Earthquake Arrival Association with Backprojection and Graph Theory

by Ian W. McBrearty, Joan Gomberg, Andrew A. Delorey, and Paul A. Johnson

Abstract The association of seismic-wave arrivals with causative earthquakes becomes progressively more challenging as arrival detection methods become more sensitive, and particularly when earthquake rates are high. For instance, seismic waves arriving across a monitoring network from several sources may overlap in time, false arrivals may be detected, and some arrivals may be of unknown phase (e.g., P or S waves). We propose an automated method to associate arrivals with earthquake sources and obtain source locations applicable to such situations. To do so, we use a pattern detection metric based on the principle of backprojection to reveal candidate sources followed by graph-theory-based clustering and an integer linear optimization routine to associate arrivals with the minimum number of sources necessary to explain the data. This method solves for all sources and phase assignments simultaneously, rather than in a sequential greedy procedure as is common in other association routines. We demonstrate our method on both synthetic and real data from the Integrated Plate Boundary Observatory Chile seismic network of northern Chile. For the synthetic tests, we report results for cases with varying complexity, including rates of 500 earthquakes/day and 500 false arrivals/station/day, for which we measure true positive detection accuracy of >95%. For the real data, we develop a new catalog between 1 January 2010 and 31 December 2017 containing 817,548 earthquakes, with detection rates on average 279 earthquakes/day and a magnitude-of-completion of $M \sim 1.8$. A subset of detections are identified as sources related to quarry and industrial site activity, and we also detect thousands of foreshocks and aftershocks of the 1 April 2014 $M_w 8.2$ Iquique earthquake. During the highest rate of aftershock activity, >600 earthquakes/day are detected in the vicinity of the Iquique earthquake rupture zone.

Supplemental Content: Additional details regarding the methods used in this article, along with a catalog of northern Chilean seismicity between 1 January 2010 and 31 December 2017.

Introduction

Earthquake catalogs are invaluable for monitoring seismic activity and for developing geophysical images of the Earth structure. Cataloging earthquakes generally follows the detection of impulsive arrivals followed by the association of arrivals between stations to a common source and then inferring the location, origin time, and magnitude of each earthquake. As improved instrumentation and increasing quantities of data have become available, arrival picking algorithms have markedly progressed (Küperkoch et al., 2010; Baillard et al., 2013; Ross and Ben-Zion, 2014; Ross et al., 2018; Zhu and Beroza, 2018). Along with these improvements, the number of detected arrivals generally increases approximately by an order of magnitude per decreasing magnitude unit of sensitivity, introducing new challenges into the earthquake detection pipeline. Converting high rates of arrival picks into an accurate earthquake catalog can be invaluable in seismology because dense catalogs improve our understanding of seismogenic processes occurring at plate boundaries (Kato and Nakagawa, 2014; Delorey et al., 2015), allow for monitoring rate changes of seismicity (Montoya-Noguera and Wang, 2017; Fiedler et al., 2018), improve resolution of tomographic images (Peng and Ben-Zion, 2005; Watkins et al., 2018), reveal dynamic triggering and anthropogenic-induced seismicity (Shapiro et al., 2006; Hill and Prejean, 2015; Peng et al., 2009; Ellsworth, 2013), and may contain information regarding the timing of future earthquakes (Rouet-Leduc et al., 2017; Lubbers et al., 2018). Dense catalogs also provide new datasets that can be incorporated into increasingly popular machine-learning approaches for a variety
of applications in seismology (DeVries et al., 2018; Perol et al., 2018; Ross et al., 2018; Zhu and Beroza, 2018; Zhu et al., 2018).

Earthquake association algorithms often struggle when arrival rates are high and there is ambiguity regarding the number of earthquakes that produce the arrival-time measurements (Johnson et al., 1997). False arrivals and unknown phase types of arrivals also introduce uncertainties. These issues have historically been most problematic during aftershock and foreshock sequences; however, in the new era of sensitive arrival picking methods, these same issues are encountered for even normal background seismicity rates. As such, there is a need for improved association algorithms to enable the development of high-fidelity, low-magnitude-of-completion earthquake catalogs. A general-purpose association algorithm should be able to handle a few interrelated problems, all connected to the key problem of earthquake detection. These include (1) an unknown number of sources in a time window, (2) false arrivals, (3) missing arrivals, (4) unknown phase types (e.g., P or S waves), (5) uncertainties in the velocity model, and (6) the scenario in which arrivals from several earthquakes overlap in time across a network (Fig. 1).

In this article, we solve the association problem when all of these uncertainties (1–6) are present using a backprojection (BP) and graph theory framework. The method processes a dataset of arrival-time picks assumed to be of unknown phase and that can include an unknown number of false arrivals and errors in arrival-time estimates. The technique is generalizable to arbitrary spatial scales (from laboratory to Earth) and recording network geometries. To perform BP, we rely on a velocity model; however, the method can tolerate uncertainty in the velocity model and travel-time calculations used in the analysis. In the process of inferring associations, estimates for hypocenter locations and origin times are also determined, which can be used as starting points for higher-precision methods. Effort has been taken to maintain a computationally efficient and scalable algorithm that can be used for catalog development in a wide range of applications. In conjunction with this work, we also introduce a newly developed phase picker, referred to as the “Norm Picker,” a single-station energy-based picking algorithm that is similar to a short-term average/long-term average (STA/LTA; Allen, 1978) but that can pick P and S waves nearly equally as well and also pick low-signal-to-noise ratio (SNR) arrivals.

The structure of this article is as follows: (1) Introduction and Background, (2) Methods, (3) Synthetic Tests, (4) Application, (5) Discussion, and (6) Conclusion. We perform synthetic tests of our method on simulated arrival-time data from 17 stations of the Integrated Plate Boundary Observatory Chile (IPOC) of northern Chile (Table S1, available in the supplemental content to this article) and use real data from the same network for our application. In the synthetic tests, we analyze the performance of the method as the rate of earthquakes, the rate of false arrivals, and errors on arrival-time data all change. We apply the method up to the relatively challenging synthetic case in which 700 earthquakes/day and 700 false arrivals/station/day occur across the network. For our application to real seismic data, we develop a catalog for northern Chile from 2010 through the end of 2017, which has a magnitude-of-completion of M ∼ 1.8 and detects on average 279 earthquakes/day.

### Background

Previous studies solved the association problem in several different ways with varying degrees of generalization and
Tables and Figures

### Table 1

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>Spatial coordinate (latitude, longitude, depth)</td>
</tr>
<tr>
<td>( T_r(\mathbf{X}) )</td>
<td>Travel time to station ( i ) from position ( \mathbf{X} ) of phase type ( r )</td>
</tr>
<tr>
<td>( \vec{T}_r(\mathbf{X}) )</td>
<td>Relative travel time to station ( i ) from position ( \mathbf{X} ) of phase type ( r )</td>
</tr>
<tr>
<td>( C(\mathbf{X}, i) )</td>
<td>Backprojection space</td>
</tr>
<tr>
<td>( \vec{C}(\mathbf{X}, i) )</td>
<td>Theoretical backprojection space for dataset of source ((\mathbf{X}, i) = 0)</td>
</tr>
<tr>
<td>( \theta(\mathbf{X}) )</td>
<td>Prior likelihood distribution of sources</td>
</tr>
<tr>
<td>( \Omega )</td>
<td>Spatial-temporal domain of sources</td>
</tr>
<tr>
<td>( \Omega_X )</td>
<td>Spatial domain of sources</td>
</tr>
<tr>
<td>( \Omega_T )</td>
<td>Temporal domain of sources</td>
</tr>
<tr>
<td>( T_0, \tilde{T} )</td>
<td>Set of all moveout vectors and set of all relative moveout vectors</td>
</tr>
<tr>
<td>( R )</td>
<td>Residual distribution between template moveout vectors and moveout vectors</td>
</tr>
<tr>
<td>( D )</td>
<td>Arrival-time dataset</td>
</tr>
<tr>
<td>( \tilde{D}(\mathbf{X}, i) )</td>
<td>Theoretical arrival-time dataset of source ((\mathbf{X}, i))</td>
</tr>
<tr>
<td>( S )</td>
<td>Candidate source dataset</td>
</tr>
<tr>
<td>( \mathcal{X} )</td>
<td>Set of all template vector locations, ( \mathbf{X}_k ) for ( k = 1, \ldots, P )</td>
</tr>
<tr>
<td>( \tau_{i,j} )</td>
<td>( j )th arrival time of station ( i )</td>
</tr>
<tr>
<td>( t_{i,k} )</td>
<td>( k )th triggering time of template ( k )</td>
</tr>
<tr>
<td>( r_{obs} )</td>
<td>Average rate of observed arrivals, per station, per day</td>
</tr>
<tr>
<td>( w_{i,j} )</td>
<td>Edge weights between ( r_{i,j} ) arrival and ( t_{i,j} ) triggering times (phase type ( r ))</td>
</tr>
<tr>
<td>( N, P )</td>
<td>( N ) stations, ( P ) templates</td>
</tr>
<tr>
<td>( A, b )</td>
<td>Constraint matrix and constraint vector of competitive assignment</td>
</tr>
<tr>
<td>( c, x )</td>
<td>Optimization vector and solution vector of competitive assignment</td>
</tr>
<tr>
<td>( \sigma, \gamma, \phi )</td>
<td>RBF-kernel width, triggering threshold, and cost term</td>
</tr>
</tbody>
</table>

RBF, radial basis function.

Computational complexity. Several approaches use heuristic-based search procedures that involve choosing a set of arrivals occurring nearby in time, removing arrivals with high misfit to the inferred best-fitting location to these picks and iteratively adding and removing new trial arrivals until convergence with a stable set of arrivals is obtained (Sippl et al., 2013; Gibbons et al., 2016). These techniques generally suffer from high computational cost and can scale with the squared or factorial of the number of arrivals considered (Johnson et al., 1997). Using waveforms (rather than arrival times), network template matching can be used to resolve the association step implicitly (Gibbons and Ringdal, 2006). Drawbacks of template matching include that templates must be known a priori, there is a high computational cost, and generally only earthquakes that are similar to a template are detected. A related class of techniques transforms raw seismic time series into characteristic functions such as analytical envelopes or STA/LTA traces before correlating the characteristic functions between stations to detect when a common source moveout crosses the network (for a review, see Li et al., 2017). Rule-based, fuzzy-logic, and “expert” systems have also been introduced (Bache et al., 1993), as well as a Bayesian-based probabilistic association method (Arora et al., 2013). Other recent developments in earthquake association include a waveform-based pairwise association technique built around the fingerprint and similarity thresholding method (Bergen and Beroza, 2018), a method using deep recurrent neural networks applied to arrival-time data (Ross et al., 2019), and a convolutional neural network approach for waveform-based pairwise associations (McBrearty et al., 2019).

