# The Outlier Description Problem – A Combinatorial Optimization Perspective

Ian Davidson
University of California – Davis
EU IAS Fellow - Collegium Lyon

Acknowledgements
NSF 1422218 – Functional Network Discovery for Brains,
ONR N000141812485 – Deep Graph Models

Past work with my students Xiang Wang, Tom Kuo and future work with S.S. Ravi

# Outline

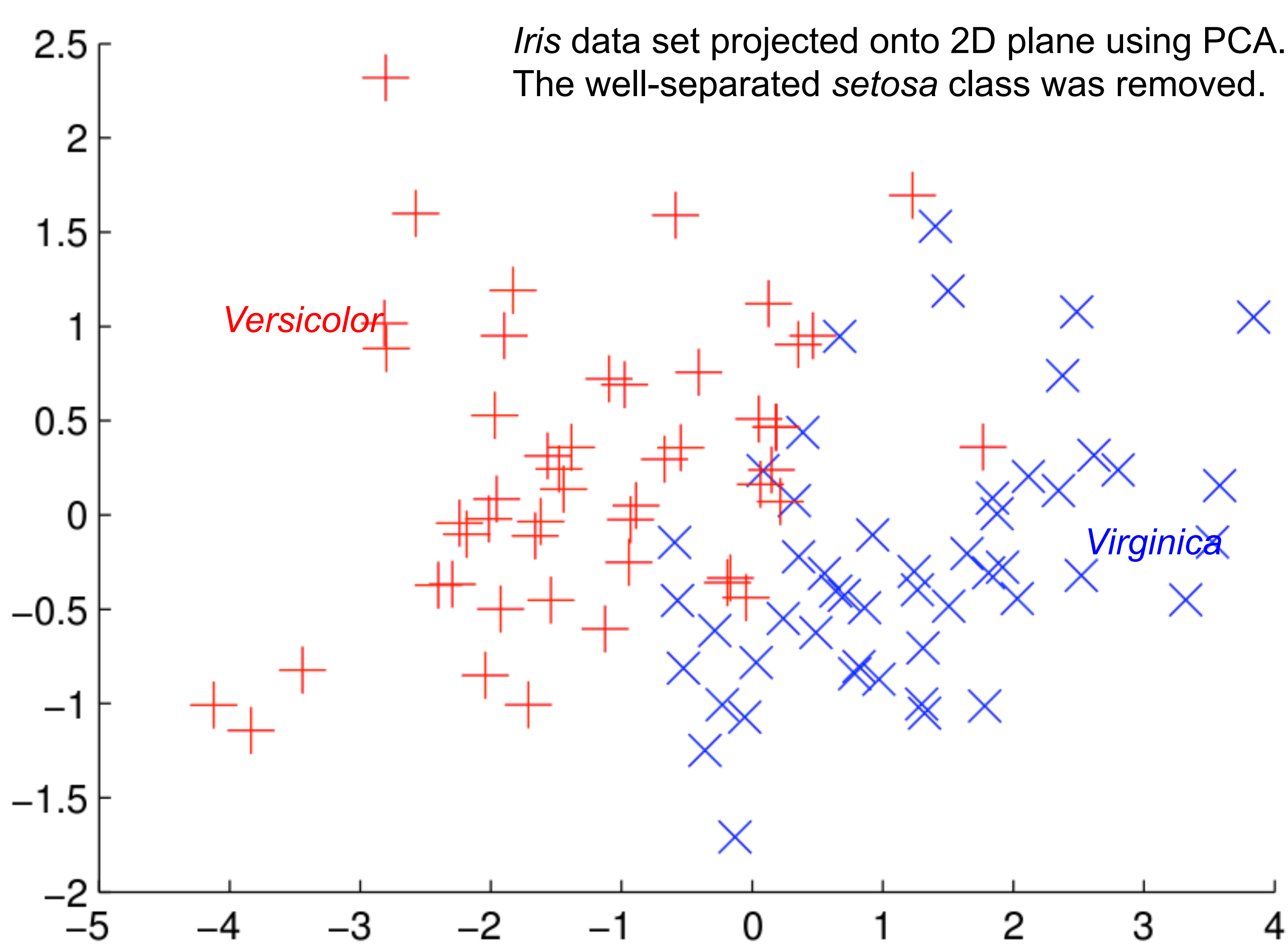
Most Existing Work - Kuo, Chia-Tung, and Ian Davidson. "A Framework for Outlier Description Using Constraint Programming." *AAAI*. 2016. and this year's IJCAI and last year's AAAI papers

- Description and Motivation - An unhappy start
- Work in Progress: Set Coverage Formulation
  - Complexity results, ILP formulations
- AAAI 16: Density Based Formulation:
  - **Constraint programming** formulations
  - Applications in Neuroscience
- Future work
  - Scalability

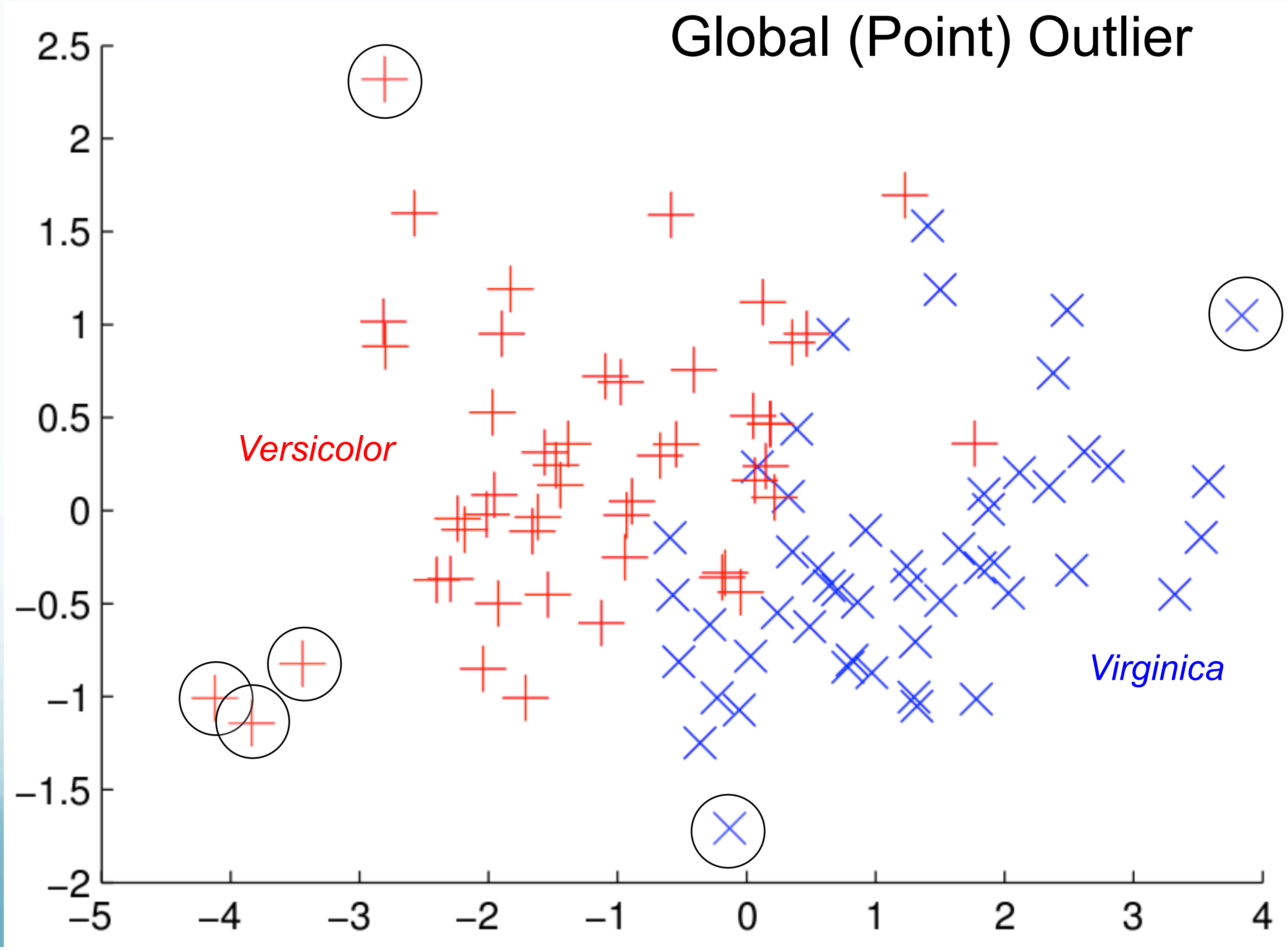
# My Motivation

- Outlier **detection** is well studied
  - Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys, 2009: 5000+ citations
  - Graph based anomaly detection and description: a survey, L Akoglu, H Tong, D Koutra DMKD 2015
  - Group Deviation Detection Methods: A Survey, E Toth, S Chawla ACM Computing Surveys
- A few years ago I was invited to give a talk in Silicon Valley about "Explanation and Outliers". The response was not what I expected …

