Anomaly Detection

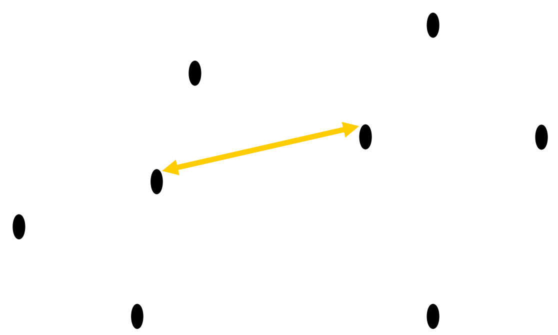
- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Limitations of statistical approaches:

- In many cases, data distribution is not normal or it may not be known
- High dimensional data does not usually follow a known multivariate distribution

Unsupervised Methods

- **Data is represented as a vector of features.**
  We have a distance measure to evaluate nearness between two points
- **Nearest Neighbor or Distance based methods**:
  Given a database D, and a data point $x \in D$, the method assigns an anomaly score to x, based on the distance of x to the other points



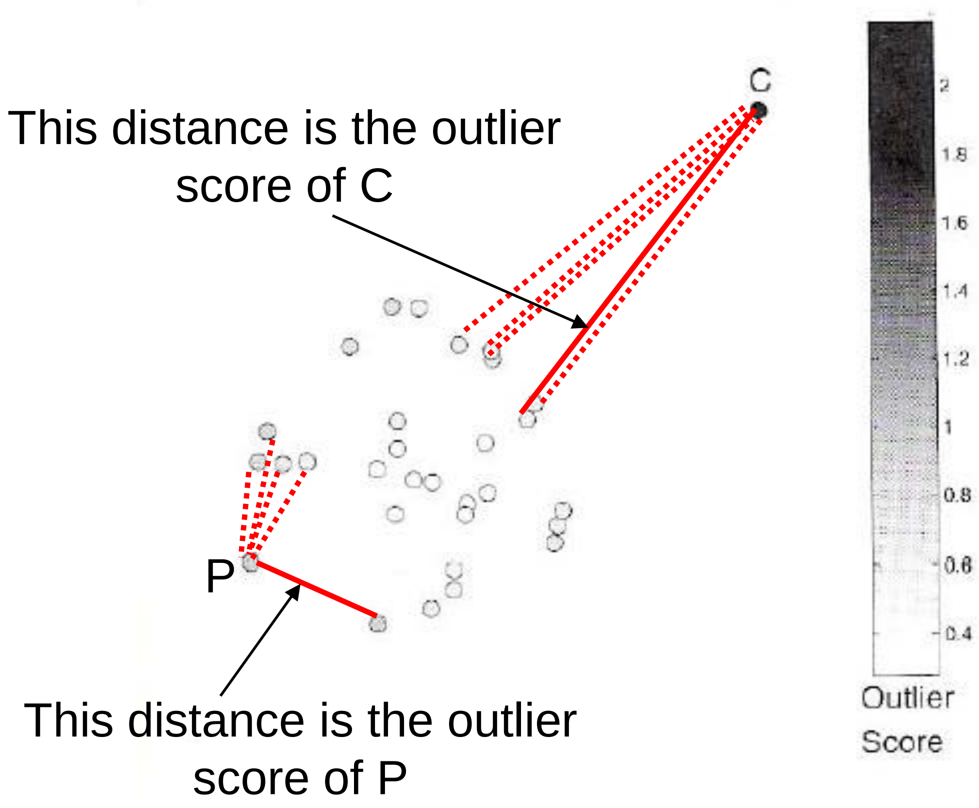
- Nearest neighbor approaches are score-based: Given a database D, and a data point $x \in D$, the method assigns an anomaly score to x
  - Given a database D, find all the data points $x \in D$ with anomaly scores **greater than** some threshold t
  - Given a database D, find all the data points $x \in D$ having the **top-n** largest anomaly scores f(x)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a **test point** x, compute the anomaly score of x with respect to D

