

Collaborative Filtering: Fundamental Limits and Good Practices

Onkar Dabeer
School of Technology and Computer Science
Tata Institute of Fundamental Research
Mumbai, India

Beyond Search

Social Network



Media



Shopping



Events



Services



Travel

Beyond Search

Social Network



Media



Shopping



Events



Services

Web Search:
Long list of
related items



Travel

Beyond Search

Social Network



Media



Shopping



Events



Services

Web Search:
Long list of
related items



Travel

Recommendations:
Few “likable” items
Limited domain

Recommendations in Action

- Amazon
 - People who bought this also bought...
- Google News: recommended stories
- iTunes - Genius sidebar
- Netflix
 - Suggests movies using rating matrix
- Facebook, LinkedIn
 - Suggest connections
- RichRelevance recommendation engine
 - Disney Stores, Sears, Office Depot, etc.

Collaborative Filtering

Items

	2	?	1	3
	4	4	?	?
Users	?	3	1	?
	1	?	2	?

Recommend items based on
available ratings

No Content Information

State of the Art

State of the Art

- Committee Methods
 - Strong empirical and some theoretical support in classification, regression

State of the Art

- Committee Methods
 - Strong empirical and some theoretical support in classification, regression
- The Netflix prize winner blends 100+ algorithms
 - SVD inspired algorithms, neighborhood methods, exploiting time stamps, movie names
 - Typical individual RMSE > 0.9 (on test data)
 - After blending, RMSE is about 0.85

State of the Art

- Committee Methods
 - Strong empirical and some theoretical support in classification, regression
- The Netflix prize winner blends 100+ algorithms
 - SVD inspired algorithms, neighborhood methods, exploiting time stamps, movie names
 - Typical individual RMSE > 0.9 (on test data)
 - After blending, RMSE is about 0.85
- RichRelevance blends about 40 algorithms

Some Questions

Some Questions

- Is a reduction of RMSE from 0.95 to 0.85 relevant on a scale of 1-5?
 - Poor confidence
 - Why not 'probability of error'?

Some Questions

- Is a reduction of RMSE from 0.95 to 0.85 relevant on a scale of 1-5?
 - Poor confidence
 - Why not 'probability of error'?
- Do we need so many non-trivial predictors?
 - Diminishing returns; few important principles?

Some Questions

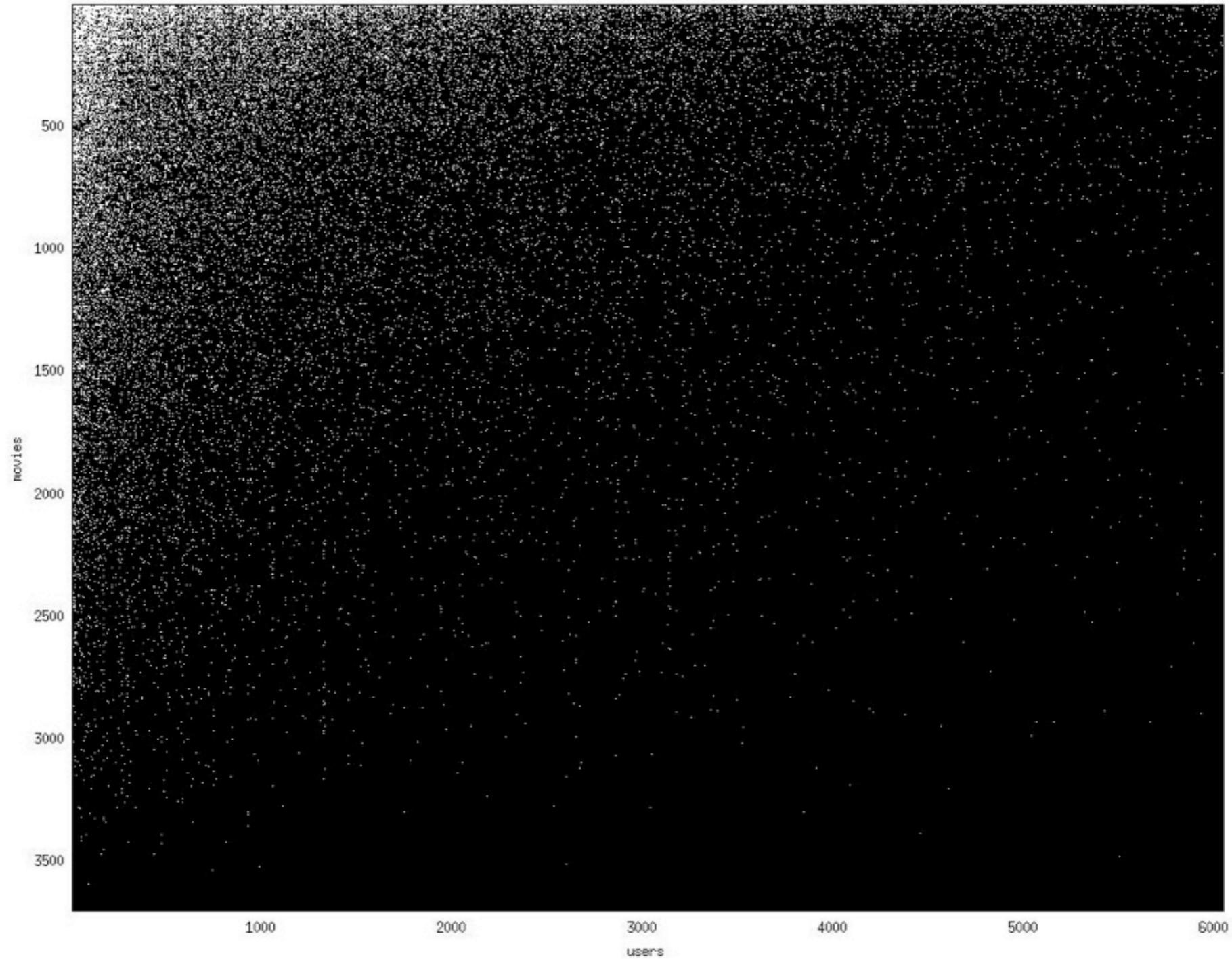
- Is a reduction of RMSE from 0.95 to 0.85 relevant on a scale of 1-5?
 - Poor confidence
 - Why not 'probability of error'?
- Do we need so many non-trivial predictors?
 - Diminishing returns; few important principles?
- Any 'provably good' principles that also show good empirical performance?

Some Questions

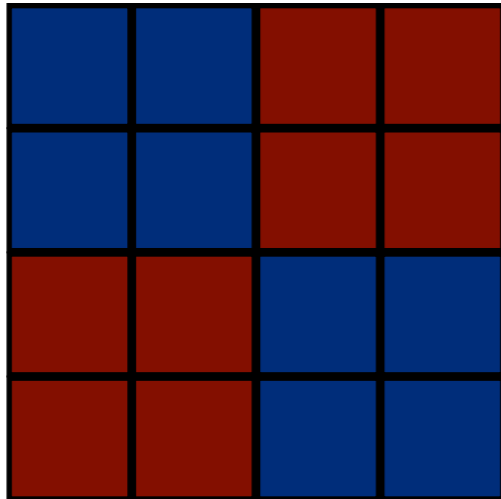
- Is a reduction of RMSE from 0.95 to 0.85 relevant on a scale of 1-5?
 - Poor confidence
 - Why not ‘probability of error’?
- Do we need so many non-trivial predictors?
 - Diminishing returns; few important principles?
- Any ‘provably good’ principles that also show good empirical performance?
- In this talk, we explore these questions

