Clustering Stream Data by Exploring the Evolution of Density Mountain

Shufeng Gong, Yanfeng Zhang, Ge Yu

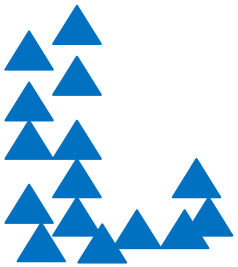
Northeastern University, China

Outline

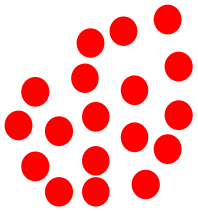
- Motivation
- EDMStream: Basic Idea
- EDMStream: Detail
- Evolution

Clustering

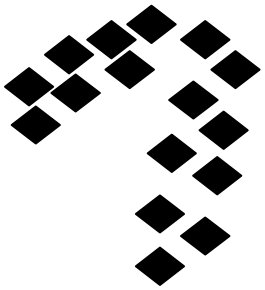
- Group the data on the basis of their similarity



cluster1



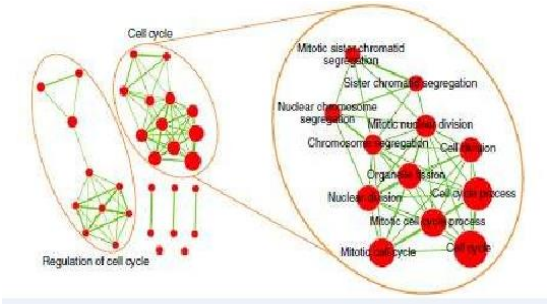
cluster2



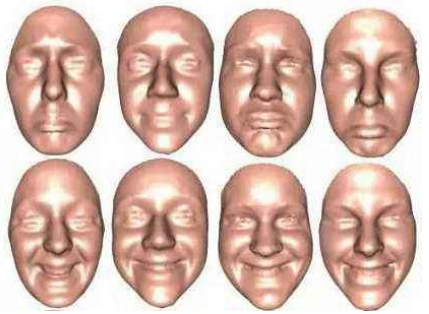
cluster3

- Clustering is widely used in many applications

intelligent business



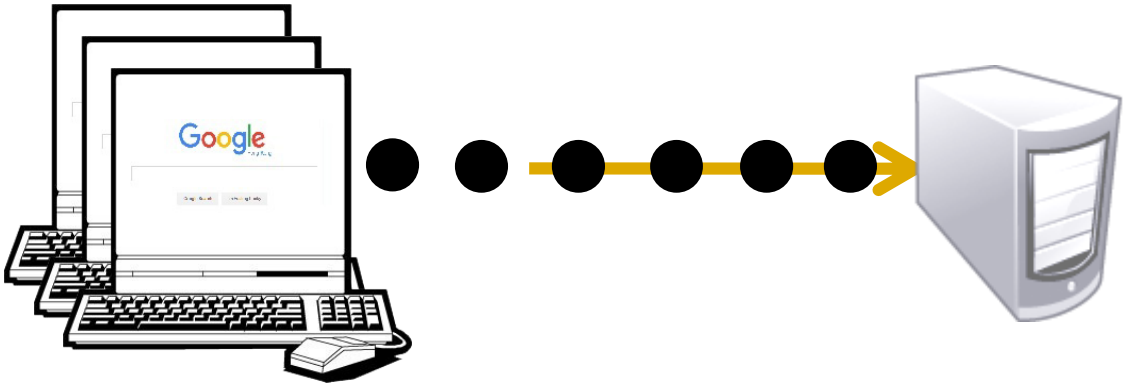
gene expression



pattern recognition

Stream & Stream Clustering

- A sequence of data points with timestamp information $p_1^{t1}, p_2^{t2} \ldots, p_N^{tN}$,where $N \rightarrow \infty$, *i.e. news tweets*.



- Group stream data on the basis of their similarity.

Challenges

- How to incrementally update clusters efficiently?
- How to track the evolutions of clusters?

Related Work

- Stream clustering
  - Clustream
  - DenStream
  - D-Stream
  - MR-Stream

Related Work

- Stream clustering
  - Clustream
  - DenStream
  - D-Stream
  - MR-Stream

update clusters online 

track evolution 

Related Work

- Stream clustering
  - Clustream
  - DenStream
  - D-Stream
  - MR-Stream
- Track evolution
  - MONIC
  - MEC
  - TRACDS

update clusters online 

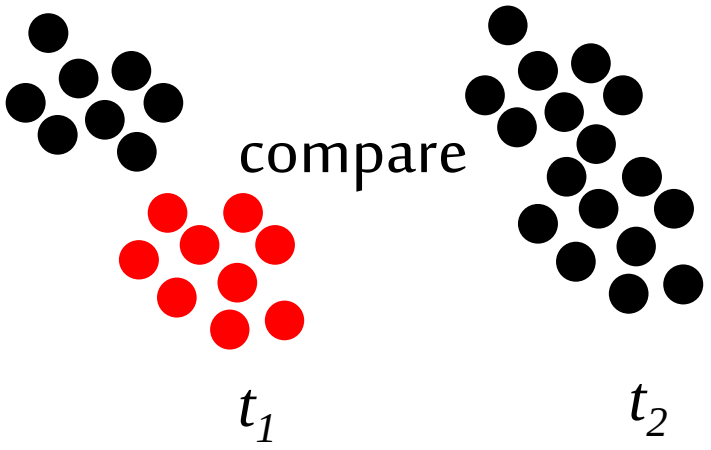
track evolution 

Related Work

- Stream clustering
  - Clustream
  - DenStream
  - D-Stream
  - MR-Stream
- Track evolution
  - MONIC
  - MEC
  - TRACDS

update clusters online ✗

track evolution ✗



Outline

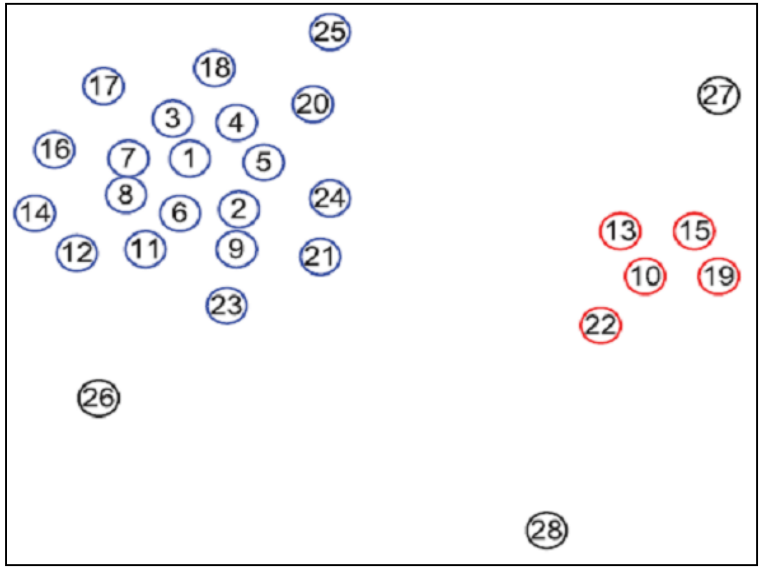
- Motivation
- EDMStream: Basic Idea
- EDMStream: Detail
- Evolution

EDMStream

Clustering: DPClustering[1]
Update online: DP-Tree
Track Evolution: Density Mountain

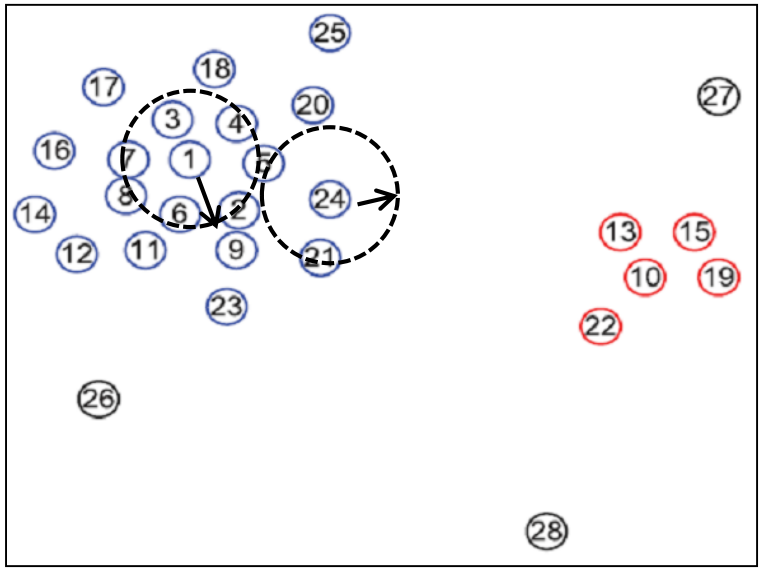
1. Rodriguez, Alex, and Alessandro Laio. "Clustering by fast search and find of density peaks." *Science* 344.6191 (2014): 1492-1496.

DPClustering



plane view

DPClustering

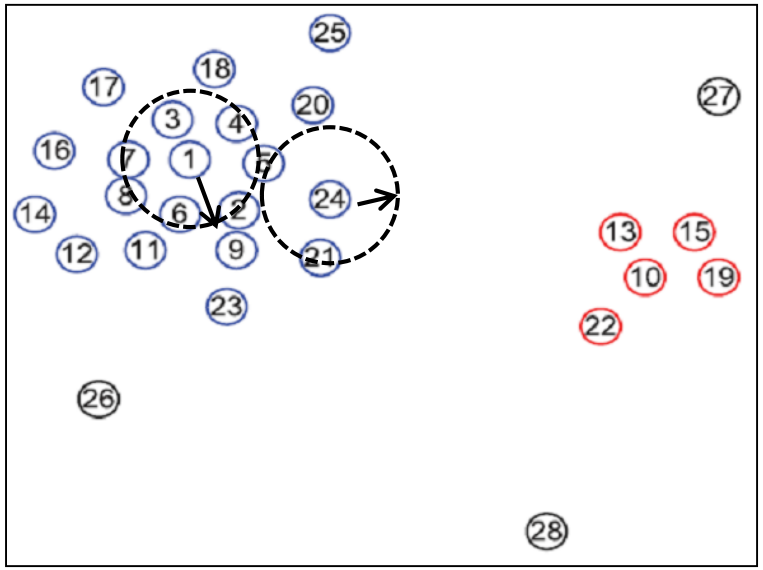


plane view

$$\rho_i = \sum_j \chi(d_{ij} - d_C)$$

$$\chi(x) = 0, if \;\; x < 0,$$
$$\chi(x) = 1, if \;\; x >= 0;$$

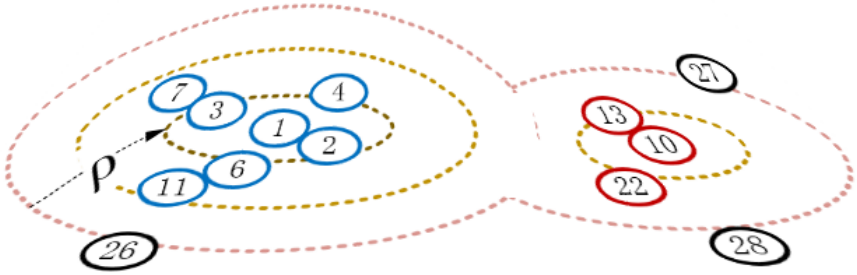
DPClustering



plane view

$$\rho_i = \sum_j \chi(d_{ij} - d_C)$$

$$\chi(x) = 0, if \;\; x < 0,$$
$$\chi(x) = 1, if \;\; x >= 0;$$

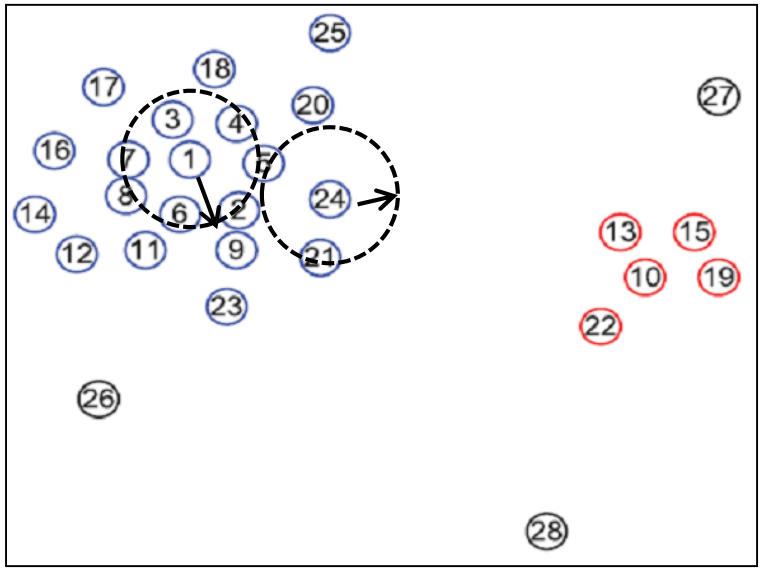


density contour

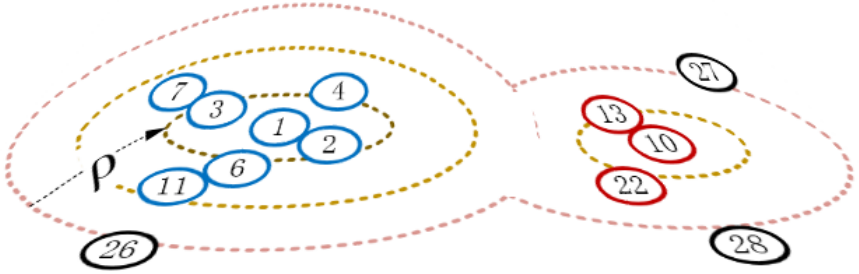
mountain ⟷ cluster

mountain peak ⟷ density peak

DPClustering



plane view



density contour

$$\rho_i = \sum_j \chi(d_{ij} - d_c)$$

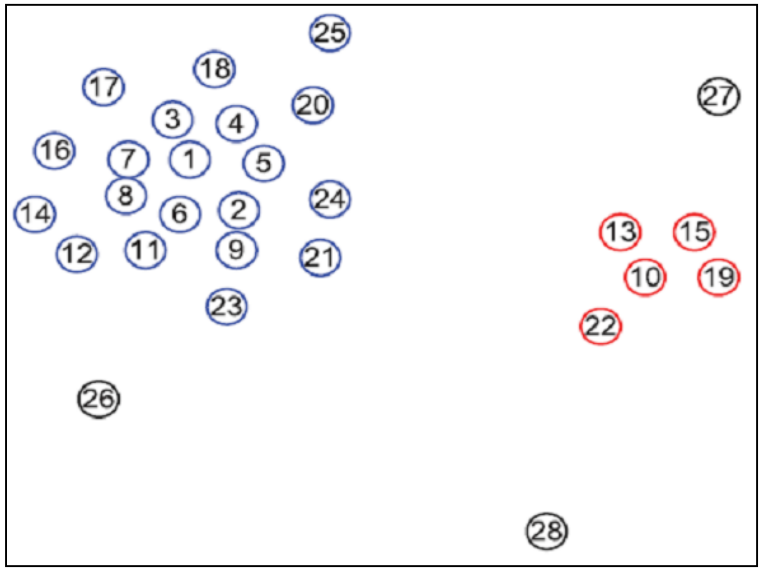
$$\chi(x) = 0, if \;\; x < 0,$$
$$\chi(x) = 1, if \;\; x >= 0;$$

How to find the density peaks and clusters?

mountain ⟷ cluster

mountain peak ⟷ density peak

DPClustering

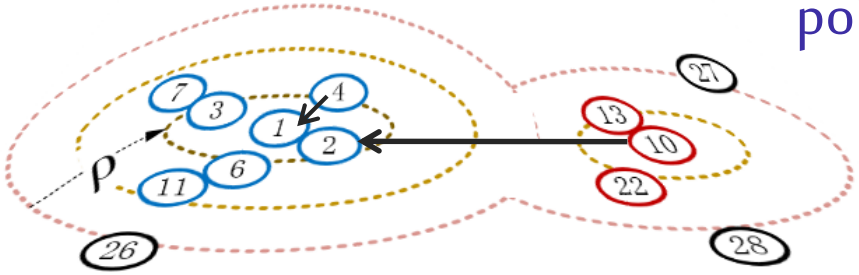


plane view

$$\rho_i = \sum_j \chi(d_{ij} - d_C) \qquad \chi(x) = 0, if \;\; x < 0, \qquad \chi(x) = 1, if \;\; x >= 0;$$

