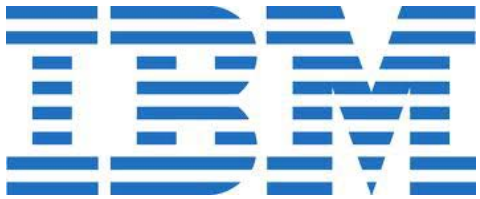
Outlier Detection for Temporal Data

Manish Gupta — UIUC

Jing Gao — SUNY

Charu Aggarwal — IBM

Jiawei Han — UIUC



SDM 2013

Austin, Texas

Tutorial Outline

Part 1

10 min	• Introduction
40 min	• Outlier Detection for Time Series Data
20 min	• Outlier Detection for Stream Data
20 min	• Outlier Detection for Stream Data in Distributed Scenarios
30 min	• Break

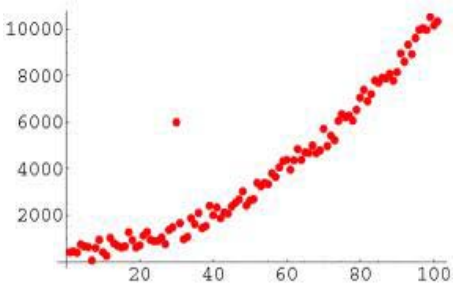
Part 2

25 min	• Outlier Detection for Spatio-temporal Data
30 min	• Outlier Detection for Temporal Networks
25 min	• Applications of Temporal Outlier Detection Techniques
10 min	• Summary, Q&A

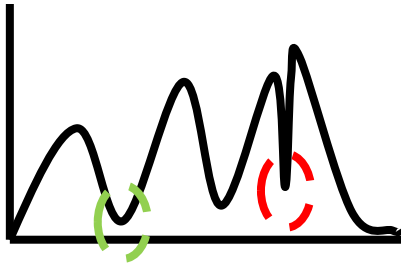
gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Outlier Detection

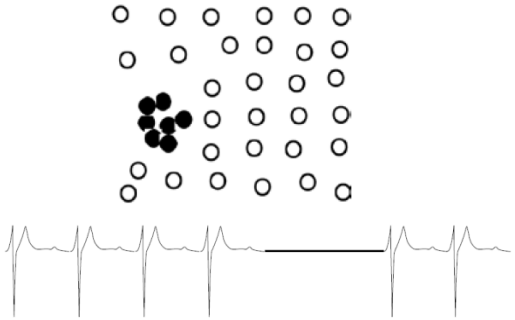
- Also called anomaly detection, event detection, novelty detection, deviant discovery, change point detection, fault detection, intrusion detection or misuse detection
- Three types



Point Outliers



Contextual Outliers



Collective Outliers

- Techniques: classification, clustering, nearest neighbor, density, statistical, information theory, spectral decomposition, visualization, depth, and signal processing
- Outlier packages:

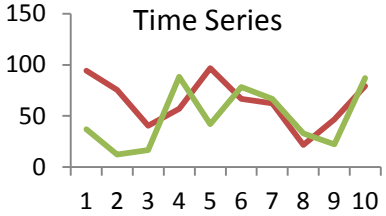

- Data types: high-dimensional data, uncertain data, stream data, network data, time series data

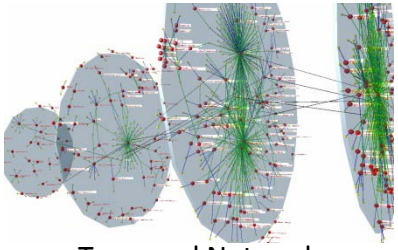
Time Series and other Temporal Data

- First work on outlier detection: [Fox, 1972]

Additive Outlier Innovative Outlier Temporary Change Level Shift

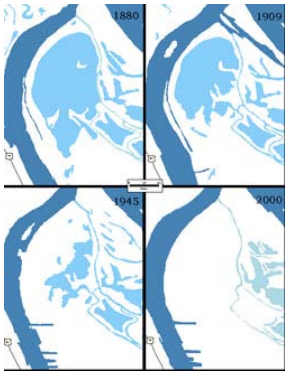
- ARIMA, ARMA, VAR, VARMA, CUSUM, etc.
- [Barnett and Lewis, 1978; Hawkins, 1980; Rousseeuw and Leroy, 1987]
- Temporal data: Social network data, astronomy data, sensor data, computer network traffic, commercial transactions, medical records, judicial records, ...
- New temporal data types





Temporal Networks

Spatio-temporal Data



Distributed Data Streams

Temporal Databases

EmpID	Name	Department	Valid Start Time	Valid End Time
10	Eric	CS	1985	2001
20	Bill	ECE	1990	2000
30	Sam	Biochem	1987	1999
40	Marina	Civil	1997	2004

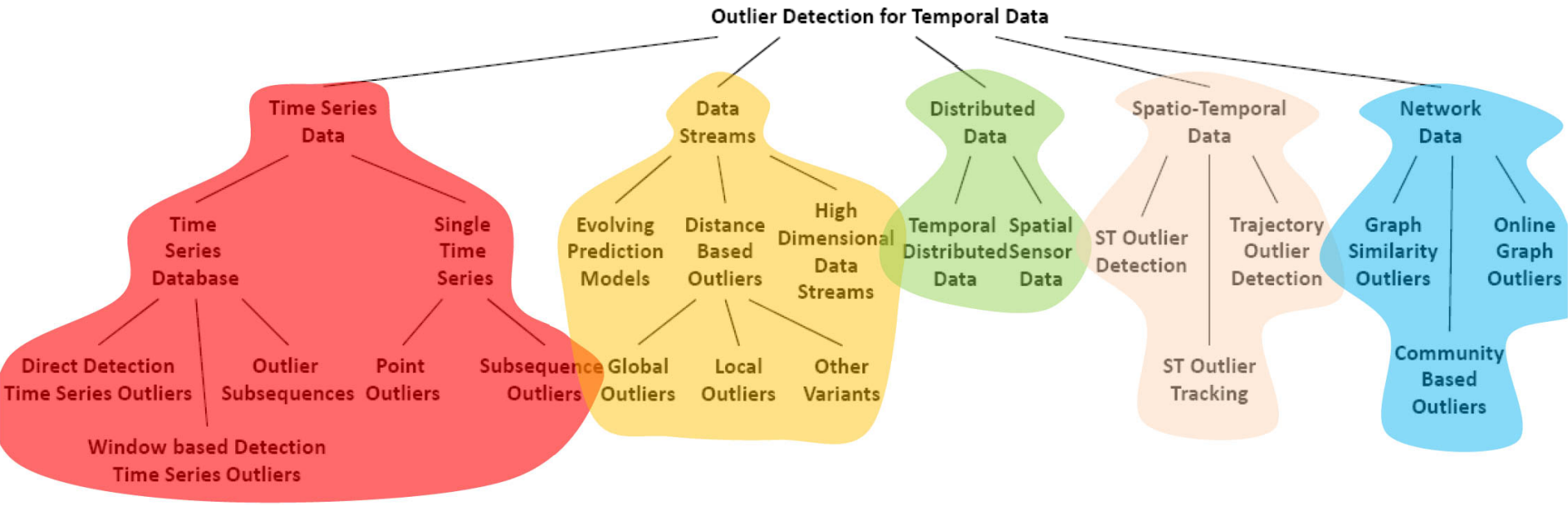
Data Streams

gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Challenges for Outlier Detection for Temporal Data

- Definition of outliers such that it captures
  - Properties of the data
  - Properties of the network
  - Space and time dimensions
- Massive scale
- Data trend drift detection and handling
- Time efficient, single scan
- Distributed data streams
  - Minimize communication overhead
  - Minimize computational load

Tutorial Organization



40 min | **20 min** | **20 min** | **25 min** | **30 min**

Outlier Detection for Time Series Data

- Outliers in Time Series Databases
  - Direct Detection of Outlier Time Series
    - Unsupervised Discriminative Approaches
    - Unsupervised Parametric Approaches
    - Supervised Approaches
  - Window-based Detection of Outlier Time Series
  - Outlier Subsequences in a Test Time Series
- Outliers Within a Given Time Series
  - Points as Outliers
  - Subsequences as Outliers

gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Time Series vs. Discrete Sequences

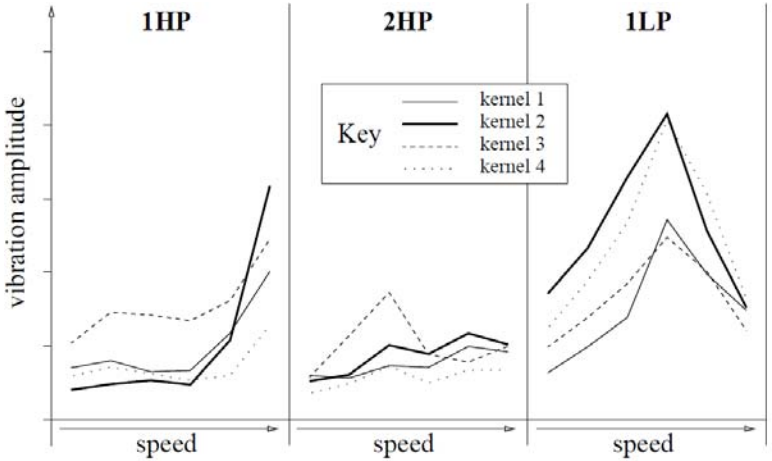
- Time series data
  - Numeric data across time
  - Studied more in the statistics community
  - Techniques include AR, VAR, ARIMA models, etc.
  - Outliers are mainly detected based on forecasts
- Discrete Sequences
  - Labels across time
  - Studied more in the data mining community
  - Techniques include Markov models, etc.
  - Outliers are defined in various ways like distance outliers, etc.
- Some techniques like clustering are shared for both types of data
- Some techniques discretize time series to sequences

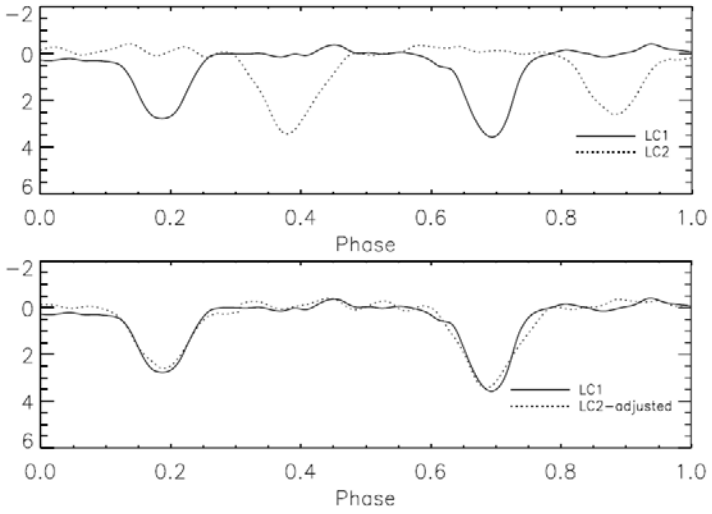
Unsupervised Discriminative Approaches

- Problem: Given a time series database, find anomalous time series sequences
- Compute the clustering space (features)
- Define a similarity function to compare two time series
- Cluster the sequences using some clustering algo
- Anomaly score = distance to closest centroid
- Most popular similarity measure: length of the longest common subsequence
  - t_1=XMJYAUZ and t_2=MZJAWXU, LCS=MJAU, nLCS=4

Unsupervised Discriminative Approaches

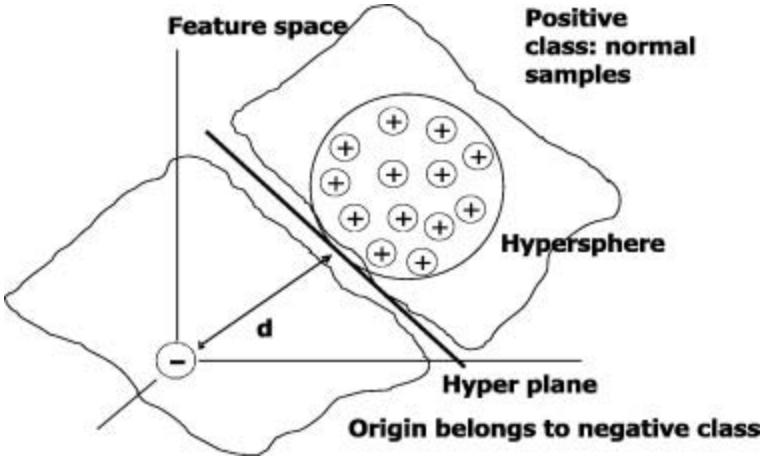
- nLCS with kmedoids [Budalakoti et al., 2006; Budalakoti et al., 2009]
- nLCS with a variant of kNN [Chandola et al., 2008]
  - Anomaly score = Distance to the k^{th} closest sequence
- Windowed subsequences with match count based sequence similarity with a threshold [Lane et al., 1997]
- Windowed subsequences with K-Means [Nairac et al., 1999]
- Duration, protocol type, #bytes for TCP connection with Euclidean distance and single-linkage clustering [Portnoy et al., 2001]
- Phased K-Means [Rebbapragada et al., 2009]

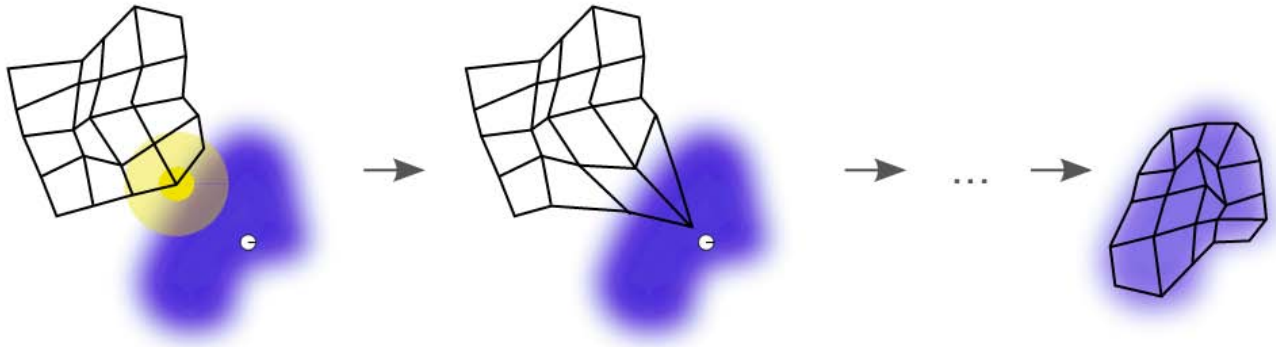




Unsupervised Discriminative Approaches

- 1-class SVM with discretized data [Szymanski and Zhang, 2004]
- Kernel-based feature maps of raw data with kNN or 1-class SVM [Eskin et al., 2002]
- 1-class SVM [Evangelista et al., 2005; Ma and Perkins, 2003b]
- Self organizing maps with windowed subsequences [González and Dasgupta, 2003]





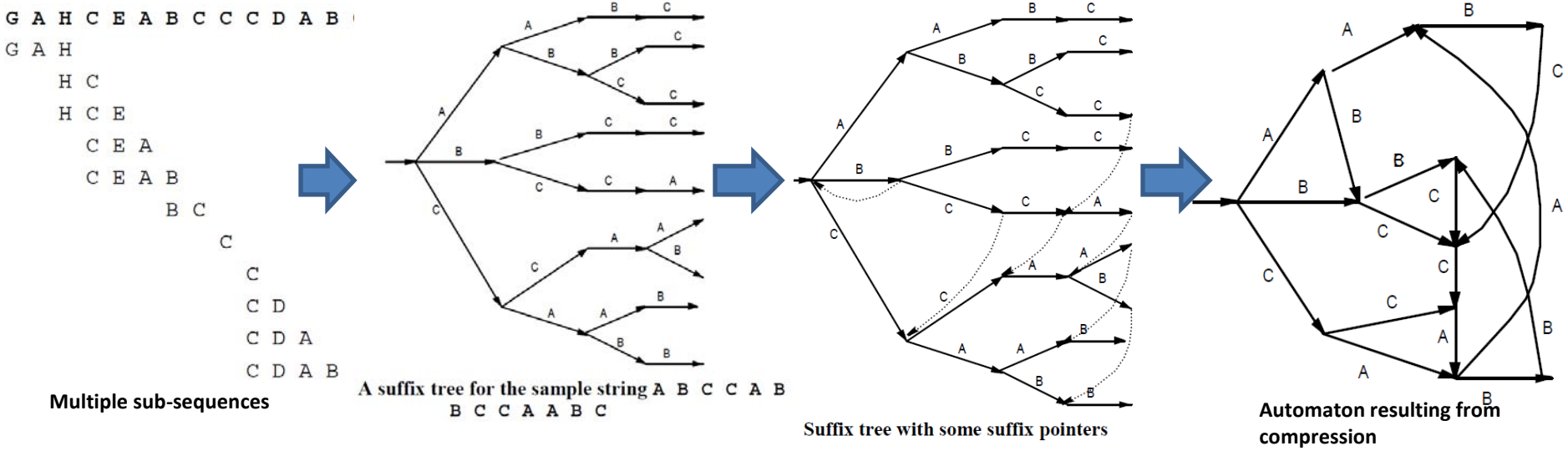
gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Unsupervised Parametric Approaches

- Value at any time point in the series depends on the values at previous few time points
- Anomaly score for time series is a function of the anomaly score of its values at each time point
- Markov models
  - Fixed history size
  - Variable history size
  - Selective history (Sparse Markovian)
- Hidden Markov models

Unsupervised Parametric Approaches

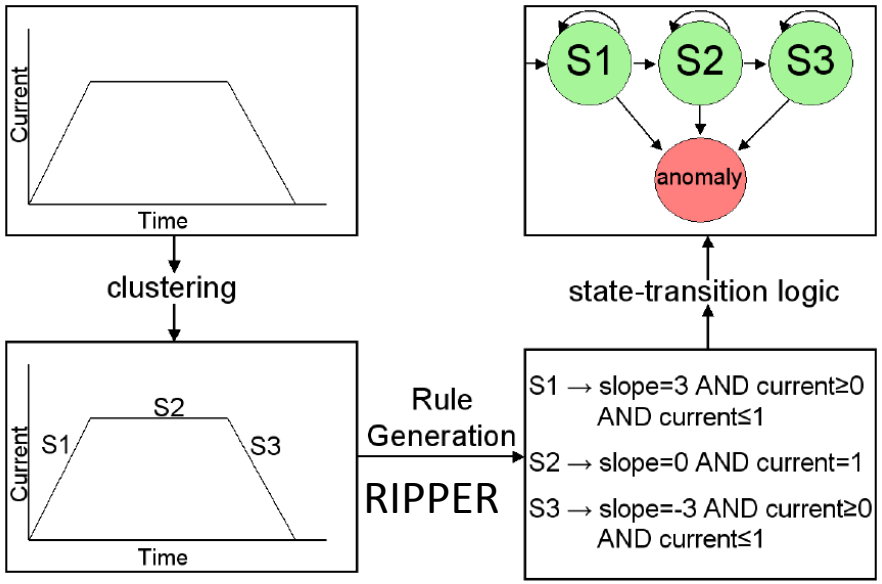
- Markov model with k=1 [Ye, 2000]
- Finite state automaton (FSA) [Marceau, 2000]
  - N-grams (length-N system call sequences)



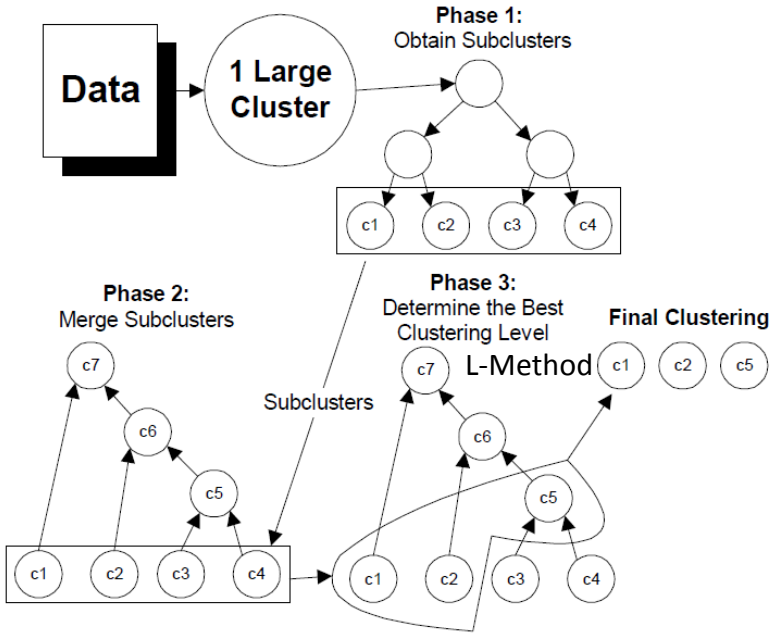
- FSA is also used by [Salvador and Chan, 2005; Chandola et al., 2008; Michael and Ghosh, 2000]

Unsupervised Parametric Approaches

- Finite state automaton (FSA) [Salvador and Chan, 2005]



Main steps in anomaly detection



Clustering methodology

- Input matches current state's characteristics ⇒ remain in current state
- Input matches the next state's characteristics ⇒ transition to the next state
- Input matches neither the current state's nor the next state's characteristics ⇒ transition to an anomaly state

Supervised Approaches

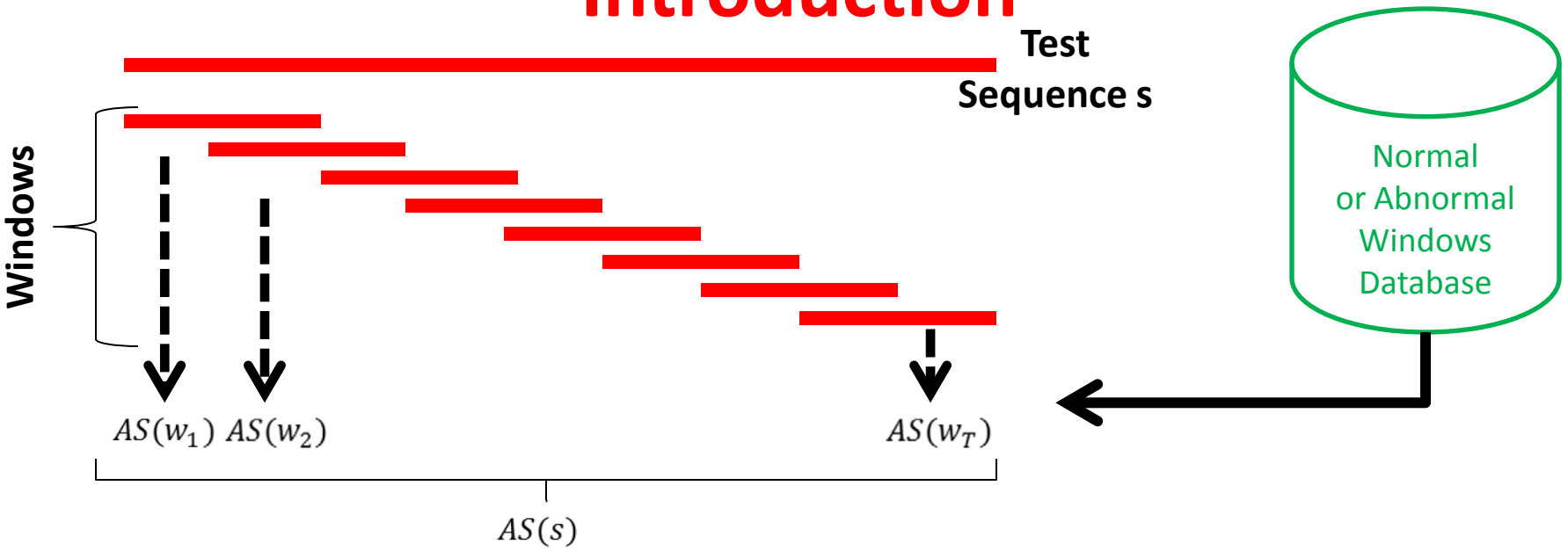
- Positional system calls features with the RIPPER classifier [Lee and Stolfo, 1998]
- Subsequences of positive and negative strings of behavior as features with
  - String matching classifier [Cabrera et al., 2001; González and Dasgupta, 2003]
  - Neural networks [Dasgupta and Nino, 2000; Endler, 1998; Gosh et al., 1998; Ghosh et al., 1999a; Ghosh and Schwartzbard, 1999]
  - Elman network [Ghosh et al., 1999a]
- Motion features with SVMs [Li et al., 2006]
- Bag of system calls features with decision tree, Naive Bayes, SVMs [Kang et al., 2005]
- Sliding window subsequences with
  - SVMs [Tian et al., 2007; Wang et al., 2006],
  - Rule based classifiers (Classification using Hierarchical Prediction Rules (CHIP)) [Li et al., 2007]
  - HMMs [Gao et al., 2002]

Outlier Detection for Time Series Data

- Outliers in Time Series Databases
  - Direct Detection of Outlier Time Series
  - Window-based Detection of Outlier Time Series
    - Normal Pattern Database Approaches
    - Negative/Mixed Pattern Database Approaches
  - Outlier Subsequences in a Test Time Series
- Outliers Within a Given Time Series
  - Points as Outliers
  - Subsequences as Outliers

gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Introduction



- Advantage: Better localization of anomalies compared to techniques that compute time series outlier score directly
- Disadvantage: New parameter -- window length parameter
- Windows: Also called fingerprints, pattern fragments, detectors, sliding windows, motifs, n-grams

Normal Pattern Database Approaches

- Sequence TIme-Delay Embedding (STIDE) [Hofmeyr et al., 1998]
  - Normal sequences are divided into size k overlapping windows
  - For a test sequence, again size k subsequences are obtained and those subsequences that do not occur in normal database are considered as mismatches
  - If a test sequence has large number of mismatches, it is marked as an anomaly
  - If a subsequence is not in the database, the mismatch score is computed as the minimum Hamming distance between the window and any of the subsequences in normal database normalized by k
- Also used in [Cabrera et al., 2001; Endler, 1998; Gao et al., 2002; Ghosh et al., 1999a; Ghosh et al., 1999b]

