Master Degree in Artificial Intelligence for Science and Technology

# Anomaly Detection: Proximity Based Approaches


Fabio Stella

Department of Informatics, Systems and Communication

University of Milan-Bicocca

fabio.stella@unimib.it

# OUTLOOK

- Proximity Based Approaches
  - distance based
  - density based
- Local Outlier Factor
- Connectivity Outlier Factor

# PROXIMITY BASED APPROACHES

- KEY ASSUMPTION: normal points have close neighbors while anomalies are located far from other points

- GENERAL TWO-STEP APPROACH
  - compute neighborhood for each data record
  - analyze the neighborhood to determine whether data record is anomaly or not

- CATEGORIES:
  - distance based methods
    - anomalies are data points most distant from other points
  - density based methods
    - anomalies are data points in low density regions

# PROXIMITY BASED APPROACHES

- **ADVANTAGE**
  - can be used in unsupervised or semi-supervised setting (do not make any assumptions about data distribution)

- **DRAWBACKS**
  - if normal points do not have sufficient number of neighbors the techniques may fail
  - computationally expensive
  - in high dimensional spaces, data is sparse and the concept of similarity may not be meaningful anymore.

    due to the sparseness, distances between any two data records may become quite similar $\Rightarrow$ each data record may be considered as potential outlier!

# PROXIMITY BASED APPROACHES

- **DISTANCE BASED APPROACHES**
  - a point $O^*$ in a dataset is a distance based $DB(p, d)$ outlier if at least a fraction $p$ of the points in the data set lies greater than distance $d$ from the point $O^*$

- **DENSITY BASED APPROACHES**
  - compute local densities of particular regions and declare instances in low density regions as potential anomalies
  - approaches
    - Local Outlier Factor (LOF)
    - Connectivity Outlier Factor (COF)
    - Multi-Granularity Deviation Factor (MDEF)

# PROXIMITY BASED APPROACHES: DISTANCE BASED OUTLIER DETECTION

- **NEAREST NEIGHBOR (NN) APPROACH*,****
  - for each data point $O$ compute the distance to the $k^{th}$ nearest neighbor $d_k$
  - sort all data points according to the distance $d_k$
  - outliers are points that have the largest distance $d_k$ and therefore are located in the more sparse neighborhoods
  - usually data points that have top $n\%$ distance $d_k$ are identified as outliers
    - $n$ is a user parameter
  - not suitable for datasets that have modes with varying density

* Knorr, Ng, Algorithms for Mining Distance-Based Outliers in Large Datasets, VLDB98

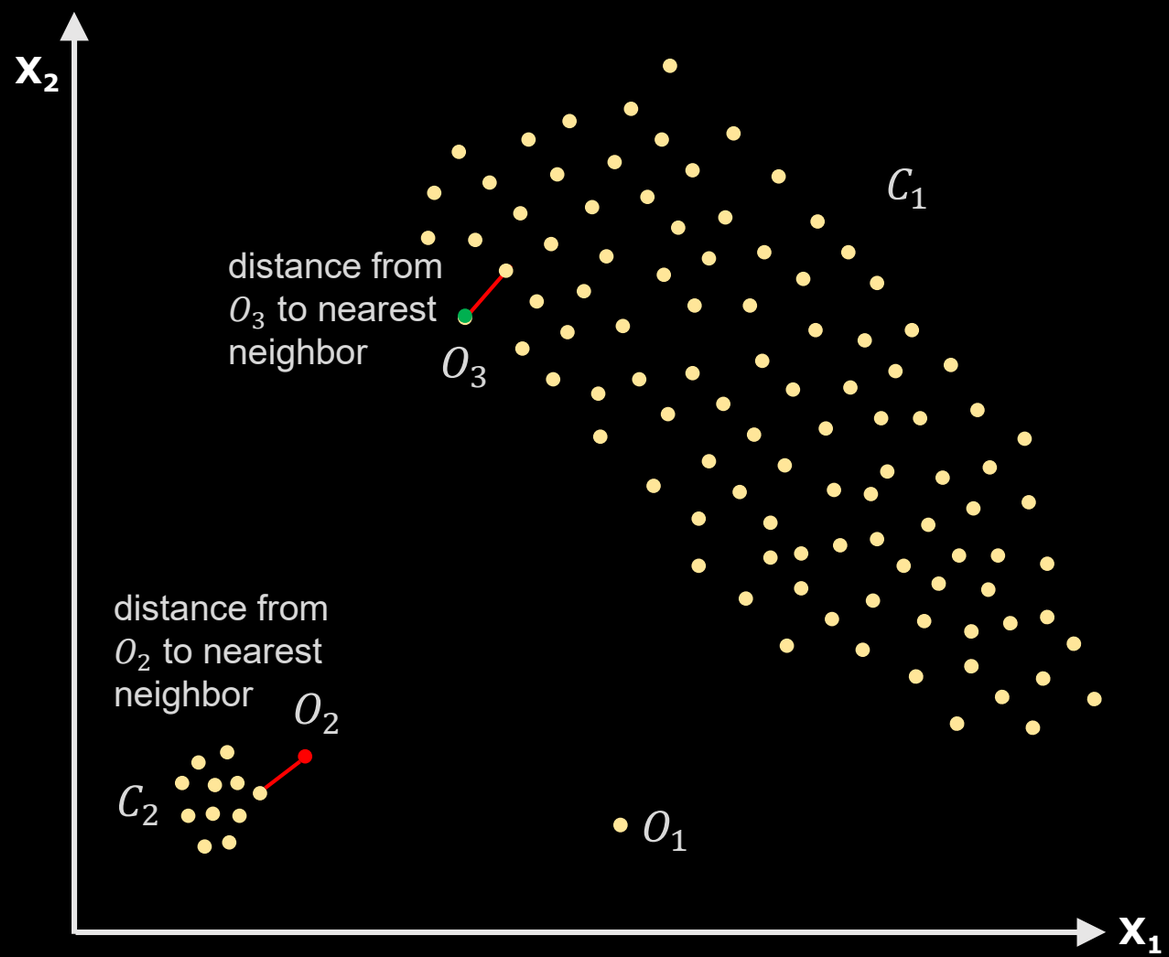
** S. Ramaswamy, R. Rastogi, S. Kyuseok: Efficient Algorithms for Mining Outliers from Large Data Sets, ACM SIGMOD Conf. On Management of Data, 2000.

