Role Mining

Mario Frank and Ian Molloy*

17th ACM Conference on Computer and Communications Security

Problem Definition

Role Engineering

- **Top Down**
  - From use cases and business properties
- **Bottom Up**
  - From existing access control data
- Bottom Up (automatic): **Role Mining** [KSS03]
- **Hybrid Role Mining**
  - Include business information to the role mining process

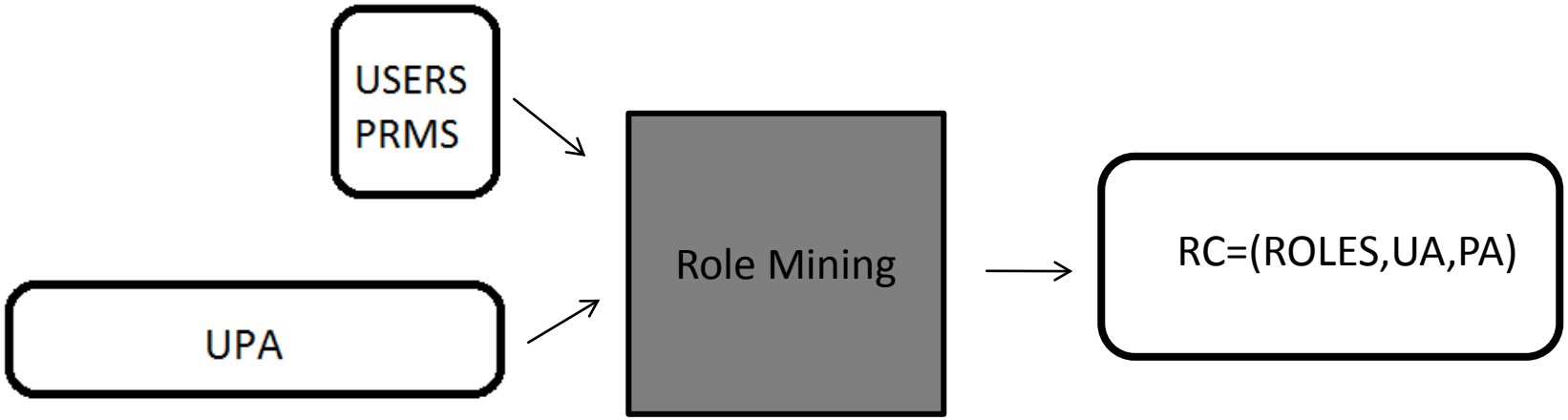
Top Down

- Manual analysis
- **Requires expertise** from security and business
- Conduct interviews, use cases, etc.
- Reluctant to outsource
- **Error prone**
- **Expensive**
- **Slow** (months)