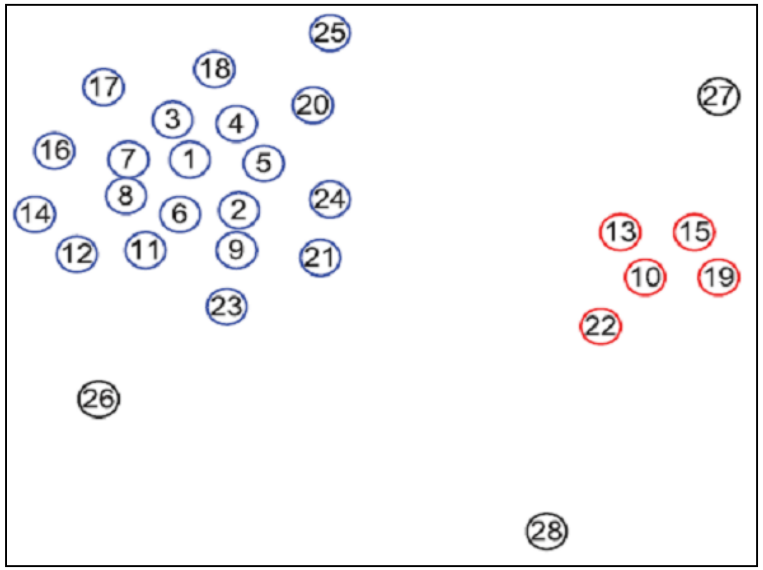
$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

the dependent distance of point 10 is |10,2|
point 2 is the dependent point of point 10



density contour

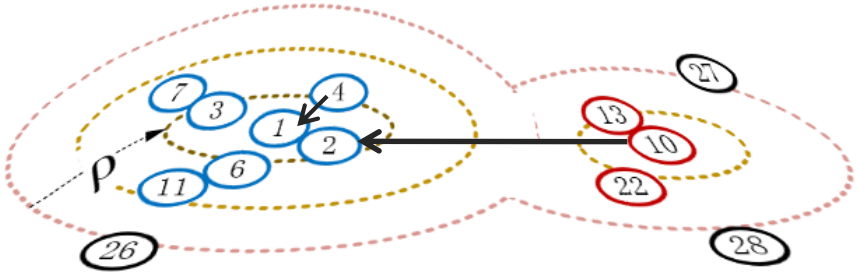
DPClustering



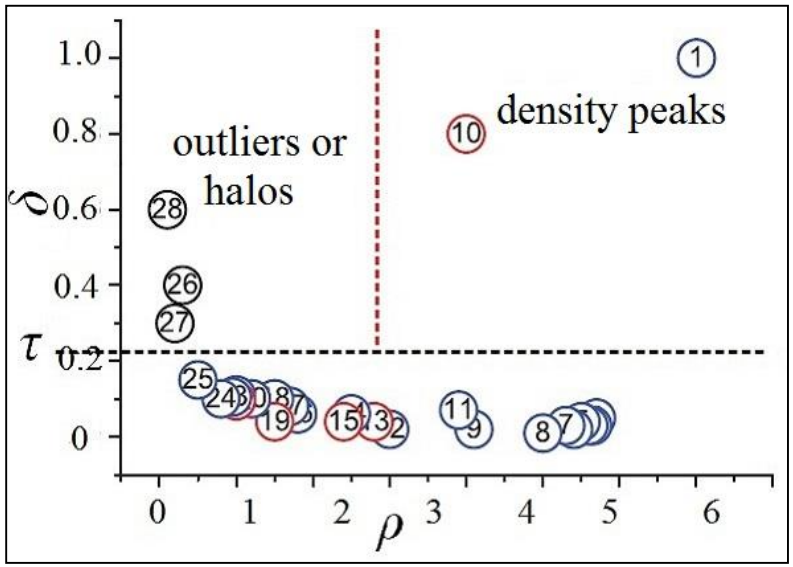
plane view

$$\rho_i = \sum_j \chi(d_{ij} - d_C) \qquad \chi(x) = 0, if \ \ x < 0, \quad \chi(x) = 1, if \ \ x >= 0;$$

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

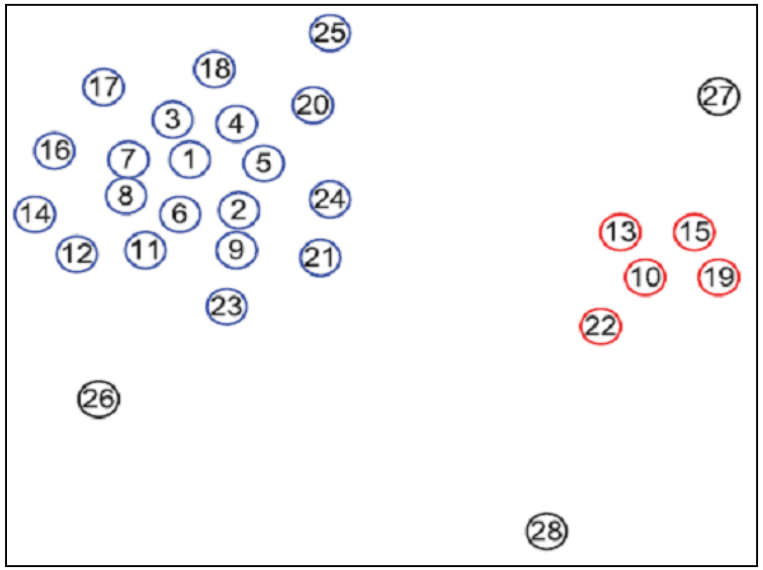


density contour



decision graph

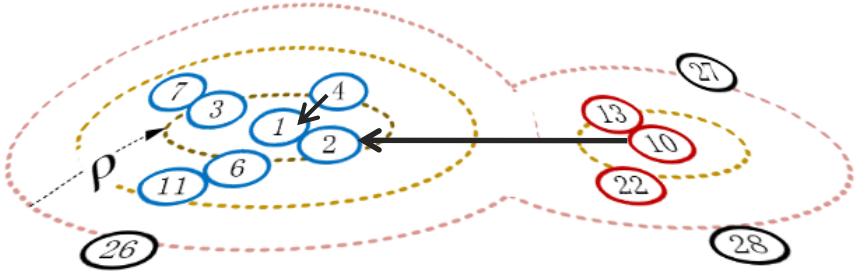
DPClustering



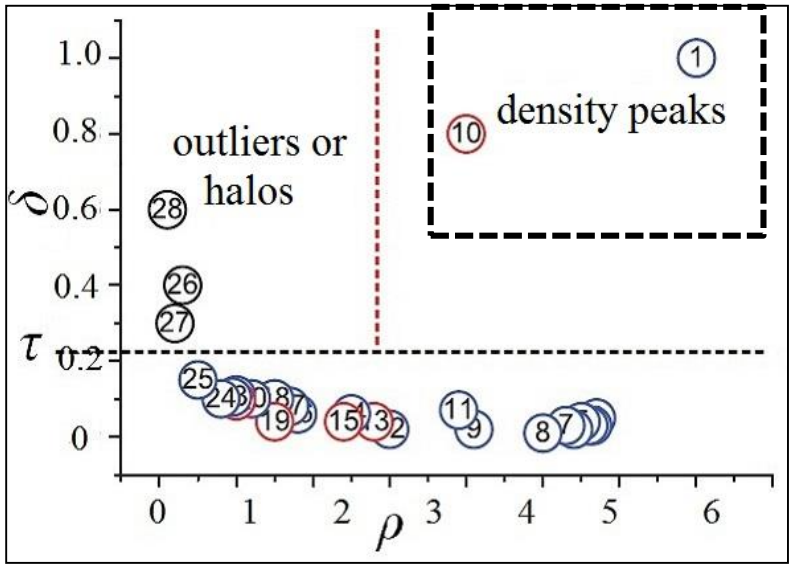
plane view

$$\rho_i = \sum_j \chi(d_{ij} - d_C) \qquad \chi(x) = 0, if \;\; x < 0, \quad \chi(x) = 1, if \;\; x >= 0;$$

