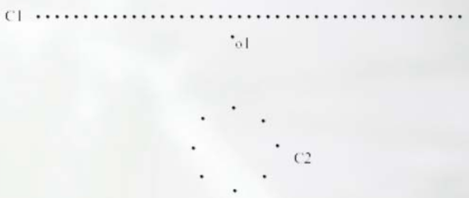
Enhancing Effectiveness of Outlier Detections for Low Density Patterns

J. Tang, Z. Chen, A. Fu, D. Cheung



John Sheldon
CMPUT 695
November 23rd, 2004

Overview

- Previous Outlier Detection Schemes
- The LOF Scheme
- Motivation for an Improved Scheme
- COF Scheme
- Examples
- Summary

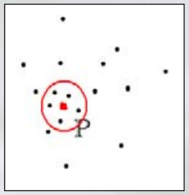
What is an outlier?

"An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism." Hawkins

- Application of outlier detection would be credit card fraud

Previous Outlier Detection Schemes

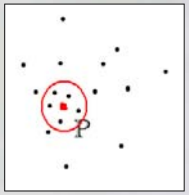
- Clustering
  - Generate outliers as a by-product
  - Outliers are highly dependant on algorithm
- Statistics
  - Examines deviations of individual data objects
  - Assumes prior knowledge of data distribution
- Distance Based Schemes
  - Based on number of other objects in neighborhood
  - More appropriate for detecting outliers w/o previous knowledge
    - DB(n,q)-outlier (Knorr and Ng)
    - (t,k)-nearest neighbor (Ramaswamy et al.)



LOF Detection Scheme

Density Based Scheme – Local Outlier Factor (LOF) (Breunig, et al.)

- Idea of k-distance = set of objects whose distance from point P is not greater then a distance k
- Local reachability density of P = density of P



$$LOF_k(p) = \frac{\sum_{o \in N_{k\text{-}distance(p)}(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_{k\text{-}distance(p)}(p)|}.$$

- measures how strong an object can be an outlier
- determined by comparing its density with those in its neighborhood

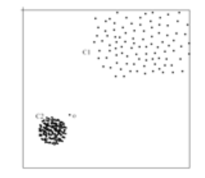
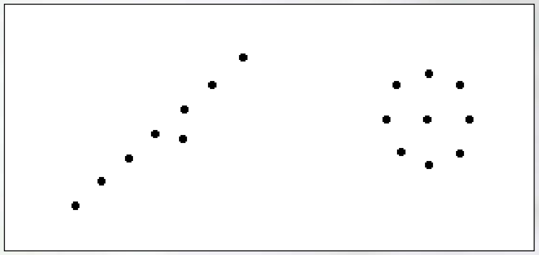


Figure 1: A Data Set showing the strength of LOF

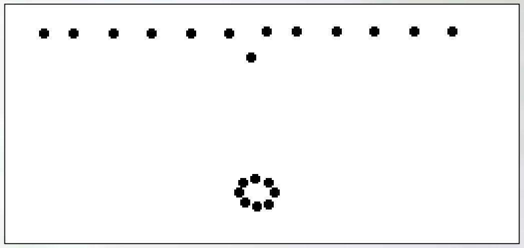
Weakness Inherent in LOF

- The weakness of LOF is that it may rule out outliers close to some non-outliers pattern that have similar low density



Connectivity-Based Outlier Factor (COF)

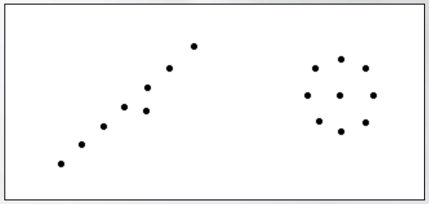
Based on the idea that an outlier does not always need to be of a lower density then a pattern it deviates from.



The LOF Scheme would not successfully find the outlier!

