Analyzing Global Climate System Using Graph Based Anomaly Detection

Kamalika Das,[†] Saurabh Agrawal,[‡] Gowtham Atluri,[‡] Stefan Leiss,[‡] Vipin Kumar[‡]

[†]UARC, NASA Ames Research Center

[‡]University of Minnesota, Twin Cities

AGU Fall Meeting 2014

Roadmap

Problem statement

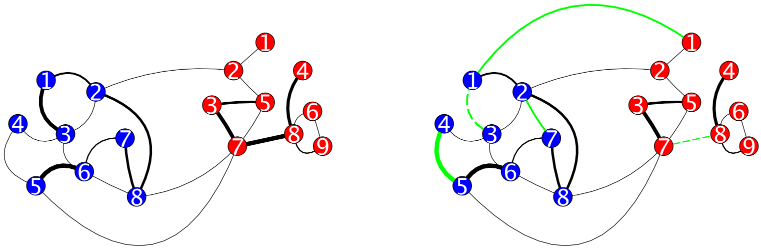
Given a time-varying sequence of weighted graphs $G_1, G_2, \ldots, G_T$:

1. Identify if any single transition G_t to G_{t+1} is anomalous
2. If yes, identify which edge relationship changes were responsible for the anomalous transition
3. Identifying abnormal climate patterns over time by analyzing anomalous nodes and edges in time graphs

What do we mean by anomalous edge changes?

- **Case 1:** high magnitude change (increase or decrease) in edge weight from time t to $t+1$.
- **Case 2**: new edges that bring distant nodes closer.
- **Case 3:** decrease in edge weight (or deletion of edges) between central or bridge nodes in the graph that push proximal nodes far apart.

Running example



- *S1: New edge between* b_1, r_1 (refers to Case 2)
- *S2: Small decrease in edge weight between* r_7, r_8 (refers to Case 3)
- *S3: Large increase in edge weight between* b_4, b_5 (refers to Case 1)
- *S4: Small decrease in edge weight between* b_1, b_3
- *S5: New edge between* b_2, b_7

Distance function

- $\bar{d}_S(G, H)$: a generic notion of distance that captures structural differences due to abnormal changes in the edges in the complimentary set $E - S$
- For a dissimilarity threshold δ, G and H considered similar with respect to edge set $E - S$ at level δ if $\bar{d}_S(G, H) < \delta$
- If $\bar{d}_S(G_t, G_{t+1}) < \delta$ for some subset S, then $E_t \subseteq S$

Optimization problem

$$E_t \quad := \quad \arg\min_S \quad |S| \qquad \text{subject to} \quad \bar{d}_S(G_t, G_{t+1}) < \delta. \tag{1}$$

Polynomial time solution

- (1) is a combinatorial optimization problem. Intractable for large graphs.
- Can be reduced to polynomial time if, for any $S \subseteq E$:

$$\bar{d}_S(G_t, G_{t+1}) = \sum_{e \in E-S} \Delta E_t(e), \tag{2}$$

where $\Delta E_t(e)$ is a non-negative functional of the graphs G_t and G_{t+1} independent of the set S

Proposed metric

$$\bar{d}_S^{(0)}(G_t, G_{t+1}) = \sum_{e \in E-S} \Delta E_t(e),$$

where $\Delta E_t(e)$ for $e = e_{i,j}$is given by

$$\Delta E_t(e_{i,j}) = |A_{t+1}(i,j) - A_t(i,j)| \times |d_{t+1}(i,j) - d_t(i,j)|.$$

Proposal for distance function

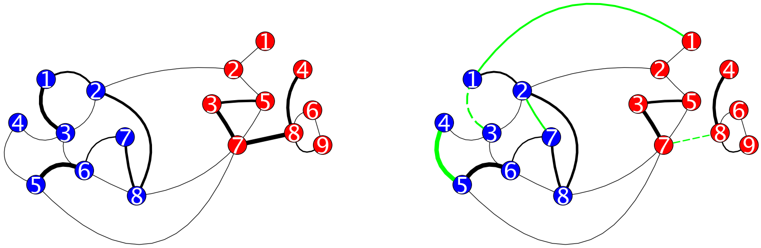
Anomalous edges:

1. Large changes in magnitude (Case 1):
   - $|A_{t+1}(i,j) - A_t(i,j)|$ will be large
   - Will result in $|d_{t+1}(i,j) - d_t(i,j)|$ being large
2. New edges / dissolving edges (Case 2/3):
   - $|d_{t+1}(i,j) - d_t(i,j)|$ will be large
   - $A_{t+1}(i,j) - A_t(i,j)$ will be non-zero

Non-anomalous edges:

1. Small magnitude changes between node-pairs i,j that are tightly coupled:
   - $|A_{t+1}(i,j) - A_t(i,j)|$ will be small
   - $|d_{t+1}(i,j) - d_t(i,j)|$ will also be small
2. Neighboring edges of new edges / dissolving edges (Case 2/3):
   - For some neighboring node of i (say n_i) and j (say n_j):
   - $|d_{t+1}(i,j) - d_t(i,j)|$ will be large
   - But, $|A_{t+1}(i,j) - A_t(i,j)|$ will be small (possibly 0)