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

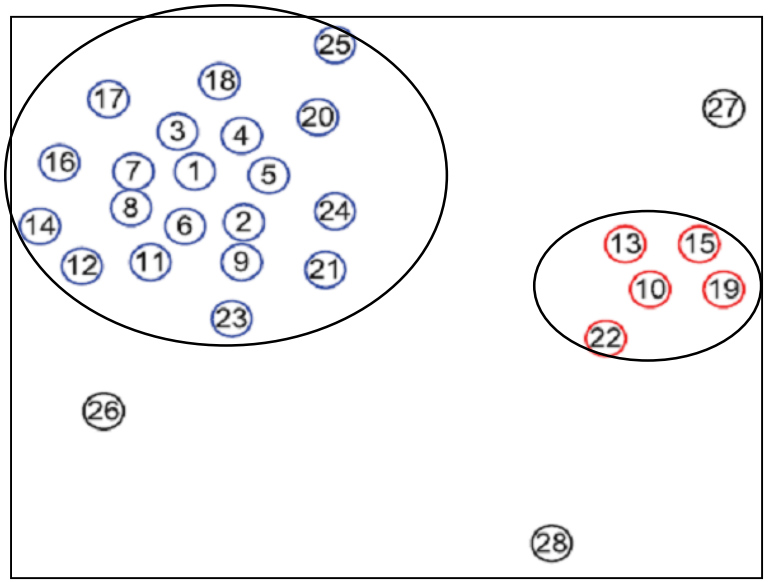


density contour



decision graph

DPClustering



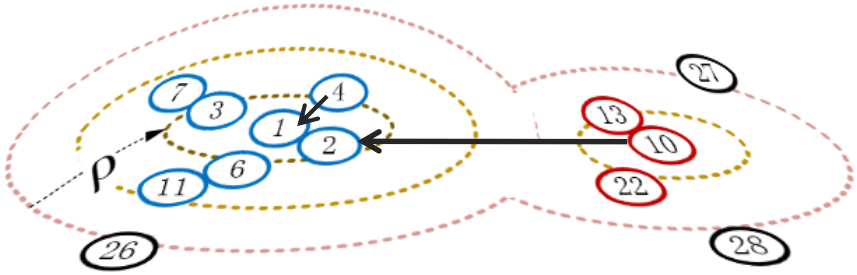
plane view

$$\rho_i = \sum_j \chi(d_{ij} - d_C)$$

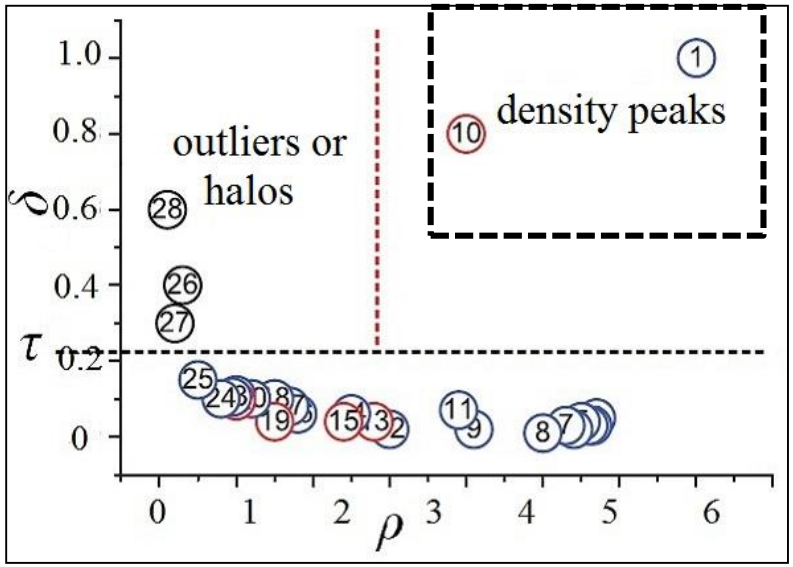
$$\chi(x) = 0, \text{ if } x < 0,$$

$$\chi(x) = 1, \text{ if } x >= 0;$$

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

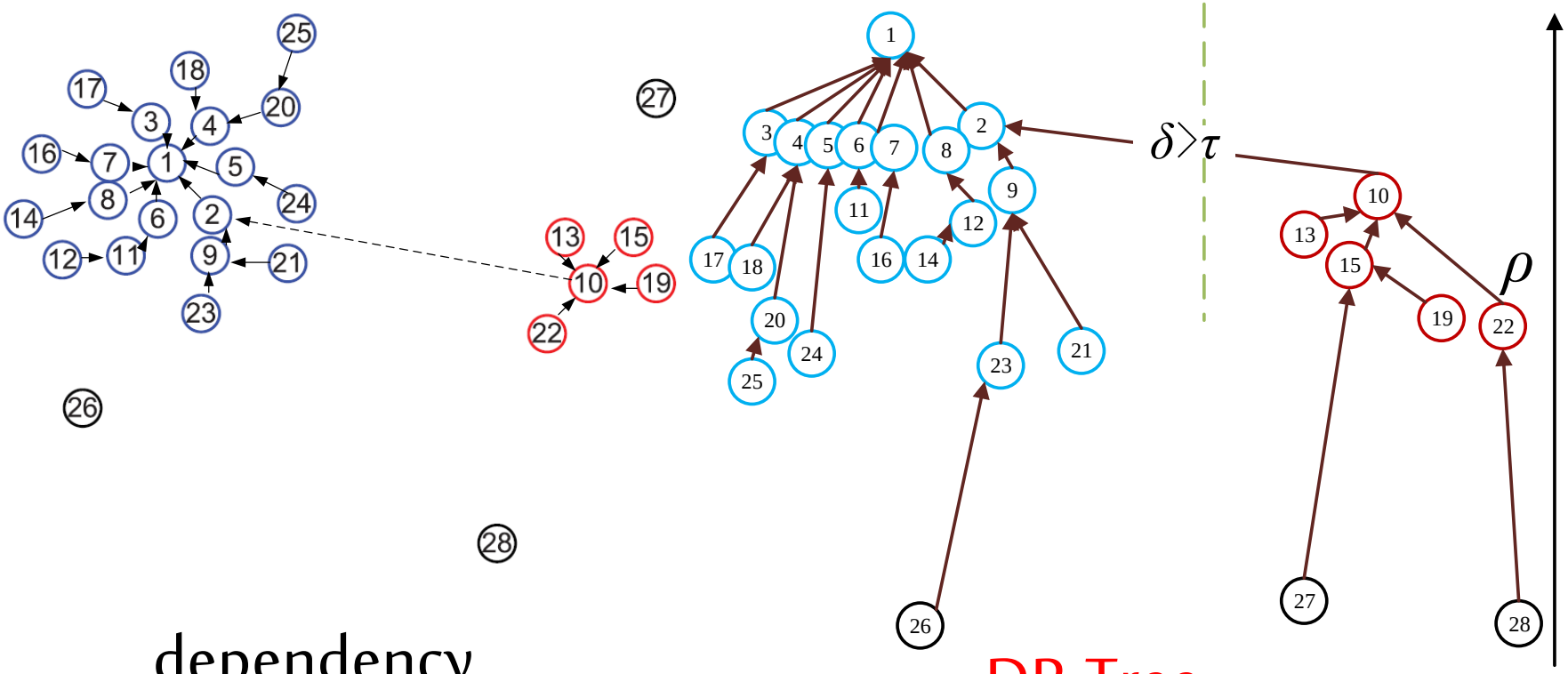


density contour

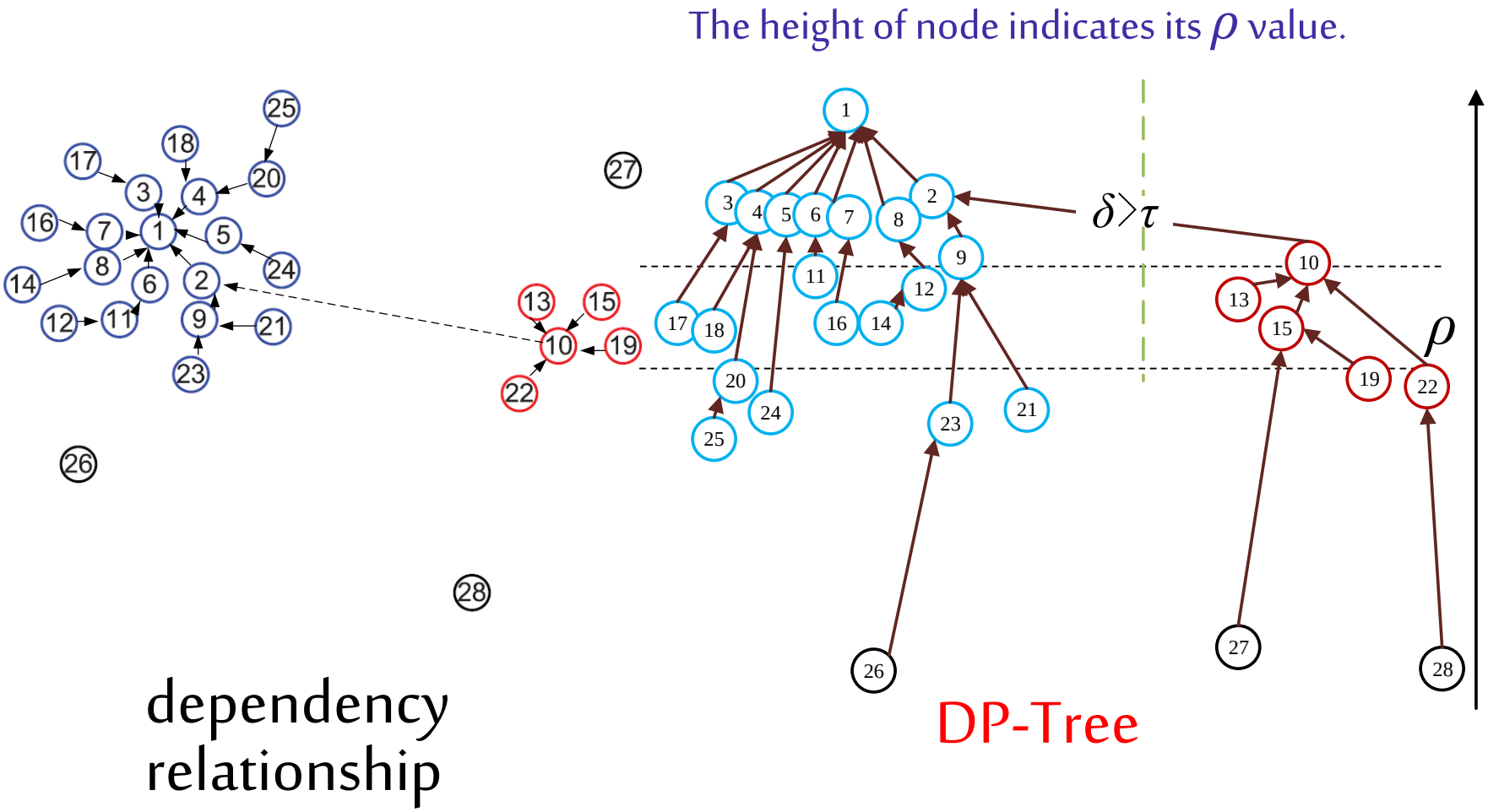


decision graph

DP-Tree

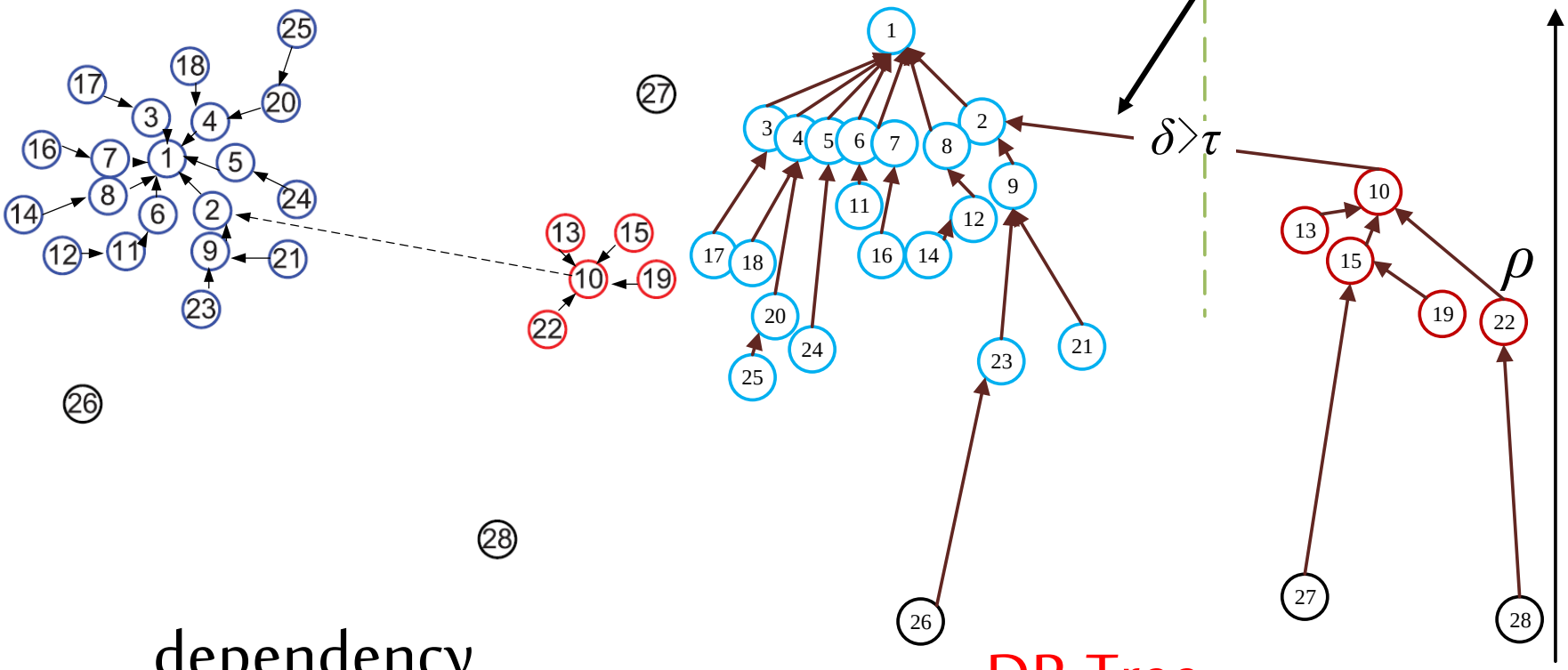


DP-Tree


The height of node indicates its ρ value.
δ>τ
ρ
dependency relationship
DP-Tree

DP-Tree

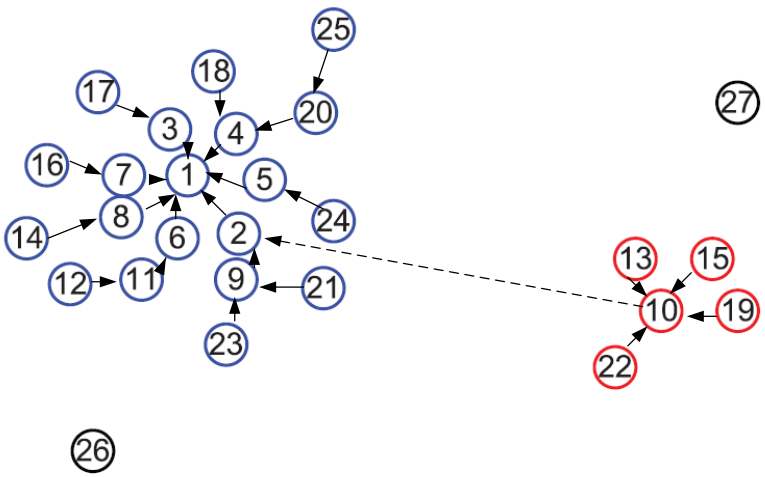
The length of a directed link indicates the source point's δ value.



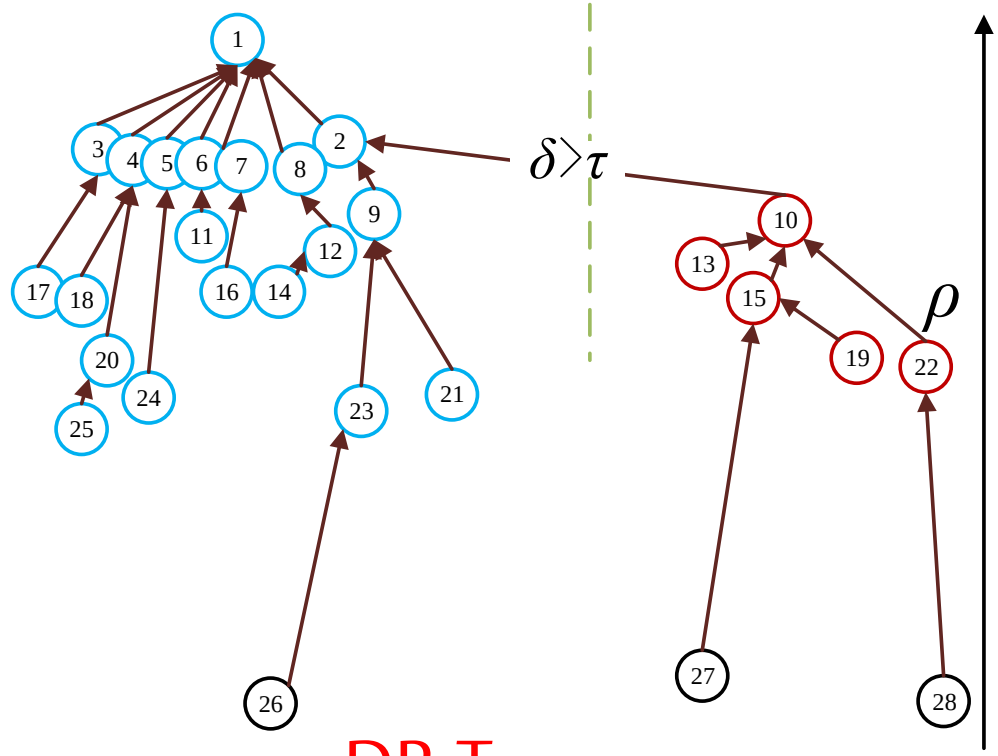
dependency relationship

DP-Tree

The changes of a point's δ and its dependent point lead to changes of clustering results.



dependency relationship

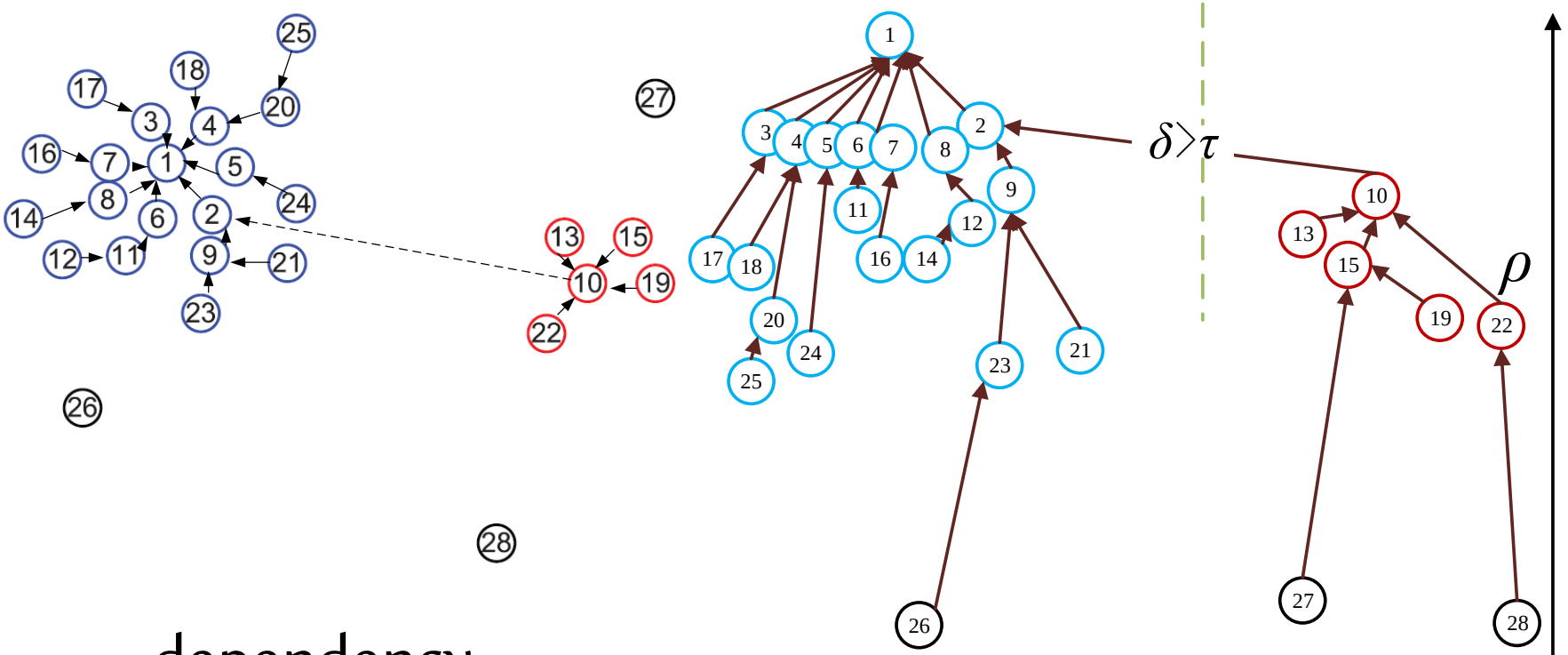


DP-Tree

DP-Tree

update online

we can update clustering results online by updating the relationship between points



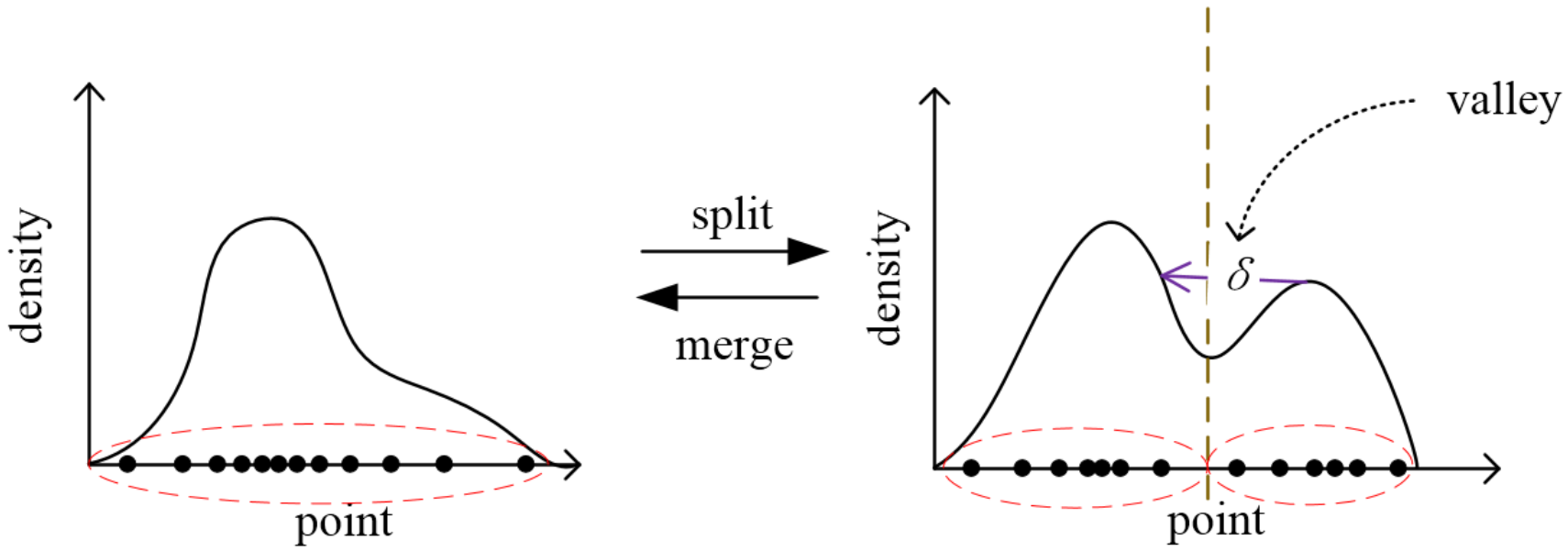
Density Mountain

When clusters evolve, we can capture evolution by monitoring the evolution of density mountain.