Connectivity-Based Outlier Factor (COF)

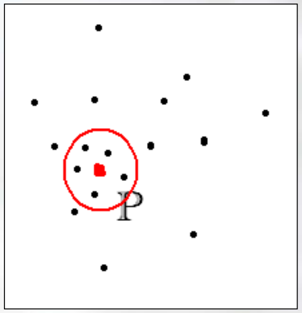
- Differentiates between “low-density” and “isolativity”
  - Low-density = number of objects in a close neighborhood
  - Isolativity = refers to the degree that an object is “connected” to other objects



- Observe that patterns with low density usually exhibit low dimensional structures

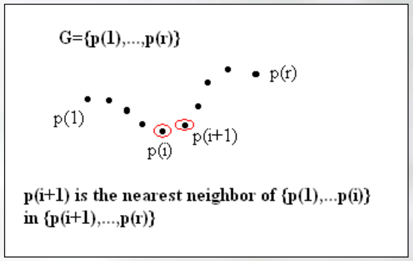
Connectivity-Based Outlier Factor (COF)

- Definitions…
  - k-nearest neighborhood - radius of a circle encompassing the k nearest objects or points.



Connectivity-Based Outlier Factor (COF)

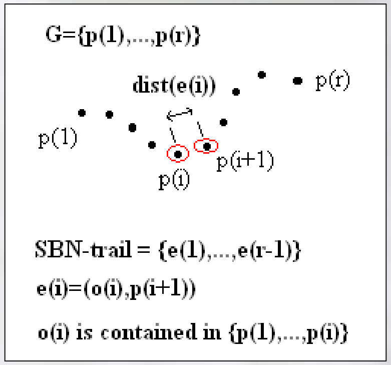
- Definitions…
  - Set based nearest path (SBN-path) – indicates order in which the nearest objects are presented
  - If next item is not unique, impose pre-defined order among its neighbors to break tie



p(i+1) is the nearest neighbor of {p(1),...p(i)} in {p(i+1),...,p(r)}

Connectivity-Based Outlier Factor (COF)

- Definitions…
  - Set based nearest trail (SBN-trail) – a sequence of edges based on the set based nearest path
  - The distances of these edges is called the cost description of the SBN-trail



SBN-trail = {e(1),...,e(r-1)}

e(i)=(o(i),p(i+1))

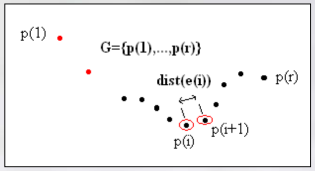
o(i) is contained in {p(1),...,p(i)}

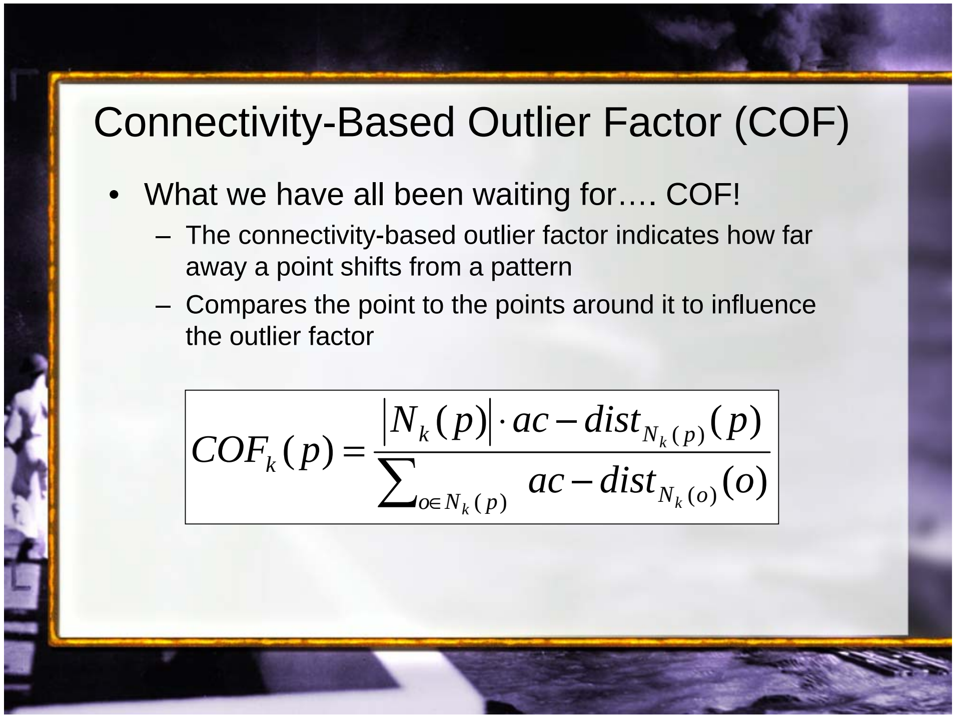
Connectivity-Based Outlier Factor (COF)

