Tripoles: A New Class of Relationships in Time Series Data

Saurabh Agrawal
agraw066@umn.edu
University of Minnesota

Gowtham Atluri
atlurigm@ucmail.uc.edu
University of Cincinnati

Anuj Karpatne,
William Haltom, Stefan Liess,
Snigdhansu Chatterjee,
Vipin Kumar
karpa009,halto004,liess,chatt019,
kumar001@umn.edu
University of Minnesota

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.

KEYWORDS

multivariate linear patterns; correlation mining; spatio-temporal; climate teleconnections; fMRI

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].

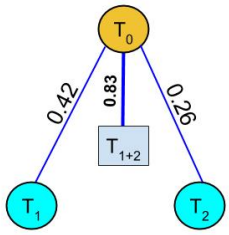


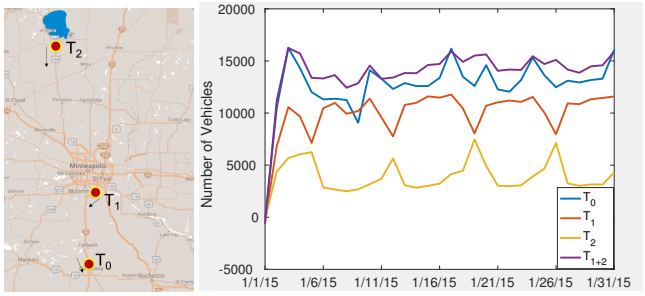
Figure 1: Graphical representation of an example tripole involving three time series, T_0, T_1, and T_2. The edge weights represent the pair-wise correlations between time series.

In this paper, we define a novel relationship across three time series, T_0, T_1, and T_2, as shown in the graphical representation of Figure 1. The weight of an edge in Figure 1 represents the pair-wise correlation among the time series of the corresponding nodes. We can see that the sum of the two time series T_1 and T_2 (denoted as T_{1+2}) provides a much higher correlation with the third time series (T_0), compared to the correlation of T_0 with either of the two time series considered individually. We refer to the group of three time-series as a *tripole*, where T_0 can be termed as the **root node**, while T_1 and T_2 can be termed as the **leaf nodes**.

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 (T_1 and T_2) and the root node (T_0) are shown as red, yellow, and blue curves, respectively, in Figure 2(b). The three time series indicate daily volume of southbound traffic crossing three different road stations in Minnesota during

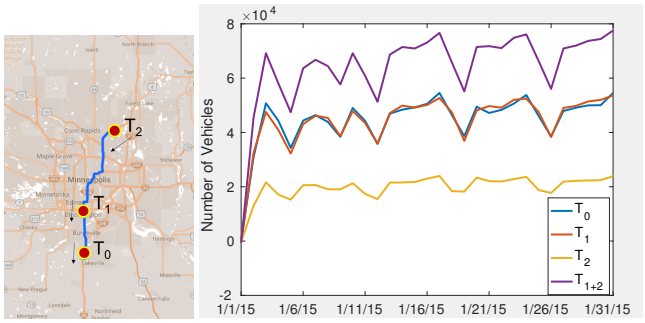
KDD '17, August 13-17, 2017, Halifax, NS, Canada

DOI: 10.1145/3097983.3098099



(a) Spatial Map (b) Time-series

Figure 2: Example of a tripole in transportation data showing different modes of traffic on a highway near Minneapolis.

