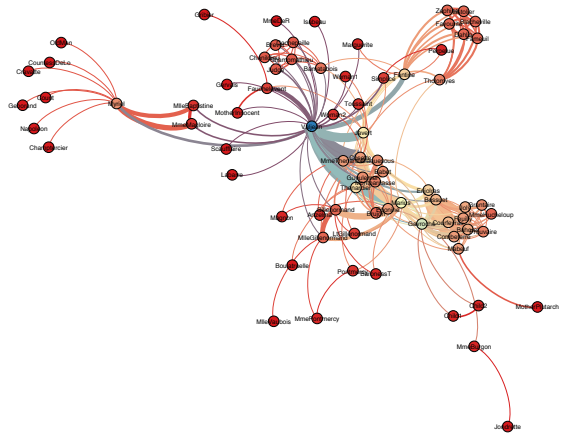
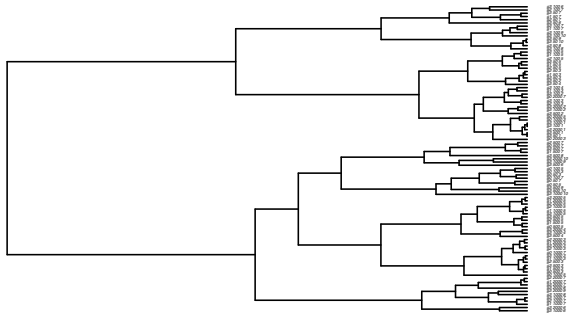
SAND2015-8684PE

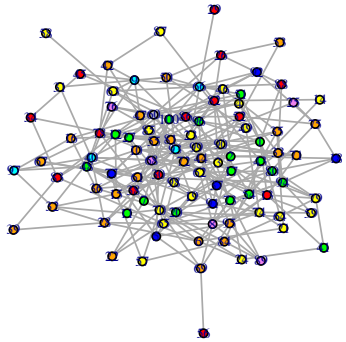
Exceptional service in the national interest

Sandia National Laboratories

G0 Network





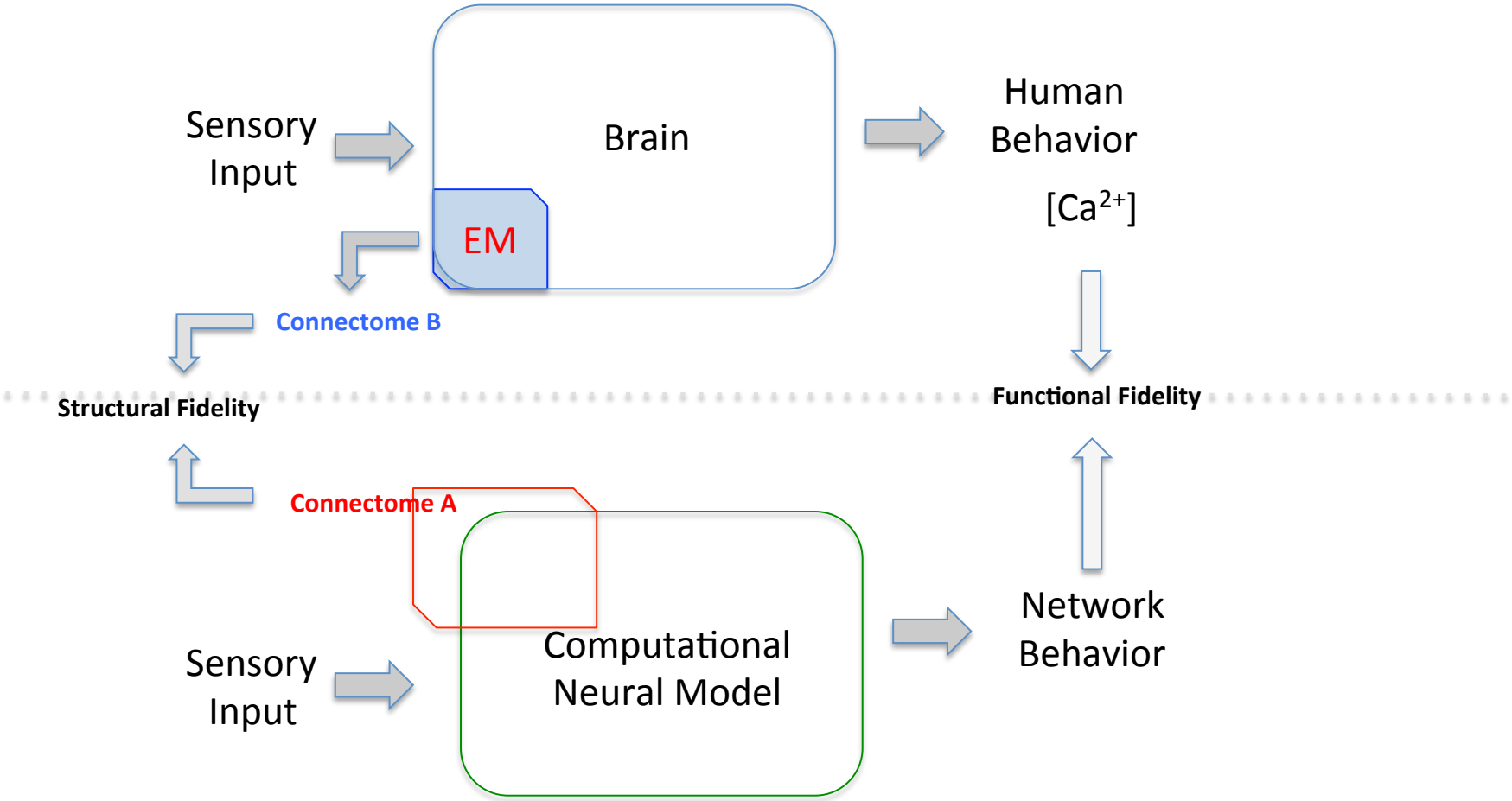


Graphical Bertillonage to Assess Structural and Functional Fidelity of Neuronal Models

David G. Robinson, PhD

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.



- Connectome B is an *individual* characteristic derived from Electromagnetic field data
- Connectome A is a *population* characteristic derived from B
- Not all of Connectome A is used in the CNM

One characterization of Structural Fidelity of CNM is a structural comparisons between Connectome B, Connectome A, and the portion of Connectome A employed by the CNM

Approximate Graph Matching

Sandia National Laboratories

- **Variety of methods, for example:**
  - Kroenecker product
  - Graduated assignment
  - Eigen-decomposition
  - Edit distance, e.g. Hamming, A*, Hausdorff (labeled, attributed, graphs)
  - Spherical approximate matching
  - etc.

- **Methods typically require a *correspondence between nodes*:**
  - Nodes are labeled and the labels have meaning.
  - Node labels are not the same as node attributes.

- **Characteristics of current problems:**
  - Unlabeled nodes
  - Directed/undirected edges
  - Node and/or edge attributes
  - Nodes or edges missing: *graphs may be incomplete*

- **Exact match between graphs is assumed to not be possible; must look for graphs that are similar.**

Uses for Similarity Measures

- **Classification**
  - Is it a cat?
- **Image Retrieval**
  - Show me pictures of cats.
- **Unsupervised segmentation**
  - Which parts of the image are a cat box?

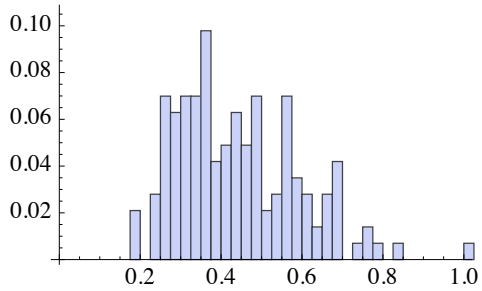
Need a similarity measure based on distribution of features or attributes: shape, color, structure, texture…

Distribution of features will be captured in a histogram or signature.