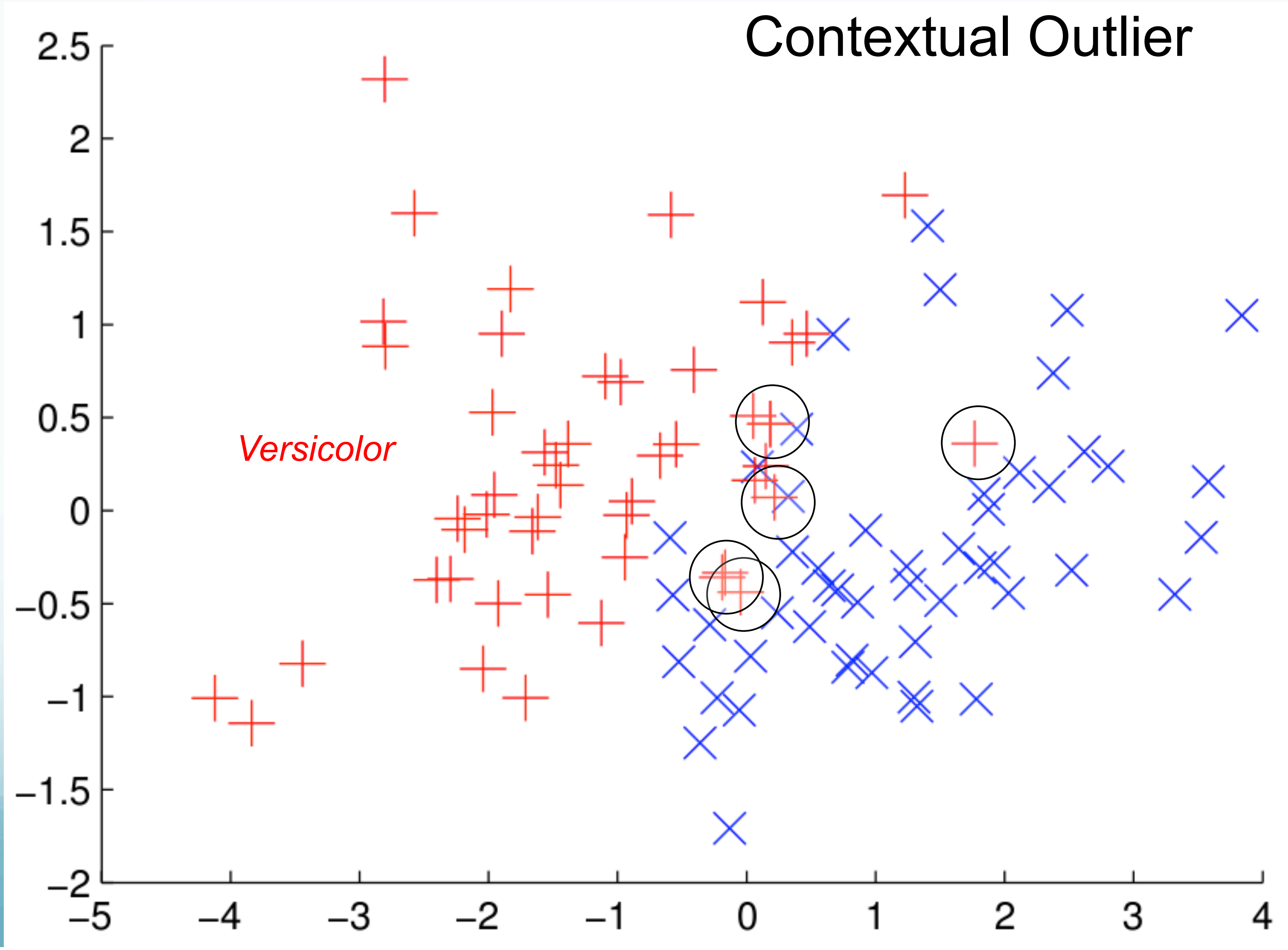
# Contextual Outlier Detection

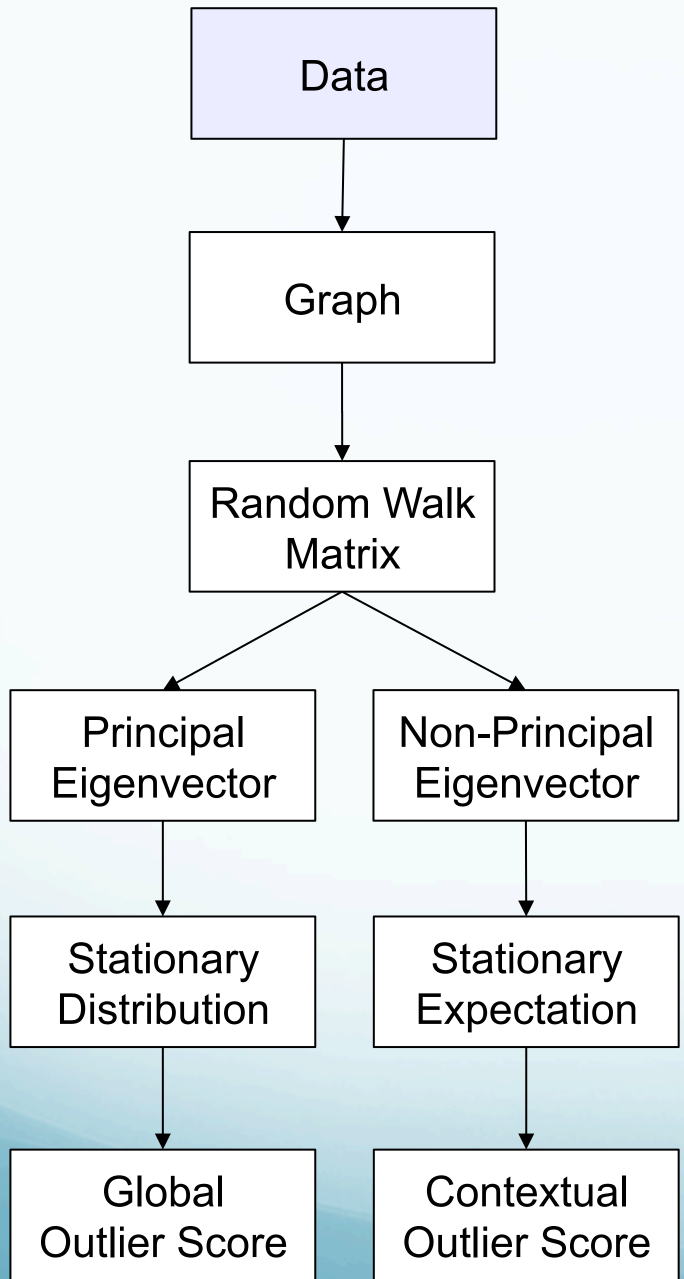


# Contextual Outlier Detection



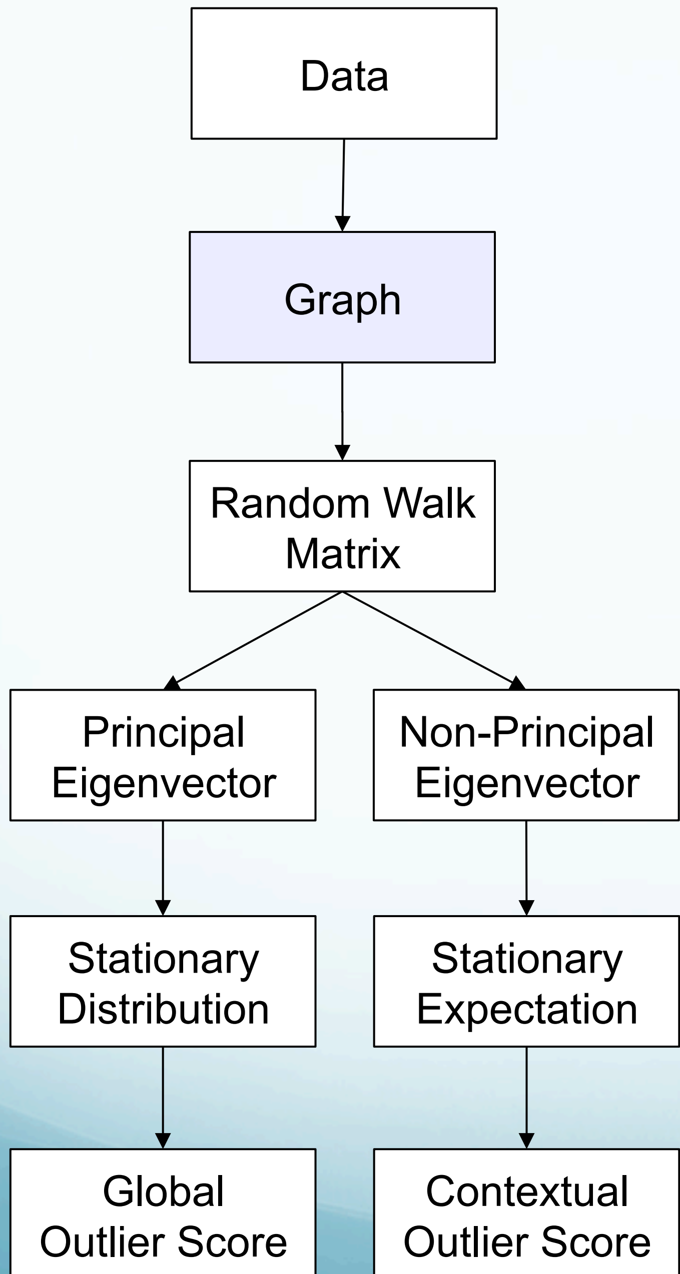
# Contextual Outlier Detection

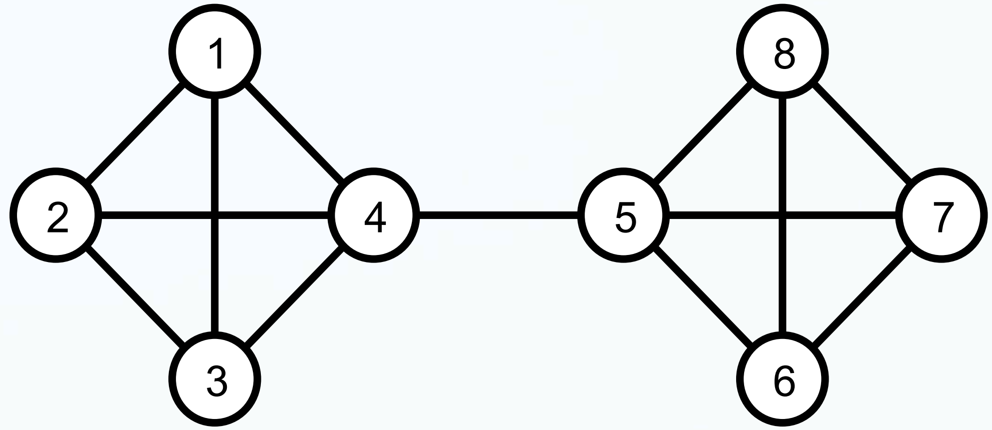


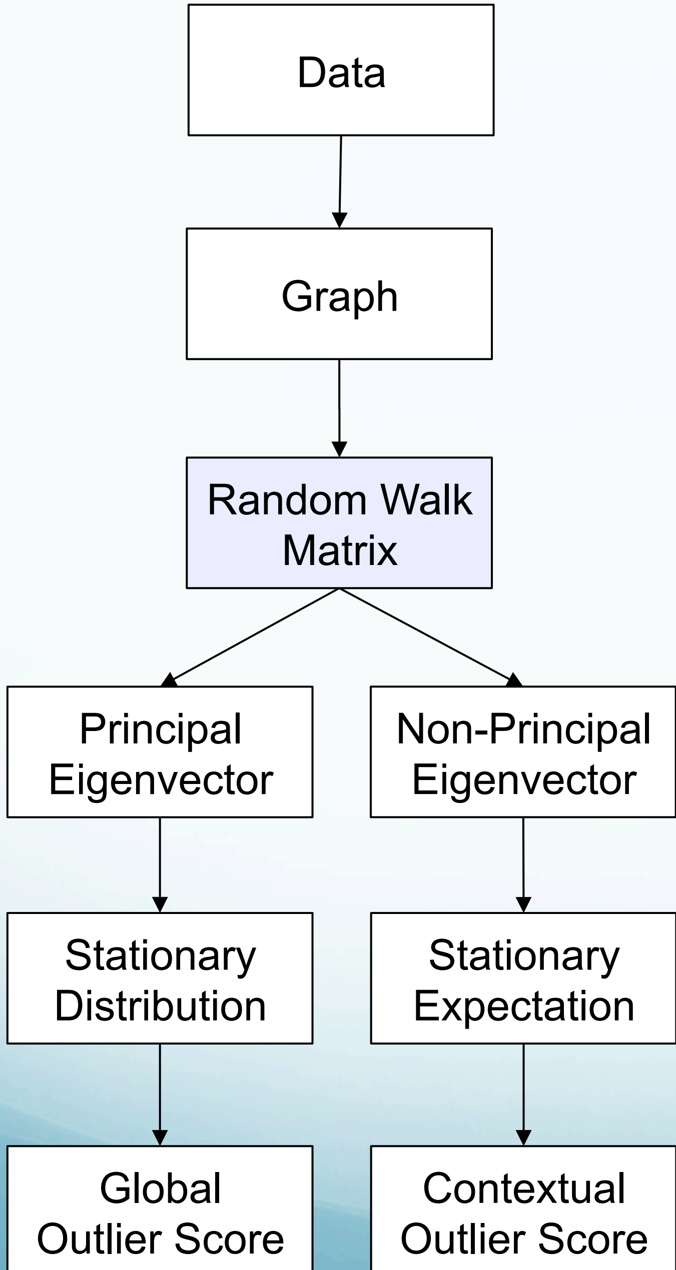


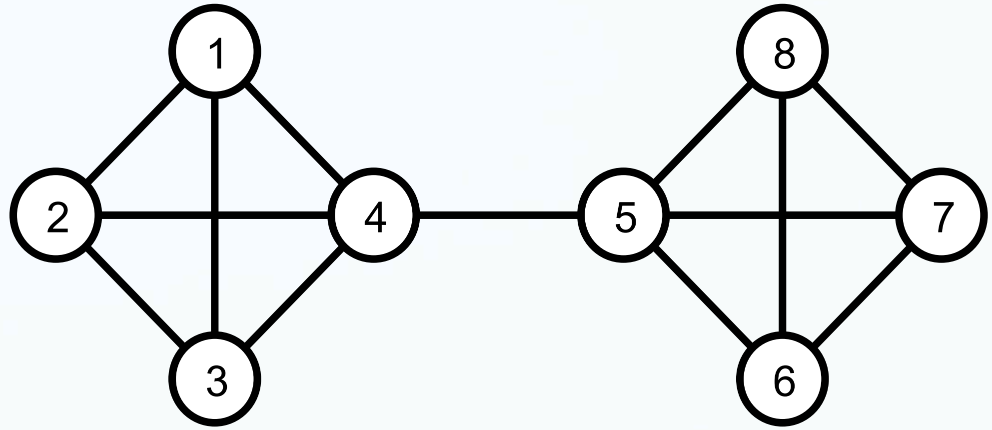

- Challenge
  - **Simultaneously** identify contexts and contextual outliers therein