Train sequence

(open, read, mmap, mmap, open, read, mmap)

Normal Database

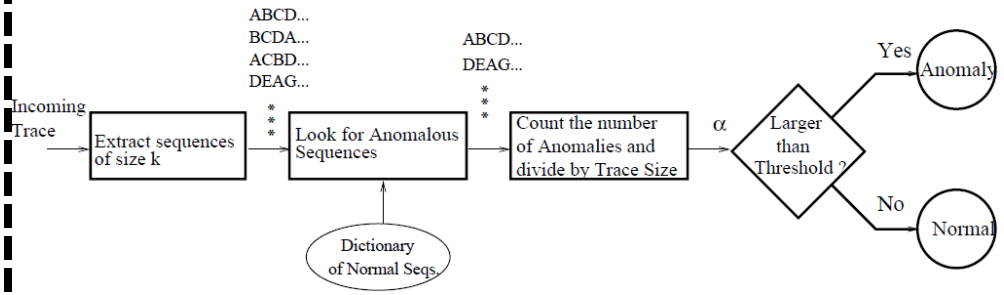
1. open, read, mmap 2. read, mmap, mmap
3. mmap, mmap, open 4. mmap, open, read

Test sequence

(open, read, mmap, mmap, open, *mmap*, mmap)

2 mismatches

1. mmap, open, *mmap* *2.* open, *mmap*, mmap



Normal Pattern Database Approaches

- TIme-Delay Embedding (TIDE) [Forrest et al., 1996]
  - For every element in every normal sequence, the elements occurring at distance 1,2,. . ., k in the sequence are noted
  - A normal database of such occurrences is created
  - Given a new test sequence, again a lookahead of the same size k is used
  - Each pair of element occurrence is checked with the normal database and the number of mismatches is computed
  - Anomaly score of the test sequence is the number of mismatches normalized by the total number of such occurrence pairs

Train: open, read, mmap, mmap, open, getrlimit, mmap, close

call	position 1	position 2	position 3
open	read, getrlimit	mmap	mmap, close
read	mmap	mmap	open
mmap	mmap, open, close	open, getrlimit	getrlimit, mmap
getrlimit	mmap	close	
close			

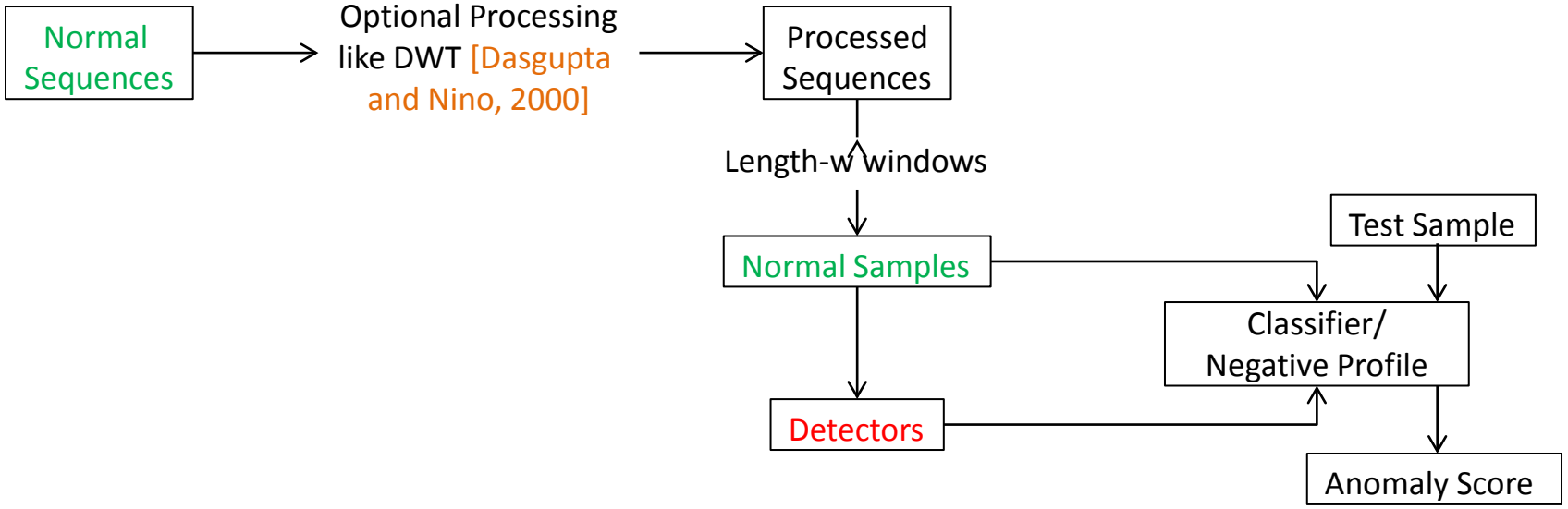
Test: open, read, mmap, open, open, getrlimit, mmap, close

4 mismatches

- open is not followed by open at position 3
- read is not followed by open at position 2
- open is not followed by open at position 1
- open is not followed by getrlimit at position 2

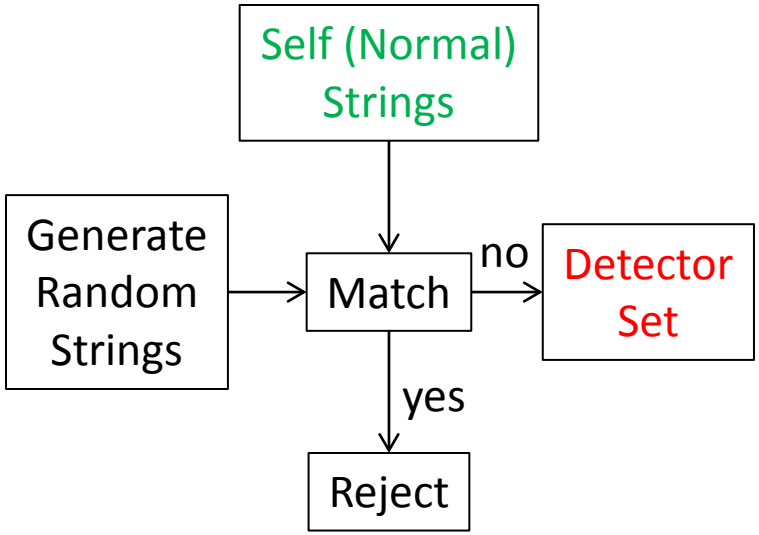
Negative and Mixed Pattern DB Approaches

- Anomaly dictionaries are used in [Dasgupta and Nino, 2000; Dasgupta and Majumdar, 2002; D'haeseleer et al., 1996; Forrest et al., 1994; González and Dasgupta, 2003]



Negative and Mixed Pattern DB Approaches

- Finding detectors or negative subsequences
  - Not in the normal set
  - Differ in at least r contiguous positions [D'haeseleer et al., 1996; Dasgupta and Majumdar, 2002]
    - Naïve [Forrest et al., 1994]
    - Dynamic programming algorithm
    - Greedy algorithm
  - Differ in at least r contiguous chunks [Dasgupta and Nino, 2000]
  - Real-valued space: Detector is a hypersphere in n-dimension space with radius <1 [González and Dasgupta, 2003]



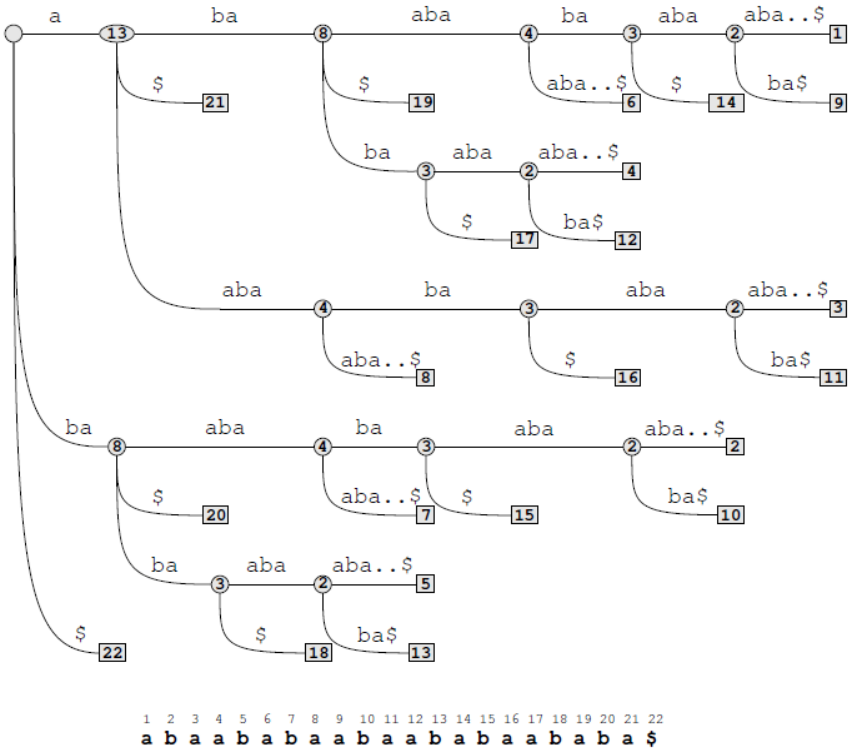
- Detectors can be generated randomly or using some domain knowledge of situations that are not expected to occur in normal sequences
  - Should be far from normal
  - Should maximize the covering of non-self space [González and Dasgupta, 2003]

Outlier Detection for Time Series Data

- Outliers in Time Series Databases
  - Direct Detection of Outlier Time Series
  - Window-based Detection of Outlier Time Series
  - Outlier Subsequences in a Test Time Series
- Outliers Within a Given Time Series
  - Points as Outliers
  - Subsequences as Outliers

Outlier Subsequences in a Test Time Series

- Given a test time series, a subsequence p is reported as an outlier if its frequency in the test time series t is substantially different from its expected frequency estimated using a time series database D
- TARZAN algorithm exploits suffix trees [Keogh et al., 2002; Lin et al., 2003; Lin et al., 2007]
  - Discretize time series
  - Build the suffix tree for the reference string and test string
  - For each substring in test string, calculate its frequency of occurrence in both reference string and test string
  - Adjust the frequency in reference string to get estimated frequency in test string
  - If substring w in test string does not occur in reference string, use suffix tree for reference string to look for the longest set of strings from reference string that cover w and then use Markov method to estimate its freq in test string
  - If the difference is greater than threshold, mark the substring as anomaly



gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Outlier Subsequences in a Test Time Series

- [Keogh et al., 2002] define a soft match version of the problem where the frequency of pattern p in the database D is defined using the largest number l such that every subsequence of p of length l occurs at least once in D
- Another form of soft match is defined in [Atallah et al., 2004] where rather than a match of pattern p, any permutation of p is also considered to be a match
- [Gwadera et al., 2005b; Gwadera et al., 2005a] have proposed Interpolated Markov Models (IMM) to efficiently compute the match score of a pattern or its permutations with any time series

Outlier Detection for Time Series Data

- Outliers in Time Series Databases
  - Direct Detection of Outlier Time Series
  - Window-based Detection of Outlier Time Series
  - Outlier Subsequences in a Test Time Series
- Outliers Within a Given Time Series
  - Points as Outliers
    - Prediction Models
    - Profile Similarity-based Approaches
    - Deviants
  - Subsequences as Outliers

gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Prediction Models

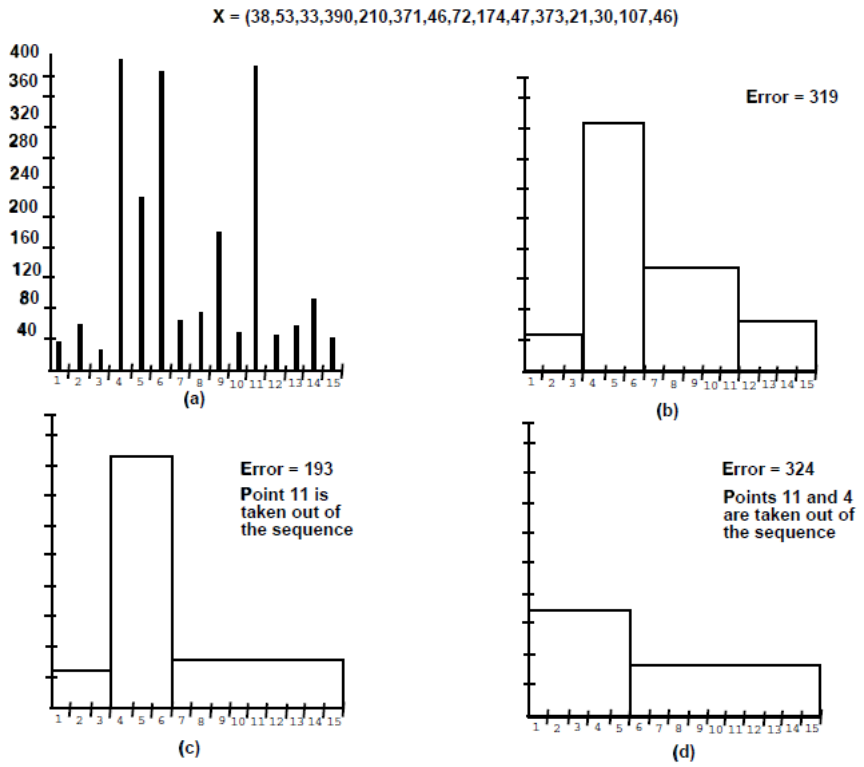
- Given a time series, one can predict the value at time t as
  - A median of the values in the size-2k window from $t-k$ to $t+k$ [Basu and Meckesheimer, 2007]
  - An average of all the points in the cluster that the value at time t maps to [Hill and Minsker, 2010]
  - Using Single-layer linear network predictor (or AR model) [Hill and Minsker, 2010]
  - Using MultiLayer Perceptron (MLP) predictor [Hill and Minsker, 2010]
  - Using Support Vector Regression [Ma and Perkins, 2003a]

Profile Similarity-based Approaches

- These approaches maintain a normal profile and then compare a new time point against this profile to decide whether it is an outlier
- For multiple OS performance metric time series, the Tiresias system [Williams et al., 2007] maintains a normal profile and also a variance vector
  - Any new data point is compared both with the normal profile and the variance vector to compute its anomaly score
  - Profile is the actual smoothed time series data from past datasets
- In [Silvestri et al., 1994], a neural network is used to maintain the normal profile and an estimation is made for the next value in the sensor stream based on this profile

Deviant Detection

- The problem is to find points in a given time series whose removal from the series results in a histogram representation with a lower error bound than the original, even after the number of buckets has been reduced to account for the separate storage of these deviant points [Jagadish et al., 1999]
- They propose a dynamic programming mechanism to solve the problem
- [Muthukrishnan et al., 2004]
  - For any bucket, the optimal set of k deviants within the bin always consists of the l highest and remaining k-l lowest values for some l <= k
  - Propose an approximation to the dynamic programming based solution that maintains a partial solution only for a few interspersed indexes of the time series rather than for each value



390 and 373 are deviants

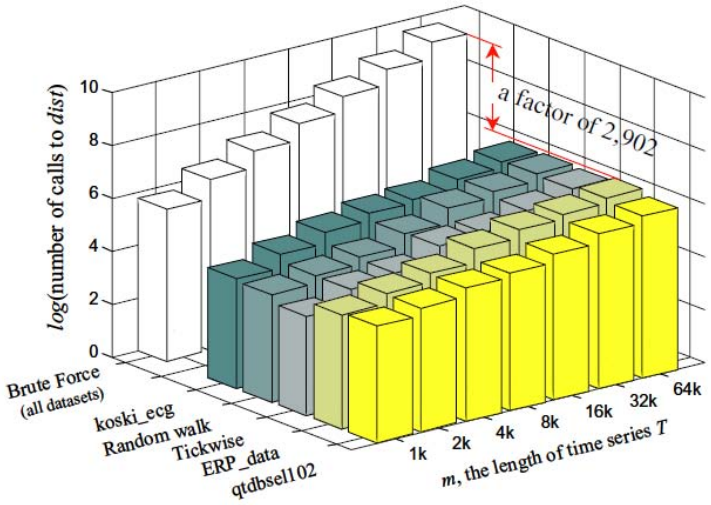
Outlier Detection for Time Series Data

- Outliers in Time Series Databases
  - Direct Detection of Outlier Time Series
  - Window-based Detection of Outlier Time Series
  - Outlier Subsequences in a Test Time Series
- Outliers Within a Given Time Series
  - Points as Outliers
  - Subsequences as Outliers
    - Discords
    - Multi-Scale Anomaly Detection

gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

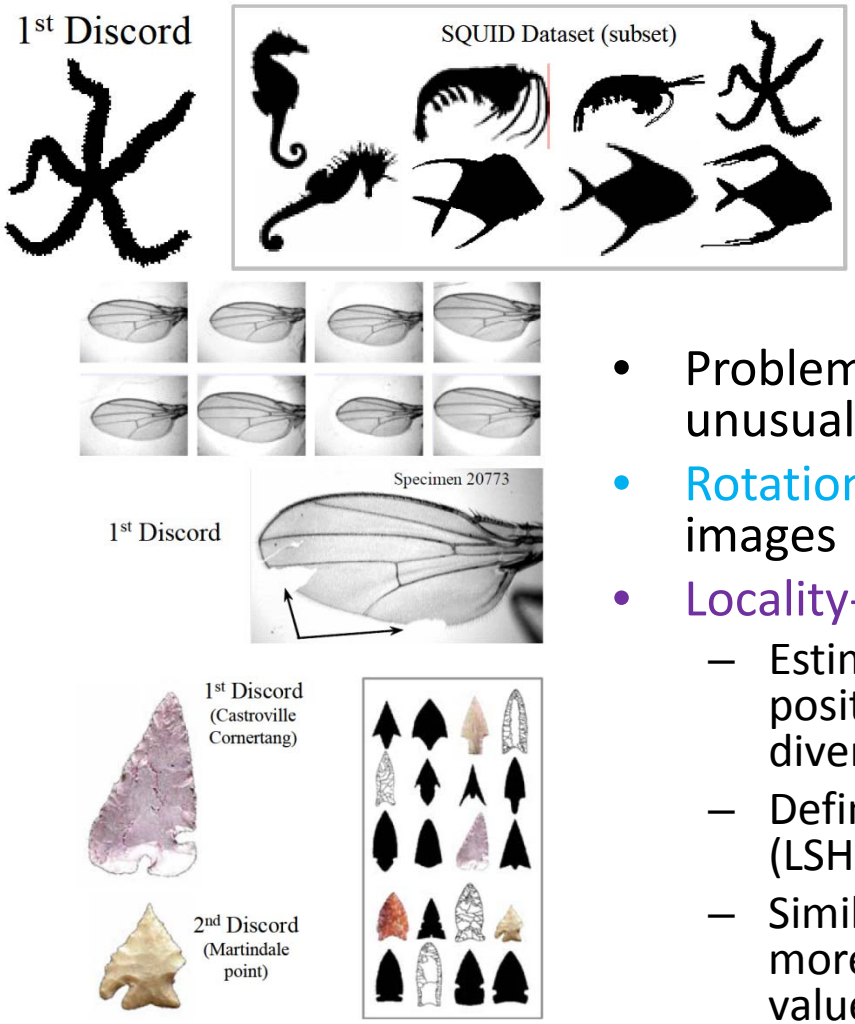
Discord Discovery: Outlier Subsequences

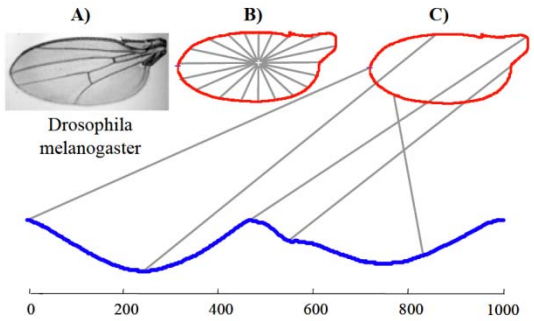
- Given a time series T, the subsequence D of length n beginning at position l is said to be the discord (or outlier) of T if D has the largest distance to its nearest non-overlapping match [Keogh et al., 2006]
- The brute force solution is to consider all possible subsequences $s \in S$ of length n in T and compute the distance of each such s with each other non-overlapping $s' \in S$
- Top-K pruning can be used to make this computation efficient
- Subsequences comparisons can be smartly ordered for effective pruning using various heuristics like
  - Heuristic reordering of candidate subsequences [Keogh et al., 2005] using Symbolic Aggregate ApproXimation (SAX)
  - Locality sensitive hashing [Wei et al., 2006]



The number of calls to the Euclidean distance function required by brute force and heuristic search over a range of data sizes for 5 representative datasets

Shape Discords [Wei et al., 2006]



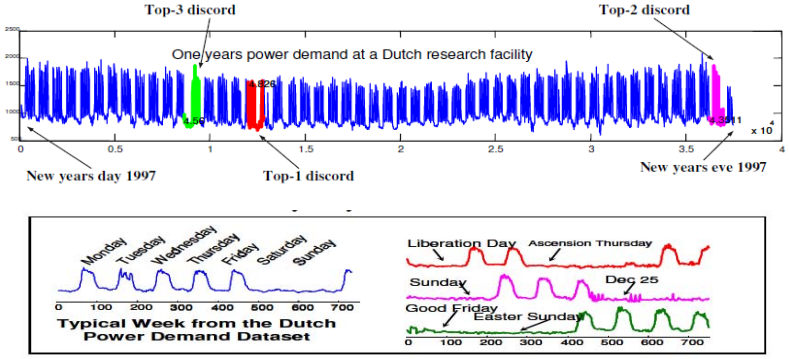


- Problem of finding shape discords, the most unusual shapes in a collection
- Rotation invariant Euclidean Distance between images
- Locality-Sensitive Hashing
  - Estimation of similarity based on sparse sampling of positions from feature vectors has been used in diverse areas for different purposes
  - Define a rotation invariant locality-sensitive hash (LSH) function
  - Similar shapes (even with different orientations) are more likely to be mapped together to same LSH value

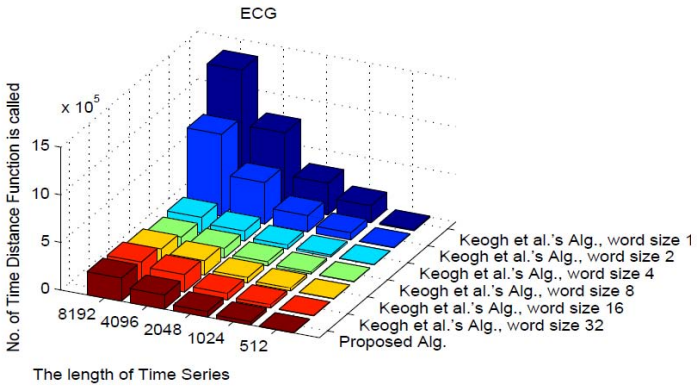
gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Discords: Haar Wavelet & Augmented Tries

- [Bu et al., 2007; Fu et al., 2006]
- A Haar wavelet and augmented trie based algorithm to mine the top-K discords from a time series database
- Haar wavelet tranform
  - provides better pruning power
  - can dynamically determine the word size (unlike SAX)
- Steps
  - Time series
  - Haar wavelet transform
  - Normalize Haar wavelet coefficients
  - Identify cutpoints by treating this distribution as Gaussian such that area between any 2 cutpoints is same
  - Map Haar coefficients to symbols based on cutpoints
  - Construct a trie from these words
  - Use the trie to perform appropriate ordering of subsequences to use as candidates and compare

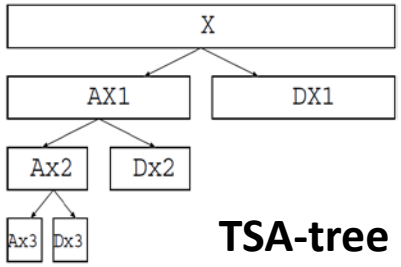


Top-3 discords in power consumption history of a Dutch research facility in year 1997

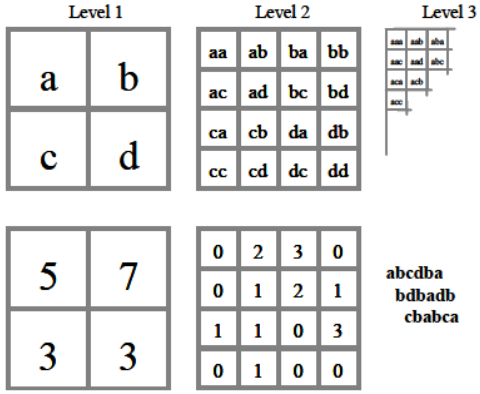


Multi-scale Anomaly Detection

- [Chen and Zhan, 2008] define subsequence outlier detection problem for an unequal interval time series
  - $X = \langle v_1 = (x_1, t_1), v_2 = (x_2, t_2), \ldots, v_n = (x_n, t_n) \rangle$
  - X is unequal when $t_{i+1} - t_i$ is not the same for all i
  - A pattern is defined as a subsequence of two consecutive points $p = \langle v_t, v_{t+1} \rangle$
  - Pattern p is anomalous if there are very few other patterns with
    - the same slope $\frac{x_{i+1} - x_i}{t_{i+1} - t_i}$ and
    - the same length $\sqrt{(x_{i+1} - x_i)^2 + (t_{i+1} - t_i)^2}$
  - To identify anomalies at multiple scales, Haar transform is used
- [Shahabi et al., 2000] propose Trend and Surprise Abstractions (TSA) tree to store trends and surprises in terms of Haar wavelet coefficients
  - Find the cities where temperature has sudden changes (daily) during the last month
  - Find the cities where temperature has sudden changes (monthly) during the last decade
- [Wei et al., 2005] define a subsequence as anomalous if its similarity with a fixed previous part of the sequence is low
  - Similarity is measured using Chaos bitmaps
  - Symbols are obtained using SAX



TSA-tree



Chaos Bitmaps

Outlier Detection for Stream Data

- Evolving Prediction Models
  - Online Sequential Discounting Algorithms
  - Dynamic Bayesian Networks
- Distance Based Outliers
  - Global Outliers
  - Local Outliers
  - Other Variants