# PROXIMITY BASED APPROACHES: DISTANCE BASED OUTLIER DETECTION

- **NEAREST NEIGHBOR (NN) APPROACH: STRENGTHS/WEAKNESSES**
  - simple
  - expensive – $O(n^2)$
  - sensitive to parameters ($k$ and $n\%$)
  - sensitive to variations in density
  - distance becomes less meaningful in high-dimensional space

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **LOCAL OUTLIER FACTOR (LOF) APPROACH**
  - Example
    - in the distance based approach, $O_2$ is not considered as outlier, while the LOF approach finds both $O_1$ and $O_2$ as outliers
    - distance based approach may consider $O_3$ as outlier, but LOF approach does not



# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **LOCAL OUTLIER FACTOR (LOF) APPROACH**
  - for each data point $p$ compute the distance to the k[th] nearest neighbor $d_k$
  - compute reachability distance for each data example $q$ with respect to data example $p$ as

$$reach - dist(p,q) = max\{d_k, d(p,q)\}$$

  - compute local reachability density of data example $q$ as inverse of the average reachability distance based on the $MinPts$ nearest neighbors of data example $q$

$$lrd(q) = \frac{MinPts}{\sum_p reach - dist_{MinPts}(p,q)}$$

  - $LOF(q)$ is the ratio of average local reachability density of $q$'s k-nearest neighbors and local reachability density of the data record $q$

$$LOF(q) = \frac{1}{MinPts}\sum_p \frac{lrd(p)}{lrd(q)}$$

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **LOCAL OUTLIER FACTOR (LOF) APPROACH: STRENGTHS/WEAKNESSES**
  - simple
  - expensive – $O(n^2)$
  - sensitive to parameters $k$ and $MinPts$
  - density becomes less meaningful in high-dimensional space

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **CONNECTIVITY OUTLIER FACTOR (COF)**