- Definitions…
  - Average Chaining Distance = average of the weighted distances in the cost description of the SBN-trail

$$ac-dist_G(p_1) = \frac{1}{r-1} \cdot \sum_{i=1}^{r-1} \frac{2(r-i)}{r} \cdot dist(e_i)$$

  - This means that if edges close to p_i are larger then those further away, then they contribute more to the average chaining distance

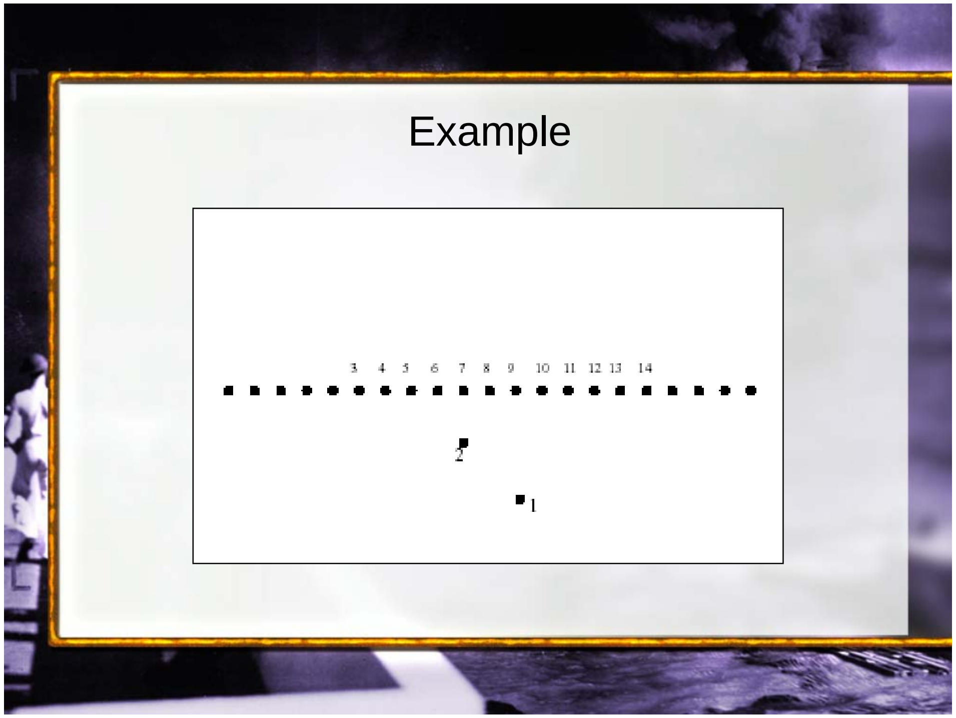




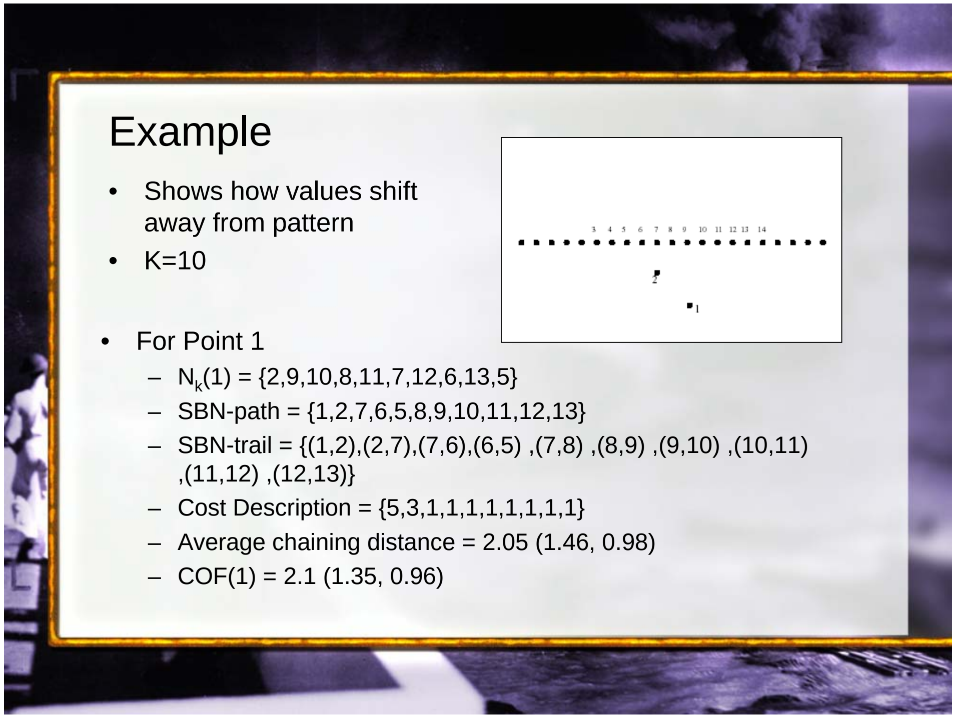
Connectivity-Based Outlier Factor (COF)

- What we have all been waiting for…. COF!
  - The connectivity-based outlier factor indicates how far away a point shifts from a pattern
  - Compares the point to the points around it to influence the outlier factor

$$COF_k(p) = \frac{|N_k(p)| \cdot ac-dist_{N_k(p)}(p)}{\sum_{o \in N_k(p)} ac-dist_{N_k(o)}(o)}$$



