ABSTRACT

Mining relationships in time series data is of immense interest to several disciplines such as neuroscience, climate science, and transportation. Traditional approaches for mining relationships focus on discovering pair-wise relationships in the data. In this work, we define a novel relationship pattern involving three interacting time series, which we refer to as a tripole. We show that tripoles capture interesting relationship patterns in the data that are not possible to be captured using traditionally studied pair-wise relationships. We demonstrate the utility of tripoles in multiple real-world datasets from various domains including climate science and neuroscience. In particular, our approach is able to discover tripoles that are statistically significant, reproducible across multiple independent data sets, and lead to novel domain insights.

ACM Reference format:

1 INTRODUCTION

Time series data is generated in a large number of real-world applications such as neuroscience, climate science, and transportation. Discovery of complex patterns of relationships between individual time-series, using data-driven approaches can improve our understanding of real-world systems, e.g., how does the brain conduct basic cognitive functions and what are the physical processes operating in the global climate system. Such information can help us devise solutions to critical real-world problems such as climate change, mental disorders, and traffic congestion.

A common type of relationship in time series data is pairs of time-series with strong Pearson correlations, which have been studied in diverse application domains. For example, in the field of neuroscience, positively correlated pairs of time series of fMRI data, obtained at two regions in the brain, provide vital signs of the brain’s connectivity and mental health [3, 4]. As another example in climate science, correlated pairs of time series of a climate variable (e.g., Sea Level Pressure (SLP)) observed at two distant regions have been extensively studied under the label of ‘teleconnections’ [8, 17]. One of the most widely-studied category of teleconnections is the dipole [8, 9], which represents a pair of regions on the Earth’s surface that are negatively correlated in their climate anomaly time-series (seasonality-removed SLP time-series). Dipoles capture underlying processes of the Earth’s climate system that are related to a number of climatic phenomena, such as floods, droughts, and forest fires [15, 18].

In this paper, we define a novel relationship across three time series, T0, T1, and T2, as shown in the graphical representation of Figure 1. The weight of an edge in Figure 1 represents the pair-wise correlation between time series. In particular, our approach is able to discover tripoles that are statistically significant, reproducible across multiple independent data sets, and lead to novel domain insights.

As a real-world example of the tripole pattern, consider the traffic data set from the Minnesota Department of Transportation [1] where the volume of traffic crossing a road section is represented as a daily time series. Using this data, one may be interested in finding non-trivial relationships among the traffic activity at three road sections. Figure 2 shows an example of such a tripole where the traffic time series of the two leaf nodes (T1 and T2) and the root node (T0) are shown as red, yellow, and blue curves respectively in Figure

Figure 1: Graphical representation of an example tripole involving three time series, T0, T1, and T2. The edge weights represent the pair-wise correlations between time series.
The three time series indicate daily volume of southbound traffic crossing three different road stations in Minnesota during January 2015. Time series $T_1$ and $T_2$ indicate the traffic volume at the roads that contribute to the traffic at the main highway where $T_0$ is being observed. This is evident from the high correlation (0.83) of their sum ($T_{1+2} = T_1 + T_2$, shown as magenta curve in Figure 2(b)) with $T_0$. However, note that the traffic volumes in $T_1$ and $T_2$ are inconsistent across days. While $T_1$ is more dominant during weekdays (possibly indicating weekday commute traffic of passengers going to work), $T_2$ is much more dominant on Sundays (see the weekly spikes in yellow curve). Due to these inconsistencies, their individual correlations with $T_0$ are much weaker (0.42 for $T_1$ and 0.29 for $T_2$). Interestingly, spikes in $T_1$ are accompanied by dips in $T_2$ and as a result, they cancel each other out in their addition. Hence, the resultant sum time series $T_{1+2}$ shows a much stronger correlation with $T_0$ and captures the true relationship between the traffic volumes at the three stations.

Note that a tripole is interesting only if correlation of $T_0$ with $T_{1+2}$ is much stronger than correlation of $T_0$ with $T_1$ or $T_2$ individually. For example, Figure 3 shows another tripole in the transportation data where the sum of the leaves has a high correlation value of 0.97 with the root. However, this high correlation can be explained by the high individual correlation between $T_1$ and $T_0$ (as they correspond to neighboring road sections on the same highway). Hence this tripole is not interesting.

As another example, consider the tripole in Sea level Pressure (SLP) data between SLP time series of three regions that are shown in Figure 4(a) using differently colored patches on the world map. We can see from Figure 4(b) that $T_0$ shows a correlation of $-0.53$ with $T_{1+2}$, which is significantly stronger than the correlation it shows with either of the two leaves, $T_1$ and $T_2$ (-0.33 and -0.21 respectively). This tripole indeed represents a physically relevant but previously unknown phenomenon (atmospheric waves that flow from Siberian Region (root) towards the two leaves in the Pacific Ocean) that was discovered using the approach described in this paper[11].