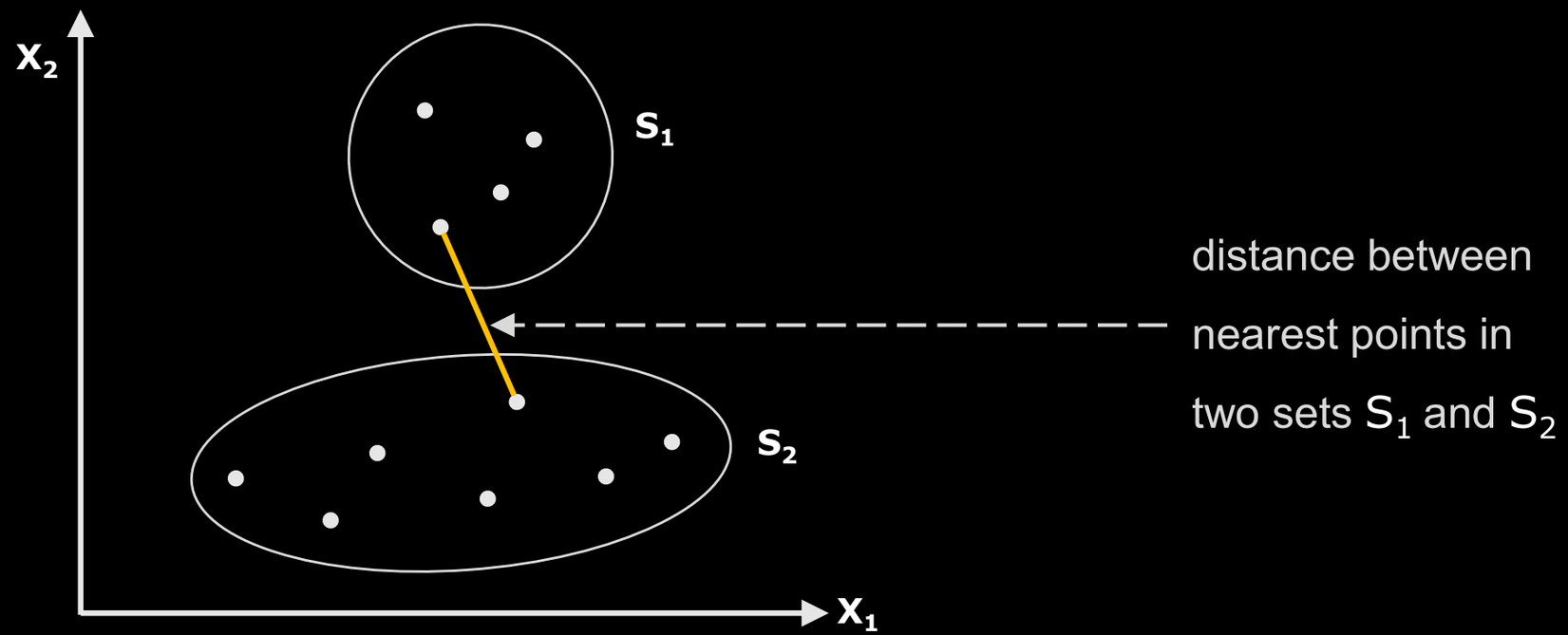
  — outliers are points $p$ where average chaining distance $ac - dist_{kNN(p)}(p)$ is larger than the average chaining distance of their k-nearest neighborhood $kNN(p)$

$$ac - dist_{kNN(p)}(p) = \sum_{i=1}^{r} \frac{2(r-i)}{r(r-i)} dist(e_i)$$

  — COF identifies outliers as points whose neighborhoods is sparser than the neighborhoods of their neighbors

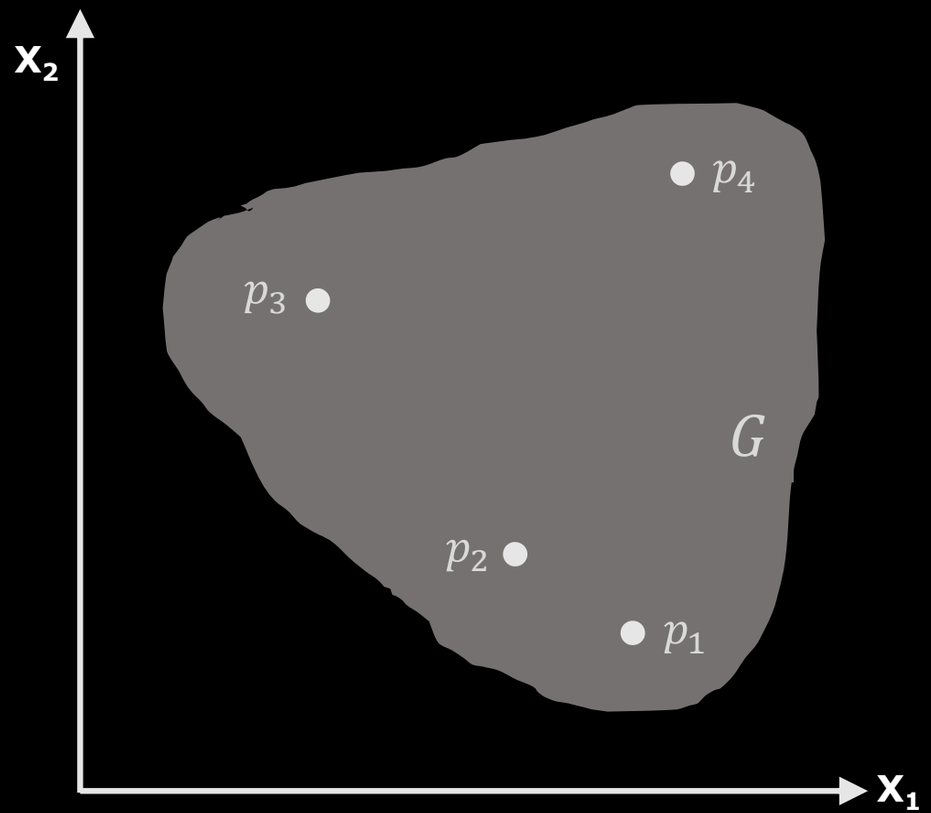
# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **CONNECTIVITY OUTLIER FACTOR (COF): DISTANCE BETWEEN TWO SETS**
  - distance between nearest points in two sets