Example

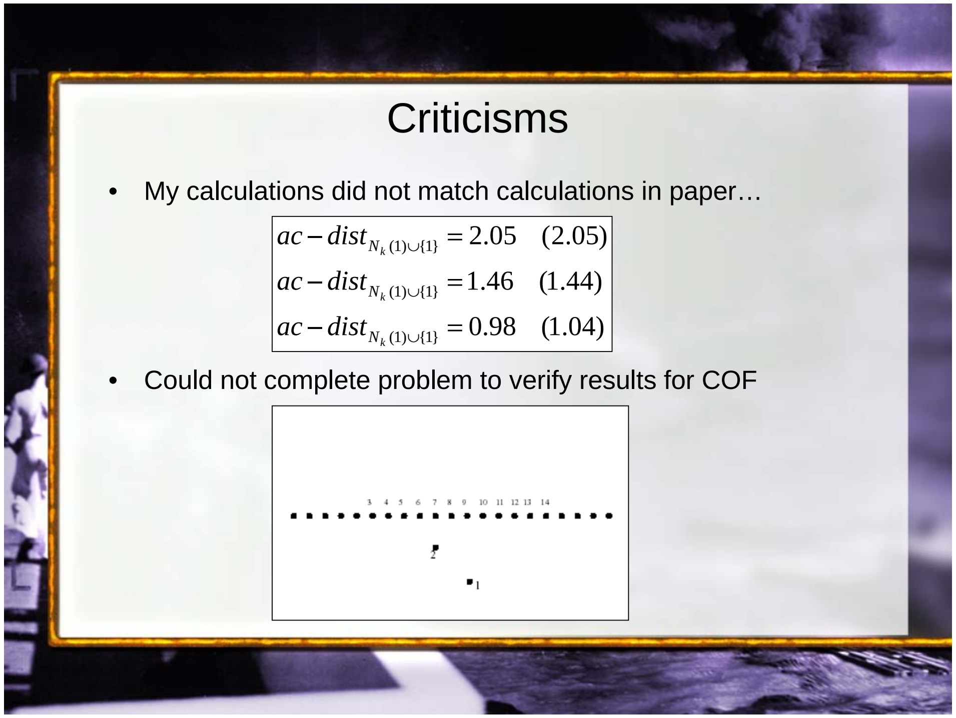


Example

- Shows how values shift away from pattern
- K=10

3 4 5 6 7 8 9 10 11 12 13 14
2
1

- For Point 1
  - $N_k(1)$ = {2,9,10,8,11,7,12,6,13,5}
  - SBN-path = {1,2,7,6,5,8,9,10,11,12,13}
  - SBN-trail = {(1,2),(2,7),(7,6),(6,5) ,(7,8) ,(8,9) ,(9,10) ,(10,11) ,(11,12) ,(12,13)}
  - Cost Description = {5,3,1,1,1,1,1,1,1,1}