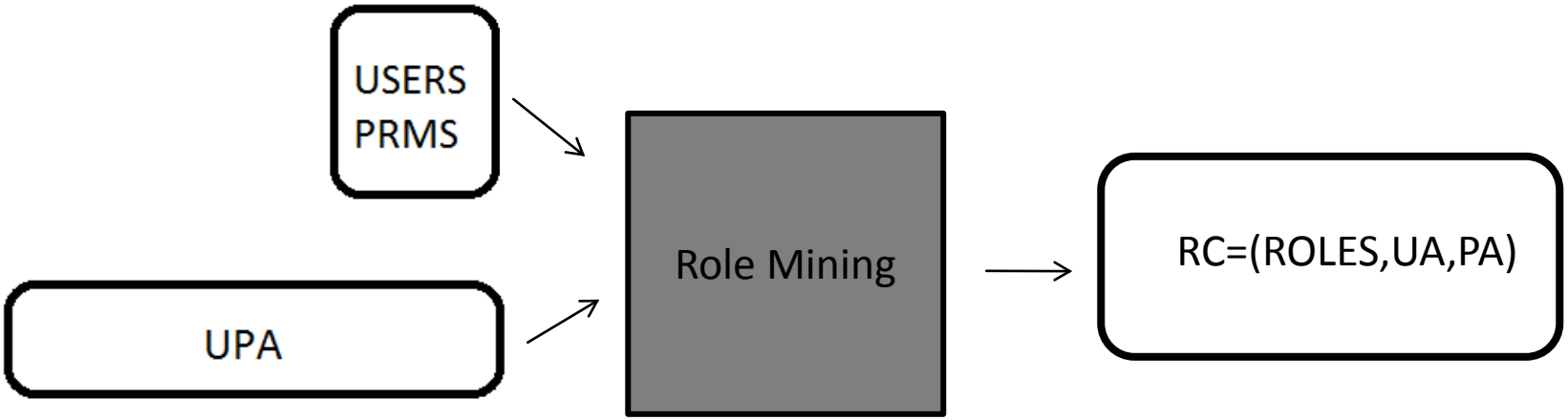
Bottom Up

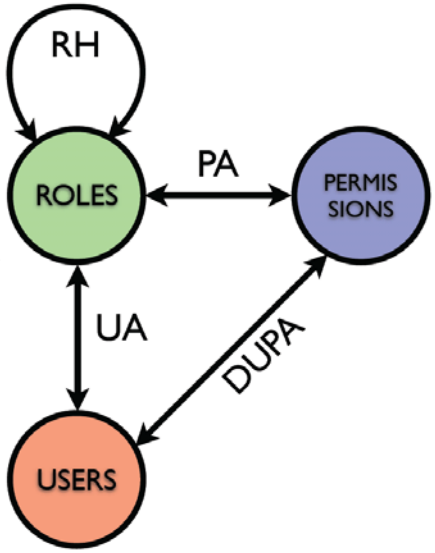
- Analyze the **existing** *data*
  - **User-Permission assignments**, attributes, usage logs, etc.
  - Apply **data mining** techniques to automate
- **Fast** (minutes-hours)
- Cheap
- Garbage in, garbage out
- Roles can be **less intuitive** than top-down engineering
  - Manual postprocessing (expensive)
  - Hybrid role mining