SmartSifter - Sequential Discounting for Categorical Variables

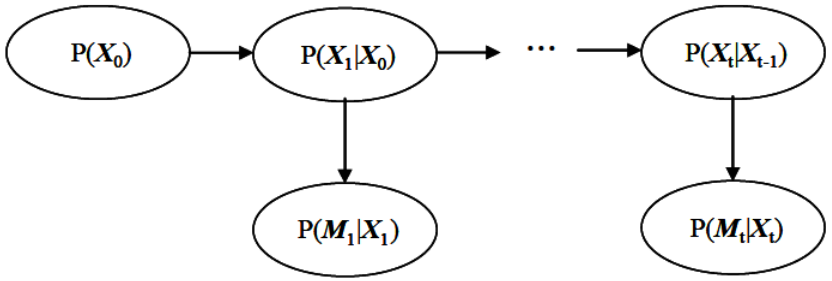
- [Yamanishi and Takeuchi, 2002; Yamanishi et al., 2004]
- Online Sequential Discounting Algorithms
  - Incrementally learn the probabilistic model of the data source every time a data item is input
  - The model forgets the effect of past data gradually
  - Outlier score is high for a data point if its score based on the learned model is high
- For categorical variables
  - SDLE (Sequential Discounting Laplace Estimation)
    - Cells are created by partitioning the space of all values of all categorical variables
    - The probability p(x) of a cell is the number of values in that cell divided by total data points with the Laplace smoothing
    - When a new data point arrives, the count for all the cells are adjusted with temporal discounting and appropriate Laplace smoothing is applied

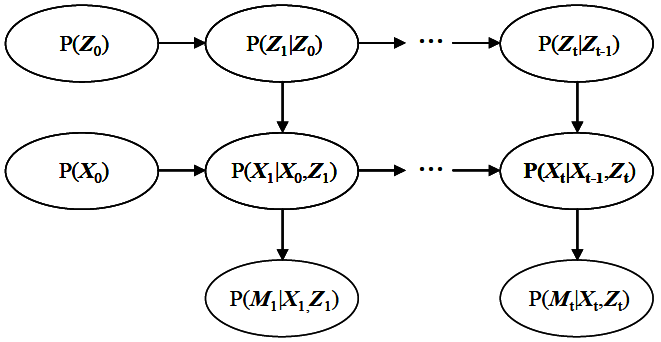
SmartSifter - Sequentially Discounting Model for Continuous Variables

- Independent Model
  - Parametric case
    - Gaussian mixture model
    - Sequentially Discounting EM (SDEM) algorithm
      - Incremental EM with discounting of effect of past examples
      - Iteratively learns coefficients of the mixture and the mean and variance of Gaussians
      - [Guo et al., 2006] propose a variant for computer network data
  - Non-parametric case
    - Kernel mixture model
    - Sequentially Discounting Prototype Updating (SDPU) algorithm
      - Coefficients of the mixture and the variance matrix are fixed
      - Iteratively learns only the means of the kernels or the prototypes
- Time Series Model
  - Sequentially Discounting AR (SDAR) algorithm
    - Learns the AR model parameters iteratively with time discounting
    - [Jiang et al., 2006] propose an ARX variant for computer network data

Dynamic Bayesian Networks

- The model can be changed by itself to incorporate drifts in the data stream
- [Hill et al., 2007] present an approach that uses Dynamic Bayesian networks which are Bayesian networks with network topology that evolves over time, adding new state variables to represent the system state at the current time t





- Kalman filtering used for inference
- Posterior distribution of the observed state variables is then used to construct a Bayesian credible interval (BCI)
- Measurements outside of the p% BCI are anomalies

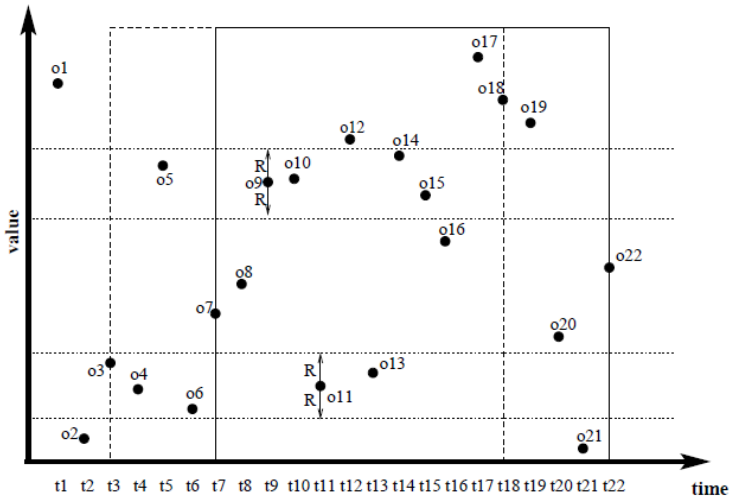
- Status (e.g. normal/anomalous) of each measurement is also modeled as a hidden state
- Maximum a posteriori measurement status (MAP-ms) of hidden state variable is used to classify the measurement as normal or anomalous

Outlier Detection for Stream Data

- Evolving Prediction Models
  - Online Sequential Discounting Algorithms
  - Dynamic Bayesian Networks
- Distance Based Outliers
  - Global Outliers
  - Local Outliers
  - Other Variants

Distance based Outliers in Data Streams

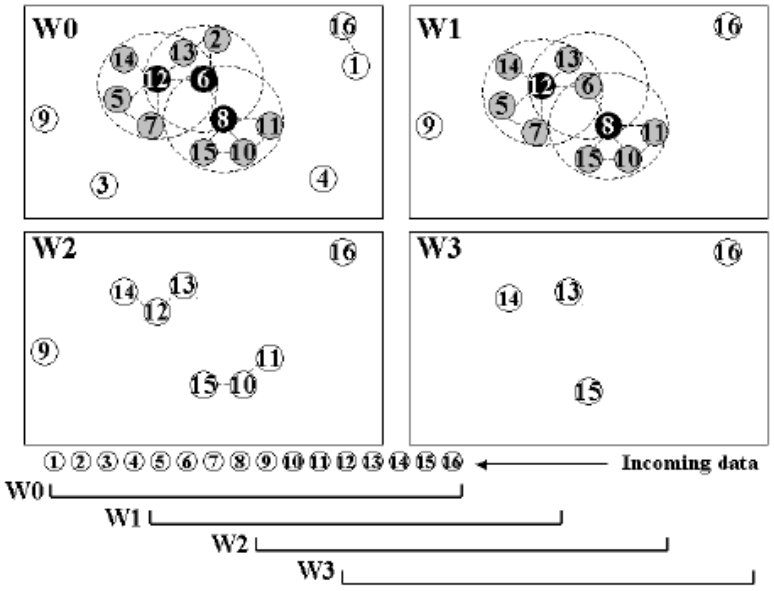
- Given a data stream, the problem is to find distance based outliers at any time window [Angiulli and Fassetti, 2007; Yang et al., 2009]
- As stream evolves
  - Old objects expire, new come in
  - #preceding neighbors of any object decreases
  - Object could become an outlier



- Since any object will expire before its succeeding neighbors, inliers having at least k succeeding neighbors will be always safe inliers
- [Angiulli and Fassetti, 2007] propose an exact algorithm to efficiently compute such outliers using a new data structure called Indexed Stream Buffer (ISB) which supports a range query search
- Exploit two heuristics
  - It is sufficient to retain in ISB only a fraction of safe inliers
  - Rather than storing the list of k most recent preceding neighbors, it is enough to store only the fraction of preceding neighbors which are safe inliers to the total number of safe inliers

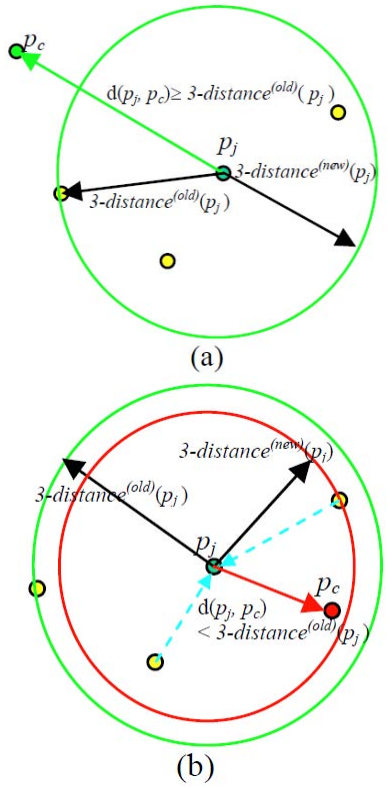
Exploiting Predictability of Object Expiration

- [Yang et al., 2009] propose that maintaining all neighbor relationships across time may be very expensive
- So abstracted neighbor relationships can be maintained
- But maintaining such cluster abstractions is expensive too
- So, they exploit the "predictability" of the expiration of existing objects
- Based on expiration of objects, they create "predicted views" of each future window
- Abstract-C algorithm makes use of the predicted views to compute distance based outliers



Incremental Local Outlier Detection for Streams

- Static Local Outlier Factor (LOF) [Breunig et al., 2000] can be applied to the incremental LOF problem as
  - Periodic LOF: Apply LOF on entire data set periodically
  - Supervised LOF: Compute the k-distances, local reachability density (LRD) and LOF using training data and use them to find outliers in test data
  - Iterated LOF: Re-apply static LOF every time a new data record is inserted into the data set
- [Pokrajac et al., 2007] propose an incremental LOF algorithm
- When a new data point arrives
  - Insertion of new record
    - Compute reachability distance, LRD and LOF of new point
  - Maintenance
    - Update k-distances, reachability distance, LRD and LOF for affected existing points (which is independent of #objects)
- Maintained statistics are used to instantly determine whether inserted data record is outlier



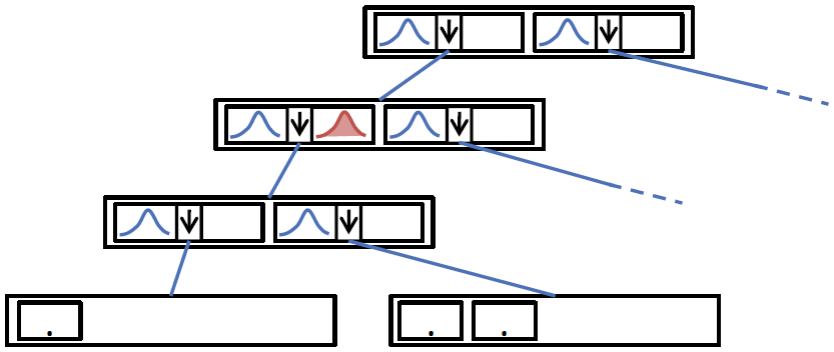
Update of kNN distance on insertion of new record (k=3) (a) 3NN does not change (b) 3NN changes

Relative Neighborhood Dissimilarity

- Let D_p= average distance of a point p from its kNN
- Let μ_p and σ_p = the mean and the standard deviation of D_p, for all neighbors p' of p
- Relative outlier score ROS(p)$= 1 - \frac{D_p}{\mu_p}$
- p is outlier if $|\text{ROS(p)}| > \frac{3\sigma_p}{\mu_p}$ [Vu et al., 2008]
- Relative Neighborhood Dissimilarity (ReND) algorithm
  - Maintains a set of n clusters each of which stores
    - Time ordered list of L data points with their kNN and distances to each of the kNNs
    - Support and centroid of cluster
  - Use this maintained cluster information to find outliers

AnyOut

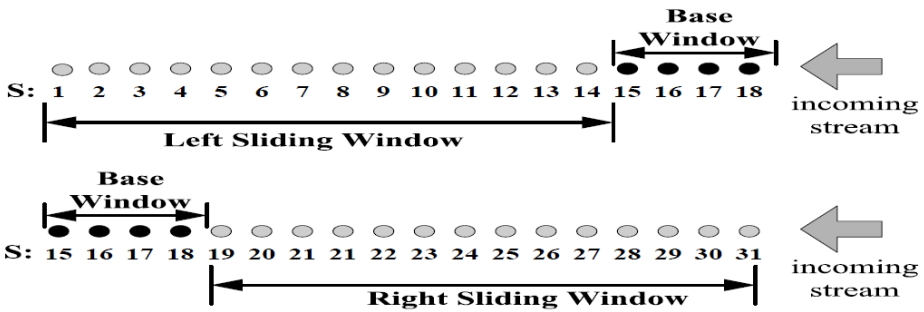
- [Assent et al., 2012] present AnyOut algorithm based on a hierarchical cluster tree
- The cluster tree is maintained in an incremental way
  - Each cluster stores the cluster feature: CF = (n, LS, SS)
- As long as time is available, the hierarchy is traversed to determine
  - Mean Outlier Score: Deviation between the object and a cluster center
  - Density Outlier Score: Gaussian probability density for the object
- Until interrupted, more fine grained resolution of the clustering structure is analyzed



ClusTree: Node entries represent clusters (descriptor, child pointer, buffer); child nodes are more fine grained clusters of the parent data

Discords in Data Streams

- Consider three windows: base window B, left window L and right window R of sizes w_b, w_l and w_r



- B is an anomaly if there are < k subsequences in L and R with distance from B < d
- [Bu et al., 2009] build local clusters on stream data and monitor anomalies via efficient pruning
- A local cluster is a subsequence containing a base window B such that all other base windows in the cluster are within a distance τ from B
- Besides clustering, they propose a piecewise metric index structure (Vantage Point trees) to reschedule the joining order of local clusters for faster computations

Outlier Detection for Stream Data in Distributed Scenarios

- Challenges and Introduction
- Outliers in Temporal Distributed Data
  - By Sharing Data
  - By Sharing Outliers
  - By Sharing Distributions
- Outliers in Spatial Sensor Data

Challenges

- Resource constraints: Energy, memory, computational capacity and communication bandwidth
- High communication cost: Communication cost is orders of magnitude more than computation costs
- Distributed streaming data: Processing data online coming at different rates from multiple distributed sensors
- Dynamic nature: Dynamic network topology, frequent communication failures, mobility and heterogeneity of nodes
- Large-scale deployment: Scalability issues with traditional outlier detection algorithms
- Identifying outlier sources: Make distinction between errors, events and malicious attacks

Introduction

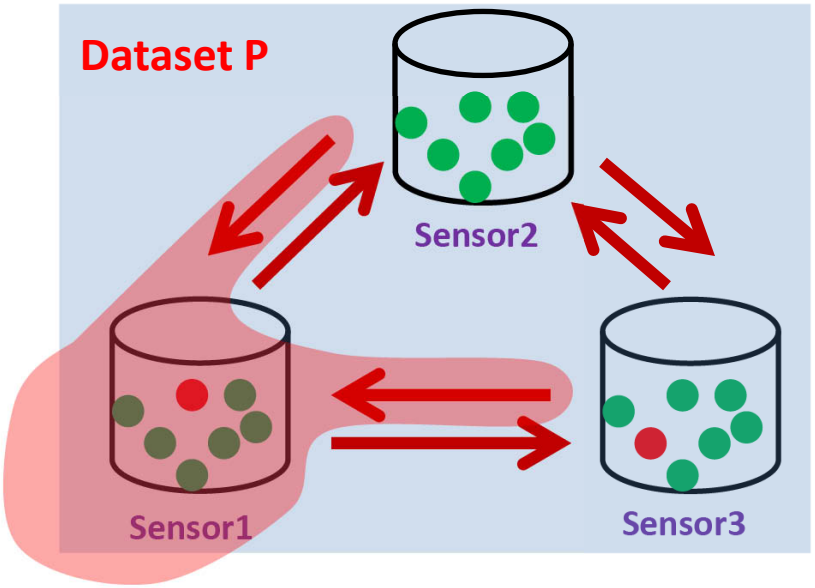
- In a distributed stream setting, points are distributed across various nodes (sensors)
- Each sensor has an associated stream of incoming points
- The aim is to find top few outliers based on the global data
  - With least amount of communication between nodes
- In the distributed spatio sensor setting, the position of sensors is also important when computing outliers

Outlier Detection for Stream Data in Distributed Scenarios

- Challenges and Introduction
- Outliers in Temporal Distributed Data
  - By Sharing Data
  - By Sharing Outliers
  - By Sharing Distributions
- Outliers in Spatial Sensor Data

Outlier Computation by Sharing Data

- [Branch et al., 2006] describe an algorithm for a distributed static setting
- R: outlier ranking function
  - Anti-monotonic wrt data and smooth, e.g., the distance to K^{th} NN, average distance to KNN, etc.
- Each sensor can compute top-K outliers (using R) for its knowledge
- Let support set of point x be smallest set of points to compute R(x)
- To compute global outliers between sensors p_i and p_j, messages must contain
  - Outliers computed locally
  - Support set of their own outliers and other's outliers



Knowledge of sensor1

Outlier Computation by Sharing Data

- In a streaming setting, when a new point is sampled, data changes at the local sensor itself
  - Change in knowledge of the sensor just like when it receives a new message
  - Recompute R on new knowledge and hence topK local outliers
- Sliding window scenario
  - Each node can retire old points regardless of where they were sampled and at no communication cost at all
  - Addition of sensors during system operation is possible
  - If sensors are removed then their contribution to the computation gets annulled when those points retire with time

Outlier Computation by Sharing Outliers

- Transmitting the entire data or even the model to every node could be expensive [Otey et al., 2006]
- Only the local outliers be exchanged between nodes
- If all nodes agree that a point is an outlier, then we can assume that the point is a global outlier
- The sites only communicate when some user-specified event occurs
- Events include
  - a user's query for the global outliers
  - when a node finishes processing a fixed number of points
  - when a node finds a fixed number of outliers

Outlier Computation by Sharing Distributions

- [Subramaniam et al., 2006] compute distance-based or density-based outliers in a hierarchical sensor network
  - Global distance based outliers

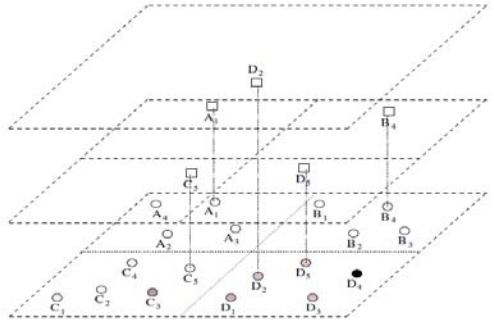
Parent
↑ Local Model + Local Outliers
Sensor

  - For density based method, Multi Granularity Deviation Factor (MDEF) of point p is a measure of how the neighborhood count of p compares with that of the values in its sampling neighborhood

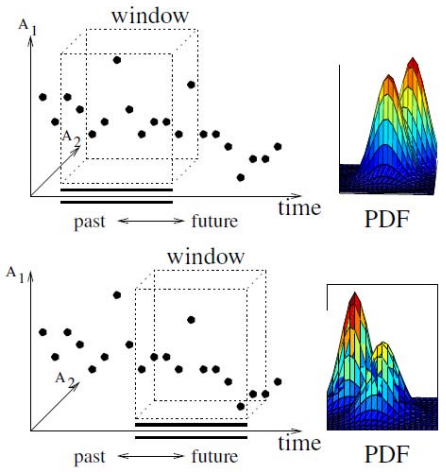
Parent
↑ New point with prob p
Sensor

Parent
↓ New Model
Sensor

- [Palpanas et al., 2003] also propose a hierarchical architecture



Hierarchical organization of a sensor network



Estimation of data distribution in the sliding window

Outlier Detection for Stream Data in Distributed Scenarios

- Challenges and Introduction
- Outliers in Temporal Distributed Data
  - By Sharing Data
  - By Sharing Outliers
  - By Sharing Distributions
- Outliers in Spatial Sensor Data

Distributed Spatio-temporal Outlier Detection

- [Jun et al., 2005] propose a spatio-temporal framework where outliers are modeled as an α-stable distribution
- Signal in sensor networks is modeled as
  - $u(x, y; t) = v(x, y; t) + n_1(x, y; t) + n_2(x, y; t)$
  - (x,y) is the Cartesian coordinates, t is time and v(x,y;t) is a source signal
  - $n_1(x, y; t)$ and $n_2(x, y; t)$ denote the additive white Gaussian noise and the symmetric alpha-stable noise
  - Signal source is assumed to be time-space separable, i.e., $v(x, y; t) = v_1(x, y) \times v_2(t)$, where $v_1(x, y)$ and $v_2(t)$ are spatial and temporal components
  - $v_1(x, y)$ is a propagation decay function
  - $v_2(t)$ is modeled as a repetitive signal

Distributed Spatio-temporal Outlier Detection

- Outlier detection process consists of four phases
  - Clustering of sensors
  - Temporal outlier detection with sensors in parallel
  - Removal of the spatial variation
    - Signal power gets weaker in proportion to the distance from the source
    - If this bias is not accounted properly, sensors near to the source, that receive more energy, might be falsely regarded as outliers, even though they are not
    - If the data has some effect, the distribution might be heavily tailed with outliers located near the tail of the distribution
    - The min and max values are removed as outliers iteratively using the property of α-stable distributions
  - Spatial outlier detection using the variogram method
    - Variogram method shows the spatial variance between sensors
    - Outliers are ones that deviate remarkably from other normal data

Break

Outlier Detection for Spatio-Temporal Data

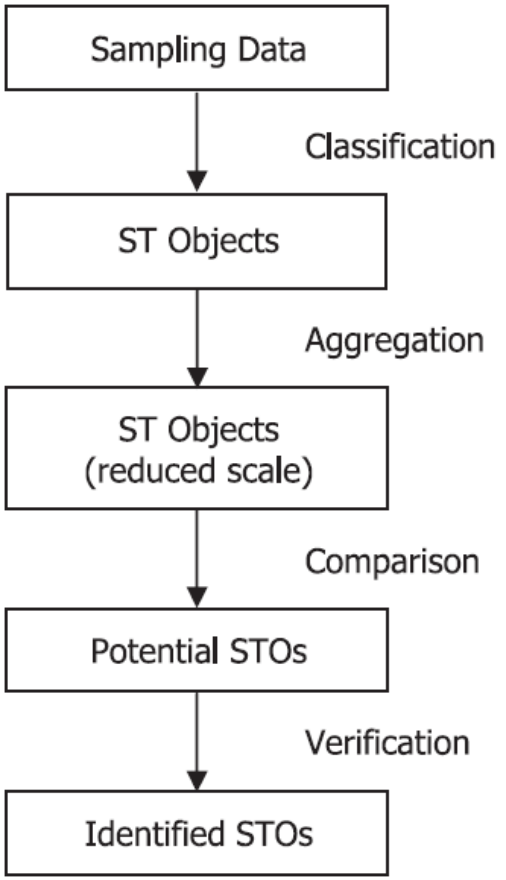
- ST-Outlier Detection
- ST-Outlier Tracking
- Trajectory Outlier Detection

DBSCAN based ST-Outlier

- A spatio-temporal outlier (ST-Outlier) is a spatio-temporal object whose thematic (non-spatial and non-temporal) attributes are significantly different from those of other objects in its spatial and temporal neighborhoods
- [Birant and Kut, 2006] propose a density based ST-Outlier detection mechanism with 3 steps
  - Cluster using a modified DBSCAN [Ester et al., 1996]
    - To support the temporal aspect, a tree is traversed to find both spatial and temporal neighbors of any object within a given radius
    - To find outliers when clusters have different densities, the algorithm assigns a density factor to each cluster and compares the average value of a cluster with the new coming value
  - Check the spatial neighbors to verify whether these potential outliers are spatial outliers too
  - Check the temporal neighbors of the spatial outliers

ST-Outliers using Cluster Differentiation

- [Cheng and Li, 2004; Cheng and Li, 2006] propose a four step approach to detect ST-Outliers
- (1) Classification (Clustering) to form regions
- (2) Aggregation: Spatial resolution (scale) of the data is reduced for clustering
- (3) Comparison (Detecting): Results obtained at two spatial scales are compared in order to detect the potential spatial outliers
  - Objects found in step 1 but not in Step 2 are potential ST-Outliers
  - Comparison can be done either by exploratory visualization analysis or visual data mining
- (4) Verification (Checking): The temporal neighbors of the suspected ST-Outliers detected in the previous step are checked to detect ST-Outliers



PCA, Rough Set and Temporal Logic

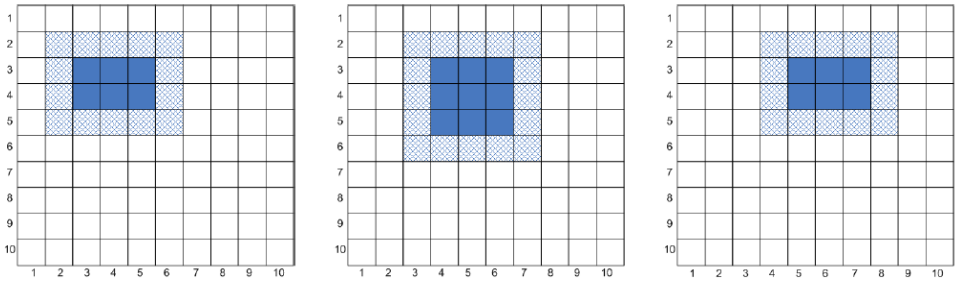
- Principal component analysis has been used to extract ST-Outliers [Lasaponara, 2006]
- [Drosdowsky, 1993] use the rotated PCA (RPCA) method in both S and T modes
- [Albanese et al., 2011] describe a rough set approach called Rough Outlier Set Extraction (ROSE)
- [Jakkula and Cook, 2008] propose various temporal logic rules to capture normal activities
  - Temporal relations for anomaly detection: before, contains, overlaps, meets, starts, started-by, finishes, finished-by, and equals
  - Conditional probability of an event Z given an event Y is P(Z|Y) = (|Before(Z,Y)| + |Contains(Z,Y)| + |Overlaps(Z,Y)| + |Meets(Z,Y)| + |Starts(Z,Y)| + |StartedBy(Z,Y)| + |Finishes(Z,Y)| + |FinishedBy(Z,Y)| + |Equals(Z,Y)|)/|Y|
  - Likelihood of event Z occurring can be computed based on every event observed on a given day to that point in time
  - Anomaly score for event Z can be computed as 1-P(Z)