  - Average chaining distance = 2.05 (1.46, 0.98)
  - COF(1) = 2.1 (1.35, 0.96)



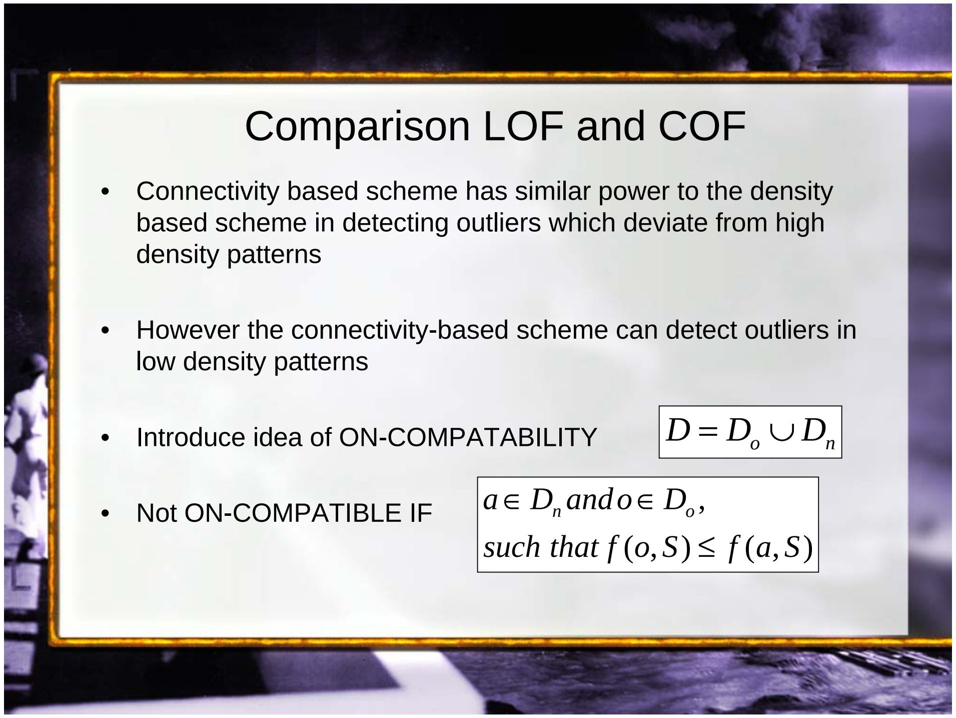
Criticisms

- My calculations did not match calculations in paper…

$$ac-dist_{N_k(1)\cup\{1\}} = 2.05 \quad (2.05)$$
$$ac-dist_{N_k(1)\cup\{1\}} = 1.46 \quad (1.44)$$
$$ac-dist_{N_k(1)\cup\{1\}} = 0.98 \quad (1.04)$$

