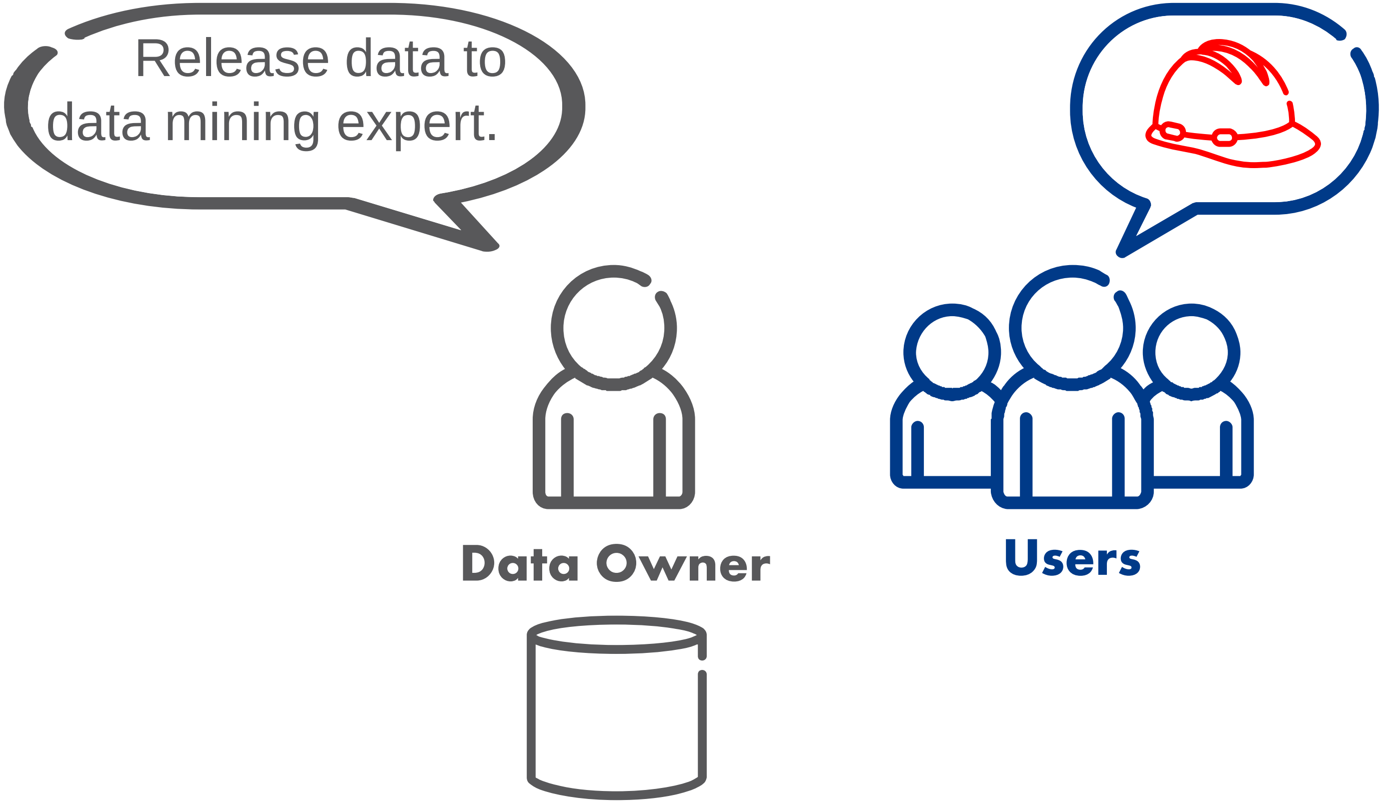
PRIVACY PRESERVING DATA PUBLISHING FOR RECOMMENDER SYSTEM

XIAOQIANG CHEN, VINCENT HUANG

OUTLINE

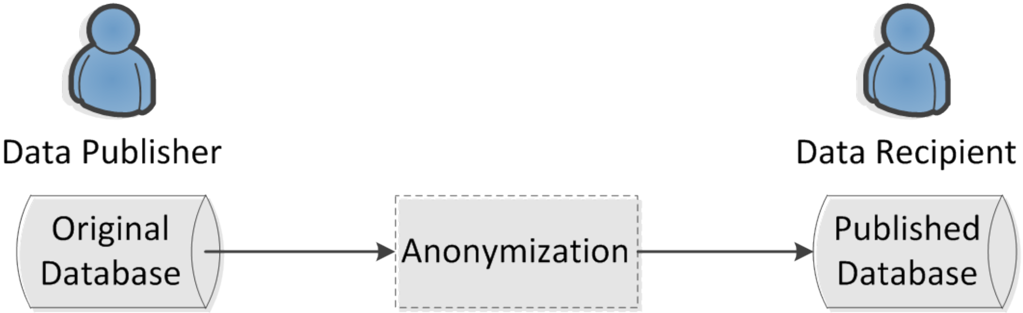
- Motivation
- Related Work
- Proposed Approach
- Bisecting K-Gather (BKG) Algorithm
- BKG Algorithm Evaluation
- Bisecting One-K-Gather (BOKG) Algorithm
- BOKG Evaluation
- Conclusions

MOTIVATION



PROBLEM STATEMENT

- Privacy Preserving Data Publishing

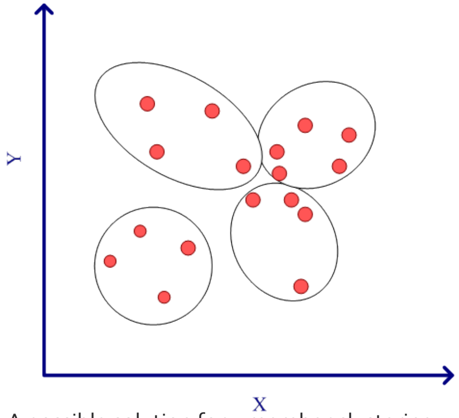


- Recommender system:
  - Ordinal data
  - High dimensionality
  - Sparsity

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7	Movie 8	Movie 9	...
User 1	1	0	2	0	0	5	5	0	0	...
User 2	2	0	0	0	0	4	5	0	0	...
User 3	5	4	0	4	0	0	0	0	0	...
User 4	5	4	0	0	0	3	0	0	0	...
User 5	4	5	0	4	0	0	0	0	0	...