Input / Output

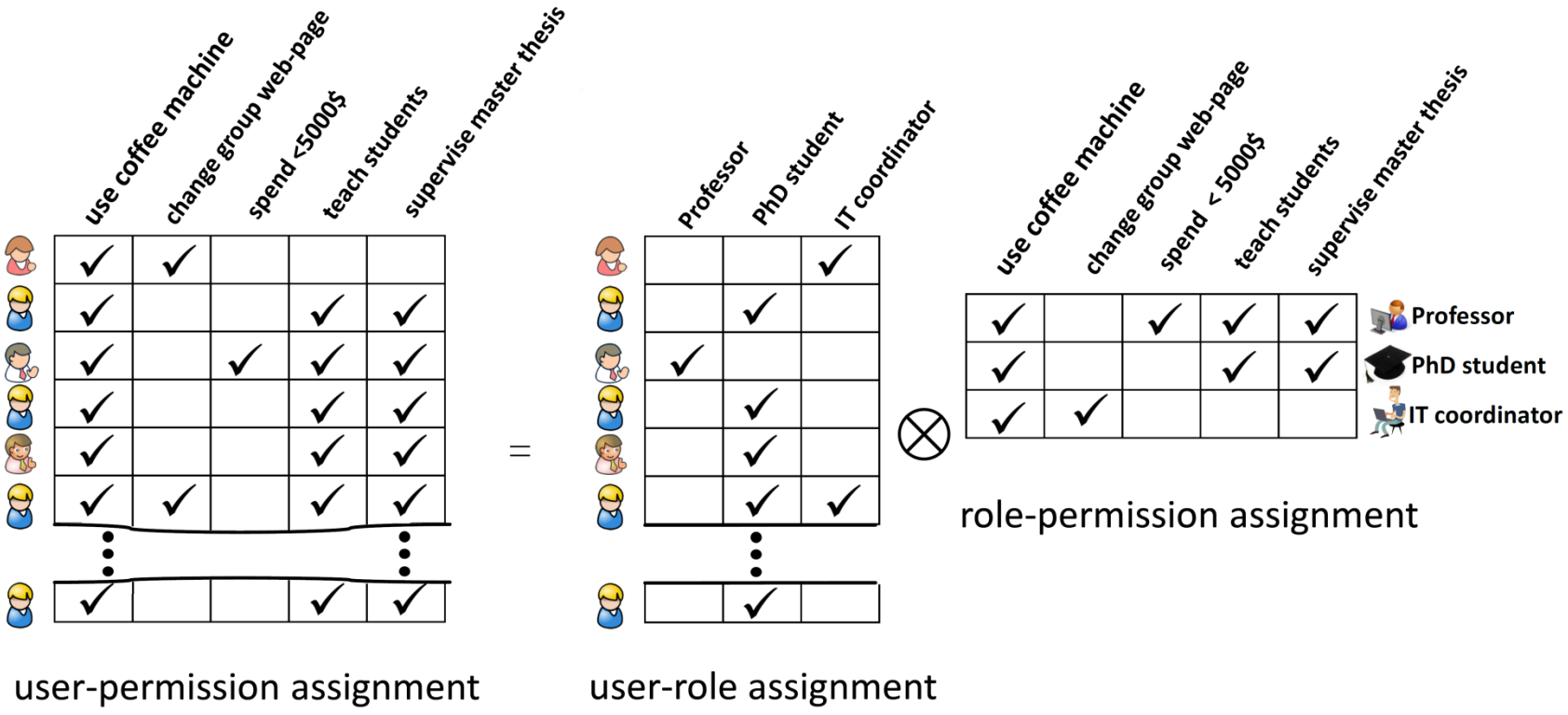


Input / Output





Example



user-permission assignment

user-role assignment

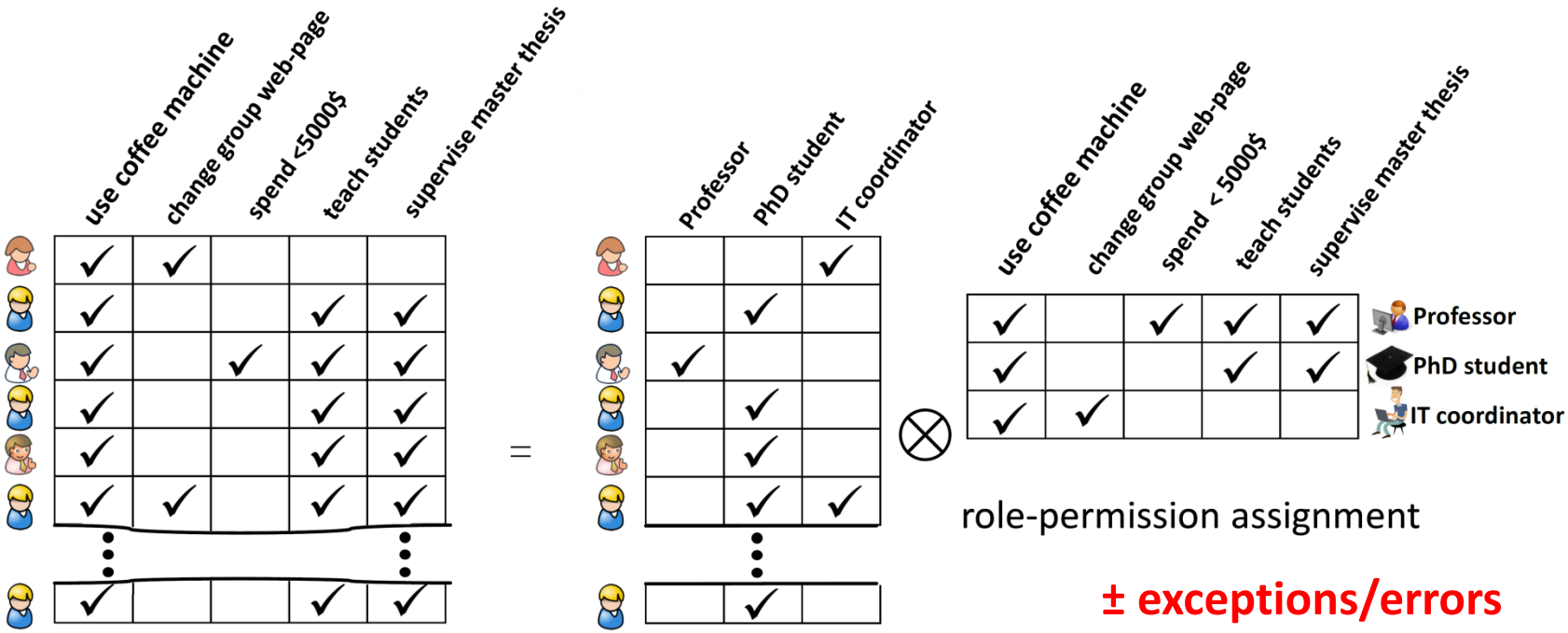
role-permission assignment

user-permission assignment
UPA

Role Mining

Role-Based Acces Control
RBAC

Example

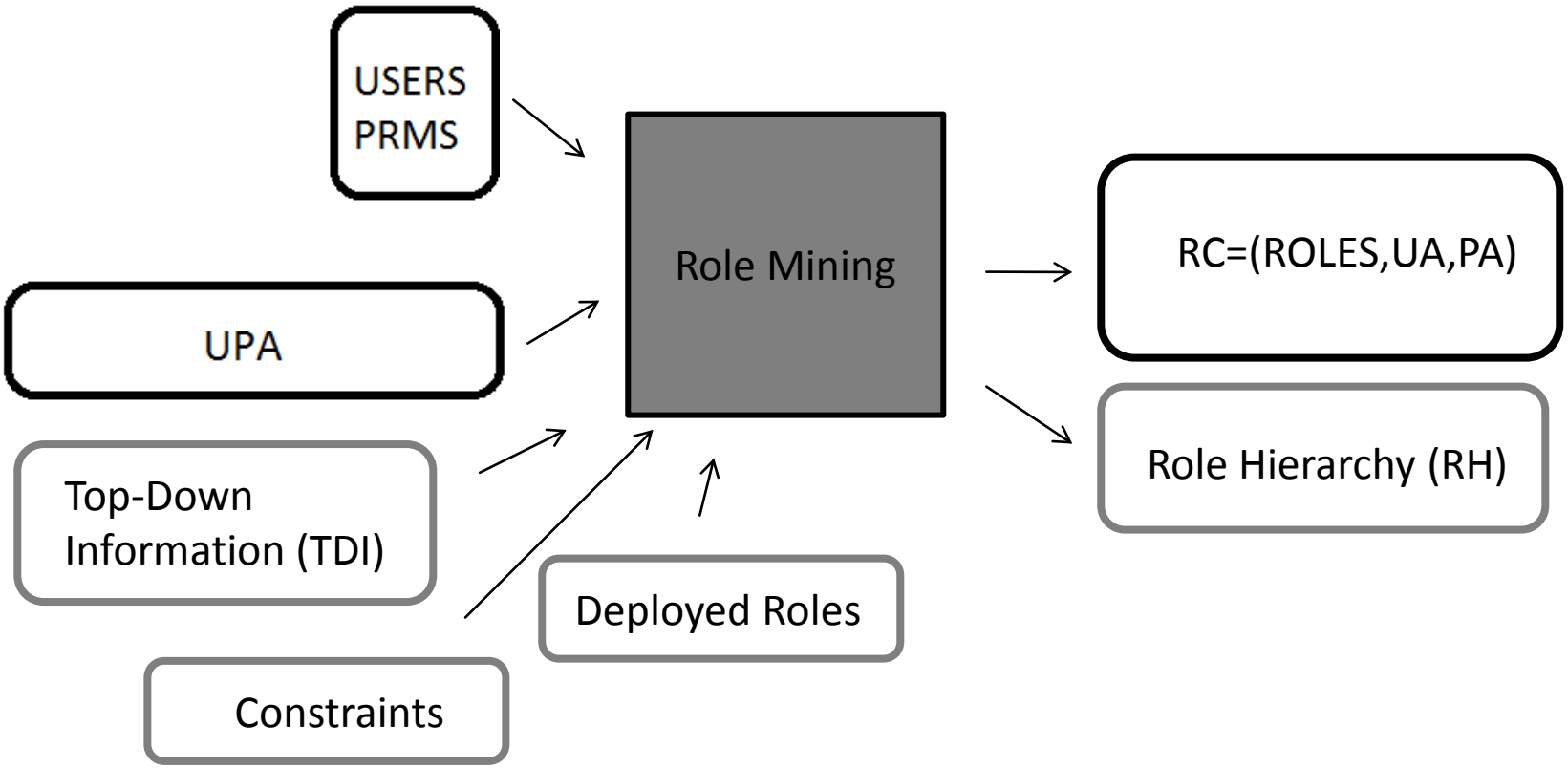


user-permission assignment
UPA

Role Mining

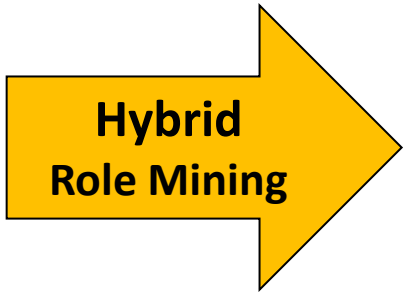
Role-Based Acces Control
RBAC

Input / Output



Hybrid Role Mining

UPA
+
Top-down information



RBAC
where roles reflect **business meaning**

- **Top-down information**: e.g. organizational unit, seniority, location, …

- **Business meaning**:
  a) Same set of roles ⇨ similar business features