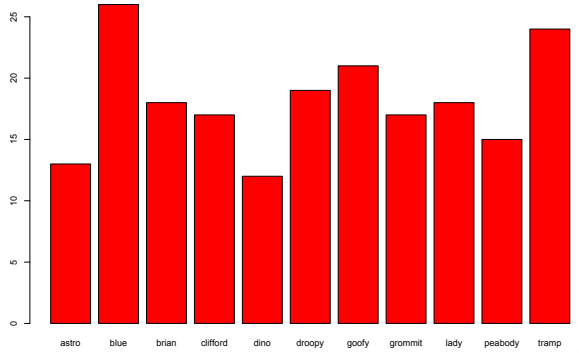
Signatures

- Signatures
  - Similar to histograms but efficiently captures more information
  - Signature: $\{s_j=(m_j, w_{mj})\}$ is a set of graph features where
    - m_j = median/centroid of feature cluster
    - w_{mj} = weight/frequency/count of feature within cluster with centroid m_j
- Signatures can be associated with:
  - Structural properties of graph G, e.g. centrality of all induced subgraphs of radius 2

$$c(G) = \frac{\sum_{i=1}^{n}(d_{max} - d_i)}{(n-2)(n-1)}$$



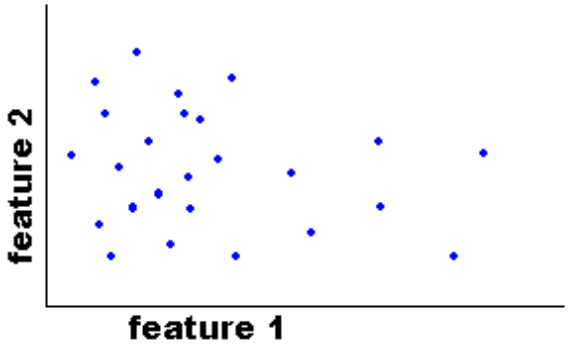
  - Attribute properties of node or edge:
    $a_{1,1...n}$= {Tree Canopy, Grass/shrub, Bare soil, Water, Buildings, Railroad, Other, Area, Aspect ratio}



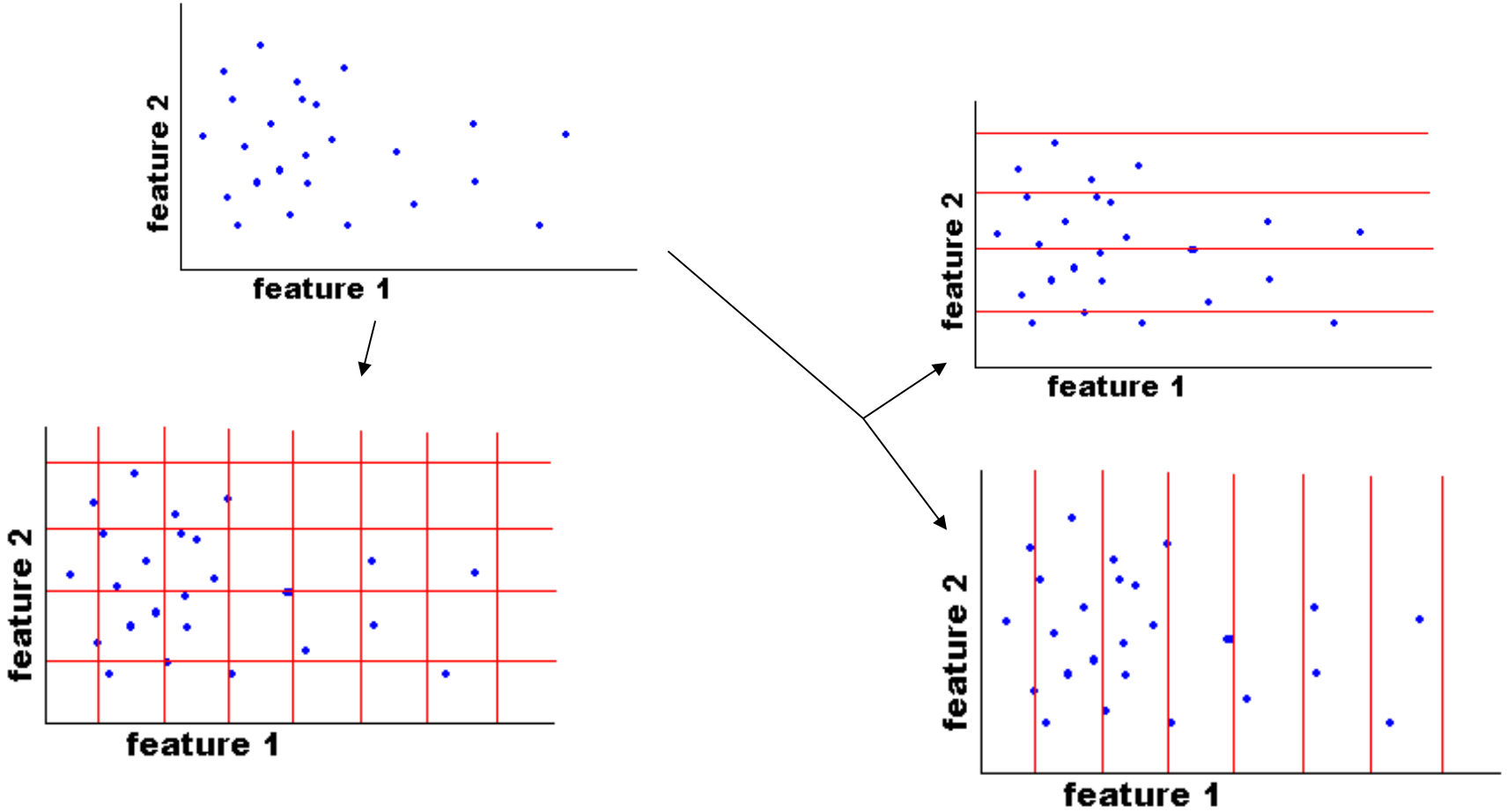
Goal: to compare graphs we need to find a distance metric that characterizes the similarity between signatures.