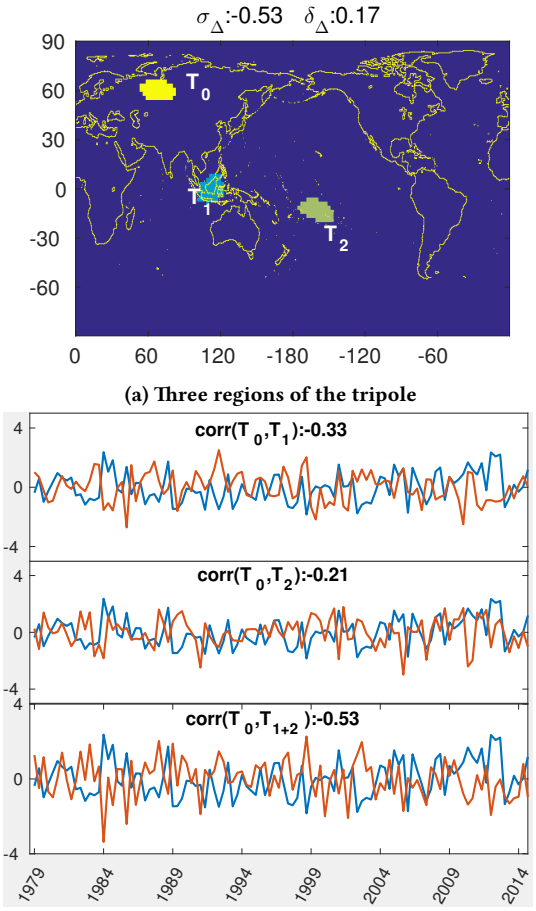


(a) Spatial Map (b) Time-series

Figure 3: Example of a trivial tripole in transportation data corresponding to adjacent road sections on a highway.

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 the computation of T_{1+2}. 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.



(a) Three regions of the tripole

(b) Time series

Figure 4: An example of a tripole in Sea Level Pressure (SLP) data that led to discovery of a new climate teleconnection.

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 $\mathcal{D}$ represent a time-series data that contains N time-series, $\{T_1, T_2, ..., T_N\}$. 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 σ_Δ, 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

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

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 σ_Δ signifies a strong relationship among the time series participating in Δ. 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 δ_Δ, can be defined as:

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

δ_Δ 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 triplets 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 δ, we can define the set of all interesting tripoles, $\mathcal{U}$ as:

$$\mathcal{U} = \{\Delta |\ \delta_\Delta > \delta\} \tag{3}$$

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 $\mathcal{U}$ as $\mathcal{U}_+$ and $\mathcal{U}_-$, 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 τ, 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 τ. 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:

$$corr(T_0, T_0') \geq \tau,\ corr(T_1, T_1') \geq \tau, \text{ and } corr(T_2, T_2') \geq \tau. \tag{4}$$

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, $\mathcal{U}$, and a similarity threshold, τ, 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 $\Delta_U \in \mathcal{U}$, there exists a tripole $\Delta_C \in C$ such that Δ_U and Δ_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 δ, a redundancy threshold τ, 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 naïve 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: **CONTRa-Complete** 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, δ, 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 Naïve 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 Δ^* 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 Δ^* for a given similarity threshold, τ, 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 naïve 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 naïve approach approach. This motivates the development of computationally efficient approaches for tripole discovery that can reduce the number of candiate tripoles that are evaluated for interestingness.

Algorithm 1 Naïve Approach

Input Dataset:$\mathcal{D}$, Parameters: δ,τ
Output A non-redundant and complete set of interesting tripoles, C.
1: Initialize $C_0 \leftarrow \phi$
2: **for** each tripole $\Delta \in \mathcal{D}$ **do**
3: **if** $\delta_\Delta > \delta$ **then**
4: $C_0 \leftarrow C_0 \cup \Delta$
5: **end if**
6: **end for**
7: $C \leftarrow$ GET NON-REDUNDANT TRIPOLES(C_0,τ)
8: **return** C

Algorithm 2 GET NON-REDUNDANT TRIPOLES

Input A set of tripoles: C_0, Similarity threshold: τ
Output A set of non-redundant tripoles C
1: $C = \phi$
2: **while** $C_0 \neq \phi$ **do**
3: $\Delta^* \leftarrow$ Tripole with highest jump in C_0
4: $C \leftarrow C \cup \Delta^*$
5: Remove all tripoles Δ^s from C_0 that are similar to Δ^*
6: **end while**
7: **return** P