Outlier Detection for Spatio-Temporal Data

- ST-Outlier Detection
- ST-Outlier Tracking
- Trajectory Outlier Detection

OutStretch Algorithm to Detect Outlier Solids

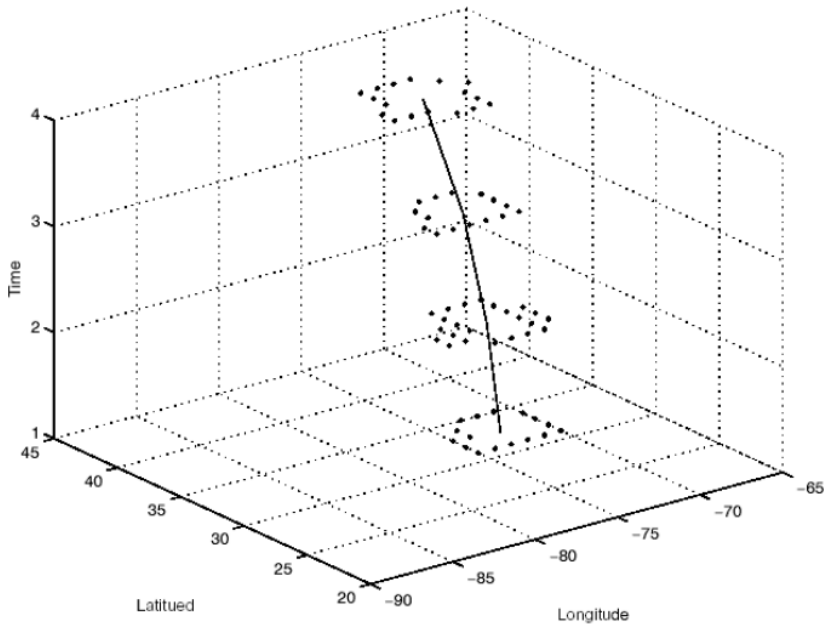
- An ST-Outlier could be considered as a solid with its base in the XY space dimension and volume across the time dimension
  - If there is higher than average precipitation in Peru over the years 1998-2002, then the solid in three dimensional (X, Y and time) space is an outlier
- Outstretch [Wu et al., 2010] tracks the outlier movement patterns of the top-K spatial outliers over several time periods
- Input: Top-K spatial outliers for each year and a variable r, the region stretch, which is the number of grids to 'stretch' by on each side of an outlier
- For all the years, each of the outliers from the current year are examined to see if they are framed by any of the stretched regions from the previous year
- If they are, the item is added to the end of the previous years child list
- As a result, all possible sequences over all years get stored into the outlier tree and can be retrieved for analysis



(a) Outlier at t=1 (b) Outlier at t=2 (c) Outlier at t=3

Wavelets for Tracking ST Outliers

- [Lu and Liang, 2004]
- Wavelet transform is applied to meteorological data to bring up distinct patterns that might be hidden within the original data
- Edge detection with competitive fuzzy classifier, is extended to identify the boundary of region outlier
- To determine the center of the region outlier, the fuzzy-weighted average of the longitudes and latitudes of the boundary locations is computed
- By linking the centers of the outlier regions within consecutive frames, the movement of a region outlier can be captured and traced



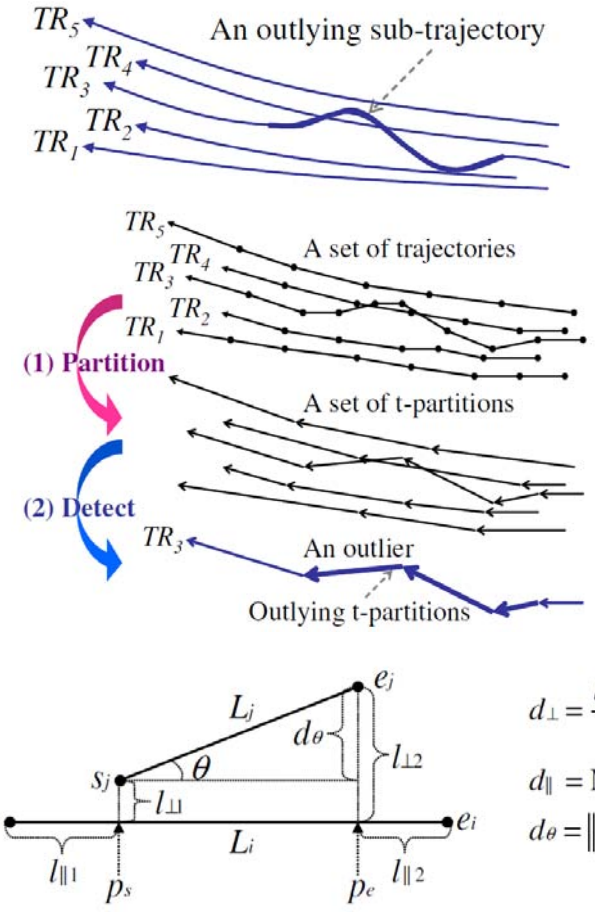
3-D center trajectory for the detected region outlier within each time frame at 0 AM -18 PM, September 18, 2003

Outlier Detection for Spatio-Temporal Data

- ST-Outlier Detection
- ST-Outlier Tracking
- Trajectory Outlier Detection

TRAjectory Outlier Detection (TRAOD)

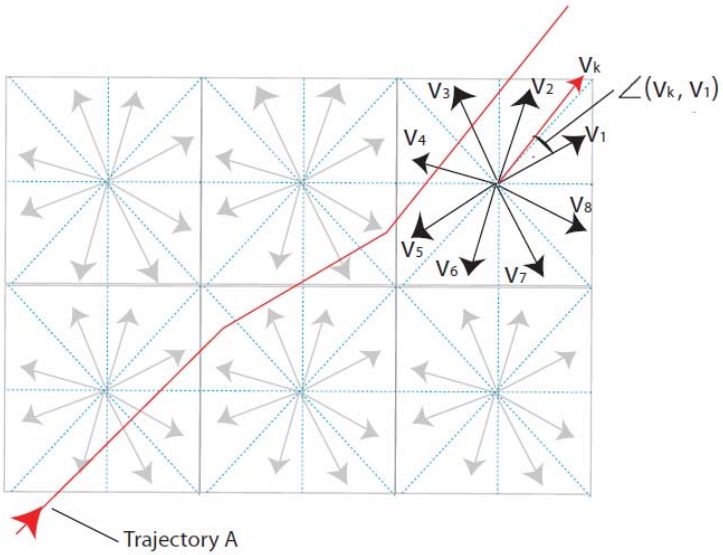
- Given a set of trajectories, find the most anomalous trajectories
- [Lee et al., 2008] propose TRAjectory Outlier Detection (TRAOD) algorithm which consists of two phases
  - 2-level Partitioning Phase
  - Detection Phase
    - A trajectory partition is outlying if it does not have sufficient number of similar neighbors
    - A trajectory is an outlier if sum of length of its outlying partitions is at least F times the sum of lengths of all of its partitions



Three components of distance function for line segments

Evolving Trajectory Outliers (TOP-EYE)

- TOP-EYE [Ge et al., 2010], an evolving trajectory outlier detection method continuously computes the outlying score for each trajectory in an accumulating way
- A decay function enables the evolving computation of accumulated outlying scores along the trajectories
- They consider two types of outlying trajectories: outliers in terms of direction and outliers in terms of density
- The continuous space is discretized into small grids
- Direction based Outliers
  - A probabilistic model is used to turn the direction information of trajectories in a grid into a vector with eight values to indicate the probabilities of moving towards eight directions within this grid
- Density based Outliers
  - The trajectory density within each grid is estimated as the number of trajectories across this grid

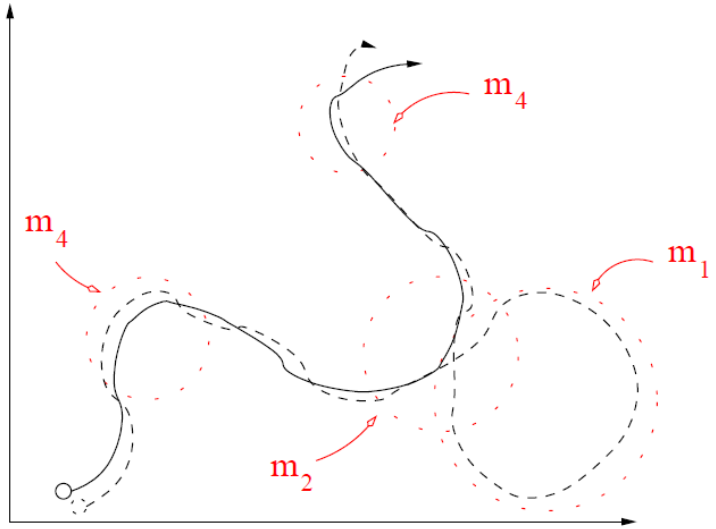


Outlier Road Segments

- [Li et al., 2009] propose a method for detecting temporal outliers with an emphasis on historical similarity trends between data points (rather than spatial continuity)
- At each time step, each road segment checks its similarity vs. other road segments, and historical similarity values are recorded in a temporal neighborhood vector at each road segment
- Outliers are calculated from drastic changes in these vectors
- Each edge is given an exponential reward or penalty each day based on
  - Whether it is historically similar to other road segments
  - Whether it is instantaneously similar to other road segments
- The outlier score of an edge on a particular day is then equal to the sum of rewards and penalties
- The power of this method vs. a method that measures only the singular road segment is that it is robust to population shifts

Motif based Trajectory Outliers

- [Li et al., 2006] propose motion-classifier for trajectory outlier detection with 3 steps
- (1) Motif (feature, time , location) extraction from the object paths
- (2) Motif-based generalization to cluster similar object movement fragments
- (3) Objects are classified by a classifier that can handle high-dimensional generalized motif feature space to discriminate anomalous trajectories from normal ones



Extracting motif expressions in raw paths

Motif Oriented Database

	Motif Expressions	Class
1	(Right-Turn, $3am, l_7$), (U-Turn, $4pm, l_2$), (U-Turn, $10pm, l_1$)	–
2	(U-Turn, $10am, l_2$)	–
3	(Left-Turn, $6am, l_7$), (U-Turn, $11am, l_3$), (Right-Turn, $1pm, l_7$), (Right-Turn, $4pm, l_7$)	+
4	(Right-Turn, $1am, l_1$), (U-Turn, $9am, l_1$), (Left-Turn, $3pm, l_3$), (U-Turn, $3pm, l_3$)	–
5	(Right-Turn, $2am, l_1$), (Left-Turn, $9pm, l_3$)	–

Outlier Detection for Temporal Network Data

- Graph Similarity Outliers
- Community Based Outliers
- Online Graph Outliers

Introduction

- Given a series of graph snapshots
- Time series of graph distance metrics can be individually modeled using univariate autoregressive moving average (ARMA) model
- Outliers are time points where the actual and predicted values differ greater than a threshold
- A large variety of similarity/distance measures have been proposed to compare two graph snapshots
- Notations
  - V_G and V_H are vertex sets for G and H resp. If $V_G = V_H$, V is used
  - E_G and E_H are edges in graphs G and H

Graph Similarity/Distance Measures (1)

1. Weight Distance [Papadimitriou et al., 2010; Pincombe, 2005]

$$d(G,H) = \frac{\sum_{u,v \in V} \frac{\left| w_E^G(u,v) - w_E^H(u,v) \right|}{\max\left(w_E^G(u,v), w_E^H(u,v) \right)}}{|E_G \cup E_H|}$$

2. MCS Weight Distance [Pincombe, 2005]
   - Same as weight distance but only for edges in MCS where the maximum common subgraph (MCS) F of G and H is the common subgraph with the most vertices
3. MCS Edge Distance [Pincombe, 2005]

$$d(G,H) = 1 - \frac{|mcs(E_G, E_H)|}{max(|E_G|, |E_H|)}$$

Graph Similarity/Distance Measures (2)

4. MCS Vertex Distance [Pincombe, 2005]

$$d(G,H) = 1 - \frac{|mcs(V_G, V_H)|}{max(|V_G|, |V_H|)}$$

5. Median Graph Edit Distance [Dickinson et al., 2002; Pincombe, 2005]

6. Modality Distance [Pincombe, 2005]

   – Absolute value of the difference between the Perron vectors (principal eigen vector of adjacency matrix) of these graphs

Graph Similarity/Distance Measures (3)

7. Graph Edit Distance [Papadimitriou et al., 2008; Pincombe 2005; Shoubridge et al., 1999]

$d(G,G') = |V|+|V'|-2|V \cap V'|+|E|+|E'|-2|E \cap E'|$

$$d_1(G,H) = \sum_{n \in V_G \setminus (V_G \cap V_H)} C_{nd}(n) + \sum_{n \in V_H \setminus (V_G \cap V_H)} C_{ni}(n) + \sum_{e \in E_G \cap E_H} C_{es}(e) + \sum_{e \in E_G \setminus (E_G \cap E_H)} C_{ed}(e) + \sum_{e \in E_H \setminus (E_G \cap E_H)} C_{ei}(e),$$

$$d_2(G,H) = c\,[|V_G| + |V_H| - 2|V_G \cap V_H|] + \sum_{e \in E_G \cap E_H} |\beta_G(e) - \beta_H(e)| + \sum_{e \in E_G \setminus (E_G \cap E_H)} \beta_G(e) + \sum_{e \in E_H \setminus (E_G \cap E_H)} \beta_H(e).$$

$$d_3(G,H) = c\,[|V_G| + |V_H| - 2|V_G \cap V_H|] + \sum_{e \in E_G \cap E_H} \frac{|(\beta_G(e)+\epsilon) - (\beta_H(e)+\epsilon)|^2}{(\min(\beta_G(e), \beta_H(e)) + \epsilon)^2} + \sum_{e \in E_G \setminus (E_G \cap E_H)} (\beta_G(e)+\epsilon)^2 + \sum_{e \in E_H \setminus (E_G \cap E_H)} (\beta_H(e)+\epsilon)^2$$

$C_{nd}(n)$=cost of deleting node n
$C_{ni}(n)$=cost of inserting node n
$C_{es}(n)$=cost of substituting an edge weight for edge e
$C_{ed}(n)$=cost of deleting edge e
$C_{ei}(n)$=cost of inserting edge e
C=tradeoff parameter
$\beta(e)$=weight of edge e
ϵ=smoothing parameter (set to 1)

Graph Similarity/Distance Measures (4)

8. Diameter Distance [Gaston et al., 2006; Pincombe, 2005]

– difference in the diameters for each graph

9. Entropy Distance [Gaston et al., 2006; Pincombe, 2005]

$$E(*) = -\sum_{e \in E_*} (\hat{w}_*^e - ln\hat{w}_*^e) \text{ where } \hat{w}_*^e = \frac{w_*^e}{\sum_{e \in E_*} w_*^e}$$

10. Spectral Distance [Gaston et al., 2006; Pincombe, 2005]

$$d(G, H) = \sqrt{\frac{\sum_{i=1}^{k} (\lambda_i - \mu_i)^2}{min(\sum_{i=1}^{k} \lambda_i^2, \sum_{i=1}^{k} \mu_i^2)}}$$

Graph Similarity/Distance Measures (5)

11. Umeyama graph distance [Dickinson and Kraetzl, 2003]

$$dist(G, H) = \sum_{u,v \in V} [w_E^G(u, v) - w_E^H(u, v)]^2$$

12. The Euclidean distance between the principal eigenvectors of the graph adjacency matrices (Vector Similarity) [Papadimitriou et al., 2008]

13. Spearman's correlation coefficient [Papadimitriou et al., 2010]

– rank correlation between sorted (based on PageRank) lists of vertices of the two graphs

Graph Similarity/Distance Measures (6)

14. Sequence similarity [Papadimitriou et al., 2010; Papadimitriou et al., 2008]
    - Similarity of vertex sequences of the graphs that are obtained through a graph serialization algorithm
15. Signature similarity [Papadimitriou et al., 2010; Papadimitriou et al., 2008]
    - Hamming distance between appropriate fingerprints of two graphs
16. Vertex/edge overlap (VEO) [Papadimitriou et al., 2010]

$$sim_{VEO}(G,H) = 2\frac{|V_G \cap V_H| + |E_G \cap E_H|}{|V_G| + |V_H| + |E_G| + |E_H|}$$

17. Vertex ranking (VR) [Papadimitriou et al., 2010]

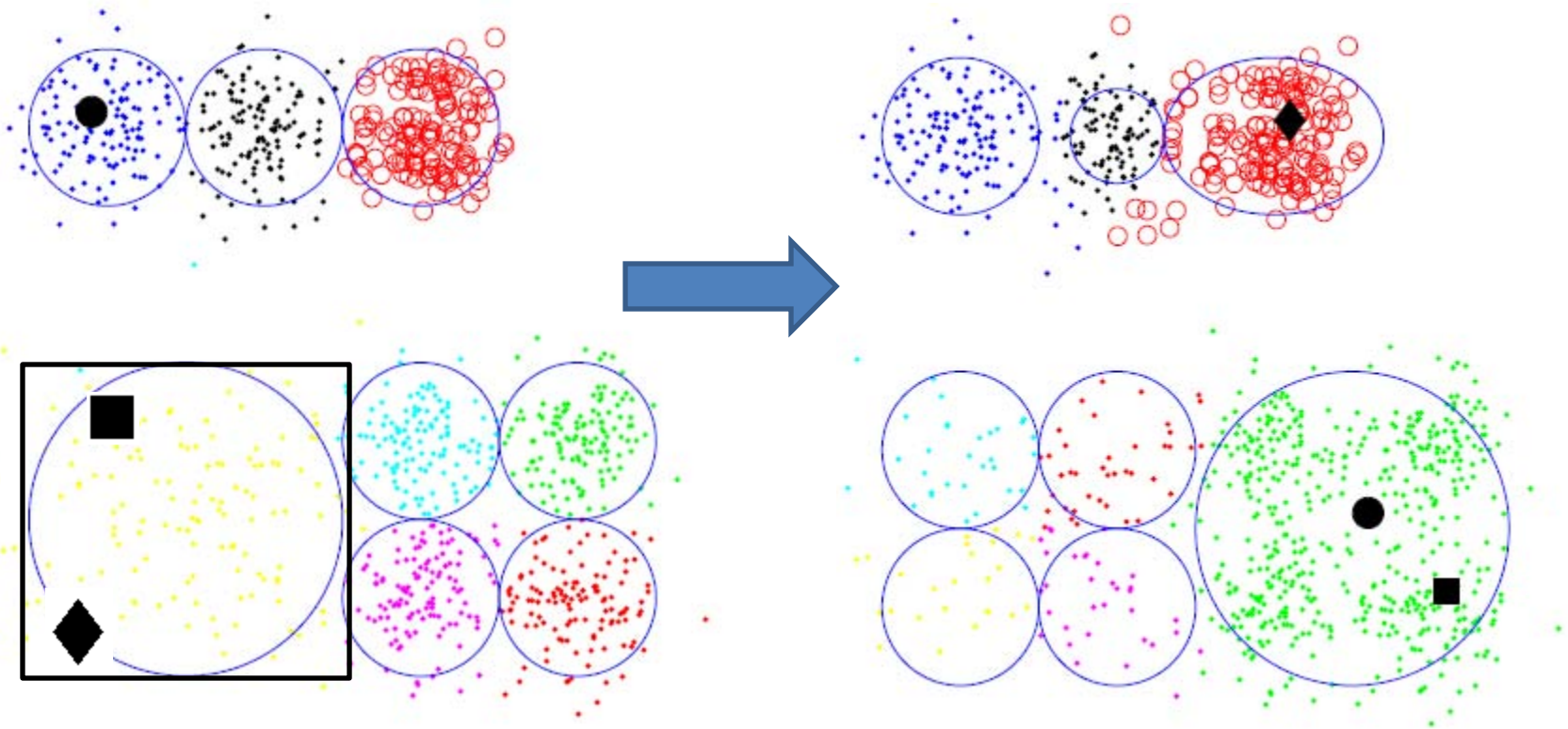
$$sim_{VR}(G,H) = 1 - \frac{2\sum_{v \in V_G \cup V_H} w_v \times (\pi_v - \pi'_v)^2}{D}$$

w is PageRank value, π is the vertex rank, D is normalization constant

Outlier Detection for Temporal Network Data

- Graph Similarity Outliers
- Community Based Outliers
- Online Graph Outliers

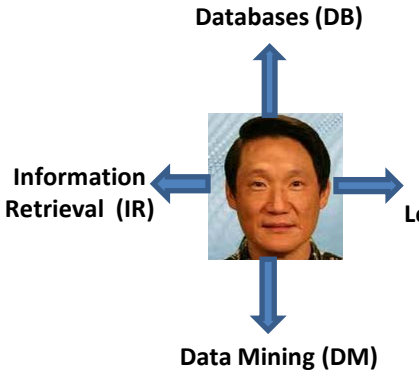
Evolutionary Community Outliers (ECOutliers)



gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

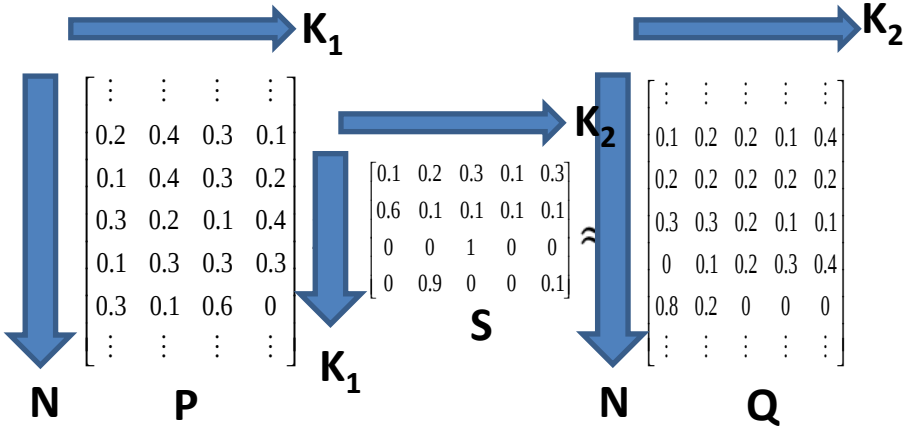
ECOutlier: Definitions

Belongingness Matrix



$$\begin{array}{c} \text{DM} \quad \text{IR} \quad \text{ML} \quad \text{DB} \\ \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ 0.2 & 0.4 & 0.3 & 0.1 \\ 0.1 & 0.4 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.1 & 0.4 \\ 0.1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.1 & 0.6 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