- Two major approaches
  - Nearest-neighbor based
  - Nearest-neighbor density based

- Approach:
  - Compute the distance between every pair of data points
  - Fix a magic number k representing the k-th nearest point to another point
  - For a given point P, compute its *outlier score* as the distance of P to its k-nearest neighbor.
    There are no clusters. Neighbor refers to a point

Unsupervised Methods

k = 5



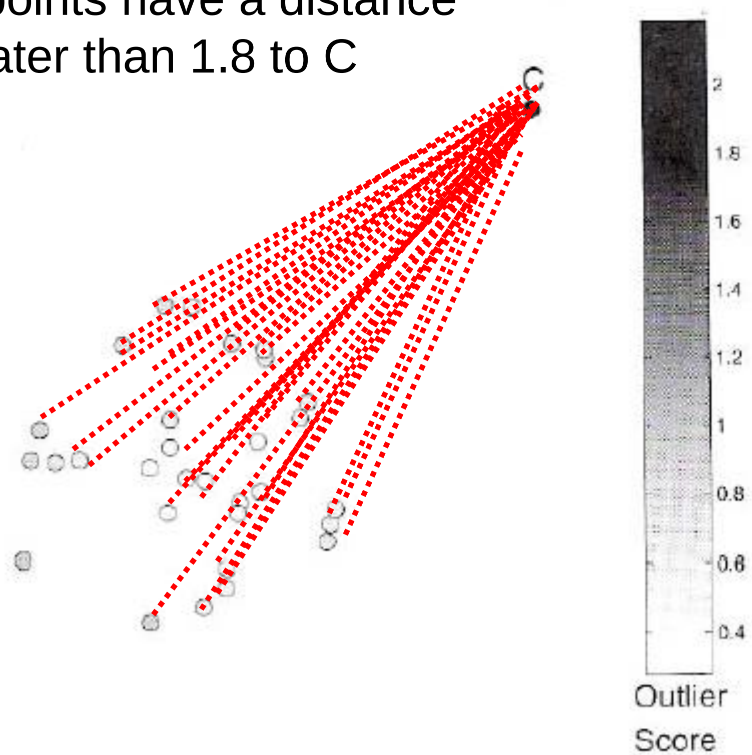
Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. **2000**. Distance-based outliers: algorithms and applications. *The VLDB Journal* 8, 3-4 (February 2000), 237-253. DOI=10.1007/s007780050006 http://dx.doi.org/10.1007/s007780050006

- Similar Approach:
  - Instead of fixing k, a distance D is fixed. Then, the method consider the percentage of points which are far away from the outlier.
  - An object O in a dataset T is a distance based DB(p,D) outlier if at least fraction p of the objects in T is located at a distance from O greater than D.

Unsupervised Methods

Knorr et al.