BACKGROUND

- K-anonymity
  - intuitively, hide each individual among k-1 others
  - K-gather clustering



A possible solution for 4 member clustering

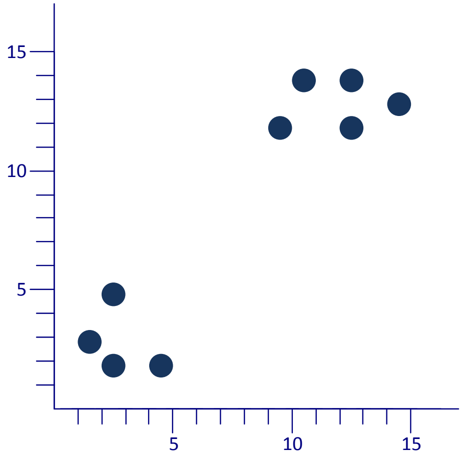
		Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7
🟩	User 1	1		2			5	5
	User 2	2					4	5
🟧	User 3	5	4		4			
	User 4	5	4				3	
	User 5	4	5		4			

homogenized

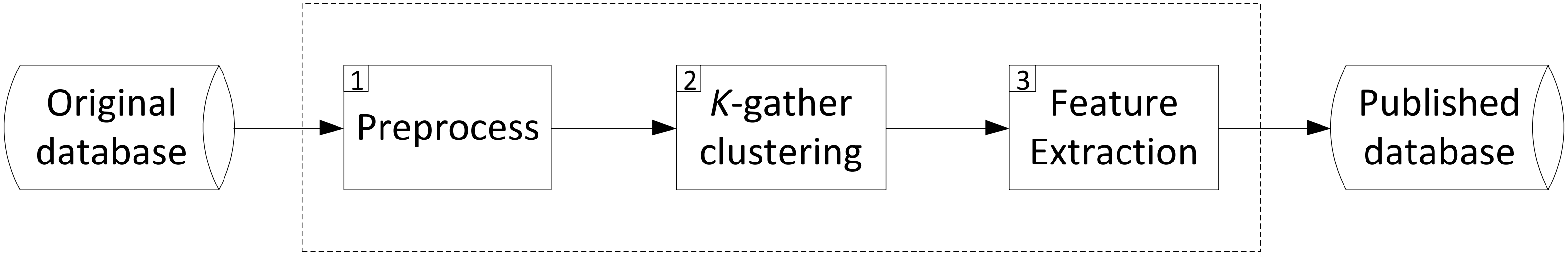
	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7	#
Group 1	1.5		2			4.5	5	2
Group 2	4.67	4.33		4		3		3

CHALLENGES

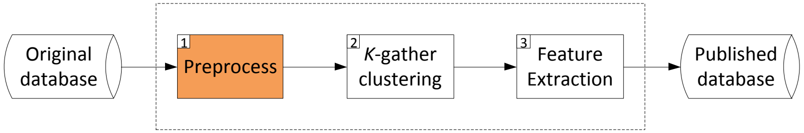
- Inherent features of recommender system
  - High dimensionality
  - Sparsity
- Drawback of fixed k-gather algorithms
  - $\lfloor \frac{n}{k} \rfloor$ clusters all of size k.



SOLUTION


Original database
1
Preprocess
2
K-gather clustering
3
Feature Extraction
Published database

SOLUTION

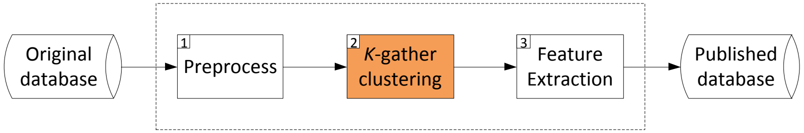


	Action	Animation	Comedy	Crime	Romance	Sci-Fi	Thriller	…
User 1	3	-0.3	0.1	2.1	-3	1.5	1	

	Action	Animation	Comedy	Crime	Romance	Sci-Fi	Thriller	…
Movie 1	1	0.1	1.4	-2	2	-0.2	-0.2	