     [MCL+08] , [CDO+09] , [MLL+10]
  b) Same business features ⇨ similar set of roles
     [FSB+09]

Overview

- Problem definitions
- Quality measures
- Role Mining solutions / algorithms
  - Discrete optimization techniques
  - Probabilistic techniques
  - Hybrid role mining
- Open problems / future research

Definitions 1/3

Definition **BASIC RMP** [VAG07] :
Given a set of users USERS, a set of permissions PRMS and a user-permission assignment UPA, find an RBAC configuration RC that **minimizes the number of roles k** and **does not deviate from UPA**.

Definition **δ -APPROX. RMP** [VAG07] :
Given a set of users USERS, a set of permissions PRMS and a user-permission assignment UPA, find an RBAC configuration RC that **minimizes the number of roles k** and **deviates from UPA with less than δ assignmets**.

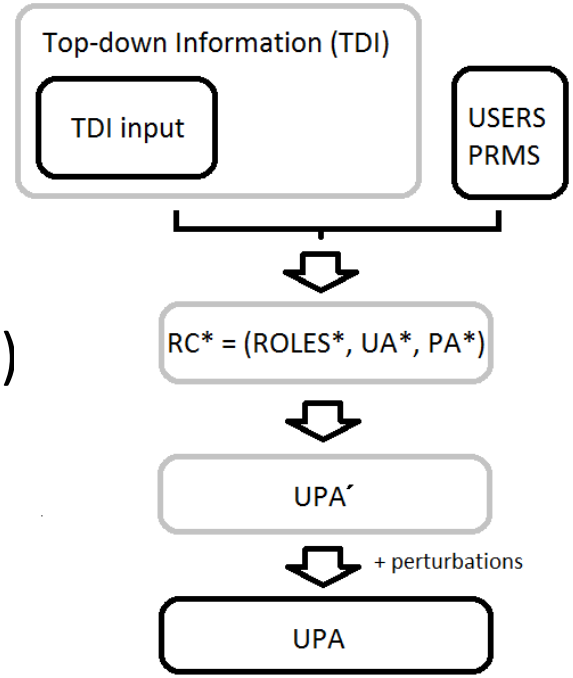
Definitions 2/3

Definition **MIN-NOISE RMP** [VAG07] :
Given a set of users USERS, a set of permissions PRMS and a user-permission assignment UPA, and the **number of roles k**, find an RBAC configuration RC with k roles, **minimizing the deviation between UPA and RC**.

Definition Min-Edge **RMP** [LVA08] :
Given a set of users USERS, a set of permissions PRMS and a user-permission assignment UPA, find an RBAC configuration RC that is **consistent** with UPA and **minimizes the number user-role assignments and role-permission assignments**.