Above examples are quite encouraging to pursue a formal study of tripoles and explore their utility in different domains. To the best of our knowledge, such relationships have never been defined nor studied systematically in the previous literature. Therefore, the very first challenge in studying tripoles is to define them formally and devise measures to assess their interestingness. Further, discovering tripoles is computationally challenging since a brute-force method that enumerates every candidate tripole would require computing $(\binom{N}{3})$ correlations, where $N$ is the number of time-series.
Furthermore, evaluating discovered tripoles is not straightforward due to lack of ground truth.

To this end, in this paper we formally define the concept of tripole and propose measures to assess its interestingness. We present a novel approach for discovering tripoles in a time series dataset that is highly computationally efficient compared to the brute-force search. Our proposed approach is based on 1) pre-pruning of the search space for interesting tripoles, and 2) exploiting the dense relationship structure between different time series (e.g. spatial autocorrelation in spatio-temporal datasets). We demonstrate the computational efficiency of proposed approach with respect to brute-force approach on two real-world datasets from neuroscience and climate science domain. Further, we define a notion of statistical significance of tripole and demonstrate the significance of the discovered tripoles in both the real-world datasets. In addition, we also show that many of the discovered tripoles are reproducible in multiple independent datasets and could potentially reveal previously unknown phenomenon.

The rest of the paper is organized as follows. Section 2 introduces definitions and describes the problem formulation for tripole discovery. Section 3 presents our proposed approach for discovering tripoles. In Section 4, we evaluate our proposed approach and the validity of discovered tripoles on two real-world datasets from climate science and neuroscience. In Section 5, we present two case studies on the discovered tripoles and discuss their physical interpretation. Section 6 presents the related work on mining time series relationships. Finally, we conclude by presenting ideas for future work in Section 7.

2 DEFINITIONS

In this section, we formally define the notion of a tripole and present some metrics to capture its useful properties, which will be used in the remainder of the paper. Using these definitions, we describe the problem formulation for discovering tripoles in a time-series data set.

Let $D$ represent a time-series data that contains $N$ time-series, $\{T_1, T_2, \ldots, T_N\}$. We assume that every time-series in this data has a zero mean and unit variance (note that this can always be ensured using a pre-processing step). We provide the following basic definition of a tripole in a time-series data as follows.

Definition 2.1 (Tripole). A tripole, $\Delta \equiv (T_0 : T_1, T_2)$, is a collection of three time-series, $T_0$, $T_1$, and $T_2$, where $T_0$ is referred to as the root of the tripole, while $T_1$ and $T_2$ are referred to as the leaves of the tripole.

Note that the order of the leaves in the notation of a tripole is unimportant, i.e., $(T_0 : T_1, T_2)$ and $(T_0 : T_2, T_1)$ are equivalent to each other. Next, we define two metrics, strength and jump, to capture the characteristics of a tripole in the following.

Definition 2.2 (Strength). The strength of a tripole $\Delta \equiv (T_0 : T_1, T_2)$, denoted by $\sigma_\Delta$, measures the correlation between the time-series at the root $T_0$, with the sum time-series of the leaves, $T_{1+2} = (T_1 + T_2)$, as follows

$$\sigma_\Delta = corr(T_0, T_{1+2})$$

(1)

The strength of a tripole basically measures the amount of variability in $T_0$ that can be explained by the combination of the leaves, represented using their sum: $T_{1+2}$. Note that while there could be other ways to combine the information in the two leaves, we chose the sum for its simplicity and ease of interpretation. This also necessitates that the time-series are pre-processed to have zero mean and unit variance.

A high magnitude of $\sigma_\Delta$ signifies a strong relationship among the time series participating in $\Delta$. Note that the strength could either be positive or negative depending on the sign of $corr(T_0, T_{1+2})$, both of which could be interesting in different applications. In this paper, tripoles with positive strength are referred to as positive tripoles, and those with negative strength are referred to as negative tripoles.

Definition 2.3 (Jump). The jump of a tripole $\Delta \equiv (T_0 : T_1, T_2)$, denoted by $\delta_\Delta$, can be defined as:

$$\delta_\Delta = corr(T_0, T_{1+2})^2 - \max\{corr(T_0, T_1)^2, corr(T_0, T_2)^2\}$$

(2)

$\delta_\Delta$ represents the additional variance of the root that is explained by the sum of two leaves as opposed to any of them. The jump thus helps in capturing interactions among tripoles of time-series that cannot be expressed as a mere combination of pair-wise relationships. For example, a tripole where each of the two leaves are individually highly correlated with the root may show a high value of tripole strength. However, such a tripole would show a low value of jump as the sum of the leaves would not provide any additional advantage in explaining the root than the individual leaves considered alone. Hence, the jump can be used as a useful measure to identify interesting tripoles in time-series data.