Wang, Xiang, and Ian Davidson. "Discovering contexts and contextual outliers using random walks in graphs." *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference*. IEEE, 2009
Kind of followup work at KDD 10, DMKD 14

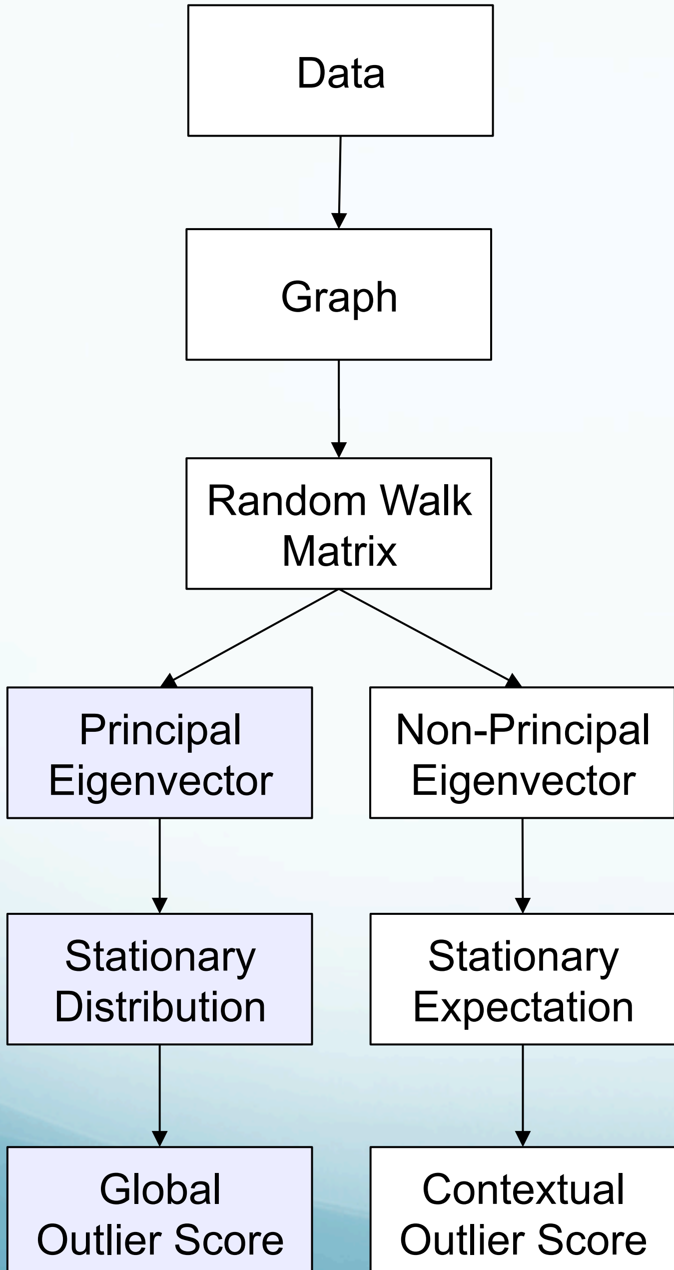

Data
Graph
Random Walk Matrix
Principal Eigenvector
Non-Principal Eigenvector
Stationary Distribution
Stationary Expectation
Global Outlier Score
Contextual Outlier Score