Definitions 3/3

Definition **INFERENCE RMP** [FBB10] :
Let a set of users USERS, a set of permissions PRMS, a user-permission relation UPA, and, optionally, part of the **top-down information TDI** be given. Under Assumption 1-3, **infer** the unknown RBAC configuration **RC***=(ROLES*, UA*, PA*).



Assumptions:

1. **RC* generated UPA**
2. **RC* reflects top-down information** (TDI)
3. **Exceptions** (errors) might **exist**.

Quality Measures

Reconstruction Accuracy

Closeness of RBAC configuration RC to user-permission assignment UPA [VAG07], [LVA08] .

- **Coverage** of UPA assignments with RC
- **Hamming distance** between UPA and RC

Size measures

Compute how well RBAC configuration **compresses** the given access-control system.

- Number of roles |R| [VAG07]
- Number of assignments |UA|+|PA| [LVA08]
- **Weighted structural complexity (wsc)** [LMQ+07]

$wsc(RC,w) = w_1|R|+w_2|UA|+w_3|PA|+w_4|DUPA|+w_5|t(RH)|$
with weights (w_1, w_2, w_3, w_4 , w_5)

Comparison between true roles and inferred roles
(true roles must be known!)

Compare number of roles [KSS03] **:** does not tell too much

Pairwise distance:

- distance measure of your choice:
  exact match, Hamming distance , Jaccard coefficient.
  Caution! Avoid repeated comparison**.**