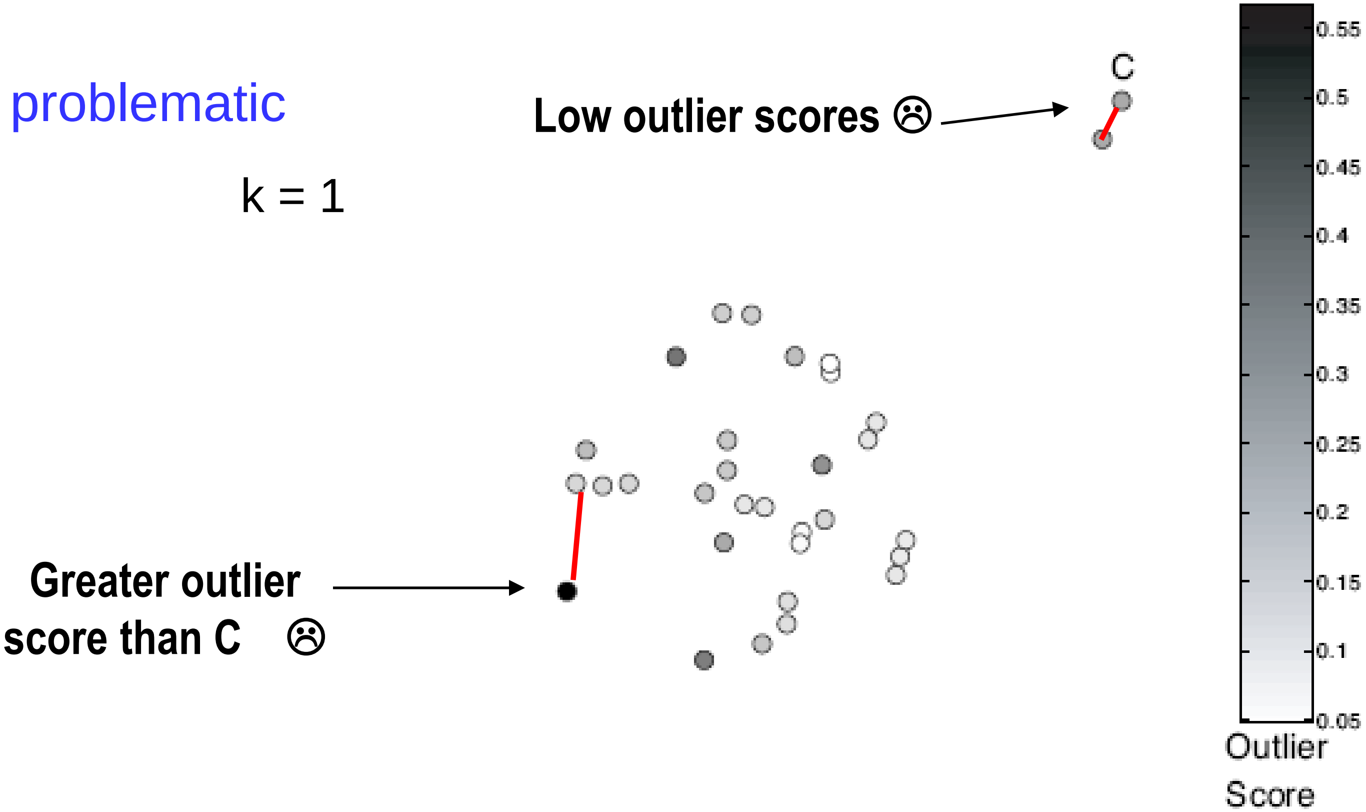
97% of points have a distance greater than 1.8 to C



