ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Multi-Assignment Clustering for Boolean Data

Andreas P. Streich, Mario Frank, David Basin and Joachim M. Buhmann

{andreas.streich, mario.frank, basin, jbuhmann}@inf.ethz.ch

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Multi-Assignment Clustering for Boolean Data

Andreas P. Streich **Mario Frank, David Basin and Joachim M. Buhmann**

{andreas.streich, mario.frank, basin, jbuhmann}@inf.ethz.ch

Why Multi-Assignments? – an Example

Cluster A: The Grad Students

- ✓ smart
- ✓ thinks about research 24/7
- ✓ prone to procrastination

Cluster B: The Comic Characters

- ✓ has less than 10 colors
- ✓ never gets older
- ✓ makes balloons when speaking

?

?

x_i.= Cecilia*

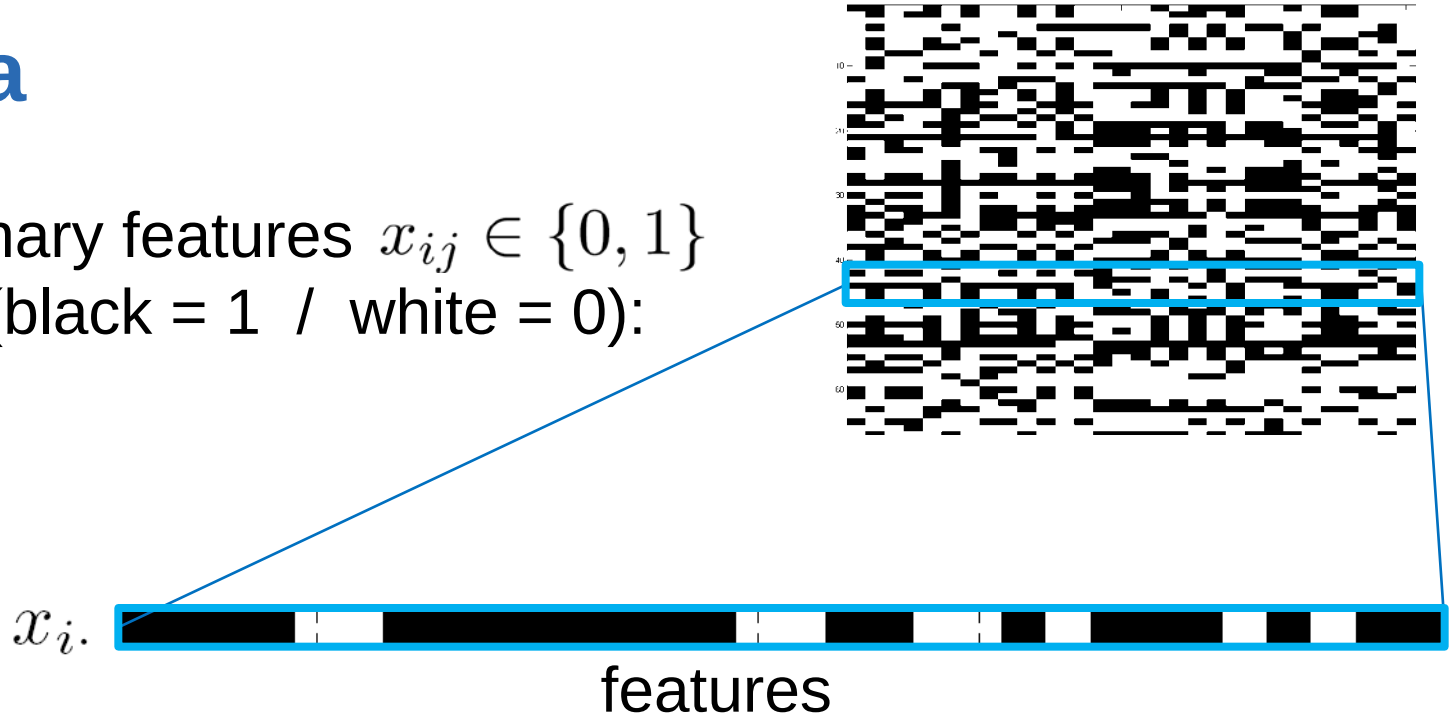
Assign her to:

- A only ?
- B only ?
- 42% A and 58% B ?
- A and B !