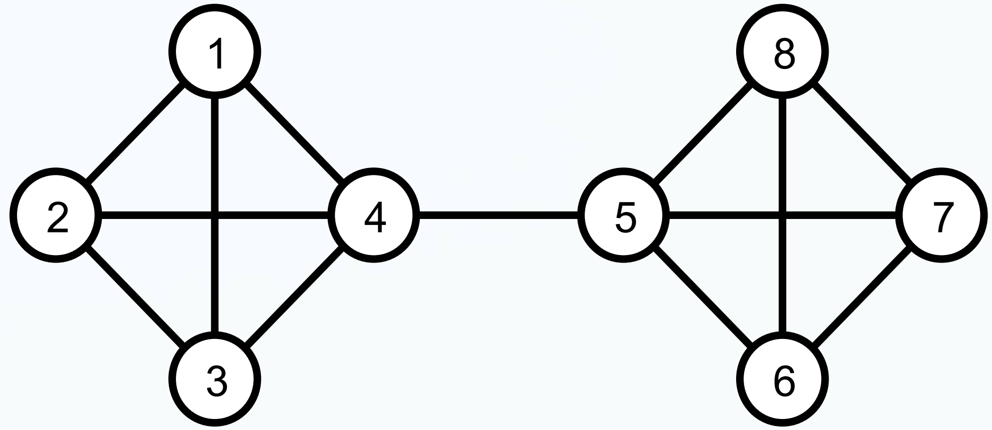

1
2
3
4
5
6
7
8



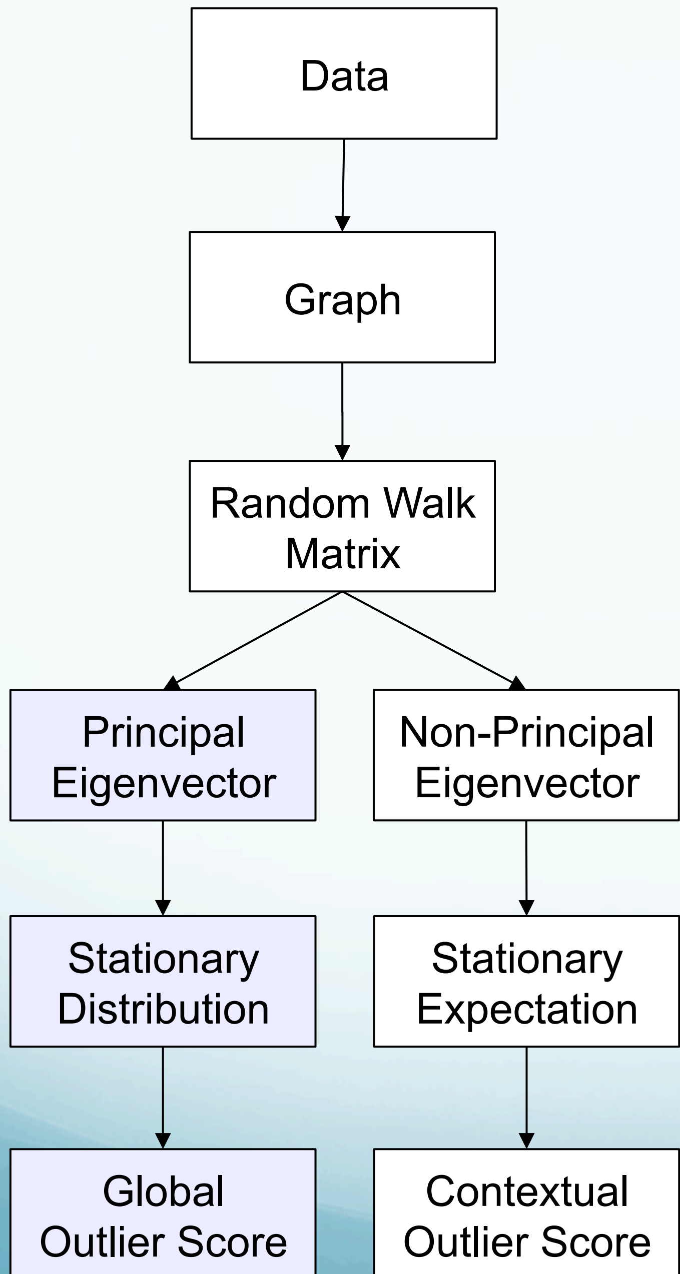

$$W = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 1/3 & 0 \end{pmatrix}$$

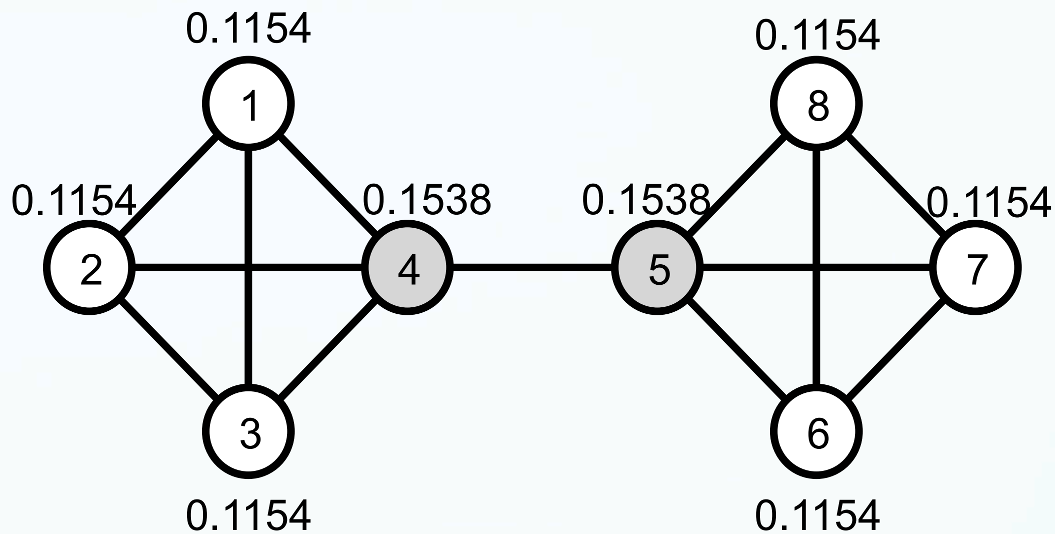



$$W = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 1/3 & 0 \end{pmatrix}$$