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 $corr(T_0, T_1), corr(T_1, T_2)$, and $corr(T_0, T_2)$ be denoted by a, b, and c respectively, and let $|a| \geq |b|$, without loss of generality. Then, the jump δ_Δ 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 σ_Δ as

$$\sigma_\Delta = corr(T_0, T_{1+2}) = \frac{cov(T_0, T_1) + cov(T_0, T_2)}{\sqrt{var(T_0) + var(T_1) + 2cov(T_1, T_2))}} \quad (5)$$

Since we assume the three time series T_0,T_1, and T_2 are normalized, the above expression reduces to

$$\sigma_\Delta = \frac{a+b}{\sqrt{2(1+c)}} \quad (6)$$

Since $|a| \geq |b|$, the expression for δ_Δ can 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 Δ, let us refer to the one with maximum absolute correlation as the **super pair**

of Δ, and let the absolute correlation strength of the super-pair be denoted by s. Therefore,

$$s = max(|a|, |b|, |c|) = max(|a|, |c|) \tag{7}$$

Using Theorem 3.1, we next show that an upper bound on the jump of a tripole could be obtained 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, $\Delta \equiv (T_0 : T_1, T_2)$. Then,*

$$\delta_\Delta \leq \frac{s^2(1+s)}{(1-s)}$$

Proof. From Theorem 3.1, we get

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

Also from Theorem 3.1, we have $|b| \leq |a|$. Therefore,

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

By definition, $s^2 \geq a^2$. Therefore,

$$\delta_\Delta \quad \leq \quad s^2\left[\frac{2}{1+c} - 1\right] \tag{8}$$

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

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

□

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_δ be the lowest value in $[0, 1]$ such that $\frac{s_\delta^2(1+s_\delta)}{(1-s_\delta)} \geq \delta$. Then, the magnitude of correlation of the super pair of an interesting tripole, s, is lower bounded by s_δ, i.e. $s \geq s_\delta$.*

The above property can, in principle, be used to avoid the enumeration of any tripole Δ 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. *Consider a negative tripole Δ such that its super pair is positively correlated. Then, $\delta_\Delta \leq 0.0903$.*

Proof. From Equation 6 and definition of negative tripoles, it is clear that a should be negative for negative tripoles. Therefore, the pair corresponding to a cannot be the super-pair since we know that super-pair is positively correlated. Hence, the super-pair will be the pair that corresponds to c and thus, $c = s$. Using this value of c in Equation 8, we get

$$\delta_\Delta \leq s^2\left[\frac{2}{1+s} - 1\right] = s^2\left[\frac{1-s}{1+s}\right]$$

For $s \in [0, 1]$, the expression $s^2\left[\frac{1-s}{1+s}\right]$ attains its maximum value of 0.0903 at $s = \frac{-1+\sqrt{5}}{2}$. □

The above two corollaries form the core components of the CONTRaComplete 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). This step ensures that the trio of T_A, T_B, and T_i is visited exactly once where (T_A, T_B) is the super-pair. Depending on the choice of root, there could be three tripoles constructed from the above trio: $(T_A : T_B, T_i)$, $(T_B : T_A, T_i)$, and $(T_i : T_A, T_B)$. For each one of these tripoles, 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 CONTRaComplete

Input Dataset:$\mathcal{D}$, Parameters: δ,τ
Output A non-redundant and complete set of interesting tripoles, C.
1: Initialize $C \leftarrow \phi$
2: Select lowest value of $s_\delta \in [0, 1]$ s.t. $\frac{s_\delta^2(1+s_\delta)}{(1-s_\delta)} \geq \delta$ ▷ Using Corollary 3.3
3: $L \leftarrow$ FIND CANDIDATE SUPER PAIRS($\mathcal{D}$,s_δ)
4: **for** each super pair $(T_A, T_B) \in L$ **do**
5: $P \leftarrow$ FIND TRIPOLES FOR SUPER PAIR($\mathcal{D}$,(T_A, T_B),δ)
6: $C \leftarrow C \cup P$
7: **end for**
8: $C \leftarrow$ GET NON-REDUNDANT TRIPOLES(C,τ)
9: **return** C