Rating Matrix Models

Movielens Rating Matrix

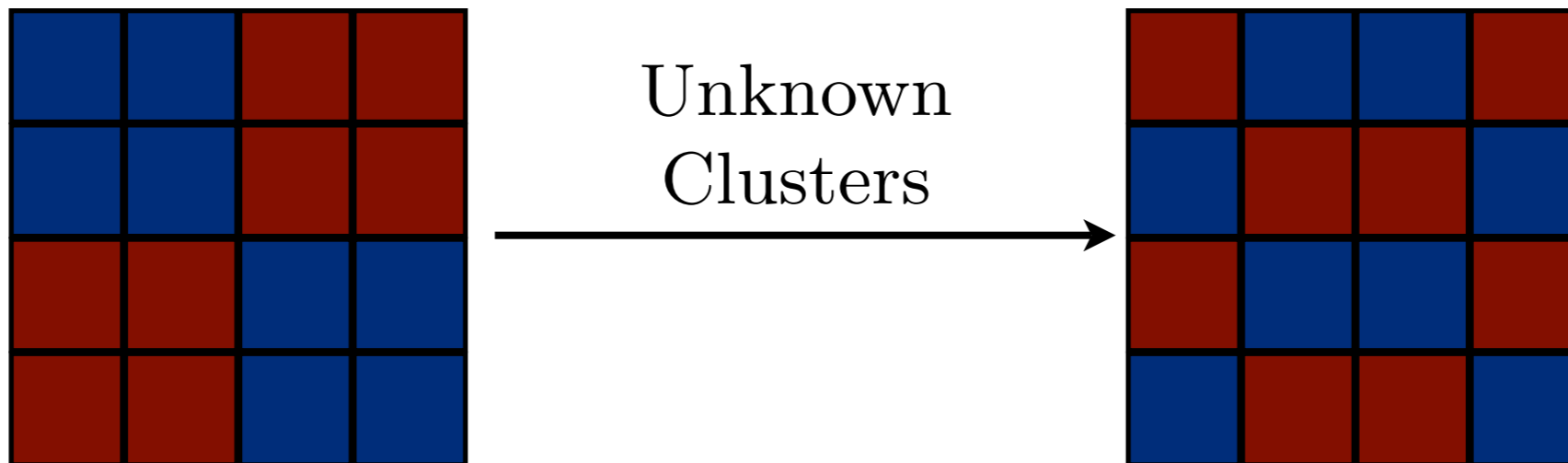


Estimation of Rearranged Smooth Processes



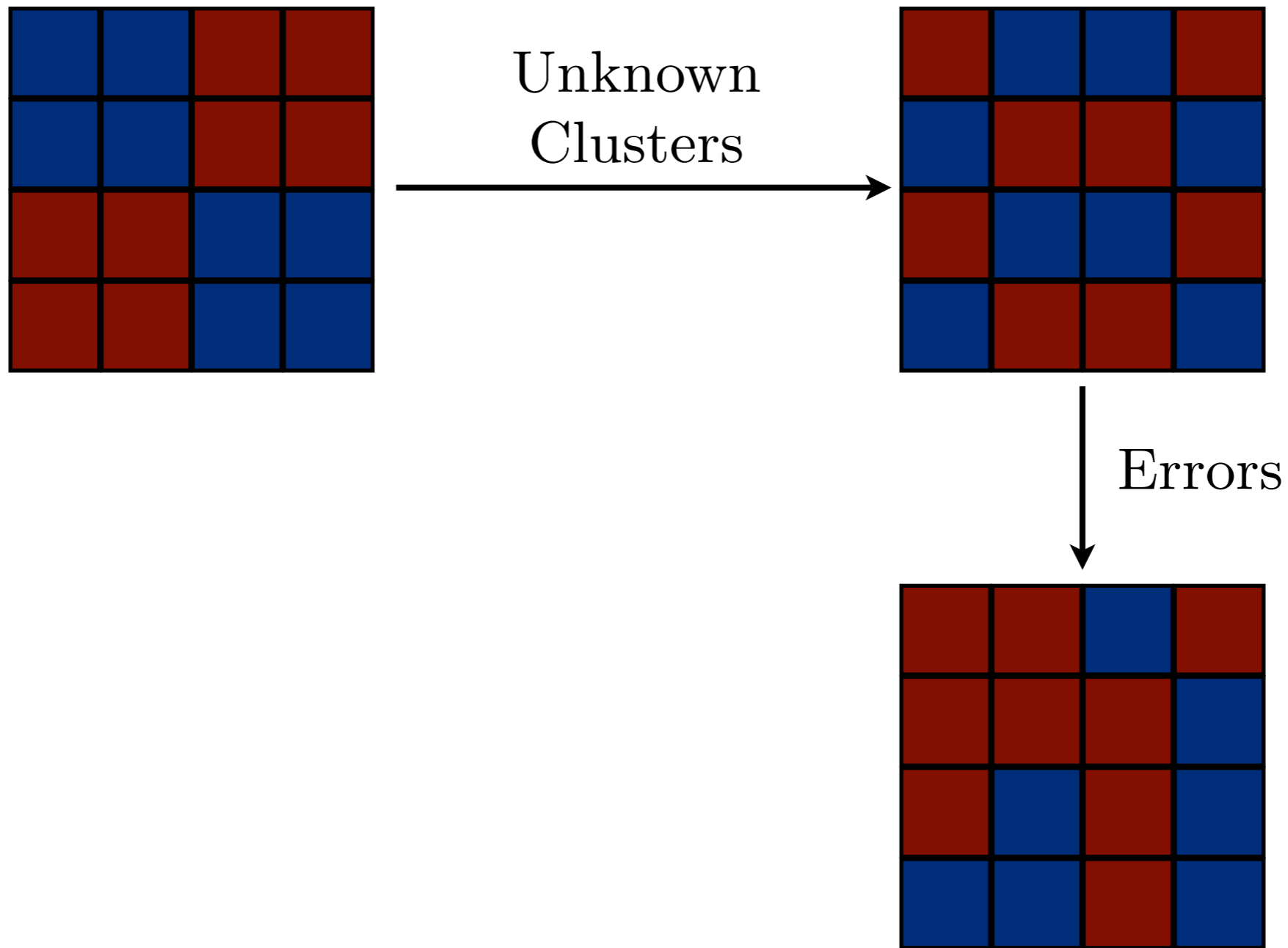
Introduced in Aditya, D, Dey (ISIT 2009)

Estimation of Rearranged Smooth Processes



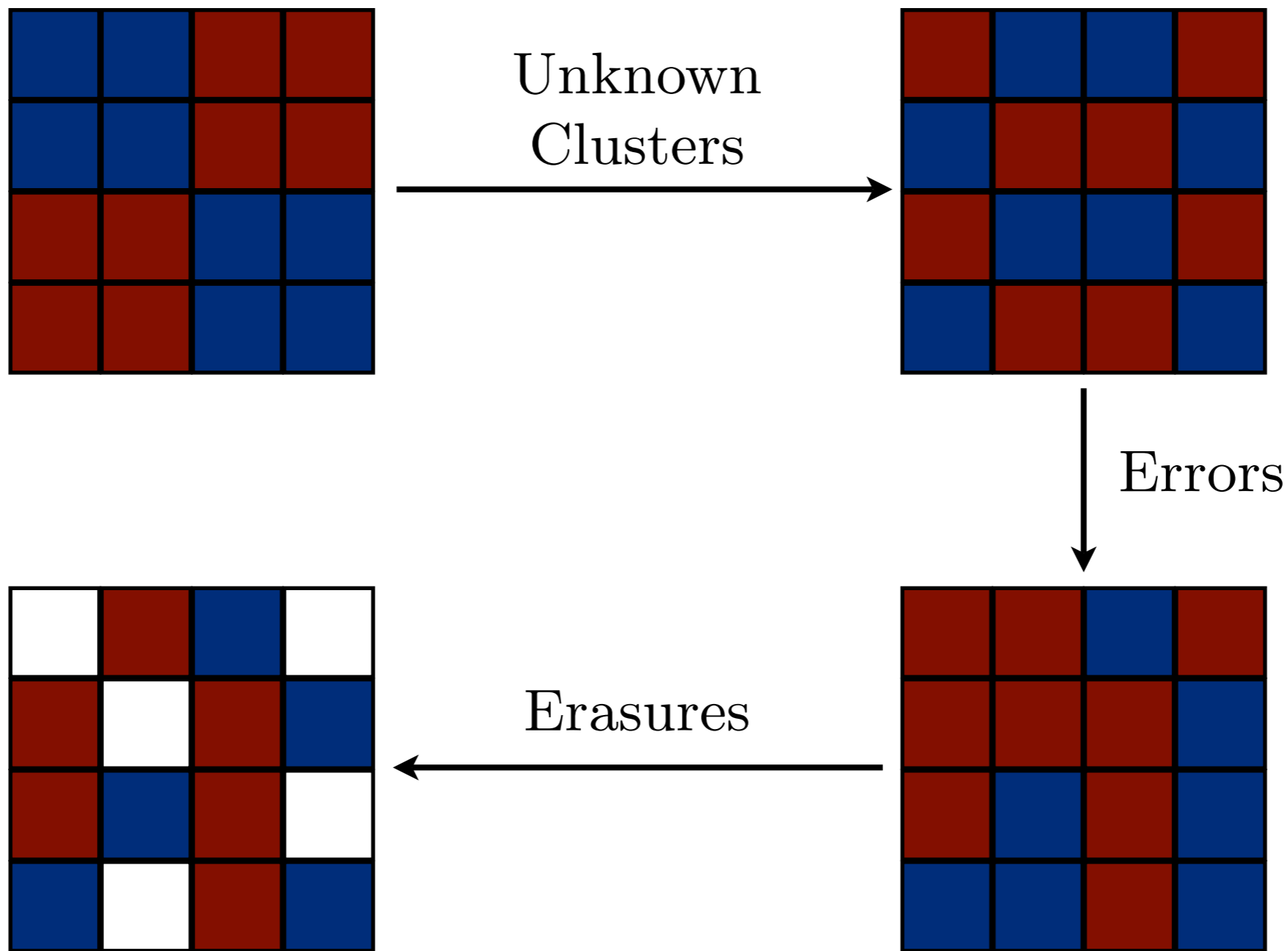
Introduced in Aditya, D, Dey (ISIT 2009)

Estimation of Rearranged Smooth Processes



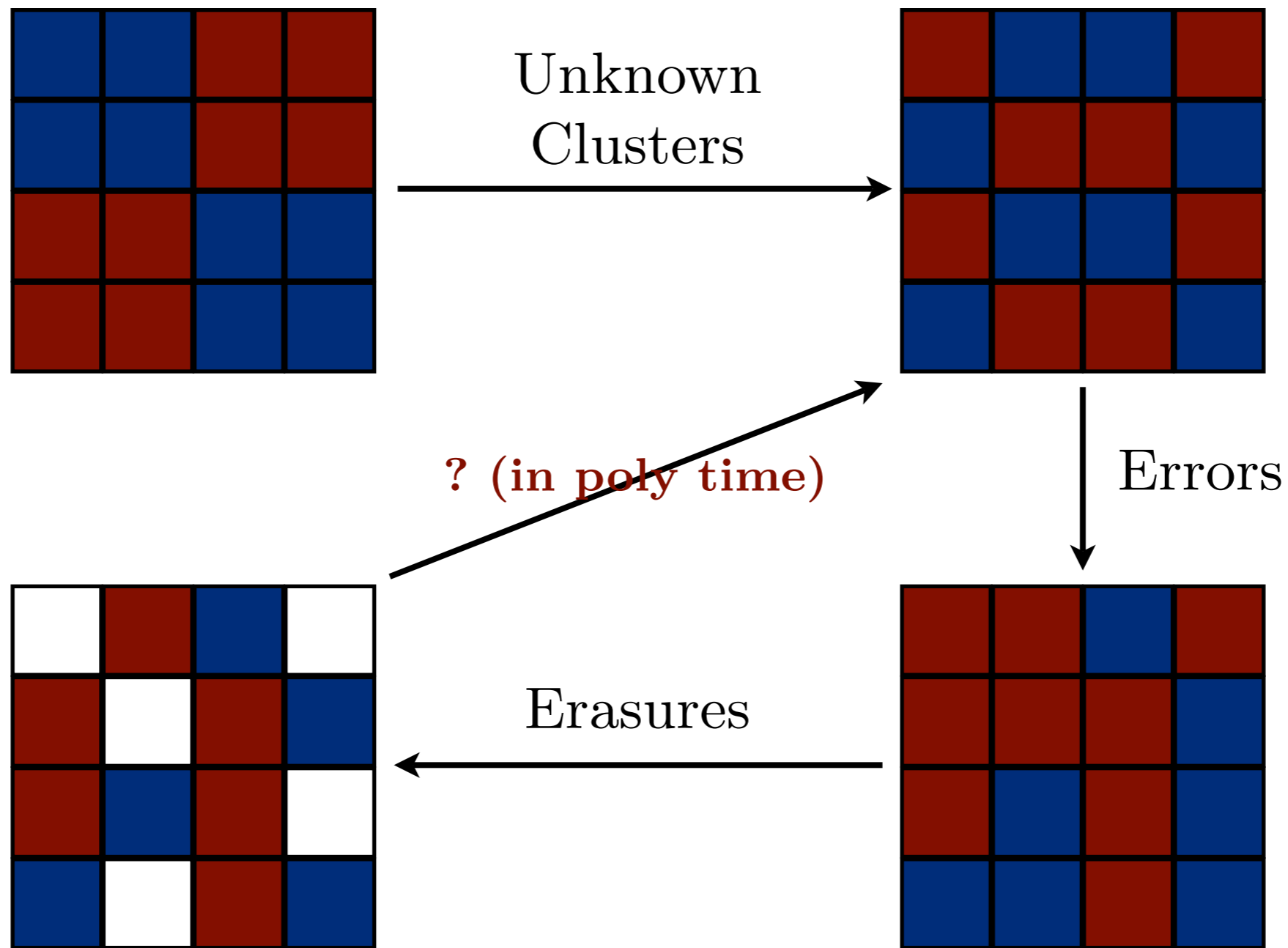
Introduced in Aditya, D, Dey (ISIT 2009)