Joint vs Marginal Histograms

Sandia National Laboratories

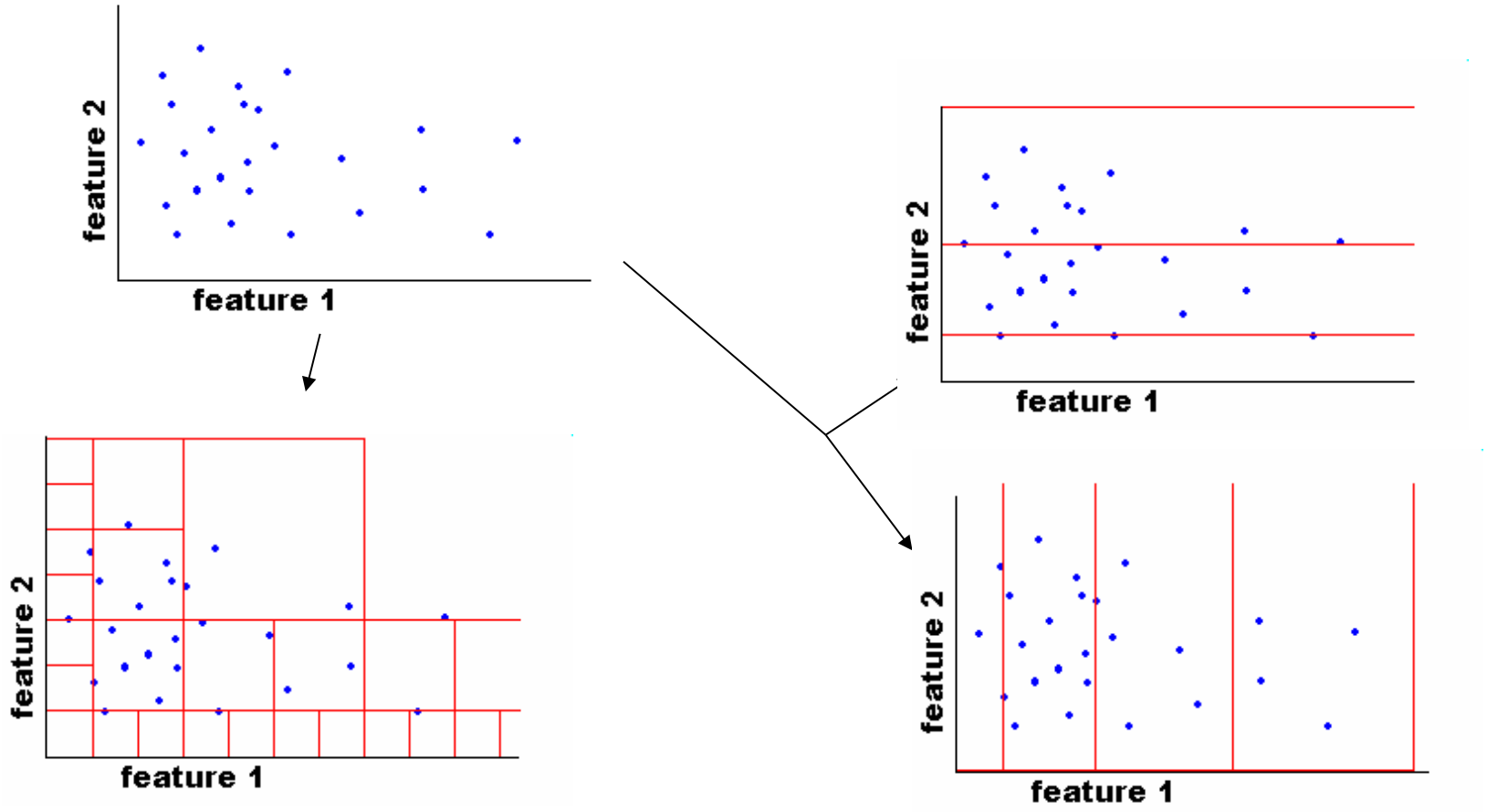


Joint vs Marginal Histograms

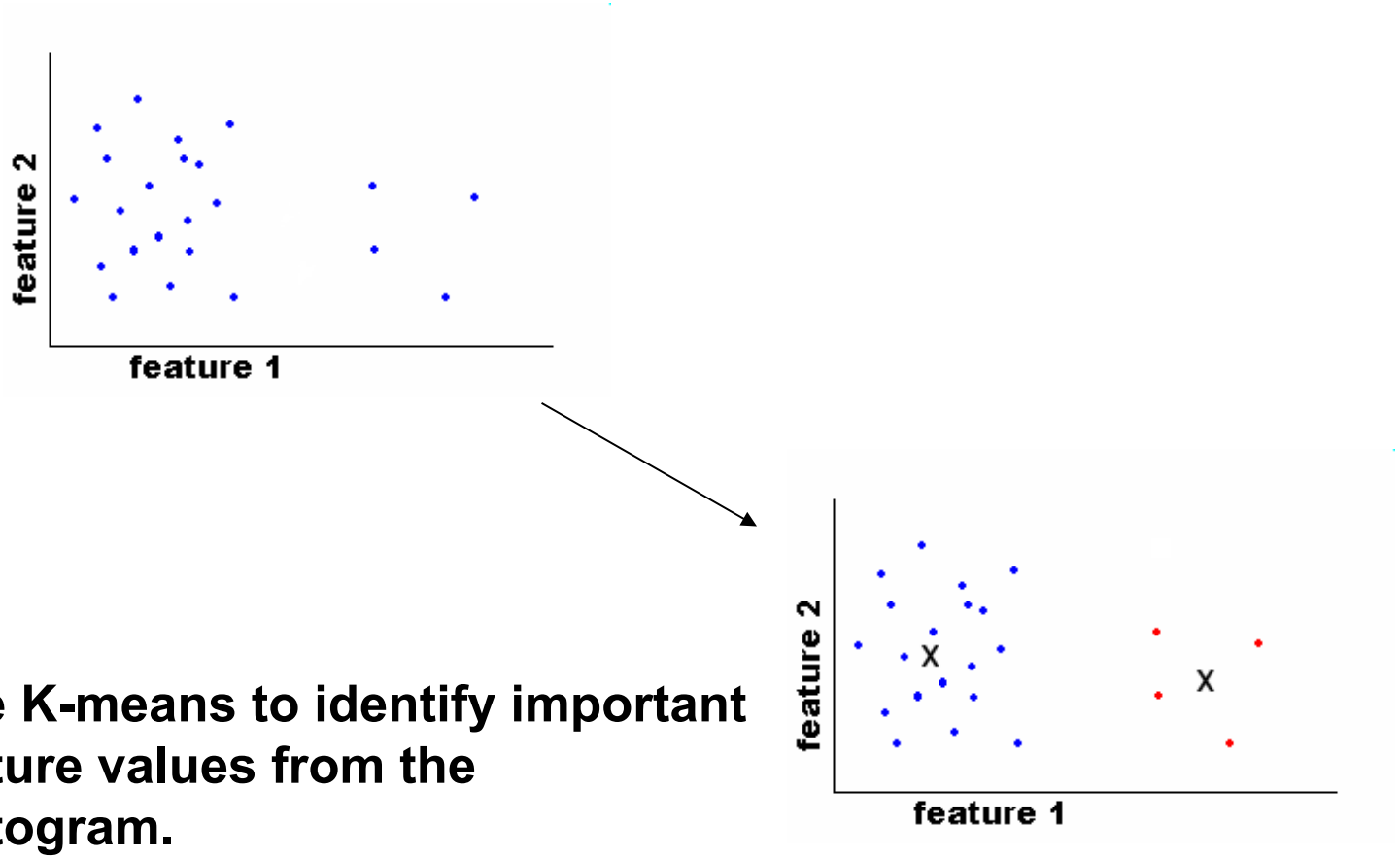


However, only really need bins to be associated with significant elements of the features/attributes

Adaptive Binning



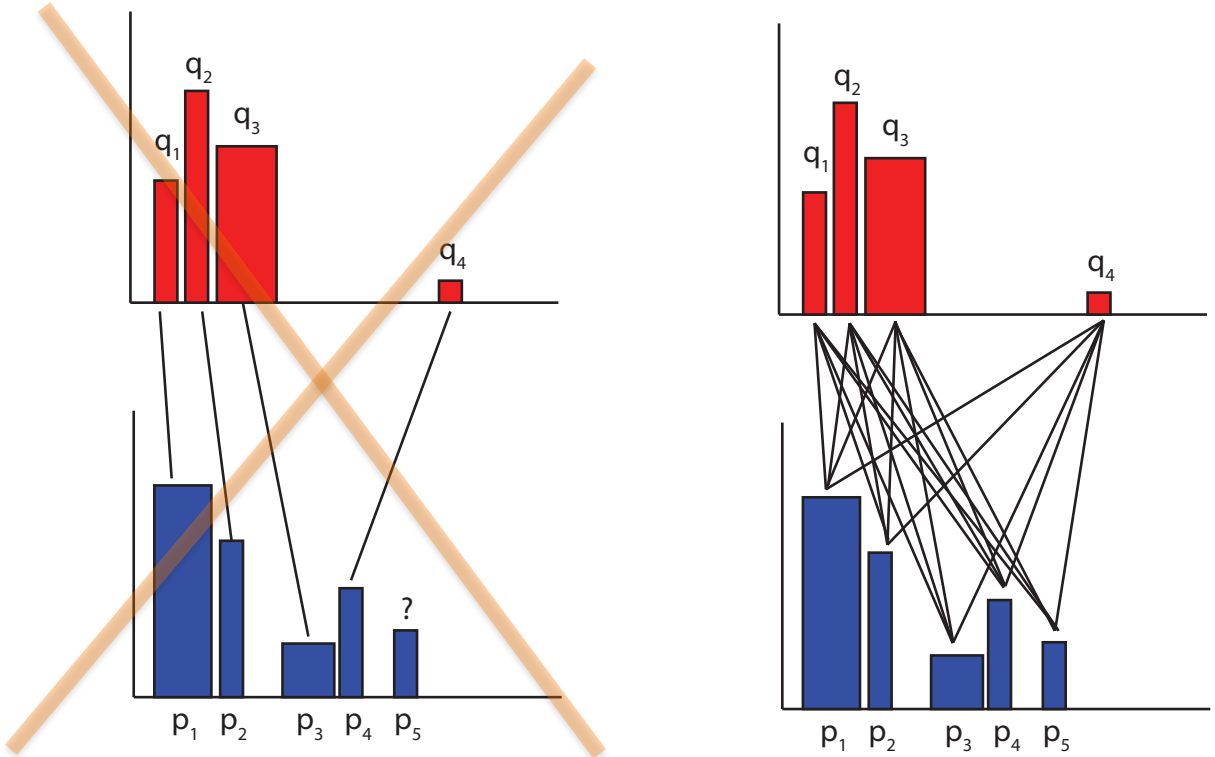
Clusters -> Signatures



- **Use K-means to identify important feature values from the histogram.**
- **Histogram converted to signature.**