Density Mountain

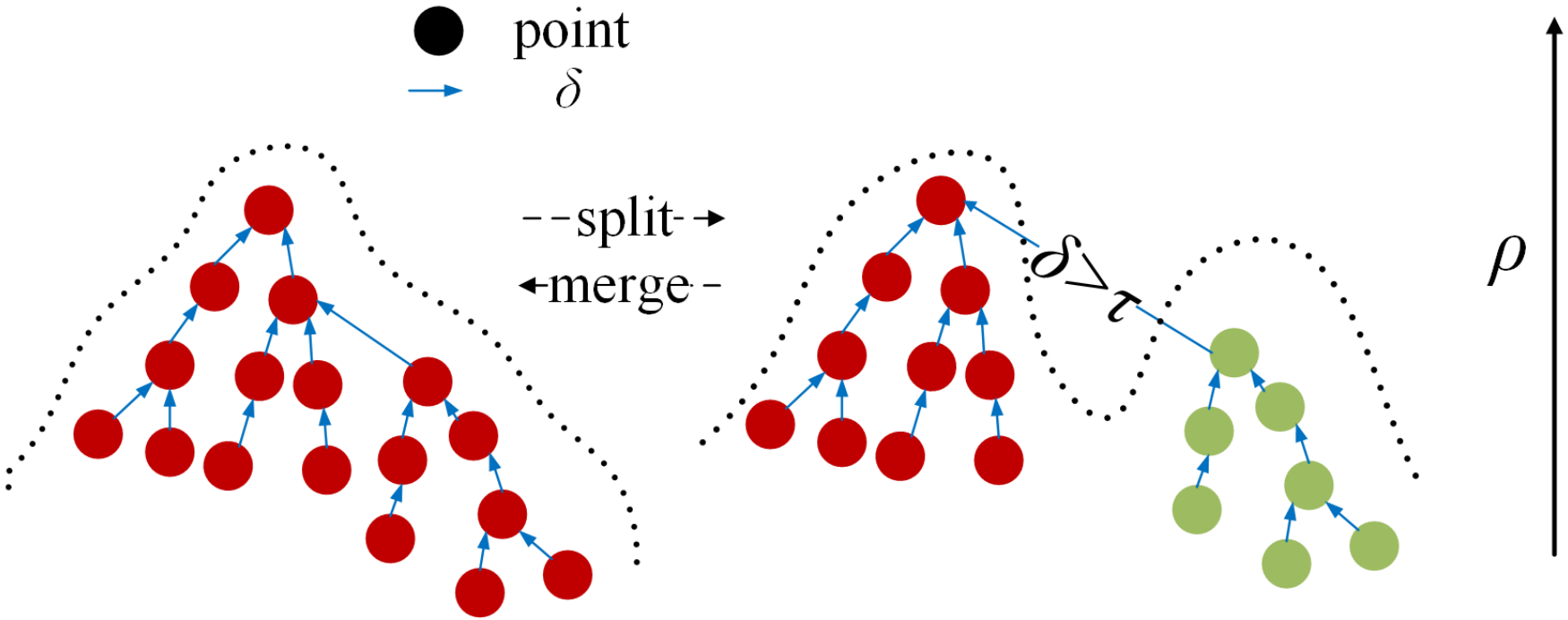
Points are in one dimensional space.

The curve that depicts point's density looks like mountain, where density peak is mountain peak.



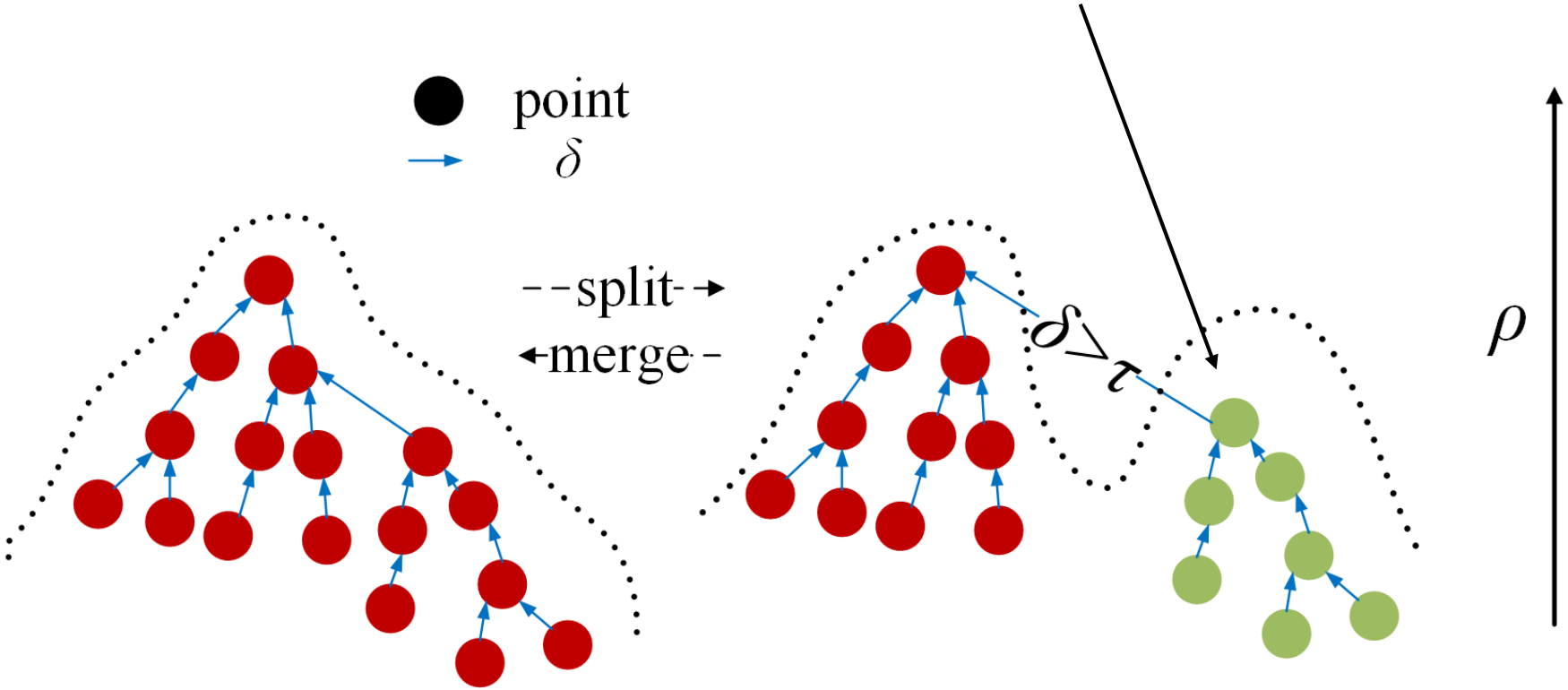
Density Mountain

We use DP-Tree to abstract density mountain.



Density Mountain

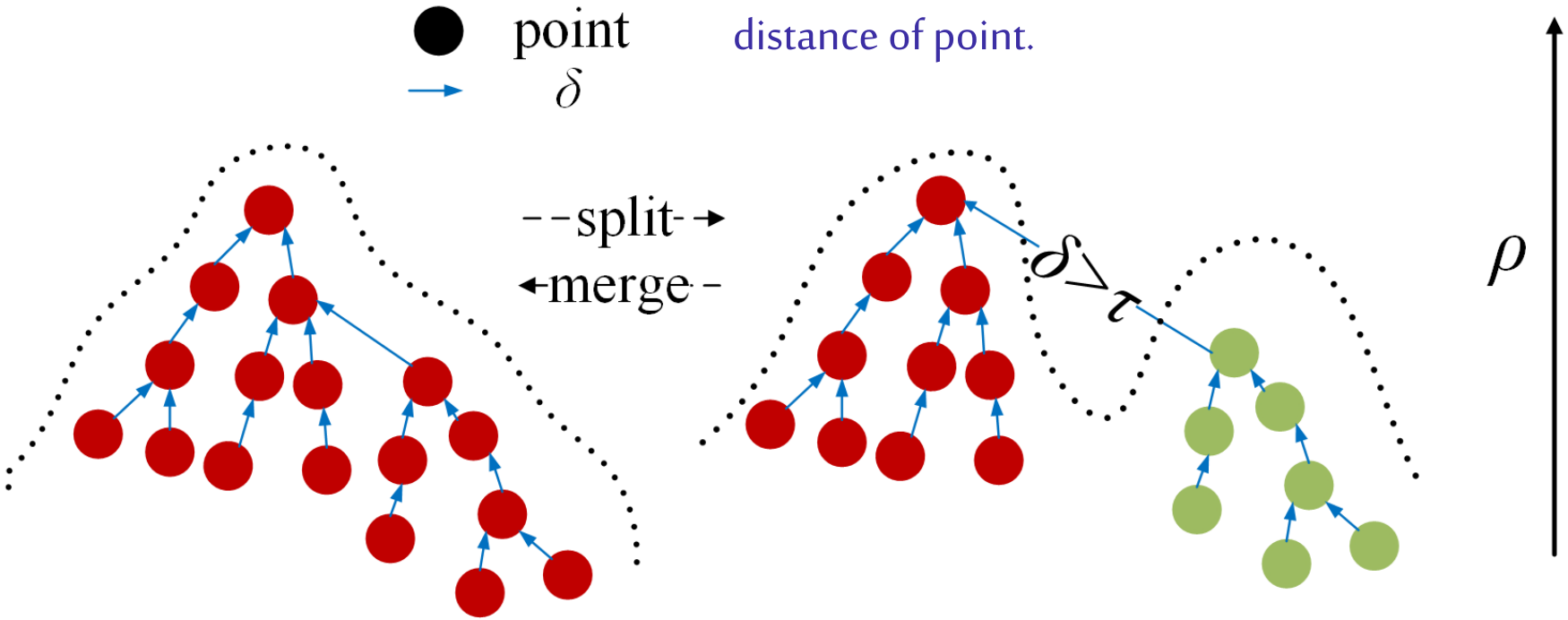
The subtree whose root's dependent distance is larger than threshold τ can be treated as a cluster.



Density Mountain

track evolution

we can track evolution of cluster in realtime by monitoring the evolution of dependent distance of point.



Outline

- Motivation
- EDMStream: Basic Idea
- EDMStream: Detail
- Evolution

Basic conception

The recent information from a stream reflects the emerging of new trends

Basic conception

The recent information from a stream reflects the emerging of new trends

- Decay function $f_i^t = a^{\lambda(t-t_i)}, a^{\lambda} < 1$

  t_i is the arrival time of point i
  t is the current time.

Basic conception

The recent information from a stream reflects the emerging of new trends

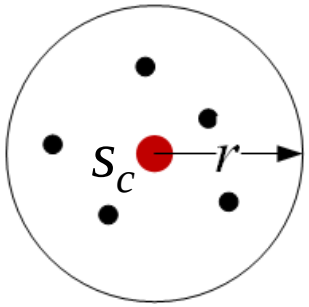
- Decay function $f_i^t = a^{\lambda(t-t_i)}, a^{\lambda} < 1$

D-Stream, DenStream....

Basic conception

It is difficult to maintain all streaming data in memory. We summarize the stream by cluster-cell (basic operation and storage unit).

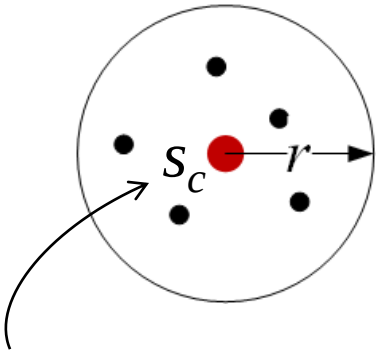
- Cluster-cell



Basic conception

It is difficult to maintain all streaming data in memory. We summarize the stream by cluster-cell (basic operation and storage unit).

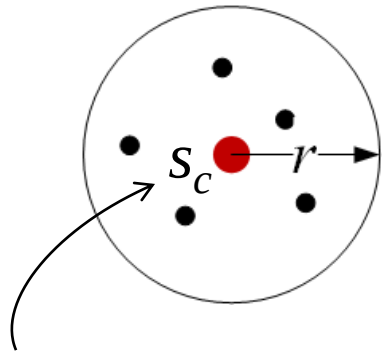
- Cluster-cell



Basic conception

It is difficult to maintain all streaming data in memory. We summarize the stream by cluster-cell (basic operation and storage unit).