Definition 2.4 (Interesting Tripoles). Given a jump threshold $\delta$, we can define the set of all interesting tripoles, $U$, as:

$$U = \{\Delta | \delta_\Delta > \delta\}$$

(3)

Further, depending on the application, we may be selectively interested in discovering the set of all positive interesting tripoles ($\sigma_\Delta > 0$) or all negative interesting tripoles ($\sigma_\Delta < 0$). We will refer to both these subsets of $U$ as $U_+$ and $U_-$, respectively.

In order to ensure the discovery of a distinct set of tripoles, we next define a notion of similarity between any two tripoles.

Definition 2.5 (Tripole Similarity). For a user-specified positive threshold $\tau$, termed as the similarity threshold, two tripoles, $\Delta \equiv (T_0 : T_1, T_2)$ and $\Delta' \equiv (T'_0 : T'_1, T'_2)$ are similar if the corresponding roots and leaves of the two tripoles have a correlation greater than $\tau$. Formally, two tripoles are similar if we can find a suitable rearrangement of the leaves of the two tripoles such that the following three conditions hold:

$$corr(T_0, T'_0) \geq \tau, \ corr(T_1, T'_1) \geq \tau, \text{ and } corr(T_2, T'_2) \geq \tau.$$  

(4)

Using the above definition of tripole similarity, we can now define the notion of a non-redundant and complete set of interesting tripoles as follows.

Definition 2.6. Given a set of all interesting tripoles, $U$, and a similarity threshold, $\tau$, we can call a collection of interesting tripoles, $C$, to be non-redundant and complete if the following two conditions hold:

- No two tripoles in $C$ are similar.
- For any tripole $\Delta_U \in U$, there exists a tripole $\Delta_C \in C$ such that $\Delta_U$ and $\Delta_C$ are similar.
The above definition ensures that (i) \( C \) does not contain redundant tripoles that are trivial copies of one another, and (ii) \( C \) covers every interesting tripole in \( \mathcal{U} \). Note that there can be multiple sets of non-redundant and complete tripoles for the same set of interesting tripoles. Using the above definitions, we can formally describe the problem of tripole discovery in time-series data as follows:

**Definition 2.7 (Problem Formulation).** Given a time-series data set \( \mathcal{D} \), a jump threshold \( \delta \), a redundancy threshold \( \tau \), and an optional choice of discovering positive or negative tripoles, the objective of tripole discovery is to output a non-redundant and complete set of interesting tripoles, \( C \).

### 3 PROPOSED APPROACH

In this section, we present a family of computationally efficient approaches for discovering tripoles in time-series data, termed as Cost-efficient Tripole Finding (CONTRA) approaches. To motivate the usefulness of CONTRa, we first describe a naive approach for producing a non-redundant and complete set of tripoles that performs a brute-force enumeration of all possible tripoles. We then provide two different implementations of CONTRa: CONTRaComplete and CONTRaFast, that leverage key characteristics of tripoles and the nature of real-world time-series data sets to significantly reduce the number of candidate tripoles that are considered for interestingness. Specifically, the CONTRaComplete approach uses an important mathematical relationship between the pair-wise correlations of an interesting tripole and the jump threshold, \( \delta \), which helps in pruning a large number of non-interesting tripoles from being enumerated. The CONTRaFast approach provides further improvements in running time by pruning out a large number of redundant tripoles, although with some loss of completeness in the generated set of interesting tripoles. We describe all of these algorithms in detail below.

#### 3.1 Naive Approach

A basic approach for discovering interesting tripoles is described in Algorithm 1, which involves enumerating all possible tripoles and checking the interestingness of every candidate tripole enumerated by the algorithm (lines 2 to 6). This results in a complete set of interesting tripoles, \( C_0 \). We then remove redundant tripoles in \( C_0 \) (line 7) using the simple procedure described in Algorithm 2. In this procedure, we first select a tripole \( \Delta' \) with the strongest jump and insert it to the set \( C \) (lines 3-4). Next, in line 5, we remove all the tripoles in \( C_0 \) that are similar to \( \Delta' \) for a given similarity threshold, \( \tau \), and repeat the entire process (lines 3-5) until \( C_0 \) is empty. Finally, the set \( C \) is returned as the non-redundant and complete set of interesting tripoles.

Because this naive approach requires enumerating every possible candidate tripole, it would have a computational complexity of \( O(N^3) \), and can be highly computationally demanding especially for large real-world datasets. For example, the climate SLP data set contains more than 10,000 time-series globally at a spatial resolution of 2.5 degrees, which results in more than \( 10^{12} \) candidate tripoles to be enumerated by the naive approach. This motivates the development of computationally efficient approaches for tripole discovery that can reduce the number of candidate tripoles that are evaluated for interestingness.