Unsupervised Methods

Choice of k is problematic

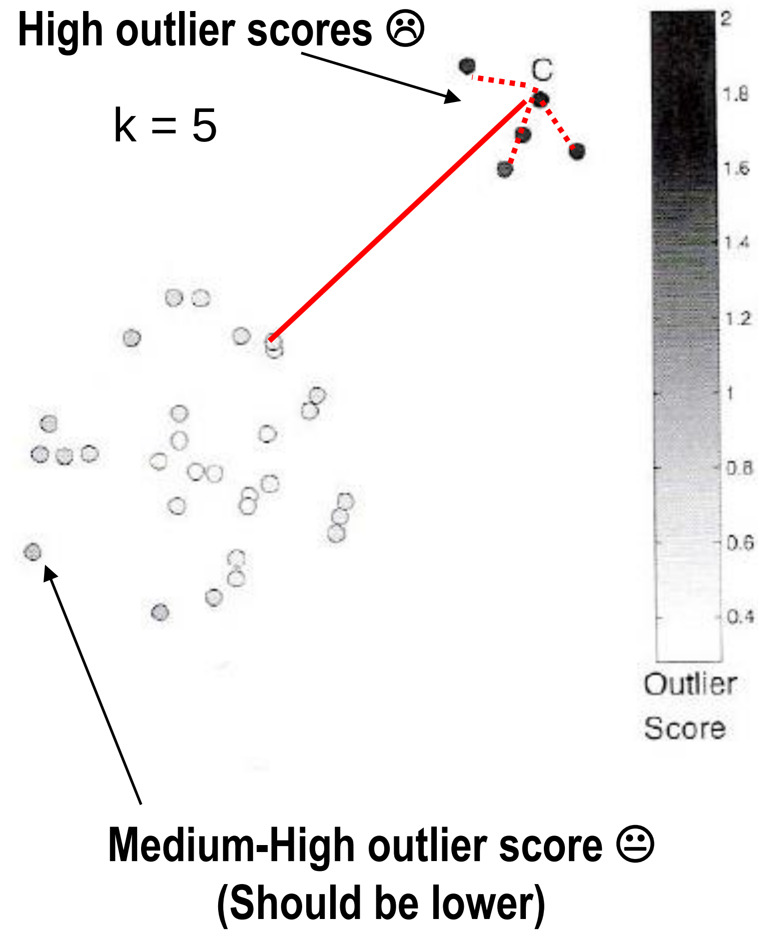
k = 1



Unsupervised Methods

Choice of k is problematic

All the points in any isolated natural cluster with fewer points than k, have high outlier score



Unsupervised Methods

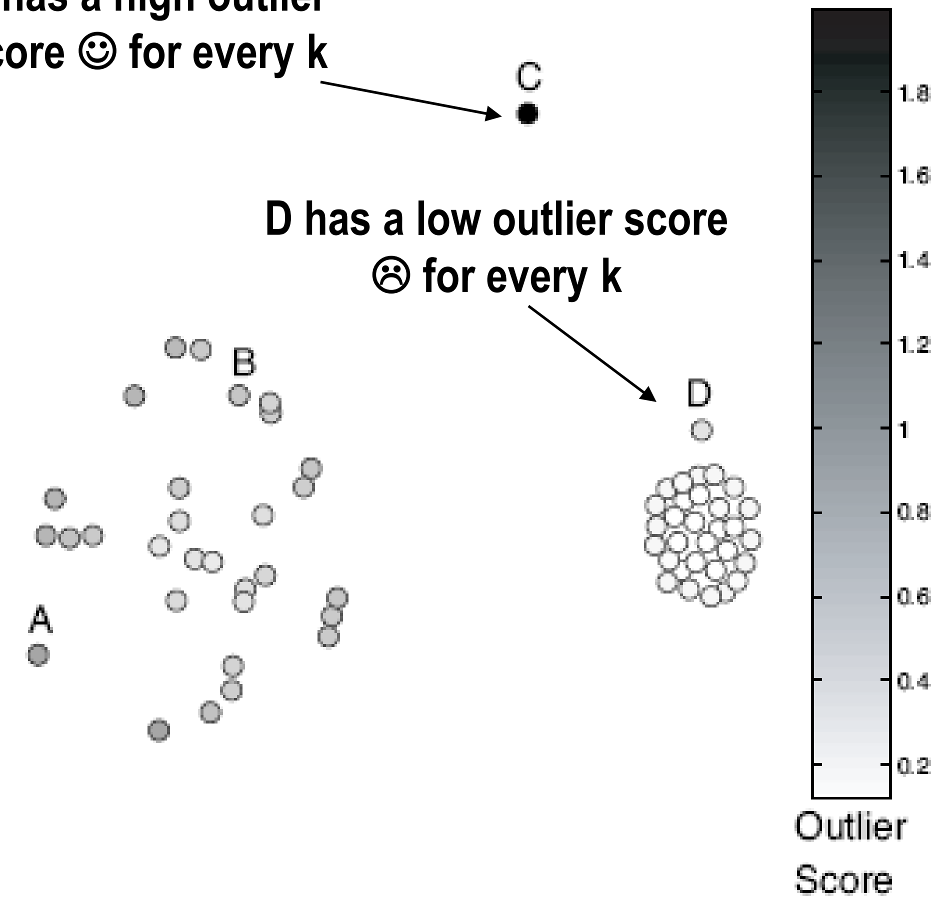
Choice of k is problematic

We could mitigate the problem by taking the average distance to the k-nearest neighbors but is still poor.