What can go wrong

True roles

⋮

What can go wrong

Mined roles

True roles

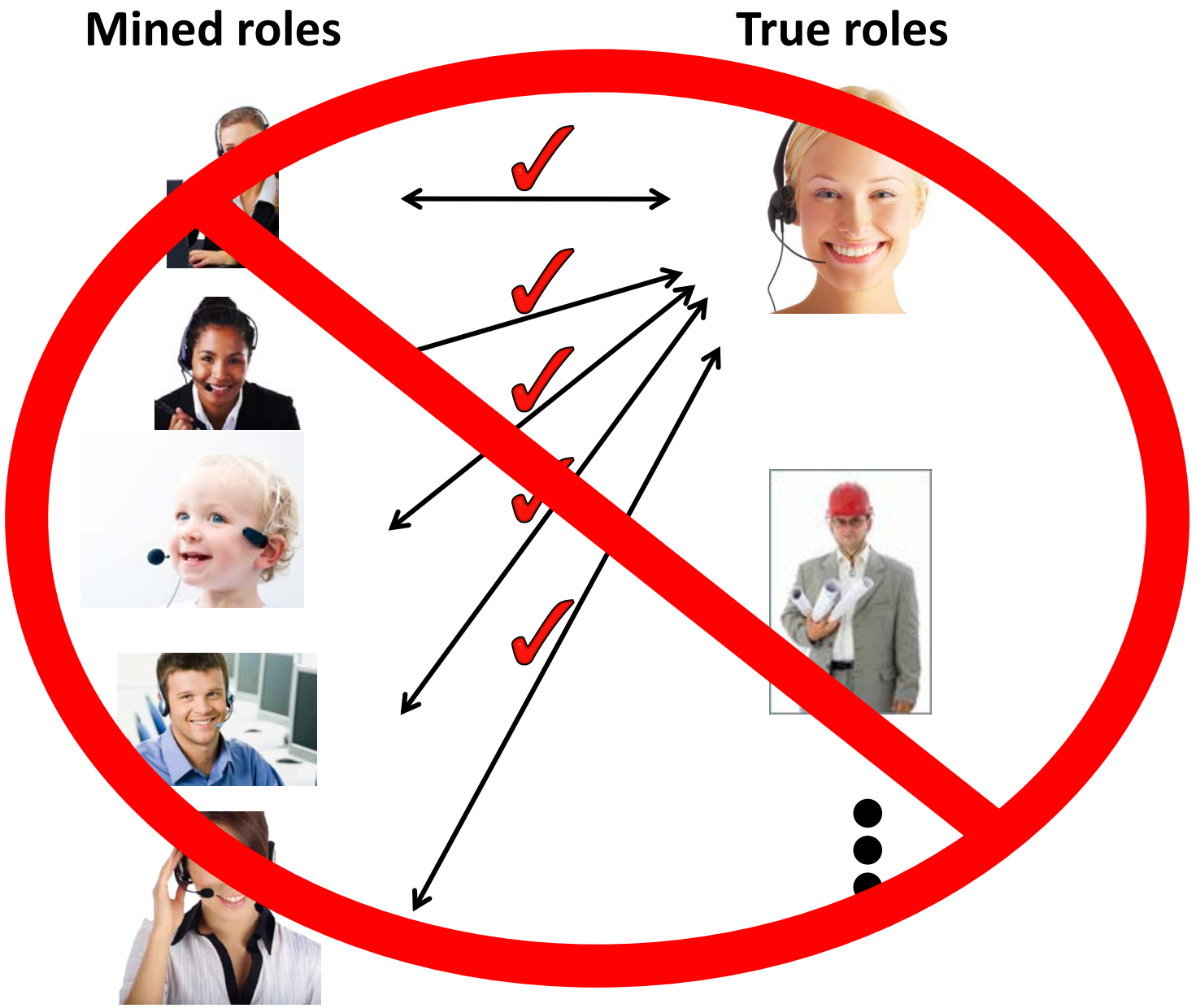
⋮

What can go wrong

Mined roles

True roles

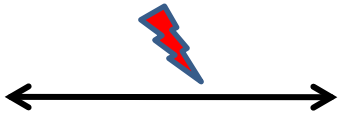
What can go wrong



Find a global permutation of roles [SFB+09], [FBB10], [MLL+10]

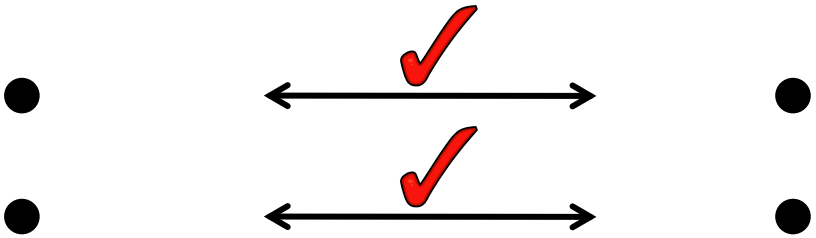
Mined roles

True roles



Find a global permutation of roles [SFB+09], [FBB10], [MLL+10]

Mined roles **True roles**



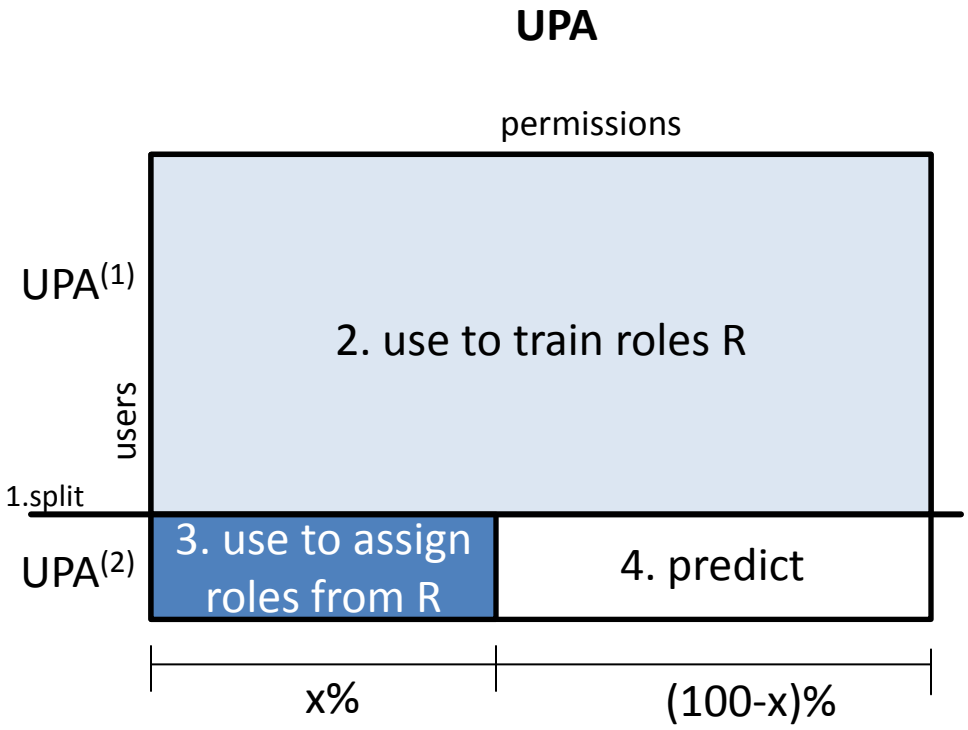
Generalization Test [SFB+09], [FBB10]

