are confirmed by the template matching and local catalogues are shown in Figs 6 and 7, respectively.

For any event, we can obtain a list of similar events, as identified by FAST. We can use the links between events to determine which detections are similar to one or more events in the template matching or local seismicity catalogues. Specifically we identify new detections that are not included in the existing catalogues, but in which we have high confidence due to their similarity to a catalogue event at four stations. Waveforms of these events are inspected visually, and those that do not have clear *P* and *S* arrivals across the network are subject to additional scrutiny (described below). There are 779 multistation FAST detections that are similar to catalogue events.

The events detected by multistation FAST that are similar to, but not included in, the template-matching catalogue reflect one advantage of our method. FAST finds additional detections similar to events in Kato *et al.* catalogue because it searches for similar waveform fingerprints, which are derived from the time-frequency representation of the signal. This allows FAST to identify events that are similar in the frequency domain but not well-aligned in the time domain.

FAST also detects 722 events that are neither in an existing catalogue, nor similar to catalogue events. These events may represent true events or false detections. Examples of waveforms of events in this category are shown in Supporting Information Fig. S5. We divide the 722 events that are not similar to known catalogue events into three groups: candidate events that are likely to originate outside of the Iquique region due to the relative arrival times (3c in Table 1), candidate events with clear *P* and *S* arrivals at multiple stations in the network (4a in Table 1), and ambiguous events that require more careful inspection (4b and c in Table 1). To confirm these ambiguous waveforms without clear *P* and *S* arrivals, we use both visual inspection and waveform cross-correlation, in particular comparing the waveforms to those of similar events (as identified by FAST).

There are 14 detections that remain unconfirmed, requiring additional evidence and analysis for a determination, though none of these events are clear false detections. Whether these detections are real events or false positives, 14 events represent only 0.5 per cent of the those identified by FAST. Thus, we can conclude that the false detection rate of FAST with pairwise pseudo-association over a network of five stations is very low.

5 DISCUSSION AND CONCLUSIONS

We have proposed three extensions to the FAST detection pipeline to enable multistation detection with high confidence and low false-detection rate in long-duration data sets over a seismic network. These extensions include event-pair extraction for improved thresholding, pairwise pseudo-association for detecting pairs of events with similar sources observed at multiple stations in a seismic network, and event resolution for producing the final list of candidate detections.

5.1 Event-pair extraction

Event-pair extraction is a deduplication routine that converts a sparse similarity matrix, the output format from FAST similarity search, into a list of pairs of events with similar waveforms. We require event-pair extraction as a means of obtaining event-pairs as inputs for pairwise pseudo-association, but event-pair extraction has advantages as a stand-alone post-processing routine, making it useful even for single-station detection.

Blind similarity-based detection is able to detect events with previously unknown signatures, but the trade-off is more false detections compared with informed similarity-based detectors. Event-pair extraction can aid in removing excessive false detections, prior to network detection, with improved thresholding. Event-pair extraction provides the ability to accept or reject detections at the level of event-pairs rather than individual fingerprint-pairs, based on more relevant criteria such as event duration, cluster shape and several similarity measures (by contrast, only a single measure, similarity, is

Table 1. Catalogues include Kato *et al.* template matching catalogue, and local catalogue (CSN) in Iquique region only, unless specified.

Total FAST candidate event detections	2788		Categories
Total FAST event detections, confirmed		2774	(1), (2), (3), (4ab)
Total FAST event detections, unconfirmed		14	(4c)
(1) FAST event detections in catalogues	1166		
(a) In Kato catalogue		894	(of 2372)
(b) In CSN catalogue		545	(of 571)
(2) FAST event detections similar to catalogue events	779		(clear, checked)
(a) Similar to events in Kato catalogue only		142	(127, 15)
(b) Similar to events in CSN catalogue only		160	(146, 14)
(c) Similar to events in both Kato and CSN catalogues		477	(472, 5)
(3) FAST event detections outside of Iquique region	256		
(a) In CSN catalogue		51	
(b) Similar to events in CSN catalogue only		70	
(c) Additional out-of-region detections		135	
(4) Additional FAST detections	587		
(a) Confirmed, clear *P* and *S* arrivals over network		252	
(b) Confirmed, by visual inspection, cross-correlation		321	
(c) Unconfirmed detections (possible false detections)		14	