- Cluster-cell

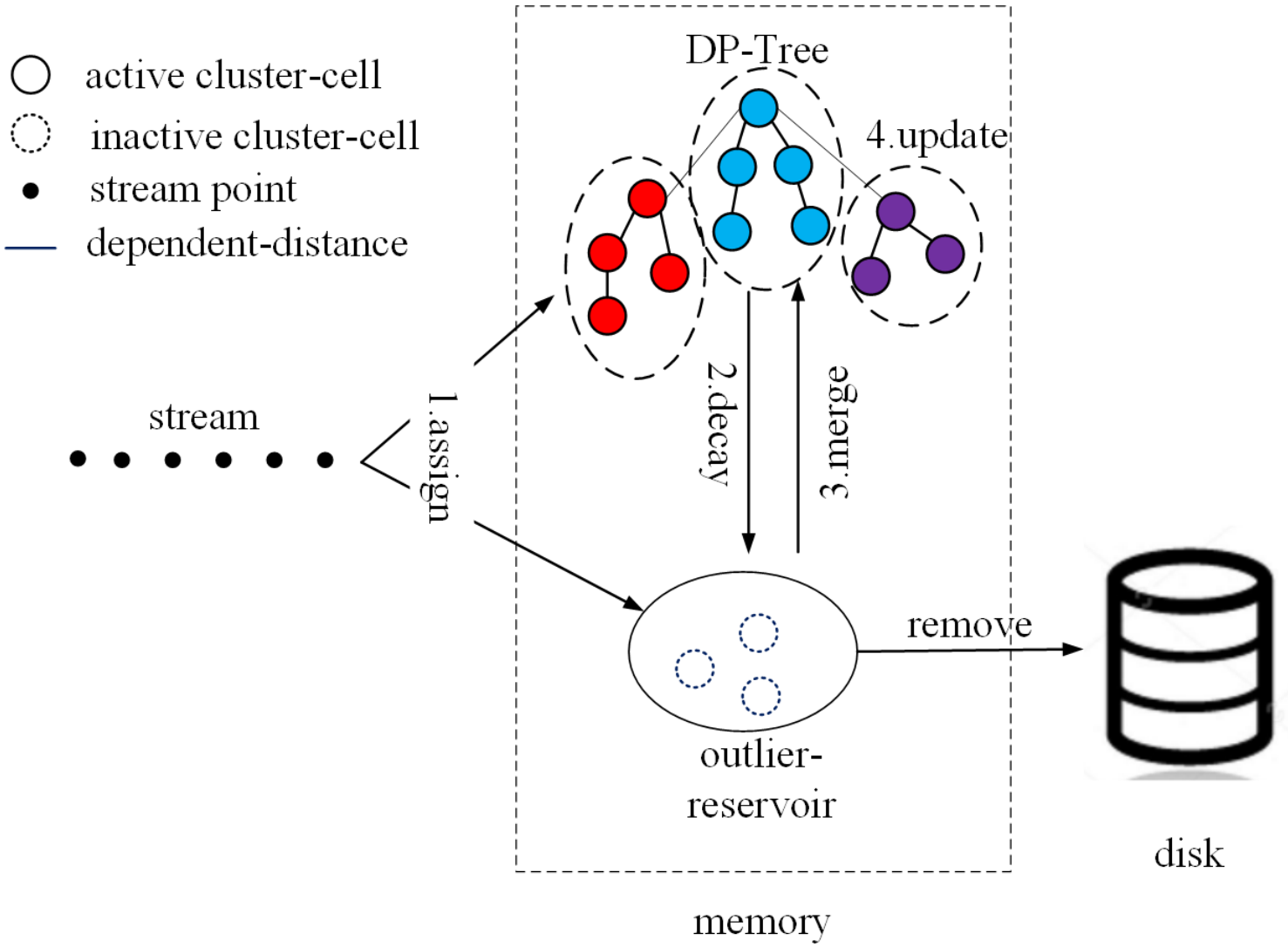


seed of cluster-cell

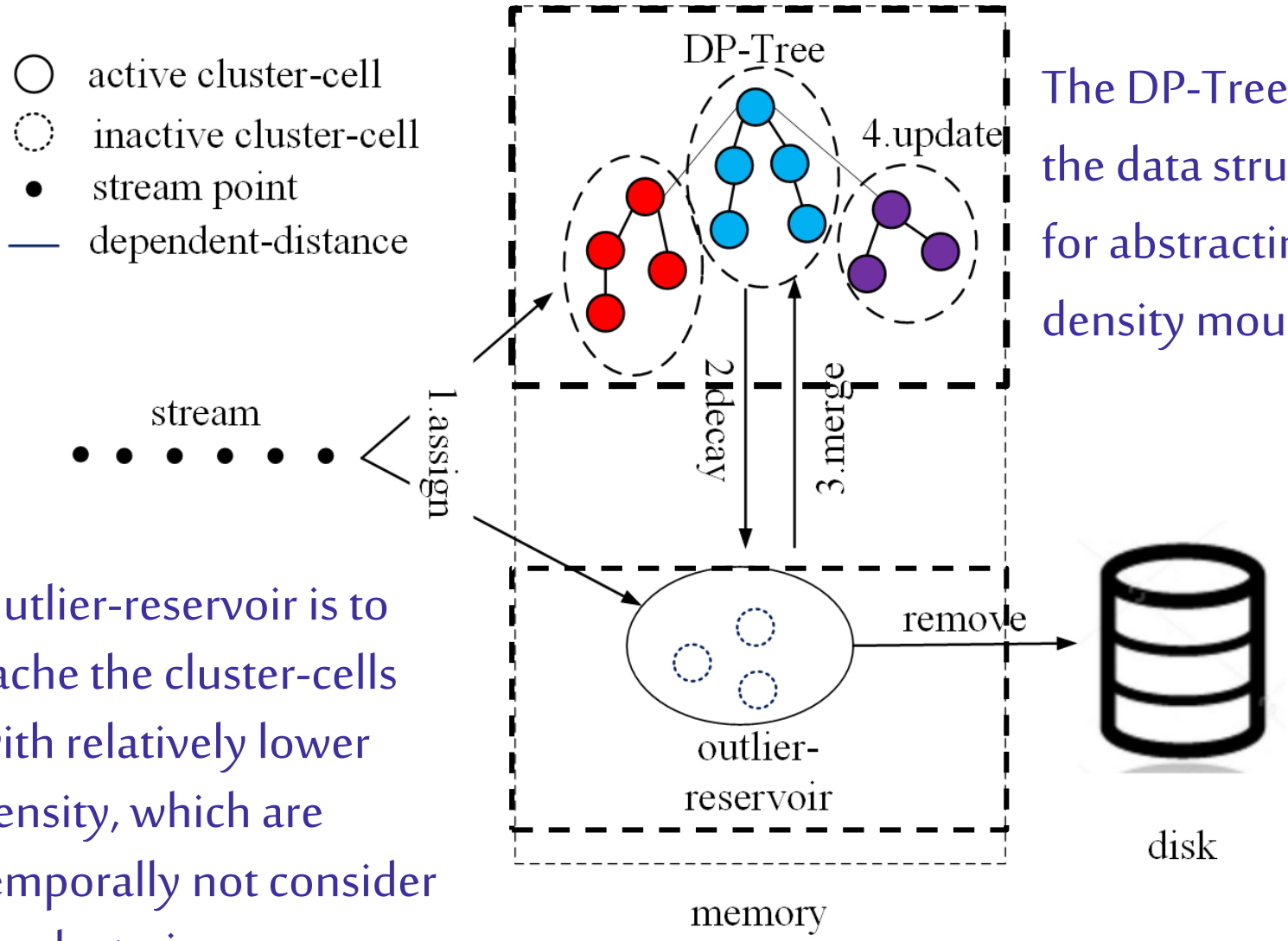
$$\rho_c^t = \sum_{p_i : |p_i, s_c| \leq r} f_i^t$$

$$\delta_c^t = \min_{c' : \rho_{c'}^t > \rho_c^t} \left(|s_c, s_{c'}| \right)$$

Overview


active cluster-cell
inactive cluster-cell
stream point
dependent-distance
stream
1.assign
DP-Tree
4.update
2.decay
3.merge
outlier-
reservoir
remove
disk
memory

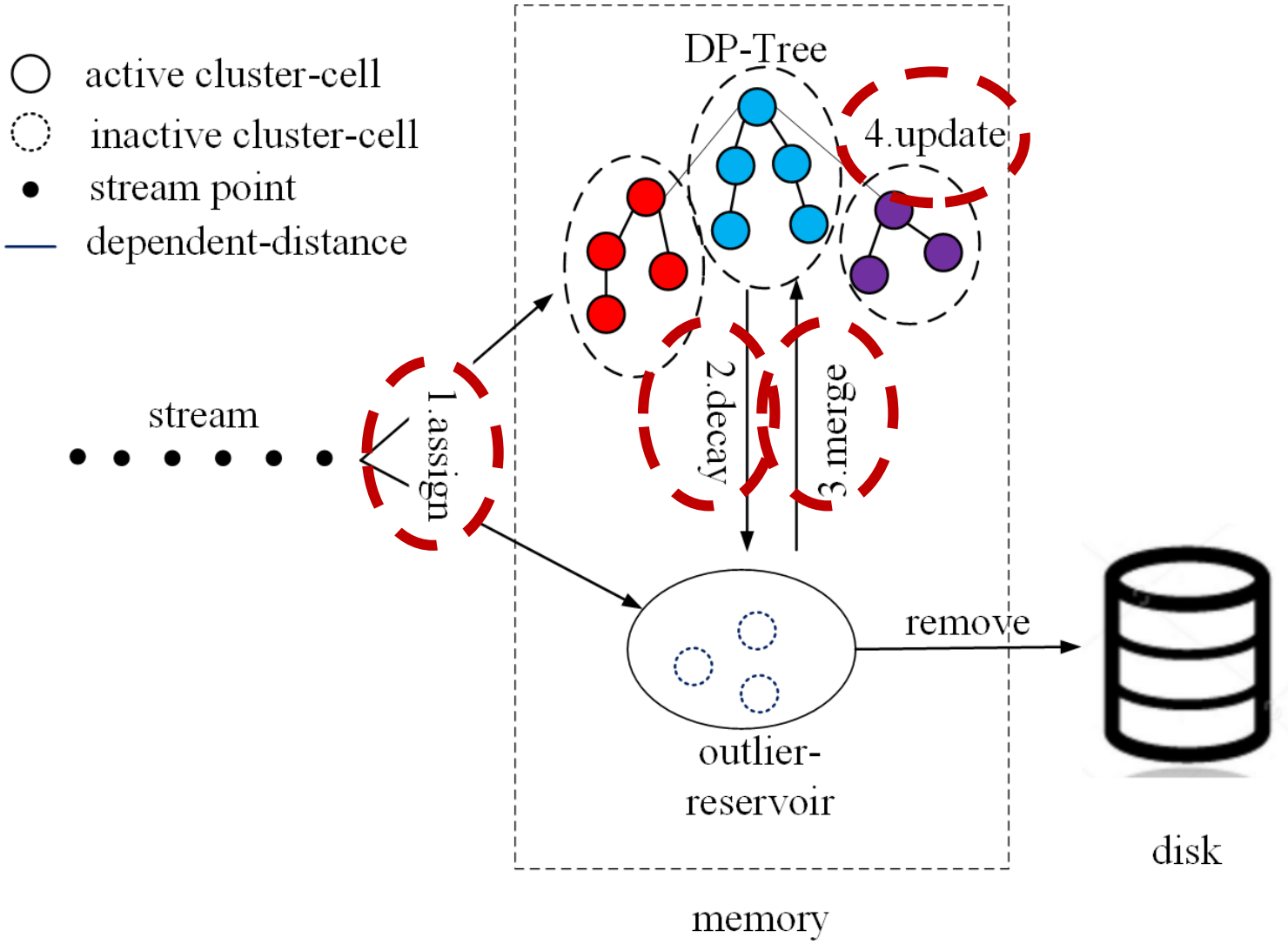
Overview



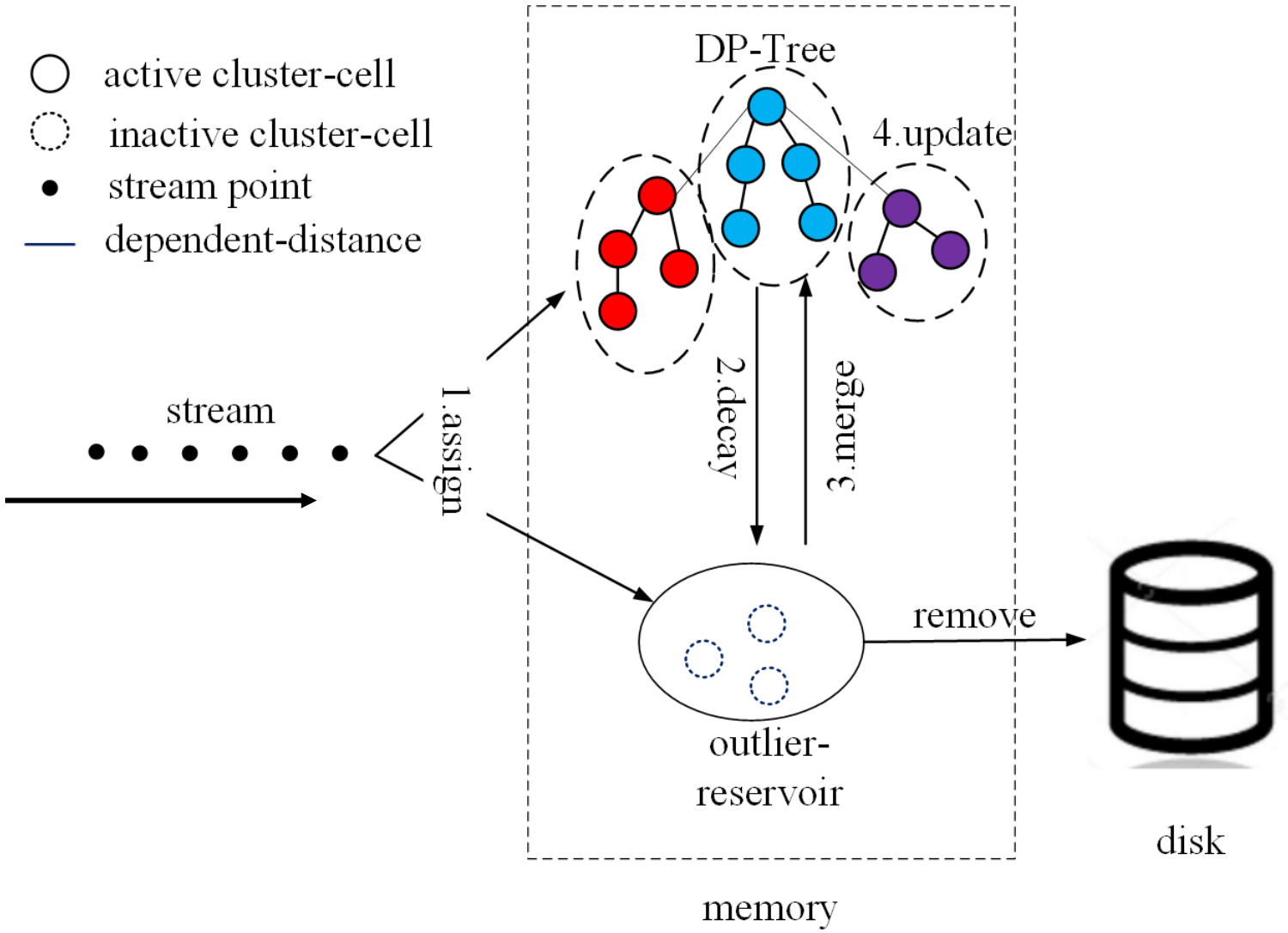
The DP-Tree is the data structure for abstracting density mountain

Outlier-reservoir is to cache the cluster-cells with relatively lower density, which are temporally not consider for clustering

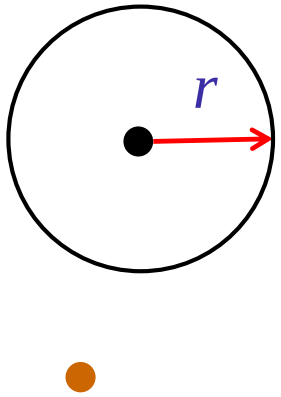
Overview


active cluster-cell
inactive cluster-cell
stream point
dependent-distance
DP-Tree
4.update
stream
1.assign
2.decay
3.merge
remove
outlier-
reservoir
disk
memory

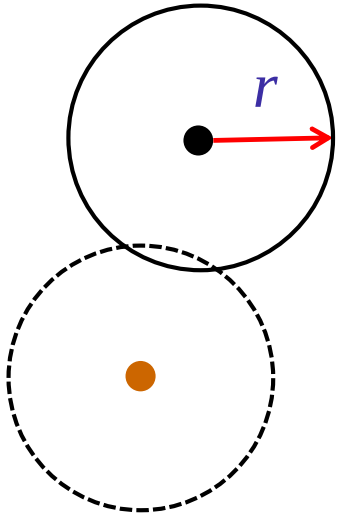
Overview


active cluster-cell
inactive cluster-cell
stream point
dependent-distance
stream
1.assign
DP-Tree
4.update
2.decay
3.merge
outlier-
reservoir
remove
disk
memory