Estimation of Rearranged Smooth Processes



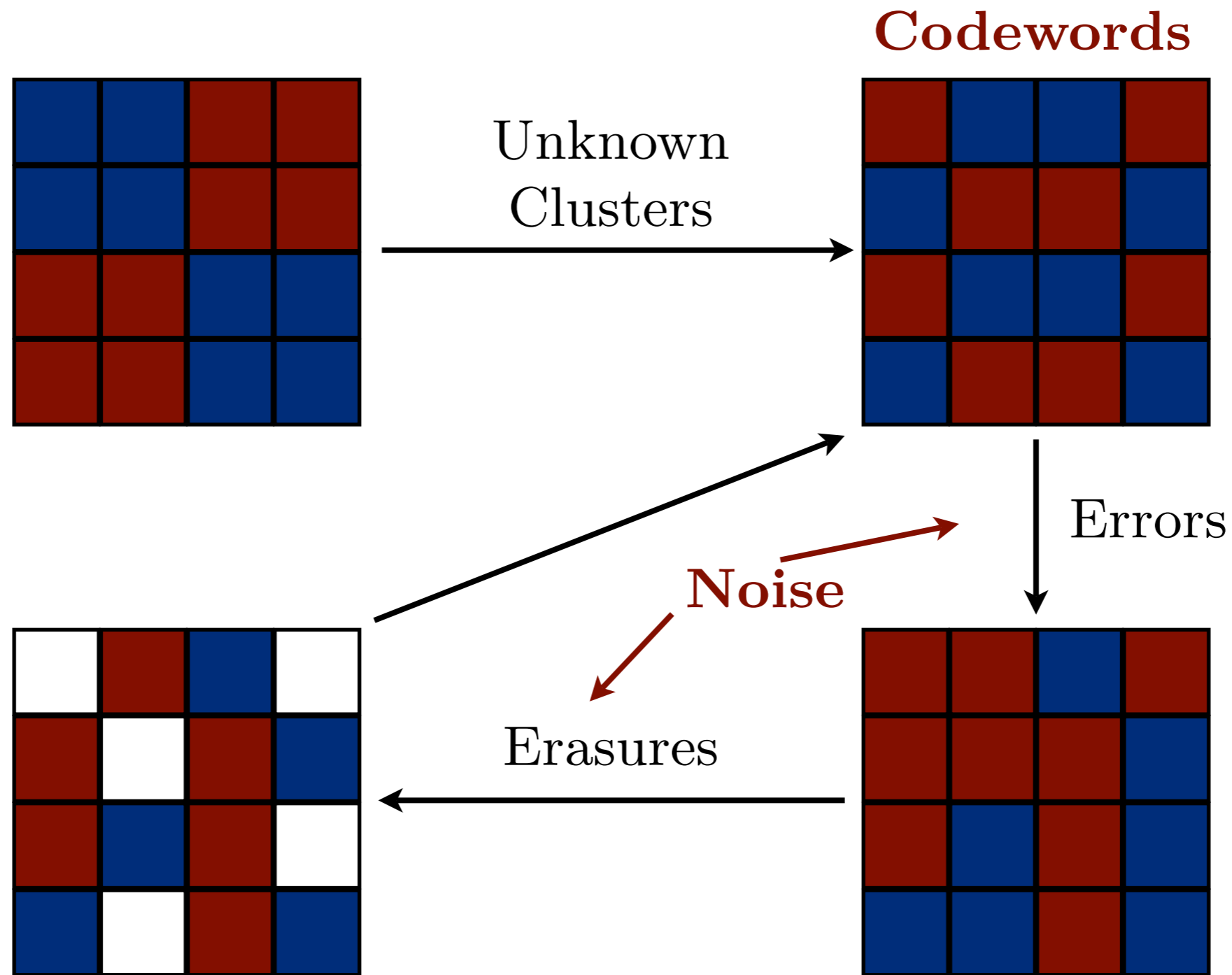
Introduced in Aditya, D, Dey (ISIT 2009)

Estimation of Rearranged Smooth Processes



Introduced in Aditya, D, Dey (ISIT 2009)

A Channel Coding Analogy



Threshold result? Structure of schemes operating at threshold?

Some Assumptions

- Number of rows, columns $\rightarrow \infty$
- Errors: Ratings flip with probability $p < 1/2$
- Erasure probability

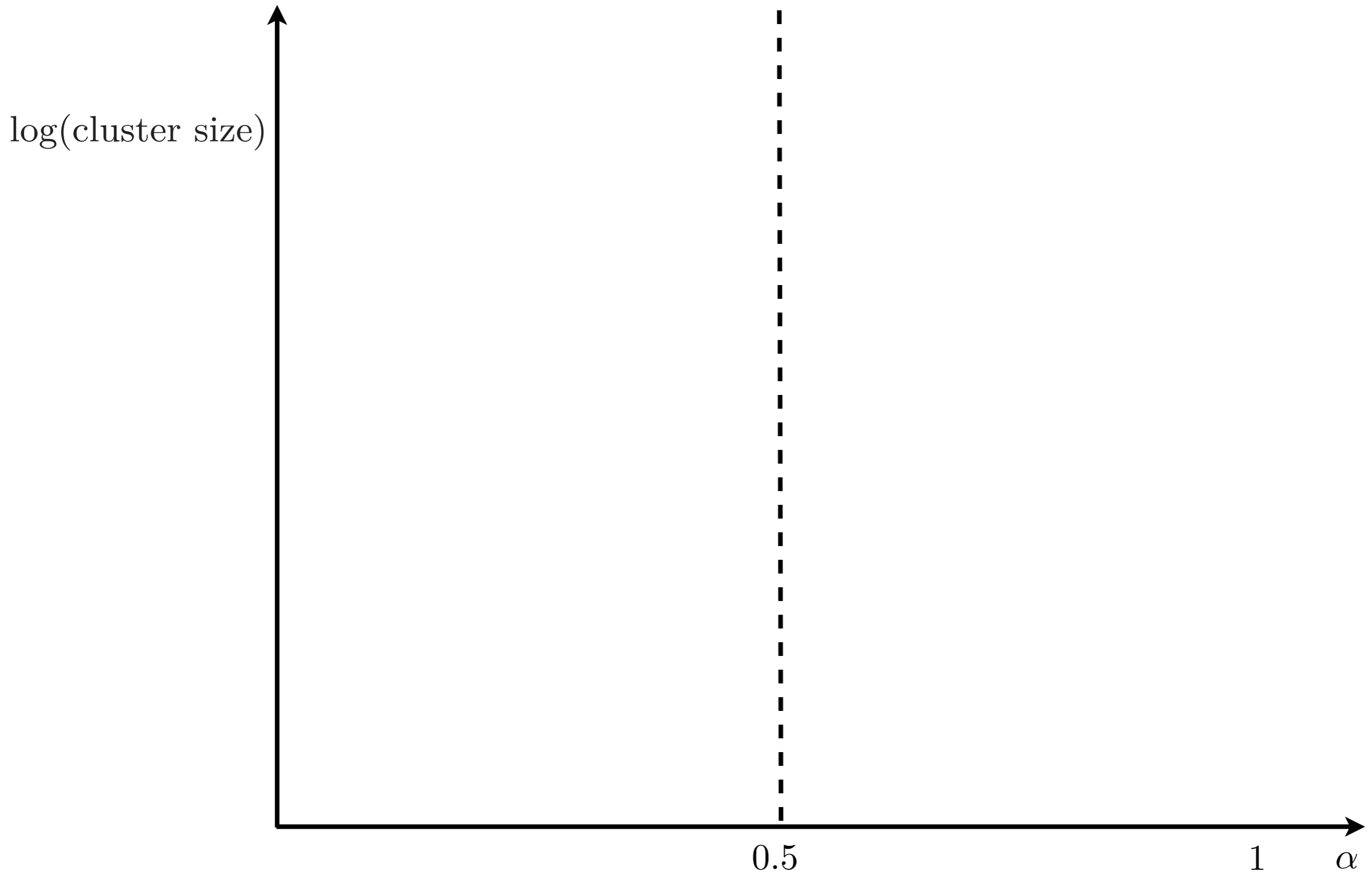
$$\epsilon = 1 - \frac{c}{n^\alpha}, \quad 0 \leq \alpha \leq 1$$

$\alpha < 1/2$: Near-quadratic regime

$\alpha > 1/2$: Near-linear regime

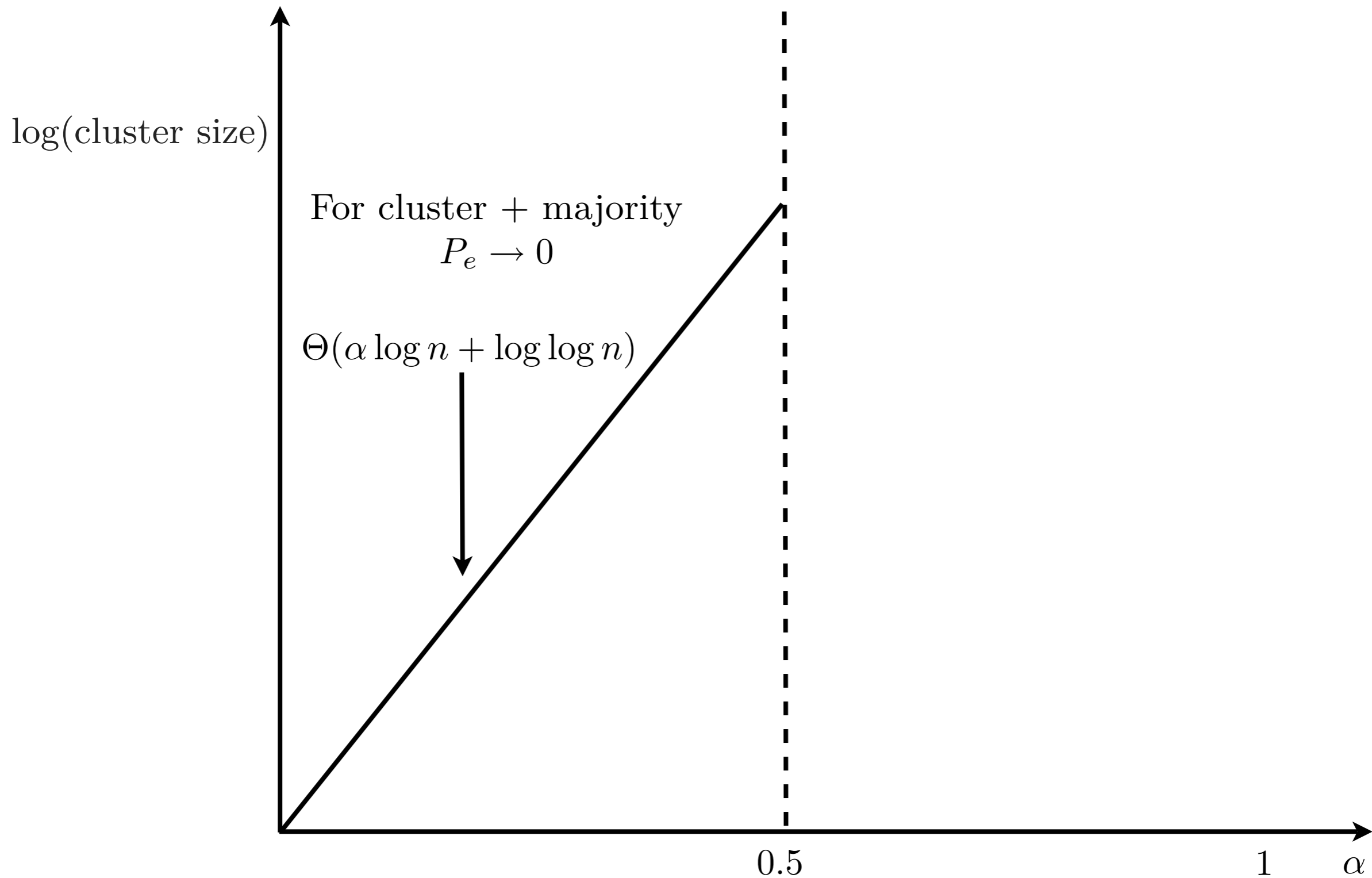
Recover Entire Matrix

(Aditya, Dabeer, Dey, IEEE Trans. Inform. Theory, Apr 2011)