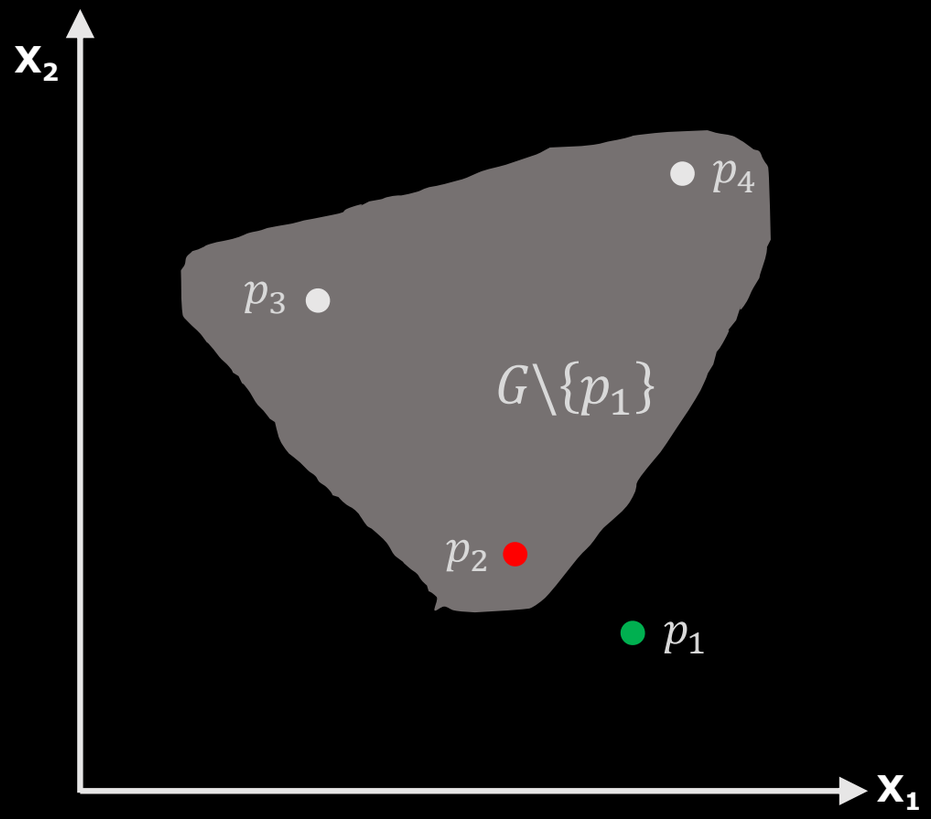
# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- CONNECTIVITY OUTLIER FACTOR (COF): SET BASED PATH



# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

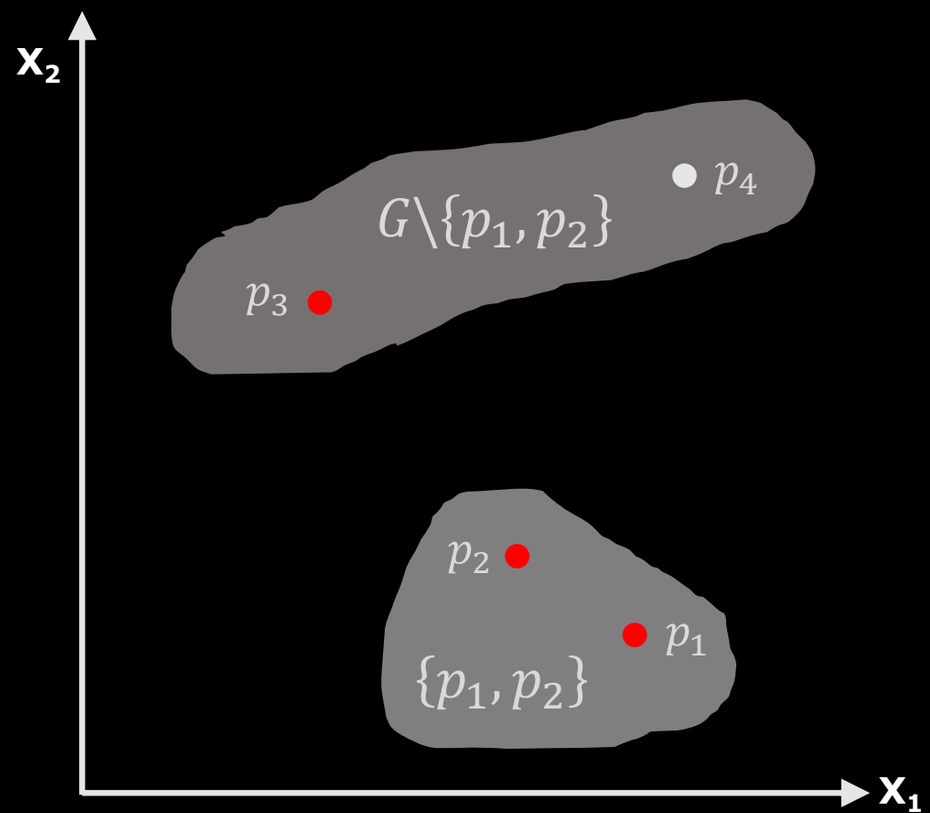
- CONNECTIVITY OUTLIER FACTOR (COF): SET BASED PATH




- $p_2$ is nearest neighbor of set $\{p_1\}$ in $G\backslash\{p_1\}$

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

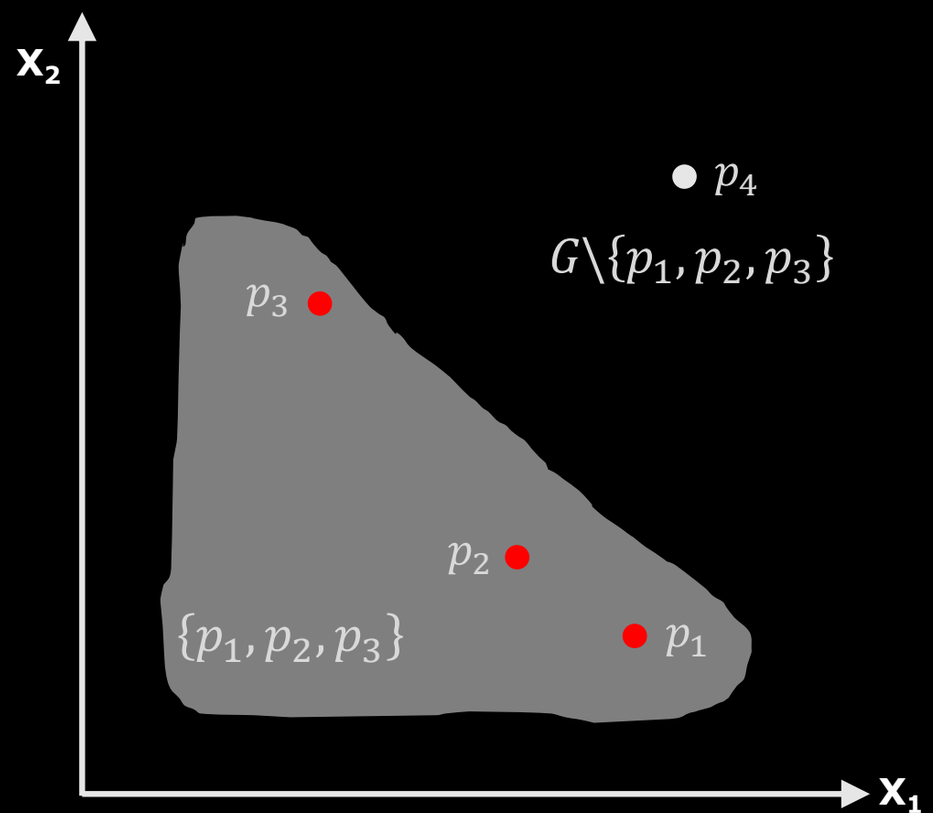
- CONNECTIVITY OUTLIER FACTOR (COF): SET BASED PATH




- $p_2$ is nearest neighbor of set $\{p_1\}$ in $G\backslash\{p_1\}$
- $p_3$ is nearest neighbor of set $\{p_1, p_2\}$ in $G\backslash\{p_1, p_2\}$