Assign

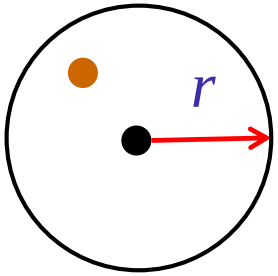


Assign



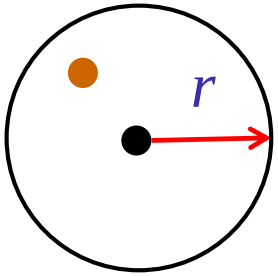
Update

- Update ρ



Update

- Update ρ

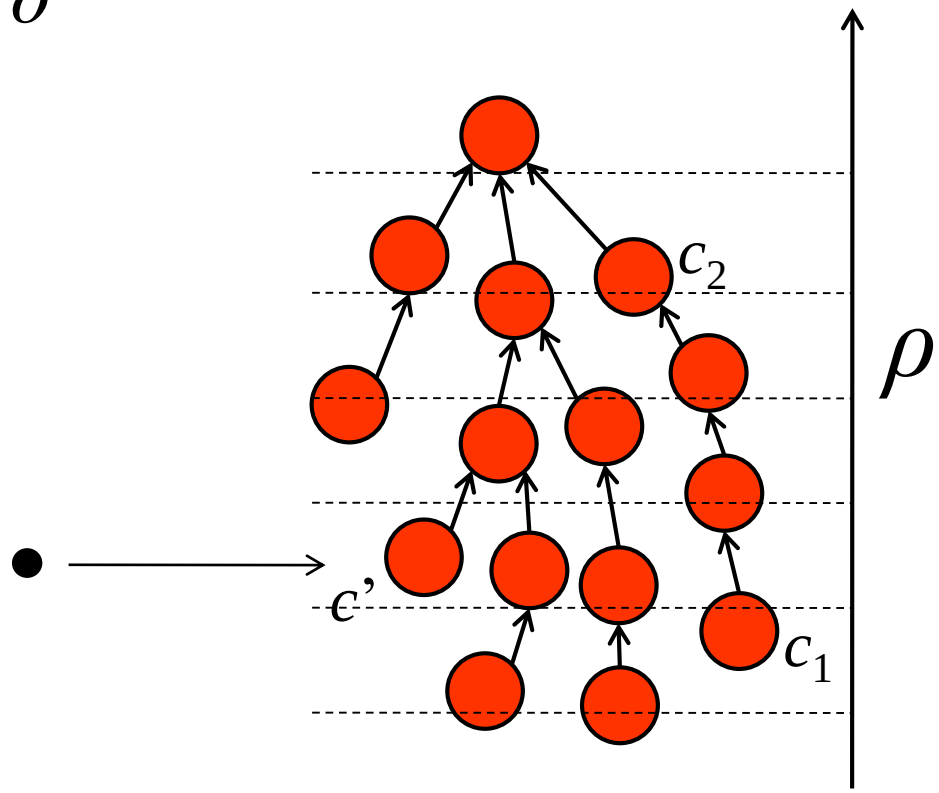


Density will be updated

$$\rho_c^{t_{j+1}} = a^{\lambda(t_{j+1}-t_j)}\rho_c^{t_j} + 1$$

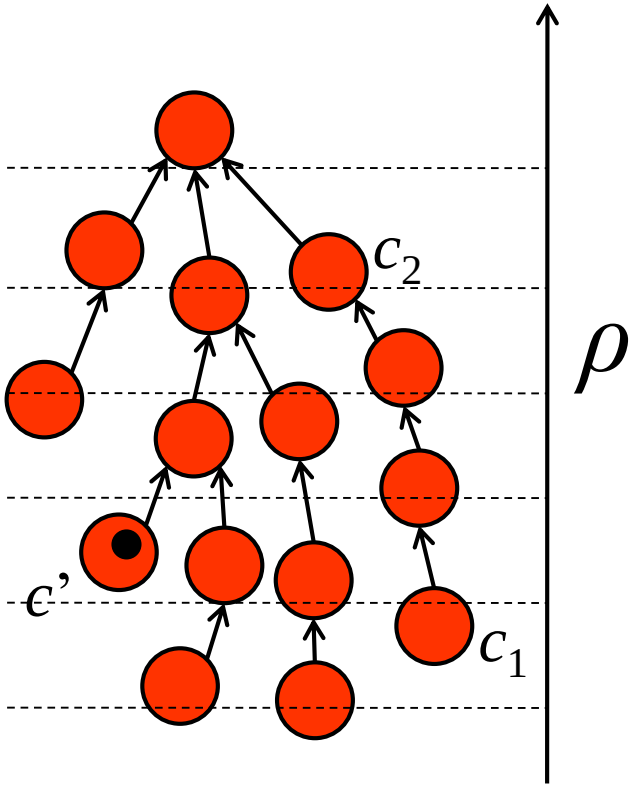
Update

- Update δ



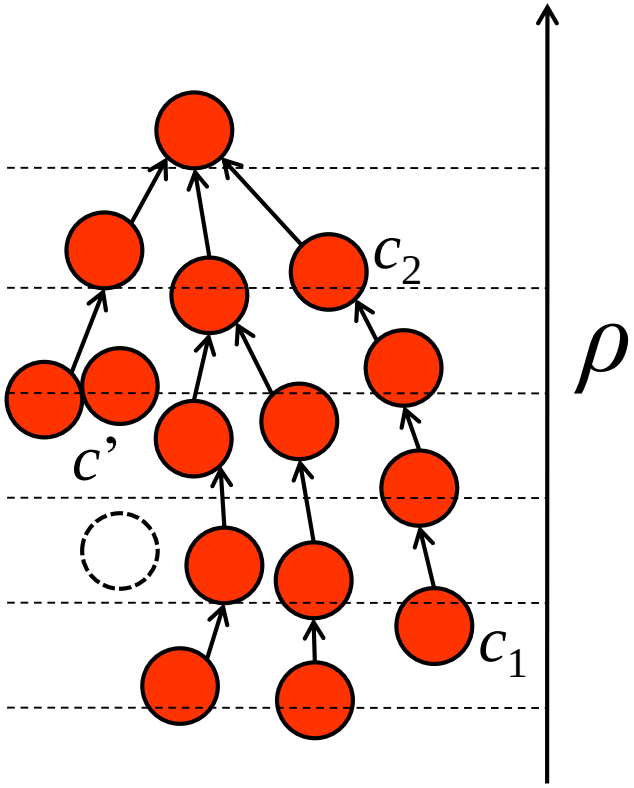
Update

- Update δ



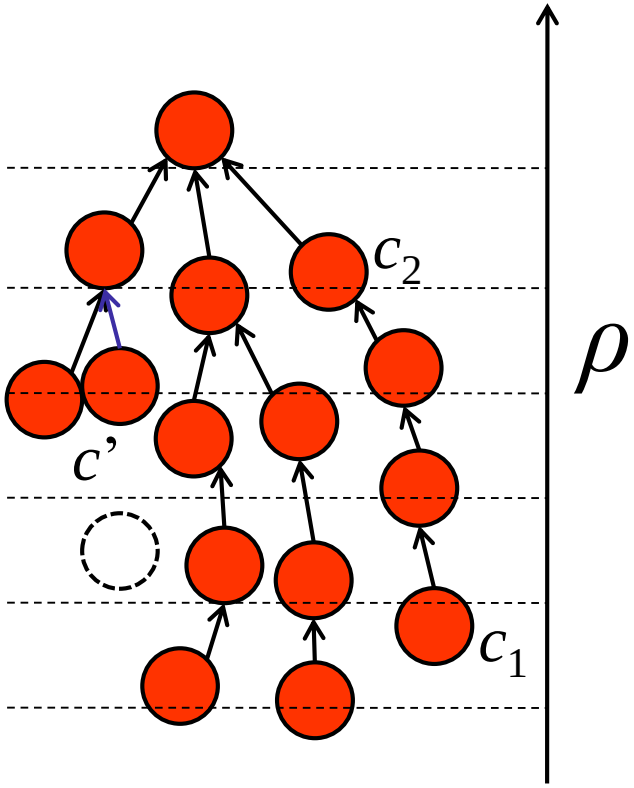
Update

- Update δ



Update

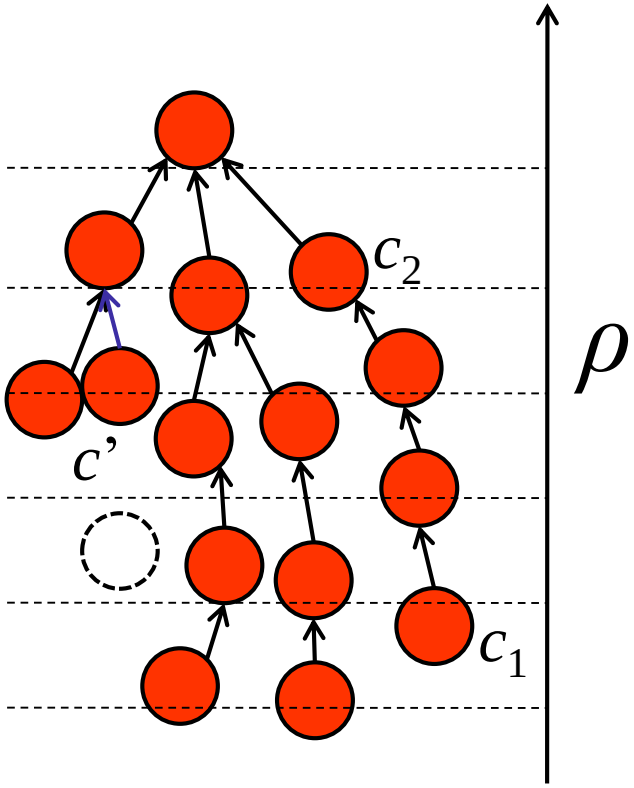
- Update δ



Update

- Update δ

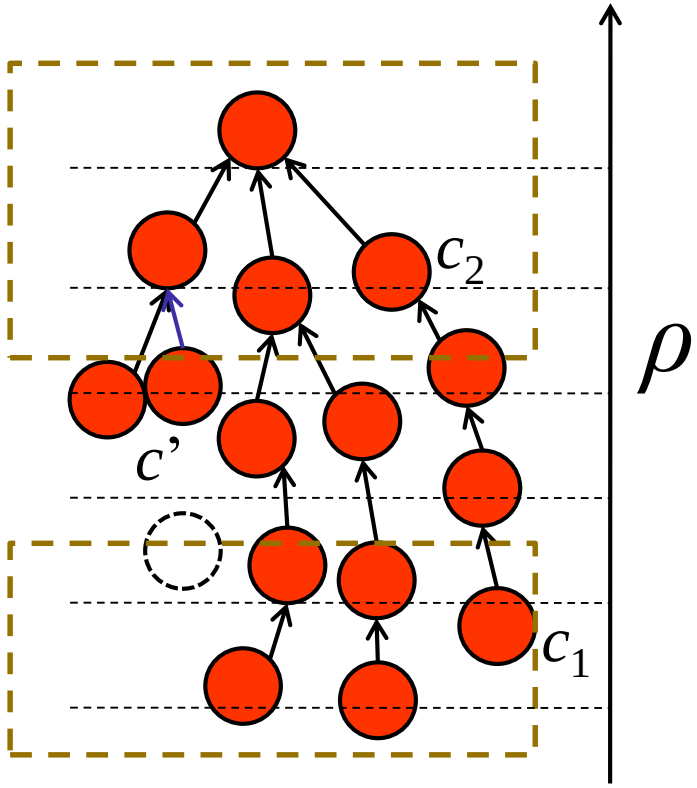
With the increasing of c′ density, c′ may become dependent cluster-cell of other cluster-cells.



Update

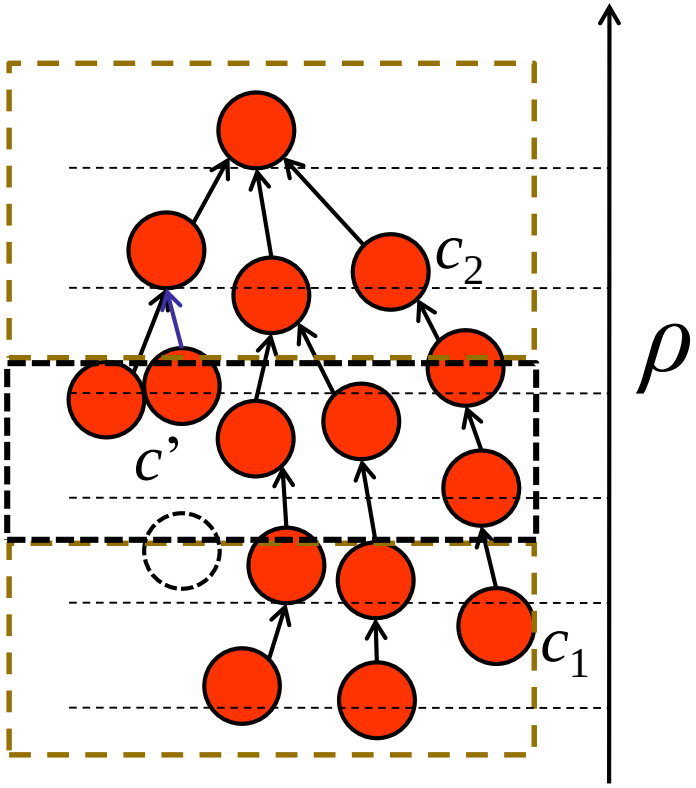
- Update δ

The cluster-cells whose density are larger than c1 and c2 are not changed.



Update

- Update δ

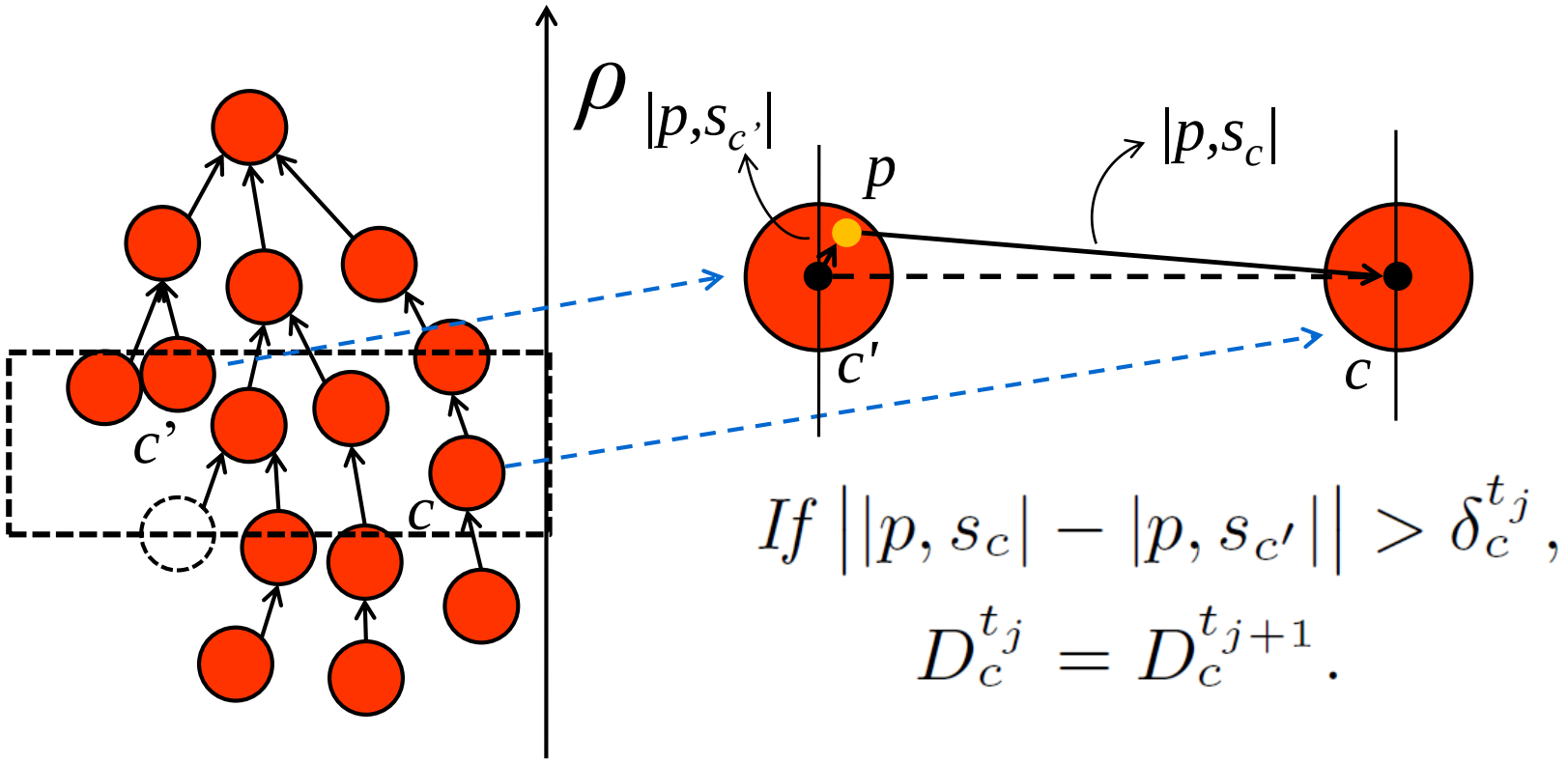


$$If\ \rho_c^{t_j} < \rho_{c'}^{t_j}\ \ or\ \rho_c^{t_{j+1}} > \rho_{c'}^{t_{j+1}},$$

$$D_c^{t_j} = D_c^{t_{j+1}}.$$

Update

- Update δ

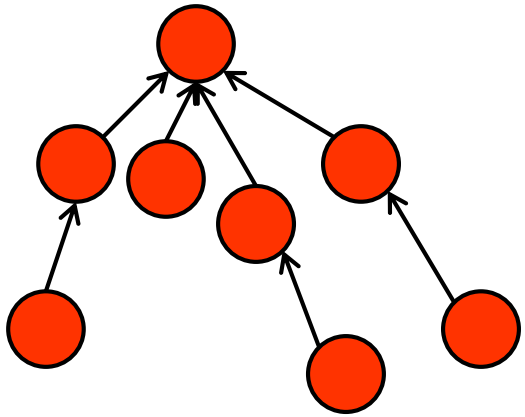


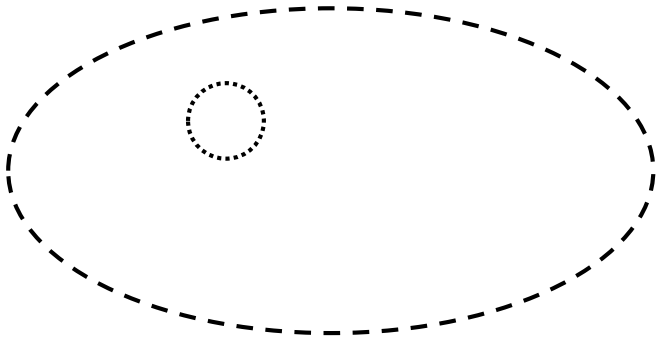
$$If \left| |p, s_c| - |p, s_{c'}| \right| > \delta_c^{t_j},$$

$$D_c^{t_j} = D_c^{t_{j+1}}.$$

Merge

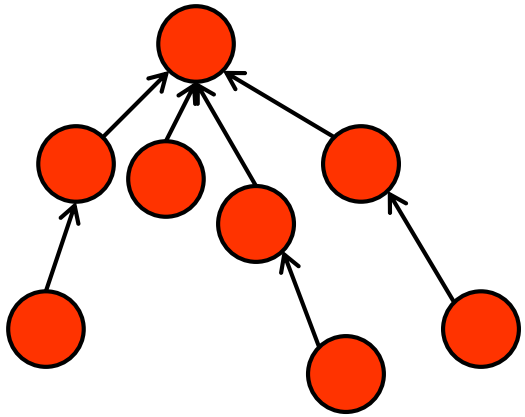
- The cluster-cell with low density may become dense, as it absorbs points.