*by courtesy of **"Piled Higher and Deeper" by Jorge Cham, www.phdcomics.com**

Data

- Binary features $x_{ij} \in \{0, 1\}$
  (black = 1 / white = 0):

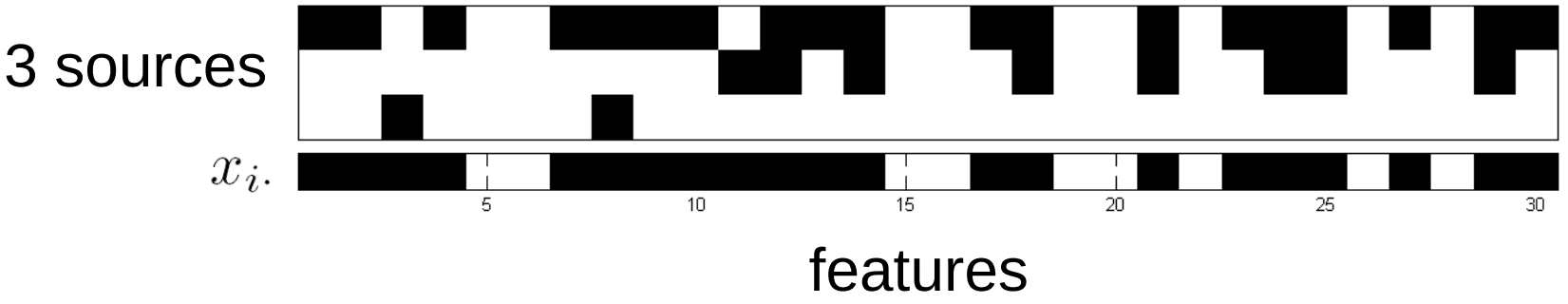


- Each object $x_{i\cdot}$ (e.g. Cecilia) is a binary vector.

Model - Logical Structure

- Objects are characterized by disjunctions of Boolean emissions:

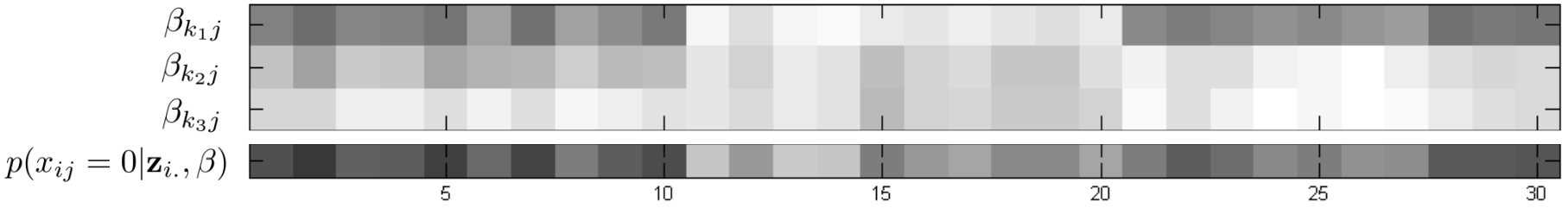
- Example for one object $x_{i\cdot}$:



Model – Probabilistic Representation

- Assignments: $z_{ik} = 1$: object i belongs to k
- Centroids: $\beta_{kj} = p(\text{cluster } k \text{ emits a } 0 \text{ at dimension } j)$
- Signal distribution

$$p_S\left(x_{ij} \mid \mathbf{z}, \beta\right) = \left[1 - \prod_{k=1}^{K} \beta_{kj}^{z_{ik}}\right]^{x_{ij}} \left[\prod_{k=1}^{K} \beta_{kj}^{z_{ik}}\right]^{1-x_{ij}}$$



Model – Combination with Noise Distribution

- Noise distribution: $\text{Bernoulli}(x_{ij}|r)$
- Combined model:

$$p_M(x_{ij} \mid \mathbf{z}_{i.}, \beta, r, \xi_{ij}) = \xi_{ij}\text{Bernoulli}(x_{ij}|r) + (1-\xi_{ij})p_S(x_{ij} \mid \mathbf{z}_{i.}, \beta)$$

- The binary ξ_{ij} indicate for each dyad x_{ij} whether it is drawn from the signal distribution or from the noise distribution. With $p(\xi_{ij}|\epsilon) = \text{Bernoulli}(\xi_{ij}|\epsilon)$