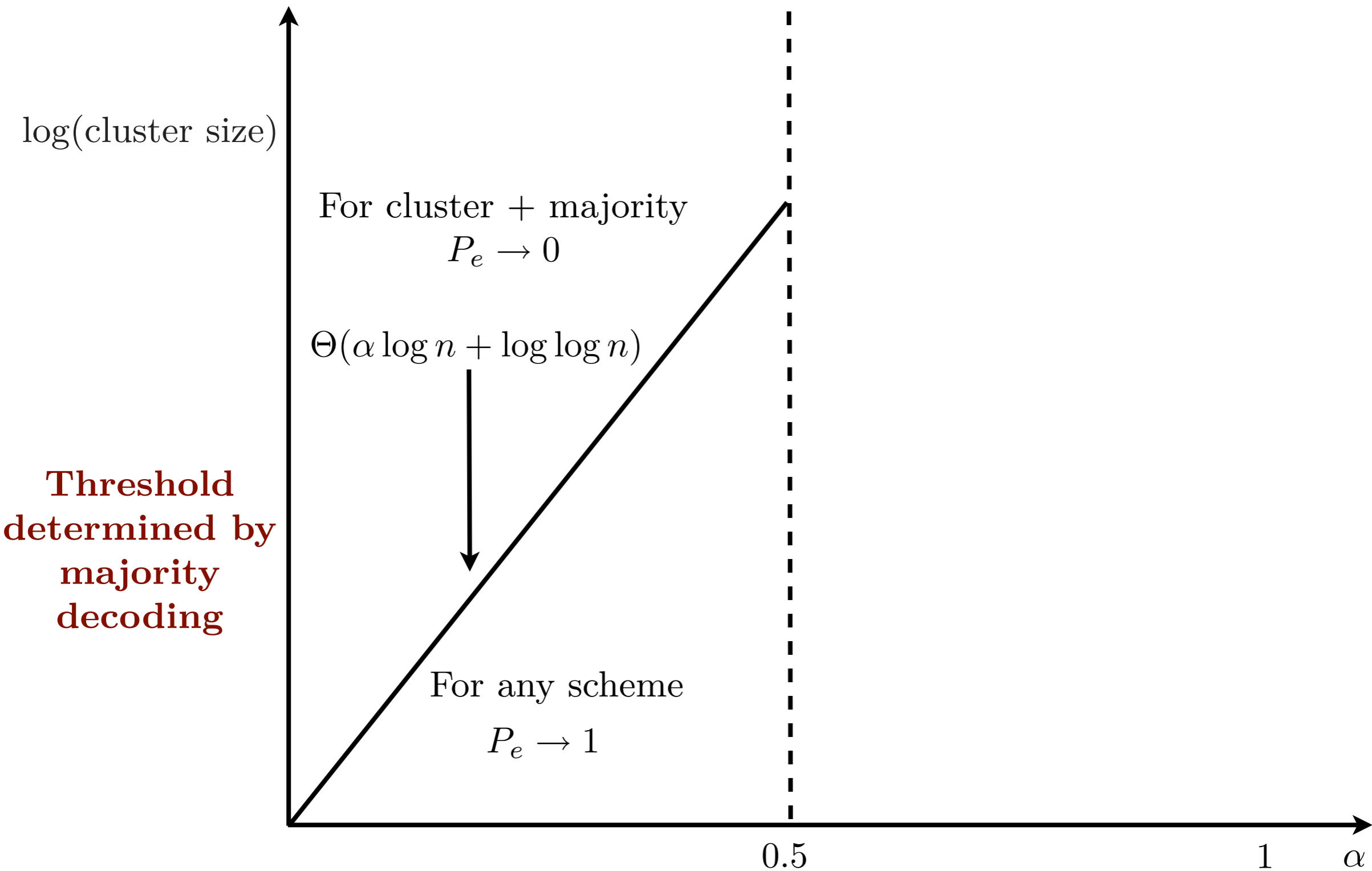
Recover Entire Matrix

(Aditya, Dabeer, Dey, IEEE Trans. Inform. Theory, Apr 2011)



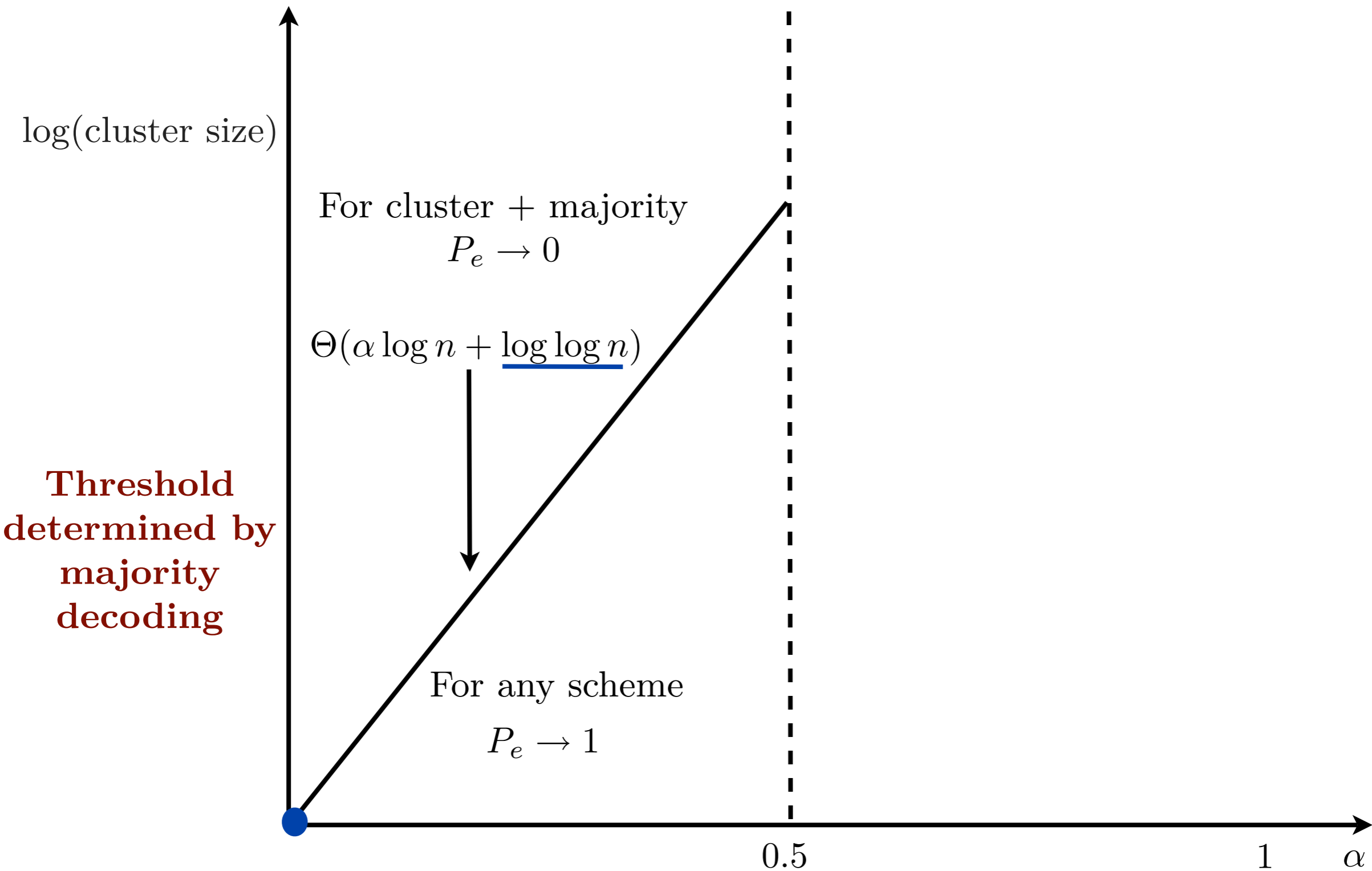
Recover Entire Matrix

(Aditya, Dabeer, Dey, IEEE Trans. Inform. Theory, Apr 2011)



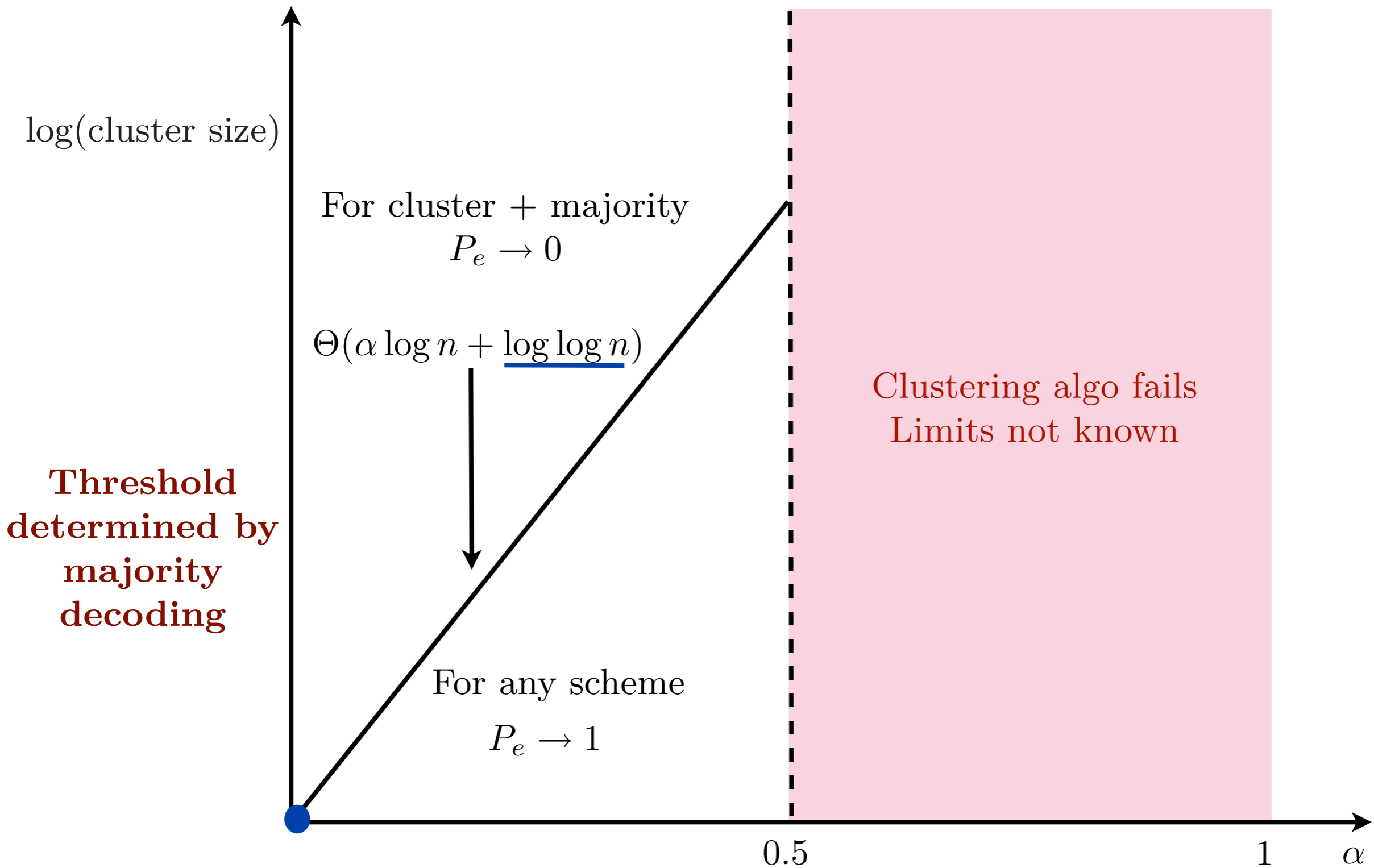
Recover Entire Matrix

(Aditya, Dabeer, Dey, IEEE Trans. Inform. Theory, Apr 2011)