#### 3.2 CONTRaComplete

The basic idea of CONTRaComplete approach is to prune the search space of interesting tripoles by exploiting a key mathematical relation between the jump of a tripole and the correlation strengths of the three pairs formed between its root and leaves. Before describing the CONTRaComplete algorithm, we first present this key relation in the following theorem:

**Theorem 3.1.** In a tripole \( \Delta \equiv (T_0 : T_1, T_2) \), let the pair-wise correlation strengths \( \text{corr}(T_0, T_1), \text{corr}(T_0, T_2), \) and \( \text{corr}(T_1, T_2) \) be denoted by \( a, b, \) and \( c \) respectively, and let \(|a| \geq |b|\), without losing generality. Then, the jump \( \delta_\Delta \) of the tripole is given by

\[
\delta_\Delta = \frac{(a+b)^2}{2(1+c)} - a^2
\]

Proof. Following the basic formulae of Pearson correlation for the sum of two variables, we can obtain an expression for \( \sigma_\Delta \) as

\[
\sigma_\Delta = \text{corr}(T_0, T_{1+z}) = \frac{\text{cov}(T_0, T_1) + \text{cov}(T_0, T_2)}{\sqrt{\text{var}(T_0) + \text{var}(T_1) + 2\text{cov}(T_1, T_2)}}
\]

Since we assume the three time series \( T_0, T_1, \) and \( T_2 \) to be normalized in this paper, therefore, the above expression reduces to

\[
\sigma_\Delta = \frac{a+b}{\sqrt{2(1+c)}}
\]

Since \(|a| \geq |b|\), the expression for \( \delta_\Delta \) could be obtained using Equation (6) as

\[
\delta_\Delta = \sigma_\Delta^2 - |a|^2 = \frac{(a+b)^2}{2(1+c)} - a^2
\]

Among the three pairs of time series in a tripole \( \Delta \), let us refer the one with maximum absolute correlation as the super pair of
also from tripoles: a new class of relationships in time series data kdd '17, august 13-17, 2017, halifax, ns, canada

λ, and let the absolute correlation strength of the super-pair be denoted by s. then using theorem 3.1, we next show that an upper bound on the jump of a tripole could be expressed in terms of s as the following corollary.

**corollary 3.2.** let s denote the absolute correlation value of the super pair of a tripole, λ ⇝ (t0 : t1, t2). then,

\[ \delta_\lambda \leq \frac{s^2(1 + s)}{(1 - s)} \]

**proof.** from theorem 3.1, we get

\[ \delta_\lambda = \frac{(a + b)^2}{2(1 + c)} - a^2 \]

also from theorem 3.1, we have \(|b| \leq |a|\). therefore,

\[ \delta_\lambda \leq \frac{4a^2}{2(1 + c)} - a^2 = a^2 \left[ \frac{2}{1 + c} - 1 \right] \]

by definition, \(s^2 \geq a^2\). therefore,

\[ \delta_\lambda \leq s^2 \left[ \frac{2}{1 + c} - 1 \right] \]

further, note that the right hand side (r.h.s) in the above inequality increases as c decreases. by definition, \(|c| \leq s\), and thus \(c \geq -s\). therefore, we get

\[ \delta_\lambda \leq s^2 \left[ \frac{2}{1 + (-s)} - 1 \right] = \frac{s^2(1 + s)}{(1 - s)} \]

an intuitive corollary of corollary 3.2 helps us obtain a lower bound on the maximum absolute pair-correlation in an interesting tripole as follows.

**corollary 3.3.** given a jump threshold δ, let \(s_\delta\) be the lowest value in [0, 1] such that \[\frac{s^2_\delta(1 + s_\delta)}{(1 - s_\delta)} \geq \delta\]. then, the magnitude of correlation of the super pair of an interesting tripole, s, is lower bounded by \(s_\delta\), i.e. \(s \geq s_\delta\).

the above property can, in principle, be used to avoid the enumeration of any tripole \(\lambda\) with maximum absolute correlation, \(s < s_\delta\). for the problem of discovering interesting negative tripoles (where the strength of every discovered tripole is negative), we can use the following additional corollary on the super pair of a negative tripole.

**corollary 3.4.** a super pair in a negative tripole is always negatively correlated.

**proof.** in the proof of corollary 3.2, notice that the jump of a \(\lambda\) attains its upper bound for the case when \(|a| = |b| = s\), and \(c = -s\). also from equation 6, it is clear that \(a\) should be negative for negative tripoles. therefore, the jump of a negative tripole attains its upper bound when \(a = b = c = -s\).