Model – Full Data Likelihood

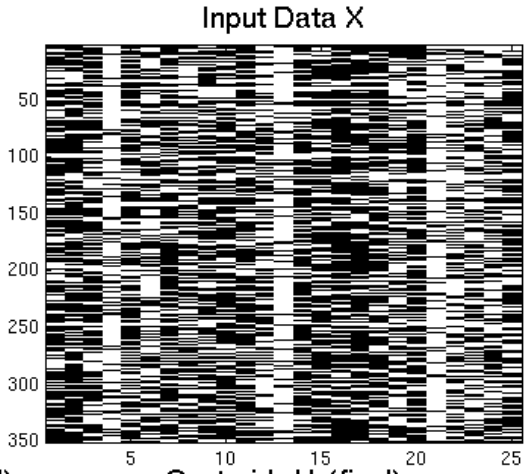
- Marginalizing out the ξ_{ij}s we get the full data likelihood:

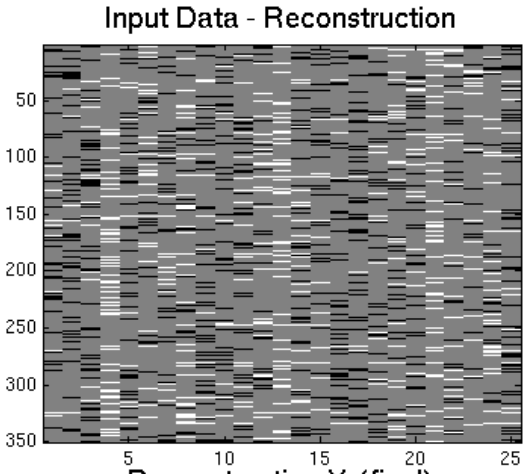
$$p_M(\mathbf{x} \mid \mathbf{z}, \beta, r, \epsilon) = \sum_{\{\xi\}} p_M(\mathbf{x}, \xi \mid \mathbf{z}, \beta, r, \epsilon)$$