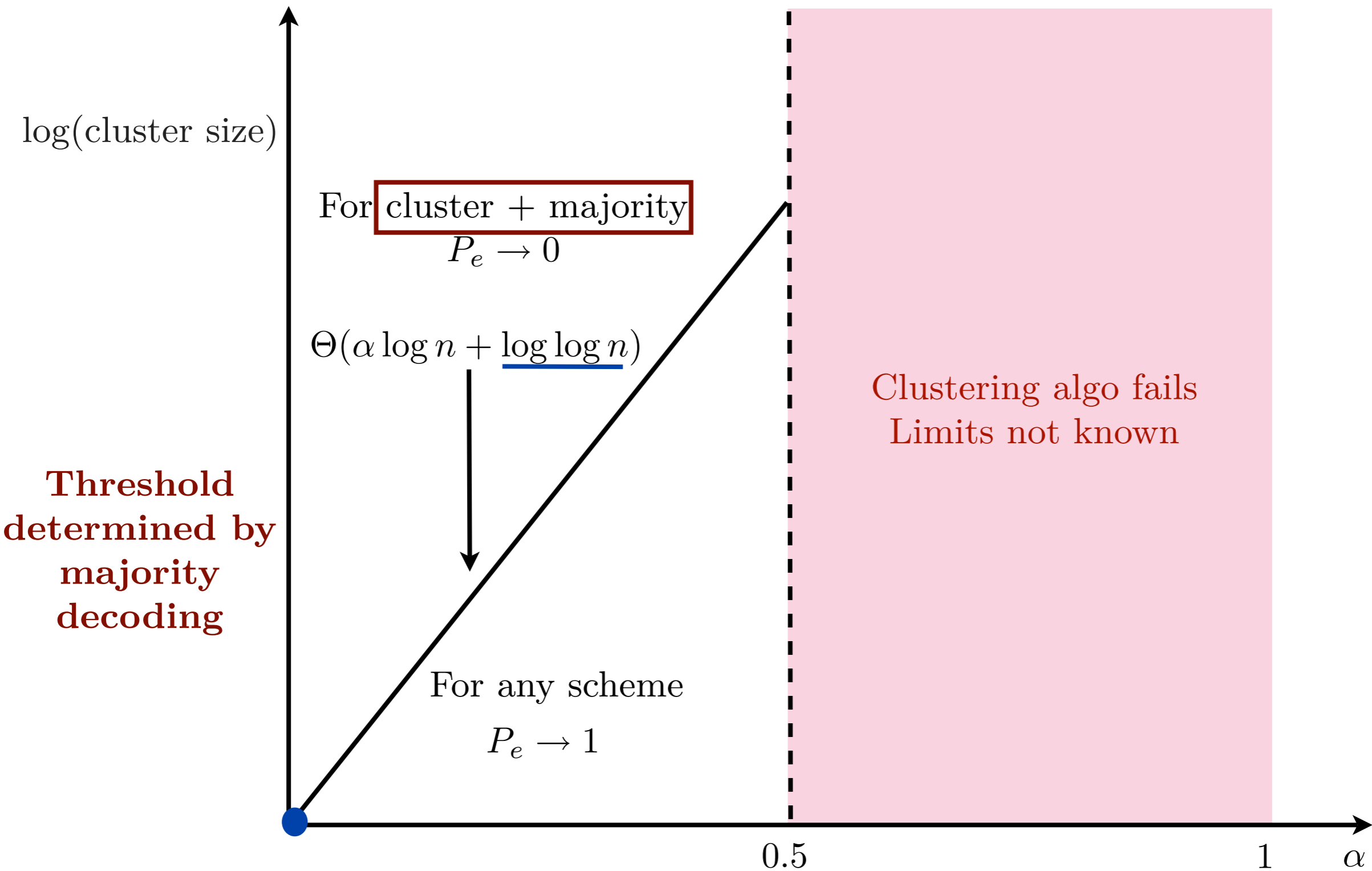
Recover Entire Matrix

(Aditya, Dabeer, Dey, IEEE Trans. Inform. Theory, Apr 2011)



Recover Entire Matrix

(Aditya, Dabeer, Dey, IEEE Trans. Inform. Theory, Apr 2011)



Implementation on Real Data

Implementation on Real Data

- Our clustering algo:
 - Similarity = $\frac{\# \text{commonly agreed entries}}{\# \text{common entries}}$
 - Threshold similarities

Implementation on Real Data

- Our clustering algo:
 - Similarity = $\frac{\# \text{commonly agreed entries}}{\# \text{common entries}}$
 - Threshold similarities
- Good for analysis

Implementation on Real Data

- Our clustering algo:
 - Similarity = $\frac{\# \text{commonly agreed entries}}{\# \text{common entries}}$
 - Threshold similarities
- Good for analysis
- But bad for implementation as may not lead to clusters on finite data

Implementation on Real Data

- Our clustering algo:
 - Similarity = $\frac{\# \text{commonly agreed entries}}{\# \text{common entries}}$
 - Threshold similarities
- Good for analysis
- But bad for implementation as may not lead to clusters on finite data
- Clustering both rows and columns is computationally intensive

The Popularity Amongst Friends (PAF) Algorithm

(Barman, D, ISIT 2010; submitted to IEEE Trans. Inform. Theory)

PAF for Binary Ratings

PAF for Binary Ratings

- **Algorithm:**

- For user 1, find top K similar users
- Similarity = # agreements in available ratings
- Recommend an unseen item that is most popular amongst these K users

PAF for Binary Ratings

- **Algorithm:**
 - For user 1, find top K similar users
 - Similarity = # agreements in available ratings
 - Recommend an unseen item that is most popular amongst these K users
- Related to practice (example: Amazon)

PAF for Binary Ratings

- **Algorithm:**
 - For user 1, find top K similar users
 - Similarity = # agreements in available ratings
 - Recommend an unseen item that is most popular amongst these K users
- Related to practice (example: Amazon)
- Not matrix completion

PAF for Binary Ratings

- **Algorithm:**
 - For user 1, find top K similar users
 - Similarity = # agreements in available ratings
 - Recommend an unseen item that is most popular amongst these K users
- Related to practice (example: Amazon)
- Not matrix completion
- Low complexity
 - User node degree \ll Total number of users
 - Simple updates

Experimental Setup

Experimental Setup

- Rating quantization
 - 4,5 mapped to 1 (yes), 1-3 mapped to 0 (no)

Experimental Setup

- Rating quantization
 - 4,5 mapped to 1 (yes), 1-3 mapped to 0 (no)
- Hide 30% of data per user; can compute metrics only when recommended item is in the hidden list

Experimental Setup

- Rating quantization
 - 4,5 mapped to 1 (yes), 1-3 mapped to 0 (no)
- Hide 30% of data per user; can compute metrics only when recommended item is in the hidden list
- Comparison with OptSpace (Keshavan et al, 2008)
 - Representative of matrix algorithms
 - Evaluated only on items recommended by local algorithm
 - Unquantized input; output quantized for BER

Empirical Performance

(Barman, Dabeer, ISIT 2010; submitted to IEEE Trans. Inform. Theory)

Movielens

	BER	MAE	RMSE
PAF	0.103	0.627	0.748
OptSpace	0.108	0.581	0.733

**Naive Estimate:
For Local Algorithm,
to compute
RMSE, MAE
1 is mapped to 4.5**

Empirical Performance

(Barman, Dabeer, ISIT 2010; submitted to IEEE Trans. Inform. Theory)

Movielens

	BER	MAE	RMSE
PAF	0.103	0.627	0.748
OptSpace	0.108	0.581	0.733

**Naive Estimate:
For Local Algorithm,
to compute
RMSE, MAE
1 is mapped to 4.5**

Snapshot of Netflix (2000)

	BER	MAE	RMSE
PAF	0.18	0.742	0.942
OptSpace	0.19	0.590	0.742

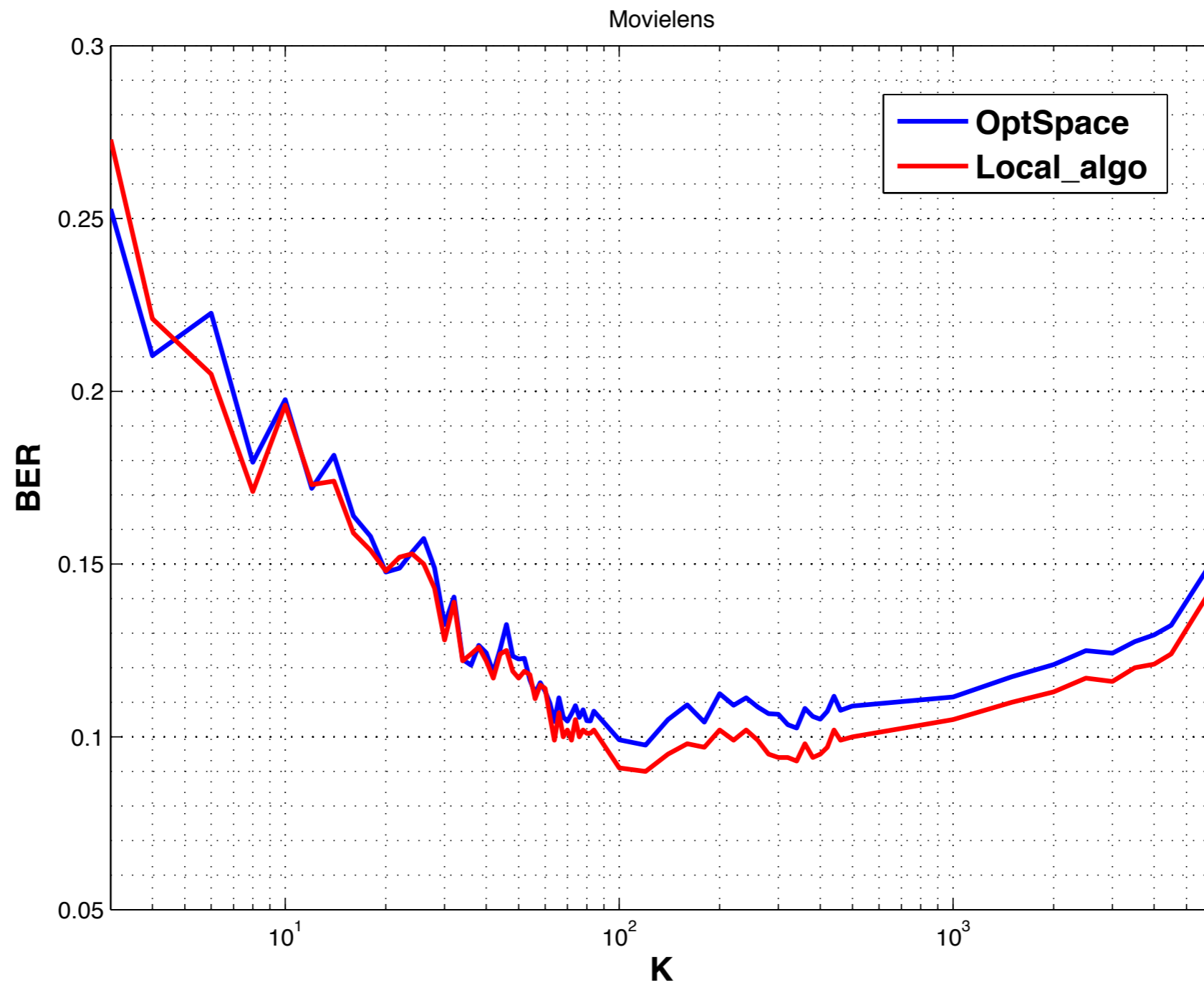
**Netflix has higher
percentage of
low ratings**

Empirical Performance (Contd.)