Backprojection, beamforming, and source scanning are other methods for detecting earthquakes and implicitly associating arrivals (Kao and Shan, 2004; de Hoop and Tittelffitz, 2015; Nakata and Beroza, 2016). Although details vary, in general, these methods backpropagate waveforms from several stations into the source domain, and locations with high coalescence are identified as possible sources. Some variants replace full waveforms with finite-width kernels such as boxcars or scaled Gaussians centered about arrival times prior to implementing BP (Ringdal and Kvaerna, 1989; Weber et al., 2007; Friberg et al., 2010; Yeck et al., 2019). Several well-established software codes for automatic earthquake processing (e.g., Earthworm, SeisComP3, Early-Est, Antelope, and GLASS3) all use variants of BP of arrivals as a key component of their association routines. In these approaches, in general, high coalescence in the BP domain is used to identify candidate sources, and then a series of iterative processing steps are taken to identify the minimum number of sources necessary to explain the data. Although effective in many scenarios, BP-based techniques generally yield difficult-to-interpret results when event rates are high and earthquakes occur closely together in time. In our work, we seek to resolve this ambiguity by interpreting BP results in a graph theory context. We use the remainder of this section to describe BP (and its subtle nuances) in greater detail.

### Notation Conventions

We define some notation conventions for clarity. The physical source space is given by \( \Omega_X \subseteq \mathbb{R}^3 \), the origin time space is given by \( \Omega_t \subseteq \mathbb{R} \), and the combined spatial–temporal domain is given by \( \Omega = \Omega_X \times \Omega_t \). The travel time to station \( i \) from coordinate \( \mathbf{X} \) of phase type \( r \) is given by \( T_i(\mathbf{X}) \); the relative travel time with respect to the network is given by \( \tilde{T}_i(\mathbf{X}) = T_i(\mathbf{X}) - \min_{\mathbf{X} \in \Omega_X} T_i(\mathbf{X}) \). The set of all arrival-time vectors is given by \( \mathcal{T} = \{ T_i(\mathbf{X}) : \mathbf{X} \in \Omega_X \} \); the set of all relative arrival-time vectors by \( \tilde{\mathcal{T}} = \{ \tilde{T}_i(\mathbf{X}) : \mathbf{X} \in \Omega_X \} \). We let \( \mathcal{D} = \{ \tau_{i,j} \} \) represent the observed set of all \( j \)th arrivals on each \( i \)th station and let \( \tilde{\mathcal{D}}(\mathbf{X}, i) = \{ t + \tilde{T}_i(\mathbf{X}) \} \) represent a set of theoretical arrivals recorded across the entire seismic network, for each \( i \)th station and \( r \)th phase type, from a source at \( \mathbf{X} \); \( i \). A complete summary of the key notation used throughout the article is given in Table 1.

The BP Space

Arrival–time based BP may be summarized as the reverse time correction and stacking of recorded seismic arrivals by the relative theoretical travel time from a source to each station. When applied to a set of discrete arrival times from a seismic network \( \mathcal{D} \), let
Earthquake Arrival Association with Backprojection and Graph Theory

\[ C(X, t) = \sum_{\tau_{i,j} \in \Omega} \sum_{r} f(t - (\tau_{i,j} - T_i(X))) \quad (1) \]

denote the BP space, \( C(X, t) \), in which \( (X, t) \in \Omega \) represent possible spatial coordinates and origin times of sources. \( f(\cdot) \) represents any type of locally supported kernel (centered about zero) used to map discrete arrivals onto continuous space and time; for example, whenever \( |t - (\tau_{i,j} - T_i(X))| < \epsilon \) for small \( \epsilon \), \( f(\cdot) \) is nonzero and represents placing a kernel of some finite width along a manifold \( \mathbb{M} \subseteq \Omega \), highlighting possible spatial-temporal sources of arrival \( \tau_{i,j} \). The key feature of this mapping is that whenever a subset of arrivals in \( D \) are from true source-receiver(s) path(s) (e.g., \( \tilde{D}(X_{i,j}) \subseteq D \) for \((X_{i,j}, t_j)\) \( \in \Omega \)), then \( C(X, t) \) will have a significant degree of coalescence at \((X_{i,j}, t_j)\) because several manifolds will all intersect and constructively interfere at this coordinate (i.e., in equation 1, all of the related terms satisfy \( t_s - (\tau_{i,j} - T_i(X_s)) = 0 \) for \( \tau_{i,j} \in \tilde{D}(X_{i,j}) \)). Equation (1) is also written for the general case in which the directivity and phase types of arrivals are initially unknown, and each arrival is backprojected in all directions for all of the phase types under consideration.

It is insightful to note that although the global maximum of \( C(X, t) \) would identify the source location (and origin time) when \( D = \tilde{D}(X_{i,j}) \), elsewhere throughout \( \Omega \), a complex interference pattern is produced in which subsets of the backprojected manifolds intersect. The coordinates of partial coalescence can form isolated, local maxima in \( C(X, t) \), producing a topology in the BP space that may be complex for even simple arrival-time datasets of a single earthquake (Fig. 2b,c). The complex interference patterns of the BP space shown in Figure 2 are typical for earthquakes in the interior of a seismic network. In teleseismic applications or when the azimuthal coverage is poor, backprojected wavefronts are significantly more coplanar in the vicinity of the source location, reducing the number of subsidiary wavefront intersections.

As a revealing example, we consider a case in which two earthquakes occur nearby in time, with one earthquake producing 10 P-wave arrivals and the second earthquake producing seven P-wave arrivals across the IPOC network (Fig. 2a). Taken individually, equation (1) produces a BP space with well-defined global maxima (Fig. 2b,c) at each of the respective source coordinates. When the arrival datasets are combined, because equation (1) is a linearly superposable field with respect to \( \tau_{i,j} \), the interference patterns overlay one another and the identification of each individual source is more ambiguous (Fig. 2d). The global maximum remains essentially invariant (from Fig. 2b) at the location of the first source (with 10 arrivals); however, the location of the second source — although remaining a local maximum — now has a coherency value on the order of many of the other local maxima in this space. In this case, it is no longer evident whether or not it or one of the other local maxima is a true source when using only the BP space. This result is general: the coherency value of subsidiary modes (or local maxima) in the BP space may not be diagnostic of which local maxima are true; for example, a maxima in \( C(X, t) \) may be a false source yet have higher coherency than another maxima that is itself a true source. This latter case is common when one earthquake has more arrivals than another nearby earthquake because the secondary modes of the larger source may have higher coherency than even the primary maxima of the smaller source. We also note that the BP image space maps shown in Figure 2 are for the fixed correct values of depth and origin time, and in practice, the BP space has four degrees of freedom (latitude, longitude, depth, and origin time), each of which introduces additional structure, local maxima, and variability.

The BP Space for Theoretical Data

One helpful insight into the ambiguity of the BP space when multiple earthquakes are present in a short time interval is that equation (1) predicts a priori what the interference
pattern will look like throughout all of $\Omega$ for any given choice of $\mathcal{D}$. This type of information has utility in determining which of many local maxima in $C(X, t)$ may be true because it can indicate which set of local maxima are expected for any given source. If, for example, a source at $(X_s, t_r)$ and its associated dataset $\mathcal{D}(X_s, t_r)$ has a predictably large coherency value at a local maxima $(X_1, t_1) \neq (X_s, t_r)$ underneath the mapping (equation 1), we might be skeptical that the locally coherent point $(X_1, t_1)$ is a true source when compared against the point of higher coherency $(X_s, t_r)$. Conversely, if a local maxima existed at $(X_2, t_2) \neq (X_s, t_r)$ where the BP space is predictably low given $\mathcal{D}(X_s, t_r)$, we would have stronger evidence that $(X_2, t_2)$ may be a second, true source, in addition to that of $(X_s, t_r)$.

This type of duality between the joint likelihood of pairs of sources will be used in a later step of our method. Analogously to the theoretical dataset $\mathcal{D}(X_s, t_r)$, we define the theoretical BP space as

$$\tilde{C}_{(X_s)}(X, t) = \sum_{t_r \in \mathcal{D}} \sum_{D} f(t - (t_{i,j} - T_i(X))) ,$$

which is equal to equation (1) computed for the case in which $\mathcal{D} = \mathcal{D}(x, t = 0)$. We note that equation (2) assumes all of the arrivals of theoretical source $\tilde{X} = (x, t = 0)$ are contained in $\mathcal{D}(x, t = 0)$ and that equation (2) is symmetric with respect to arbitrary pairs of sources $\{(X_1, t_1), (X_2, t_2)\}$ in the sense that $\tilde{C}_{(X_s)}(X_2, t_2 - t_1) = \tilde{C}_{(X_s)}(X_1, t_1 - t_2)$ whenever the kernel $f(\cdot)$ is symmetric.

Association Methods

The primary objective of our method is to process the BP space produced by any of the complications considered (1–6) in the Introduction under the mapping (equation 1) and still infer the correct number of sources, the associations of arrivals to their correct source, and the identification of phase types and of false arrivals. The basic outline of our method is to detect coalescent points in the BP space with an efficient search procedure. With these detections, we assemble a mathematical graph which connects all coalescent points (candidate sources) to arrivals consistent with those sources. We then formulate and apply a constrained linear optimization on the graph itself, which imposes physical constraints on the allowed connections (called “edges” in graph theory parlance) between candidate sources and arrivals. This step enables us to extract the true sources from the full graph and discard the false coalescent sources. In this process, arrivals are assigned to their optimal sources, phase types of arrivals are determined, and most false arrivals are discarded. For numerical efficiency, we also describe a graph pruning technique that can be applied prior to the optimization routine, which is useful to keep the optimization problem tractable and allow the method to scale over continuous time (streams of arrival times) applications. The pruning routine is implemented by combining information contained in both the observed and theoretical BP space between pairs of sources into a mathematical graph that connects all sources to one another, and that allows for the rapid removal of the least optimal sources using the Markov clustering algorithm (MCA; Dongen, 2000). All of the steps of the method are outlined in the flowchart titled Algorithm 1.