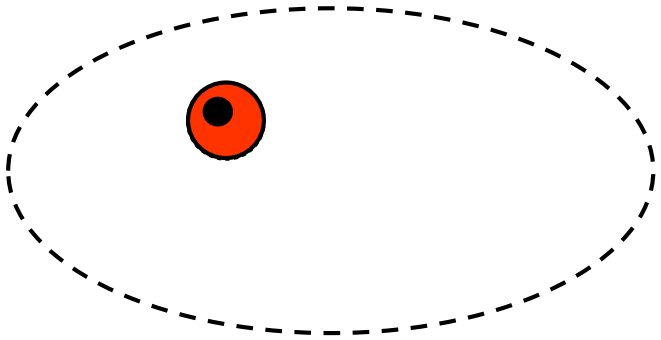




Merge

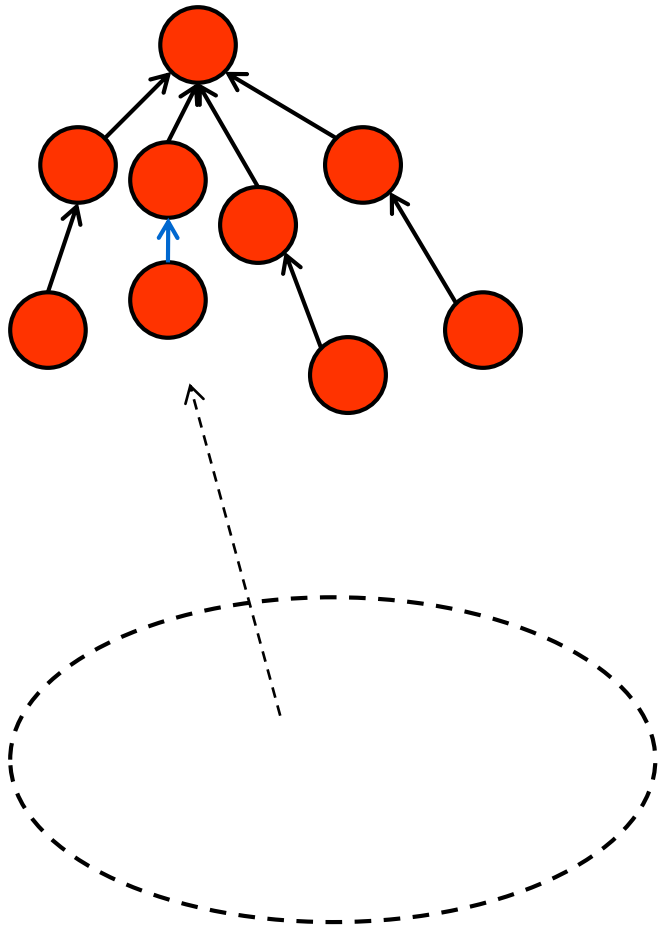
- The cluster-cell with low density may become dense, as it absorbs points.





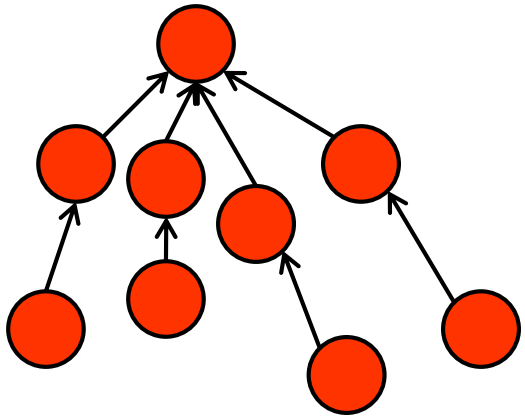
Merge

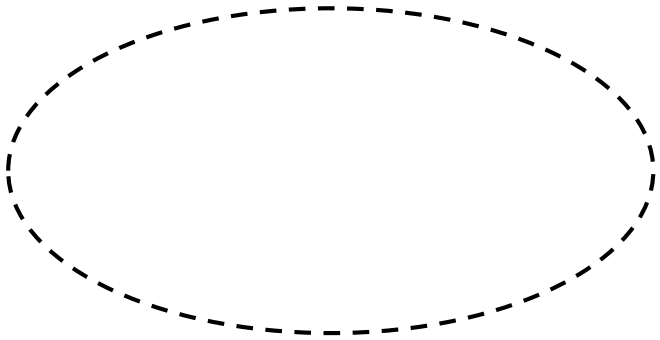
- Then it will be merged into DP-Tree.



Decay

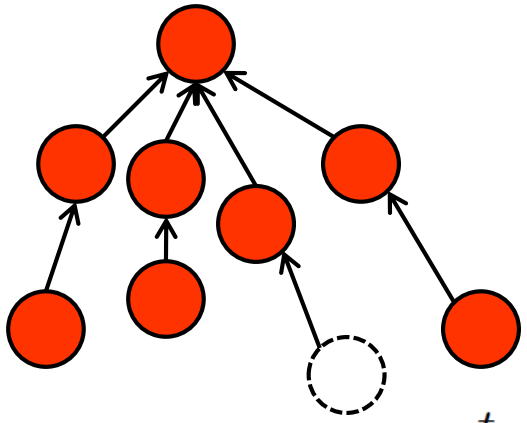
- As time goes on, the density of cluster-cell will decay, if it has not absorbed points for long time.



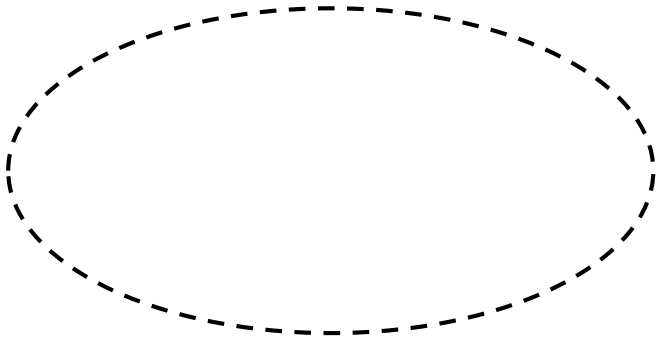


Decay

- As time goes on, the density of cluster-cell will decay, if it has not absorbed points for long time.

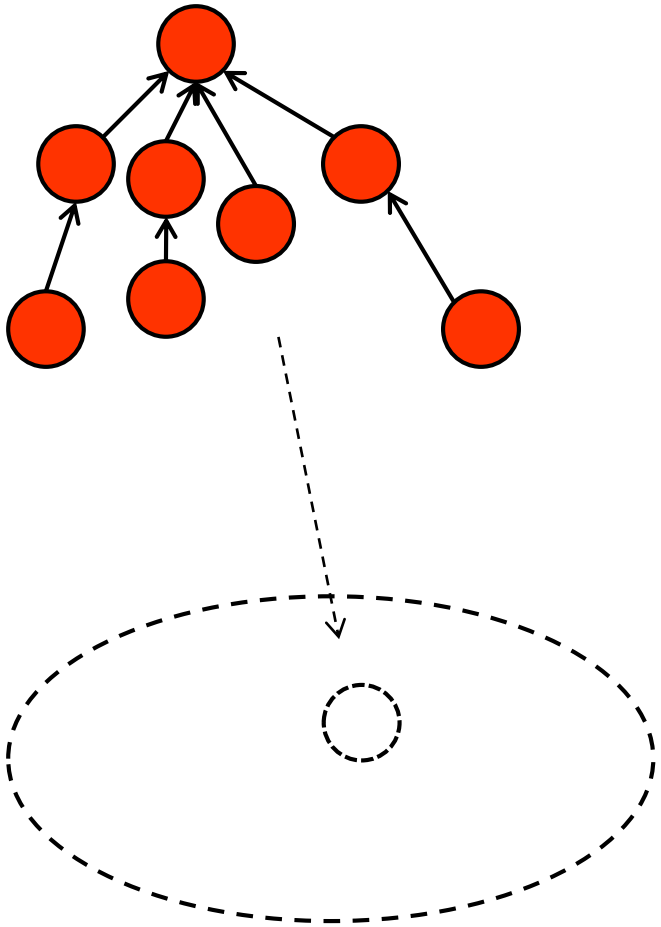


$$\rho_c^t < \frac{\beta \cdot \overline{v}}{1 - a^{\lambda}}$$

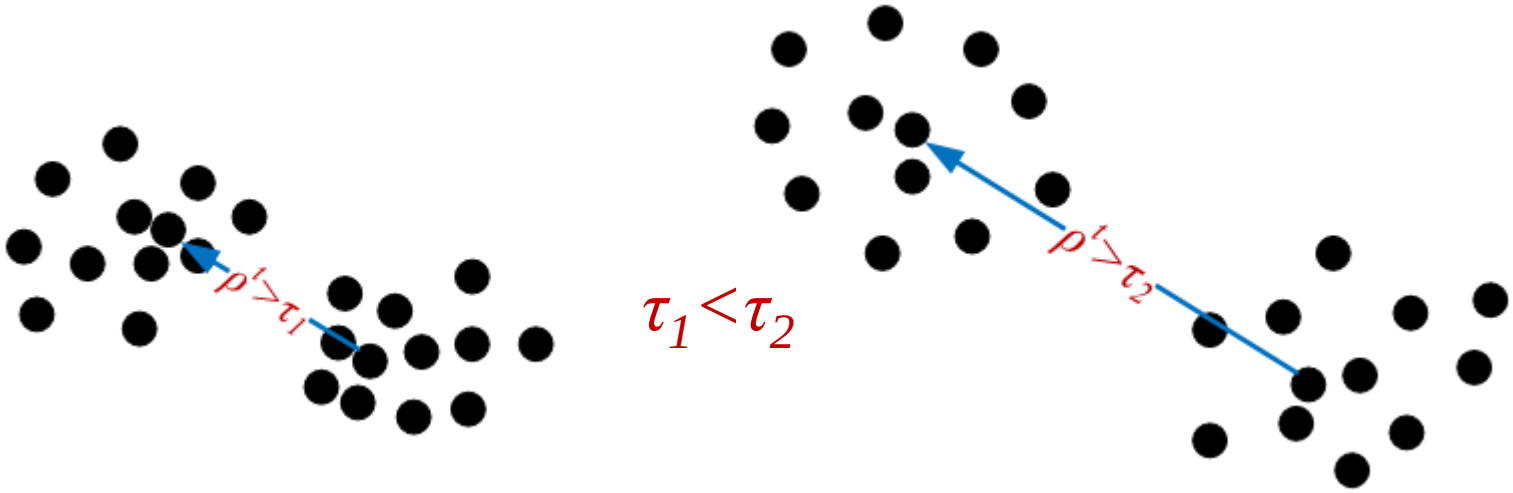


Decay

- The cluster-cell with lower density is moved to outlier-reservoir.

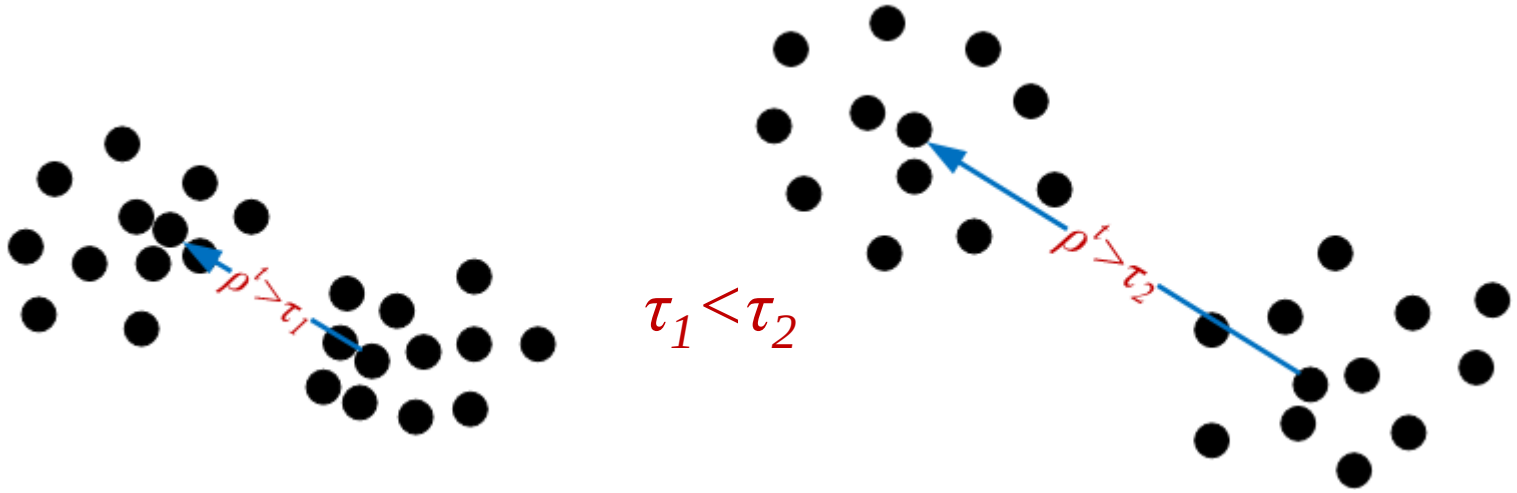


Adaptive tuning τ



The points may be denser or sparser, Therefore, the adaptive tuning τ is very important for us.

Adaptive tuning τ

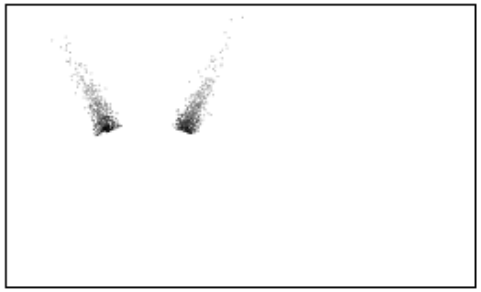


$$\mathcal{F}(\tau^t) = \alpha \cdot \frac{n \cdot \overline{\delta}}{\sum_{c:\delta_c^t > \tau^t} \delta_c^t} + (1-\alpha) \cdot \frac{\sum_{c:\delta_c^t \leq \tau^t} \delta_c^t}{m \cdot \overline{\delta}}$$