Community-Community Correspondence Matrix

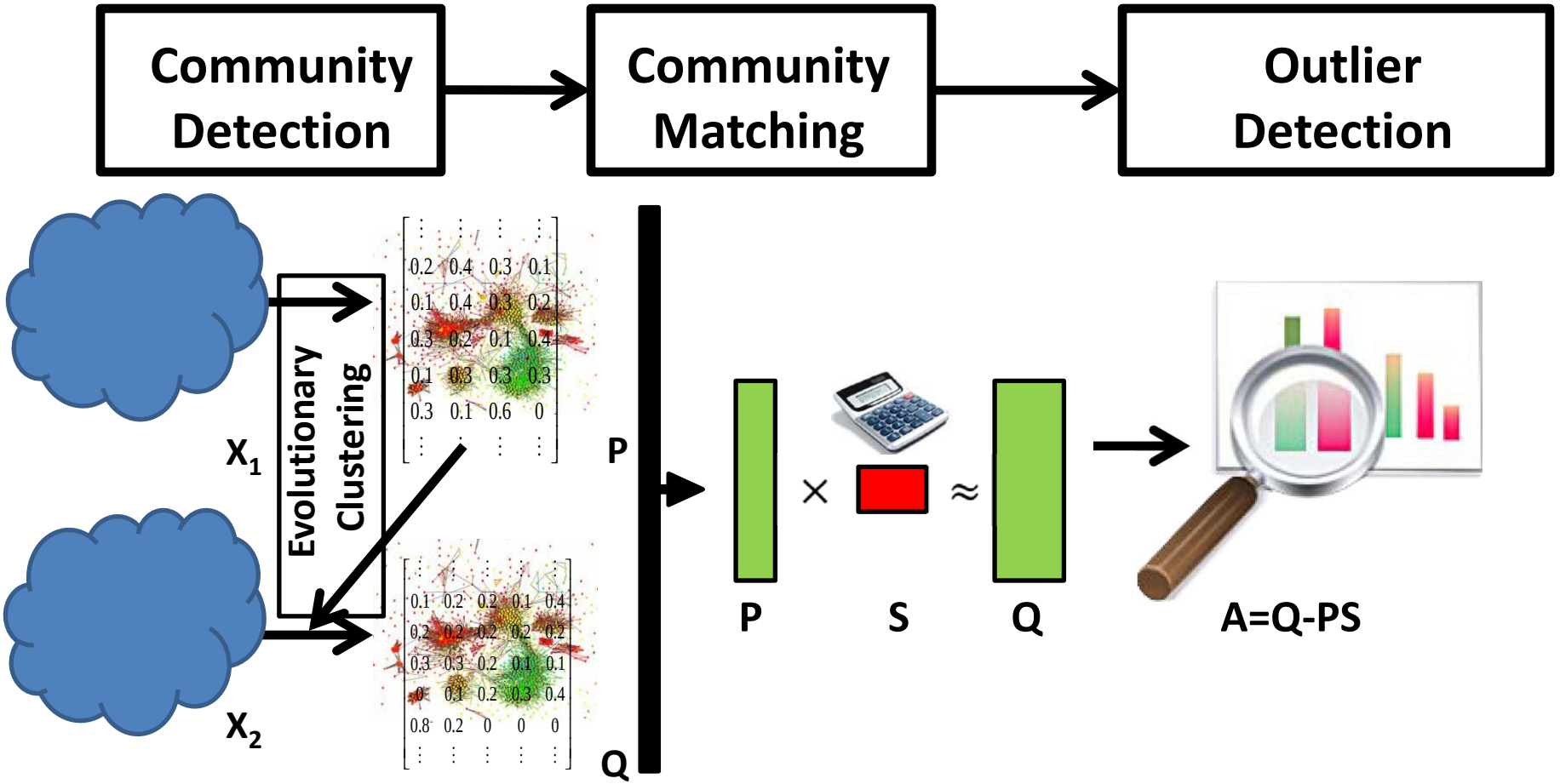


$$P \in [0,1]^{N \times K_1} \qquad \sum_{i=1}^{K_1} p_{oi} = 1$$

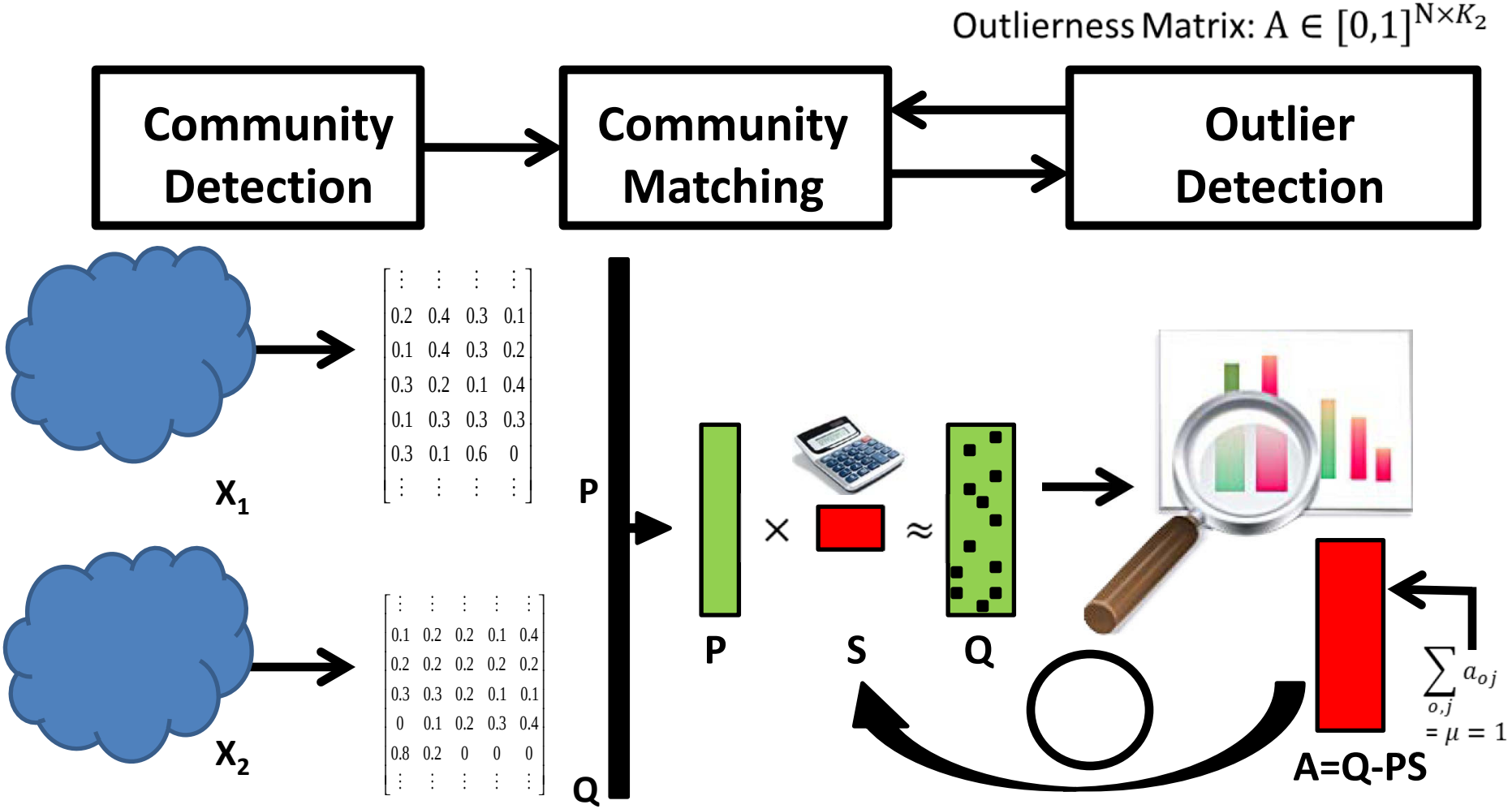
$$Q \in [0,1]^{N \times K_2} \qquad \sum_{j=1}^{K_2} q_{oj} = 1$$

$$S \in [0,1]^{K_1 \times K_2} \qquad \sum_{j=1}^{K_2} s_{ij} = 1$$

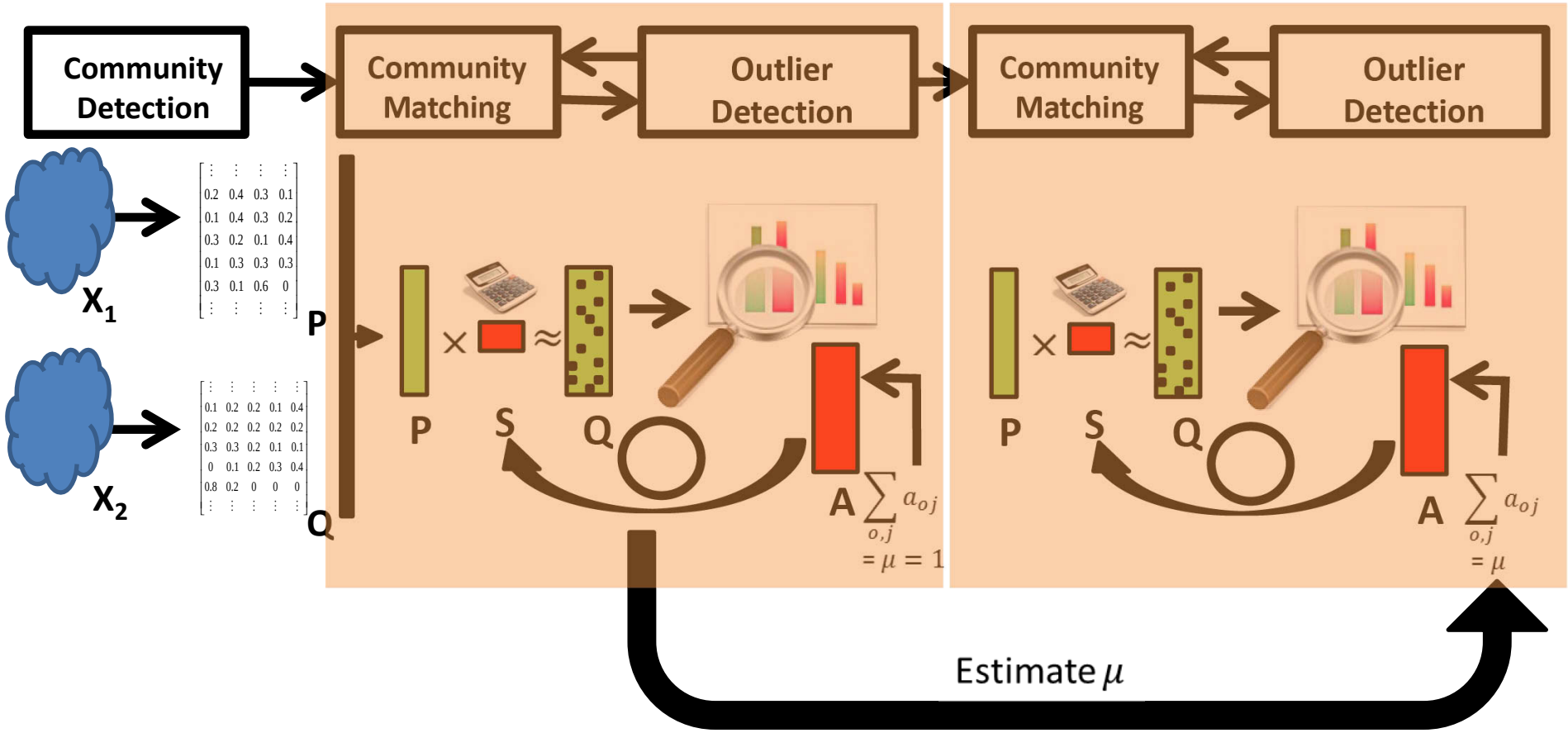
TwoStage ECOutlier Detection Framework



OneStage ECOutlier Detection Framework



OneStageμ ECOutlier Detection Framework



ECOutlier: Community Matching and Outlier Detection Together

$$P \times S \approx Q$$

Given P and Q, estimate S and A

Minimize element-wise distance between Q and PS wrt outlier-weighted elements

subject to the following constraints

Correspondence matrix should be right stochastic

Maximum amount of outlierness should be controlled

- N = #objects
- K_1 = #clusters in X_1
- K_2 = #clusters in X_2
- P^{NXK_1} = belongingness matrix for X_1
- Q^{NXK_2} = belongingness matrix for X_2
- $S^{K_1XK_2}$ = correspondence matrix
- A^{NXK_2} = outlierness matrix
- μ = maximum level of overall outlierness

gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

ECOutlier: Community Matching and Outlier Detection Together

$$P \times S \approx Q$$

Given P and Q, estimate S and A

- $\min \sum_{o=1}^{N} \sum_{j=1}^{K_2} \log\left(\frac{1}{a_{oj}}\right) (q_{oj} - \overrightarrow{p_{o.}} \cdot \overrightarrow{s_{.j}})^2$
- subject to the following constraints
- $\sum_{j=1}^{K_2} s_{ij} = 1 \quad (\forall i = 1 \dots K_1)$

Inequality

- $\sum_{o=1}^{N} \sum_{j=1}^{K_2} a_{oj} \leqq \mu$
- $s_{ij} \geq 0 \quad (\forall i = 1 \dots K_1, \forall j = 1 \dots K_2)$
- $1 \geq a_{oj} \geq 0 \quad (\forall o = 1 \dots N, \forall j = 1 \dots K_2)$

- N = #objects
- K_1 = #clusters in X_1
- K_2 = #clusters in X_2
- P^{NXK_1} = belongingness matrix for X_1
- Q^{NXK_2} = belongingness matrix for X_2
- $S^{K_1XK_2}$ = correspondence matrix
- A^{NXK_2} = outlierness matrix
- μ = maximum level of overall outlierness

ECOutlier: Community Matching and Outlier Detection Together

Convex Optimization

Negative Log ensures outliers will be within small range. Hence, convex in a_{oj}

Quadratic in s_{ij}. Hence, convex in s_{ij}

$$\min \sum_{o=1}^{N} \sum_{j=1}^{K_2} \log\left(\frac{1}{a_{oj}}\right)\left(q_{oj} - \overrightarrow{p_{o.}} \cdot \overrightarrow{s_{.j}}\right)^2$$

subject to the following constraints

$$\sum_{j=1}^{K_2} s_{ij} = 1 \quad (\forall i = 1 \dots K_1)$$

$$\sum_{o=1}^{N} \sum_{j=1}^{K_2} a_{oj} = \mu$$

$$s_{ij} \geq 0 \quad (\forall i = 1 \dots K_1, \forall j = 1 \dots K_2)$$

$$1 \geq a_{oj} \geq 0 \quad (\forall o = 1 \dots N, \forall j = 1 \dots K_2)$$

$$P \times S \approx Q$$

Given P and Q, estimate S and A

- N = #objects
- K_1 = #clusters in X_1
- K_2 = #clusters in X_2
- P^{NXK_1} = belongingness matrix for X_1
- Q^{NXK_2} = belongingness matrix for X_2
- $S^{K_1XK_2}$ = correspondence matrix
- A^{NXK_2} = outlierness matrix
- μ = maximum level of overall outlierness

ECOutliers: Derivation of Update Rules

- Differentiate wrt a_{oj} and set it to 0

Community Matching Error for the $(o,j)^{th}$ entry

Outlierness for the $(o,j)^{th}$ entry

$$a_{oj} = \frac{\left(q_{oj} - \overrightarrow{p_{o.}}.\overrightarrow{s_{.j}}\right)^2 \mu}{\sum_{o'=1}^{N}\sum_{j'=1}^{K}\left(q_{o'j'} - \overrightarrow{p_{o'.}}.\overrightarrow{s_{.j'}}\right)^2}$$

Total outlierness in the snapshot

Total Matching Error

- Differentiate wrt s_{ij} and set it to 0

Correspondence for community i in X_1 and j in X_2

$$s_{ij} = \frac{\sum_{o'=1}^{N} 2\log\left(\frac{1}{a_{o'j}}\right) p_{o'i}\left[q_{o'j} - \sum_{\substack{k=1 \\ k\neq i}}^{K} p_{o'k}s_{kj}\right] - \beta_i}{\sum_{o'=1}^{N} 2\log\left(\frac{1}{a_{o'i}}\right) p_{o'i}^2}$$

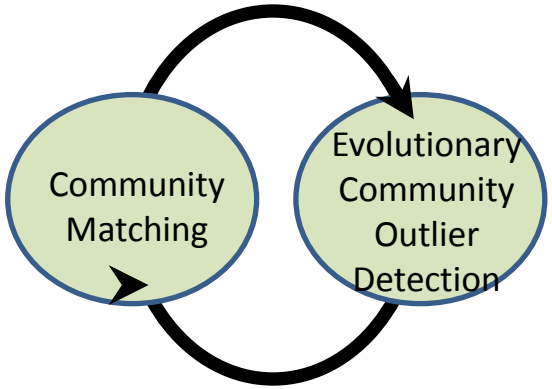
Part of $q_{o'j}$ that needs to be explained by $p_{o'i}$ s_{ij}

Give outliers lower weight when matching

Normalization to make sure that $\sum_i s_{ij} = 1$

ECOutlier Detection Algorithm (OneStageμ)

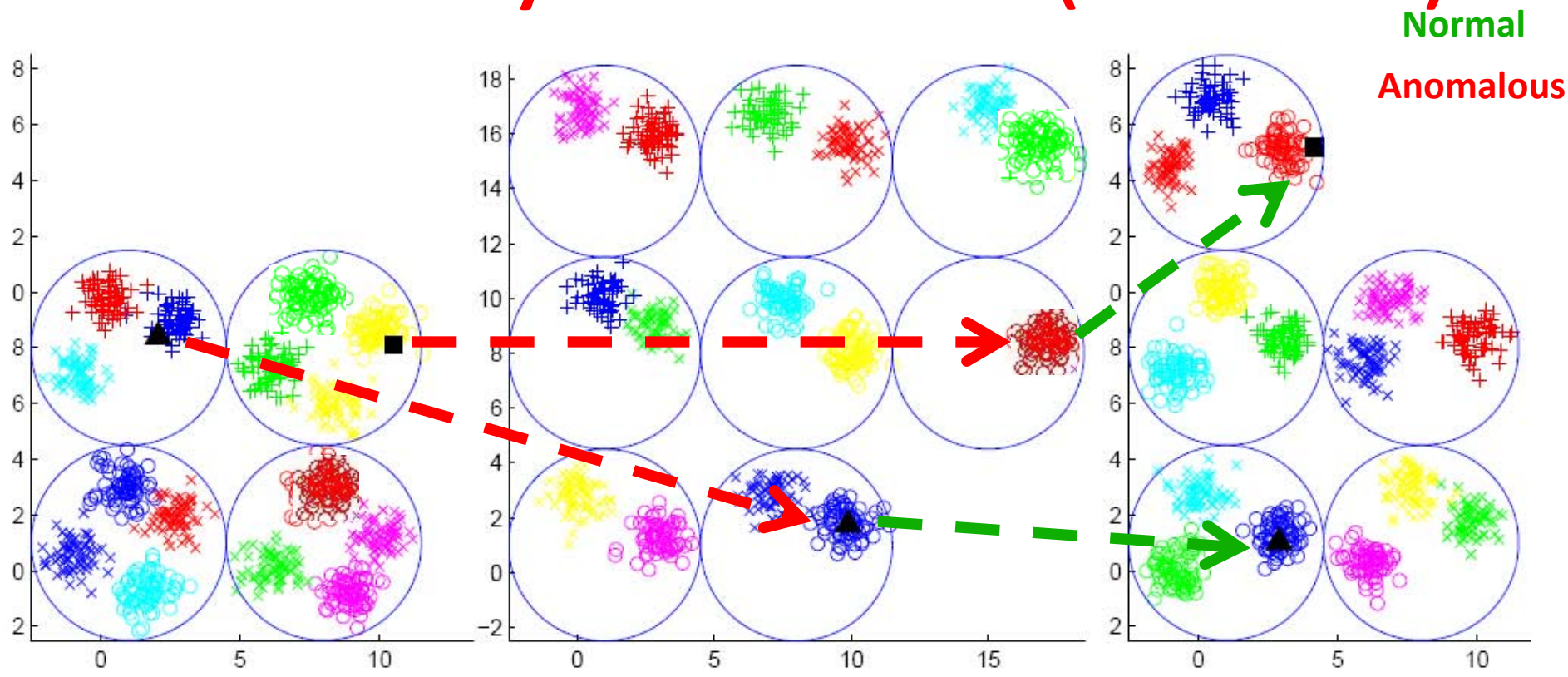
- Input: P and Q
- Output: Estimates of S and A
- Initialize μ, all s_{ij} to $\frac{1}{K_2}$ and all a_{oj} to $\frac{1}{NK_2}$
- While (not converged)
  - Compute A (Outlier Detection step)
  - Compute S (Community Matching step)
- Estimate $\mu = \frac{\sum_{o'=1}^{N}\sum_{j'=1}^{K_2}(q_{o'j'} - \overrightarrow{p_{o'.}}.\overrightarrow{s_{.j'}})^2}{\max_{o,j}(q_{oj}^2)}$
- While (not converged)
  - Compute A (Outlier Detection step)
  - Compute S (Community Matching step)



$O(NK_1^2K_2t)$
N=#objects
K_1=#clusters in X_1
K_2=#clusters in X_2
t=#iterations

- Two pass algorithm
- Coordinate descent iterative computation of S and A

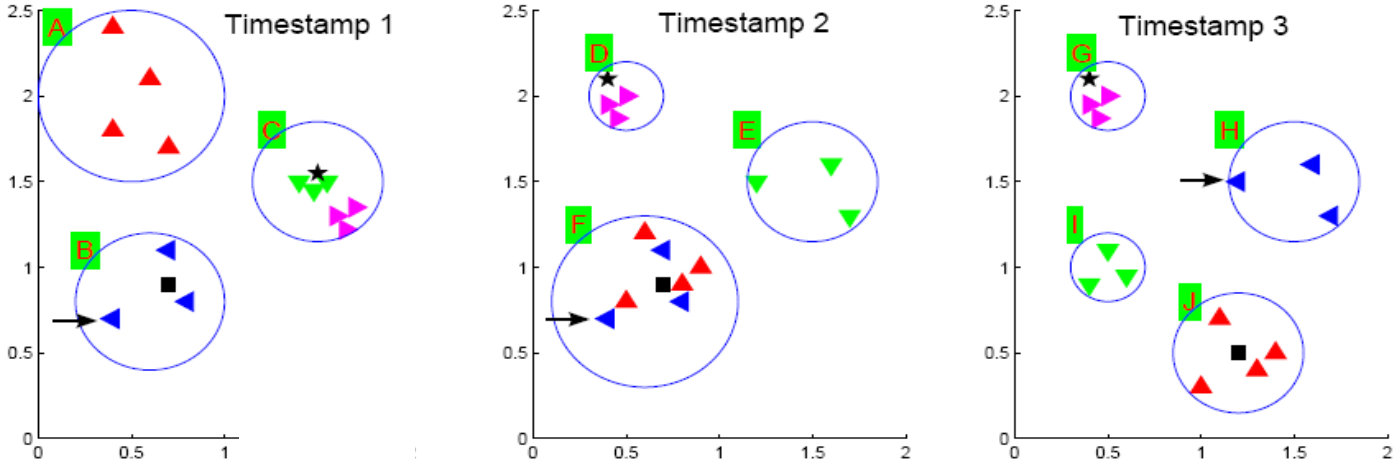
Community Trend Outliers (CTOutliers)



Community Trend Outliers: Nodes for which evolutionary behaviour across a series of snapshots is quite different from that of its community members

CTOutliers: Soft Sequence Representation

- Every object has a distribution associated with it across time
  - In a co-authorship network, an author has a distribution of research areas associated with it across years



Soft sequence for object denoted by (→)

<1: (A:0.1 , B:0.8 , C:0.1) , 2: (D:0.07 , E:0.08 , F:0.85) , 3: (G:0.08 , H:0.8 , I:0.08 , J:0.04)>

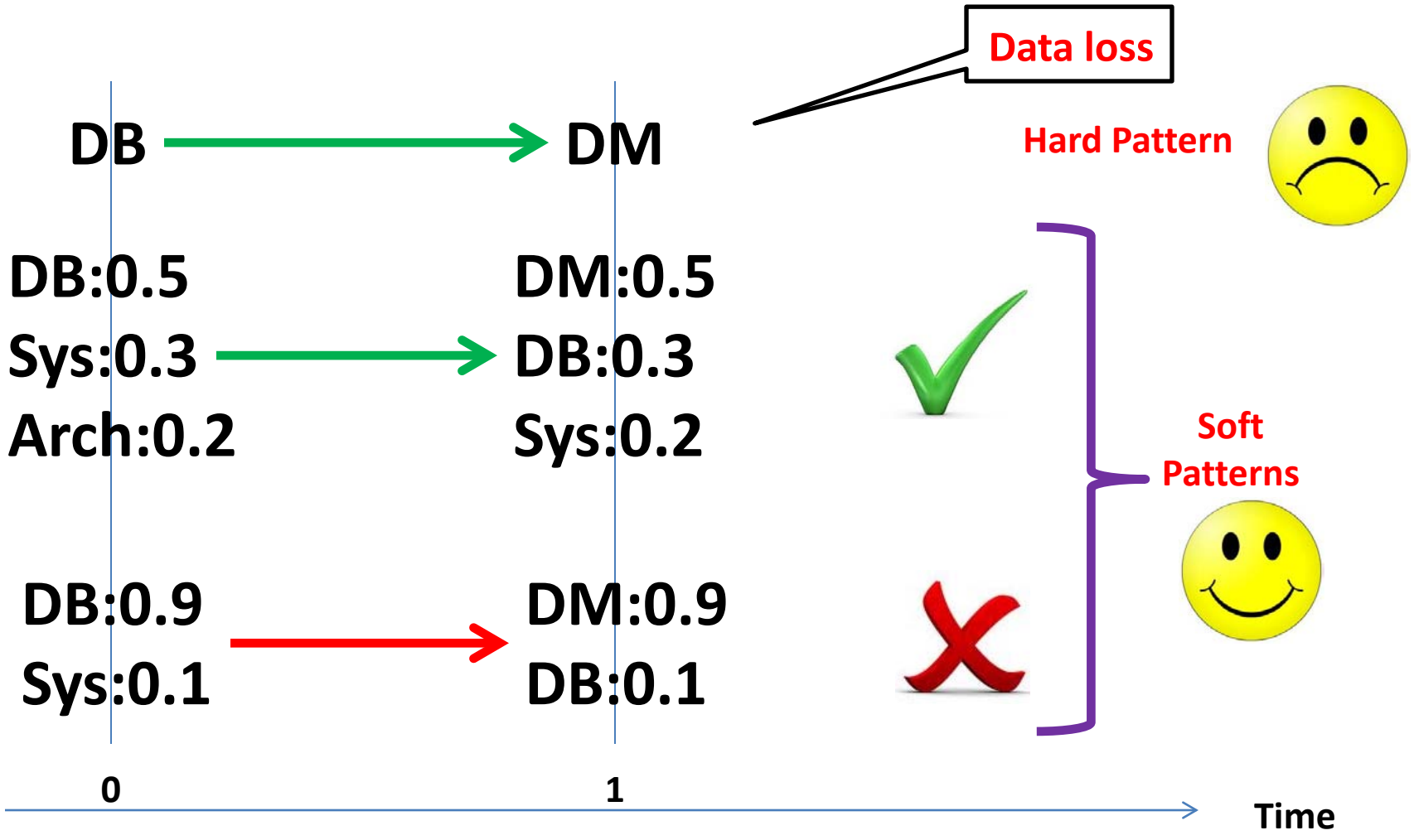
Hard sequence is <1:B, 2:F, 3:H>

Outliers: ■ and ★

CTOutliers: Problem Formulation

- Problem
  - Input: Soft sequences (each of length T) for N objects, denoted by matrix S
  - Output: Set of *CTOutlier* objects
- SubProblems
  - Pattern Extraction
    - Input: Soft sequences (S)
    - Output: Frequent soft patterns (P)
  - Outlier Detection
    - Input: Frequent soft patterns (P)
    - Output: Set of *CTOutlier* objects

CTOutliers: Benefits of Soft Patterns



gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

CTOutliers: Support for Soft Patterns

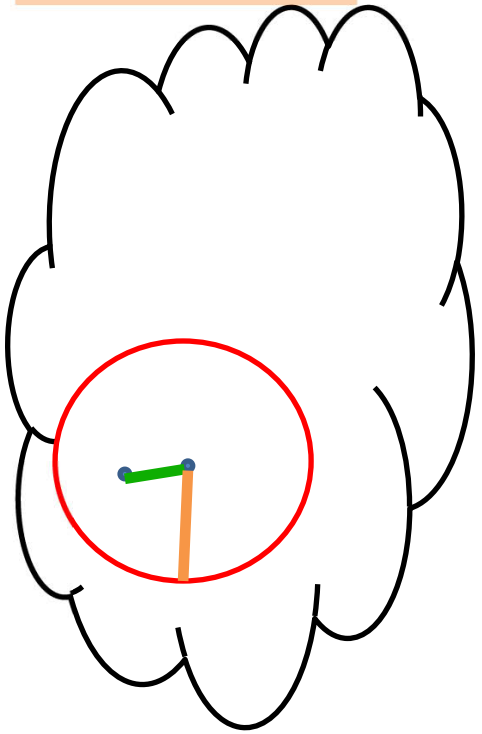
Notation	Meaning
min_sup	Minimum Support
t	Index for timestamps
o	Index for objects
p	Index for patterns
N	Total number of objects
T	Total number of timestamps
S_{t_o}	Distribution for object o at time t
P_{t_p}	Distribution for pattern p at time t
TS_p	Set of timestamps for pattern p

$$sup\left(P_{t_p}\right) = \sum_{o=1}^{N}\left[1 - \frac{Dist\left(S_{t_o}, P_{t_p}\right)}{maxDist\left(P_{t_p}\right)}\right]$$