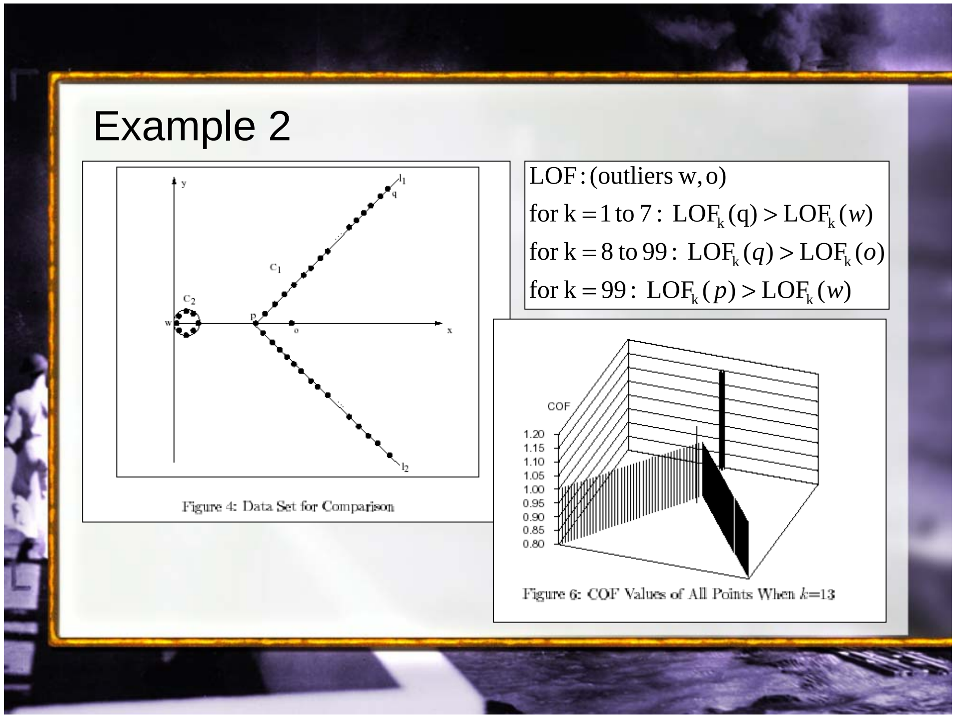
- Could not complete problem to verify results for COF

3 4 5 6 7 8 9 10 11 12 13 14
2
1



Comparison LOF and COF

- Connectivity based scheme has similar power to the density based scheme in detecting outliers which deviate from high density patterns
- However the connectivity-based scheme can detect outliers in low density patterns
- Introduce idea of ON-COMPATABILITY $D = D_o \cup D_n$
- Not ON-COMPATIBLE IF $a \in D_n \, and \, o \in D_o,$ $such \ that \ f(o,S) \leq f(a,S)$