3.3 CONTRaFast

The CONTRaFast approach introduces an additional step to remove redundant tripoles from being enumerated, that provides further

Algorithm 4 FIND CANDIDATE SUPER PAIRS

Input Dataset:$\mathcal{D}$,Parameters:s_δ
Output A set L of candidate super pairs
1: Initialize $L \leftarrow \phi$
2: **if** negative tripoles are to be only found and $\delta \geq 0.0903$ **then**
3: $E \leftarrow$ All pairs of time series with correlation below 0 and magnitude $\geq s_\delta$ ▷ Using Corollary 3.4
4: **else**
5: $E \leftarrow$ All pairs of time series with correlation magnitude $\geq s_\delta$ ▷ Using Corollary 3.3
6: **end if**
7: **return** L

Algorithm 5 FIND TRIPOLES FOR SUPER PAIR

Input Dataset:D, Super Pair (T_A, T_B), Parameters:δ
Output A set P of interesting tripoles with (T_A, T_B) as super pair
1: $Q \leftarrow$ All time series T_i s.t. $max(|corr(T_i, T_A)|, |corr(T_i, T_B)|) < |corr(T_A, T_B)|$
2: $C_1 \leftarrow$ All tripoles $\Delta_i \equiv (T_i : T_A, T_B)$ s.t. $T_i \in Q$ and $\delta_{\Delta_i} \geq \delta$
3: $C_2 \leftarrow$ All tripoles $\Delta_i \equiv (T_A : T_i, T_B)$ s.t. $T_i \in Q$ and $\delta_{\Delta_i} \geq \delta$
4: $C_3 \leftarrow$ All tripoles $\Delta_i \equiv (T_B : T_i, T_A)$ s.t. $T_i \in Q$ and $\delta_{\Delta_i} \geq \delta$
5: $P = C_1 \cup C_2 \cup C_3$
6: **return** P

reduction in the running time however with some loss of completeness in the discovered interesting tripoles. Algorithm 6 provides an outline of the CONTRaFast approach. The only difference between CONTRaFast and CONTRaComplete 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, denoted by $Z(T_A)$ and $Z(T_B)$, which are the sets of time-series that have a correlation greater than κ with T_A and T_B. Note that κ is an algorithm-specific threshold of CONTRaFast, that helps in pruning redundant super pairs. The basic intuition of using κ is that two super pairs that are highly similar tend to participate in similar tripole patterns. However, by using κ, we may miss the discovery of some interesting tripoles at the boundary that are deemed similar but are actually non-redundant tripoles. A high value of $\kappa > \tau$ ensures a better completeness in the discovered tripoles. In fact, using $\kappa = 1$ in CONTRaFast is equivalent to the CONTRaComplete algorithm, which is guaranteed to produce a complete set of interesting tripoles. We show in Section 4.3 that the use of κ provides significant reductions in compute time. In the remainder of this paper, we will refer to the CONTRaFast approach with the generic name CONTRa.

Algorithm 6 CONTRaFast

Input Dataset:$\mathcal{D}$, Parameters: δ,τ,κ
Output A non-redundant and complete set of interesting tripoles, C.
1: Initialize $C \leftarrow \phi$
2: Select lowest value of $s_\delta \in [0, 1]$ s.t. $\frac{s_\delta^2(1+s_\delta)}{(1-s_\delta)} \geq \delta$ ▷ Using Corollary 3.3
3: $L \leftarrow$ FIND CANDIDATE SUPER PAIRS($\mathcal{D}$,s_δ)
4: $L \leftarrow$ FIND NON-REDUNDANT SUPER PAIRS(L,κ)
5: **for** each super pair $(T_A, T_B) \in L$ **do**
6: $P \leftarrow$ FIND TRIPOLES FOR SUPER PAIR($\mathcal{D}$,(T_A, T_B),δ)
7: $C \leftarrow C \cup P$
8: **end for**
9: $C \leftarrow$ GET NON-REDUNDANT TRIPOLES(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.