Performance of distance metric



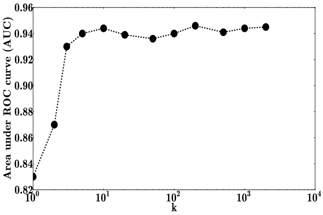
Edge	b_1, r_1	b_4, b_5	r_7, r_8
$\Delta E_i(.)$	10.6	9.56	8.99
Edge	b_1, b_3	b_2, b_7	Rest
$\Delta E_i(.)$	0.1	0.22	0

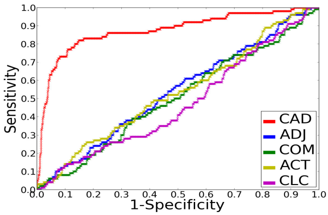
Table : Table listing the values of $\Delta E_t(.)$ for edges in the illustrative example.

Anomaly detection performance on synthetic data set

- Random realization of 4-component Gaussian mixtures (matrix P) at time t
- Sum of random perturbation of P (matrix Q) with matrix R, where

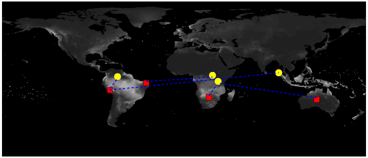
$$R(i,j) = \begin{cases} 0 & \text{with probability } p = 0.95 \\ \text{u(i,j)} & \text{with probability } p = 0.05, \end{cases}$$





Results on precipitation (PRE) network for different time transitions

- 67,420 nodes, monthly precipitation aggregates, top-10 neighbor graph
- Analysis results for January for the 1994-1995 transition



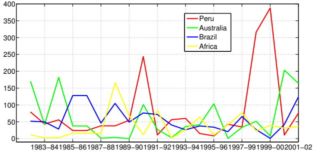
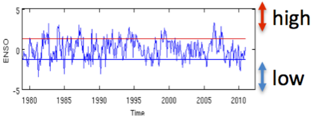


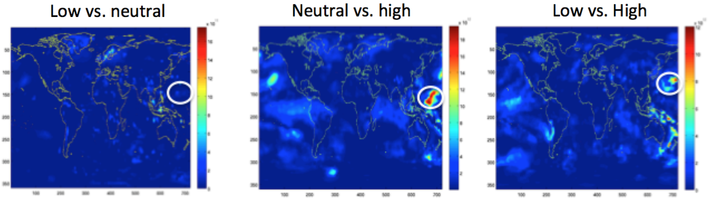
Figure : Heat map of rainfall for January 1995. Red squares and yellow circles are nodes associated with anomalous edges (indicated by blue dotted lines) found by CAD

Results on temperature (TAS) network for SOI phases

- Monthly Temperature At Surface (TAS), 1980-2010, 2.5°x2.5° resolution (10512 nodes)
- Preprocessing: Removal of annual seasonality and linear trends followed by z-scoring
  - High ENSO($>$ 1 std. dev)
  - Neutral (within 1 std. dev)
  - Low ENSO($<$ 1 std.dev)



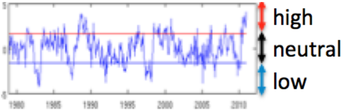
- For each phase, network constructed by computing Pearson correlation between the time series of two grid cells. Only the edges with negative correlations ($<$ -0.3) were retained.



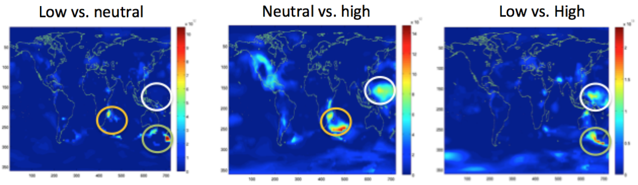
- Most of the anomalous nodes in neutral vs. high and low vs. high found were concentrated in equatorial pacific where ENSO's impact is found
- These nodes are not anomalous in low vs. neutral

Results on pressure (PSL) networks for SOI phases

- Monthly Pressure at Sea level (PSL), 1980-2010, 2.5°x2.5° resolution
- Preprocessing : Removal of annual seasonality and linear trends followed by z-scoring
  - High ENSO($>$ 1 std. dev)
  - Neutral (within 1 std. dev)
  - Low ENSO($<$ 1 std.dev)



- For each phase, network was constructed by computing Pearson correlation between the time series of two grid cells. Only the edges with negative correlations ($<$ -0.3) were retained.



- Region near South Africa behaves similarly in low and high phases of SOI, but not in neutral.
- Region in equatorial pacific behaves similarly in low and neutral phases of SOI, not in high.
- Region near south-east of Australia behaves similarly in neutral and high phases of SOI, not in low phase.

Summary

- We proposed a novel method for localizing abnormal changes in edges that are responsible for anomalous change in structure in dynamic graphs
- CAD tracks changes in edge strength and structure (via commute time distance) in order to determine these anomalies.
- CAD has an $O(n \log n)$ run-time complexity per graph instance for sparse graphs, making it scalable
- Experimental studies on synthetic and large climate datasets showed that CAD consistently and efficiently localizes anomalous edges and associated nodes responsible for anomalous changes in graph structure
- Ongoing work includes more systematic study of the SOI phase transitions honoring the time component of these climate phenomena