Template Moveout Vectors

To detect all local maxima in $C(X, t)$ for an arbitrary dataset $\mathcal{D}$, one could theoretically compute $C$ over a dense grid search of discretized $(X, t) \in \Omega$ and robustly find all local maxima. This process, however, would be computationally inefficient. A more practical alternative is to select a small finite set of representative spatial BP coordinates, $\mathcal{X} = \{X_k\}$, and compute $C(X_k, t_r)$ for all $X_k \in \mathcal{X}$, and over all time $t \in \Omega$. Because absolute seismic travel times to a fixed station change only gradually as source coordinates change, the

Algorithm 1

The Association and Detection Method. This flowchart summarizes the essential steps to the method: (1) extracting template moveout vectors, (2) running discrete backprojection and detecting all candidate sources, (3) computing the source-arrival graph edge weights, (4) simplifying the source-arrival graphs by extracting disconnected components, running the graph pruning routine (MCA), running competitive assignment (CA) to compute the final set of detections, and (5) estimating earthquake locations, origin times, and magnitudes.

1: procedure INITIALIZE
2: Input: $\tilde{T}_i(X)$ for $X \in \Omega$
3: Choose $P$
4: Run K-Means to infer $P$ optimal $\{X_i\}$ template coordinates given $\tilde{T}_i(X)$
5: procedure run discrete backprojection
6: Input: $\mathcal{D}$ and $\{\tilde{T}_i(X)\}$
7: Choose $\sigma$ and $\gamma$ (or adaptively choose $\gamma$ given $D$)
8: Calculate $C(X_i, t)$ using equation (4)
9: Record all triggering times $(t_{ij})$ where $C(X_i, t) \geq \gamma$
10: Convert triggering times into source coordinates; $(X_{i}, t_{ij} - \min_{\ell} \tilde{T}_i(X) \in \Omega)$
11: Optionally optimize sources in $S$ in continuous $\Omega$ with local gradient ascent
12: procedure compute graph
13: Input: $\mathcal{D}$ and $\mathcal{S}$
14: For each source-arrival pair $(t_{ij}, t_{kl})$, calculate edge weights $w_{ijkl}$ using equation (5)
15: procedure simplify and solve each graph
16: Input: $w_{ijkl}$
17: Choose $r$ and $\phi$
18: Extract all disconnected subgraphs from $w_{ijkl}$
19: for all disconnected subgraphs do
20: Run MCA with parameter $r$ to simplify each subgraph
21: Run CA with parameter $\phi$ on the reduced subgraph and record solution
22: procedure characterize each earthquake
23: Input: All sources and associated arrivals returned from CA
24: for all earthquakes do
25: Relocate with probabilistic formulation (equation S4)
26: Estimate origin time (equation S5)
27: Estimate magnitude (equation S6)
space of all moveout vectors generally has low-order complexity (i.e., varies smoothly), which ensures that a set of optimally chosen representative coordinates \( \{ \mathbf{X}_k \} \) with their corresponding relative template travel-time vectors \( \{ \hat{T}_i' (\mathbf{X}_k) \} \) can capture the broad characteristics of the space of all relative travel-time vectors \( \hat{T} \). \( \& \) Figure S1 illustrates the essence of this partitioning approach, in which we subdivide the region surrounding the IPOC seismic network into eight subregions, in which for each subregion, a single moveout vector approximates the true moveout vector for any source within that subregion. In this case, eight subregions are shown for clarity, though in practice, the optimal number is generally in the hundreds, which produce many higher resolution subdivisions than shown in \( \& \) Figure S1.

To find a representative set of template vectors that optimally captures the main characteristics of \( \hat{T} \), we implement vector quantization (Vasuki and Vanathi, 2006) with the \( K \)-means clustering algorithm (Jain, 2010). \( K \)-means is an efficient algorithm that can identify \( P \) optimal cluster centroids (e.g., template vectors) by iteratively sampling vectors from \( \hat{T} \) and updating a set of template vectors \( \mathbf{v}_k \) \((k = 1, \ldots, P)\). The updates ensure that, statistically, arbitrary vectors in \( \hat{T} \) have at least one relatively similar template vector in \( \{ \mathbf{v}_k \} \). That is, the algorithm updates the templates to iteratively minimize the distribution of nearest-neighbor residuals

\[
\mathcal{R} = \{ \min_k || \hat{T}_i' (\mathbf{X}) - \mathbf{v}_k ||_{\text{rms}} : \mathbf{X} \in \Omega \} \tag{3}
\]

between vectors in \( \{ \mathbf{v}_k \} \) and those in \( \hat{T} \). As can be confirmed with empirical tests and by evaluating \( \mathcal{R} \) for arbitrary choices of \( \{ \mathbf{v}_k \} \), this type of clustering shows that \( \{ \mathbf{v}_k \} \) can be chosen more optimally than simply uniformly subdividing space. This is because more cluster centroids can locate in areas with high travel-time variability, whereas a sparse set of cluster centroids can occupy regions with highly similar sets of moveout vectors (e.g., in areas with poor seismic network azimuthal coverage). In practice, the template moveout vectors returned from \( K \)-means may not have an exact corresponding travel-time vector from any given \( \mathbf{X} \in \Omega \). However, \( \mathbf{X}_k \) can be taken as the nearest matching source coordinate to the template vectors; that is, \( \mathbf{X}_k = \arg \min_k || \hat{T}_i' (\mathbf{X}) - \mathbf{v}_k ||_{\text{rms}} \) for each \( k = 1, \ldots, P \).

In this application, the minimum travel time was subtracted from each travel-time vector prior to running \( K \)-means because \( \hat{T} \) has a simpler structure than \( T \), and only relative arrival times are important when applying BP techniques. We note that in practice, \( \hat{T} \) can either be known on a discrete grid (such as the case for most 3D velocity model raytracing algorithms) or can be sampled from continuous space during the clustering algorithms iterations if analytical formula of \( \hat{T}_i' (\mathbf{X}) \) are available. The fidelity with which the obtained template moveout vectors characterize all of space can be evaluated by estimating the residual distribution (equation 3), which can help guide the choice of \( P \) that should be used for any given application.

Discrete BP with Template Moveout Vectors

Rather than backproject to all of continuous \((\mathbf{X}, t)\), for efficiency, we backproject to the discrete templates located at \( \mathbf{X}_k \in \chi \), with the fixed template moveout vectors obtained from \( K \)-means clustering. This procedure, which we refer to as “discrete BP,” can be succinctly written as a slight modification of equation (1):

\[
C(\mathbf{X}_k, t) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[ \sum_{j} f(t - (\tau_{i,j} - \hat{T}_i' (\mathbf{X}_k))) \right] \tag{4a}
\]

\[
f(t) = \frac{1}{2} \exp \left( -\frac{t^2}{2(\sigma^2 + \sigma^2(\cdot))} \right) \tag{4b}
\]

for \( \tau_{i,j} \in D \) and \( \mathbf{X}_k \in \chi \).

The discrete BP algorithm stacks the backprojected finitely supported radial basis function (RBF) kernel \( f(\cdot) \), with standard deviation \( \sigma \), for all arrivals \( \tau_{i,j} \in D \), templates \( \mathbf{X}_k \in \chi \), and phases \( r \). The number of stations is given by \( N \), and the additional operator \([\cdot]^\perp\) is used to clip the maximum contribution to \( C_k(t) \) of any given station to a maximum of 0.5 for any given fixed template and phase type. This is useful to prevent an individual station from saturating \( C_k(t) \) when, for example, the station has many closely spaced false arrivals in \( D \). Here, we have written equation (4a) explicitly for the case of two phases (e.g., \( P \) and \( S \) waves), and the scaling is designed such that whenever \( D \) contains a set of \( P \)- and \( S \)-wave arrivals recorded across all of the network equal to a template moveout vector (i.e., \( \{ t_s + \hat{T}_i' (\mathbf{X}_k) \} \subset D \) ), then \( C_k(t) = 1 \). The \( \sigma \) width in equation (4b) is chosen to account for uncertainties in phase picking and travel-time calculations, and to account for the sparseness of template moveout vectors, which only sample \( C(\mathbf{X}, t) \) at the discrete \( \mathbf{X}_k \) coordinates. In general, reasonable \( \sigma \) values can be chosen by considering these uncertainties and then refined through trial and error, such as by computing \( \mathcal{C}(\mathbf{X}, t) \) for representative earthquakes and inferring the necessary level of smoothing required to localize sources in the BP space. For completeness, we have also included the optional term \( \hat{\sigma}(\cdot) \) in equation (4b) to indicate that RBF-kernel widths can adapt for different pairs of source coordinates and stations if uncertainties in travel times vary across the seismic network and monitoring region.

It is straightforward to compute equation (4a) for any arbitrary dataset \( D \). At this step in the method, we are interested in detecting all local maxima in \( C(\mathbf{X}_k, t) \) (referred to as candidate sources), regardless if many of the maxima are not true sources. To achieve this, we choose a hyperparameter \( \gamma \), referred to as the “triggering threshold,” and detect all local peaks in \( C_k(t) \geq \gamma \) using basic 1D peak-finding algorithms. Although the maximum value that can be obtained is 1, in general, it is more practical to choose a relatively low \( \gamma \). This will allow for the detection of sources at intermediate coordinates to any of the \( \mathbf{X}_k \), detection of sources with noisy arrival-time estimates (or uncertainties in travel-time calculations), and the detection of small magnitude sources that do
not produce arrivals on all stations. Because each template has a characteristic source location \( X_k \in \mathcal{X} \), the set of all local maxima in equation (4a) can be assigned equivalent sources in \( \Omega \); the set of all candidate sources is given by \[ S = \{(X_k, t_{k,l}) - \min_{i} T'_i(X_k)\} \] for each \( l \)th triggering time of the \( k \)th template \( t_{k,l} \). The time-offset \( \min_{i} T'_i(X_k) \) from the triggering time \( t_{k,l} \) is required because the template move-out vectors encode relative arrival times (i.e., have the minimum travel time subtracted), and the correction is needed to return the template triggering time back into the absolute time domain of origin times in \( \Omega \).