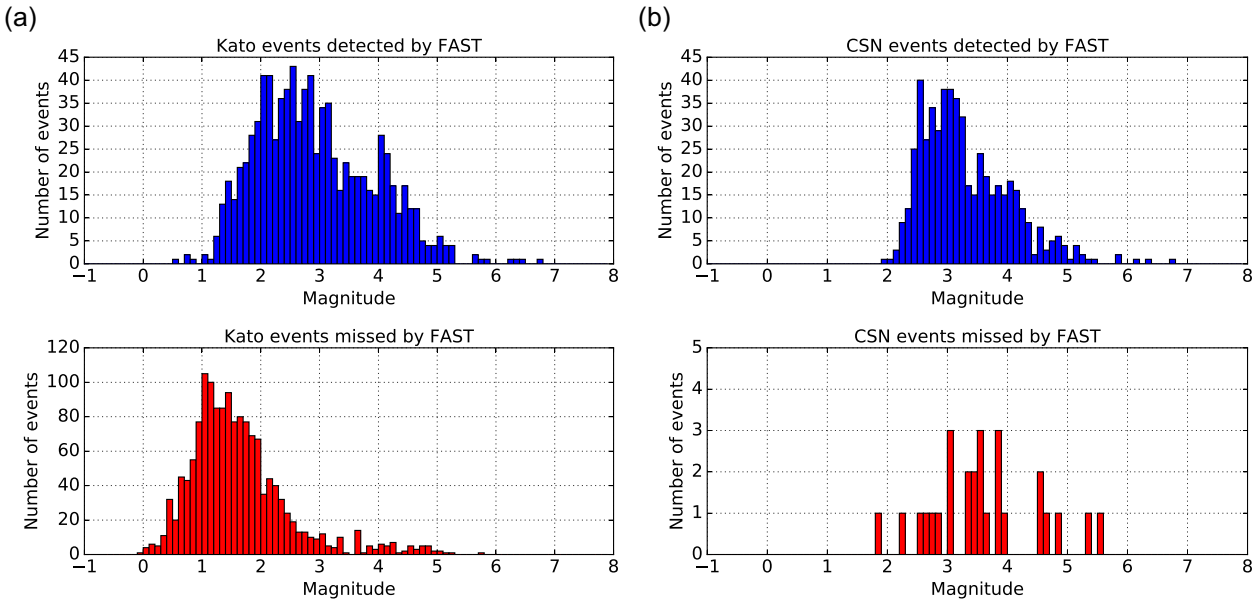

Figure 7. Histograms of the magnitudes of catalogue events detected (top row, blue) or missed (bottom row, red) by multistation FAST. All plots have the same *x*-axis scale, but *y*-axis scale varies. (a) Among events included in Kato *et al.* template matching catalogue, FAST detects 60 per cent of events *M*2.0–2.9, and 79 per cent of events *M*3.0+, but only 12 per cent (160 events) of those below *M*1.9. (b) Most of the events in the local (Chilean National Seismological Center, CSN) catalogue have magnitudes greater than 2, and multistation FAST detects 545 of 571, or 95.4 per cent, of the CSN catalogue events in the Iquique region.

available for each fingerprint-pair). Event-pair extraction also maintains the pairwise similarity information about detections, another source of insight for reducing false detections, including multiplicity, clustering or PageRank scores (Aguiar & Beroza 2014). In a partially informed similarity search, known events or waveform characteristics can be used to guide the selection of thresholding parameters.

The output size from blind similarity search will increase as it is extended to larger data sets, and event-pair extraction can play a useful role in reducing the file sizes generated by similarity search. Each non-zero (fingerprint-pair) in the sparse similarity matrix output is represented with three integers (two coordinates, one similarity statistic), while each event-pair can be represented by 4–7 integer values (3–4 coordinates, and 1–3 similarity statistics). In the Iquique foreshock data set, the number of event-pairs compared to fingerprint-pairs is 35–50 per cent, and after pruning isolated detections, this reduces to 1.8–2.4 per cent; thus, if we apply event-pair extraction and pruning to the sparse similarity search output, we can reduce the size of the output file by more than a factor of 20 while maintaining all key detection results. Event-pair extraction and pairwise pseudo-association can both be extended to long-duration data sets, for example by parallelizing over contiguous blocks of the sparse similarity matrix.

In the case where only one station is available, event-pair extraction and event resolution can be applied together, skipping the pairwise pseudo-association step, to generate the final list of candidate detections. These two techniques together can function as an update to the single-station post-processing steps in Yoon *et al.* (2015) that convert the sparse similarity matrix, generated by the efficient similarity search step, into a list of event detections.

5.2 Pairwise pseudo-association

Pairwise pseudo-association is a network detection method designed specifically to combine the results of blind similarity search at each station in a seismic network. Our approach allows us to reduce significantly the number of single-station false detections by identifying a set of high-confidence candidate events observed at multiple stations in the network for further analysis and processing (e.g. phase-picking, source location).