Algorithm 7 FIND NON-REDUNDANT SUPER PAIRS

Input A set of super-pairs:E, Parameter:κ
Output A set L of non-redundant super pairs
1: Initialize $L \leftarrow \phi$
2: **while** $E \neq \phi$ **do**
3: Let $(T_A, T_B) \leftarrow$ most negatively correlated pair in E
4: Add (T_A, T_B) to L
5: $Z(T_A)$, $Z(T_B) \leftarrow$ time series having correlation $\geq \kappa$ with T_A,T_B
6: Remove all cross pairs between $Z(T_A)$ and $Z(T_B)$ from E
7: **end while**
8: **return** L

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 internal parameter κ used in the CONTRaFast algorithm. 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 will allow one to obtain tripoles with lower jump, but this will also increase the computational cost as the search space of interesting tripoles is expected to be relatively large. Redundancy threshold τ controls the redundancy among the discovered tripoles and could vary across applications. In a spatio-temporal data that inherently exhibits spatial autocorrelation, setting τ to higher (lower) values will result in interesting tripoles at a finer (coarser) spatial resolution, but will also require higher (lower) computational time. The internal parameter κ of CONTRaFast determines the completeness of the search, as setting κ to 1 ensures that the search is complete, while the lower values κ will result in a faster search that is only approximately complete. In our experiments, the jump threshold δ and the redundancy threshold τ were set to 0.15 and 0.7 respectively for both the datasets. For brain fMRI dataset, we set κ to 1, whereas for SLP dataset being much larger, we set it to 0.99.

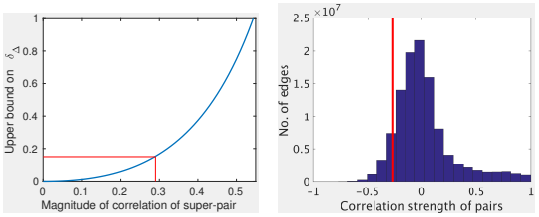
For the above parameter setting, we obtained 506 interesting negative tripoles in SLP dataset and 1015 interesting positive tripoles in one of the brain fMRI scans collected when the subject was resting.

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 Corollary 3.3) 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 Corollary 3.2, 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 for $\delta \geq 0.0903$. Therefore, for $\delta = 0.15$, for a negative tripole to show jump ≥ 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 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 δ 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.

4.3.2 Impact of parameter κ. : We next demonstrate the impact of parameter κ on computational efficiency and completeness of



(a) Relationship between s and upper bound on δ_Δ **(b) Distribution of pairwise correlations in SLP dataset**

Figure 5: Impact of super-pair pruning

CONTRa using the following the 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.

κ	*Cost*	*MissFrac*
0.7	9 sec	0.35
0.8	11 sec	0.27
0.9	26 sec	0.13
0.95	82 sec	0.07
0.97	220 sec	0.05
0.99	5978 sec	0.01
0.999	> 4 days	0 **(reference)**

Table 1: *Cost* and *MissFrac* of CONTRa for finding interesting negative tripoles in SLP dataset at different values of κ

Table 1 shows the *Cost* and *MissFrac* of CONTRa for finding interesting negative tripoles in SLP dataset for different values of κ. As shown in the table, as κ 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 κ 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 $\kappa = 0.99$ relative to the output set obtained 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 κ to 0.99 allows us to find almost all the interesting tripoles in a much smaller time.