Movielens After Removing Popular Movies
(those with 60% or more 1's)

	BER	MAE	RMSE
PAF	0.335	0.709	1.010
OptSpace	0.327	0.718	0.901

Empirical Performance (Contd.)



Remarks

Remarks

- PAF competitive for BER

Remarks

- PAF competitive for BER
- Are MAE, RMSE relevant?
 - Scale 1-5, RMSE 0.7+ - poor confidence intervals
 - Noisy data and difficult to squeeze out more than 1 bit information

Remarks

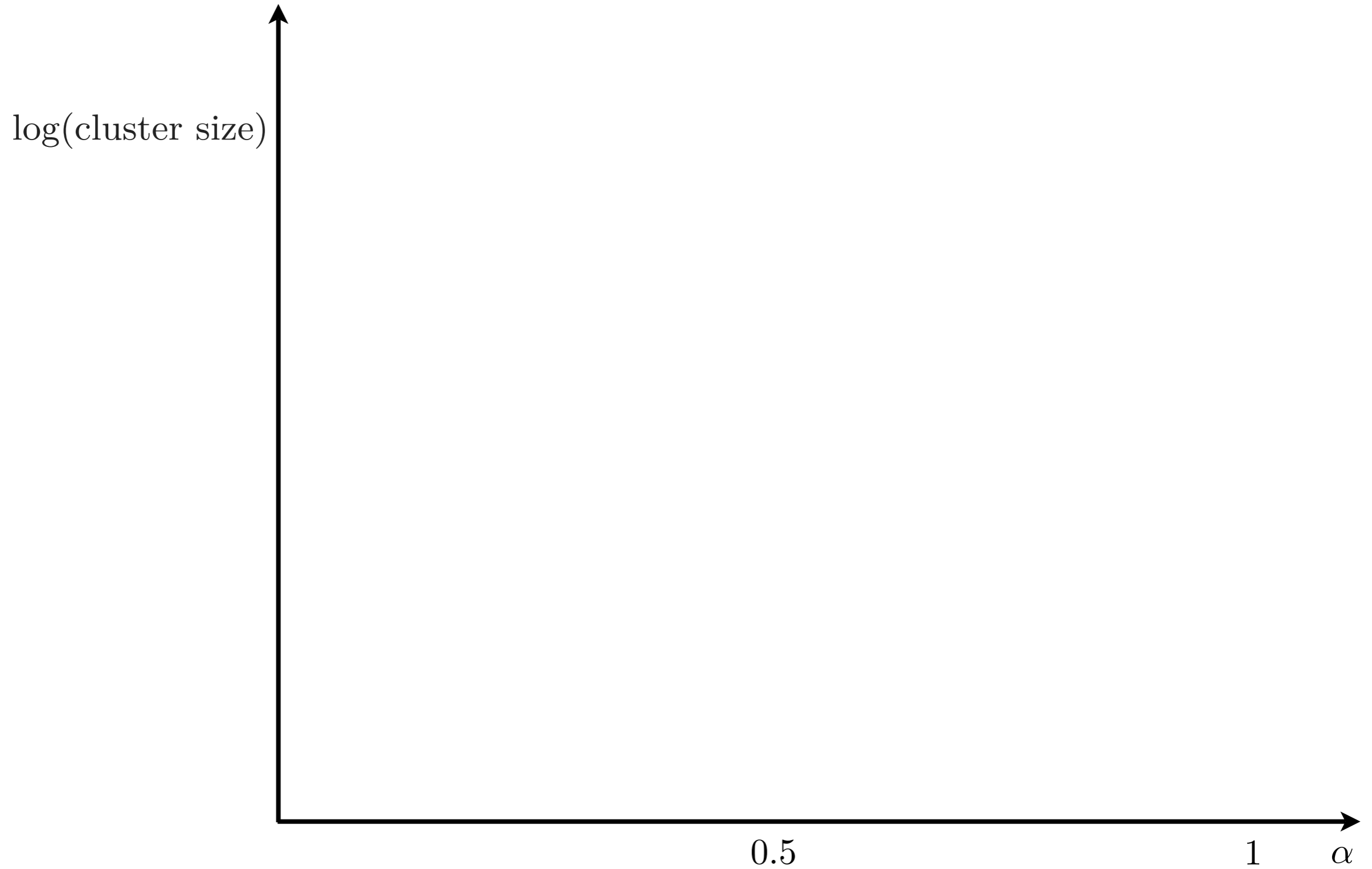
- PAF competitive for BER
- Are MAE, RMSE relevant?
 - Scale 1-5, RMSE 0.7+ - poor confidence intervals
 - Noisy data and difficult to squeeze out more than 1 bit information
- 2-10 times faster than OptSpace
 - Also amenable to recursive update

Remarks

- PAF competitive for BER
- Are MAE, RMSE relevant?
 - Scale 1-5, RMSE 0.7+ - poor confidence intervals
 - Noisy data and difficult to squeeze out more than 1 bit information
- 2-10 times faster than OptSpace
 - Also amenable to recursive update
- Any provably good properties?

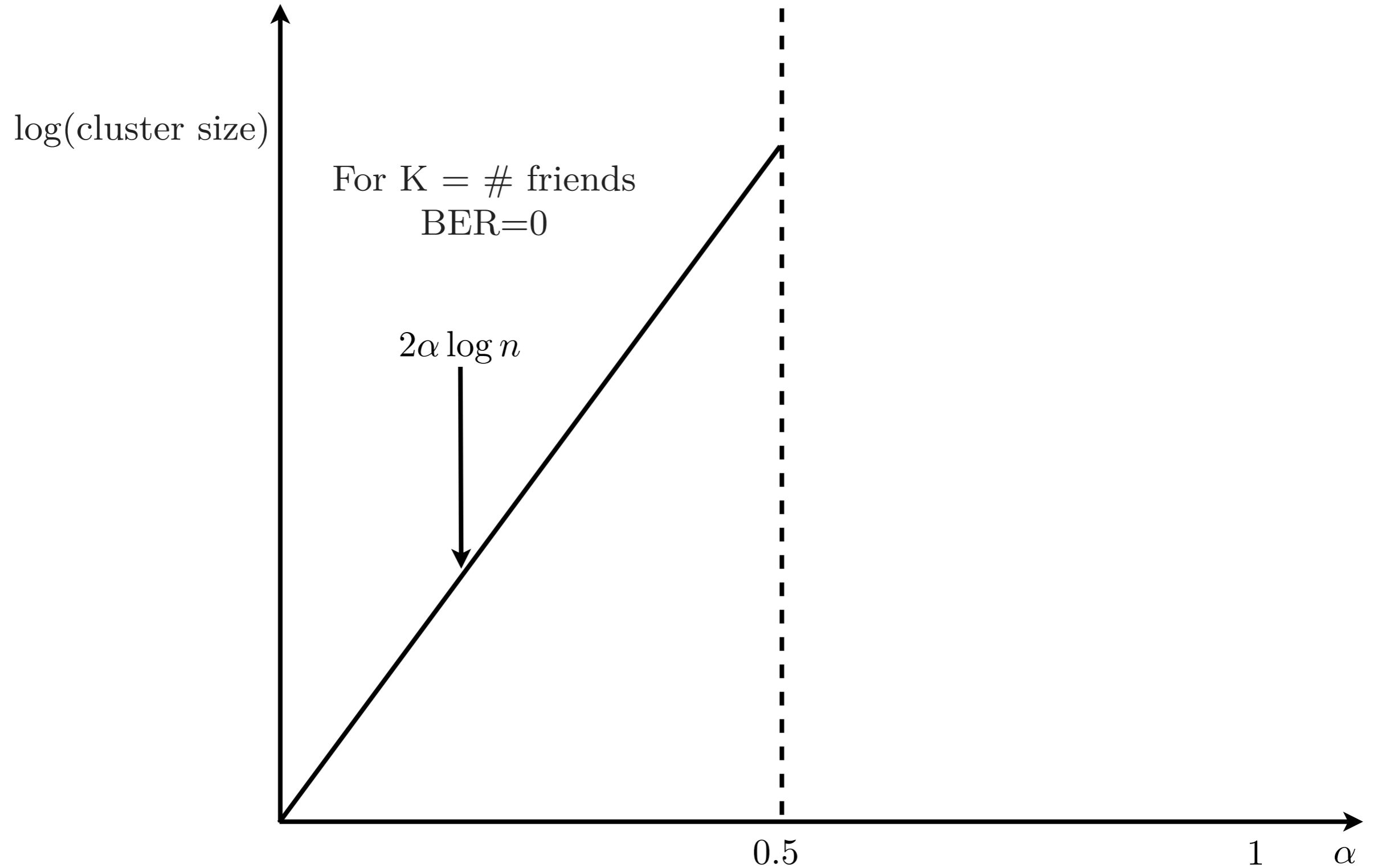
Asymptotic BER of PAF

(Barman, Dabeer, ISIT 2010; submitted to IEEE Trans. Inform. Theory)