Calculating equation (4a) with a fixed kernel \( f(\cdot) \) and a finite set of template vectors (in which usually it is sufficient to set \( P < 1000 \)) makes discrete BP and the identification of candidate sources highly computationally efficient. This replaces the need for a dense grid search over the entire source domain. However, a valid criticism may be that computing equation (4a) only yields a sparse and suboptimal sampling of \( C(X, t) \) because in reality, none of the sources in \( S \) are actually at maxima in the continuous \( C(X, t) \) space (e.g., a maxima in \( C(X_k, t) \) restricted to \( X = X_k \) may be on the flank of a true peak in continuous \( \Omega \) space). Our experiments have shown that this effect on the remainder of our algorithm is relatively small as long as \( P \) is large enough; however, to circumvent this issue, we introduce an optional step that significantly reduces the resulting sensitivity to moveout templates. Namely, we take all candidate sources \((X_k, t_{k,l}) - \min_{i} T'_i(X_k) \in S \) and update their locations in continuous \( C(X, t) \) until a local maxima is reached using standard local gradient ascent. Such an approach is allowed because equation (4) is differentiable whenever \( \partial T'_i(X)/\partial X \) is available. To account for the general case in which \( T'_i(X) \) is known only on a discrete grid and \( \partial T'_i(X)/\partial X \) is unknown, we propose fitting \( T'_i(X) \) with a simple fully connected neural network that will ensure \( T'_i(X) \) is continuous and differentiable inside \( \Omega \) (for a demonstration, see the \( \mathbb{E} \) supplemental content). In this framework, locally optimizing sources in \( C(X, t) \) is efficient, especially when using automatic differentiation software (e.g., Paszke et al., 2017), and the computations can also be sped up with graphic processing units (GPU) acceleration.

Graph Construction of Source–Arrival Pairs

Following discrete BP (and the optional local gradient ascent), a set of candidate sources \( S \) has been obtained. As discussed in the Introduction, the relative coherency values of maxima in the BP space are not themselves sufficient to determine which sources are real or false (Fig. 2). Furthermore, the BP space has a strong interference pattern, and a fixed (low) triggering threshold \( \gamma \) is used to identify all candidate sources. This results in only a small subset of the total number of candidate sources in \( S \) being true sources. The key component of our method is to make use of more information that is available than simply the coherency value of maxima in the BP space. The main insight is that after the detection of all candidate sources, we know which (and by how much) each arrival contributed to the coherency value of each source. Namely, by inspection of equation (4a) and disregarding the scaling factor \( 1/N \), we see that for any pair \((\tau_{i,j}, t_{k,l})\), the arrival \( \tau_{i,j} \) contributed

\[
\begin{align*}
w_{i,j,k,l} = \frac{1}{N} \left( \tau_{i,j} - (\hat{T}'(X_k) - \hat{T}'(X_k)) \right) \\
= \exp \left( -\frac{(\tau_{i,j} - (\hat{T}'(X_k) - \hat{T}'(X_k)))^2}{2(\sigma^2 + \sigma^2)} \right)
\end{align*}
\]

to \( C(X_k, t_{k,l}) \) for phase type \( r \). Because this quantity can be computed for all pairs of arrival(s)–source(s), the set of all \( w_{i,j,k,l} \) can also be used to represent the edge weights of a mathematical graph that links all source–arrival pairs by a weight proportional to how much each arrival contributed to the detection of each source (which is also dependent on the phase type \( r \)). A large edge weight quantity for a source–arrival (and phase dependent) pair indicates the arrival constructively interfered with several other backprojected arrivals to elucidate the source in question. Conversely, when the edge weight goes to zero, the arrival did not contribute to the detection of the source with that phase type. Over continuous-time applications, a graph of this kind will be highly sparse because each arrival will only be connected to a small subset of the total number of candidate sources. We also note that equation (5) is written for the case in which sources have not been locally optimized in \( C(X, t) \) space; however, if they have been, the only change is that \( t_{k,l} \) becomes the new origin time, \( X_k \) becomes the new source coordinate in continuous \( \Omega \) (rather than being fixed at the original \( t_{k,l} \) and \( X_k \) values), and \( T'_i(X_k) \) is used instead of \( \hat{T}'(X_k) \).

Given the properties of the source–arrival graph outlined, it is clear such a graph will contain useful information for interpreting the results of BP (Fig. 3). For instance, for an arrival dataset of a single earthquake and its associated collection of candidate sources obtained from discrete BP \( \{D(X_{i,j}, S)\} \), the true source in \( S \) will have large edge weights to all of the arrivals in \( D(X_{i,j}) \), whereas each false candidate source will have large edge weights to only a subset of arrivals (each candidate source being connected to different subsets of the full set of all arrivals). Moreover, because each edge weight is recorded as a phase-dependent measurement, the true source will have large edge weights of phase type \( r = 1 \) to the \( P \) waves in the dataset and large edge weights of phase type \( r = 2 \) to the \( S \) waves in the dataset but not vice versa. In the Competitive Assignment section, we describe a constrained optimization routine that can be applied to arbitrary \( w_{i,j,k,l} \) graphs, which can extract these inferences directly, and which also generalizes to multiple earthquakes being contained in the dataset and other complicating factors such as missing arrivals and false arrivals.

Competitive Assignment

When viewed on short time scales, the source–arrival graph \( w_{i,j,k,l} \) generally has a highly entangled set of edge
connections because any number of arrivals may coalesce to cause a source to trigger (i.e., to exceed the threshold for declaring a candidate source), and each arrival may itself contribute to several candidate sources. Our optimization scheme imposes constraints on the connections in $w_{ijkl}$ that force arrivals to be linked to at most one candidate source (using only one phase type), while maximizing the number of source–arrival edges retained and applying a negative penalty for all retained sources from $S$. This causes candidate sources to compete with others to obtain the most connections as possible and results in many candidate sources being discarded, so that only an optimal (small) suite of inferred true sources are retained. In the trivial case of a single earthquakes dataset and its collection of candidate sources, such an optimization would result in the one true source “winning” the connections to each arrival, and all of the false candidate sources being discarded because this would maximize the number of assigned arrivals while minimizing the number of active sources. The success obtained on the trivial problem, however, extends to more complicated cases by use of the full optimization algorithm, referred to as “competitive assignment,” which we now define.

The optimization problem is defined as a constrained integer linear programming problem (ILP) applied directly to the graph edge weights contained in $w_{ijkl}$. ILPs are a type of linear optimization problem in which the solution vector $x$ is composed of all integers (Nemhauser and Wolsey, 1988; Schrijver, 1998) and in this case is binary and has an entry for...
each edge in the graph specifying whether to keep \((x_i = 1)\) or delete \((x_i = 0)\) that edge. The optimization seeks to maximize \(c^T x\), in which vector \(c\) is composed of all of the edge weights in \(w_{ijkl}^c\), with a constraint matrix \(A\) acting on \(x\) that ensures that:

1. each arrival is assigned to (associated with) at most one source and one phase type,
2. each station contributes at most one of each phase type to each source, and
3. a user-specified penalty \(-\phi\) is applied for each source that retains \(>0\) connections.

The competitive assignment problem can then be written succinctly as

\[
\max_{x} \ c^T x \tag{6a}
\]

\[
\text{such that } Ax \leq b \tag{6b}
\]

\[
x \in \{0, 1\} \tag{6c}
\]

\[
\{A, b, c\} \xrightarrow{\text{Algs. S1--S3}} \{D, S\}, \tag{6d}
\]

in which the precise development of the constraints \(\{A, b\}\), optimization vector \(c\), and how to read the result from the binary solution vector \(x\) only involve an indexing scheme. The scheme can be automated for any given dataset \(D\) and set of candidate sources \(S\). The notation in equation (6d) indicates that \(\text{Algs. S1--S3}\) summarize the steps of this scheme, and details may be found in the \(\text{Algs.} S1--S3\) supplemental content. The binary solution vector \(x\) obtained from maximizing equation (6a) encodes the assignment of all arrivals to at most one source, the phase type of their assignment, and the active and inactive state of all candidate sources. Arrivals not assigned to a source are flagged as possible false arrivals. The solution represents a structured and physically realistic subgroup that is extracted from the input graph that maximizes the number and strength of all source–arrival connections retained while minimizing the number of active sources needed to do so.

The competitive assignment algorithm can succeed for even complex arrival and candidate source datasets (and their associated graphs) for intuitive reasons. If multiple earthquakes are present in \(w_{ijkl}^c\), optimizing equation (6a) will encourage the identification of all of the true sources because doing so will enable the assignment of the most arrivals. Because each arrival can only be assigned to one source, false sources are unlikely to be retained because all of the arrivals that originally caused constructive interference at the false sources in \(C(X, t)\) will instead be assigned to their true source, and the cost penalty \(-\phi\) will outweigh the benefit of keeping the false sources (even if, for instance, a few arrivals happen to be left over and could otherwise still be assigned to those false sources). Because the edge weights are measured as phase-dependent quantities, assigning each arrival to their true source with the correct phase assignment will maximize equation (6a) more than if an incorrect phase assignment is used. In addition, underneath the BP mapping, most false arrivals will not align with the moveout of a true source (as long as \(\sigma\) is not too large), and hence (true) source–false) arrival edge weights will be predominantly zero, and the majority of false arrivals will not be assigned to any source.