the above two properties form the core components of the contrast approach (formally described in algorithm 3), which is able to prune a large number of candidate tripoles that are guaranteed to be non-interesting. the basic idea is to start with the set of all pairs of time-series that can potentially serve as super pairs of an interesting tripole (line 3 of algorithm 3). note that a pair of time-series that does not meet the criteria listed in corollaries 3.3 and 3.4 can be confidently pruned out, resulting in a small set of candidate super pairs. this is described in lines 2 to 6 of algorithm 4. as will be described later in section 4.3, this helps in obtaining major reductions in the number of candidate tripoles that are enumerated for analyzing their interestingness. having obtained a set \(l\) of all candidate super pairs, we next grow every super pair, \((t_A, t_B)\), by considering a third time-series, \(t_i\), such that the two new edges, \((t_A, t_i)\) and \((t_B, t_i)\) have a lower absolute correlation than \((t_A, t_B)\) (line 1 of algorithm 5). for every one of the three tripoles that can be constructed using the three time-series, \(t_A, t_B,\) and \(t_i\), we compute their jump value and check if it is greater than the jump threshold (lines 2 to 4). this results in the final set \(p\) of all interesting tripoles that can be constructed using the super pair \((t_A, t_B)\). we add \(p\) to the set \(c\) of interesting tripoles (line 6 of algorithm 3). we finally remove similar tripoles in \(c\) (line 8 of algorithm 3) to obtain a non-redundant and complete set of interesting tripoles.

algorithm 3 contrast complete

input dataset \(d\). parameters: \(δ, \tau\)

output a non-redundant and complete set of interesting tripoles, \(c\).

1: initialize \(c \leftarrow φ\)
2: select lowest value of \(s_\delta\) ∈ \([0, 1]\) s.t. \[\frac{s^2_\delta(1 + s_\delta)}{(1 - s_\delta)} \geq \delta\] → using corollary 3.3
3: \(l \leftarrow \text{find candidate super pairs}(d, s_\delta)\)
4: for each super pair \((t_A, t_B)\) ∈ \(l\) do
5: \(p \leftarrow \text{find tripoles for super pair}(d, t_A, t_B, δ)\)
6: \(c \leftarrow c \cup p\)
7: end for
8: \(c \leftarrow \text{get non-redundant tripoles}(c, τ)\)
9: return \(c\)

3.3 contrast fast

the contrast approach introduces an additional step to remove redundant tripoles from being enumerated, that provides further reduction in the running time however with some loss of completeness in the discovered interesting tripoles. algorithm 6 provides an outline of the contrast approach. the only difference between contrast and contrast complete is the generation of a non-redundant set of super pairs (line 4). the procedure for generating the non-redundant set of super pairs is described in algorithm 7. in this algorithm, we start with the most negatively correlated pair, \((t_A, t_B)\), that is added to the set of non-redundant super pairs (lines 3 and 4). we then consider the zone of exclusions of \(t_A\) and \(t_B\).
Algorithm 5 FIND TRIPOLIES FOR SUPER PAIR

Input: Dataset D, Super Pair (TA, TB), Parameters δ
Output: A set P of interesting tripoles with (TA, TB) as super pair

1: Q ← All time series Ti s.t. max|corr(Ti, TA)|, |corr(Ti, TB)| < |corr(TA, TB)|
2: C1 ← All tripoles δi ≡ (TA : Ti, TB) s.t. Ti ∈ Q and δι ≥ δ
3: C2 ← All tripoles δi ≡ (TA : Ti, TB) s.t. Ti ∈ Q and δι ≥ δ
4: C3 ← All tripoles δi ≡ (TA : Ti, TB) s.t. Ti ∈ Q and δι ≥ δ
5: P = C1 ∪ C2 ∪ C3
6: return P

Algorithm 6 CONTRaFast

Input: Dataset D, Parameters δ, τ, κ
Output: A set C of non-redundant complete sets of interesting tripoles, C

1: Initialize C ← ∅
2: Select lowest value of xδ ∈ [0, 1] s.t. p2(xδ + xδ) ≥ δ → Using Corollary 3.3
3: L ← FIND CANDIDATE SUPER PAIRS(D, xδ)
4: L ← FIND NON-REDUNDANT SUPER PAIRS(L, κ)
5: for each super pair (TA, TB) ∈ L do
6: P ← FIND TRIPOLIES FOR SUPER PAIR(D, (TA, TB), δ)
7: C ← C ∪ P
8: end for
9: C ← GET NON-REDUNDANT TRIPOLIES(C, τ)
10: return C

4 EXPERIMENTAL RESULTS AND EVALUATION

Our evaluation framework consists of two major parts. In the first part, we evaluate the computational efficiency and completeness of the search of CONTRa on a real-world dataset from climate science. In the second part, we evaluate the validity of discovered interesting tripoles in two real-world datasets from climate science and neuroscience domain. We first present the description of the two datasets along with the data pre-processing steps that were applied to them.