4.4 Evaluation of Discovered Tripoles

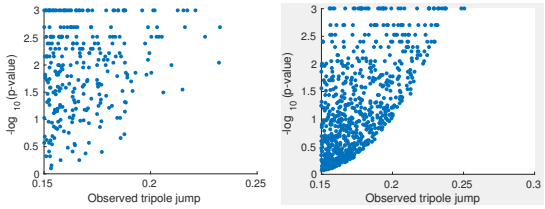
The concept of a tripole relationship, to the best of our knowledge, has not been explored before and so no ground truth is available regarding the existence of such relationships in any domain. Due to this, it is challenging to evaluate the validity of the tripole relationships. We pursued two directions 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_{rand} and recomputed the jump δ_{rand} of the tripole $\Delta_{rand} \equiv (T_0 : T_1, T_{rand})$. The random time series was generated such that the original correlation between the leaf and root time series was preserved, i.e. $corr(T_0, T_{rand}) \in (b-0.01, b+0.01)$, where $b = corr(T_0, T_2)$. This process was repeated 10000 times to obtain a distribution of the δ_{rand}, and thus obtain a p-value of the original tripole jump δ_Δ.

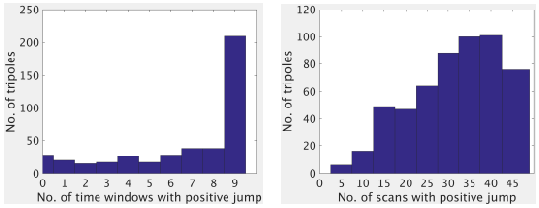
Using the above procedure, we evaluated the jumps of all discovered tripoles in SLP dataset and fMRI dataset by computing their p-values as shown in Figures 6(a) and 6(b). At a jump threshold of 0.15 and 0.05 level of significance, we found 438 out of 506 discovered negative tripoles in SLP dataset that 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.



(a) SLP Dataset **(b) fMRI dataset**

Figure 6: $-log_{10}$(p-value) (Y-axis) vs tripole jump (X-axis) for all discovered tripoles

4.4.2 ***Reproducibility of Tripoles in Independent Datasets:*** Another way to study the validity of the discovered tripoles is to check if the tripoles discovered in one dataset are also reproducible in an independent dataset, i.e. they showed positive jumps. Following this idea, we used the monthly Sea Level Pressure data provided by Hadley Center (HadSLP2) at a spatial resolution of 5×10 degree resolution (that were interpolated to the spatial resolution of 2.5×2.5 degree of the original NCEP2 dataset) for the time duration



(a) SLP Dataset **(b) fMRI dataset**

Figure 7: Number of significant tripoles showing positive jumps in k testing datasets

1901-1976. We divided this time duration into nine overlapping time windows, each of size 36 years. The different windows are: 1901-1936, 1905-1941, . . ., 1941-1976. For each of the 438 tripoles in SLP dataset that showed significant jumps, we computed its tripole jump across the nine time windows we defined in the HadSLP2 dataset. Figure 7(a) shows the number of tripoles that showed a positive jump in k windows for $k \in [0, 9]$. Almost half of the discovered significant tripoles (211 in total) showed a positive jump in all the time windows, which is unlikely to happen for a spurious tripole.

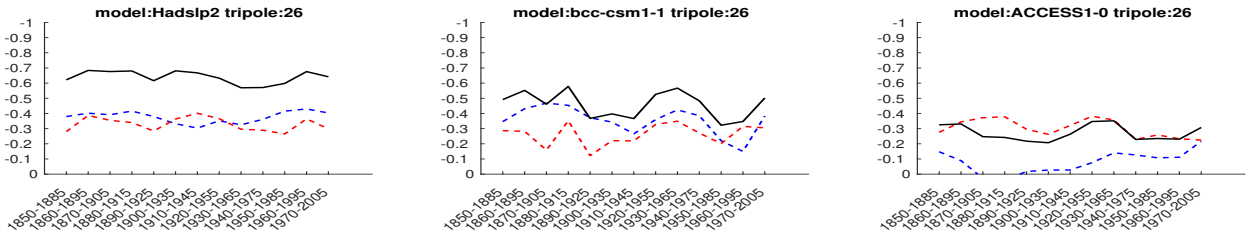
Similar analysis was performed for 547 positive tripoles that were found to show significant jumps in fMRI scan data of one of the subjects, by studying their behavior in 49 independent brain fMRI scans of the same subject that were observed in different time periods. At least 25% of discovered significant tripoles (141 in total) showed a positive jump in more than 40 out of 49 scans.