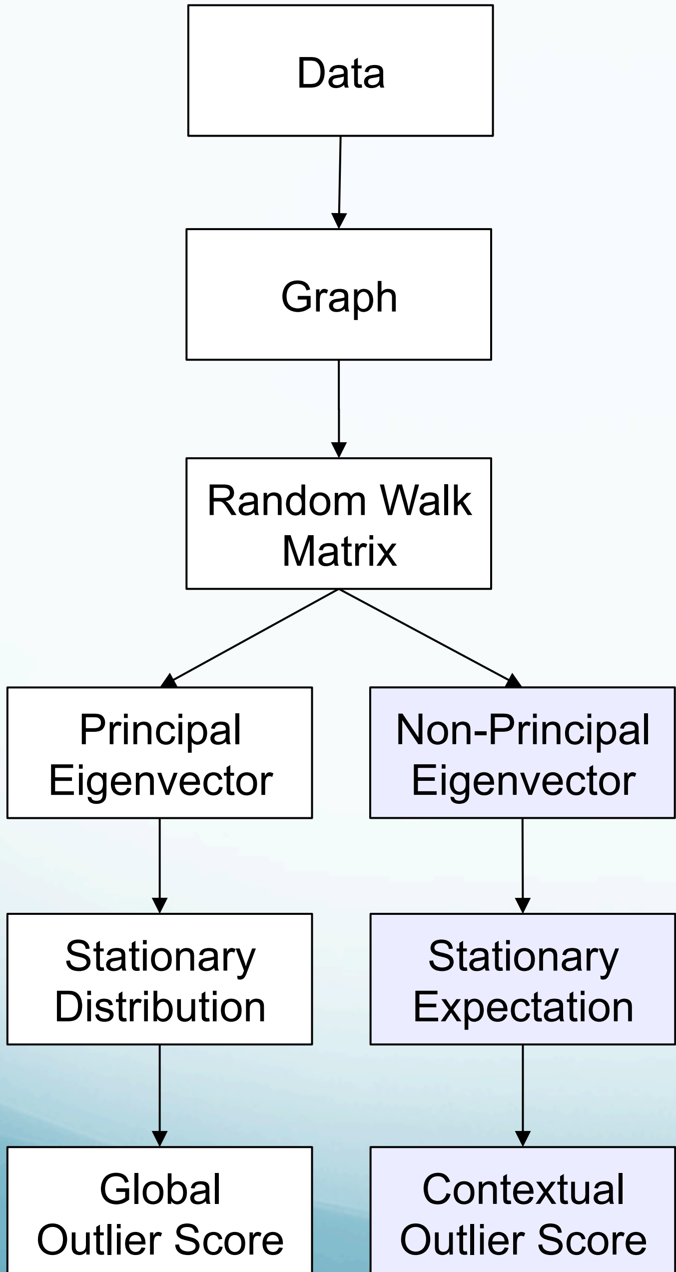
H. D. K. Moonesinghe, Pang-Ning Tan: Outlier Detection Using Random Walks. ICTAI 2006: 532-539

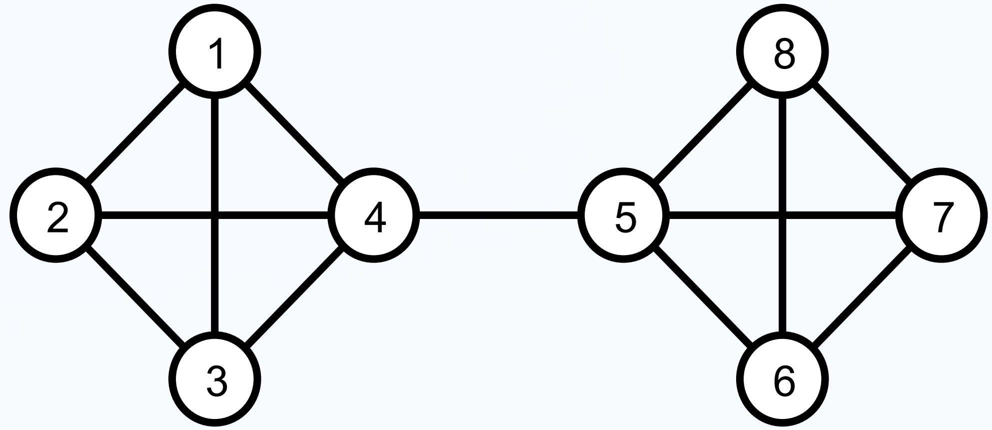



$$W = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 1/3 & 0 \end{pmatrix}$$

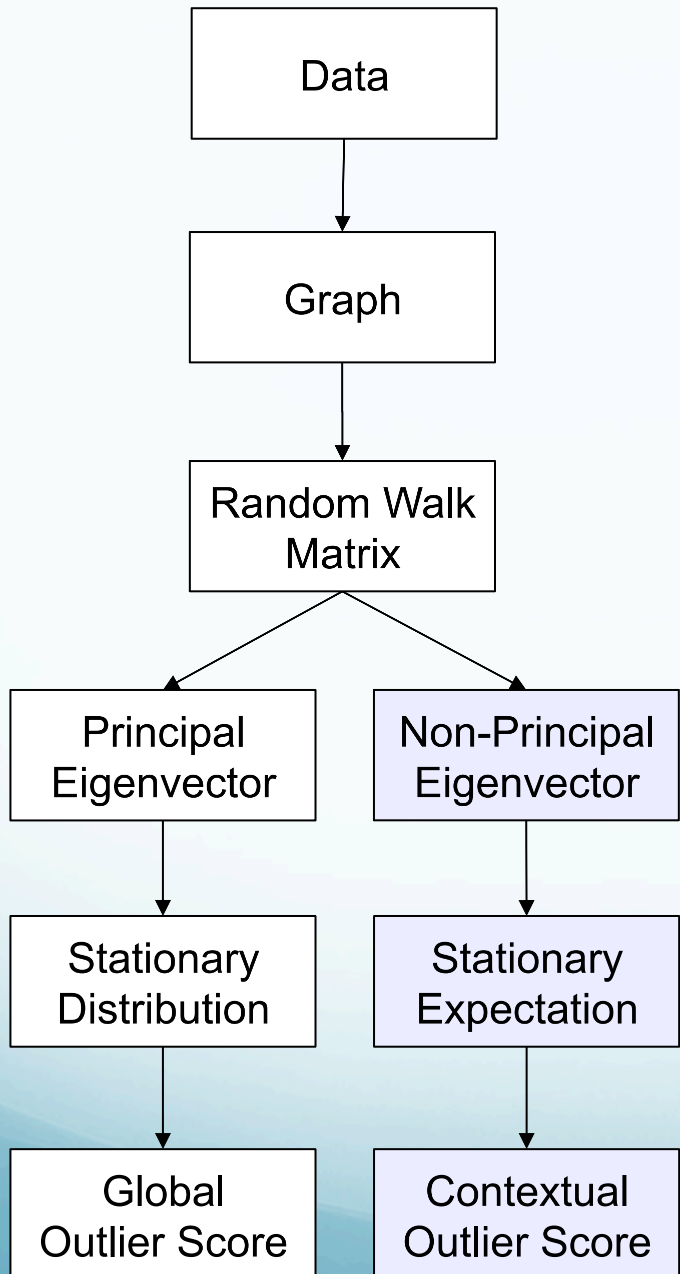
H. D. K. Moonesinghe, Pang-Ning Tan: Outlier Detection Using Random Walks. ICTAI 2006: 532-539

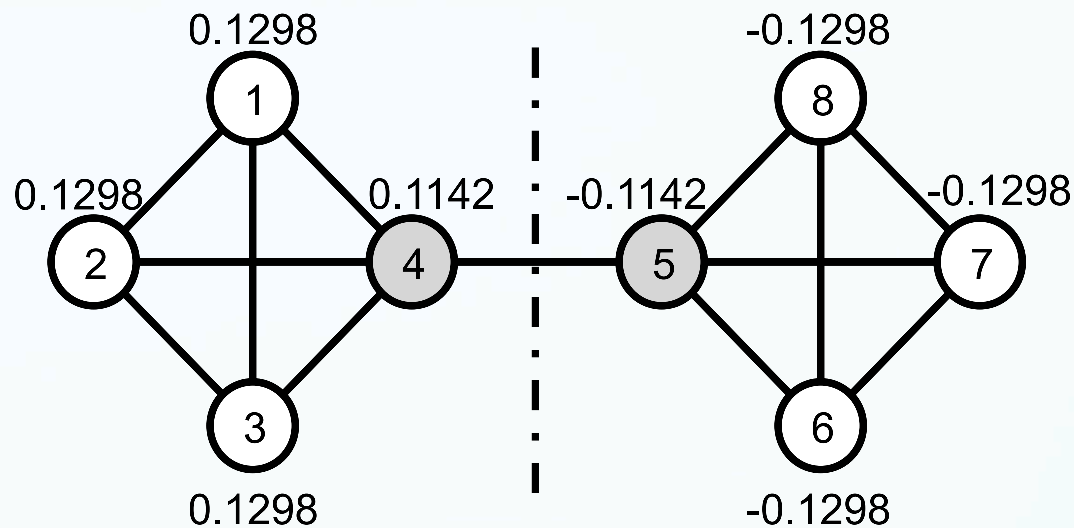



$$W = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 1/3 & 0 \end{pmatrix}$$

Wang, Xiang, and Ian Davidson. "Discovering contexts and contextual outliers using random walks in graphs." *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference*. IEEE, 2009




$$W = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/4 & 1/3 & 1/3 & 0 \end{pmatrix}$$

ICDM paper has lots of nice math, polynomial time algorithms, explanations etc

# Contextual Outlier Score-Stationary Expectations

$$\mathrm{Y}_i^t = \begin{cases} 1 & \mathrm{X}^t = i, \mathrm{X}^0 \in S^+ \\ -1 & \mathrm{X}^t = i, \mathrm{X}^0 \in S^- \\ 0 & \text{otherwise} \end{cases}$$