4.1 Data and Pre-processing

4.1.1 Global Sea Level Pressure (SLP) Data: We used monthly Sea Level Pressure (SLP) dataset provided by NCEP/National Center for Atmospheric Research (NCAR) Reanalysis Project [10] which is available from 1979-2014 (36 years) at a spatial resolution of 2.5 × 2.5 degree (10512 grid points, also referred to as locations). In this paper, we used the data from the months corresponding to winter season (December, January, and February) from each year, thereby resulting in 36 × 3 = 108 observations for each location. For each time series, we followed the standard pre-processing steps followed in climate science to remove the annual seasonality and linear trends [8]. Relationships in spatio-temporal data are preferably studied between regions (sets of spatially contiguous locations) as opposed to individual locations, since they are more reliable and stable over time. Therefore, around each of the 10,512 locations as centres, we grew a spatially contiguous region by including the top 50 locations that were most strongly correlated to the centre location and were spatially contiguous. For the resultant 10,512 regions, the representative time series were obtained as the normalized average time series of the constituent locations. To ensure that each region behaves as a single entity, we measured tightness of the region by measuring the average correlations between the centre location and all the member locations, and discarded 61 regions for which the average correlations were below 0.8. As a result, we get a set of representative time series of 10,443 regions. The choices of size of the regions and the threshold on the tightness of region were based on the suggestions given by domain experts.

4.1.2 Brain fMRI Data: We used neuroimaging data collected at the University of Utah as part of a reproducibility study [2]. In this study, a set of 50 functional-Magnetic Resonance Imaging (fMRI) scans of one subject were acquired when the subject is resting. Another set of 50 fMRI scans were collected from the same subject while the subject is involved in an audio-visual task (watching cartoons). The spatial resolution is 3mmx3mmx3mm and the temporal resolution is 2 secs. A number of fMRI pre-processing steps including motion correction, unwarping, and filtering that have been described in [2] have been performed. In addition, an Automated Anatomical Labeling Atlas [16] that maps grey matter locations to 90 anatomical regions is used to compute a mean time series of each brain region from each scan. We used these 90 time series to discover interesting tripoles from each scan.

4.2 Experimental Setup

Our approach requires input values for three parameters: jump threshold δ, the redundancy threshold τ, and the tuning parameter κ. The choices of the parameters could be specified by the user depending on the availability of computational resources and the application at hand. Setting jump threshold δ to lower values would lead one to obtain tripoles with lower jump, but would also increase the computational cost as the search space of interesting tripoles would expand. Similarly, setting τ to higher (lower) values would lead to the searching the interesting tripoles at a finer (coarser)
resolution, but would require higher (lower) computational time. As well, $\kappa$ determines the completeness of the search, as setting $\kappa$ to 1 ensures completeness of the search, while the lower values $\kappa$ would lead to approximately complete but a faster search. In our experiments, the jump threshold $\delta$ and the redundancy threshold $\tau$ were set to 0.15 and 0.7 respectively for both the datasets. For brain fMRI dataset, we set $\kappa$ to 1, whereas for SLP dataset being much larger, we set it to 0.99.

For the above parameter setting, we obtained 506 interesting negative tripoles in SLP dataset and 1015 interesting positive tripoles in one of the brain fMRI scans collected when the subject was resting.

### 4.3 Evaluation of CONTRa

In this section, we compare the computational efficiency and completeness of the search of CONTRa with respect to brute-force search. We first demonstrate the impact of super-pair thresholding (result of Theorem 1) introduced in CONTRaComplete on the computational efficiency by estimating the fraction of super-pairs that are pruned in CONTRa.

**4.3.1 Impact of super-pair thresholding.** Using Theorem 1, we can obtain a relationship between $s$, the magnitude of correlation strength of super pair $s$ and the upper bound on the jump of a tripole as shown in Figure 5(a). For $\delta = 0.15$, $s$ needs to be greater than 0.29. Further, from Corollary 3.4, the correlation strength of the super-pair in negative tripoles is always negative. Therefore, for $\delta = 0.15$, for a negative tripole to show jump $\geq 0.15$, the correlation strength of super-pair should always be lower than -0.29. Figure 5(b) shows the distribution of the correlation strengths of all pairs of time series in SLP dataset. The above threshold on $s$ prunes more than 93% of pairwise-correlations in the search of candidate super-pairs of interesting negative tripoles in SLP dataset, which consequently prunes the search space of interesting tripoles by a factor of more than 14. The size of pruning further increases for higher thresholds of $\delta$ which implies that lesser computational time is required to search for most interesting tripoles. The factor of pruning is expected to be high in many of the time series datasets, where the typical distribution of pairwise correlations between time series is centered around zero, and therefore, majority of the pairs showing poor correlation strengths tend to get pruned.

**Figure 5:** Impact of super-pair pruning

4.3.2 Impact of parameter $\kappa$. We next demonstrate the impact of parameter $\kappa$ on computational efficiency and completeness of CONTRa using the following two evaluation metrics:

**1. Computational Time (Cost):** For assessing computational efficiency, we record the computational time taken in execution of all the three stages of the algorithm.