R_{11}=3*1+(-0.3)*0,1+……

SOLUTION



Original database

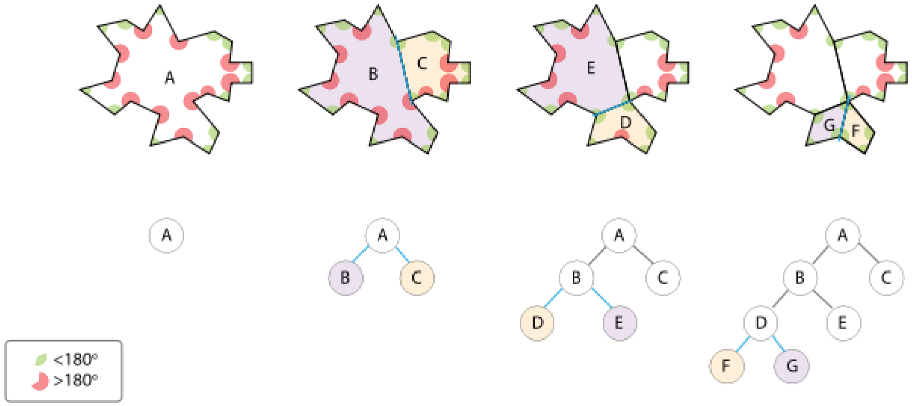
1 Preprocess

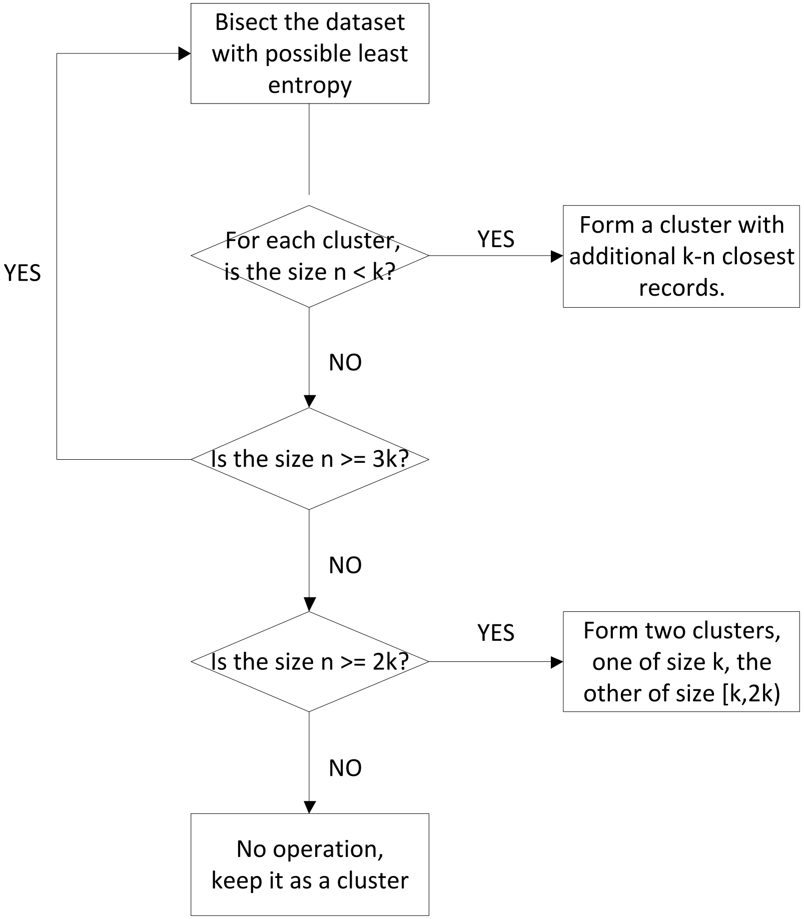
2 *K*-gather clustering

3 Feature Extraction

Published database

- K-gather clustering
  - Bisecting K-Gather
  - Non-fixed approach
  - V.S. fixed approach
    - Bounded T-Means





Bisect the dataset with possible least entropy

For each cluster, is the size n < k?

YES

Form a cluster with additional k-n closest records.

NO

Is the size n >= 3k?

YES

NO

Is the size n >= 2k?

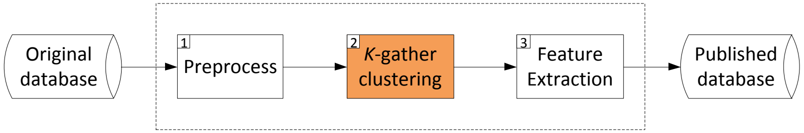
YES

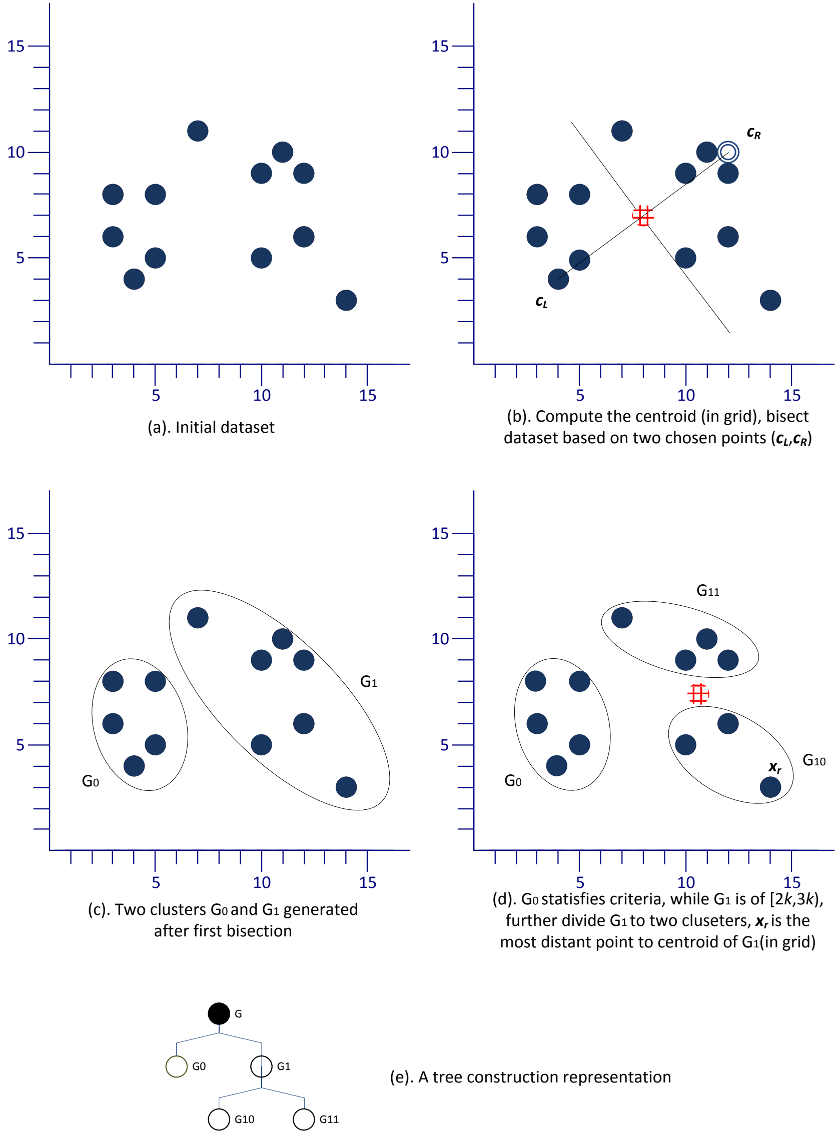
Form two clusters, one of size k, the other of size [k,2k)