C has a high outlier score ☺ for every k

D has a low outlier score ☹ for every k

A has a medium-high outlier score ☹ for every k



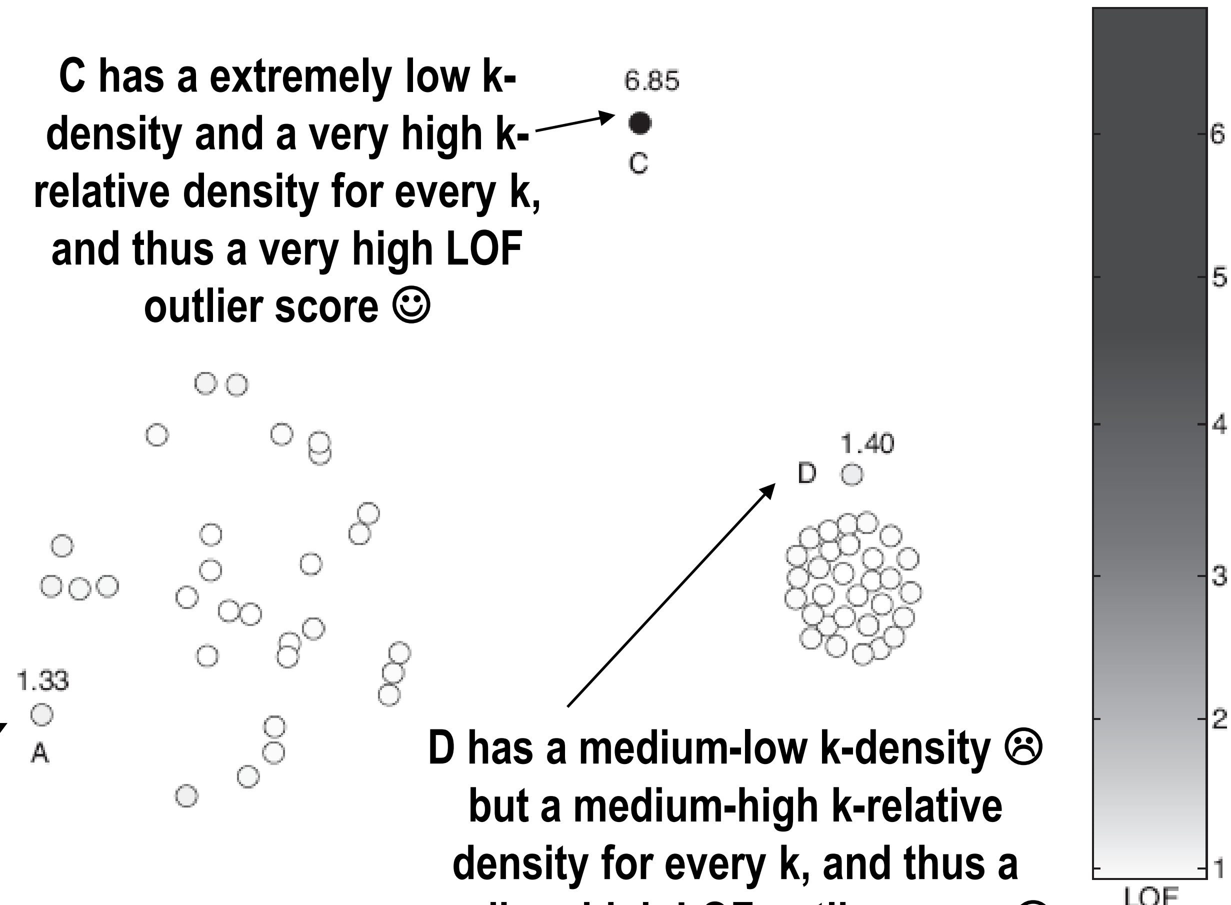
Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. **2000**. **LOF**: identifying density-based local outliers. ***SIGMOD Rec****.* 29, 2 (May 2000), 93-104. DOI=10.1145/335191.335388