**Theorem 1** (The Stationary Expectation of a Contextual Random Walk)**.** *If we set* $\mu = (\mu_1, \ldots, \mu_n)^T$ *to be*

$$\mu_i = \frac{\mathbf{v}(i)}{\sum_{j=1}^{n} |\mathbf{v}(j)|}, \forall i, 1 \leq i \leq n, \tag{17}$$

*where* $\mathbf{v}$ *is a non-principal eigenvector of* $W$ *associated*

# "The Core Problem is Often Outlier **Explanation** Not Detecting"

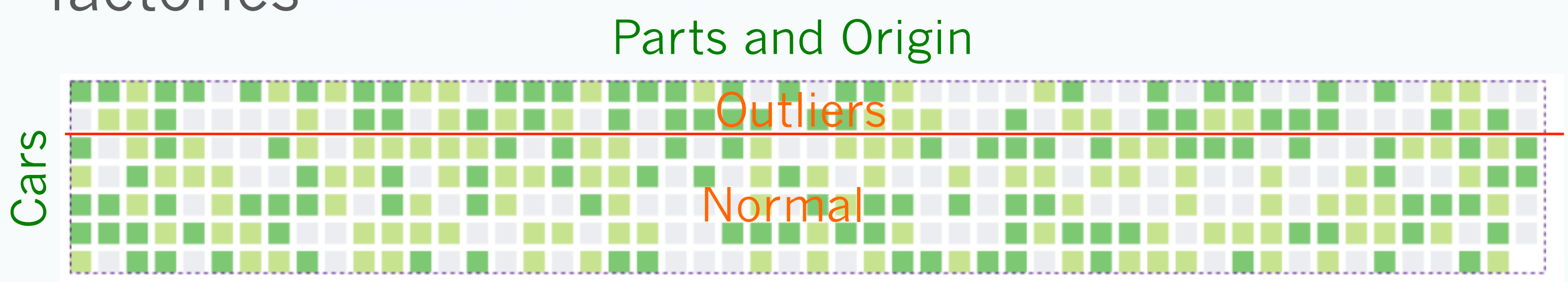
- Paraphrasing: Lots of existing processes, channels and mechanisms to capture unusual behavior. A challenge is how to explain it.
- For example
  - Automobiles identified as "lemons" despite being made of the exact same parts as non-lemons. Why?
    - *Positive explanation:* i.e. The lemons contain transmissions manufactured in France and steering columns manufactured in Belgium
  - Patients identified as demented/gifted (from various cognitive tasks scoring) and we want to identify what parts of their brain's behavior explains this.
    - *Negative explanations:* i.e. The functional connectivity between the hippocampus and pre-frontal cortex is missing for demented individuals

# There Is a Precise Definition To Identify "Lemon" Cars

- 1) Substantial defect
  - A problem covered by the warranty that impairs the car's use, value, or safety, such as faulty brakes or steering. Minor defects such as loose radio knobs and door handles do not meet the legal definition of "substantial defect."
- 2) Reasonable Repair Attempts
  - If the defect is a serious safety defect (for example, involving brakes or steering), it must remain unfixed after one repair attempt.
  - If the defect is not a serious safety defect, it must remain unfixed after three or four repair attempts (the number varies by state).
  - If the vehicle is in the shop a certain number of days (usually 30 days in a one-year period) to fix one or more substantial warranty defects, it may fit the definition of a lemon.
- If criteria is met your entitled fo a full refund

# An Illustrative Automobile Recall Example

- Each car of a given model contains the exact same parts – but they could be manufactured in different factories




- Some combination of part origins associated with continual problems (a lemon)
- Identify the patterns present in outliers but not in the rest of the population or vice versa

# Why Not Use Existing Methods?

- A) Contrast Pattern Mining
  - Outlier collection too small to find frequent patterns in with confidence?
- B) Discriminatory Learning
  - Key requirement is explanability
  - Linear discrimination is unrealistic in lower dimensional space and need to project to HD space. Explanability will be lost
- C) Generative Models
  - Hard to build an accurate generative model for P(x,abnormal)

Our approach: Formulate the problem as discrete combinatorial optimization

# A Very General Formalization of the Problem

- Let $C_N$ and $C_O$ be two disjoint sets of points in some $d$ dimensional space $S$.
- Let $f$ be a mapping from the $d$ to $k$ dimensional space:

$$f : S \rightarrow S'$$

- Let $P$ be some property calculated on each set of points

$$max_f |P(f(C_N)) - P(f(C_O))|$$

- Can be property is common in $N$ and rare in $O$ or vice-versa (application specific – i.e. neuroscience is the former)

# What We'll Cover

- Two frameworks
  - A set coverage formulation (not published based on AAAI17, IJCAI18)
  - A density formulation (published in AAAI 16)
- Both are formulated as constraint/declarative programming optimization formulations

| Property (P) | Mapping (f) |
|---|---|
| Coverage | Binary chosen sub-space |
| Density, Diameter, Connectivity | Binary chosen sub-space |

# Density Based Formulations in Constraint Programming [AAAI16]

- ILP Formulations can be limiting.
- Why Constraint Programming Implementation
  - Same benefits of ILP – discrete optimization
  - Still finds global optima
  - Allows semantically meaningful objectives
  - Solve **multiple** unknowns at once
  - Easy to model variations

# A Brief Introduction to Constraint Programming

## Constraint Satisfaction/Optimization Problem (CSP/COP)

- a set of variables $X$
- a domain $Dom(x)$ of possibles values for each variable $x \in X$
- a set of constraints $C$, each one expresses a condition on a subset of $X$
- an objective function to be optimized $f$ for a COP

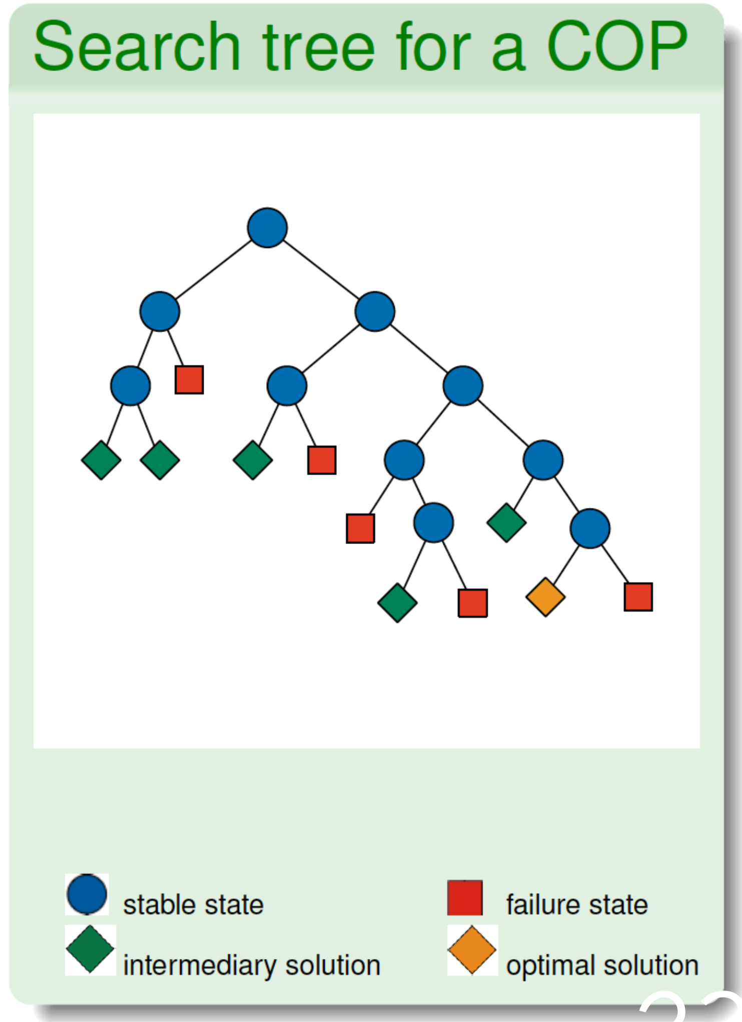
A *solution* is a complete assignment $x \in X \mapsto v \in Dom(x)$ s.t. all the constraints are satisfied and $f$ is optimized.

## Programming using constraints

- Problem must be modeled as a CSP/COP
- The solver searches for one/all/the best solution(s)
- Strategies can be given to guide the search

# Principles of CP

- Each level is for a variable (assume binary domain)
- Tree is built dynamically
- We can use the constraints to prune parts of the search space
- All done in a problem agnostic language – lots of high quality languages – i.e. GeCode, Zinc, Choco Numberjack etc.
- Lots of conferences and applications particularly in AI conferences.



# The Benefits of CP Formulation

Kuo, Chia-Tung, and Ian Davidson. "A Framework for Outlier Description Using Constraint Programming." *AAAI*. 2016.