$$= \prod_{i,j} (\epsilon \cdot p_N(x_{ij}) + (1-\epsilon) \cdot p_S(x_{ij}))$$

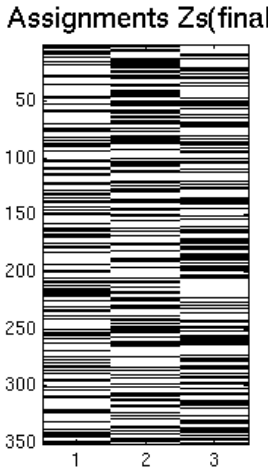
- Inference: Deterministic Annealing

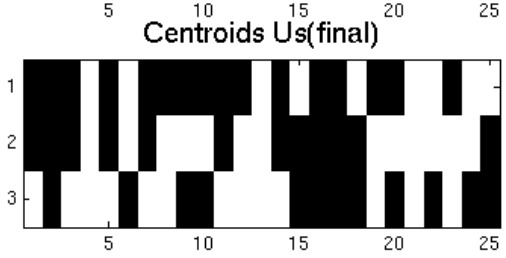
Experiments - A Clustering Result



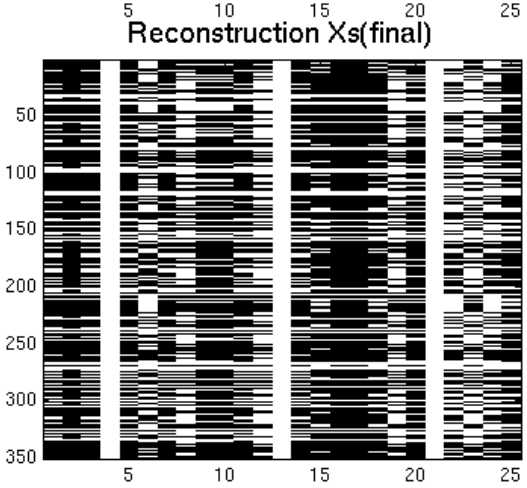


1 = black
0 = grey
-1 = white

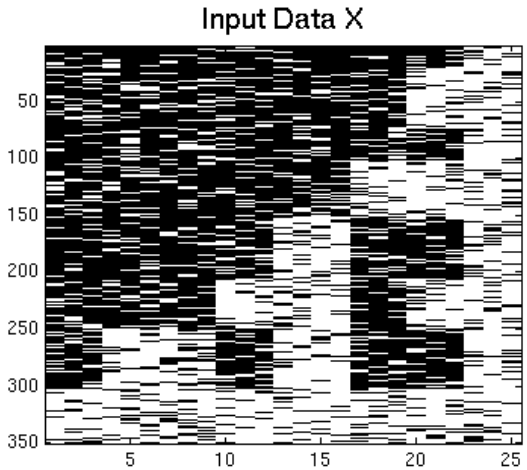


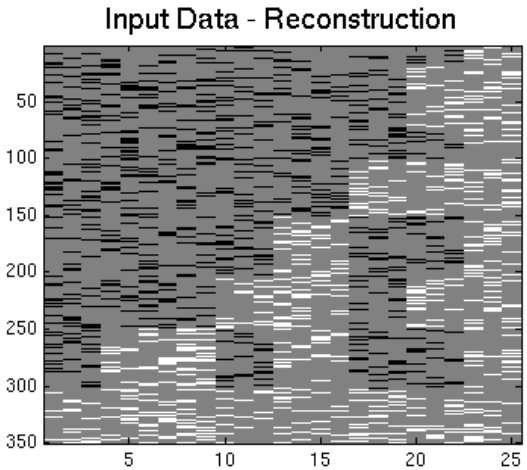


1 = black
0 = white

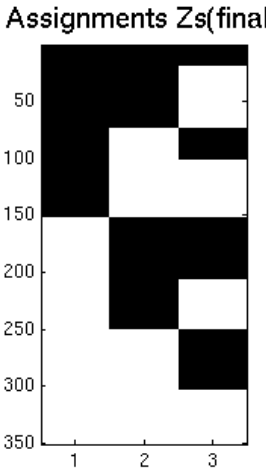


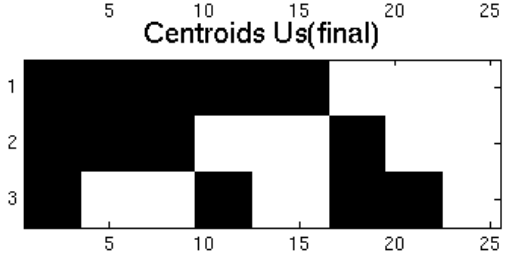
Experiments - A Clustering Result - sorted