For longer patterns

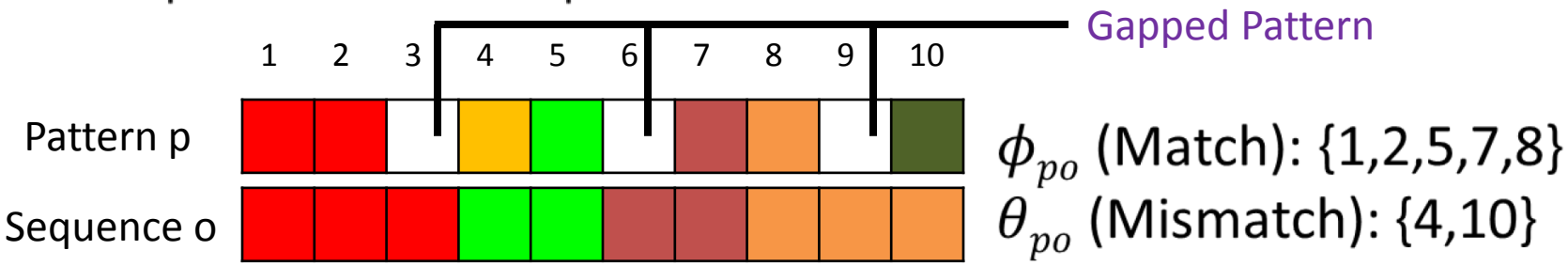
Candidate generation uses Apriori

$$sup(p) = \sum_{o=1}^{N}\prod_{t\in TS_p}\left[1 - \frac{Dist\left(S_{t_o}, P_{t_p}\right)}{maxDist\left(P_{t_p}\right)}\right]$$



CTOutlier Detection

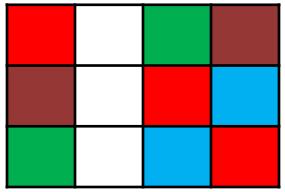
- Given: Set of soft patterns (P) and set of sequences (S)
- Output: Find outlier sequences



- $score(o,p) = \sum_{t \in \theta_{po}} \left[\sup\left(P_{t_p}\right) \times \frac{Dist(S_{t_o}, P_{t_p})}{maxDist(P_{t_p})} \right]$
- $score(o) = \sum_{p \in P} \sup(p) \times score(o,p)$
  - But object o may follow only one pattern! So, sum may be incorrect
- $score(o) = \min[\sup(p) \times score(o,p)]$
  - But generally $\sup(p) \times score(o,p)$ will be min for very short pattern p mostly of length 2

CTOutlier Score using Pattern Configurations

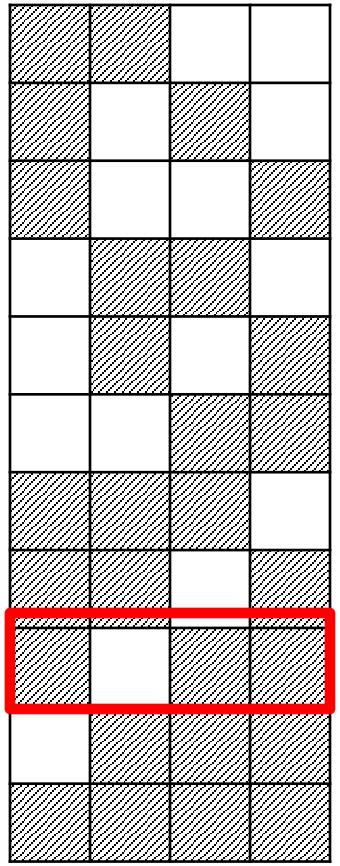
- Divide pattern space into different "projections" called configurations
- A configuration is a set of timestamps of size>1
- E.g., {1,3,4} is a configuration



$$score(o) = \sum_{c=1}^{|C|} score(c, o) = \sum_{c=1}^{|C|} score(bmp_{oc}, o)$$

where bmp_{oc} is the best matching pattern for object o given the configuration c, and C is the set of all configurations

T=4



CTOutlier: Finding Best Matching Pattern

- Find all patterns that are defined exactly for configuration c
- For each such pattern

$$match(o,p) = \sum_{t \in \phi_{po}} \frac{\sup\left(P_{t_p}\right) \times \sup(P_{t_p}, S_{t_o})}{avgDist(P_{t_p})}$$

- Match Score is high if
  - Timestamps where the o and p match are high
  - p has higher support
  - p represents compact clusters
  - o is close to the cluster centroid of p across the various timestamps
- Best matching pattern for o is pattern with highest match(o,p)

CTOutlier Score (Sequence, Best Matching Pattern)

- Given a sequence s and a configuration c
  - Compute best matching pattern q=bmp_{oc}
  - Next, we compute outlier score as

$$score(o,q) = \sup(q) \times \sum_{t \in \theta_{qo}} \left[\sup\left(P_{t_q}\right) \times \frac{Dist(S_{t_o}, P_{t_q})}{maxDist(P_{t_q})} \right]$$

Mismatch between q and o at time t

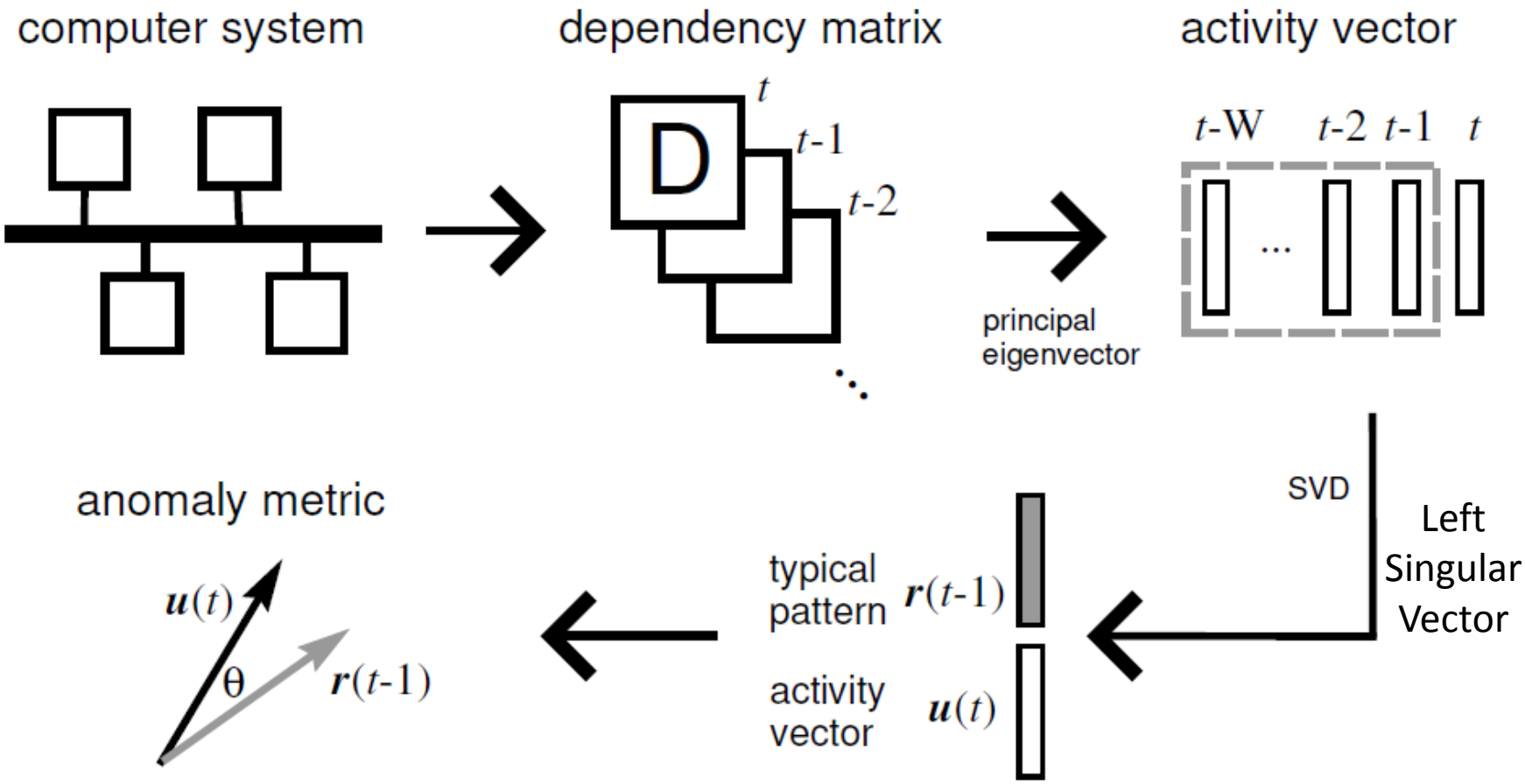
- Outlier score is high if
  - Mismatch for a large number of timestamps
  - Sequence is "far away" from patterns for many timestamps, especially if the pattern is compact for those timestamps

Outlier Detection for Temporal Network Data

- Graph Similarity Outliers
- Community Based Outliers
- Online Graph Outliers

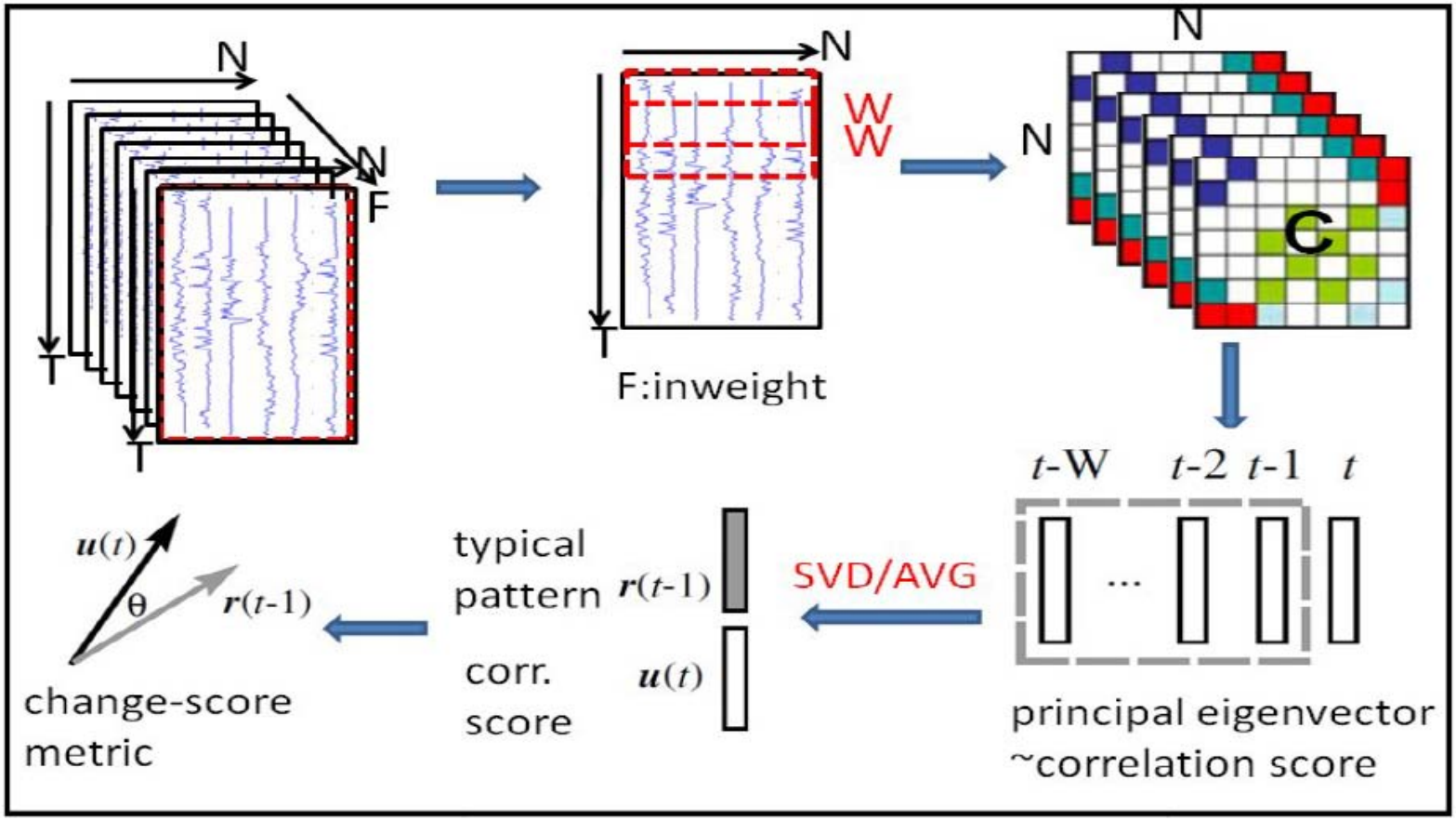
Eigenspace-based Anomaly Detection

[Ide and Kashima, 2004]



Outliers in Mobile Communication Graphs

[Akoglu and Faloutsos, 2010]



Structural Outlier Detection

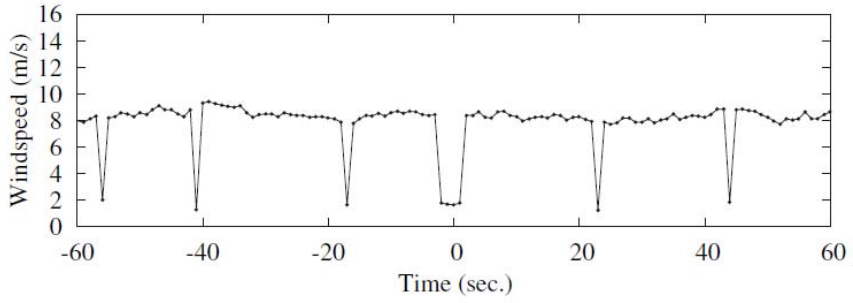
- [Aggarwal et al., 2011] propose the problem of structural outlier detection in massive network streams
- Outliers are graph objects which contain unusual bridging edges
- The network is dynamically partitioned in order to construct statistically robust models of the connectivity behavior
- For robustness, multiple such partitionings are maintained
- These models are maintained with the use of an innovative reservoir sampling approach for efficient structural compression of the underlying graph stream
- Using these models, edge generation probability is defined and then graph object likelihood fit is defined as the geometric mean of the likelihood fits of its constituent edges
- Those objects for which this fit is t standard deviations below the average of the likelihood probabilities of all objects received so far are reported as outliers

Applications of Temporal Outlier Detection Techniques

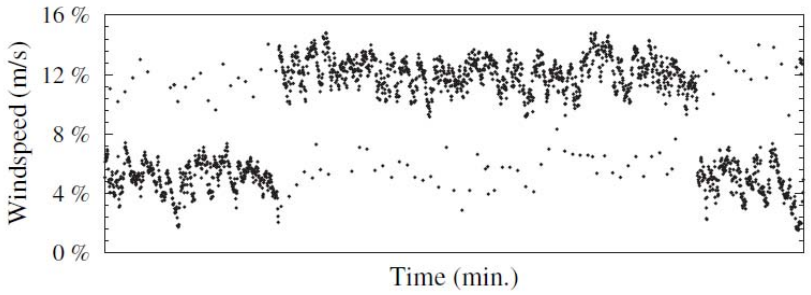
- Environmental Sensor Data
- Industrial Sensor Data
- Surveillance/Trajectory Data
- Computer Networks Data
- Biological Data
- Astronomy Data
- Web Data
- Information Networks Data
- Economics Time Series Data

Wind Speed Measurement Errors

- Temperature and humidity sensors: to detect climate change events
- [Hill and Minsker, 2010; Hill et al., 2007] identify measurement errors in a wind speed data stream from WATERS Network Corpus Christi Bay testbed, provided by the Shoreline Environmental Research Facility (SERF), Texas



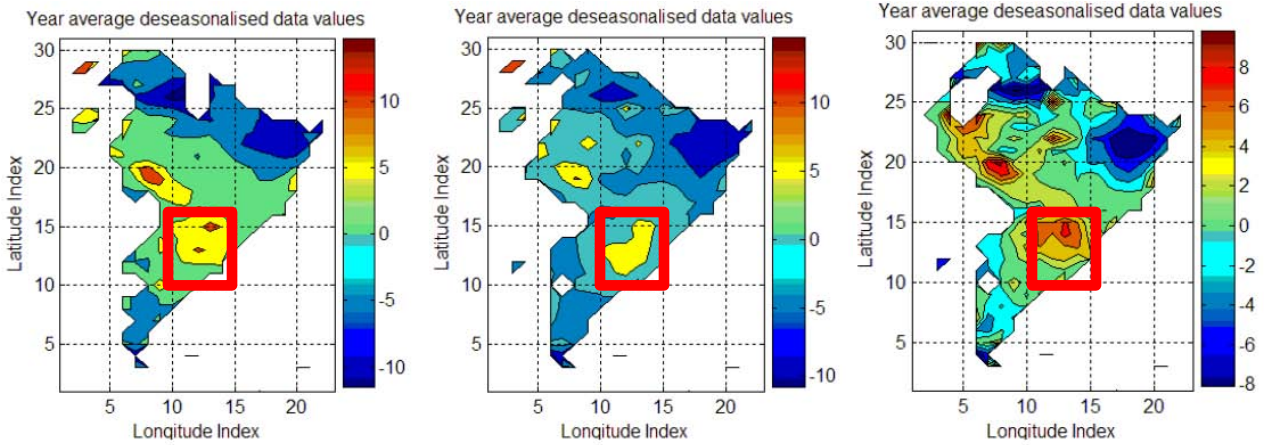
Data exhibiting errors resulting from short duration faults



Data exhibiting errors resulting from long duration faults

Hydrology Datasets

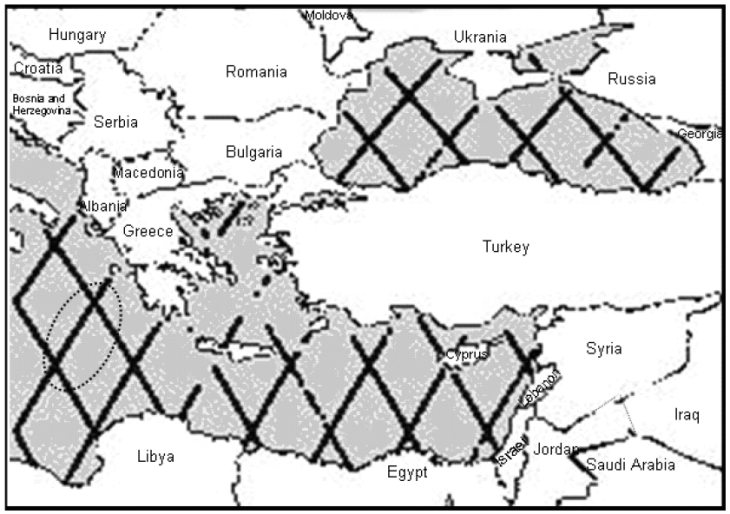
- [Sun et al., 2005] explore a 0.5 degree grid dataset over South China area monthly from 1992 to 2002 with five variables: cloud cover percentage, diurnal temperature range, precipitation, temperature, vapor and pressure
  - Outliers: locations with different temperature from their neighborhoods in recent 10 years
  - They find droughts and flood outliers like flood in Yangzi River Valley in 1998
- [Wu et al., 2010] find extreme precipitation events from South American precipitation data set obtained from the NOAA (Hydrology) for Peru over 1998-2002
  - They also compare the behavior of the outlier regions to the El Niño and La Niño phases of the El Niño Southern Oscillation (ENSO) phenomenon



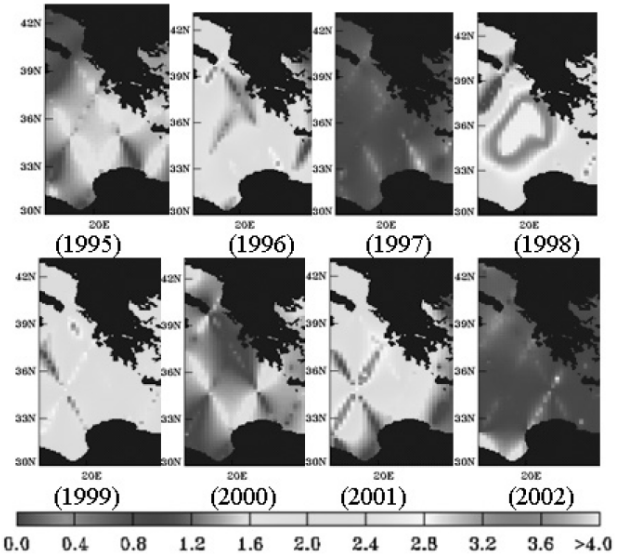
(a) 2002 (b) 2003 (c) 2004

Rain, Precipitation, Wave Heights

- [Angiulli and Fassetti, 2007] study the rain, sea surface temperature, relative humidity, precipitation time series obtained from the Pacific Marine Environmental Laboratory of the U.S. National Oceanic and Atmospheric Administration (NOAA)
- [Birant and Kut, 2006] work with the wave height values of four seas: the Black Sea, the Marmara Sea, the Aegean Sea, and the east of the Mediterranean Sea. Outliers are locations with significantly high wave height values on a particular date compared to its spatiotemporal neighbors



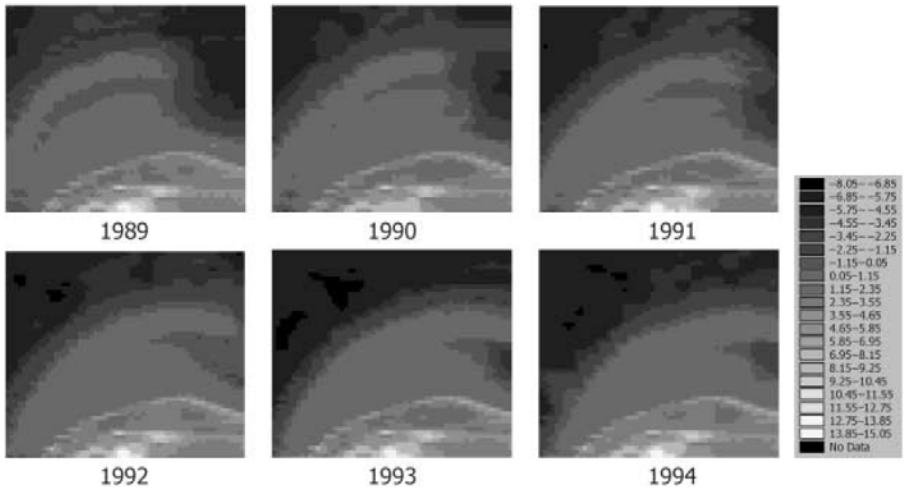
The region circled in dashed lines contains S-Outliers (January 24, 1998).