**2. Fraction of missed interesting tripoles (MissFrac):** We quantify the incompleteness of CONTRa based on the fraction of tripoles of a complete set $C$ of interesting tripoles that were missed in its output. A complete set $C$ could ideally be obtained by executing CONTRa at $\kappa = 1$. However, in practice, that takes more than a few days to complete on SLP dataset. Therefore, for evaluation purposes, we generated a pseudo-complete set $C'$ of interesting tripoles by running CONTRa at $\kappa = 0.99$. For SLP dataset at $\delta = 0.15$, we obtained $C'$ that consisted of 506 negative interesting tripoles.

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>Cost</th>
<th>MissFrac</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5 sec</td>
<td>0.15</td>
</tr>
<tr>
<td>0.05</td>
<td>30 sec</td>
<td>0.07</td>
</tr>
<tr>
<td>0.01</td>
<td>11 sec</td>
<td>0.03</td>
</tr>
<tr>
<td>0.001</td>
<td>11 sec</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table 1: Cost and MissFrac of CONTRa for finding interesting negative tripoles in SLP dataset at different values of $\kappa$

Table 1 shows the Cost and MissFrac of CONTRa for finding interesting negative tripoles in SLP dataset for different values of $\kappa$. As shown in the table, as $\kappa$ approaches towards 1, the computational time increases dramatically from order of few seconds to days. On the other hand, MissFrac decreases and the search of interesting tripoles approaches the completeness. However, note that the rate of increment in Cost shoots up dramatically as $\kappa$ is further increased beyond 0.99. For example, it took more than 4 days to complete the search at $\kappa = 0.999$. On the other hand, the fraction of missing interesting tripoles at $\kappa = 0.999$ is less than 1%. This could be attributed to the inherent spatial autocorrelation present in SLP dataset that results in a lot of redundant tripoles, and therefore, setting $\kappa$ to 0.99 allows us to find almost all the interesting tripoles in a much smaller time.

4.4 Evaluation of Discovered Tripoles

The concept of a tripole relationship, to the best of our knowledge, has not been explored before and so no ground truth is available regarding the existence of such relationships in any domain. Due to this, it is challenging to evaluate the validity of the tripole relationships. We pursued two directions to study the validity of the discovered tripoles: 1) Analysis of the statistical significance of the jump of the discovered tripoles, and 2) Reproducibility of the discovered tripoles in independent datasets.

**4.4.1 Analyzing Statistical Significance of Tripoles.** We assessed the statistical significance of the jump of the discovered tripoles using a randomization-based approach that answers the question: Can the jump of the given tripole be achieved by replacing one of the two leaves with a random time series whose correlation with the root is preserved? If this happens for a very few random time series, then the tripole is considered to be statistically significant. In particular, it suggests that the jump obtained in the discovered tripoles could not be achieved by a random chance and...
the two leaves in the discovered tripoles indeed compliment each other to obtain a stronger signal of root.

The approach was setup in the following fashion: For a given tripole \( \Delta \equiv (T_0 : T_1, T_2) \), we replaced the leaf time series (without loss of generality, we chose the leaf that shows weaker correlation with the root) by a random time series \( T_{\text{rand}} \) and recomputed the jump \( \delta_{\text{rand}} \) of the tripole \( \Lambda_{\text{rand}} \equiv (T_0 : T_1, T_{\text{rand}}) \). The random time series was generated such that the original correlation between the leaf and root time series was preserved, i.e. \( \text{corr}(T_0, T_{\text{rand}}) \in (b-0.01, b+0.01) \), where \( b = \text{corr}(T_0, T_2) \). This process was repeated 10000 times to obtain a distribution of the \( \delta_{\text{rand}} \) and thus obtain a p-value of the original tripole jump \( \delta_\Delta \).

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.

Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that showed significant jumps. Similarly, in brain fMRI dataset of one of the subjects, we found 547 out of 1015 discovered positive tripoles that showed significant jumps at a significance level of 0.05. Note that for both the datasets, tripoles with higher jump tend to have greater statistical significance.
are often used to study climate change under different greenhouse
gas emission scenarios) to reproduce these tripoles can also provide
the information about their ability to represent these processes.
Hence tripole based analysis can play a role in evaluating the quality
of different models. Figure 8 shows the tripole strengths across
different windows for observations data (HadSLP2) and two of the
climatic models used in the IPCC (Intergovernmental Panel on
Climate Change) CMIP5 evaluation. It can be seen from the figure
that models are not able to do a very good job of representing this
tripole and thus the corresponding underlying processes.

Figure 8: The negative tripole shown in Figure 4(a) represents a new climate teleconnection discovered between West Siberian
Plain and two regions in Pacific Ocean, for which Figures (a), (b), and (c) show the tripole strength (black), $corr(T_0, T_1)$ (blue),
and $corr(T_0, T_2)$ (red) in HadSLP2 data and two climate models bcc-csm and ACCESS1-0 respectively.