NO

No operation, keep it as a cluster

SOLUTION





(a). Initial dataset

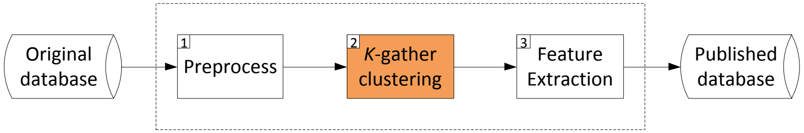
(b). Compute the centroid (in grid), bisect dataset based on two chosen points (***c**$_L$*,***c**$_R$*)

(c). Two clusters G$_0$ and G$_1$ generated after first bisection

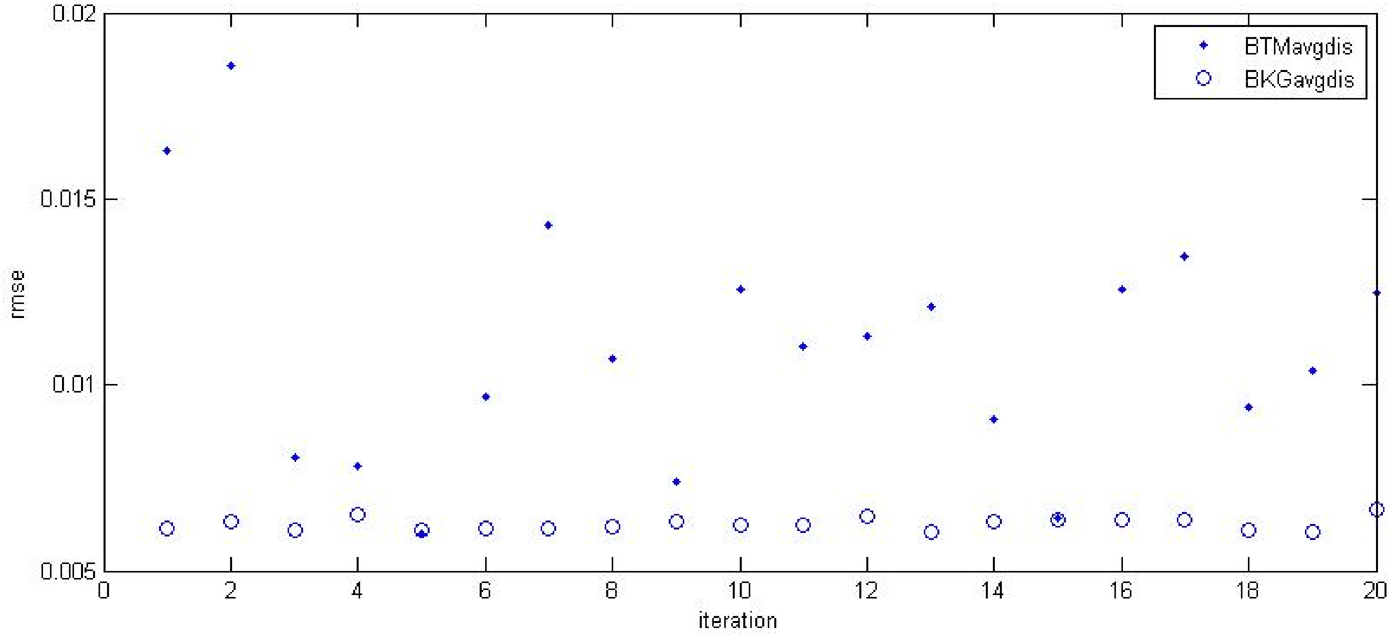
(d). G$_0$ statisfies criteria, while G$_1$ is of [2*k*,3*k*), further divide G$_1$ to two cluseters, ***x**$_r$* is the most distant point to centroid of G$_1$(in grid)

(e). A tree construction representation

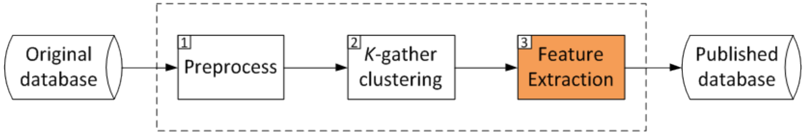
SOLUTION



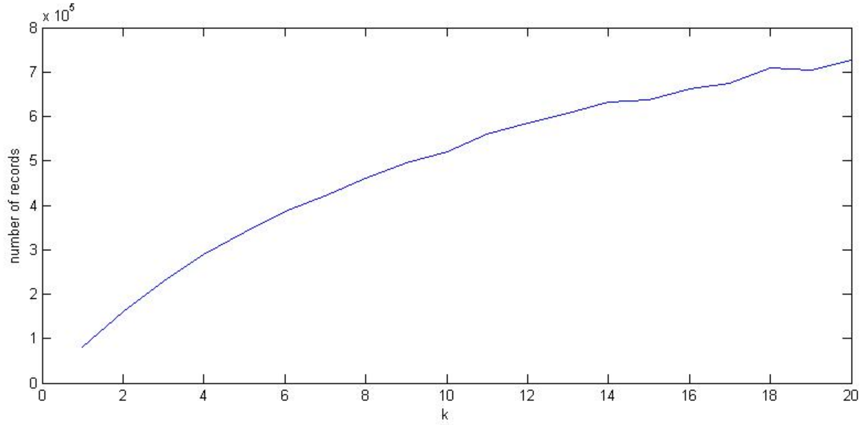
- Comparison with BTM.

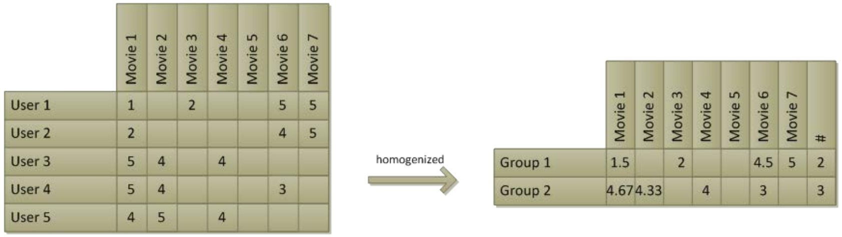


SOLUTION


Original database
1 Preprocess
2 *K*-gather clustering
3 Feature Extraction
Published database

- Feature Extraction
  - Compute average ratings over only users who rated that movie.
  - Anonymize the ratings within each cluster.
  - So, some unrated entries will get values after anonymization. the total number of entries increases with k.