Graph Similarity Metric

- Similarity between graphs will be based on the similarity of their signatures
- The distance between signatures will be a measure of their similarity
- Signature comparison is not done cell-by-cell as is done for some histogram comparison (chi-squared metric), but rather across all cells simultaneously using a distance algorithm, e.g. Earth Movers Distance.



Distance Measures

Sandia National Laboratories

- **Heuristic**
  - Minkowski-form
  - Weighted-Mean-Variance (WMV)
- **Nonparametric test statistics**
  - χ^2 (Chi Square)
  - Kolmogorov-Smirnov (KS)
  - Cramer/von Mises (CvM)
- **Information-theory divergences**
  - Kullback-Liebler (KL)
  - Jeffrey-divergence (JD)
- **Ground distance measures**
  - Histogram intersection
  - Quadratic form (QF)
  - Earth Movers Distance (EMD)

[see backup slides for more detail on each metric]

Earth Movers Distance

- EMD is defined for signatures of the form P={(x_1,p_1),...,(x_m,p_m)} and Q={(y_1,q_1),..., (y_m,q_m)} where x_i is the center of cluster *i* and represents the feature of interest, e.g. 'color', and p_i is the weight of cluster *i,* e.g. number of points of that feature type in the cluster.
- Let $F = [f_{ij}]$ represent the flow of material between P_i (supply) to Q_j (demand). Two signatures *P* and *Q* can be compared by finding the flow *F* that minimizes the transportation problem:

$$Work(P,Q;F) = \left(\min_{f_{ij}} \sum_{i,j} f_{ij} d_{ij} \right) \qquad s.t.: \qquad (1)$$

$$f_{ij} > 0 \quad \text{earth can only be moved from P to Q} \qquad (2)$$

$$\sum_j f_{ij} \leq p_i \quad \text{the earth to be moved must be less than what is in P} \qquad (3)$$

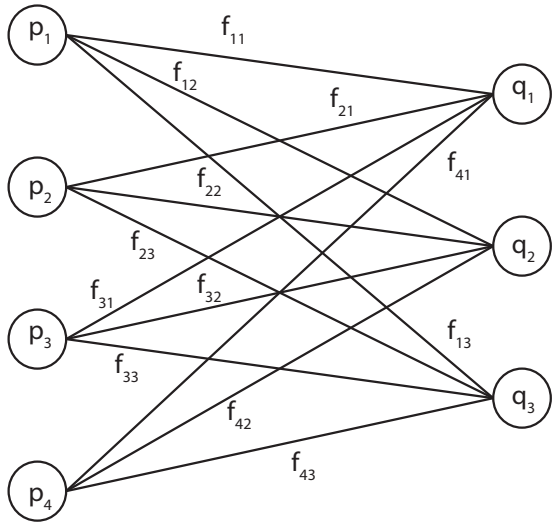
$$\sum_i f_{ij} \leq q_j \quad \text{the earth to be moved must be less than what Q can receive} \qquad (4)$$

$$\sum_i f_{ij} = \min\left(p_i, q_j\right) \quad \text{move the maximum amount of earth} \qquad (5)$$

$$EMD(P,Q;F) = \left(\min_{f_{ij}^*} \sum_{i,j} f_{ij}^* d_{ij} \right) / \sum_{i,j} f_{ij}^*$$

EMD (continued)

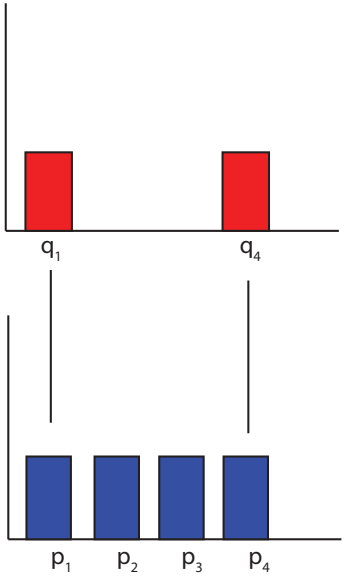
- The ground metric *d* is the distance between features, which is interpreted as the cost of turning a unit mass of one feature into a unit mass of the feature in another signature. Figure depicts a basic transportation problem where $m = 4$, $n = 3$.



- For signatures with the *same total mass* the EMD is a true metric on distributions, and it is identical to the Mallows distance. Note that normalizing signatures with the same mass does not affect their EMD. However, EMD on signatures is not invariant to weight scaling unless both signatures are scaled by the same factor.

Issues w/EMD

- All is not rosy however in the case where the signatures are not of equal mass. If one signature is a partial match for another signature, then a degenerate situation develops.
- Consider the two signatures in Figure; notice that *Q* is a partial match for *P.* the EMD remains zero even with the addition of multiple values of p_i



- Alternative formulation exists that doesn't overcome the above problem, but does account for unequal signatures mass:

$$EMD^*(P,Q;F) = \left(\min_{f_{ij}^*} \sum_{i,j} f_{ij}^* d_{ij}\right) + \left|\sum_i P_i - \sum_j Q_j\right| \times \alpha \max_{i,j}\{d_{ij}\}$$

Advantages

Sandia National Laboratories

- Uses signatures (more efficient than histograms)
- Nearness measure without quantization
- Partial matching
- A true metric (well, almost)

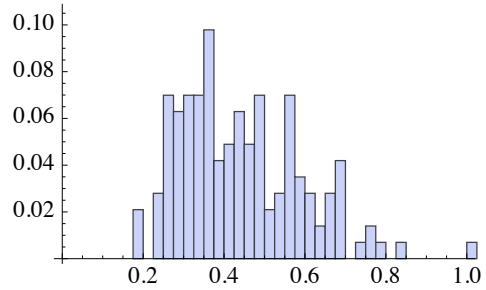
Disadvantage

- High computational cost
  - Not effective for unsupervised segmentation, etc.

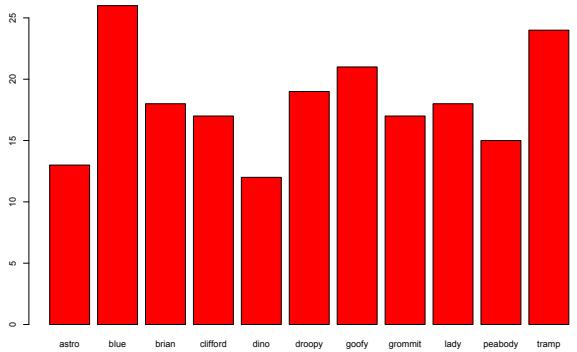
Graph Signatures

- Graph Signatures
  - Similar to histograms but efficiently captures more information
  - Signature: $\{s_j=(m_j, w_{mj})\}$ is a set of graph features where
    - m_j = median/centroid of feature cluster
    - w_{mj} = weight/frequency/count of feature within cluster with centroid m_j
- Signatures can be associated with:
  - Structural properties of graph G, e.g. **centrality of all induced subgraphs of radius 2**

$$c(G) = \frac{\sum_{i=1}^{n}(d_{max} - d_i)}{(n-2)(n-1)}$$



  - Attribute properties of node or edge:
    $a_{1,1...n}$= {Tree Canopy, Grass/shrub, Bare soil, Water, Buildings, Railroad, Other, Area, Aspect ratio}



Goal: need to find a distance metric that characterizes the similarity between signatures. Our focus (for now) is on finding graphs with similar structural characteristics.