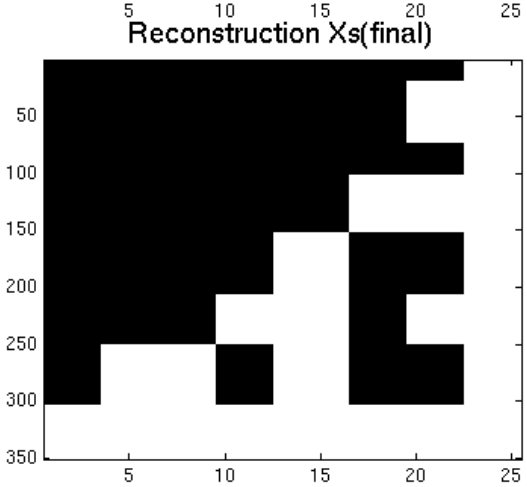


1 = black
0 = grey
-1 = white

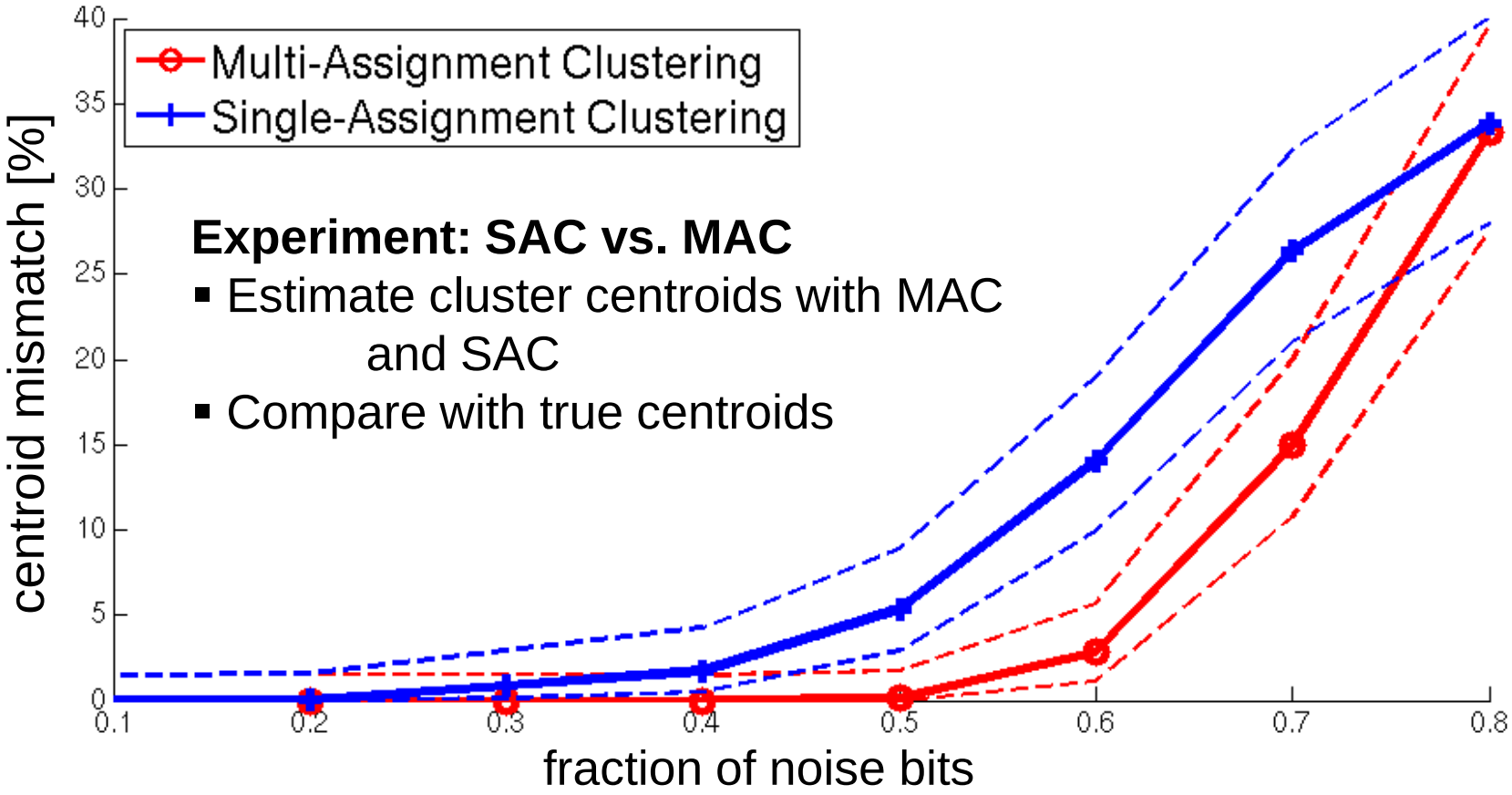




1 = black
0 = white



Experiments - Synthetic Data



Experiment: SAC vs. MAC

- Estimate cluster centroids with MAC and SAC
- Compare with true centroids

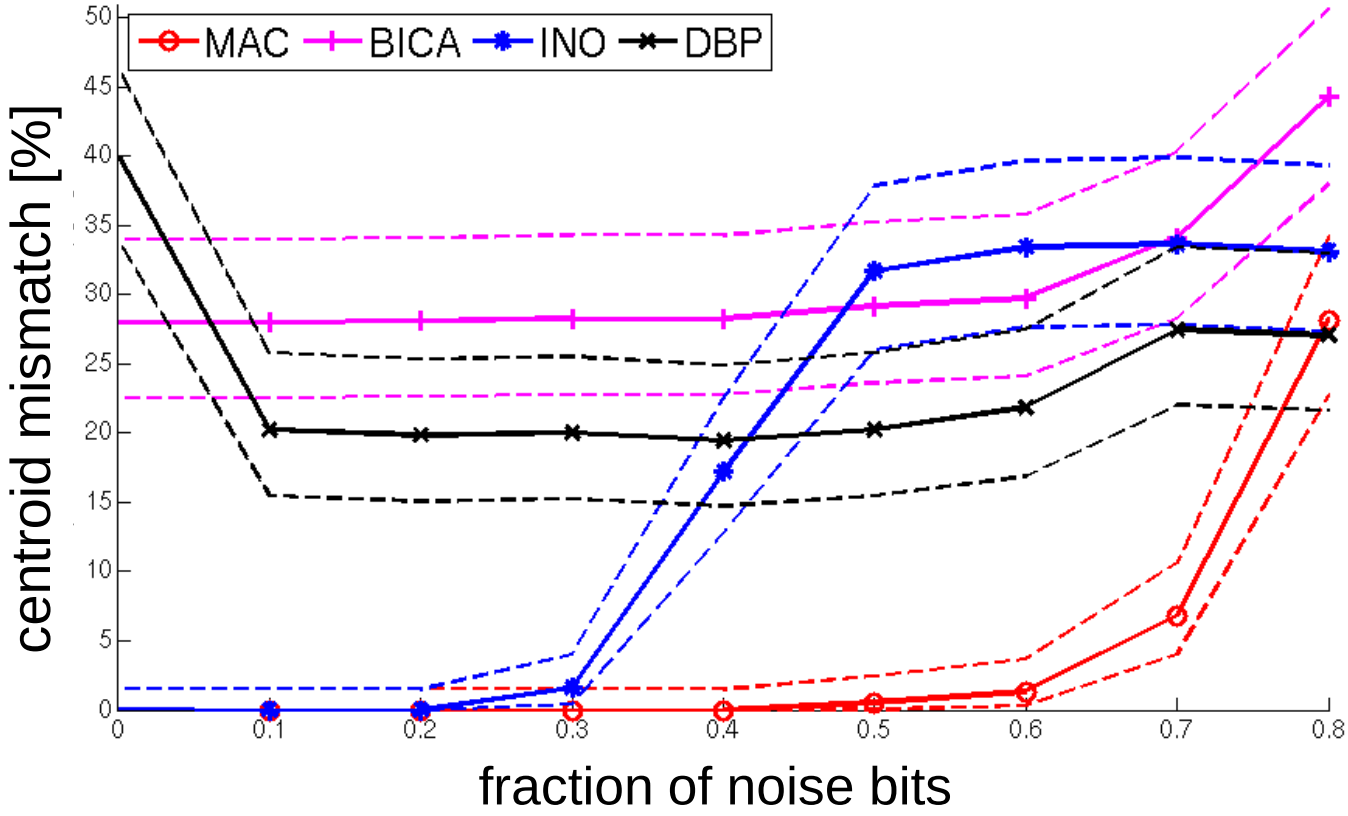
Experiments - Synthetic Data

Experiment: Comparison of MAC with other methods

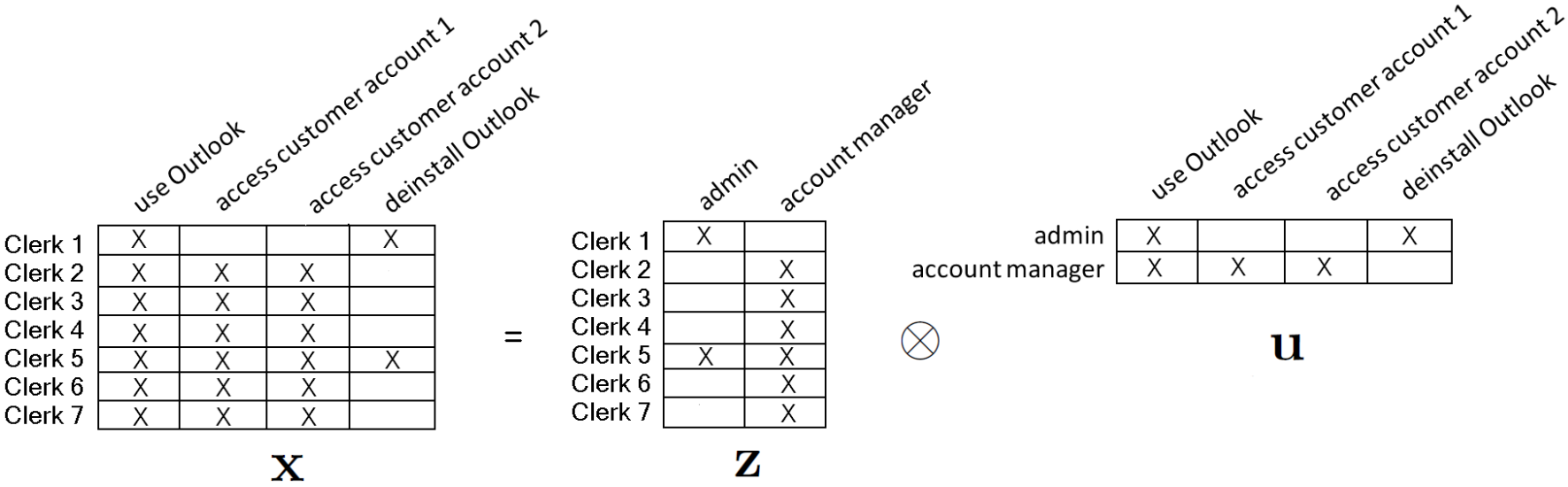
- Randomly generate synthetic data
- Estimate cluster assignments and centroids with MAC and with other multi-assignment methods:
  - Wood et al.: **A non-parametric Bayesian method for inferring hidden causes** (INO)
  - Miettinen et al.: **The Discrete Basis Problem** (DBPs)
  - Kabán et. al.: **Factorisation and denoising of 0-1 data: A variational approach** (BICA)
- Compare with true centroids