In Figure 3, the graph construction and ILP optimization steps are shown for three representative candidate sources (local maxima) detected in the BP space for the same example of two earthquakes shown in Figure 2. In this example, the ILP optimization resolves the ambiguity of the BP space (Fig. 3a) to determine the correct number of sources and the phase assignments of each arrival. Although the arrivals are all treated as having an unknown phase type during the BP step, the solution determines that all observed arrivals are \(P\) waves. Notably, the false candidate source (marked C in Fig. 3a) has higher BP coherency than a true source (marked B in Fig. 3a), yet the optimal solution of the ILP correctly identifies source C as being false and sources A and B as being true. This is possible because the simultaneous joint inversion for all source–arrival assignments reveals that the maximum number of source–arrival assignments is achieved by retaining sources A and B rather than retaining source C (in combination with either source A or B) and letting source C incorrectly split the arrival assignments between the two earthquakes.

Continuous-Time Applications and Computational Efficiency

The number of candidate sources triggered in discrete BP has a strong effect on the computational efficiency, as well as the memory requirements of the method. As the duration of the dataset \(D\) increases and more earthquakes are included, inevitably, the number of triggered sources increases without bound and practical questions surrounding how to apply the method to continuous-time applications (without artificially subdividing \(D\)) arise. A related consideration is whether any of the hyperparameters used in the method should change as seismicity conditions vary. To address both issues and maintain as automated a method as possible, we introduce a few modifications and intermediate steps (all of which are outlined in Algorithm 1 for reference). Specifically, we adaptively choose the triggering \(\gamma\) and cost \(\phi\) hyperparameters as the rate of observed seismicity (over a time window) changes, and we also introduce a few graph-based tools to reduce the complexity of the graphs passed to the competitive assignment optimization routine. The former ensures that increased rates of seismicity do not blow up the number of false candidate sources triggered, whereas the later steps are beneficial for keeping the graphs competitive assignment is applied to small and tractable. Most ILP solvers have computation time that scales with around \(O(n^3)\) for \(n\) nodes, and it is thus computationally advantageous to apply competitive assignment to as sparse and small of graphs as possible.

Adaptive Triggering Threshold \(\gamma\) and Cost Term \(\phi\). During standard operational settings, choosing a fixed triggering
threshold $\gamma$ for all time intervals is usually appropriate. However, when seismicity rates change during earthquake swarms or aftershock and foreshock sequences (among other possible situations), the number of triggered sources may increase with $O(e^n)$ for $n$ earthquakes and soon become computationally prohibitive. The cause of this is that the base level of the $C(X, t)$ space increases as more spurious arrivals are included in $D$. To maintain a computationally efficient algorithm, it is helpful to increase $\gamma$ proportionally to the background coherency level in $C(X, t)$, such that the triggered candidate sources still distinctly stand out above the background level of coherency.

To implement a simple adaptive choice for $\gamma$, we choose $\gamma$ to be linearly proportional to the average rate of arrivals occurring across the network in a time interval. That is, we set

$$\gamma = \frac{1}{N} (m_r r_{obs} + b_r),$$

in which $r_{obs}$ is the average rate of arrivals per station, per day, and the coefficients $m_r$ and $b_r$ are chosen to maintain a balance between accuracy and efficiency. In our application, we have found that even a small linear scale of $(m_r, b_r) = (0.00122, 4.9)$, allowing $\gamma$ to adapt between $6/N$ and $7/N$ as $r_{obs}$ changes between 900 and 1800 is sufficient to allow the method to run for either modest or high rates of seismicity with little changes to the resulting computation time or memory cost of the method. After $\gamma$ is decided, we set the cost as $\phi = N\gamma - 0.5$ in all cases. The window size over which to compute $r_{obs}$ is subjective; however, in our application, we compute $\gamma$ over daily seismicity rates.

Disconnected Subgraphs. Mathematical graphs can be separated into distinct subgraphs by searching for populations of nodes that are entirely disconnected from other populations of nodes. When $D$ includes arrivals spanning a time interval many times longer than travel times allow, typically $w_{ijkl}$ will be highly sparse and contain subsets of nodes that are disconnected from one another (as long as $\gamma$ is not too small). Because these divisions already separate groups of sources and arrivals from one another naturally, we extract all such disconnected subgraphs from $w_{ijkl}$ and run the competitive assignment and graph pruning subroutine (described next) on these subgraphs individually.

Graph Pruning with the MCA. To improve the efficiency of competitive assignment (which may become slow when an input graph has, e.g., >50 sources $\times$ 150 arrivals), we use a graph-theory-based tool (MCA) to discard a number of suboptimal sources contained in $S$ prior to running competitive assignment. This reduction is achieved by identifying which sources in $S$ are most likely subsidiary local maxima to different, more optimal sources in $S$. This information can itself be revealed by a particular consideration of the observational BP space $C(X, t)$ and theoretical BP space $\tilde{C}(X, t)$ for a given dataset $D$ and set of candidate sources $S$.

To explain the relationship between MCA, $\tilde{C}, C, S,$ and $D$, we briefly describe the essence of the MCA algorithm (and also define it precisely in Algorithm 8). MCA is a clustering algorithm applied to stochastic matrices. Stochastic matrices are matrices that define the discrete transition probabilities between any two nodes in a graph when following a random walk. On a graph, a random walk is the trajectory a particle takes when iterating a random sequence of transitions between its current node location to any of the other nodes it has edge connections with, in which the individual discrete transition probabilities to neighboring nodes are equal to the edge weights between those nodes. The purpose of MCA is to identify subsets of nodes (or node populations) in a graph that are more interrelated than others, as encoded by the structure of edge weights describing the transition probabilities between the nodes in the graph. Intuitively, if two populations of nodes have many interconnecting edges and only a few weak connections between the two populations, then it is apparent that random walks spend a significantly higher proportion of time walking among either cluster than compared with transitioning between the different clusters. MCA is simply an efficient analytical method that can be applied to stochastic matrices to extract these types of inferences directly (however, it generalizes to any number of clusters). In addition, it also identifies the centroid of each cluster, that is, the fixed-point node that the rest of the nodes in the cluster are most likely to visit when following a random walk.

This algorithm can be applied to identify which sources in $S$ are suboptimal and enable the simplification of $S$. The stochastic matrix we use in MCA is a source–source indexed graph and is given by

$$H_{ij} = \tilde{C}(X, t_j) - C(X, t_i),$$

which defines the transition probabilities to account for two principles: (1) the transition probabilities are high between pairs of sources $(X, t_i), (X, t_j) \in S$ whenever the theoretical pairwise BP space value of these sources is high (which is a symmetric quantity and observation data independent), and (2) the transition is asymmetrically skewed toward the source of higher observational BP coherency. This type of stochastic matrix encourages transitions between sources that are intrinsically likely to be related (as defined by $\tilde{C}$), while biasing transitions away from sources of low coherency toward sources of higher coherency (given by $C$). These properties ensure that cluster centroids returned from MCA should be of higher coherency than other sources contained in $S$ that they are intrinsically related to, given the theoretical BP space. In practice, MCA also uses a hyperparameter $\epsilon$ to help control how coarse or fine the resulting clusters will be. Because we use MCA only to decrease the size of the graphs (rather than solve the detection problem completely), in general, we choose a high value of $\epsilon$ (e.g., $\epsilon = 5$), which promotes returning many clusters, and we discard all sources that are not returned as a cluster centroid. Given the nature of the graphs we typically encounter, even a single earthquake
tends to return several cluster centroids; however, for some well-isolated earthquakes, at times, only a single (correct) cluster centroid (or source) for a single earthquake is returned. In either case, we run competitive assignment on the returned reduced graphs because competitive assignment can enforce physical constraints, whereas MCA cannot.

Waveform Arrival Picking: Norm Picker

We use a single-station energy-based picking method to pick arrivals from raw seismograms in our application to data from the IPOC network. We call the algorithm the “Norm Picker,” and it is designed to reveal onsets of energy much like the STA/LTA (Allen, 1978). The picking method is modeled after a similar one used in Delorey et al. (2015), and is a technique that steps through waveform data over time by a small time-step, and at each time-step, it (1) slices a finite window \( w_1 \) of data from the mean of the envelope of all three components, (2) normalizes the slice by the root mean square (rms), (3) smooths the normalized data by a small moving average time window \( w_2 \), and (4) stacks the slice back onto its corresponding time interval from which it was sliced. Finally, a (5) moving window \( w_3 \) fit of the derivative of this time series is taken, in which local maxima above a threshold are marked as arrival picks. The algorithmic steps are defined in Figure S2. We used the Norm Picker algorithm in our application to northern Chilean seismicity primarily because of its ease of use and its ability to pick relatively low-SNR \( P \)- and \( S \)-wave arrivals; however, other recent developments in arrival picking (e.g., Bailleul et al., 2013; Ross et al., 2018; Zhu and Beroza, 2018) could also be used with our association technique.

The reason this picking algorithm is effective at detecting low-SNR arrivals, as well as being relatively insensitive to coda waves masking \( S \)-wave arrivals, is that it adapts a local view of the seismicity depending on how much energy is arriving at any given time. By normalizing the envelope of a waveform by its rms, in which \( w_1 \) is usually taken a few times longer than the duration of a typical earthquake, a high-SNR arrival will be normalized to a similar dynamic range as a low-SNR arrival will be. When a low-SNR arrival is contained inside the window \( w_1 \), it will be amplified by the rms normalization, helping enable its detection. The last step, which takes a moving window fit to the derivative of stacked normalized envelope waveforms, ensures that only when energy is sustained over a time interval on the order of \( w_3 \) will it have a high derivative value, which eliminates detecting spurious spikes of energy that do not have a substantial duration in time. Furthermore, the derivative detects the onset of energy, which helps localize the arrival time.