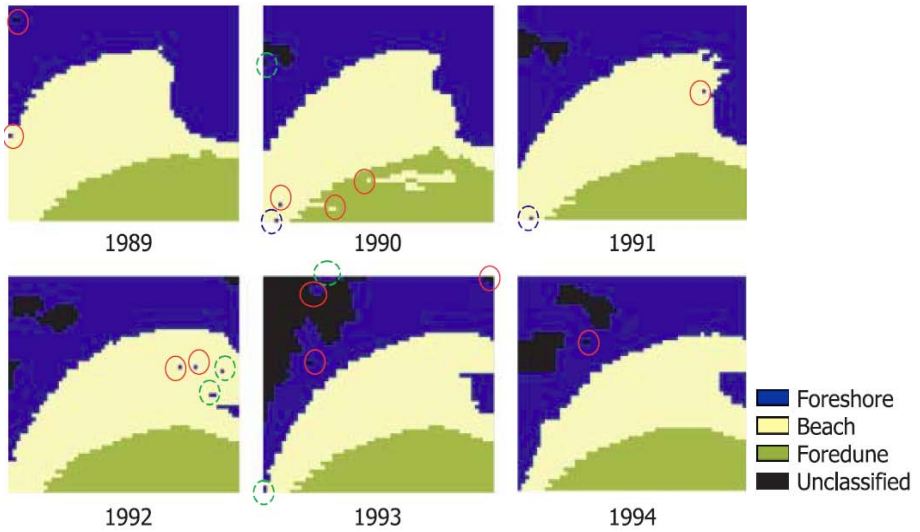
Wave height values of the same region on the same day, but in different years

Water Height based Spatio-temporal Outliers

- [Cheng and Li, 2006] compute spatio-temporal outliers from the water height data obtained from Ameland, a barrier island in the north of the Netherlands. They classify the coastal areas into foreshore, beach and foredune based on observed water height across multiple years and then find outliers

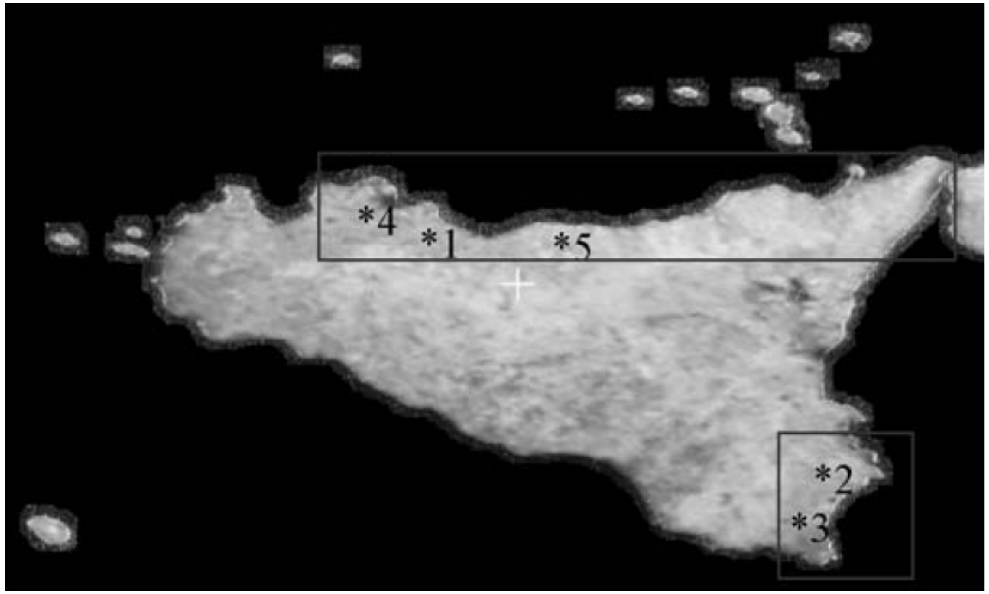


Digital Elevation Model (DEMs) of Ameland in six consecutive years

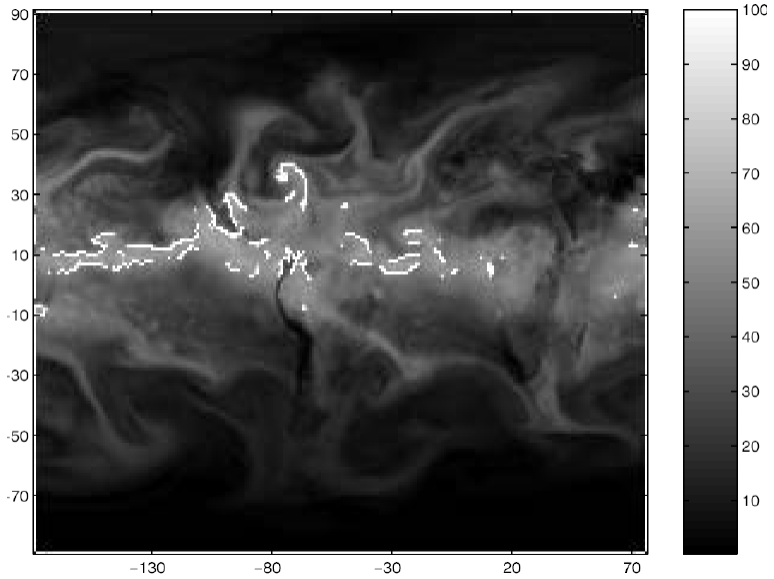


Verified ST-Outliers (those circled with dashed lines are not ST-Outliers)

Hurricanes, Floods and Rainfall Anomalies



The areas indicated in the black boxes were affected by large fires that occurred during the 1998 and 1999 fire seasons (Sicily Island) [Lasaponara, 2006]

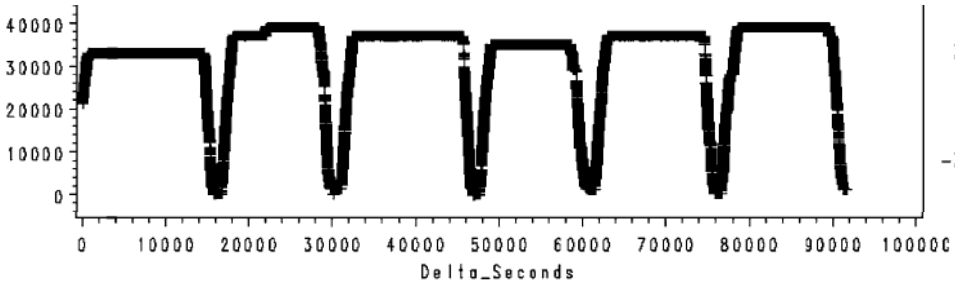


Detected boundary of Hurricane Isabel using Wavelet power distribution of Global water vapor at 18 PM, Sep. 18, 2003 [Lu and Liang, 2004]

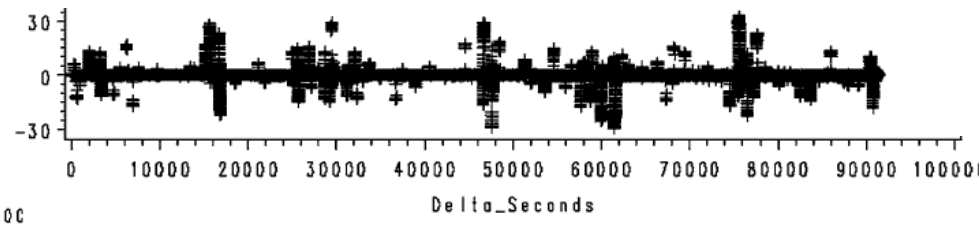
- Anomalies in Australian seasonal rainfall [Drosdowsky, 1993] using RPCA

Aircraft Time Series Data

- Two different signals obtained from a flight data recorder (FDR): altitude of the aircraft and roll angle have been used to find anomalies [Basu and Meckesheimer, 2007]
- Outliers
  - During cruise a change in altitude similar to one that is normally observed during take-off and landing, may indicate an outlier.
  - Roll angles of ±100 degrees are not realistic.



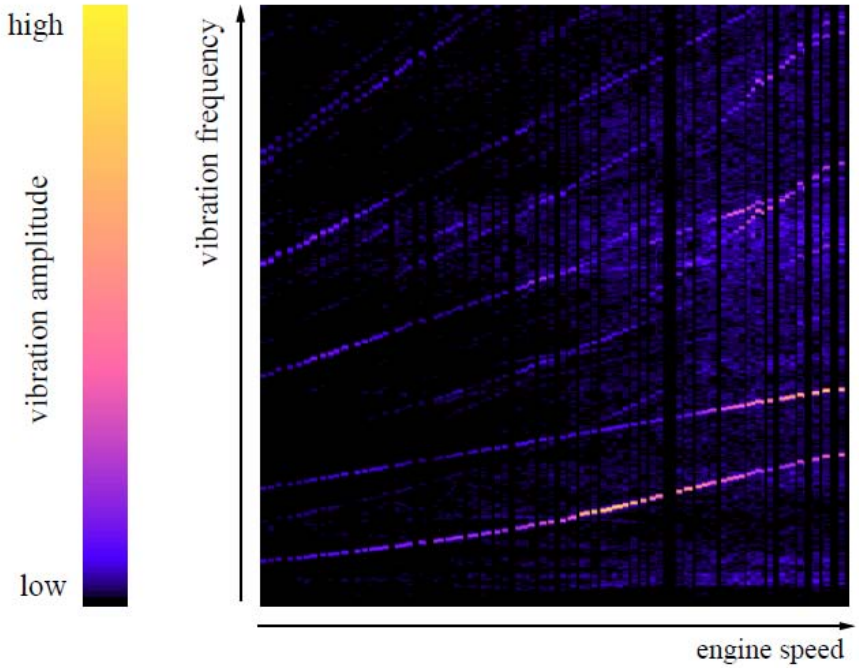
Altitude time series signal



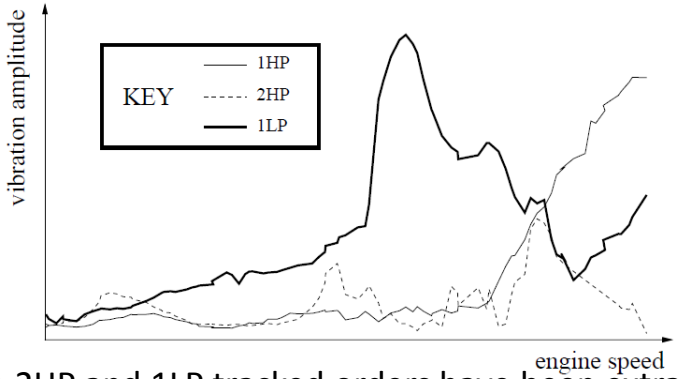
Roll Angle time series signal

Jet Engine Vibration Data

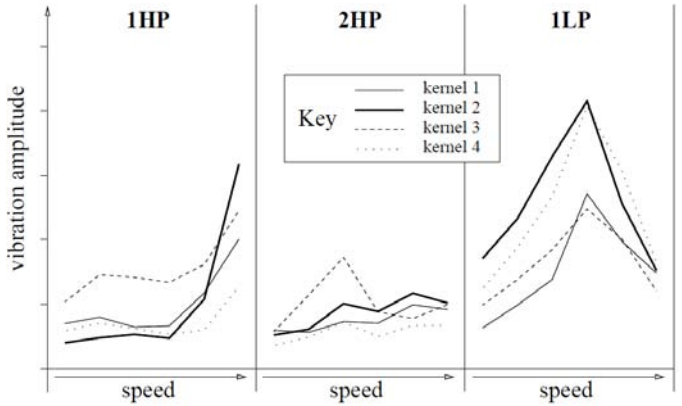
- Anomalies in jet engines have been discovered by analyzing the high/low pressure shaft harmonic frequencies in jet engine vibration data [Nairac et al., 1999]



The Zmod plot for the acceleration phase of a vibration pass-off test. The straight lines visible on the Zmod plot correspond to the main harmonics of the rotation frequency of the high pressure (HP) shaft, while the curved lines correspond to the low pressure (LP) shaft.



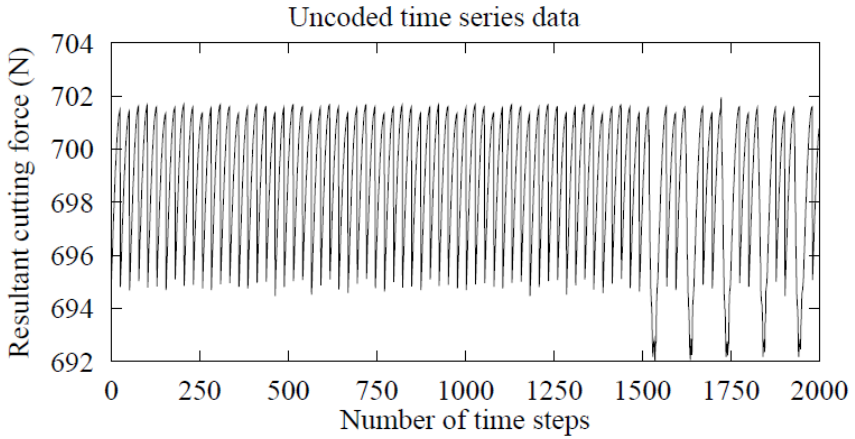
The 1HP, 2HP and 1LP tracked orders have been extracted from the Zmod plot



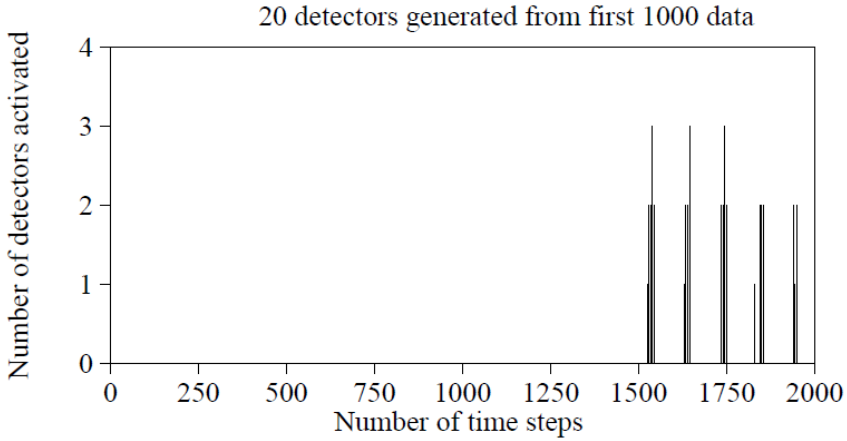
The 4 kernel centres

Tool Breakage Detection

- Anomalies like tool breakage detection have been discovered using cutting force data [Dasgupta and Forrest, 1996]
- Milling industries need online monitoring of tool conditions
- Cutting parameters include temperature, cutting force, torque, vibration, acoustic emission, motor current
- Normally, cutting force periodically varies with tooth frequency which depends on spindle speed

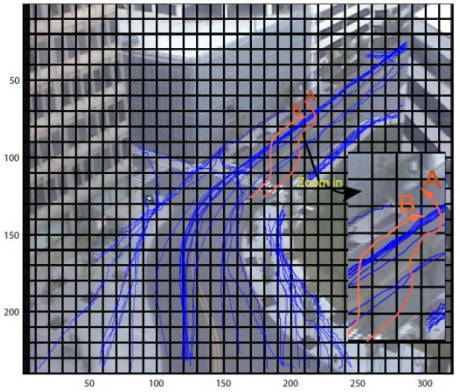


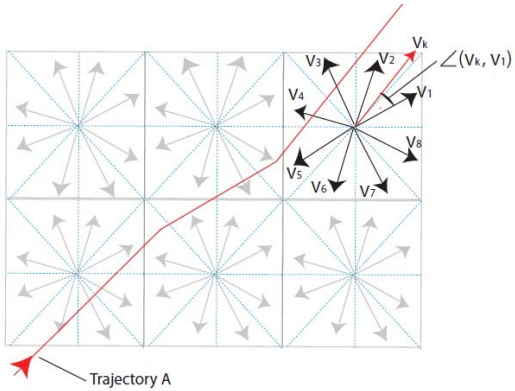
One tooth is broken after 1500 time steps

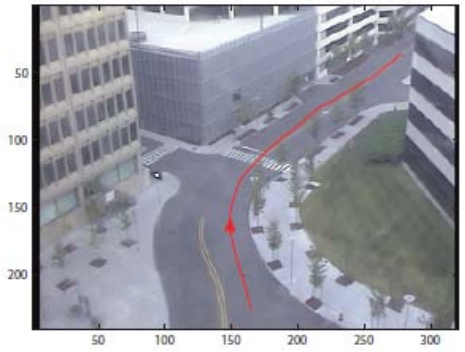


Trajectory Outliers

- [Li et al., 2009] discover anomalies from average daily speed and average daily load dataset for taxicabs in San Francisco during July 2006.
- [Ge et al., 2010] find anomalies like vehicle moving to the left side of the road (wrong way), vehicle taking a short cut to drive to the wrong way, people crossing the street illegally.



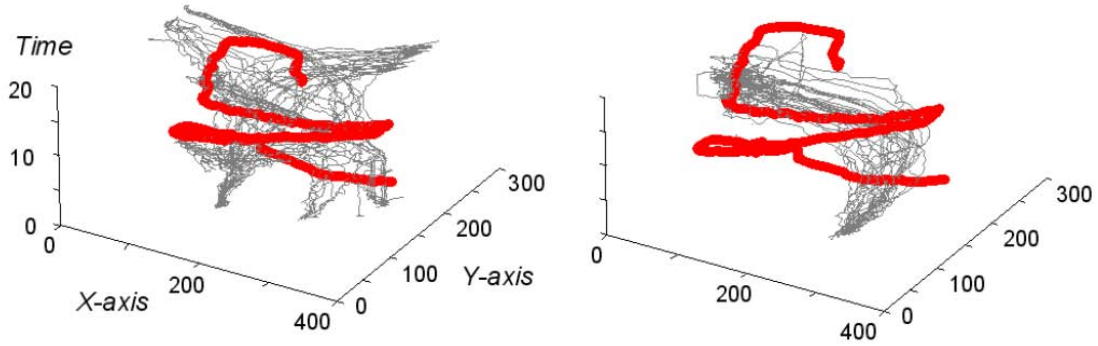




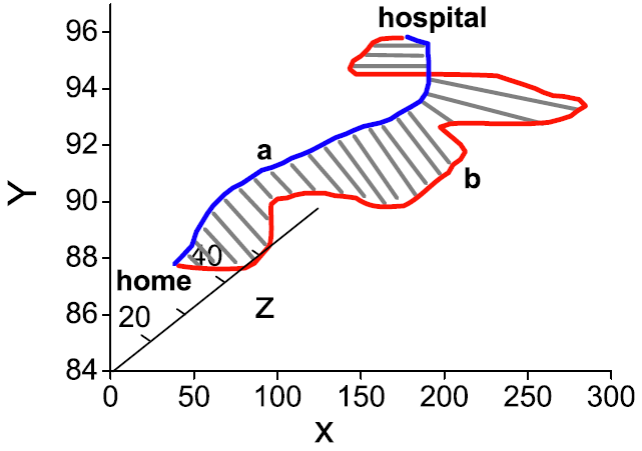
gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

Trajectory Discords

- [Yankov et al., 2008] also find discords from trajectory data
- Trajectory data can also be used for monitoring the movement of patients and senior citizens (for example, to discover events such as taking a wrong bus, having a bad fall, encountering a sudden slow-down and getting lost [Bu et al., 2009]
- Surveillance data can be useful in smart homes to discover anomaly situations like the resident turned on the bathwater, but has not turned it off before going to bed [Jakkula and Cook, 2008]



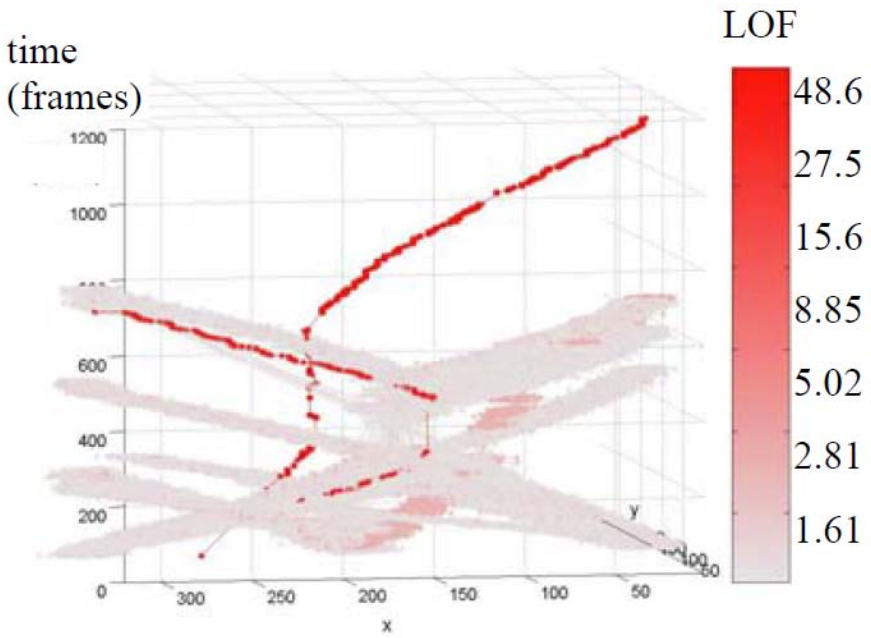
The number one discord found in a trajectory data (bold line) with 50 trajectories



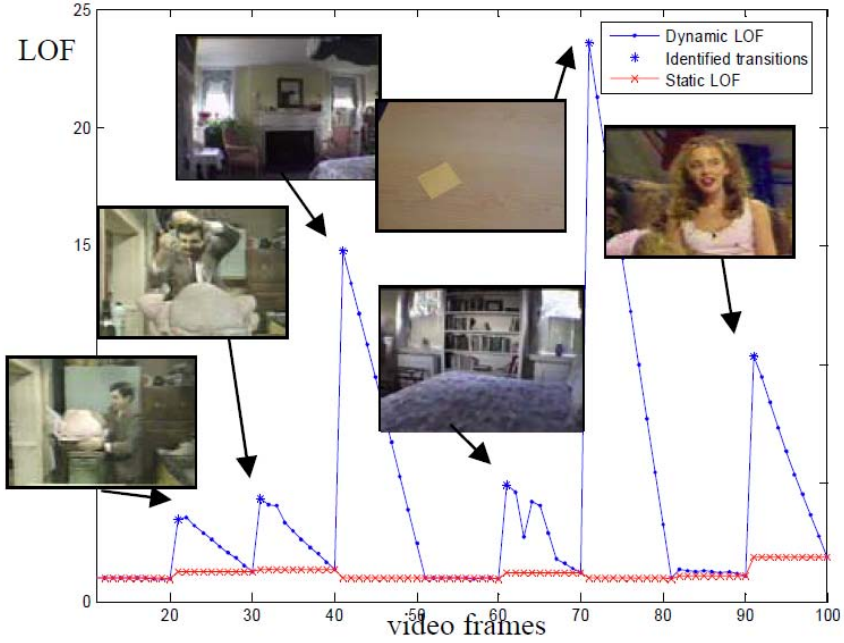
An Elder Monitoring Application

Incremental LOF based Trajectory Discords

- Surveillance videos can be explored to discover outliers like appearance of a new object, zooming objects to camera and novel video content [Pokrajac et al., 2007]
- They study trajectories from surveillance videos to identify anomalous behavior like person walking right and then back left and person walking very slowly



Unusual trajectories in motion videos



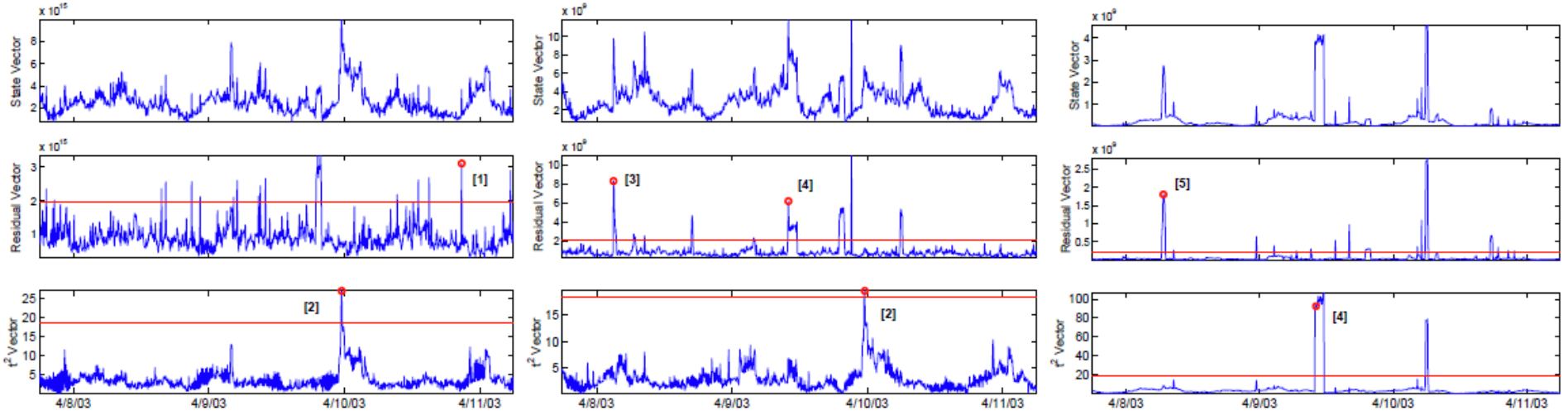
Outliers: appearance of a new object (frame 21), zooming objects to camera (frame 31) & novel video content (frames 41, 61, 71, 91)

Intrusion Detection

- Techniques for outlier detection from temporal data have been widely used for intrusion detection [Angiulli and Fassetti, 2007; Hofmeyr et al., 1998; Lane and Brodley, 1997; Lane and Brodley, 1998; Sequeira and Zaki, 2002; Warrender et al., 1999]
- [Ye, 2000] use the audit data of a Sun Solaris system from MIT Lincoln lab to detect intrusion scenarios like password guessing, using symbolic links to gain root privileges, attempts to gain an unauthorized access remotely, etc.

Origin-destination flow Anomalies

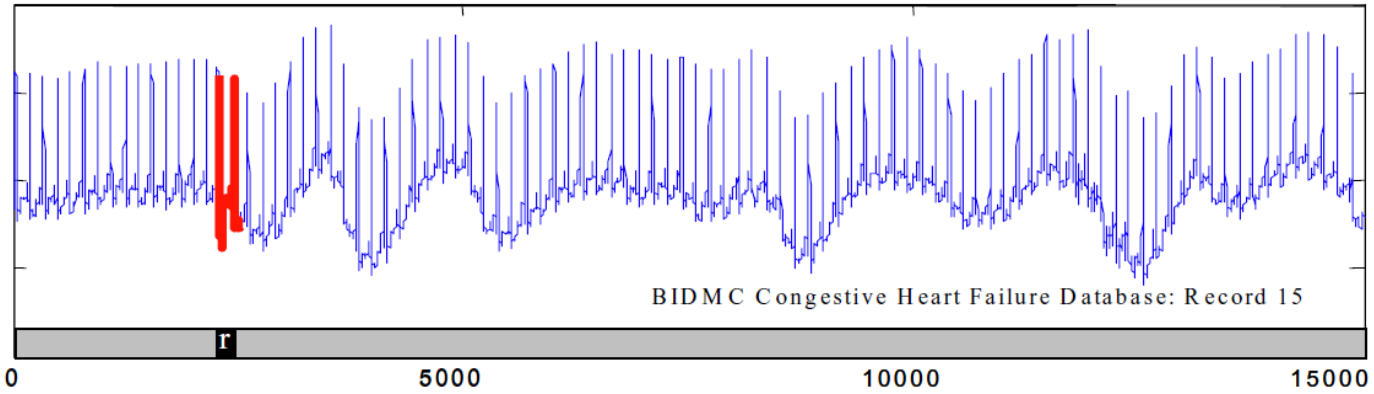
[Lakhina et al., 2004] discover anomalies like High rate point to point byte transfer, denial of service (DOS), distributed denial of service (DDOS) attacks, Flash crowd (large demand for a resource/service), scanning a host for a vulnerable port or scanning network for a target port, WORM, outage events, etc.