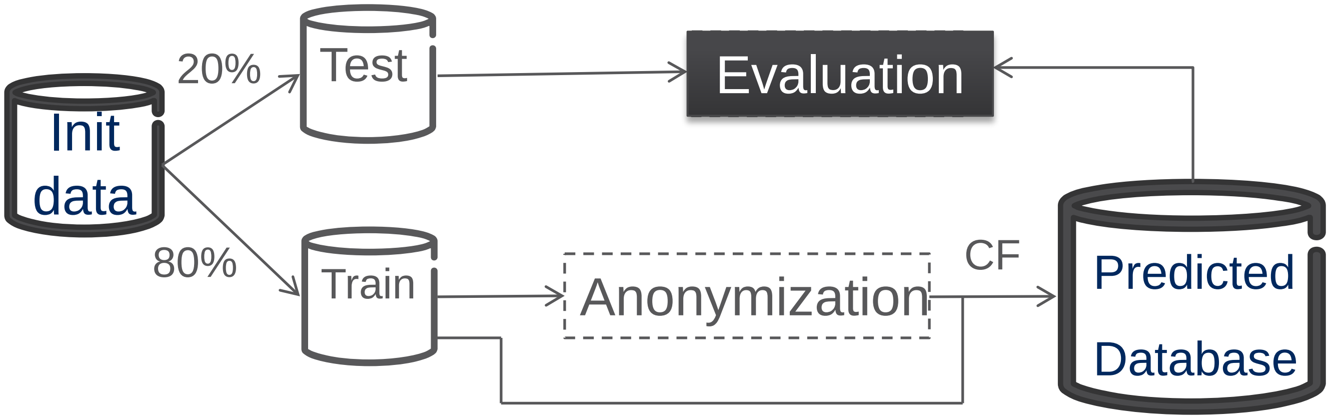
	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7
User 1	1		2			5	5
User 2	2					4	5
User 3	5	4		4			
User 4	5	4				3	
User 5	4	5		4			

homogenized

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7	#
Group 1	1.5		2			4.5	5	2
Group 2	4.67	4.33		4		3		3

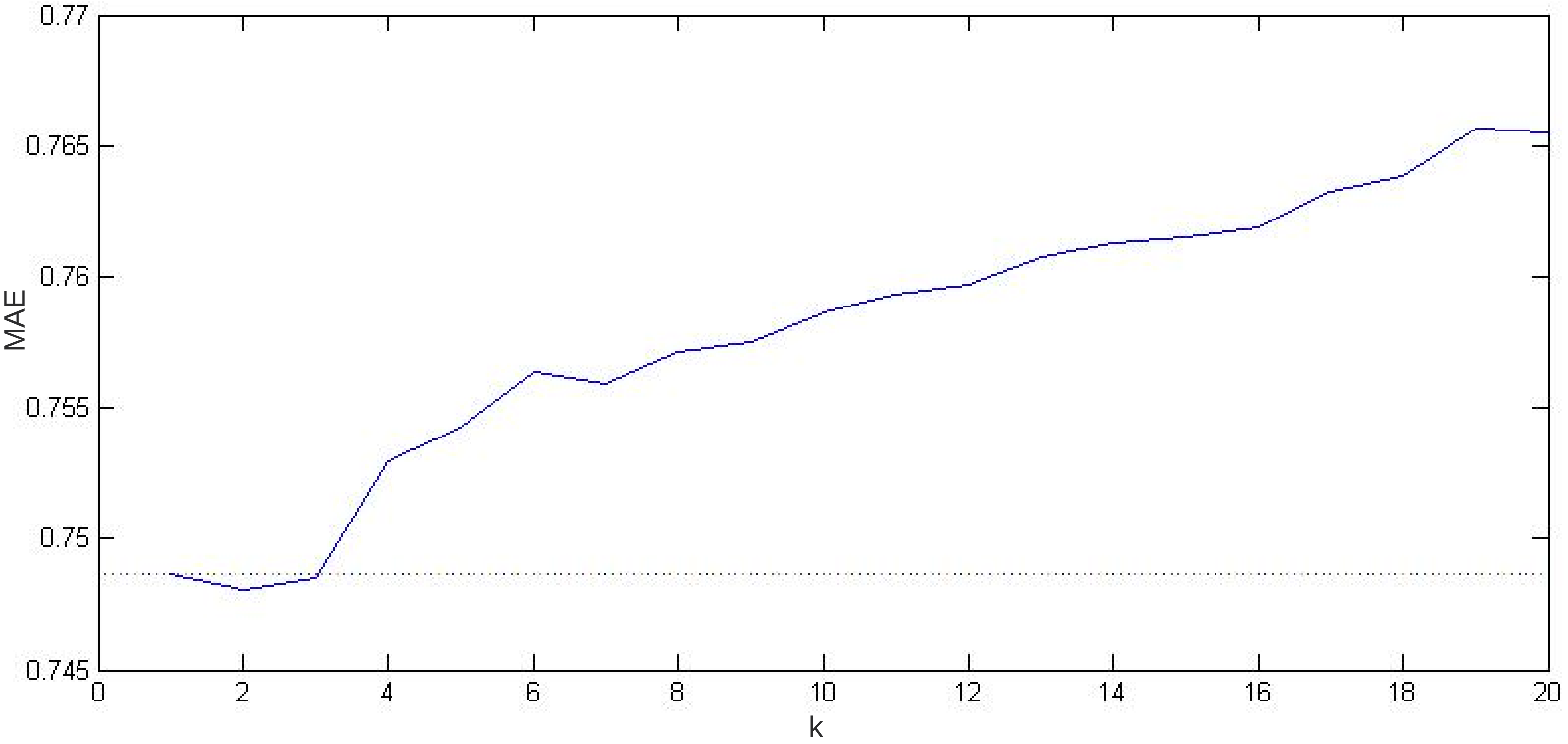
UTILITY EVALUATION

- Method
  - Database: MovieLens
    - 100K ratings of 1682 movies by 943 users.
    - Ratings follow the 1 (bad) to 5 (excellent) numerical scales.
    - The sparsity of the data set is high, at a value of 93.7%.
  - Measure of Prediction Accuracy: MAE
    - $MAE = \frac{1}{T}\sum_{t=1}^{T}\left|x_p - \hat{x}\right|$