Graph Bertillonage: Approximate Graph Matching

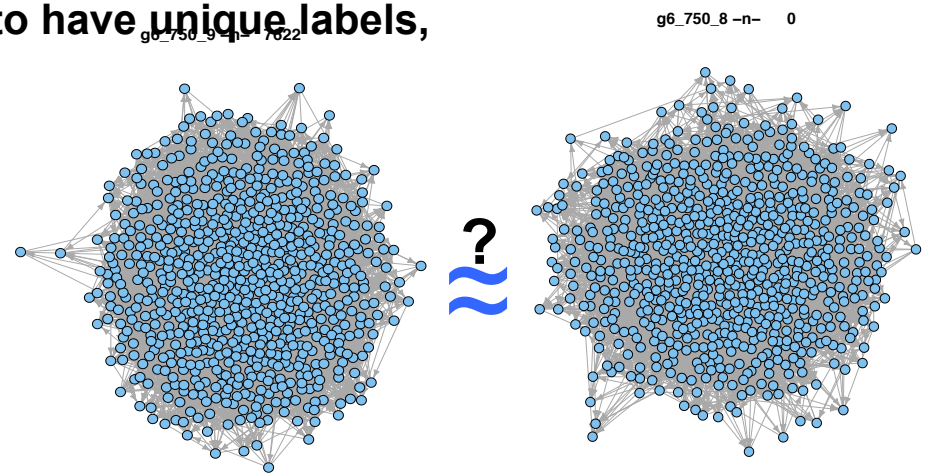
- **Problem:** *How can we statistically characterize the similarity between two graphs?*

- **Why are we interested:**
  - Does graph match something we've seen before?
    - Generation of artificial graphs important area of research for evaluating graph analysis methods. Are the simulated graphs similar to the original?
  - Can we determine what function the graph is performing? Social structure, cyber security, software algorithm ID, etc.
  - Is the graph changing over time?

- **Existing methods generally require a correspondence between nodes, i.e. nodes are required to have unique labels, and require complete graphs**

- **New method is very general**
  - Unlabeled
  - Directed/undirected
  - Only portions of graphs are needed
  - No self-loops!

- **Extension**
  - Semantic graphs: node/edge attributes

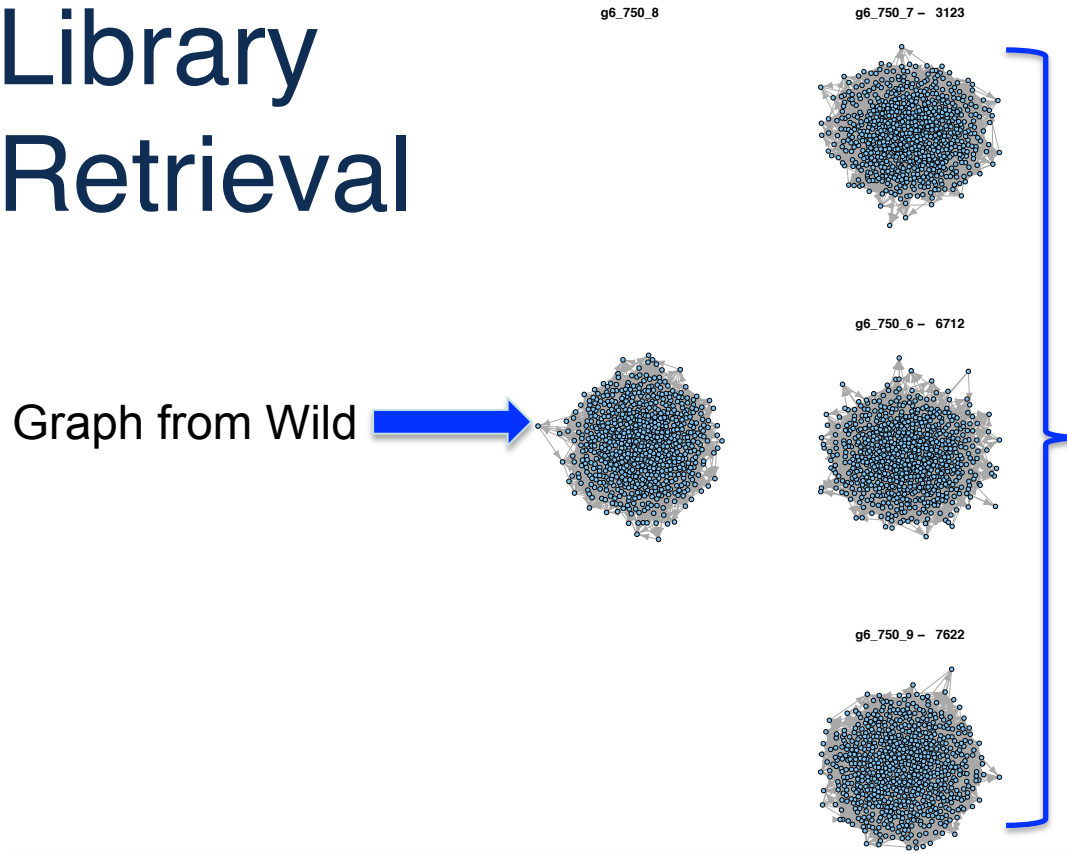


Why 'Bertillonage'?

Sandia National Laboratories

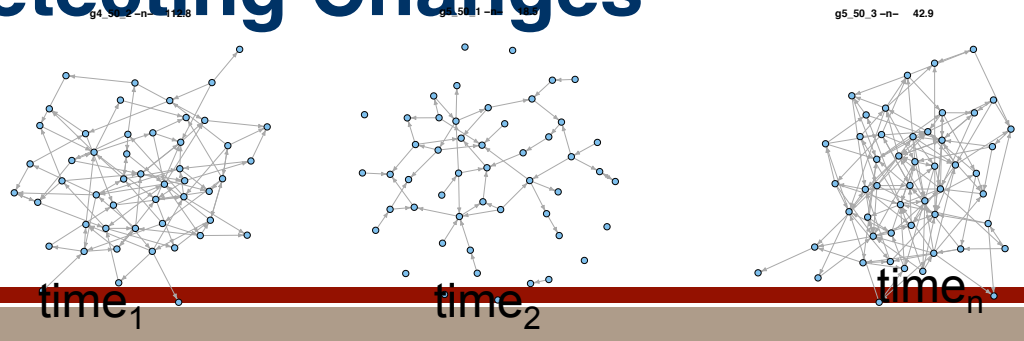
- Bertillonage is a simple forensic analysis technique based on bio-metrics that was developed in 19th century France before the advent of fingerprints.
- Alphonse Bertillon was a French criminologist and anthropologist who created the first system of physical measurements, photography, and record-keeping that police could use to identify recidivist criminals.
- Bertillon developed an anthropometric method based on measurements from head and body, shape of facial features, and individual marks (tattoos, scars, etc.). These characteristics were filed with photographs of the suspects and cross-indexed to permit quick, systematic access.
- The method, referred to as *Bertillonage*, worked well under ideal conditions, but was difficult to implement for a variety of reasons. For example, inaccurate measurements were common for untrained personnel. In addition, suspect characteristics changed with age. it was eventually abandoned in favor of fingerprints.