5 PHYSICAL INTERPRETATION OF TRIPOLES

Results on evaluation of discovered tripoles indicate the existence of a large number of tripoles that are statistically significant and are reproducible in multiple independent datasets, which strongly suggest that these tripoles could be the potential signatures of the real phenomena. Further validation and study of these tripoles by domain experts could possibly explain the phenomenon that results in the manifestation of these tripoles. In this section, we present two case studies on physical interpretation of two of the discovered tripoles in SLP and brain fMRI data.

5.1 Discovery of a New Climate Teleconnection

One of the interesting negative tripoles found by our algorithm resulted in the discovery of a new phenomenon in climate science domain[11]. This negative tripole is shown in Figure 4(a). The root of the tripole lies in the northwestern Russia region, whereas the two leaves of the tripole are in close proximity to the two centres of action of El-Nino Southern Oscillation (ENSO), one of the most widely studied dipole teleconnections in the climate domain. The strength of the tripole, measured as the correlation between the SLP anomalies observed at West Siberian Plain and the sum of SLP anomalies at the two ends of ENSO, was found to be -0.53, which is significantly stronger than either of the pairwise correlations of the Russian region with either of the two centres of ENSO. Moreover, the strength of the tripole was also found to be consistently high in all the previous time windows during 1901-1976. As discussed in



(a) HadSLP2 (b) Climate model: bcc-csm (c) Climate model: ACCESSS1-0

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.

[11], we found out that the sum of anomaly time series at two ends of ENSO potentially represent the background state of ENSO, and the tripole pattern reflects the connection between the background state of ENSO and the anomalies observed at the West Siberian Plain. This connection is attributed to a wave train which originates from subtropical Atlantic and propagates north-eastward towards the north of the West Siberian plain, where it gets further deflected southeastward and travels all the way to central Pacific ocean around Tahiti and modifies the background state of ENSO. We refer readers to [11] for further details about the phenomenon.

As discussed above, tripoles discovered in observations data often represent underlying processes. The ability of climate models (that 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 climate 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.

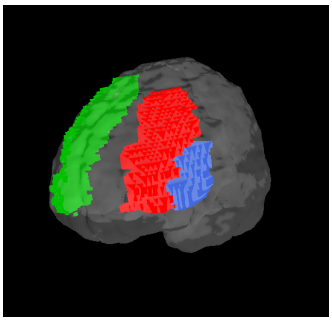


Figure 9: A tripole in brain fMRI data that is discriminative between resting state and the audio-visual task state of the subject.

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 $\sigma_\Delta \geq 0.5$ and a jump $\delta_\Delta \geq 0.2$ 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 function [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 underlying *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 time series data mining literature. Most of the existing work is focused on studying relationships between a pair of time series [8, 19, 20]. In particular, 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 Analysis (ICA) have also been applied to identify regions with strong correlations between their time series [5, 17].

Another relevant body of literature is that of time series 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 the accuracy of the prediction, that is equivalent to the strength of the relationship, and do not consider any notion of jump, which is the primary interestingness measure of the proposed definition of tripoles. 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 *tripole* 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 [11].