Pairwise pseudo-association is a useful technique for confirming FAST detections using a network-based approach. In the Iquique foreshock data set, network detection with FAST using pseudo-association identifies 2061–2829 unique events recorded at each station. The number of events detected at each station using single-channel FAST (with event-pair extraction and pruning but no network detection), is significantly higher, ranging from 9900 to 13628. Among these events are a substantial number of false detections, due to correlated noise, but without additional contextual information it is difficult to assess the detection performance and false detection rate. Examples of waveforms identified by single-station FAST are shown in Supporting Information Fig. S6. Identifying true events by computing the normalized cross-correlation of detected waveform pairs is unreliable for a single station due to (1) the possible presence of correlated noise and (2) the ability of FAST to detect events with high similarity in the spectral domain but only moderate time-domain similarity. Therefore, to confirm the single-station detections would require identifying phase arrivals and associating and locating events; this may be challenging for events with weak or non-impulsive arrivals. Pairwise pseudo-association allows us to cut down significantly on false detections due to local noise sources, but it should be noted that it will not eliminate signals from non-earthquake sources, such as quarry blasts, that are strong enough to be observed at multiple stations in the network.

Pairwise pseudo-association has built-in flexibility, including the ability to control the trade-off between detection sensitivity and tolerance of false detections. Blind similarity-based detectors such as FAST share the same challenge as other classes of detection algorithms: to detect as many weak events as possible, it is necessary to accept the possibility of more false detections in the results. Ideally we would like an algorithm that will detect relatively weak signals with a low false-detection rate. Pairwise pseudo-association returns not only a list of candidate events and event-pairs, observed at multiple stations across the network, but also provides a number of detection statistics including the number of stations that observed the event, the multiplicity of detections (total number of similar events), and several measures of pairwise event similarity over the network. These provide a greater control over the criteria used to select which events to accept in our final list of candidate detections, and in practice the criteria will be selected based on tolerance for false positives. For the Iquique case study, we use a threshold on the number of stations at which the event is detected; we selected a higher threshold to ensure few false detections, but if we had a greater tolerance for false detections we could loosen our criteria for inclusion in our final event list.

Another advantage of the proposed extensions of FAST is that they maintain the modularity of the FAST detection pipeline and allow for flexibility and customization for a particular data set or detection task. As an example, although we have not specifically investigated detection in a dense network, many of the techniques and lessons presented in this work could be extended to dense networks. The event-pair extraction and pairwise pseudo-association for sparse networks could be applied directly in the case of a dense network, possibly with additional constraints within pseudo-association related to network geometry and relative arrival times. Alternatively, we could approach detection in a dense network by dividing the stations into smaller subnetworks; detection over stations within each subnetwork could be handled using one of the techniques for the limited moveout case, and then the output from each subnetwork could be combined using pairwise pseudo-association over all subnetworks.

Finally, as a blind waveform similarity-based detector, FAST is capable of identifying uncatalogued events without the availability of templates. The extension of FAST for detection over a seismic network can be used to identify new multistation template waveforms to complement higher sensitivity, informed similarity-based detectors like template matching. When the set of known event waveforms is limited, FAST can find similar signals from weak, uncatalogued events. FAST fingerprint extraction uses time-frequency features but not phase information, also enabling detection of similar events not picked up by template matching. Template matching is more sensitive for detection when waveforms are known *a priori*. Thus, we expect that when used together, FAST for identification of multistation templates and template matching for high sensitivity-detection, these two techniques will identify more weak events than either method on its own.

ACKNOWLEDGEMENTS

This research is supported by National Science Foundation (NSF) grant number EAR-1551462. This research is supported by the Southern California Earthquake Center (Contribution No. 7955). SCEC is funded by NSF Cooperative Agreement EAR-1033462 & USGS Cooperative Agreement G12AC20038. Computing resources were provided by the Center for Computational Earth and Environmental Science (CEES), Stanford University. Local seismicity catalogue was provided by the Chilean National Seismological Center (Centro Sismológico Nacional, CSN), Universidad de Chile (http://www.sismologia.cl, last accessed 2017 May 5). Template matching earthquake catalogue was provided by Aitaro Kato (Kato & Nakagawa 2014; Kato *et al.* 2016). All waveform data used in this study is from the Integrated Plate Boundary Observatory Chile (IPOC) seismic network (GFZ German Research Centre for Geosciences and Institut des Sciences de l'Univers-Centre National de la Recherche CNRS-INSU 2006). ObsPy software (Beyreuther *et al.* 2010) was used for accessing and processing waveform data.