Library Retrieval



Example: with statistical confidence of 90%, the decompiled control flow chart for the decompiled code similar to algorithms X, Y, Z

Detecting Changes



Example: detecting statistically significant changes in network traffic

Structural Metrics (typical)

Sandia National Laboratories

Closeness centrality (normalized) measures how many steps is required to access every other vertex from a given vertex thus providing an indication of the potential for independent communication. It is the inverse of the average length of the shortest paths to/ from all the other vertices in the graph.

$$c(G) = \frac{|V| - 1}{\sum_{i \neq v} d(v,i)}$$

Betweenness centrality (normalized) measures the potential for a node to control the communication within the network.

$$C_b(G) = \sum_{b<c} \left[\frac{g_{bc(a)}}{g_{bc}(n^2 - 3n + 2)} \right]$$

Leadership measures the degree to which a particular node dominates the connections between nodes.

$$c(G) = \frac{\sum_{i=1}^{n} (d_{max} - d_i)}{(n-2)(n-1)}$$

Diversity captures the topological survivability and variable connectivity. Is the graph dominated by a small number of large, highly connected subgraphs, or is it composed of a large number of very loosely connected nodes (or small subgraphs)?

$$s(G) = \sum_{(i,j) \in E} d_i d_j = \sum_{i \in V} \sum_{j \in V} d_i a_{ij} d_j$$

Neighborhoods

Sandia National Laboratories

- For each node on a graph a subgraph is induced by finding the neighborhood of size two for that node.
- For each induced subgraph, the graph characteristics are calculated
- These graph graph characteristics are collected to form the signatures for each graph metric
- Alternative signature constructions methods are being investigated:
  - Equal bin width (current)
  - K-means to find bin centroids and bin width (testing)

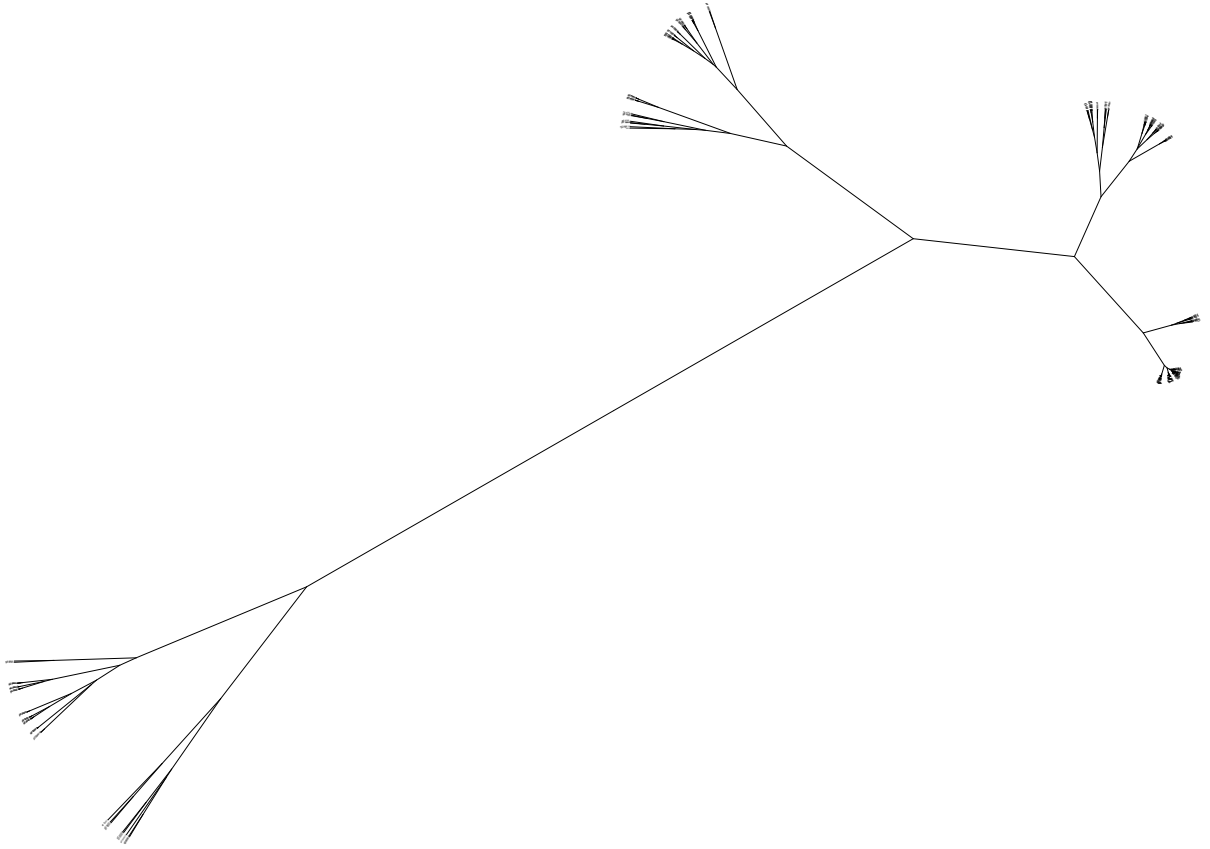
Verification

Sandia National Laboratories

- The goal was to verify the ability of the algorithm to characterize the similarity between graphs.
- A series of random graphs (Erdos-Renyi) were generated of various sizes and connectivity patterns.
- This simulation resulted in a large number of graphs across a spectrum of parameters. Each graph is represented by a code: gn_N_d
  - n= simulation number, i.e. all graphs g#_1000_1 represent random realization of the same E-R graph.
  - N = number of nodes in graph
  - d = expected degree
- Verification involved using GB to characterize the pair-wise similarity between the graphs and cluster them based on the similarity metric.
- The following few slides represent various depictions of the results
- All the graphs present the same results, just in slightly different visual forms. Different applications/users find benefit from different presentations.