- Define the *k-density of a point as the inverse of the average sum of the distances to its k-nearest neighbors.*

- Define the *k-relative density* of a point P as the ratio between its *k-density* and the average *k-densities* of its *k*-nearest neigbhors

- The outlier score of a point P (called LOF for this method) is its *k-relative density.*.

Breunig et al (**LOF**)



C has a extremely low k-density and a very high k-relative density for every k, and thus a very high LOF outlier score ☺

A has a very low k-density 😐 but a medium-low k-relative density for every k, and thus a medium-low LOF outlier score ☺

D has a medium-low k-density ☹ but a medium-high k-relative density for every k, and thus a medium-high LOF outlier score ☺

Some remarks:

- Computation requires comparing many distances.

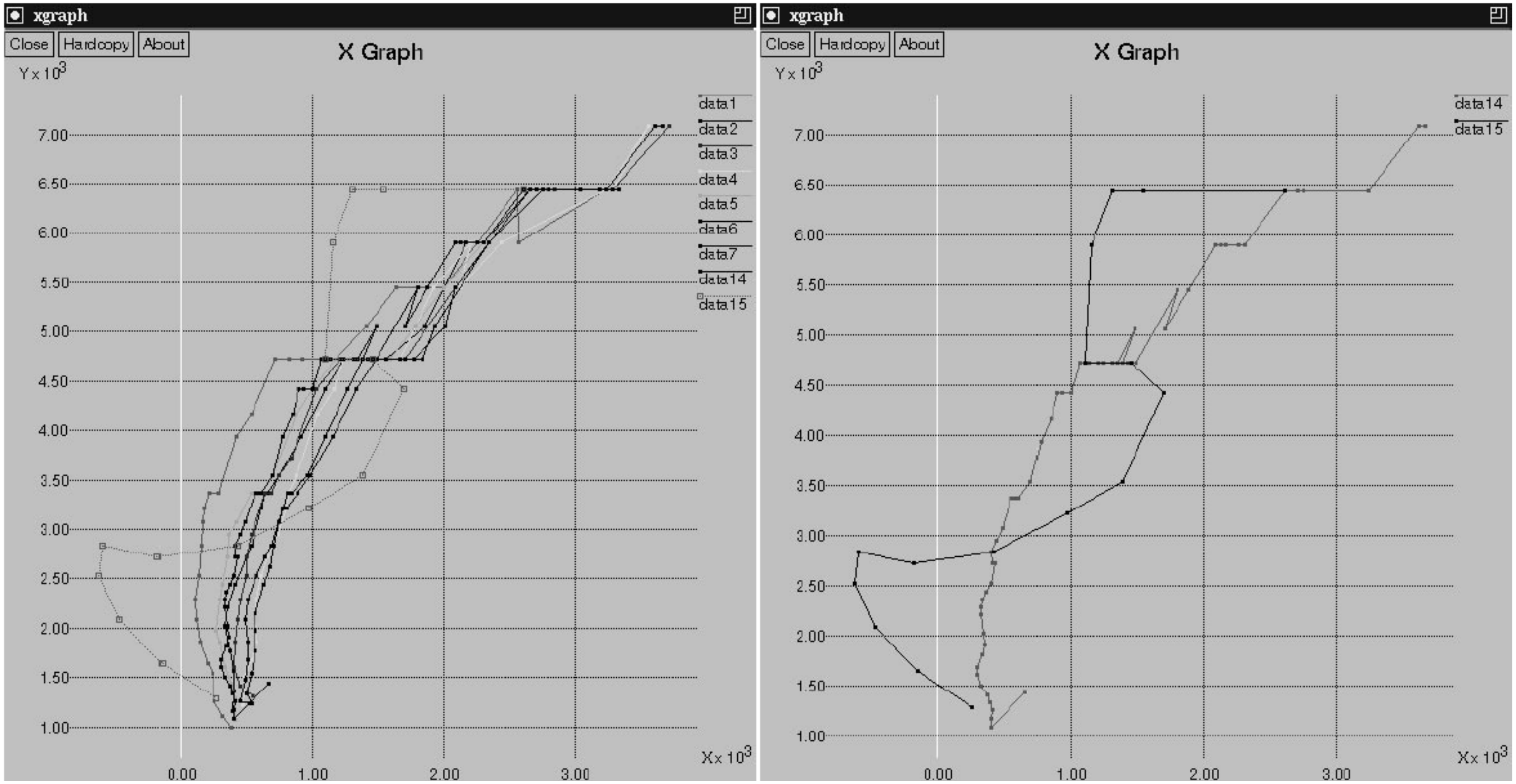
  → $O(N^2)$

  Some improvements can be considered: Divide the data space in cells and use this spatial information to prune the search → O(pN) where p is the dimensionality and N is the size of the data

- As in any problem in data mining, it's very important the preprocessing phase

Knorr et al.

Video Trajectory Surveillance



Knorr et al.

Application: Video Trajectory Surveillance

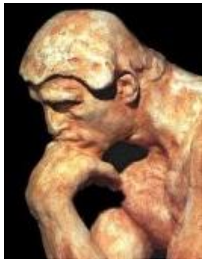
Trajectories are NOT represented in a 2D-position space.

Trajectories are summarized by the following features:

- Start and end points.
- Number of points: the length of the trajectory.
- Heading: the average, minimum, and maximum values of the directional vector of the tangent of the trajectory at each point.