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

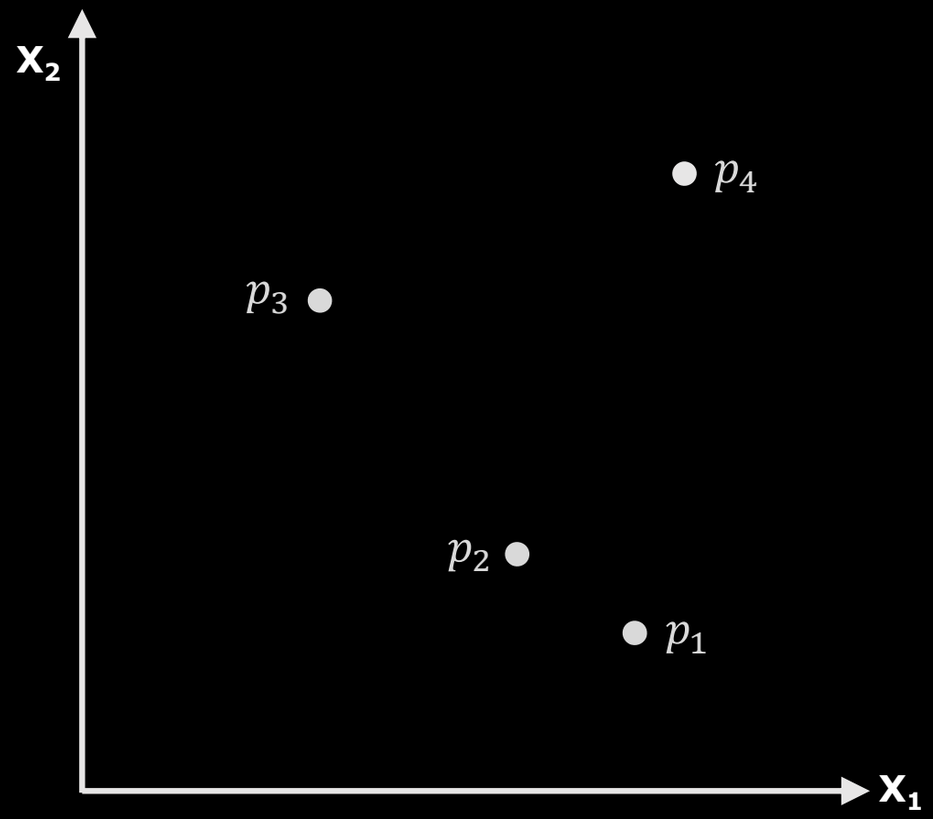
- CONNECTIVITY OUTLIER FACTOR (COF): SET BASED PATH




- $p_2$ is nearest neighbor of set $\{p_1\}$ in $G\backslash\{p_1\}$
- $p_3$ is nearest neighbor of set $\{p_1, p_2\}$ in $G\backslash\{p_1, p_2\}$
- $p_4$ is nearest neighbor of set $\{p_1, p_2, p_3\}$ in $G\backslash\{p_1, p_2, p_3\}$