UTILITY EVALUATION

- Results
  - K=1, 0.748
  - K=943, 0.815



CUSTOMIZED K-ANONYMITY

- Motivation

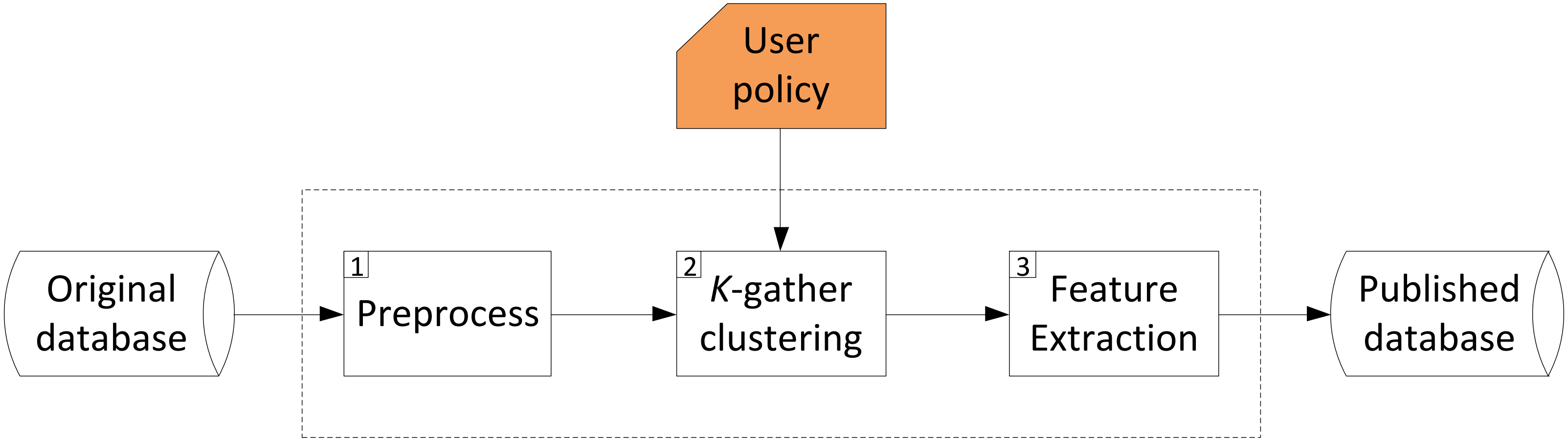
I don’t treat these information as my privacy, I just want better recommendations

Users

- Hypothesis
  - User of lower privacy level can get better recommendation.