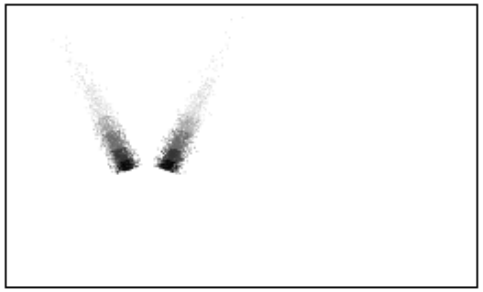
Outline

- Motivation
- EDMStream: Basic Idea
- EDMStream: Detail
- Evolution

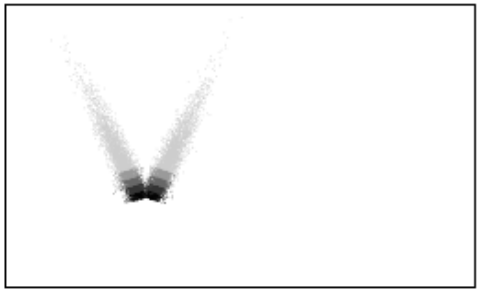
Track evolution



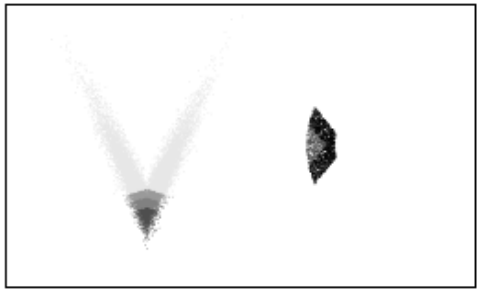
(a) $t_1 = 1s$



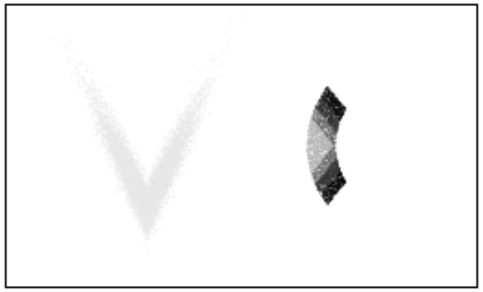
(b) $t_2 = 4s$



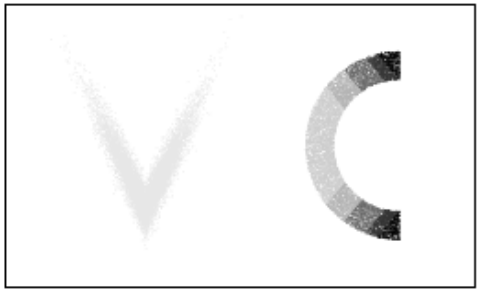
(c) $t_3 = 8s$



(d) $t_4 = 12s$

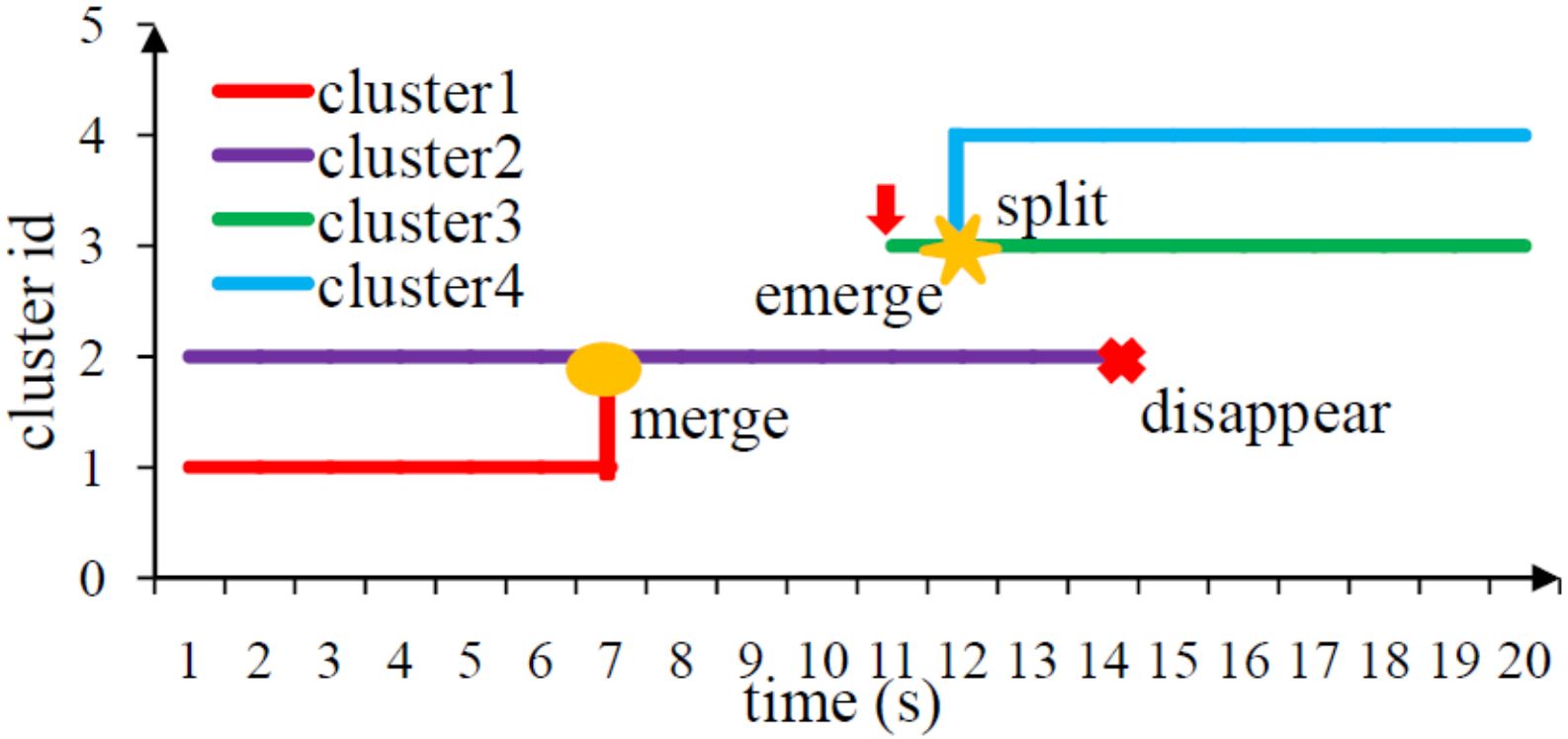


(e) $t_5 = 14s$

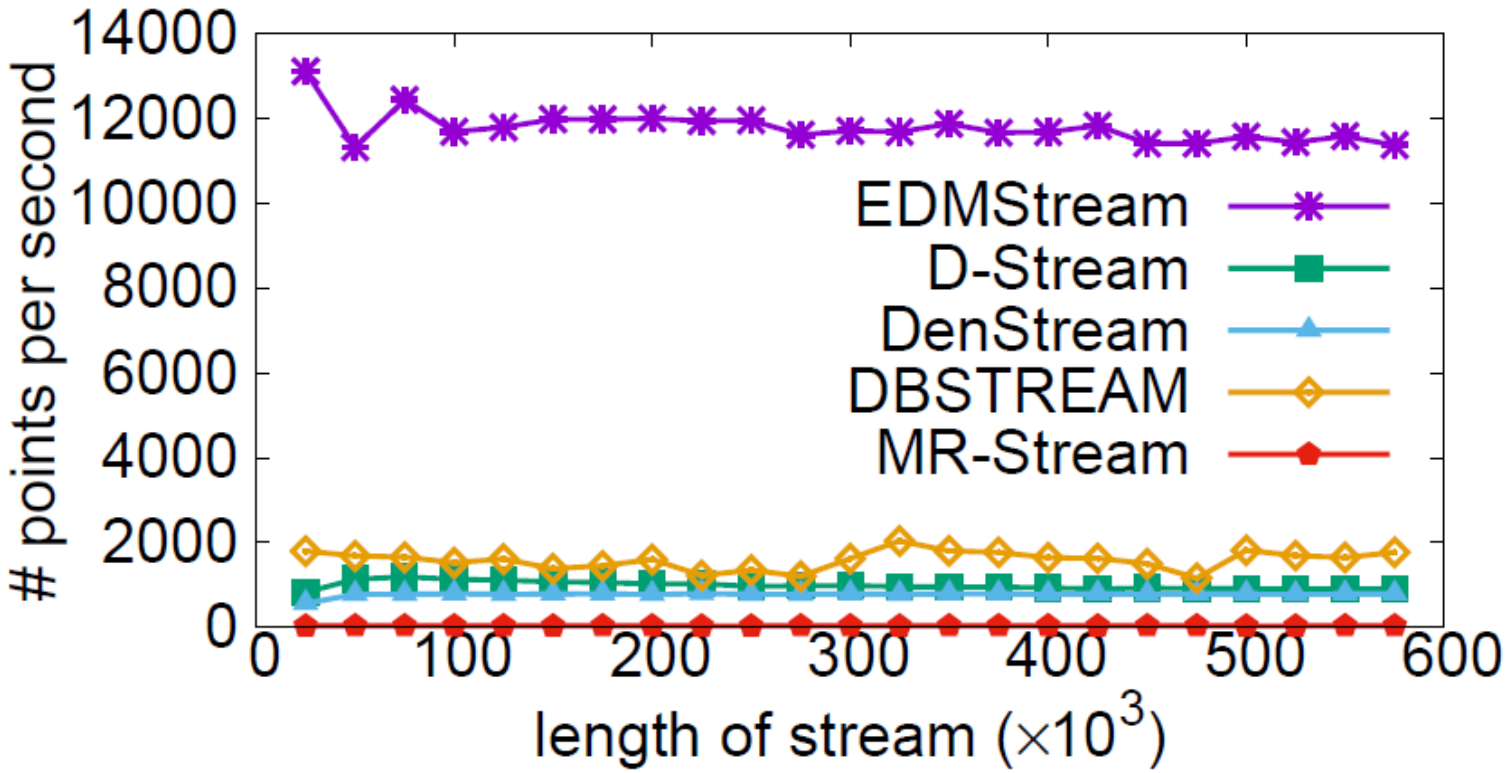


(f) $t_6 = 20s$

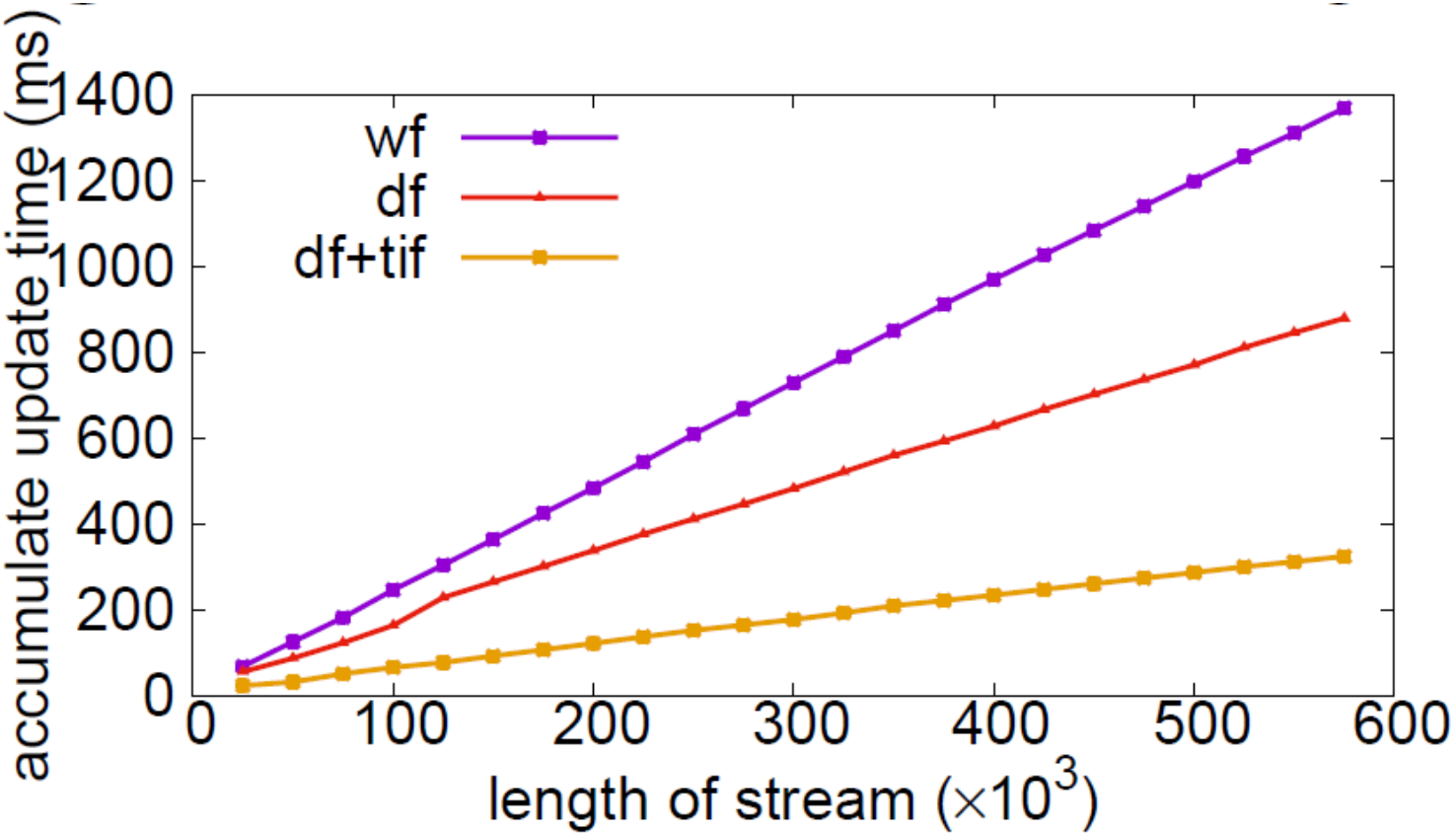
Track evolution



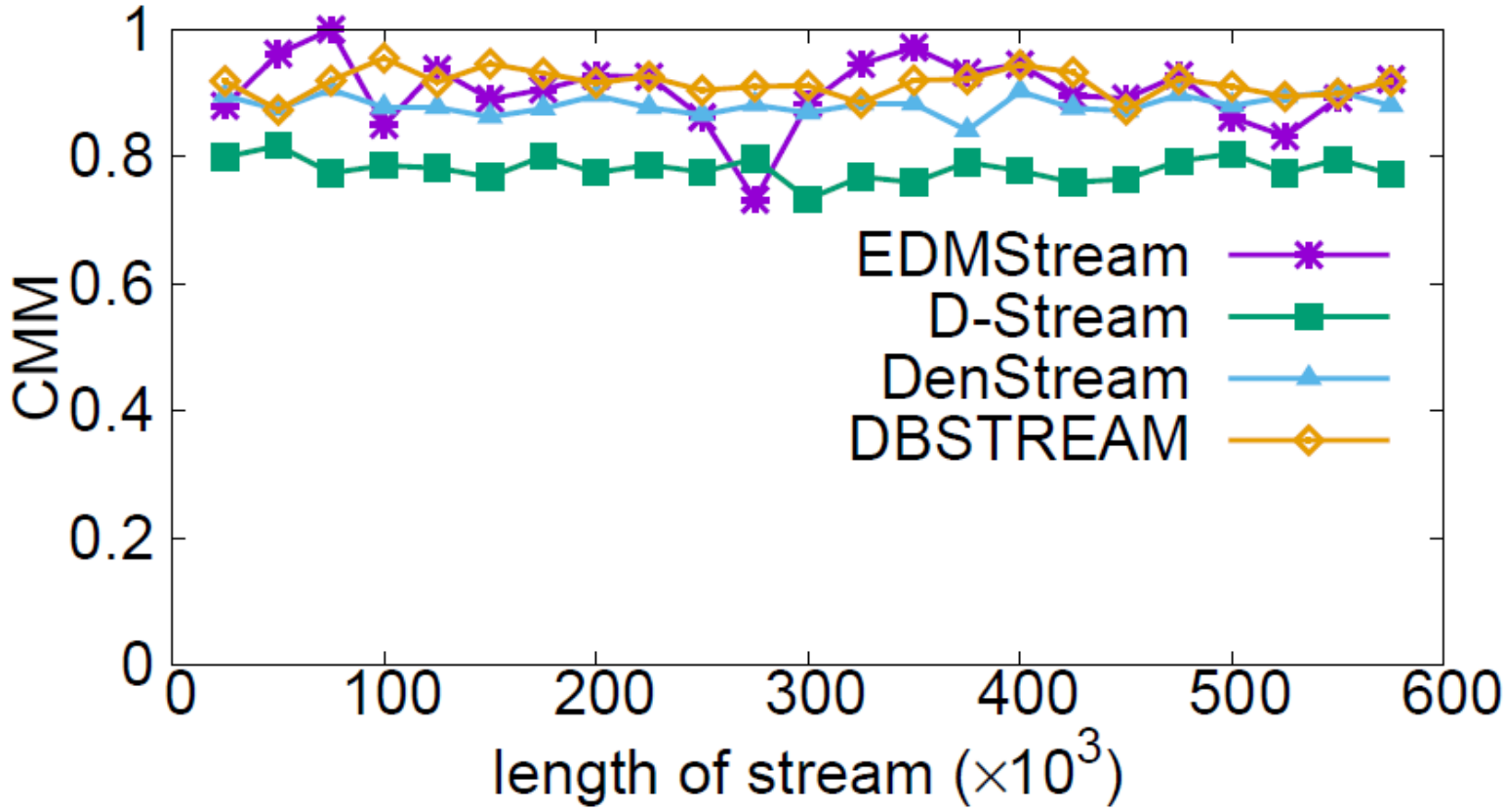
Throughput


14000
12000
10000
8000
6000
4000
2000
0
points per second
0
100
200
300
400
500
600
length of stream (×10^3)
EDMStream
D-Stream
DenStream
DBSTREAM
MR-Stream

Filter strategies



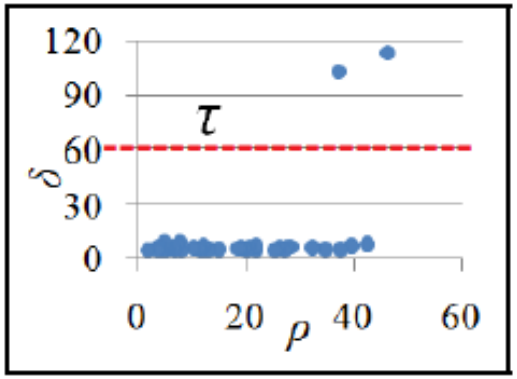
Cluster quality



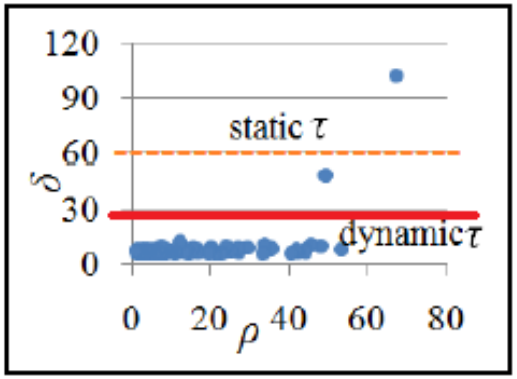
Adaptability

time point (s)	1	2	3	4	5	6	7	8	9
dynamic τ	2	2	2	2	1	1	2	3	2
static τ	2	2	2	1	1	1	2	2	2

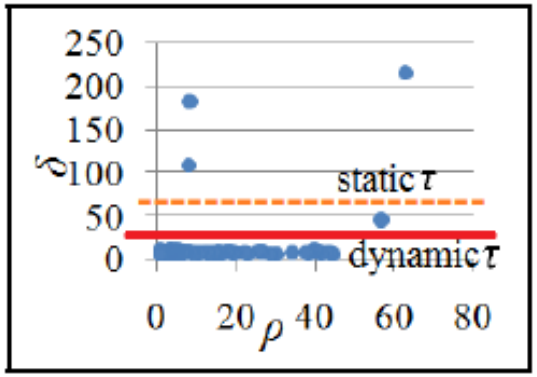
time point (s)	10	11	12	13	14	15	16	17	
dynamic τ	2	2	3	3	2	2	2	2	
static τ	2	2	3	3	1	2	2	2	



(a) init



(b) t_2=4s



(c) t_3=14s

Question & Answer