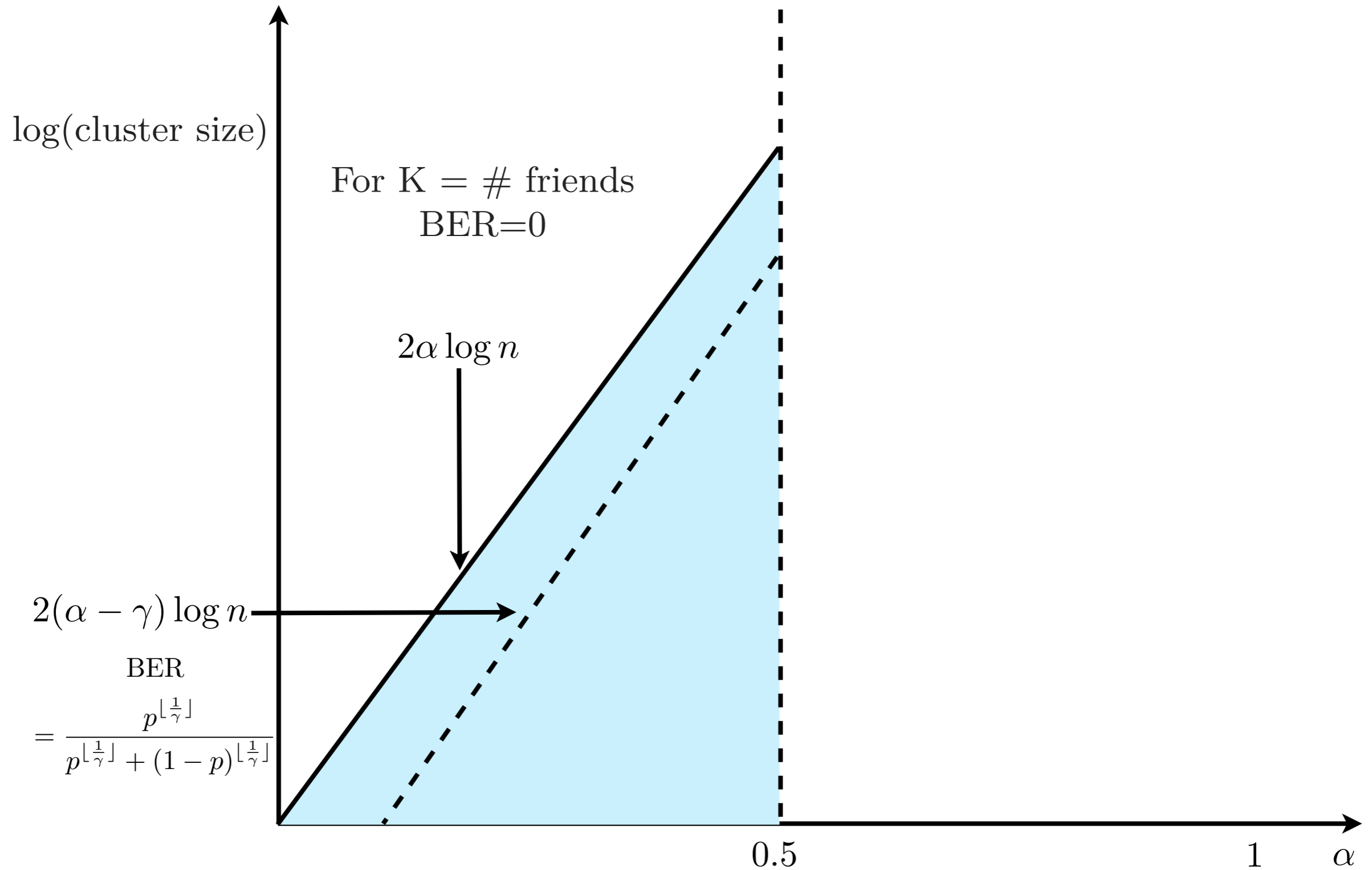
Asymptotic BER of PAF

(Barman, Dabeer, ISIT 2010; submitted to IEEE Trans. Inform. Theory)



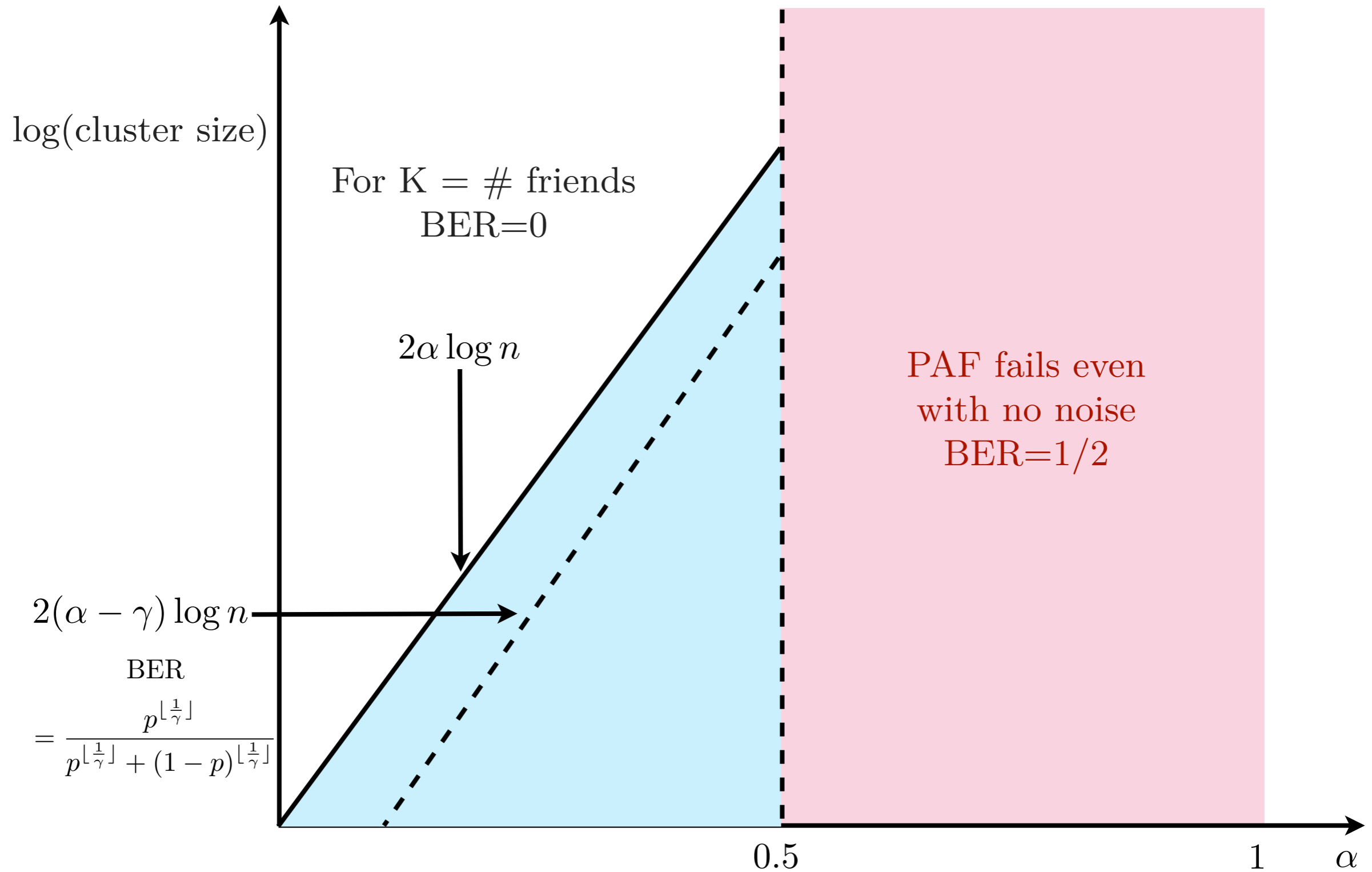
Asymptotic BER of PAF

(Barman, Dabeer, ISIT 2010; submitted to IEEE Trans. Inform. Theory)



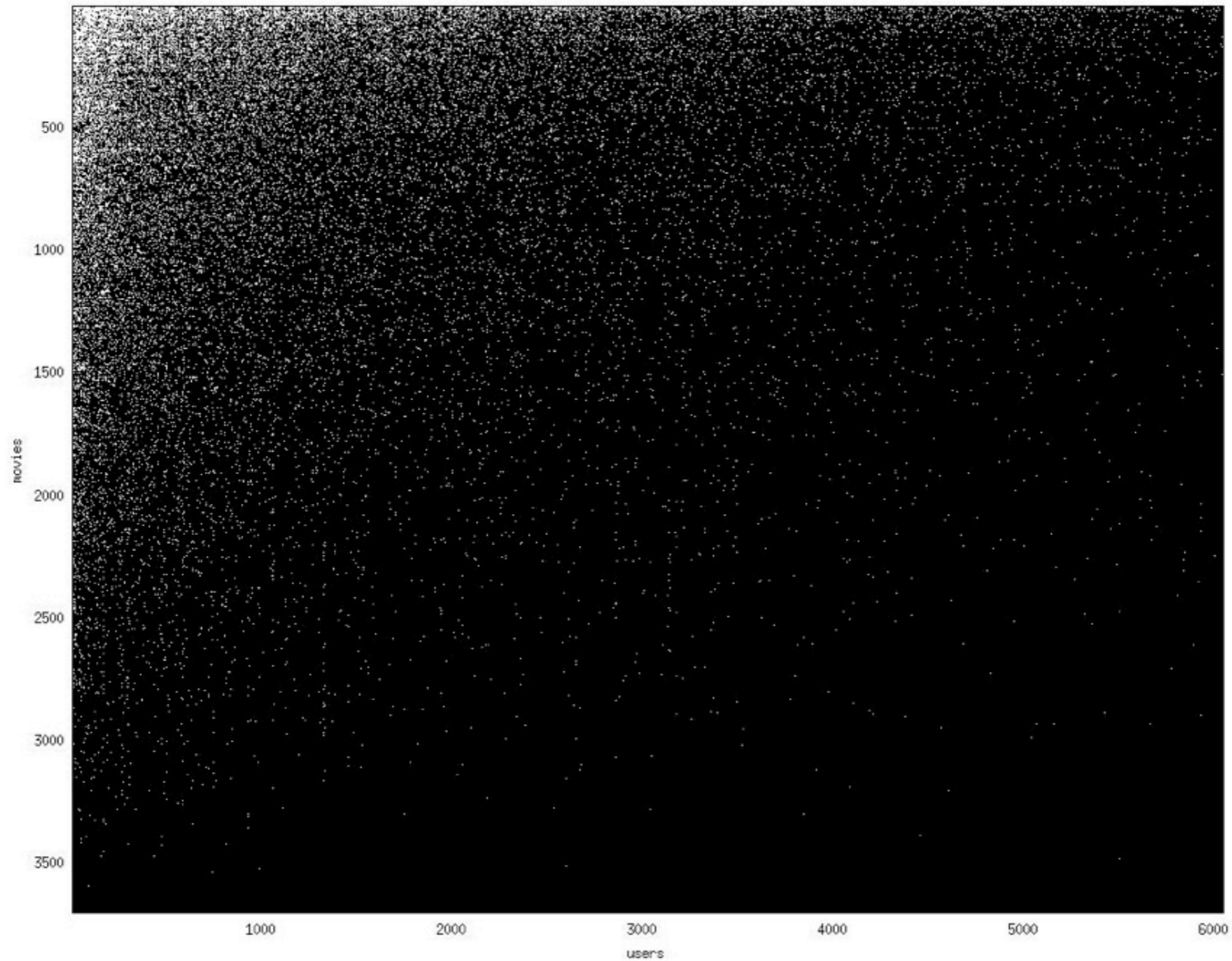
Asymptotic BER of PAF

(Barman, Dabeer, ISIT 2010; submitted to IEEE Trans. Inform. Theory)