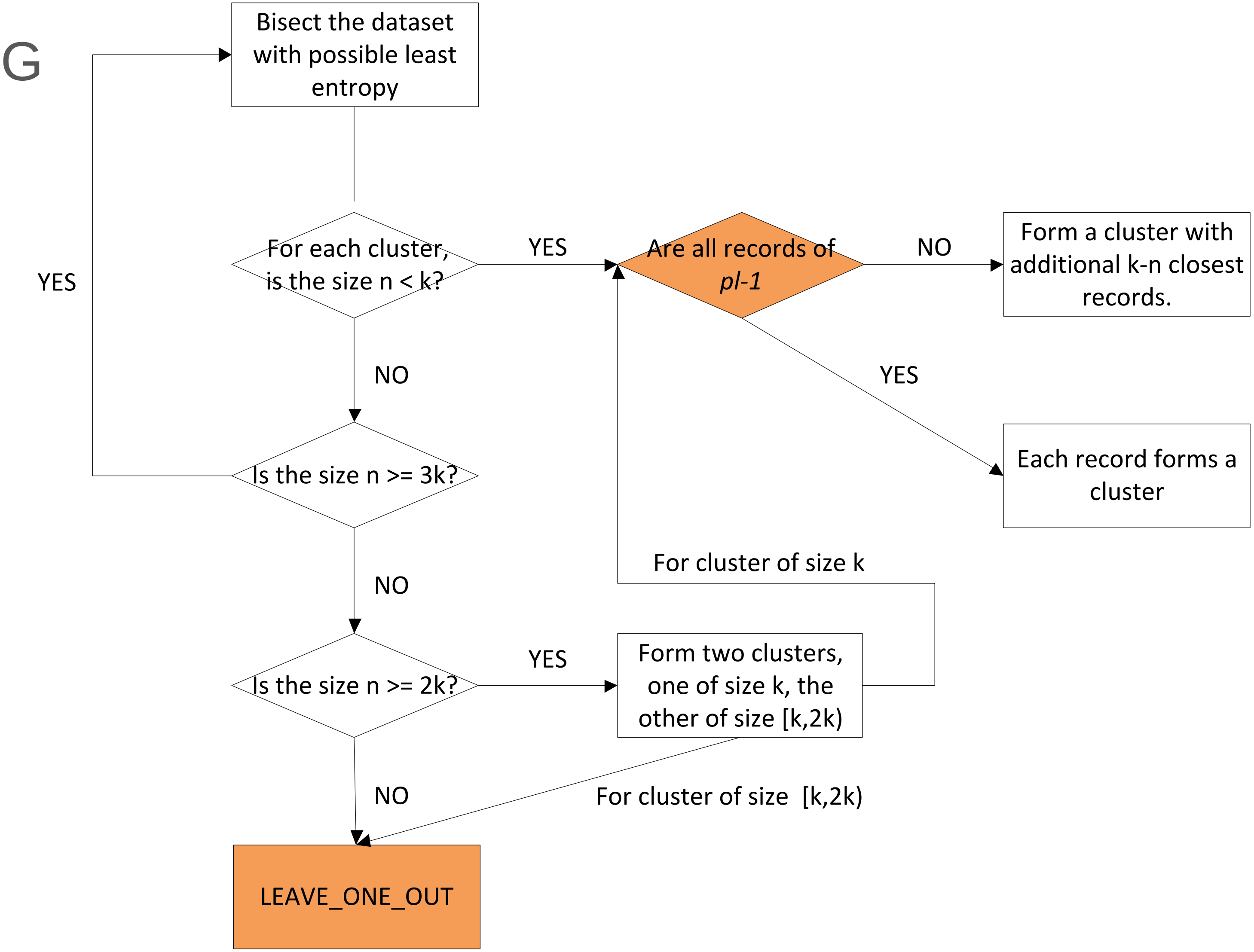
CUSTOMIZED K-ANONYMITY

- User policy



CUSTOMIZED K-ANONYMITY

› BOKG


Bisect the dataset with possible least entropy

For each cluster, is the size n < k?

YES

Are all records of *pl-1*

NO

Form a cluster with additional k-n closest records.

YES

Each record forms a cluster

NO

YES

Is the size n >= 3k?

NO

Is the size n >= 2k?

YES

Form two clusters, one of size k, the other of size [k,2k)

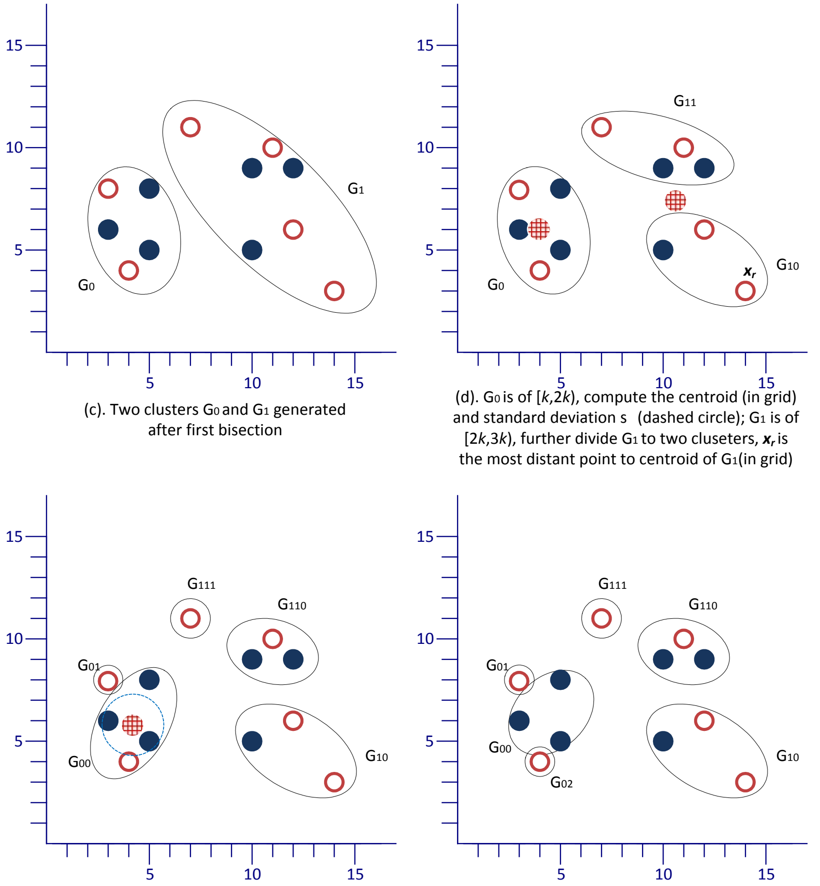
For cluster of size k

NO

For cluster of size [k,2k)

LEAVE_ONE_OUT

CUSTOMIZED K-ANONYMITY



(c). Two clusters G_0 and G_1 generated after first bisection

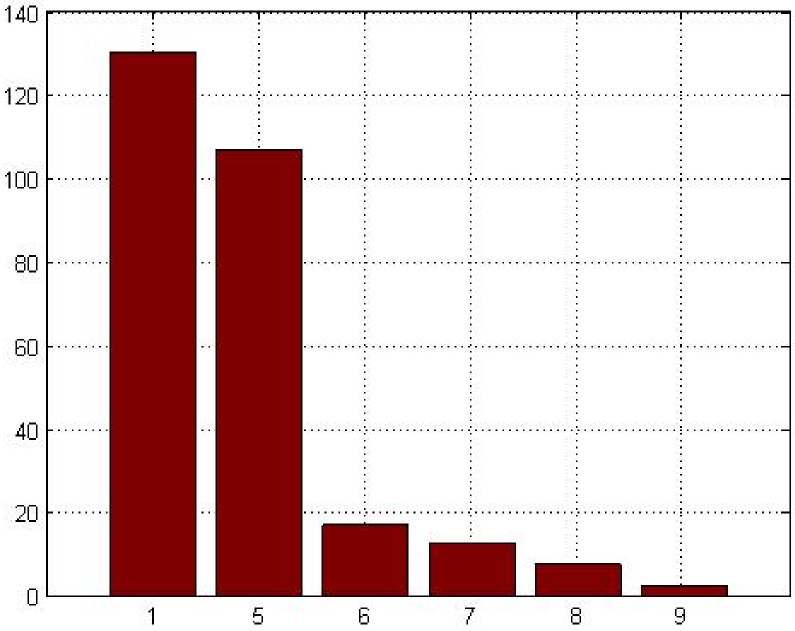
(d). G_0 is of [*k*,2*k*), compute the centroid (in grid) and standard deviation s (dashed circle); G_1 is of [2*k*,3*k*), further divide G_1 to two cluseters, $\boldsymbol{x_r}$ is the most distant point to centroid of G_1(in grid)