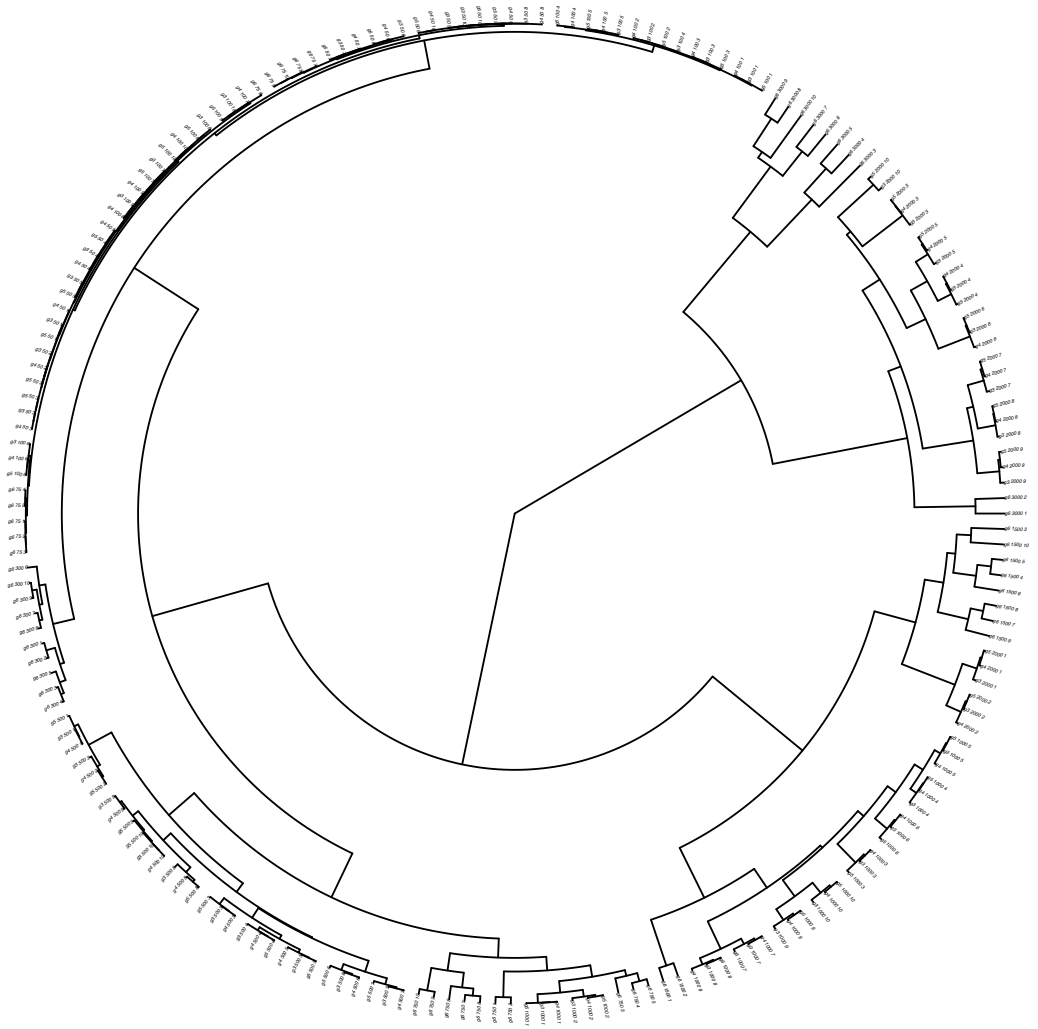
Erdos-Renyi Testing

Sandia National Laboratories



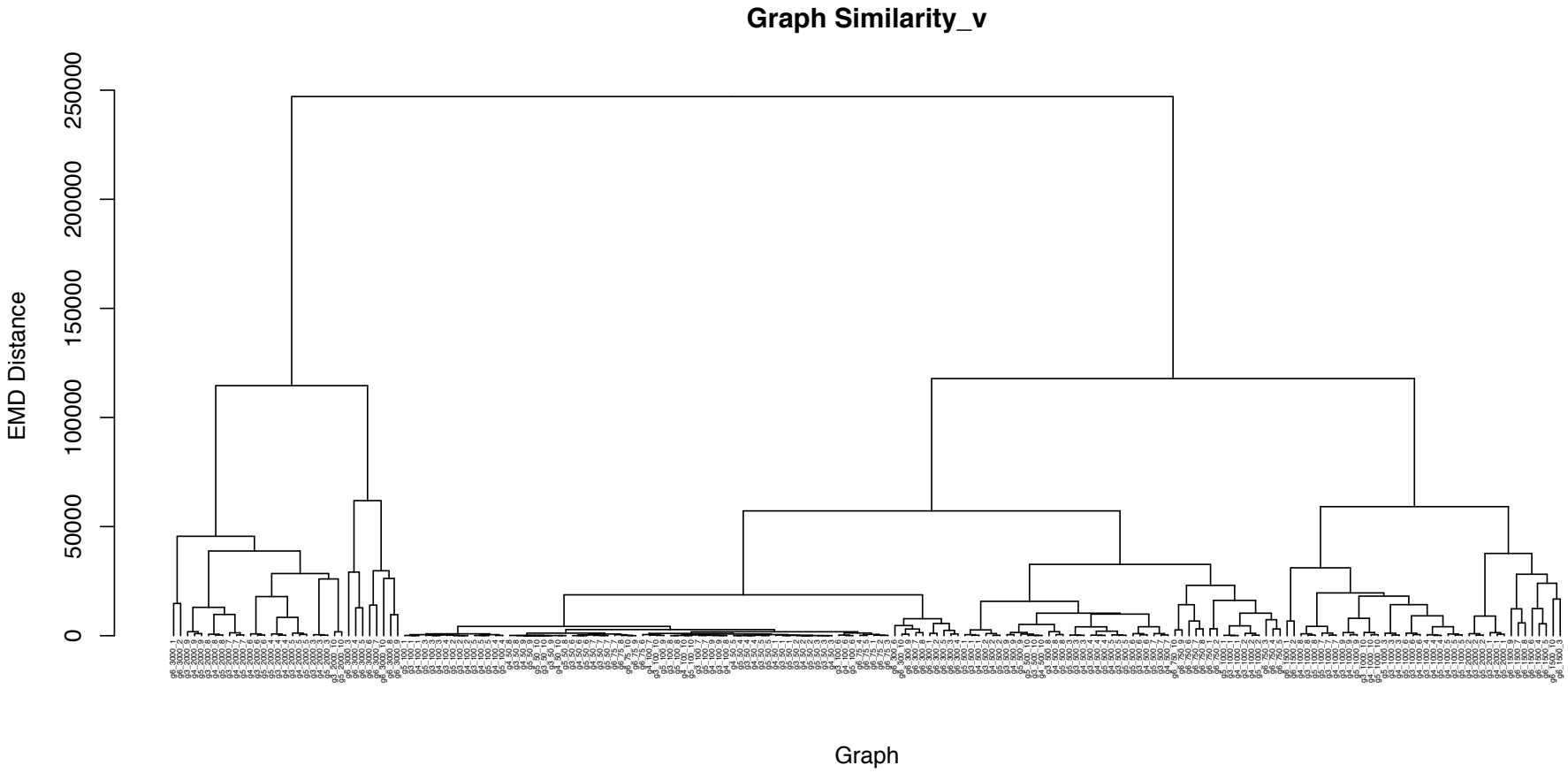
(zoom in on leaves)

Erdos-Renyi (fan)



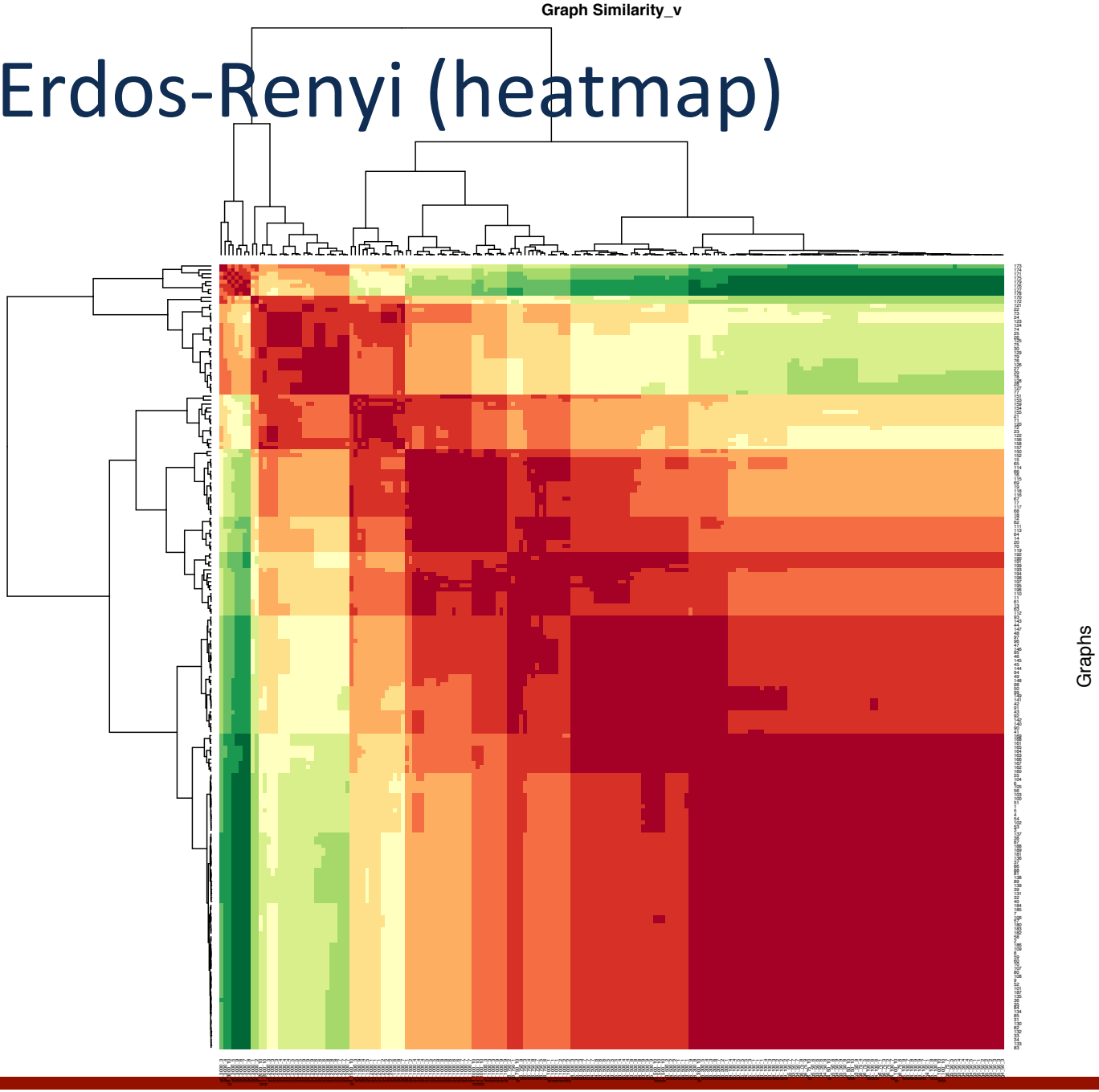
(zoom in on leaves)