Projection vector

Lower and upper bound on # NN for normal and outliers

Objective    Maximize $k_N - k_O$

Variables    $F = [f_1, f_2, \ldots, f_{|S|}] \in \{0, 1\}^{|S|}$

$k_{min} \leq k_O \leq k_N \leq k_{max}$

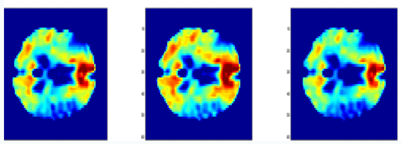
NN distance $0 \leq r \leq r_{max}$

Constraints    $\forall x \in N, \; |\mathcal{N}_F(x, r)| \geq k_N$

$\forall y \in O, \; |\mathcal{N}_F(y, r)| < k_O$

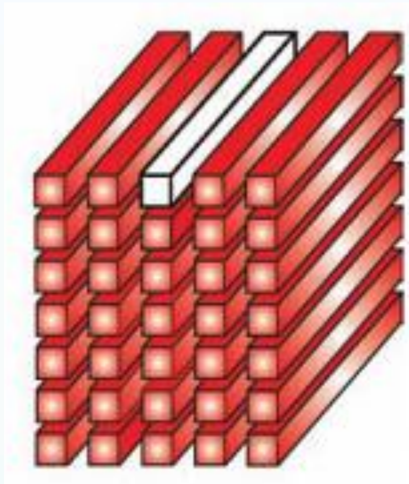
Multi-criteria optimization over k, r and F

# Ex. Functional Network Discovery
## [With NMRC, Pennington Institute]



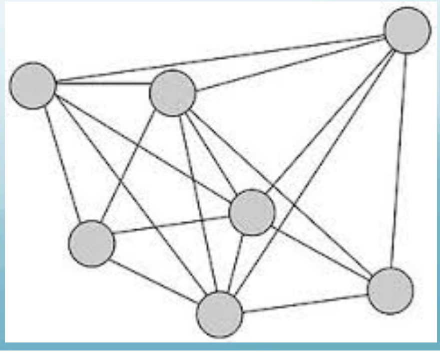
Take functional scans
Co-register with
structural scans

**Tensor Representation**



Stack Images Over Time
Each voxel is a time series

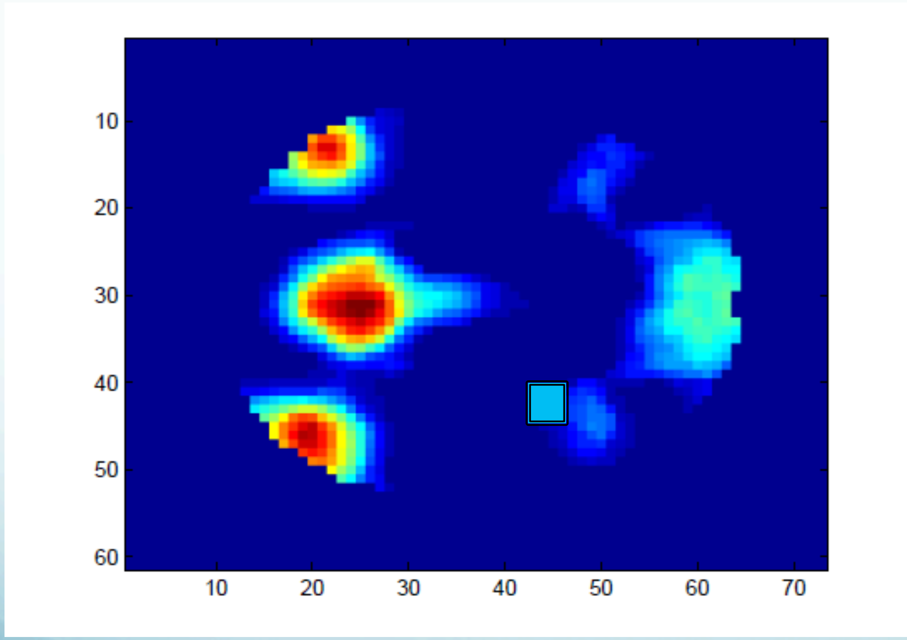
**Graph Representation**
Measure correlations over
voxels to construct edge
Weights
No need for DTW



High correlation represents
synchronized (or functional)
**connectivity**

# Functional Network Discovery

- Synchronized co-activation of spatially separated regions is associated with a **functional network**




**Neuroscience Question**

What links characterize control but not demented individuals? *Negative/Subtractive explanation*

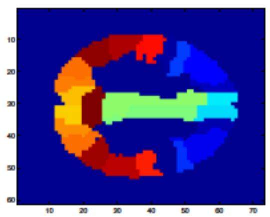

Liu et al.: Regional homogeneity, functional connectivity and imaging markers of Alzheimer's disease: A review of resting-state fMRI studies. Neuropsychologia 46, 1648-1656 (2008).
Venkataraman, A. et al.: Exploring Functional Connectivity in fMRI via Clustering. In ICASSP 2009.

# Experiment

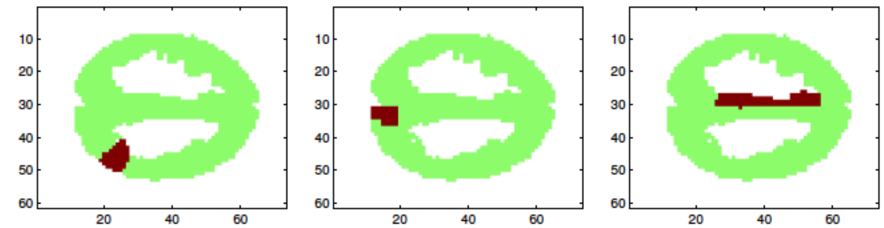
- Small data problem: 19 in N group and 21 in O group
- Each instance is represented by a fully connected graph of 27 anatomical regions/nodes (i.e. 351 edges)
  - These edges are our features
- kmin = 1; kmax = 10 and r is discretized.
- The optimum solution
  - $|F| = 17$, $k_N = 4$; $k_O$ **= 0** and $r = 2.5$.
- Explanation: the connections/edge weights in this 17 dimensional subspace are **consistent** for the *N* group but not the *O* group.

# Visualization of Results

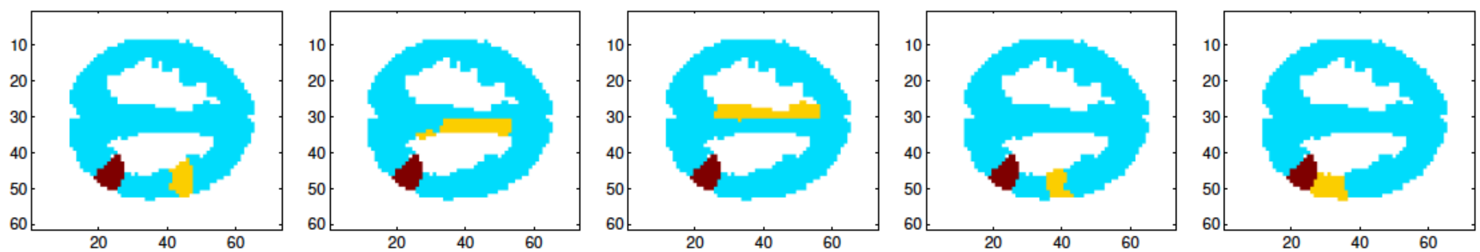
## These Connections/Regions are Very Different for Demented Individuals (i.e. they break down)



(a) Color-coded known anatomical regions in the brain.



(b) The top 3 anatomical regions appearing the most times in the pairs selected in $F$; Number of occurrences are 5 (left), 3 (middle) and 3 (right).



(c) The pairs of anatomical regions selected in $F$ that involve the most frequent occurring region (left in 2(b)).