REFERENCES

Allen, R., 1982. Automatic phase pickers: their present use and future prospects, *Bull. Seism. Soc. Am.,* **72**(6B), S225–S242.

Aguiar, A.C. & Beroza, G.C., 2014. PageRank for earthquakes, *Seism. Res. Lett.,* **85**(2), 344–350.

Aguiar, A.C., Chao, K. & Beroza, G.C., 2017. Tectonic tremor and LFEs on a reverse fault in Taiwan, *Geophys. Res. Lett.,* **44,** 6683–6691.

Andoni, A. & Indyk, P., 2006. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, *FOCS'06: 47th Annual IEEE Symposium on Foundations of Computer Science, 2006.,* IEEE, pp. 459–468.

Baluja, S. & Covell, M., 2008. Waveprint: efficient wavelet-based audio fingerprinting, *Pattern Recognit.,* **41**(11), 3467–3480.

Barrett, S.A. & Beroza, G.C., 2014. An empirical approach to subspace detection, *Seismol. Res. Lett.,* **85**(3), 594–600.

Bergen, K. & Beroza, G.C., 2016. Earthquake fingerprints: representing earthquake waveforms for similarity-based detection, *AGU Fall Meeting Abstract S53A-2794,* San Francisco, CA.

Bergen, K., Yoon, C. & Beroza, G.C., 2016. Scalable similarity search in seismology: a new approach to large-scale earthquake detection, in *SISAP'16: Proceedings of the 9th Int. Conf. on Similarity Search and Applications, Lecture Notes in Computer Science, Vol. 9939,* Springer, pp. 301–308.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y. & Wassermann, J., 2010. ObsPy: a Python toolbox for seismology, *Seism. Res. Lett.,* **81**(3), 530–533.

Broder, A.Z., Charikar, M., Frieze, A.M. & Mitzenmacher, M., 2000. Min-wise independent permutations, *J. Comput. Syst. Sci.,* **60**(3), 630–659.

Brown, J.R.,Beroza, G.C. & Shelly, D.R., 2008. An autocorrelation method to detect low frequency earthquakes within tremor, *Geophys. Res. Lett.,* **35,** L16305, doi: 10.1029/2008GL034560.

GFZ German Research Centre For, Geosciences & Institut Des Sciences De, L'Univers-Centre National De La Recherche CNRS-INSU. 2006. *IPOC Seismic Network. Integrated Plate Boundary Observatory Chile - IPOC,* doi:10.14470/pk615318.

Getoor, L. & Machanavajjhala, A., 2012. Entity resolution: theory, practice & open challenges, *Proceedings VLDB Endowment,* **5**(12), 2018–2019.

Gibbons, S.J. & Ringdal, F., 2006. The detection of low magnitude seismic events using array-based waveform correlation, *Geophys. J. Int.,* **165**(1), 149–166.

Harris, D., 2006. Subspace detectors: Theory. Lawrence Livermore National Laboratory Report UCRL-TR-222758., Lawrence Livermore National Laboratory, Livermore, CA.

Jansen, A. & Van Durme, B., 2011. Efficient spoken term discovery using randomized algorithms, in *ASRU'11: Proc. IEEE Workshop Autom. Speech Recognit. Underst.,* IEEE, pp. 401–406.

Johnson, C.E., Lindh, A.G. & Hirshorn, B.F., 1997. Robust regional phase association,*Open-File Report 94–621,* U.S. Geological Survey.

Kato, A. & Nakagawa, S., 2014. Multiple slow-slip events during a foreshock sequence of the 2014 Iquique, Chile M_w 8.1 earthquake, *Geophys. Res. Lett.,* **41**(15), 5420–5427.

Kato, A., Fukuda, J.I., Kumazawa, T. & Nakagawa, S., 2016. Accelerated nucleation of the 2014 Iquique, Chile M_w 8.2 earthquake, *Sci. Rep.,* **6,** 24792.

Mykkeltveit, S. & Bungum, H., 1984. Processing of regional seismic events using data from small-aperture arrays, *Bull. Seism. Soc. Am.,* **74**(6), 2313–2333.

Peng, Z. & Zhao, P., 2009. Migration of early aftershocks following the 2004 Parkfield earthquake, *Nat. Geosci.,* **2**(12), 877–881.

Ringdal, F. & Kværna, T., 1989. A multi-channel processing approach to real time network detection, phase association, and threshold monitoring, *Bull. Seism. Soc. Am.,* **79**(6), 1927–1940.

Shelly, D.R., Beroza, G.C. & Ide, S., 2007. Non-volcanic tremor and low-frequency earthquake swarms, *Nature,* **446,** 305–307.