Earthquake Characterization

After arrival associations and phase types have been determined, earthquakes can be relocated with standard techniques (Tarantola and Valette, 1982; Lomax et al., 2000). In the observational settings that pertain to the IPOC network, we detect earthquakes that are nearby, as well as distant to the seismic network (e.g., \( >250 \) km distance), including many earthquakes at depths \( >200 \) km. As a result, we found that the IPOC network is poorly suited to accurately constrain earthquake locations, particularly along the plate interface to the east of the network (even when both \( P \)- and \( S \)-wave detections have been made), largely because the earthquakes are far away from the seismic stations and the velocity model is poorly known in this region. For the purposes of developing a catalog that remains comparable to existing works (Bloch et al., 2014; Kato and Nakagawa, 2014; León-Ríos et al., 2016; Sippl et al., 2018), we chose to relocate events in the NonLinLoc Bayesian framework (Lomax et al., 2000), in which we implement a prior on the source locations to bias locations closer to the plate interface, which is consistent with the predominant distribution of earthquake locations reported previously (Sippl et al., 2018). Specifically, we define the (unnormalized) prior likelihood of source locations as

\[
\theta(X) = \exp\left(-\frac{||X - X_{\text{plate}}||^2}{2\sigma_X^2}\right),
\]

in which \( \sigma_X \) is set as 15 km, and \( X_{\text{plate}} \) represents the nearest point on the Slab 1.0 Earth model of the Nazca plate (Hayes et al., 2012) to the coordinate \( X \). The \( \cdot \cdot \) operator in equation (9) represents Euclidean distance between coordinates in a Cartesian (rather than elliptical) coordinate system.

Following estimating earthquake locations with maximum-likelihood estimation (equation S4), earthquake origin times are estimated with analytical formula (equation S5), and magnitudes are estimated with a local magnitude scale (equation S6) calibrated to produce magnitudes comparable with the Chilean Centro Sismológico Nacional (CSN) catalog. As shown in the Application: Northern Chile Earthquake Catalog section, the use of a prior succeeds in resolving the intrinsic uncertainties of locations (primarily of depths) to the east of the seismic network while also not having an overwhelming influence on resulting locations in other regions, as we still detect many shallow earthquakes located near the surface, away from the plate interface, in locations previously reported to have seismicity.

Synthetic Tests

We conduct a set of synthetic tests to evaluate the performance of our method under a range of complexity cases. Specifically, we test the performance of the algorithm as the rate of earthquakes \( (r_e) \), the rate of false arrivals \( (r_f) \), and the uncertainty on arrival-time picks \( (\sigma_r) \) all change. To generate synthetic arrival-time datasets \( D \) that resemble true observational settings, we use the following procedure:

1. Choose \( r_e, r_f, \) and \( \sigma_r \).
2. For a time window \( T \), sample the number of earthquakes from a Poisson distribution with rate \( r_e \times T \).
3. Sample earthquake hypocenters from $\theta(\mathbf{X})$.
4. Sample earthquake origin times from a Uniform distribution over interval $[0, T]$.
5. Calculate theoretical times of all arrivals and add additional error sampled from a Laplace distribution with mean 0 and scale $\sigma_a$.
7. For each earthquake, station, and phase–arrival pair, compute the predicted empirical amplitude of each arrival (using equation S6), and if below a threshold $A_{\text{min}}$, delete the arrival from that station.
8. For each station, sample the number of false arrivals from a Poisson distribution with rate $r_f \times T$, and sample random false arrival times from a uniform distribution over interval $[0, T]$ for each.

These steps create realistic arrival-time datasets that easily adapt between complex and simple cases by changing $r_e, r_f,$ and $\sigma_a$. In step (7), we use the predicted attenuation of sources to stations to control if arrivals are detected on a station to simulate the real situation in which small earthquakes only produce arrivals on a subset of the nearest stations. The threshold for discarding arrivals $A_{\text{min}}$ is chosen such that ~30% of all arrivals are missed. For our purposes, we use the same prior $\theta(\mathbf{X})$ applied to locate earthquakes to generate the locations of the synthetic sources, which promotes sources that locate nearby the plate interface. It is worth noting that the earthquake origin times and locations are chosen independently, and there is no enforced minimum separation of sources in time or space, which enables many sources to produce overlapping moveouts across the network (Fig. 4).

To run the synthetic tests, we set the hyperparameters to the same values used in our application to the real data. We set the RBF-kernel width to $\sigma = 2.8$ s and let the triggering threshold $\gamma$ follow the linear scaling curve of $\gamma = \frac{1}{5}(m_r r_{\text{obs}} + b_f)$, in which $(m_r, b_f) = (0.00122, 4.9)$ to adaptively adjust for the rate of observed arrivals $r_{\text{obs}}$ and keep the memory requirements of the method low. The cost term in competitive assignment is always set to $\phi = N_T - 0.5$. For the experiments, we set $T = 86,400$ s (one day) and vary $r_e, r_f \in [100, 300, 500, 700]$ events/day and $\sigma_a \in [1.0, 1.5, 2.0, 2.5]$ s.

Evaluation

The evaluation of performance is split into several components. For earthquake detections, we count the number of true detections (true positives), missed detections (false negatives), and false predicted sources (false positives). From these values, we compute the precision ($P$), recall ($R$), and $F1$ quality metrics (Powers, 2011), in which precision is given by the ratio of true positives to true positives plus false positives, recall is given by the ratio of true positives to true positives plus false negatives, and $F1$ is a summary statistic given by $F1 = 2 \times (P \times R)/(P + R)$. True detections are marked for all synthetic sources for which the nearest predicted moveout curve is within 6.5 rms residual of the true moveout curve. False negatives are marked when the nearest predicted moveout curve to a synthetic source has a residual greater than 6.5 rms. False positives are marked for all other sources we predict that are not matched to a true source. Moveout curve comparisons are used instead of thresholds on nearest location and origin times (between true sources and predicted sources) because moveout curves combine information from both locations and origin times into a single vector, and they are more closely related to the data themselves rather than the specifics of network geometry and source distribution. An additional caveat to the definition of false negatives is that we do not count an earthquake as missing if the attenuation laws in step (7) result in an earthquake having less than $2 \times \gamma + 1$ arrivals because in this case, the number of arrivals is so few that we do not expect the method to detect the event anyway (recall that each arrival can only contribute at most 0.5 to the discrete BP metric). This usually occurs for at most ~25% of the sources in any of the synthetic catalogs. It would be simple and perhaps clearer to delete these sources entirely; however, keeping them in the dataset makes the detection problem harder when these arrivals coincide with other larger magnitude sources, which is a phenomenon that happens in the real data and is thus worth including in these tests. We also count the number of correctly associated $P$ and $S$ waves and the number of false arrivals correctly identified as false versus those mistakenly assigned to an earthquake.

Results

The results of the synthetic tests are summarized in Tables 2 and 3 and reported in full in Tables S2 and S3. We find that as $r_e$ and $\sigma_a$ are increased, the $F1$ detection accuracy statistic varies between 0.99 and 0.93, when taking the median result over all $r_f$ values tested. $r_f$ is found to have a relatively small effect on detection results (Table S2), causing at most a 0.03 variation in $F1$ over the range $r_f \in [100, 300, 500, 700]$ false arrivals/station/day. In contrast, $F1$ can vary by up to 0.05 and 0.04 units as $r_e$ and $\sigma_a$ vary, respectively, over their testing ranges $r_e \in [100, 300, 500, 700]$ events/day and $\sigma_a \in [1.0, 1.5, 2.0, 2.5]$ s, in most cases, precision is systematically higher than recall (on average by 0.03 units), indicating that the most common type of error is to miss an earthquake rather than create a false source. Example synthetic earthquakes, along with the solutions obtained, are shown in Figure 4. In this example, several earthquakes occur very closely in time, with many overlapping moveouts, yet the solution still resolves this ambiguity. In Figure S3, an additional detection example is shown, in which additional randomly missed arrivals are also included in the data.

The resulting accuracy of classifying phase types and identifying false arrivals shows higher variability than the earthquake detection statistics over the range of testing parameters. We find that the proportion of correctly associated $P$ waves, $S$ waves, and false arrivals vary between 0.99 and 0.82, 0.97 and 0.78, and 0.99 and 0.94, respectively, when
Table 2
Summary of the Synthetic Test Results for Event Detection

<table>
<thead>
<tr>
<th>$r_e$</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>$\sigma_a$</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.99</td>
<td>0.97</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>300</td>
<td>1.0</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>500</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.96</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>700</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Precision, recall, and F1 of earthquake detection accuracy are shown as the rate of earthquakes ($r_e$) and arrival-time uncertainty ($\sigma_a$) change. All values are the median value over all false arrival rates ($r_f$) tested. Full set of all results is given in Table S2 (available in the supplemental content to this article).
taking the median value over all $r_f$ values tested. For classifying $P$ and $S$ waves, $r_e$ and $r_f$ cause variation of up to 0.06 and 0.05 units, respectively, whereas $\sigma_a$ has the largest effect and can induce changes of up to 0.17 units. This finding, that phase association accuracy is most affected by $\sigma_a$, results because if observed arrival times are too far from their theoretical arrival time, the graph edge weights linking them to their correct sources are often zero. In all cases, $P$ waves are more accurately identified than $S$ waves (on average by 0.03 units). False arrival identification accuracy changes by up to 0.06, 0.01, and 0.03 units as $r_e$, $r_f$, and $\sigma_a$ all vary, respectively, over their testing ranges. Notably, the rate of false arrivals $r_f$ has the least effect on the ability to accurately identify false arrivals, whereas $r_e$ has the greatest affect. We interpret the large effect $r_e$ has on false arrival identification accuracy to be that increased numbers of sources make the likelihood for spurious false arrival associations more common.

Application: Northern Chile Earthquake Catalog

To demonstrate our method, we apply it to 8 yr of data from the IPOC network between 1 January 2010 and 31 December 2017. We use data from 17 broadband seismic stations of this network ([1] Table S1) and run the Norm Picking algorithm ([1] Algorithm S9) on the raw seismic data to develop a set of arrival picks before using the association algorithm. Before making arrival picks, seismic data are band-passed between 5 and 22 Hz to help remove teleseismic arrivals and enhance the detection of small microseismic earthquakes. As in the synthetic tests, the number of templates in discrete BP is set as $P = 500$, the RBF-kernel width as $\sigma = 2.8$ s, and the triggering threshold as $\gamma = \frac{1}{N} (m_\gamma r_{\text{obs}} + b_\gamma)$, in which $(m_\gamma, b_\gamma) = (0.00122, 4.9)$ and $\sigma_{\text{obs}}$ is calculated as the average number of arrivals per station per day. The cost term is always set as $\phi = N\gamma - 0.5$.