- Velocity: average, minimum, and maximum velocity of the person during the trajectory.

An ad hoc distance measure is defined in this space

MINDS – MINnesota INtrusion Detection System (LOF based)



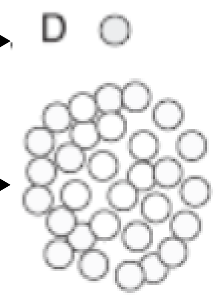
- Basic features of individual TCP connections
  - source & destination IP/port, protocol, number of bytes, duration, number of packets
- Time based features: detect fast scans -e.g: DoS attacks-
  - For the same source (destination) IP address, number of flows to unique destination (source) IP addresses inside the network *in last T seconds*
  - Number of connections from source (destination) IP to the same destination (source) port *in last T seconds*
- Connection based features: detect slow scans
  - For the same source (destination) IP address, number of flows to unique destination (source) IP addresses inside the network *in last N connections*
  - Number of connections from source (destination) IP to the same destination (source) port *in last N connections*

MINDS – MINnesota INtrusion Detection System (LOF based)

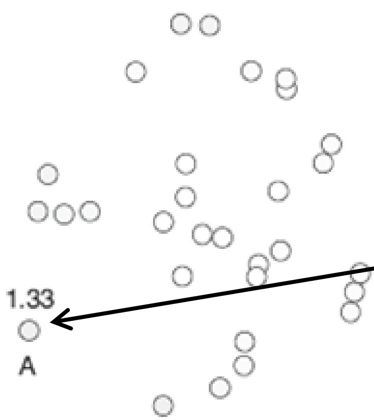
In order to avoid computation time, MINDS uses a sample of non-anomalous data entries and compare new entries with this sample (in a ”semisupervised way”)

Example: Slapper worn → Not detected with a simple distance based approach but detected with LOF:

Combination of source-destination port very rare. Detected as anomaly ☺



Similar scans



Some worms not detected ☹ : Portsweeps scans which are located in the sparse region of normal data.