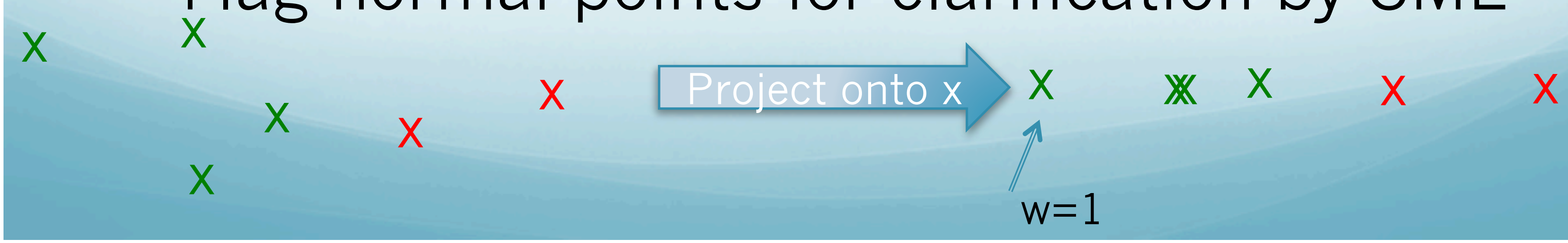
# Human in Loop Extension

Kuo, Chia-Tung, and Ian Davidson. "A Framework for Outlier Description Using Constraint Programming." *AAAI*. 2016.

| | |
|---|---|
| Objective | Maximize $k_N - k_O$ |
| Variables | $F = [f_1, f_2, \ldots, f_{\lvert S\rvert}] \in \{0,1\}^{\lvert S\rvert}$ |
| | $k_{min} \le k_O \le k_N \le k_{max}$ |
| | $0 \le r \le r_{max}$ |
| Constraints | $\forall x_i \in N,\ \lvert\mathcal{N}_F(x_i, r)\rvert \ge (1 - w_i)k_N$ |
| | $\sum_{i=1}^{n} w_i \le w_{max}$ |
| | $\forall y \in O,\ \lvert\mathcal{N}_F(y, r)\rvert < k_O$ |

I can ignore some points
From the NN constraint

Multi-criteria optimization over k, r, F and **w**
Flag normal points for clarification by SME



# Two Sub Space Explanation

Kuo, Chia-Tung, and Ian Davidson. "A Framework for Outlier Description Using Constraint Programming." *AAAI*. 2016.

| | |
|---|---|
| Objective | Maximize $k_N - k_O$ |
| Variables | $F = [f_1, \ldots, f_{\lvert S\rvert}], G = [g_1, \ldots, g_{\lvert S\rvert}] \in \{0,1\}^{\lvert S\rvert}$ |
| | $k_{min} \leq k_O \leq k_N \leq k_{max}$ |
| | $0 \leq r_F, r_G \leq r_{max}$ |
| Constraints | $\forall x \in N,\ \lvert\mathcal{N}_F(x, r_F)\rvert \geq k_N \texttt{ AND } \lvert\mathcal{N}_G(x, r_G)\rvert \geq k_N$ |
| | $\forall y \in O,\ \lvert\mathcal{N}_F(y, r_F)\rvert < k_O \texttt{ OR } \lvert\mathcal{N}_G(y, r_G)\rvert < k_O$ |

Multi-criteria optimization over k, r, F and G

Here we have say the normal group have stable features in two subspaces (F,G) and the outliers have stable features in at most one of them.

# New Directions

- General framework to do outlier explanation – Use other properties beyond density i.e. diameters
- Polynomial Time Algorithms (previous work in CP formulation was limited to 1000's of points in 100s of dimensions)
  - What happens if we want to apply these methods to huge data sets
    - **Fixed Parameter Tractability** in Terms of the Number of Dimensions
  - What happens if we want to apply these methods to very wide data sets
    - **Using Johnson-Lindenstrauss Lemma** to Reduce the Number of Dimensions

# FPT Algorithms

For a given F optimizing over w and r can be completed in Polynomial time

| | |
|---|---|
| Objective | Maximize $k_N - k_O$ |
| Variables | $F = [f_1, f_2, \ldots, f_{\lvert S \rvert}] \in \{0,1\}^{\lvert S \rvert}$ |
| | $k_{min} \le k_O \le k_N \le k_{max}$ |
| | $0 \le r \le r_{max}$ |
| Constraints | $\forall x \in N,\ \lvert \mathcal{N}_F(x, r) \rvert \ge k_N$ |
| | $\forall y \in O,\ \lvert \mathcal{N}_F(y, r) \rvert < k_O$ |

Lower and upper bound on # NN f
normal and outlier

Projection vector

NN distance

**Algorithm 2:** Algorithm for a Given Subspace

1 Project the point set $P = N \cup O$ onto the chosen subspace of dimension $d_1$. Let $P'$ denote the projected set of points and let $n = |P| = |P'|$.

2 Compute all the $O(n^2)$ pairwise distances between points in $P'$ (in dimension $d_1$).

3 Let $t$ denote the number of distinct distances computed in Step 2. Let $a_1 < a_2 < \ldots < a_t$ denote the $t$ distances sorted in increasing order.

4 CurrentMax $= 0$.

5 **for** $i = 1$ **to** $t$ **do**
6     Let radius $r = a_i$.
7     **if** $r \le r_{max}$ **then**
8         Compute $k_N$, $k_O$ for distance $r$.
9         **if** $((k_{min} \le k_O \le k_N \le k_{max})$ **and** $(k_N - k_O >$ CurrentMax$))$ **then**
10             CurrentMax $= k_N - k_O$.
11         **end**
12     **end**
13 **end**

14 Output CurrentMax and the corresponding radius $r$.

# FPT Algorithms

| | |
|---|---|
| Objective | Maximize $k_N - k_O$ |
| Variables | $F = [f_1, f_2, \ldots, f_{\lvert S\rvert}] \in \{0,1\}^{\lvert S\rvert}$ |
| | $k_{min} \leq k_O \leq k_N \leq k_{max}$ |
| | $0 \leq r \leq r_{max}$ |
| Constraints | $\forall x \in N,\ \lvert\mathcal{N}_F(x,r)\rvert \geq k_N$ |
| | $\forall y \in O,\ \lvert\mathcal{N}_F(y,r)\rvert < k_O$ |

Lower and upper bound on # NN fo
normal and outlier

Projection vector

NN distance

**Algorithm 2:** Algorithm for a Given Subspace

1 Project the point set $P = N \cup O$ onto the chosen subspace of dimension $d_1$. Let $P'$ denote the projected set of points and let $n = |P| = |P'|$.
2 Compute all the $O(n^2)$ pairwise distances between points in $P'$ (in dimension $d_1$).
3 Let $t$ denote the number of distinct distances computed in Step 2. Let $a_1 < a_2 < \ldots < a_t$ denote the $t$ distances sorted in increasing order.
4 CurrentMax = 0.
5 **for** $i = 1$ **to** $t$ **do**
6   Let radius $r = a_i$.
7   **if** $r \leq r_{max}$ **then**
8     Compute $k_N$, $k_O$ for distance $r$.
9     **if** (($k_{min} \leq k_O \leq k_N \leq k_{max}$) **and** ($k_N - k_O >$ CurrentMax)) **then**
10       CurrentMax = $k_N - k_O$.
11     **end**
12   **end**
13 **end**
14 Output CurrentMax and the corresponding radius $r$.

Can we do better that
exhaustive search over F
i.e. SDM 2010 wrapping
binary search around SAT

# JL Lemma

**Statement of Johnson-Lindenstrauss (JL) Lemma:** Let $P$ be any set of $n$ points in $\mathbf{R}^d$. Given an $\epsilon$, $0 < \epsilon < 1$, let $k$ be an integer such that

$$k \geq \frac{4 \ln n}{(\epsilon^2/2 - \epsilon^3/3)}.$$

(Note that $k$ is *independent* of $d$.) Then, there is a function $g : \mathbf{R}^d \longrightarrow \mathbf{R}^k$ such that for any pair of points $u$ and $v$ in $P$, $(1-\epsilon)[D_d(u,v)]^2 \leq [D_k(g(u),g(v))]^2 \leq (1+\epsilon)[D_d(u,v)]^2$. Moreover, the function $g$ can be computed in randomized polynomial time.

**Algorithm 1:** A Randomized Algorithm for the Transformation Implied by the JL Lemma

1. Let $P$ denote the given set of $n$ points in $\mathbf{R}^d$, represented by the $n \times d$ matrix $A$. (Each row of $A$ corresponds to a point; each column represents a dimension.)
2. Choose $\epsilon > 0$ and $\beta > 0$. Let $k = k = \left\lceil \frac{4+2\beta}{\epsilon^2/2 - \epsilon^3/3} \ln n \right\rceil$.
3. Construct $R = [r_{ij}]$, a $d \times k$ matrix, where each entry $r_{ij}$ $(1 \leq i \leq d$ and $1 \leq j \leq k)$ is chosen independently as follows: $r_{ij} = +1$ with probability $1/2$ and $r_{ij} = -1$ with probability $1/2$.
4. Construct the $n \times k$ matrix $E$ by the following equation: $E = \frac{1}{\sqrt{k}} A R$.
5. Now, $E$ specifies the transformed set of points in $\mathbf{R}^k$, where $k = O(\log n)$.

# Conclusion

- Anomaly explanation is an important problem
- I think we need new formulations to handle the problem due to i) small data, ii) need for interpretabilty
- I presented two:
  - A) A set coverage formulation
  - B) A density formulation
- I used discrete optimization formulations which have the benefit of interpretability amongst others.
- Lots of interesting directions and settings
  - Vertex labeled graphs (explanations in terms of labels)
  - Spatial and/or temporal data (explanations in terms of location and/or time).
  - Human in the loop extensions (easy active learning?)