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

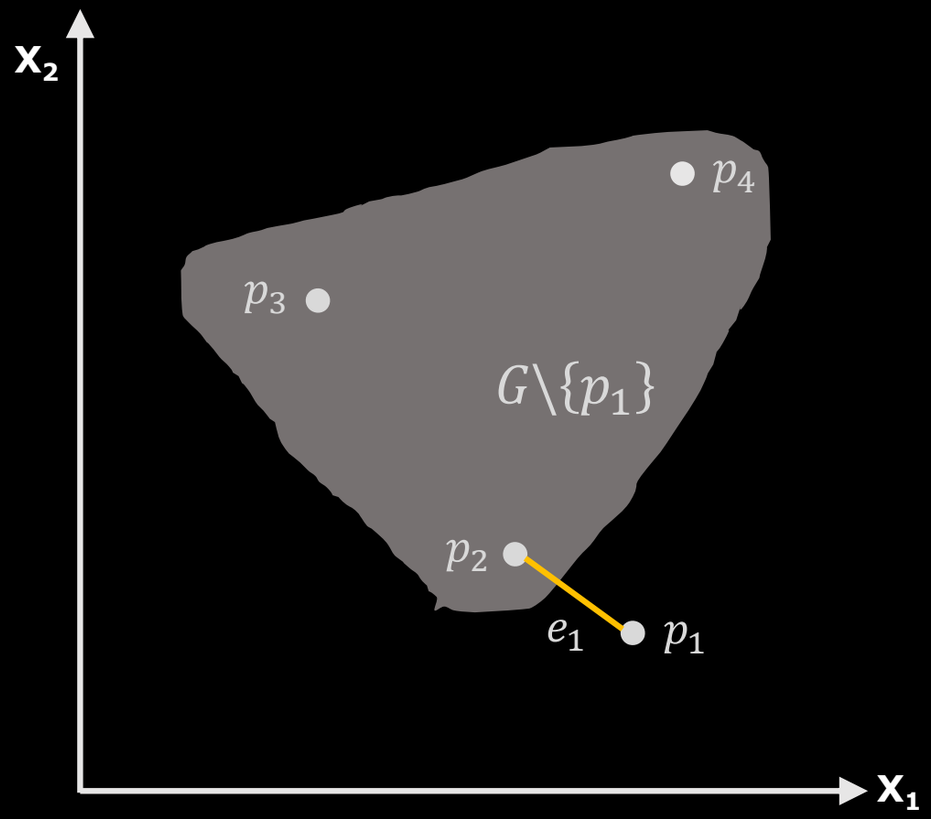
- **CONNECTIVITY OUTLIER FACTOR (COF): SET BASED PATH**




sequence $\{p_1, p_2, p_3, p_4\}$ is called Set Based Nearest Path (SBN) from $p_1$ on $G$

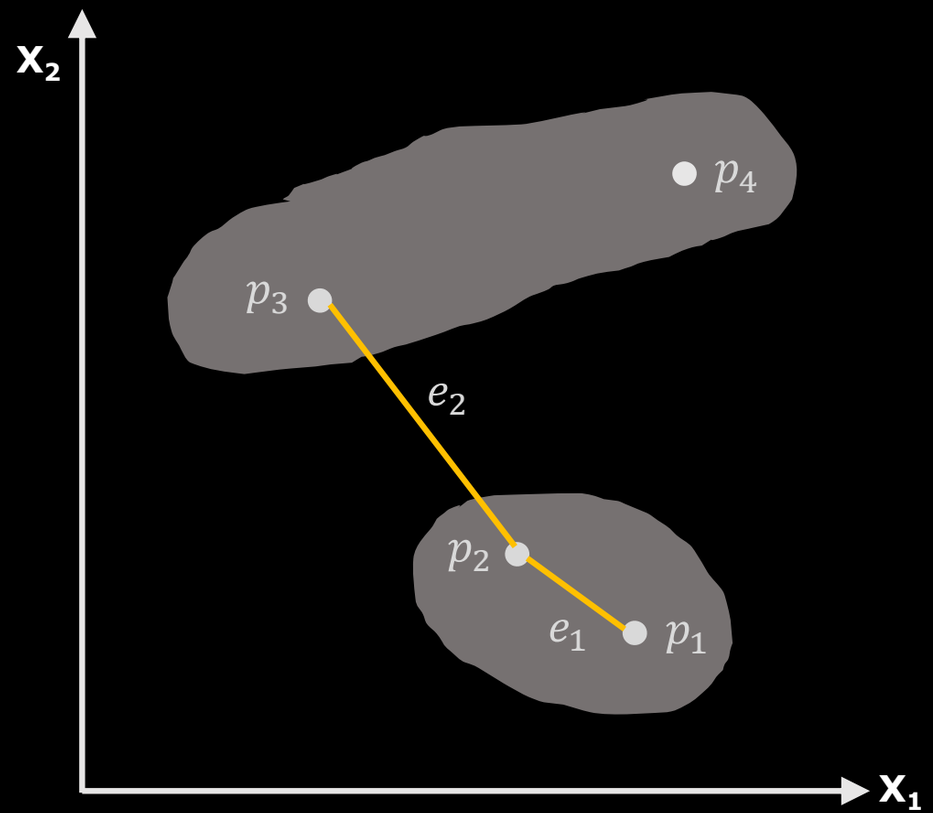
# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- CONNECTIVITY OUTLIER FACTOR (COF): COST DESCRIPTION