Figure 9: A tripole in brain fMRI data that is discriminative
between resting state and the audio-visual task state of the
subject.

5.2 Discovering discriminative relationships
between a resting and an audio-visual task
in fMRI data
Several tripoles were discovered using time series from the resting
state and the audio-visual task fMRI data, separately. While these
tripoles capture interesting phenomenon pertaining to the two
different states, here we describe a tripole that is discriminative
between resting state and the audio-visual task. Specifically, a
tripole formed by three brain regions: Left Middle Frontal ($R_0$),
Right Superior Frontal ($R_1$), and Left Inferior Frontal ($R_2$), was found
to exhibit a strength $\sigma_A \geq 0.5$ and a jump $\delta_A \geq 0.2$ in 26 (out of
50) audio-visual task scans, but only in 2 (out of 50) resting state
scans. This tripole is shown in Figure 9. The fact that this tripole
exists in a significant fraction of the audio-visual task scans and
only in a very small fraction of the resting state scans suggests
that this relationship is driven by the audio-visual stimulus. In
fact, the frontal regions of the brain, that participate in the above
tripole, are known to have neuronal cells that has the ability to
integrate signals from different perceptual regions such as auditory,
visual and somatosensory [14]. Moreover, the fact that Left Middle
Frontal is the root of the tripole suggests that it plays more of the
integrative role of assimilating information from the Right Superior
Frontal and the Left Inferior Frontal regions.

A number of task based fMRI studies are routinely conducted
by the neuroscience community to elucidate the different types of
interactive and integrative mechanisms underlying brain func-
tion [13]. Approaches we developed for discovering tripoles have
potential applications in these studies, where tripoles such as the
one discussed above can offer interesting insights about the under-
lying integrative mechanisms that are otherwise not captured in
traditional analysis.

6 RELATED WORK
To the best of our knowledge, the notion of a tripole relationship has
never been explored before in the literature of mining relationships
in time series data. Most of the existing work is focused on studying
relationships between a pair of time series [8, 19, 20]. In particu-
lar, such relationships have been studied in spatio-temporal data
across different domains. For example, [8] proposed a graph-based
approach to find ‘dipoles’ in a spatio-temporal climate data that are
characterized by pairs of regions whose corresponding time series
are negatively correlated with each other. Another such example is
of [7] that proposed a tensor factorization approach to decompose
spatio-temporal brain fMRI data into multiple factors, where each
factor captures a set of regions that are strongly related to each
other. Other factorization-based approaches including Principal
Component Analysis (PCA) and Independent Component Analy-
sis (ICA) have also been applied to identify regions with strong
correlations between their time series [5, 17].

This work could also be related with the literature of time se-
ries prediction, where various regularization-based regression
approaches (lasso, group-lasso etc.) have been proposed to find a
subset of time series (predictors) that can be combined (linearly or
non-linearly) to predict a given time series (predictand) with a high
accuracy [6, 12]. The resultant set of predictors can therefore be
considered to be showing a strong relationship with the predictand.
However, these approaches are originally designed to only optimize
We thank reviewers for helpful comments and suggestions. Therefore, such approaches are not directly suited to our problem formulation of finding all interesting tripoles.

7 CONCLUSION AND FUTURE WORK

In this paper, we presented a novel triple relationship that captures an interaction among three time series. We proposed a computationally efficient approach to discover them. We evaluated our approach on two real world datasets from climate science and neuroscience domains. Our approach performed better than a brute-force approach in terms of computational efficiency. Using a randomization-based approach we have shown that the tripoles discovered using our approach are not due to random chance (i.e., they are statistically significant). Moreover, we have shown that the tripoles discovered on the climate science dataset are reproducible by splitting the data into different time segments. Finally, we discussed the relevance of the discovered tripoles to climate science and neuroscience domains. In particular, one of the tripoles found in the climate science dataset has resulted in the discovery of a novel climatic phenomenon that provided in supplementary material and has also been published in a top-tier journal in climate science.

The triple relationship can be further generalized along multiple dimensions. A natural extension would be to pursue multi-pole relationships that involve more than three time series. In this work, we used a simple addition function to combine to combine the information in the two leaves of tripoles. In future, tripoles could be studied for a class of combination functions (e.g. linear /non-linear combinations). In addition, tripoles could also be studied for other choices of relationship measures. Further, tripoles can be studied by taking into account time-lagged relationships between the three time series. Another useful extension of this work could be to develop more advanced algorithms and obtain guarantees on the completeness of the search and the interestingness of tripoles.

ACKNOWLEDGEMENTS

We thank reviewers for helpful comments and suggestions. This work was supported by NSF grant IIS-1029771 and NASA grant 14-CMAC14-0010. Access to the computing facilities was provided by the University of Minnesota Supercomputing Institute.

REFERENCES