The tripole relationship can be further generalized along multiple dimensions. A natural extension would be to pursue multi-pole type 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 tripole. 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

[1] Traffic Forecasting and Analysis Data, Minnesota Department of Transportation. http://www.dot.state.mn.us/traffic/data/reports-hrvol-atr.html. (????). Accessed: 2017-02-16.
[2] Jeffrey S Anderson, Michael A Ferguson, Melissa Lopez-Larson, and Deborah Yurgelun-Todd. 2011. Reproducibility of single-subject functional connectivity measurements. *American journal of neuroradiology* 32, 3 (2011), 548–555.
[3] Gowtham Atluri, Angus MacDonald III, Kelvin O Lim, and Vipin Kumar. 2016. The Brain-Network Paradigm: Using Functional Imaging Data to Study How the Brain Works. *Computer* 49, 10 (2016), 65–71.
[4] Gowtham Atluri, Michael Steinbach, Kelvin O Lim, Vipin Kumar, and Angus MacDonald. 2015. Connectivity cluster analysis for discovering discriminative subnetworks in schizophrenia. *Human brain mapping* 36, 2 (2015), 756–767.
[5] Michael Barnathan, Vasileios Megalooikonomou, Christos Faloutsos, Scott Faro, and Feroze B Mohamed. 2011. TWave: high-order analysis of functional MRI. *Neuroimage* 58, 2 (2011), 537–548.
[6] Soumyadeep Chatterjee, Karsten Steinhaeuser, Arindam Banerjee, Snigdhansu Chatterjee, and Auroop R Ganguly. 2012. Sparse Group Lasso: Consistency and Climate Applications.. In *SDM*. SIAM, 47–58.
[7] Ian Davidson, Sean Gilpin, Owen Carmichael, and Peter Walker. 2013. Network discovery via constrained tensor analysis of fmri data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 194–202.
[8] Jaya Kawale, Stefan Liess, Arjun Kumar, Michael Steinbach, Peter Snyder, Vipin Kumar, Auroop R Ganguly, Nagiza F Samatova, and Fredrick Semazzi. 2013. A graph-based approach to find teleconnections in climate data. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 6, 3 (2013), 158–179.
[9] Jaya Kawale, Michael Steinbach, and Vipin Kumar. 2011. Discovering Dynamic Dipoles in Climate Data.. In *SDM*. SIAM, 107–118.
[10] Robert Kistler, William Collins, Suranjana Saha, Glenn White, John Woollen, Eugenia Kalnay, and others. 2001. The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bulletin of the American Meteorological society* 82, 2 (2001), 247–267.
[11] Stefan Liess, Saurabh Agrawal, Snigdhansu Chatterjee, and Vipin Kumar. 2017. A Teleconnection between the West Siberian Plain and the ENSO Region. *Journal of Climate* 30, 1 (2017), 301–315.
[12] Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. 2009. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 587–596.
[13] Russell A Poldrack and Krzysztof J Gorgolewski. 2015. OpenfMRI: open sharing of task fMRI data. *NeuroImage* (2015).
[14] Lizabeth M Romanski. 2007. Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. *Cerebral Cortex* 17 (2007), i61–i69.
[15] F Siegert, G Ruecker, A Hinrichs, and AA Hoffmann. 2001. Increased damage from fires in logged forests during droughts caused by El Nino. *Nature* 414, 6862 (2001), 437–440.
[16] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 1 (2002), 273–289.
[17] John M Wallace and David S Gutzler. 1981. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review* 109, 4 (1981), 784–812.
[18] Philip J Ward, Brenden Jongman, Matti Kummu, Michael D Dettinger, Frederiek C Sperna Weiland, and Hessel C Winsemius. 2014. Strong influence of El Nino Southern Oscillation on flood risk around the world. *Proceedings of the National Academy of Sciences* 111, 44 (2014), 15659–15664.
[19] Hui Xiong, Mark Brodie, and Sheng Ma. 2006. Top-cop: Mining top-k strongly correlated pairs in large databases. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 1162–1166.
[20] Hui Xiong, Shashi Shekhar, Pang-Ning Tan, and Vipin Kumar. 2004. Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 334–343.