Example 2

Figure 4: Data Set for Comparison

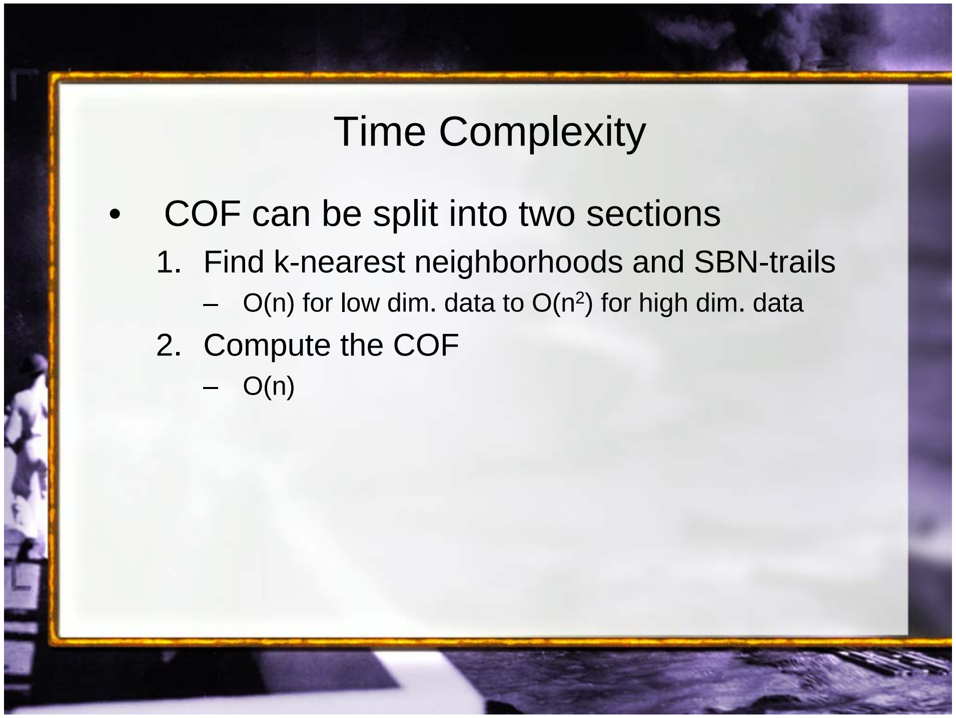
$\mathrm{LOF:(outliers\ w,o)}$

$\mathrm{for\ k=1\ to\ 7:\ LOF_k(q) > LOF_k}(w)$

$\mathrm{for\ k=8\ to\ 99:\ LOF_k}(q) > \mathrm{LOF_k}(o)$

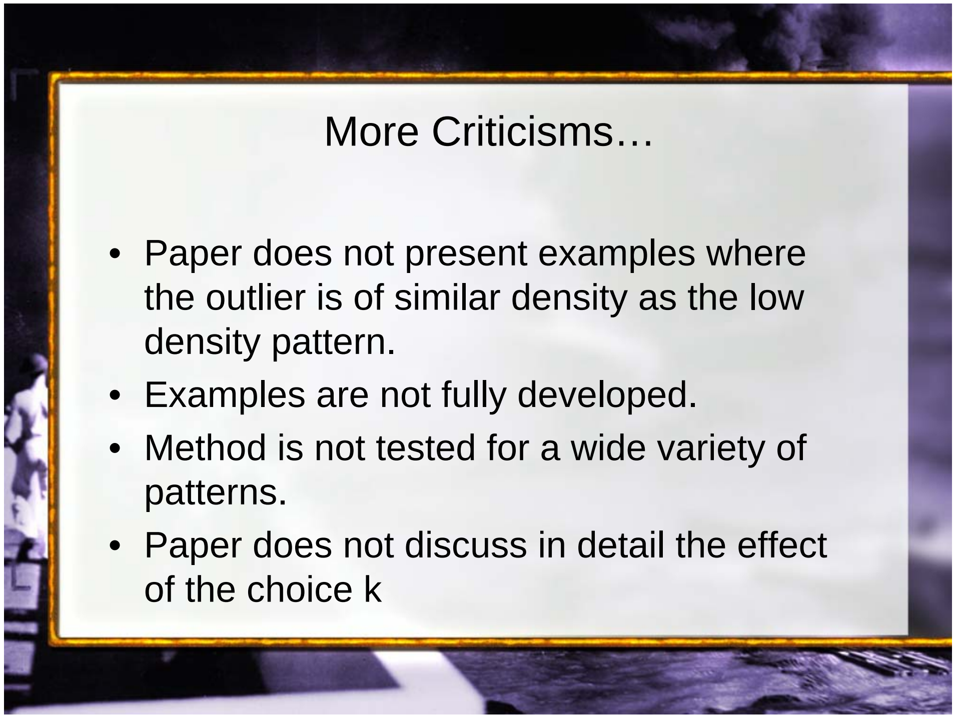
$\mathrm{for\ k=99:\ LOF_k}(p) > \mathrm{LOF_k}(w)$