Experiments - Synthetic Data

Comparing accuracy of different methods at different noise levels:



Experiments – Role Mining



Direct user permission assignment

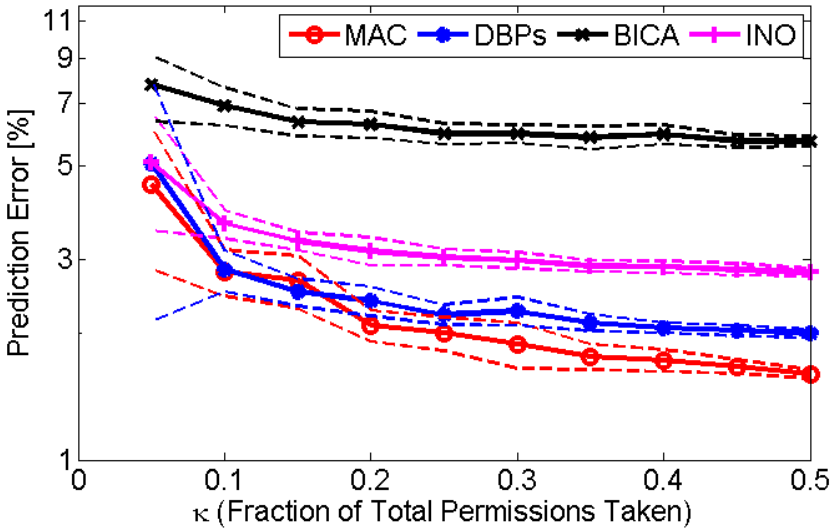
Role-Based Acces Control (RBAC)

Experiments – Role Mining

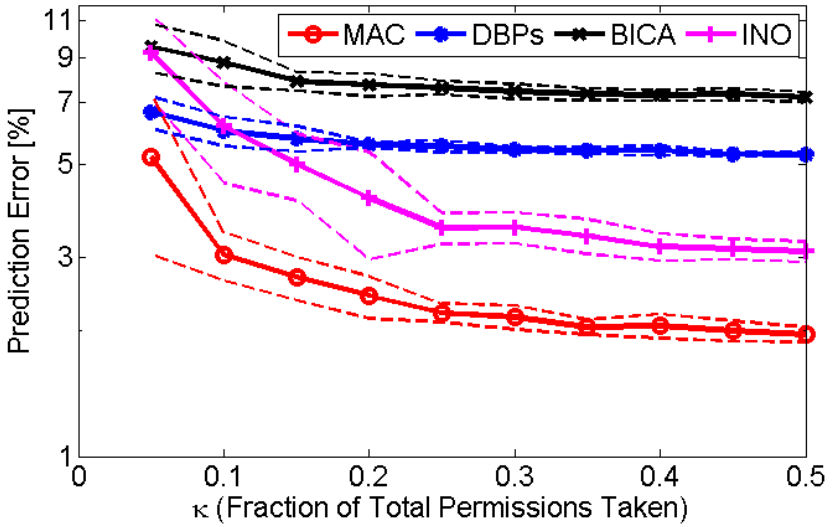
Experiment: Role Mining on 4900 users & 1300 permissions from a bank

- Estimate the underlying roles with MAC, DBPs, INO and BICA on a training set.
- Take a few permissions of the users from the test set to decide their role memberships.
- Compute the fraction of wrongly predicted permissions.

original data



+33% noise



Outlook

- Model other noise processes (asymetric noise, etc.)
- Incorporate side information to the model
  e.g. contract codes, organizational units
  - “Hybrid Role-Mining“

Thank you

for your attention