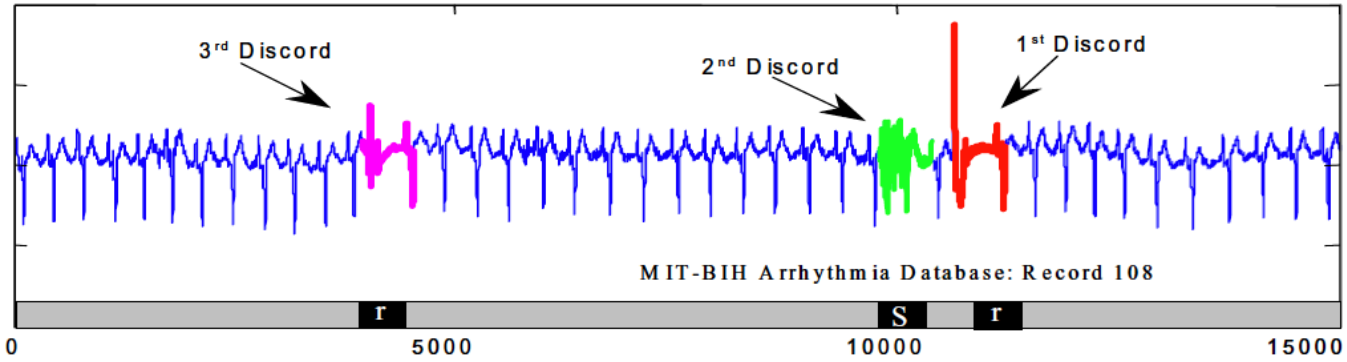
(a) # of Bytes (b) # of Packets (c) # of IP-Flows

An illustration of the PCA method on the three types of OD flow traffic. Top row: timeseries of state vector squared magnitude ($\|x\|^2$); middle row: timeseries of the PCA residual vector squared magnitude ($\|\hat{x}\|^2$); bottom row: t^2 vector

Electrocardiogram Anomalies



An ECG that has been annotated by a cardiologist as containing one premature ventricular contraction [Keogh et al., 2005]

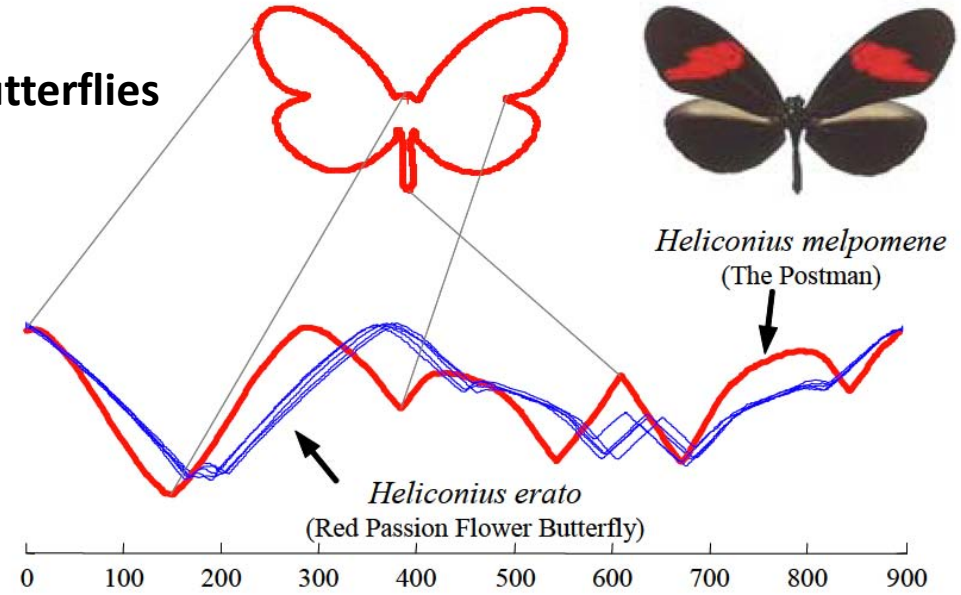


An excerpt of an ECG that has been annotated by a cardiologist as containing 3 various anomalies [Keogh et al., 2005]

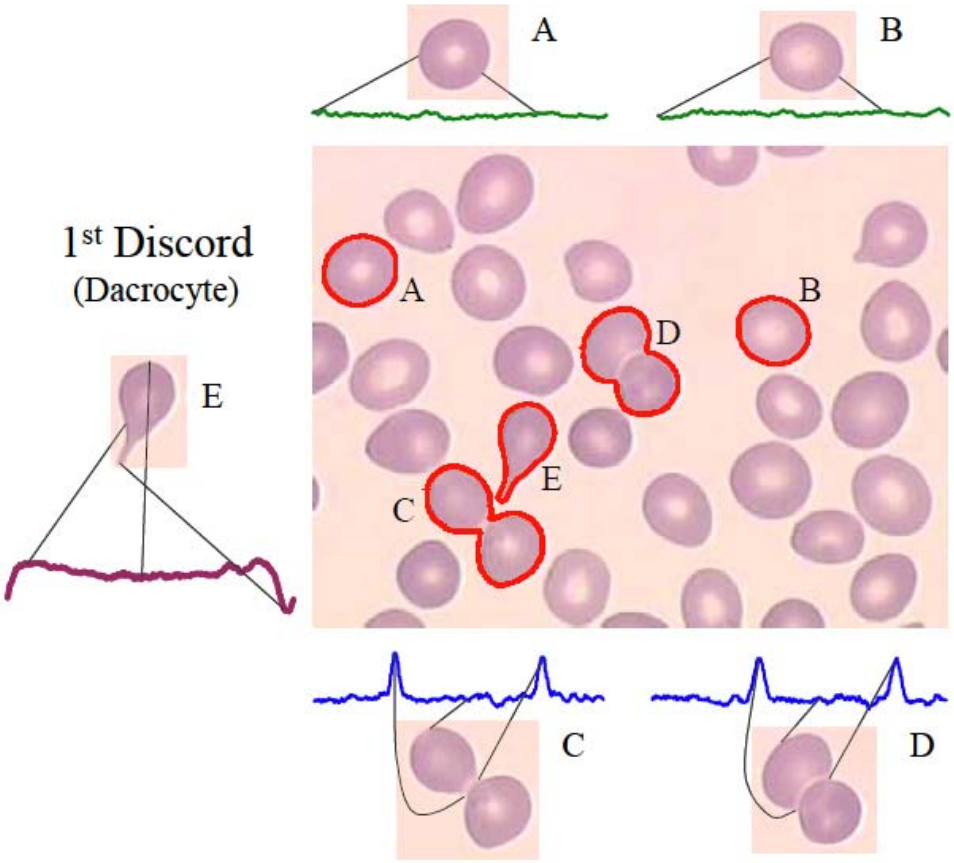
Unusual Butterfly

Six examples of butterflies, ostensibly all Heliconius erato (Red Passion Flower) Butterflies [Wei et al., 2006]

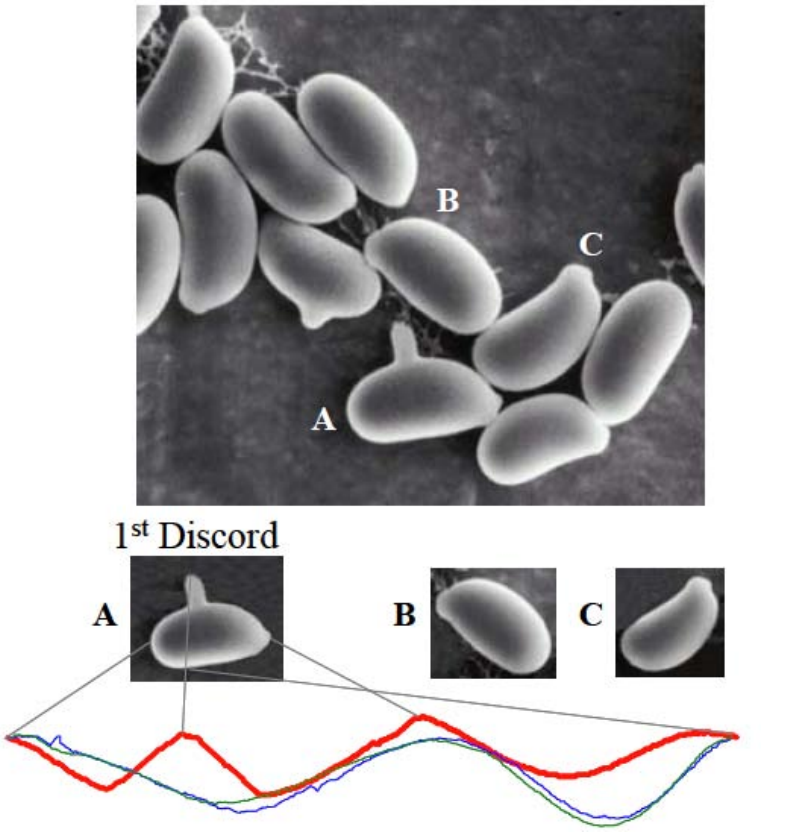
An example of Heliconius melpomene and not Heliconius erato [Wei et al., 2006]



Shape Discords [Wei et al., 2006]



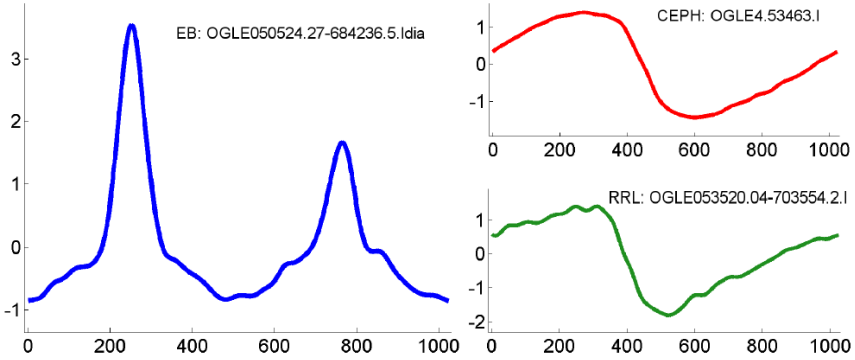
Wei et al. discover teardrop shaped cell, or dacrocyte, which is indicative of several blood disorders



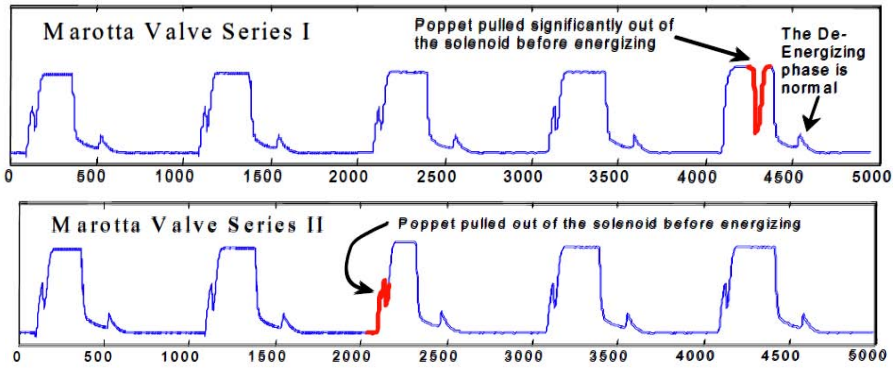
Some spores produced by a fungus. One spore is different because it has a germ tube (Image by Charles Mims, University of Georgia)

Space Telemetry, Star Light, Gamma Rays

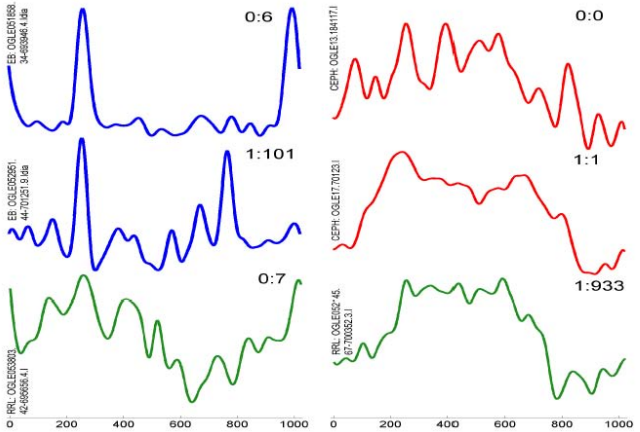
- [Keogh et al., 2005] discover discords from Space telemetry data (Space Shuttle Marotta Valve time series)
- [Yankov et al., 2008] find discords from star light-curve data consisting of light-curves produced by three classes of star objects: Eclipsed Binaries, Cepheids and RR Lyrae variables
- [Zhu and Shasha, 2003] find high gamma ray bursts from Milagro Gamma Ray data stream which consists of time series of the number of photons observed (events) every 0.1 second



Typical examples from 3 classes of lightcurves: Left) Eclipsed Binary, Right Top) Cepheid, Bottom) RR Lyrae



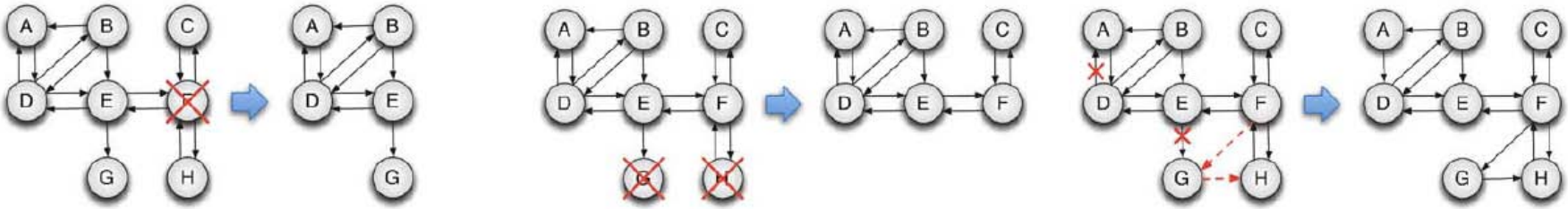
Examples of an annotated Marotta Valve time series [Keogh et al., 2005]



Top light-curve discords in each class

Outlier Web Crawl Snapshot

- Given multiple crawls of the web graph, [Papadimitriou et al., 2008; Papadimitriou et al., 2010] find a crawl graph with anomalies.
- These anomalies refer either to failures of web hosts that do not allow the crawler to access their content or to hardware/software problems in the search engine infrastructure that can corrupt parts of the crawled data.



(a) Missing Connected Subgraph. (b) Missing Random Vertices. (c) Connectivity Change.

- Signature Similarity turned out to be most important measure
- [Yankov et al., 2008] study the MSN query logs to discover both anticipated and unanticipated web queries as discords. E.g., “Full moon”

Graph Outliers in Graph Streams

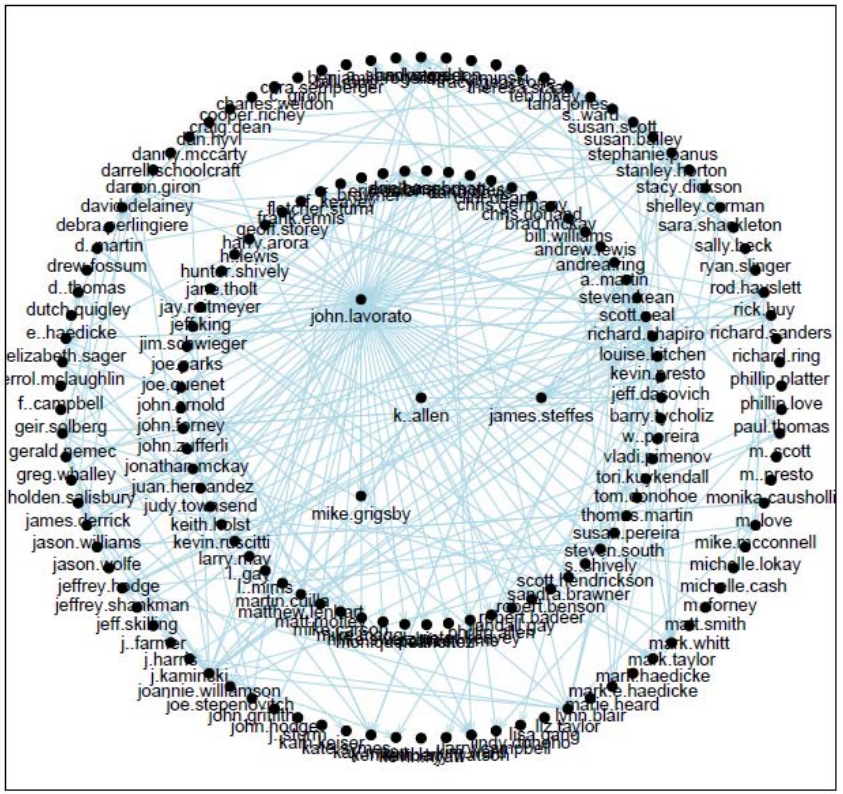
- [Aggarwal et al., 2011] discover graphs representing inter-disciplinary research papers as outliers from the DBLP dataset. They also discover movies with a cast from multiple countries as outliers from the IMDB dataset
- (DBLP) Yihong Gong, Guido Proietti, Christos Faloutsos, Image Indexing and Retrieval Based on Human Perceptual Color Clustering, CVPR 1998: 578-585
  - Yihong Gong: computer vision and multimedia processing
  - Christos Faloutsos: database and data mining
- (DBLP) Natasha Alechina, Mehdi Dastani, Brian Logan, John-Jules Ch Meyer, A Logic of Agent Programs, AAAI 2007: 795-800
  - Natasha Alechina: United Kingdom
  - John-Jules Ch Meyer: Netherlands
- (IMDB) *Movie Title:* Cradle 2 the Grave (2003)
  - Jet Li: Chinese actor
  - DMX (I): American actor

Community Trend Outliers

- [Gupta et al., 2012b] discover authors outliers from DBLP co-authorship network such that they show a change in their research areas quite different from other authors.
- DBLP: Georgios B. Giannakis
  - X_1 conferences: CISS, ICC, GLOBECOM, INFOCOM
  - X_2 conferences: ICASSP, ICRA
- They also discover outlier actors from IMDB such that they show a very unusual change in their movie genre distribution
- IMDB: Kelly Carlson (I)
  - Changed community from Sports, Thriller, Action to Drama and Music

Outliers based on Scan Statistics

- [Priebe et al., 2005] study the communication graph of the Enron data with respect to the maximum degree and digraph size scan statistics. Excessive communication outliers which can suggest insider trading scenarios are discovered.
- Using "outdegree at level 2" as scan statistic, k.allen turns out as outlier
- To discover excessive chatter based outliers, they propose a composite 2nd order scan statistic which ensures
  - Some minimum level of recent activity
  - Order 0 and order 1 scan statistics do not yield detections



Distribution Change Outliers [Gupta et al., 2012a]

- Unusual GDP Distribution Change in 1985 for Uganda

National Resistance Army (NRA) came to power and made economic reforms (1985-86)

Uganda, Canada, France, Iceland

Consumption	Net Exports
45%	45%

Uganda

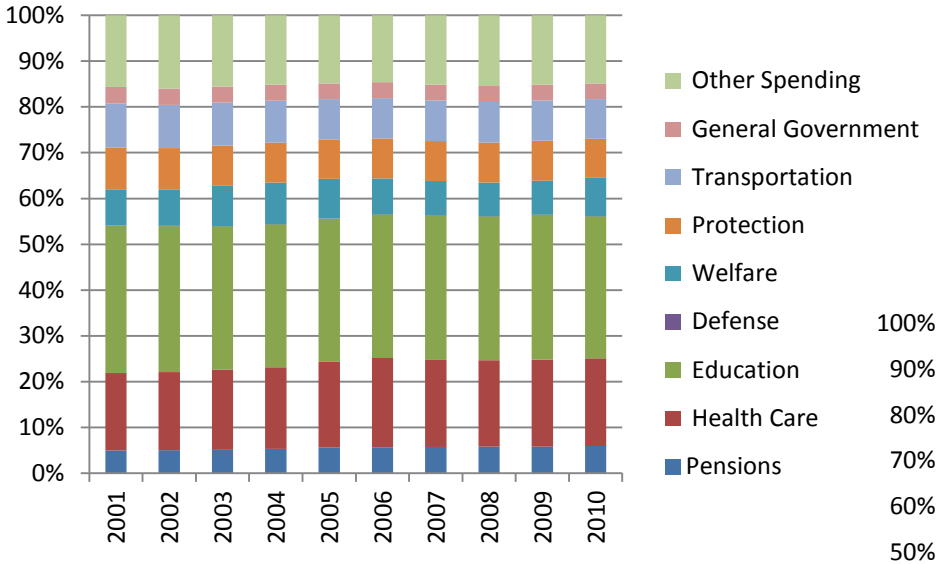
Consumption	Net Exports
80%	1-2%

Canada, France, Iceland

Consumption	Net Exports
45%	45%

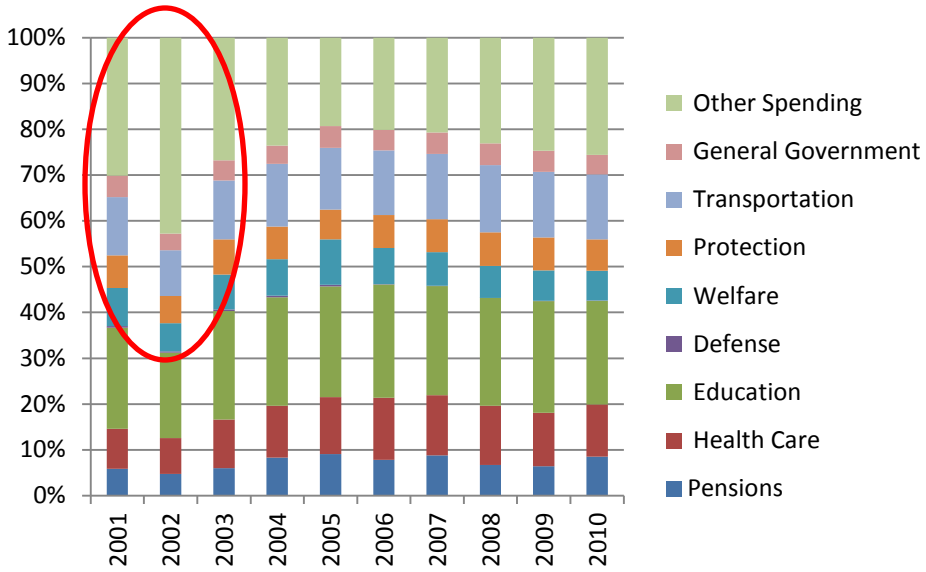
Distribution Change Outliers [Gupta et al., 2012a]

- Unusual Budget Distribution for the past decade for Arkansas



Average trend of 5 states with distributions close to that of AK for 2004-2009

Distributions of Budget Spending for AK



gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu

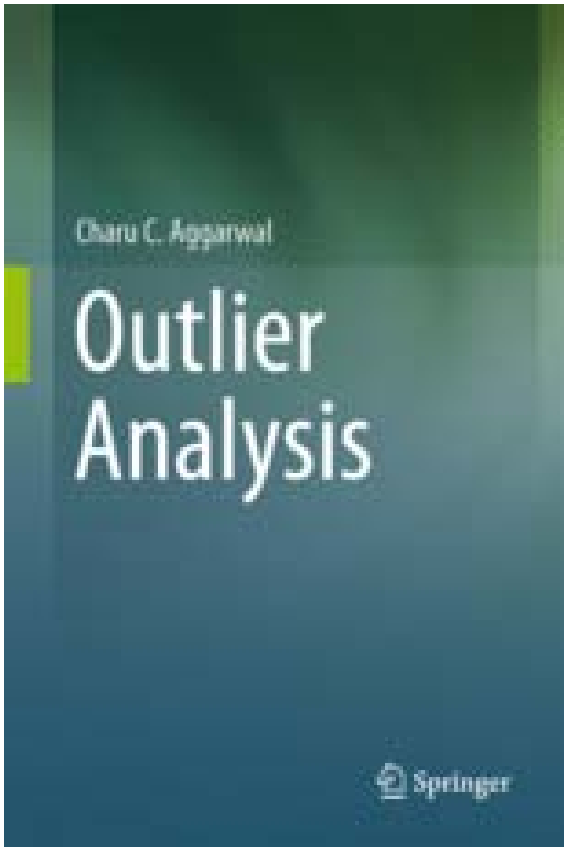
Outliers in Mixed Attribute Space

- [Otey et al., 2006]
- US Census Bureau's Income data set
  - A 39 year old self-employed male with a doctorate, working in a clerical position making less than 50,000 dollars per year
  - A 42 year old self-employed male from Iran, with a bachelors degree, working in an Executive position making less than 50,000 dollars per year
  - A 38 year old Canadian female with an Asian Pacific Islander origin working for the US Federal Government for 57 hrs per week and making more than 50,000 dollars per year
- US Congressional Voting Data
  - A Republican congressman who voted significantly differently from his party on four bills

Summary

- We presented an organized overview of the various techniques proposed for outlier detection on temporal data
- Specifically, we discussed techniques for
  - Time series data
  - Data streams
  - Distributed data streams
  - Network data
  - Spatio-temporal data
- For each of these forms of data, we presented various outlier definitions proposed in the literature and introduced in brief corresponding techniques
- Finally, we also discussed various applications for which these techniques have been successfully used
- http://dais.cs.uiuc.edu/manish/pub/gupta12_temporalOutlierDetectionSurvey.pdf

Further Reading



- Outlier Analysis (Springer) Authored by Charu Aggarwal, January 2013
- Time Series Outliers
  - Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection for Discrete Sequences: A Survey. IEEE Transactions on Knowledge and Data Engineering (TKDE), 24(5):823–839, May 2012
- Novelty Detection
  - Markos Markou and Sameer Singh. Novelty Detection: A Review - Part 1: Statistical Approaches. Signal Processing, 83(12):2481–2497, 2003
  - Markos Markou and Sameer Singh. Novelty Detection: A Review - Part 2: Neural Network Based Approaches. Signal Processing, 83(12):2499–2521, 2003
- http://dais.cs.uiuc.edu/manish/pub/gupta12_temporalOutlierDetectionSurvey.pdf

KDD 2013 Workshop on

Outlier Detection and Description

August 11th, Chicago

Topics of Interest:

- **Outlier Detection:**
  WHICH objects are highly deviating?
- **Outlier Description:**
  WHY are these objects deviating?

Confirmed Keynote Speakers:

- **Charu Aggarwal**
  (IBM T. J. Watson Research Center)
  Keynote: "Outlier Ensembles"
- **Raymond Ng**
  (University of British Columbia)
  Keynote: Title TBA

Important Date:

- **Submission deadline: May 28, 2013**

http://outlier-analytics.org/odd13kdd/

Thanks!

gupta58@illinois.edu, jing@buffalo.edu, charu@us.ibm.com, hanj@cs.uiuc.edu