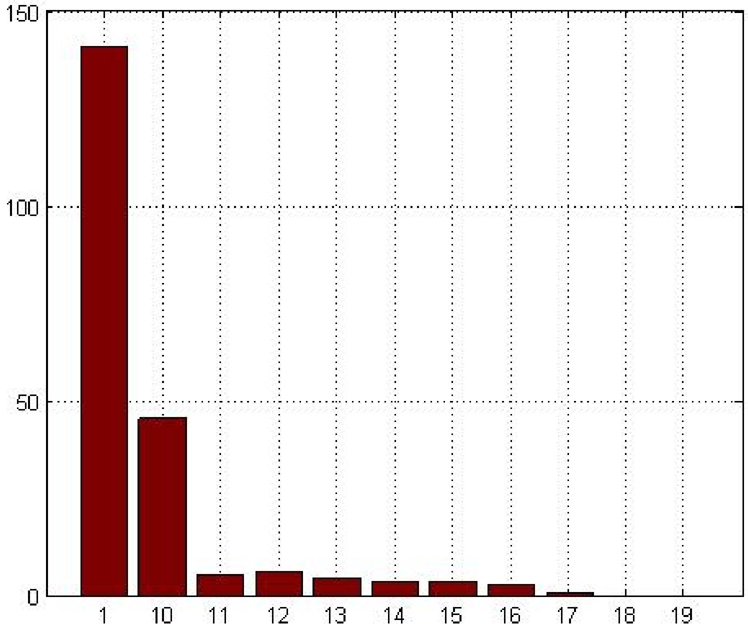
(e). Form a separate cluster of point (3.8) of *pl-1* which is most distant from centroid and out of the s circle, recompute centroid and s ; same as cluster of point (7,11)

(f). Point (4,4) stands alone, eventually, we have six clusters, three of which are one-record clusters

EXPERIMENTAL RESULTS

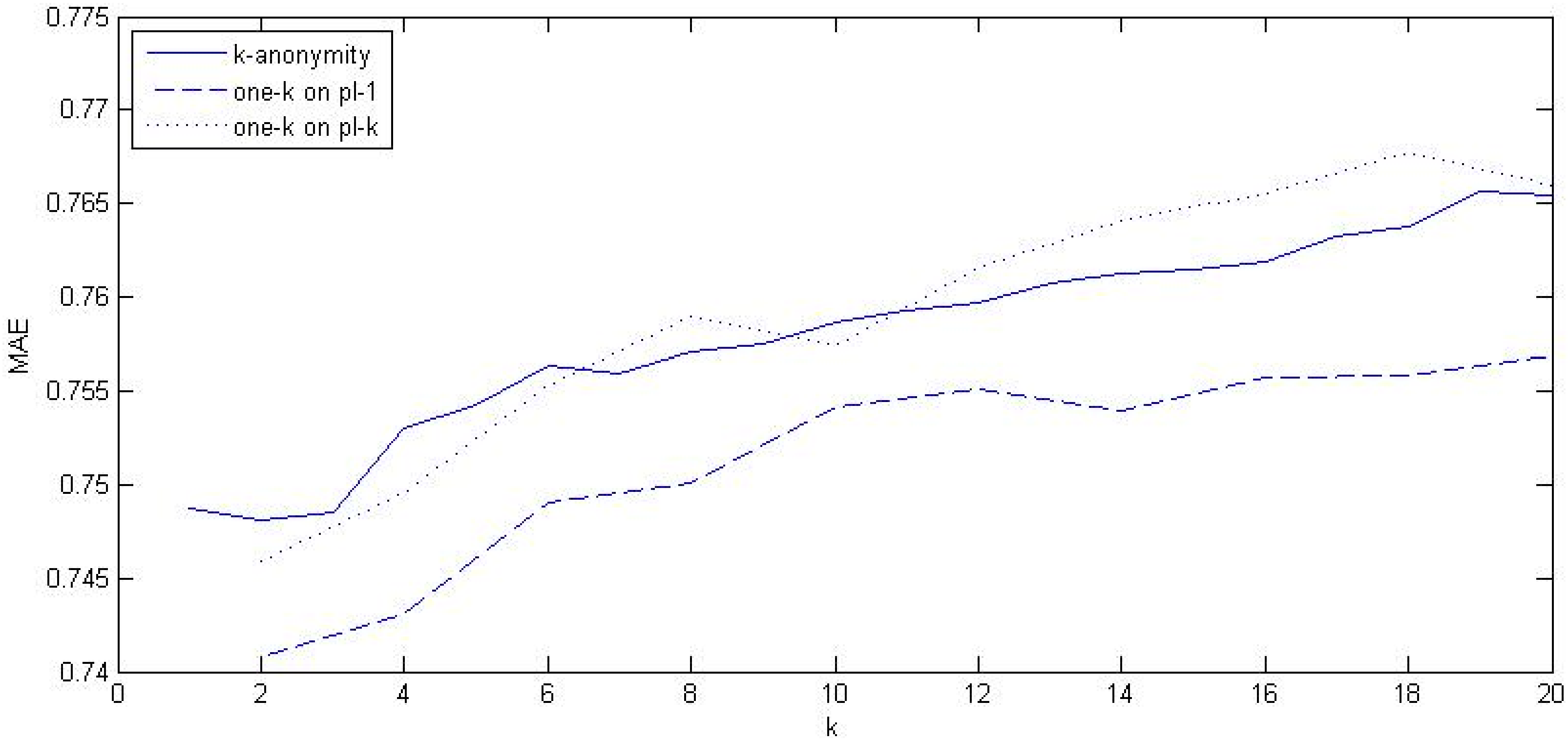
- Distributions
  - 471 pl-1 users
  - 472 pl-k users





EXPERIMENTAL RESULTS

› Results



CONCLUSIONS

- SVD solves high dimensionality and sparsity
- BKG is an efficient anonymization algorithm and preserves data utility
- BOKG supports customized privacy policies
- Better performance with less privacy requirements in mixed situations

Q & A

ERICSSON