The source region we consider is between [26.5° S, 16.5° S], [72.5° W, 66.5° W] and depths [5 km, −250 km]. For a velocity model, we use the trench-perpendicular 2D $V_p$ transect of model I2 given by Comte et al. (1994), which we extrapolate north and south throughout the study region. We further assume that $V_S$ scales by 1.76 to $V_p$ (Comte et al., 1994). Given this velocity model, we calculate $T^*_f(X)$ on a grid of (125, 85, 75) (latitude, longitude, depth) elliptically distributed points within the source region using the shortest path raytracing method (Moser, 1991; Bai et al., 2007). A two hidden-layer fully connected neural network is fit to model the travel-time grid ([1] Fig. S4) to ensure that $T^*_f(X)$ is differentiable and continuously known between the original grid points.

Results and Validation

The newly developed catalog we obtain contains a high rate of seismicity and contains 817,548 detections. We detect on average 279 earthquakes/day (Table 4) and estimate a magnitude-of-completion of $M \sim 1.8$. This is in contrast to previous works, in which 35 earthquakes/day were detected in the most complete catalog to date (Sippl et al., 2018) and 7 earthquakes/day are reported in the operational CSN catalog. Of all the detections, we find that at least ~100,000 (or ~12%) of the events in the catalog occur close enough together in time to produce overlapping moveouts across the network.

The accuracy of the detections in our catalog is validated in several ways. We (1) confirm the broad distribution of earthquakes is largely consistent with previously reported catalogs, (2) verify that we redetect a majority of the earthquakes in the CSN and Sippl et al. catalogs, (3) visually assess the accuracy of a set of representative earthquakes, (4) ensure the catalog contains the foreshock and aftershock sequences of the 1 April 2014 $M_w$ 8.2 Iquique earthquake that have previously been reported (Kato and Nakagawa, 2014; Kato et al., 2016), and (5) compare the GR curves between this studies catalog and those of CSN and Sippl et al.

A representative distribution of 16,343 earthquakes from a two month interval (1 January 2015−1 March 2015) is shown in Figure 5, which shows a spatial distribution broadly in agreement with the CSN and Sippl et al. catalogs ([1] Fig. S5). In particular, the primary bands of trench parallel seismicity, with a transition zone between ~50–80 km depth is apparent, and the clustering of deep seismicity (>125 km) is consistent with that found in both catalogs. Earthquakes at depth largely follow the trend of the subducting Nazca plate (Fig. 5b). Some events, however, also locate nearby the surface, with even fewer events locating deep in the mantle to the west. Because of the probabilistic

<table>
<thead>
<tr>
<th>$r_e$</th>
<th>$P$ Waves</th>
<th>$S$ Waves</th>
<th>False Arrivals</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.99</td>
<td>0.96</td>
<td>0.85</td>
</tr>
<tr>
<td>300</td>
<td>0.98</td>
<td>0.95</td>
<td>0.80</td>
</tr>
<tr>
<td>500</td>
<td>0.97</td>
<td>0.94</td>
<td>0.83</td>
</tr>
<tr>
<td>700</td>
<td>0.97</td>
<td>0.92</td>
<td>0.82</td>
</tr>
<tr>
<td>$\sigma_a$</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Proportion of correctly associated $P$ waves, $S$ waves, and correctly identified false arrivals are shown as the rate of earthquakes ($r_e$) and arrival-time uncertainty ($\sigma_a$) change. All values are the median value over all false arrival rates ($r_f$) tested. Full set of all results is given in [1] Table S3.
formulation used to locate earthquakes (Equation S4) that incorporates the prior $\theta(X)$, earthquakes in either of these locations are easily identified by having large negative log-likelihood values (Fig. 5). The negative log likelihoods of sources in these regions are intrinsically high (implying a low-likelihood value) primarily because the prior likelihood is very small far away from the plate interface.

By examination, we find that sources occurring nearby the surface largely cluster nearby open pit quarries and industrial sites (Fig. 6a,b). The sources occurring nearby quarries have a nonuniform distribution of origin times with respect to the local time of day (Fig. 6c) and show clustering between 12 and 9 p.m. (UTC − 8), which is consistent with the predominant blasting times of quarries as reported in previous studies (Sippl et al., 2018). By visual inspection, we have also identified that the anomalous sources occurring deep in the mantle to the west are generally true earthquakes (actually occurring deep in the mantle to the east) that simply have initial locations following discrete BP to the west because the north–south alignment of the IPOC network results in a strong symmetry, and thus ambiguity, between moveout vectors on either the east or west side of the IPOC network. The initial locations are also far enough from the nonzero regions of $\theta(X)$ that the gradient-based relocation routine cannot relocate the sources (as the gradients are always zero). In a future run of the catalog development, we expect that we can account for this discrepancy and relocate these sources by either restricting all templates $X_k$ to nonzero regions of $\theta(X)$ or by estimating maximum-likelihood earthquake locations with a global (rather than local) search procedure. Nonetheless, the fraction of events locating in this region is small enough (<0.5%) that we do not believe it impacts the majority of our findings.

We compare the three catalogs directly in terms of matched earthquakes and count an earthquake pair as a match (between our catalog and the other catalogs) when the rms residual between the nearest matching moveout curve is <6.5 s, in an analogous way as done in the synthetic tests. Doing so shows that we recover >98% of the events reported in both the CSN and Sippl et al. catalogs (Table 4). In addition, for all matched earthquakes, we calculate the residual differences between source coordinates, origin times, and magnitudes. The residuals are relatively small, with means near zero and standard deviations of ~0.08° S, 0.18° W, 21 km, 1.28 s, and 0.295 units for latitude, longitude, depth, origin time, and magnitude residuals, respectively (Table S4). Most notably, the largest discrepancy is that for deep earthquakes, we locate events ~23 and ~13 km shallower than the CSN and Sippl et al. catalogs, respectively, which we infer is caused by using different velocity models between the studies and potentially by our use of an elliptical (rather than spherical) Earth model.

We visually assess the detection quality of a representative set of 300 earthquakes on 20 January 2014 and assigned each a quality of (1, 2, or 3), where to our best judgment, 1 = acceptably accurate, 2 = approximately accurate with some ambiguous arrival assignments or phase classifications, and 3 = likely incorrect set of associations and sources. Carrying out this procedure, we found we had a quality distribution of (85%, 11%, 4%) for qualities (1, 2, 3), respectively, in which quality 3 cases nearly always occurred when two or more earthquakes were split incorrectly. By inspection, we found that the seismicity distribution present on 20 January 2014 is largely consistent with the seismicity distribution found throughout the catalogs duration, and we do not expect significant deviations from these quality distributions during other time intervals.

We further confirm that the foreshocks and aftershocks of the 1 April 2014 $M_{w}$ 8.2 Iquique earthquake are detected by our catalog. We detect strong clustering of seismicity near the $M_{w}$ 8.2 hypocenter for several weeks preceding and following the rupture (Fig. 7a,b). The number of detections in the restricted interval of [21.5° S, 18.5° S], [72° W, 69.5° W], and depths [5 km, −75 km] exceeds 600 detections per day during the highest rates of seismicity. These findings give confidence that dynamic earthquake sequences (i.e., foreshocks and aftershocks) do not impede the performance of our method, and we are capable of detecting events during both highly activate and standard background seismicity intervals with no modifications to the method. A representative set of the aftershock detections along with an inset of waveforms is shown in Figure 8.

Last, we note that our improved rates of detections compared with previous works results in a GR distribution with lower magnitude-of-completion by ~0.8–1 magnitude unit (Fig. 7c). Above the magnitudes-of-completion of the different catalogs, the GR curves of Sippl et al., CSN, and our own are largely parallel, indicating a largely consistent set of common detections between the three. To make these GR comparisons fair, we calculated the GR curves using only earthquakes that occurred in spatial and temporal regions that all of the catalogs covered. Specifically, they were computed over the region [25° S, 18° S], [71.5° W, 66.5° W], depths [5 km, −250 km], and years [2010, 2014]. Furthermore, magnitudes of Sippl et al. and our own are offset by −0.108 and −0.130 units, respectively, to align the GR curves with the CSN catalog between the 3.5 and 5.5 magnitude unit interval.

Computational Cost

In this application, we found that using 20 cores of an Intel Xeon 2.40 GHz CPU, which parallelizes over the discrete BP, local gradient optimization, and competitive assignment routines (with no GPU acceleration), we process a day’s worth of seismicity in ~135 s using an average of ~35 GB of RAM during processing. On average, each day, 11,000 arrivals across the seismic network are processed, resulting in detecting ~273 earthquakes/day and discarding two thirds of all arrivals originally detected by the Norm Picking algorithm as false arrivals (or those caused by magnitude sources small enough that we cannot catalog).
Discussion

The problem of earthquake detection is of general interest and a large body of research exists that focuses on the problem of arrival picking (Allen, 1978; Baer and Kradolfer, 1987; Sleeman and Van Eck, 1999; Saragiotis et al., 2002; Küperkoch et al., 2010; Baillard et al., 2013; Ross et al., 2018; Zhu and Beroza, 2018), with comparatively fewer works on arrival associations (Bache et al., 1993; Johnson et al., 1997; Arora et al., 2013; Bergen and Beroza, 2018; Ross et al., 2019; Yeck et al., 2019). With the rapidly improving suite of arrival picking methods, development of effective arrival association algorithms is increasingly important because associations

![Figure 5](https://pubs.geoscienceworld.org/ssa/bssa/article-pdf/109/6/2510/4891096/bssa-2019081.1.pdf)

**Figure 5.** Representative set of 16,343 earthquakes from the catalog between the two months of 1 January 2015–1 March 2015. (a) Map view of events (circles) colored by depth; background color is the depth to the Slab 1.0 Nazca plate. Pink lines are cross-section transects labeled A–A′, B–B′, C–C′, and D–D′. (b) Cross-section display of earthquakes within ±0.2° of transects A–D from (a). Circle color represents negative log-likelihood of the source location as calculated by the probabilistic earthquake relocation routine (equation S4, available in the supplemental content to this article). Pink lines denote the depth to Slab 1.0 Nazca plate along the transects.