Figure 6: COF Values of All Points When k=13



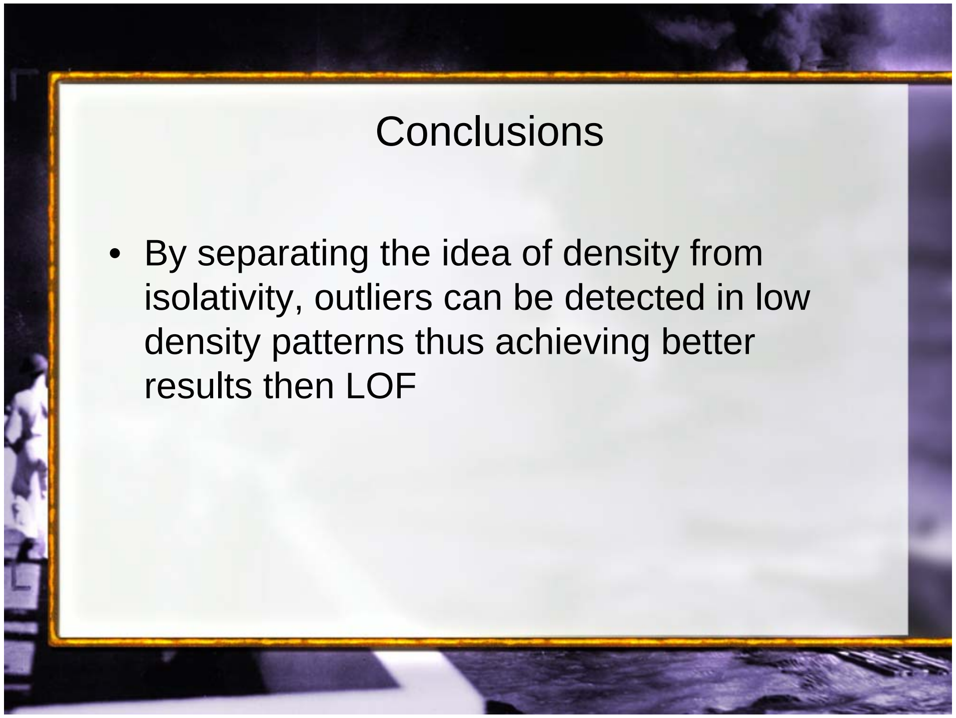
Time Complexity

- COF can be split into two sections
  1. Find k-nearest neighborhoods and SBN-trails
     – O(n) for low dim. data to O(n^2) for high dim. data
  2. Compute the COF
     – O(n)



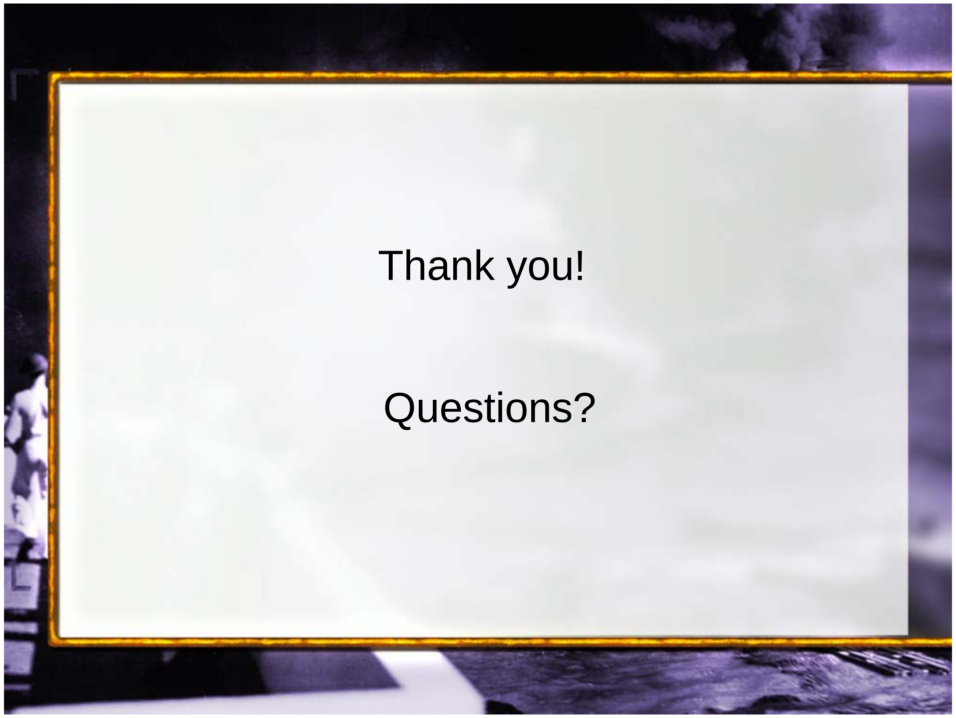
More Criticisms…

- Paper does not present examples where the outlier is of similar density as the low density pattern.
- Examples are not fully developed.
- Method is not tested for a wide variety of patterns.
- Paper does not discuss in detail the effect of the choice k



Conclusions

- By separating the idea of density from isolativity, outliers can be detected in low density patterns thus achieving better results then LOF



Thank you!

Questions?