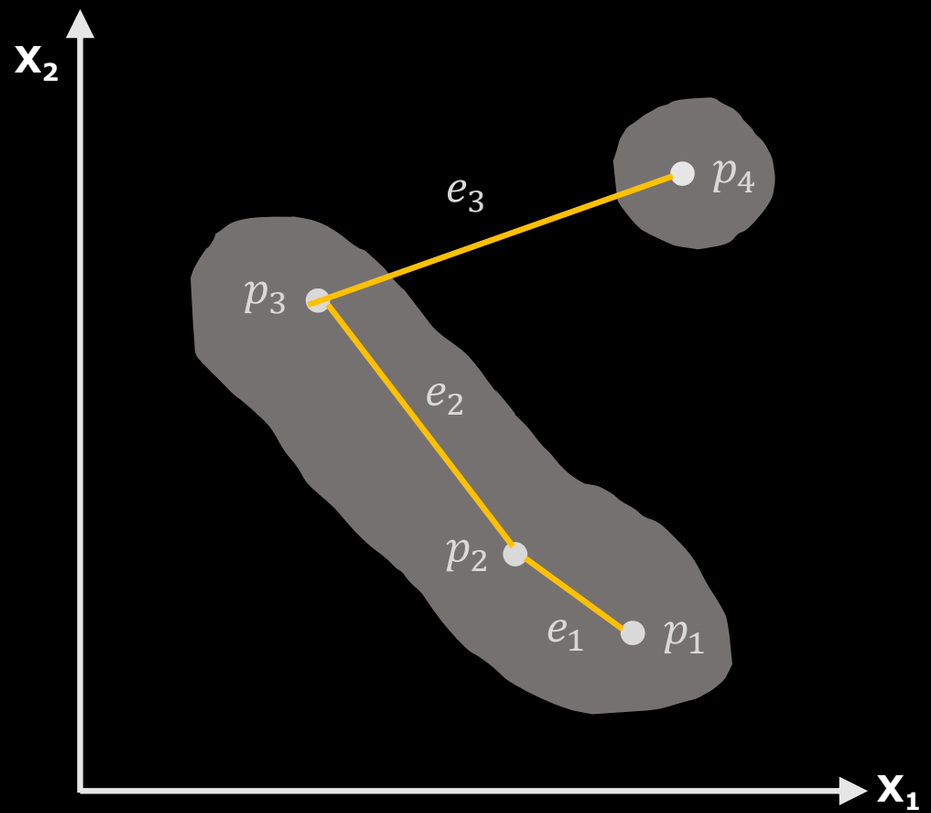
# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- CONNECTIVITY OUTLIER FACTOR (COF): COST DESCRIPTION



# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **CONNECTIVITY OUTLIER FACTOR (COF): COST DESCRIPTION**




— distances $dist(e_i)$ between two sets $\{p_1, \dots, p_i\}$ and $G\backslash\{p_1, \dots, p_i\}$ for each $i$ are called COST DESCRIPTIONS

— edges $e_i$ for each $i$ are called SBN trail

— SBN trail may not be a connected graph!

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **CONNECTIVITY OUTLIER FACTOR (COF)**
  - outliers are points $p$ where average chaining distance $ac - dist_{kNN(p)}(p)$ is larger than the average chaining distance of their k-nearest neighborhood $kNN(p)$

$$ac - dist_{kNN(p)}(p) = \sum_{i=1}^{r} \frac{2(r-i)}{r(r-i)} dist(e_i)$$

  - we average cost descriptions!
  - we would like to give more weights to points closer to the point $p_1$
  - the smaller $ac - dist$, the more compact is the neighborhood $G$ of $p$

# PROXIMITY BASED APPROACHES: DENSITY BASED OUTLIER DETECTION

- **CONNECTIVITY OUTLIER FACTOR (COF)**
  - COF is computed as the ratio of the $ac - dist$ (average chaining distance) at the point and the mean $ac - dist$ at the point's neighborhood
  - similar idea as LOF approach:
    - a point is an outlier if its neighborhood is less compact than the neighborhood of its neighbors

$$COF_k(p) = \frac{ac - dist_{kNN(p) \cup p}(p)}{\frac{1}{k}\sum_{o \in kNN(p)} ac - dist_{kNN(o) \cup o}(o)}$$

# RECAP

- Proximity Based Approaches
  - distance based
  - density based
- Local Outlier Factor
- Connectivity Outlier Factor