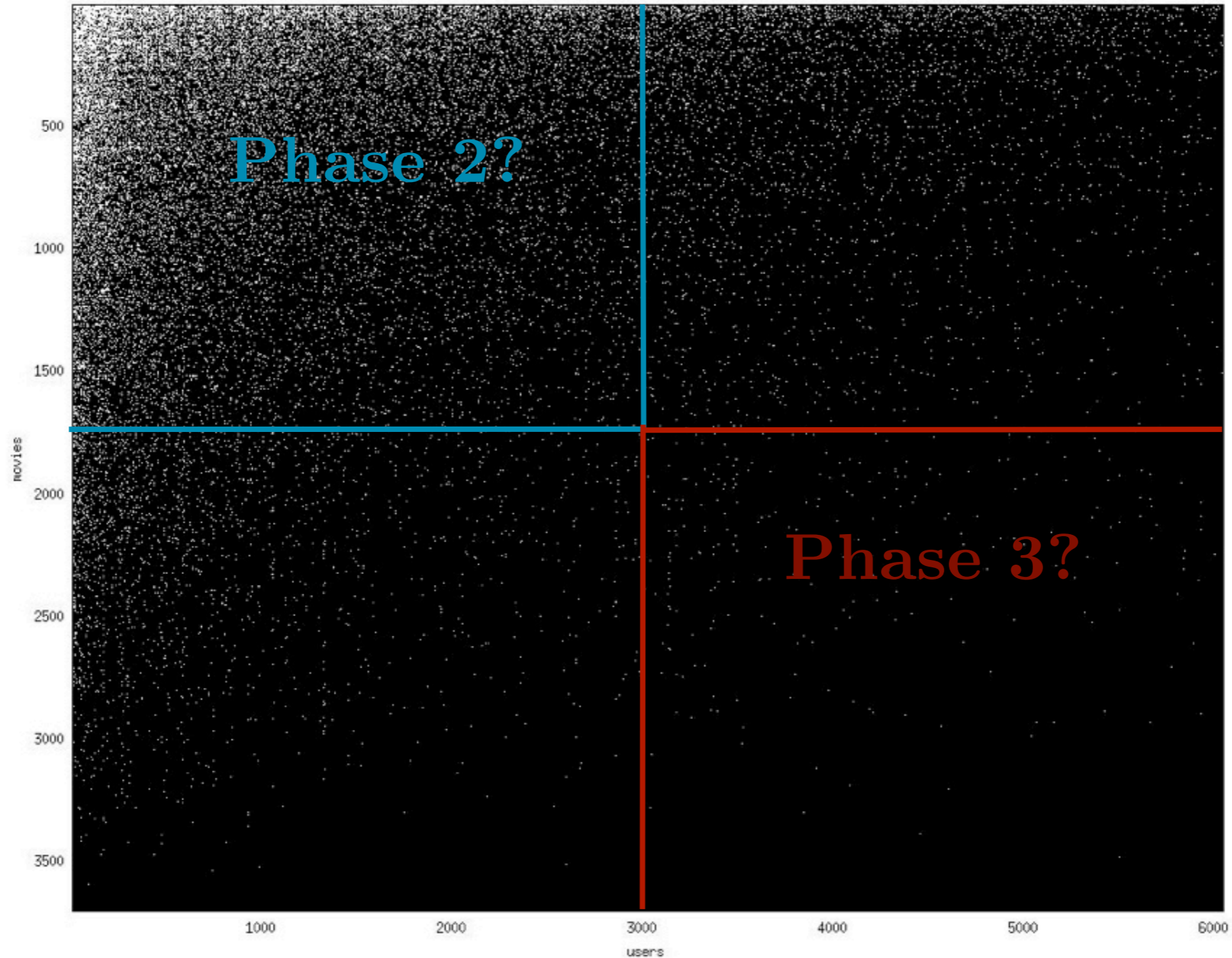


Where Does Real Data Live?

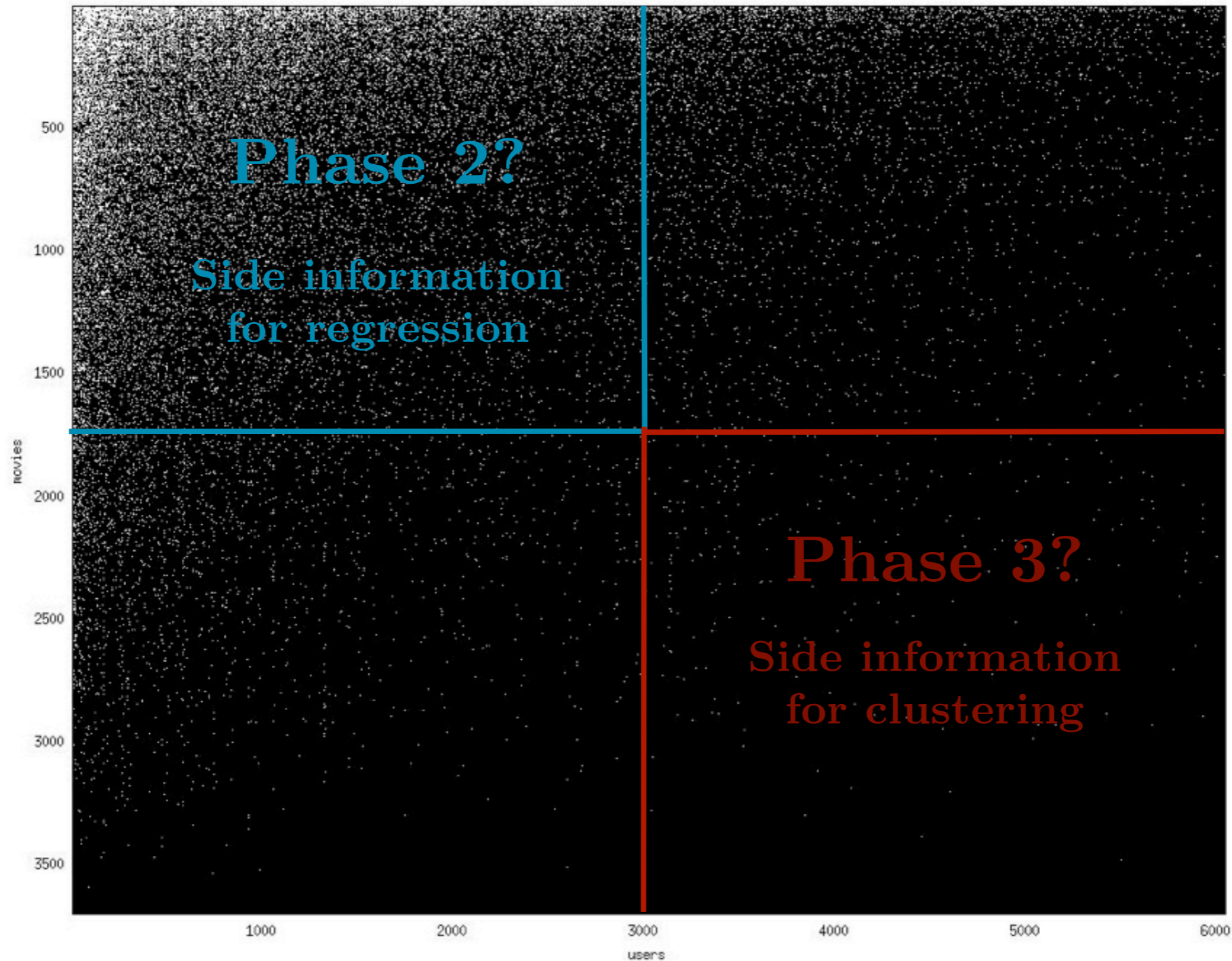
The Movielens Matrix



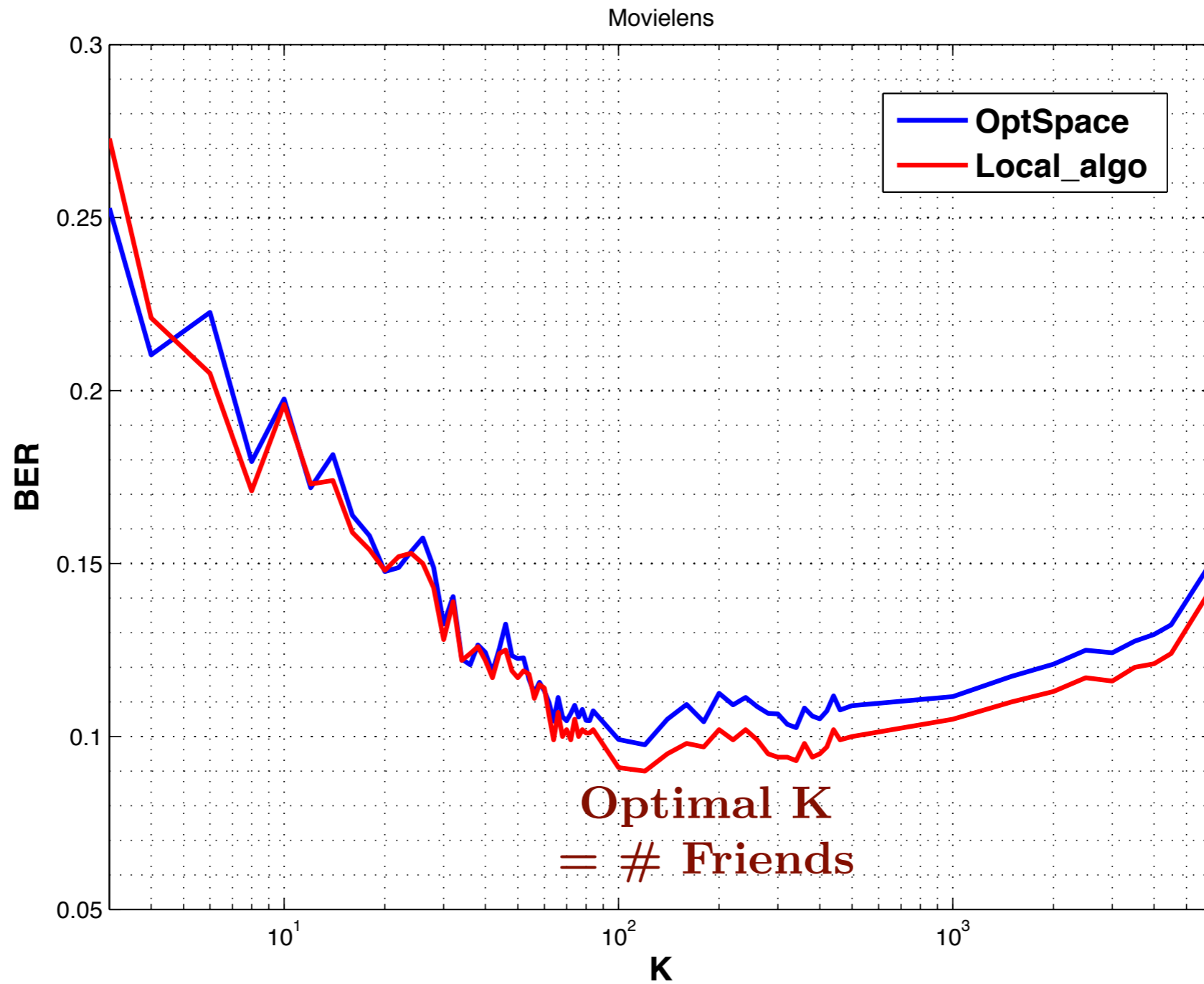
The Movielens Matrix



The Movielens Matrix



Degrees of Freedom



Remarks

Remarks

- PAF algorithm is scalable and competitive

Remarks

- PAF algorithm is scalable and competitive
- Provably good in near-quadratic regime
 - BER bounded away from $1/2$

Remarks

- PAF algorithm is scalable and competitive
- Provably good in near-quadratic regime
 - BER bounded away from $1/2$
- Near-linear regime: Blend side-information

Remarks

- PAF algorithm is scalable and competitive
- Provably good in near-quadratic regime
 - BER bounded away from $1/2$
- Near-linear regime: Blend side-information
- Future directions
 - Sampling - account for heavy tails
 - Choosing list size - expert selection?
 - Rearranged general “smooth” processes?
 - Incorporating side information?
 - Temporal dynamics

More Details

- S. T. Aditya, O. Dabeer, and B. K. Dey, “A channel coding perspective of recommendation systems,” *ISIT* 2009, June 2009, Seoul, Korea
- S. T. Aditya, O. Dabeer, and B. K. Dey, “A channel coding perspective of collaborative filters,” *IEEE Transactions on Information Theory*, April 2011
- K. Barman, O. Dabeer, “Local Popularity Based Collaborative Filters,” *ISIT*, June 2010, Austin, USA
- K. Barman, O. Dabeer, “What is Popular Amongst Your Friends?” submitted to *IEEE Transactions on Information Theory*, July 2010 (arXiv:1006.1772)
- O. Dabeer, “Collaborative Filtering,” Tutorial, *National Communications Conference*, IISc Bangalore, Jan 2011 (<http://www.ncc.org.in/ncc2011/tutorials.html>)