Skoumal, R.J., Brudzinski, M.R., Currie, B.S. & Levy, J., 2014. Optimizing multi-station earthquake template matching through re-examination of the Youngstown, Ohio, sequence, *Earth Planet. Sci. Lett.,* **405,** 274–280.

Skoumal, R.J., Brudzinski, M.R. & Currie, B.S., 2016. An efficient repeating signal detector to investigate earthquake swarms, *J. Geophys. Res. Solid Earth,* **121**(8), 5880–5897.

Tibi, R., Young, C., Gonzales, A., Ballard, S. & Encarnacao, A., 2017. Rapid and robust cross-correlation-based seismic signal identification using an approximate nearest neighbor method, *Bull. Seism. Soc. Am.,* **107**(4), 1954–1968.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S. & Trujillo, J., 1998. A comparison of select trigger algorithms for automated global seismic phase and event detection, *Bull. Seism. Soc. Am.,* **88**(1), 95–106.

Yoon, C., O'Reilly, O., Bergen, K. & Beroza, G., 2014. Computationally efficient search for similar seismic signals in continuous waveform data over a seismic network, *AGU Fall Meeting Abstracts S52A-04.,* San Francisco, CA.

Yoon, C.E., O'Reilly, O., Bergen, K.J. & Beroza, G.C., 2015. Earthquake detection through computationally efficient similarity search, *Sci. Adv.,* **1**(11), e1501057, doi:10.1126/sciadv.1501057.

Yoon, C.E., Huang, Y., Ellsworth, W.L. & Beroza, G.C., 2017. Seismicity during the initial stages of the Guy-Greenbrier, Arkansas, earthquake sequence, *J. Geophys. Res. Solid Earth,* **122**(11), 9253–9274.

Young, C.J., Woodbridge, J., Shaw, R. & Slinkard, M., 2015. Using KLSH to rapidly search large seismic signal archives on a desktop computer, *AGU Fall Meeting Abstracts S53B-2825,* San Francisco, CA.

Zhang, M. & Wen, L., 2015. An effective method for small event detection: match and locate (M&L). *Geophys. J. Int.,* **200**(3), 1523–1537.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Table S1. Parameters used for fingerprint extraction and similarity search for Iquique foreshock data set.

Table S2. Parameters used for event-pair extraction, pruning, and network pseudo-association in Iquique foreshock data set.

Table S3. Percentage of total fingerprints at each station (vertical channel) that are included in the database set for all-to-some similarity search and percentage of those used for computing the wavelet coefficient statistics for fingerprinting.

Table S4. Comparison of the output of all-to-all versus all-to-some search on the same collection of fingerprints.

Figure S1. Examples of event-pairs identified with event-pair extraction: strong event-pair detection.

Figure S2. Examples of event-pairs identified with event-pair extraction: lower signal-to-noise event-pair detections.

Figure S3. Examples of event-pairs identified with event-pair extraction: false detections.

Figure S4. Size of single-station FAST output for station PSGCX at different stages of post-processing pipeline.

Figure S5. Examples of event waveforms for events that were detected by FAST, but not in template matching (Kato) or local (CSN) seismicity catalogues.

Figure S6. Examples of waveforms identified by single-station detection with FAST.

Figure S7. Relationship between the Jaccard similarity and the probability of detection in LSH-based similarity search.

Figure S8. Diagram of all-to-some FAST search.

Figure S9. Example of data selection for 'database set' for all-to-some FAST similarity search and 'background set' for computing background statistics in fingerprint extraction.

Figure S10. Barplots illustrate the comparison between the runtime and outputs of all-to-some search with all-to-all search as the baseline.

Figure S11. Waveforms for six events that were detected using all-to-all search but were not detected by all-to-some search.

Figure S12. The reduction in runtime of the all-to-some versus all-to-all search will depend on the distribution of the data in the hash tables.

Figure S13. Comparison of data distribution in hash tables for the all-to-all search (all fingerprints in database set) versus all-to-some search (subset of fingerprints in database set) for Iquique foreshock data set.

Figure S14. Diagram illustrating fingerprint concatenation.

Figure S15. Additional candidate events identified by 5SFP-FAST, but not FAST with pairwise pseudo-association.

Figure S16. Multistation-fingerprint FAST detection: A comparison of the 1067 netFAST detections also identified by 5SFP-FAST with the 1675 netFAST detections that were missed by 5SFP-FAST.

Figure S17. Multistation-fingerprint FAST detection: A comparison of the similarity of the pairs of multistation versus pairs of single-station fingerprints.