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Catalog Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CSN</td>
</tr>
<tr>
<td>Number of events</td>
<td>30,796</td>
</tr>
<tr>
<td>Events/day</td>
<td>7</td>
</tr>
<tr>
<td>Matched events</td>
<td>98.7%</td>
</tr>
</tbody>
</table>

Time range, number, and event rates listed for CSN. Sippl et al., and this studies catalog. Proportion of matched events between other catalogs and our own with root mean square-residual between moveout vectors <6.5 s are listed. CSN, Centro Sismológico Nacional.
Figure 6. Subset of 531 detected earthquakes (or blasts) from the events shown in Figure 5. Events selected have negative log-likelihood location values $>10$. (a) Map view of events (circles) colored by depth; background color is the depth to the Slab 1.0 Nazca plate. Pink squares labeled A–F denote locations of quarries or industrial sites. (b) Satellite imagery from Google Earth of the six locations marked A–F in (a). (c) Histogram of earthquakes/blasts shown in (a) by local time of day (GMT-3). Blasting interval above background level is marked by shaded region.
Figure 7. Earthquake detections in the temporal and spatial vicinity of the $M_w$ 8.2 Iquique earthquake. (a) Each subpanel plots all epicenters of earthquakes $> -75$ km depth in the catalog, in sequential nonoverlapping windows of three days. Days above each panel list the days before or after 1 April 2014. Numbers list the number of events shown in the subpanel between the day listed and the day of the next subpanel. Pink square marks epicenter of the Iquique earthquake. Other features include black line (trench), brown line (coast), and blue and red colors of text indicating time intervals before or after the Iquique earthquake, respectively. (b) Earthquake counts between 1 January 2014 and 1 August 2014 in the spatial domain of (a) computed over moving 24 hr windows. (c) Gutenberg–Richter (GR) magnitude distribution curves for the Centro Sismológico Nacional (CSN), Sippl et al. (2018), and this studies catalog. GR curves are calculated using overlapping data intervals (between catalogs), as described in the Application: Northern Chile Earthquake Catalog section.
Figure 8. Example set of earthquake detections during 2 April 2014. (a) Arrival times and detections of 17 earthquakes, including many aftershocks of the 1 April 2014 $M_w$ 8.2 Iquique earthquake. (b) Inset of waveforms from an example earthquake shown in (a). Waveforms shown are HHE channel 50 Hz data band-passed between 5 and 22 Hz, normalized by their maximum amplitudes. In (a,b), predicted move-outs of $P$ and $S$ waves are given by blue and red lines, and predicted $P$ and $S$ arrival associations are given by blue and red circles, respectively. Gray circles represent arrivals not associated with a source.
Earthquake Arrival Association with Backprojection and Graph Theory

by Los Alamos National Lab Research Library, dtrugman

Earthquake Arrival Association with Backprojection and Graph Theory

become more difficult as rates of arrival detections increase with a power law proportional to the arrival picking sensitivity. Issues arise in associating arrivals with their correct sources, as well as in the auxiliary problems of determining the number of sources, determining the phase types of arrivals, and identifying false arrivals. By solving this problem as accurately as possible, high fidelity and low-magnitude-of-completion earthquake catalogs can be obtained. Such catalogs can enhance understanding of many types of seismogenic processes (e.g., spatial-temporal clustering, aftershock and foreshock sequences, dynamic triggering, or long-term rate changes).

In this article, we presented an algorithm that uses tools from BP and graph theory to address the issues inherent to the problem of associations and enable the development of dense catalogs after arrival-time picks have been made with a sensitive arrival picking algorithm (e.g., the Norm Picker described herein, or other alternatives; Baillard et al., 2013; Ross et al., 2018; Zhu and Beroza, 2018). We demonstrated how this technique can have high accuracy on relatively challenging synthetic and real detection problems (Figs. 4 and 8) and have also used the method to develop a catalog in northern Chile between 1 January 2010 and 31 December 2017 that has a magnitude-of-completion of M ~ 1.8, which lowers the magnitude of completion by 0.8–1 units compared with existing catalogs (Table 4; Fig. 7c). Visual verification of a representative set of newly detected events was used to validate the detections to find that we have an approximately (85%, 11%, 4%) distribution of detection qualities (1, 2, 3), respectively. Other forms of validation are that we detected the aftershocks and foreshocks of the 1 April 2014 Mw 8.2 Iquique earthquake with no modifications to the method, and we also detected a number of sources nearby local quarries and industrial sites (Fig. 6). We re-detected nearly all previously cataloged earthquakes in the most complete catalog of this region to date (Table 4) while increasing the average rate of detection from 35 events/day to 273 events/day. In the application to northern Chile, we found we could process one day’s worth of seismicity using 20 cores of an Intel Xeon 2.40 GHz CPU in ~135 s, enabling the creation of all 8 yr of the catalog in ~4.5 days of processing time.

When comparing catalogs, it is important to recognize that different groups produce earthquake catalogs for a variety of purposes. Although we use our method to lower the magnitude-of-completion and detect smaller earthquakes with an efficient, automated approach, the CSN is tasked with reporting about hazardous earthquakes throughout Chile and focuses on characterizing larger earthquakes with high accuracy. Similarly, the study of Sippl et al. (2018) focused on identifying seismogenic structures and used double-difference earthquake relocation (Waldhauser and Ellsworth, 2000) to map the surface and double-seismic zone of the subducting Nazca plate in northern Chile, with a precision and scale that was previously unreported. The different catalogs are hence complementary, and each may provide valuable insights into seismogenic processes occurring in this region.

The association method described in this article shares similarities to existing association methods used in several commonly used software packages (e.g., Earthworm, SeisComP3, Early-Est, Antelope, and GLASS3). Many of these packages use BP of arrivals to identify candidate source locations (and origin times) as coordinates with high BP coalescence. When coalescence thresholds have been exceeded, these tools extract arrivals in agreement with the maximal coalescent source and then iteratively process remaining arrivals to detect additional sources, proceeding in this way until the maximum remaining coalescence decreases below a specified threshold. The notable difference of our method is that all local maxima in the BP space above a threshold are first collected (which typically includes several true and many false candidate sources), and then the constrained ILP optimization problem is used to infer all sources and phase assignments simultaneously based on the physical principles encoded in the constraints of the ILP problem (equation 6). This approach avoids the iterative, greedy association procedure used in the other methods. By posing the problem as an explicit optimization (rather than a set of heuristic rules), our approach is flexible and can be augmented for individual problem specifications by modifying (equation 6a) to include additional optimization targets or by modifying (equation 6b) to include additional relevant constraints.

One key component of our method that could be improved would be in generalizing our use of the BP metric. Currently, equation (1) assumes that all stations can theoretically detect arrivals of all earthquakes. However, for large-aperture networks, the majority of small earthquakes will produce arrivals only on subsets of the networks stations due to attenuation. By backprojecting arrivals from the entire network at all times to all templates, the current metric would allow spurious, unrelated arrivals from distant portions of the network (which line up with the moveout of a source but are not physically related) contribute to the detection of a source and be falsely assigned to that earthquake. If it is possible to modify equation (1) such that all sources only “look” locally for the stations that can potentially observe that source (given its magnitude), then the association method could scale to large aperture networks while maintaining its ability to detect both small and large earthquakes. In addition, if the directivity of arrivals were known, equation (1) could be modified to only propagate waves in the direction of the source, which would significantly reduce the number of false local maxima in the BP space.

Other places in which the method can be improved are in adaptively adjusting the RBF-kernel widths to account for known variability in travel-time uncertainties (between different stations and candidate source regions) and in incorporating additional information into the graphs prior to applying competitive assignment. For example, edge weights could be weighted by the predicted phase likelihood of each arrival (Zhu and Beroza, 2018), and if pairs of arrivals are known with high certainty to come from a common source (Bergen and Beroza, 2018; McBrearty et al., 2019), a constraint could be added into the constraint matrix of equation (6b) to force such arrivals to be assigned to a...
common source. An avenue of development, more generally, may be in constructing arrival–arrival indexed graphs (with edge weights proportional to the likelihood those arrivals are associated), which may contain additional or complimentary information to be used with the current approach. Such graphs have the advantage that the number of nodes is always bounded by the initial dataset and cannot blow up with increasing numbers of false candidate sources.

The earthquake catalog we produced in this study may contain valuable data for future studies. A large body of work is now using machine-learning-based methods that benefit from large training datasets. By containing more than 800,000 earthquakes, the catalog presented here may be a valuable source of training data for future machine-learning-based studies. Other possible directions of future research would be to investigate the dynamics of small-scale spatial-temporal interactions of earthquakes, monitor rate changes of seismicity, or investigate for the prevalence of dynamic triggering contained in the catalog.

Conclusion

Associating earthquake arrivals remains a key component to developing dense earthquake catalogs, and it is crucial that association methods can resolve uncertainties such as unknown number of sources, unknown phase types, false arrivals, uncertainties on arrival-time picks, travel-time calculations, and the situation in which several earthquakes occur nearby in time and produce overlapping moveouts across the network. We have shown that by combining tools from BP and graph theory, a robust solution to all of these problems can be obtained. Our solution solves for all sources and their phase assignments simultaneously, rather than inferring each source sequentially in a greedy fashion, as is common in other association routines. We have demonstrated our method on synthetic tests as well as in the development of a new catalog in northern Chile between 1 January 2010 and 31 December 2017 that has a magnitude-of-completion of $M \sim 1.8$. The new catalog in northern Chile contains approximately eight times more earthquakes from previously reported catalogs in this region and may contain valuable insights into seismogenic processes occurring at the active Nazca–South American plate subduction zone yet to be discovered.

Data and Resources

This work included data from the CX seismic network, obtained from the GFZ Data Service. The catalog developed in this work is given in the † supplemental content to this article. For inquiries regarding code availability and future updates to the catalog, please contact the corresponding author.

Acknowledgments

This work was funded by Institutional Support (LDRD) at the Los Alamos National Laboratory. The authors thank Alex Hutko, Ben Baker, and four anonymous reviewers for helpful feedback on our work. The authors also thank Bertrand Rouet-Leduc, Claudia Hulbert, Christopher Ren, James Theiler, Nicholas Lubbers, Marian Anghel, and Daniel Trugman for many helpful discussions.

References