Erdos-Renyi (phylo)

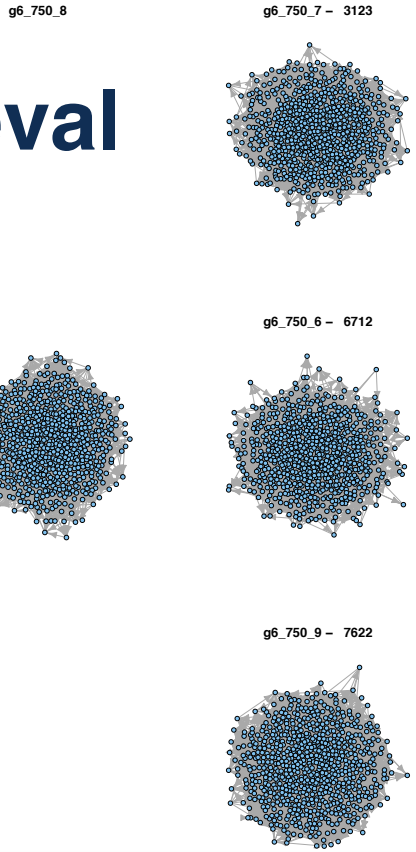


(zoom in on leaves)

Erdos-Renyi (heatmap)



Library Retrieval



Graph from Wild

Similar Graphs found in Library

Example: with statistical confidence of 90%, the decompiled control flow chart for the decompiled code similar to algorithms X, Y, Z

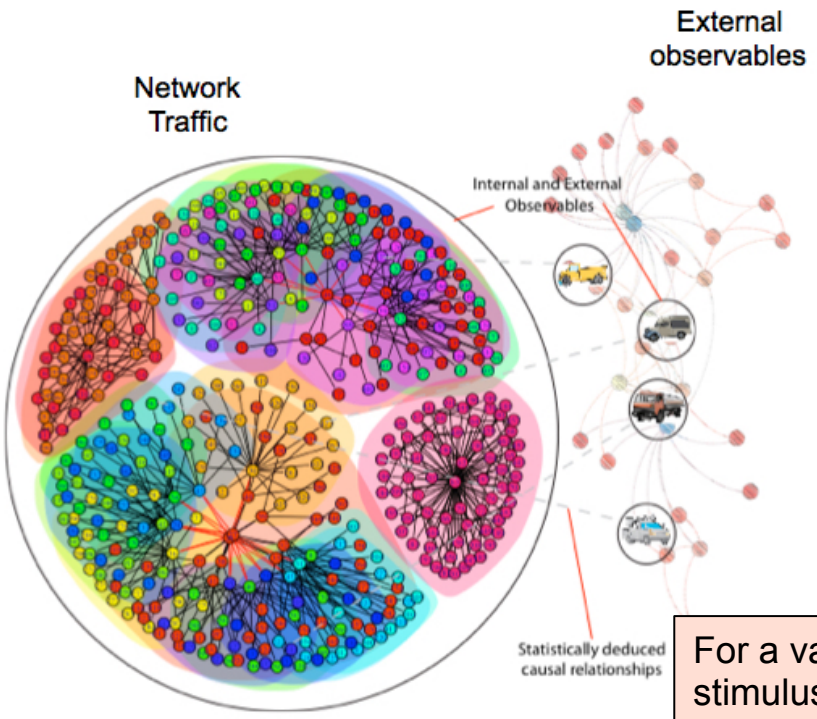
Given an new graph from the wild we can query a historical library to identify graphs that have the similar structure.

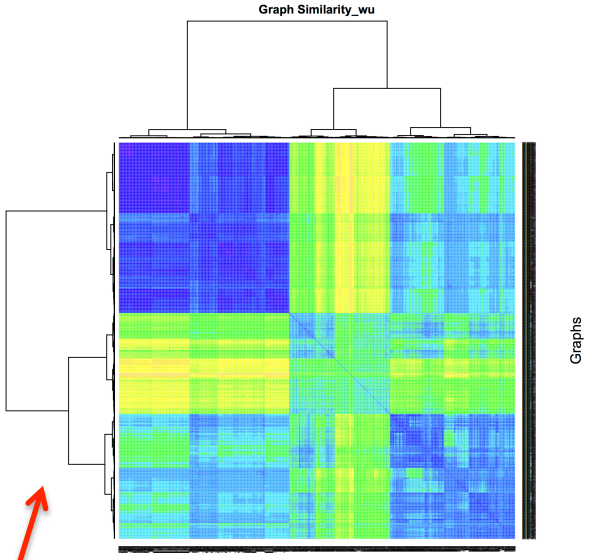
Detecting Changes

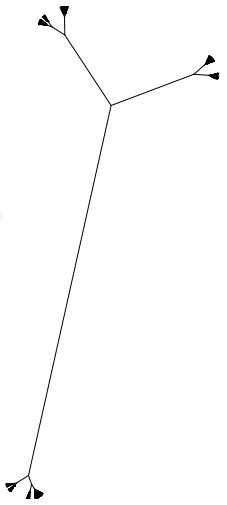
$time_1$ $time_2$ $time_n$

Example: detecting statistically significant changes in network traffic

Network Reaction to External Stimulus (static)







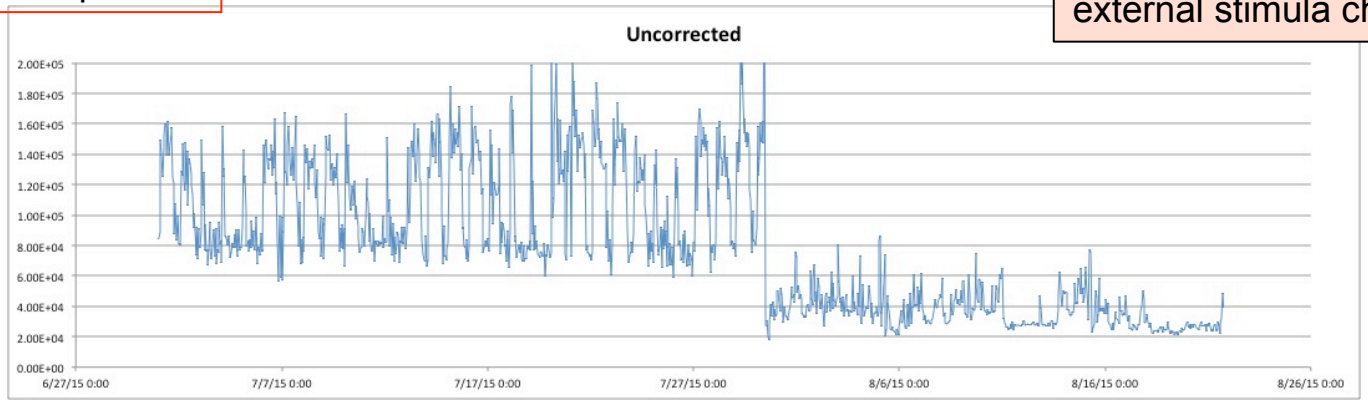
For a variety of static external stimulus, it is possible cluster or group the network reactions for specific stimulus

Network Reaction to External Stimulus (dynamic)

Sandia National Laboratories

It is also possible to monitor the temporal change in a network structure, possibly as external stimula change

Example A



Example B