(true roles unknown)

Exploit that undelying structure **RC* reproduces over the users**, whereas the **noise does not**.

Generalization test:

1. randomly split UPA in $UPA^{(1)}$ and $UPA^{(2)}$
2. train roles R on $UPA^{(1)}$
3. assign users from $UPA^{(2)}$ to roles based on x% of their permissions
4. predict remaining (100-x)% of permissions
5. compute prediction error



Closer to RC* ⇨ better prediction error

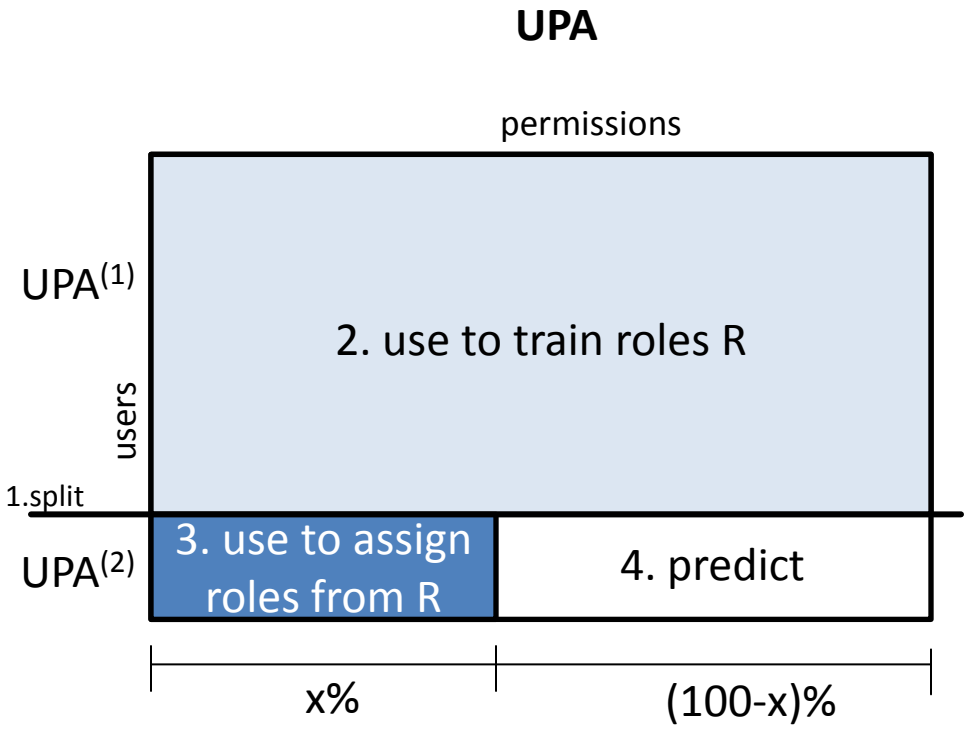
Generalization Test with TDI [FSB+09]

(true roles unknown)

Exploit that undelying structure **RC* reproduces over the users**, whereas the **noise does not**.

Generalization test (when TDI is given):

1. randomly split UPA in $UPA^{(1)}$ and $UPA^{(2)}$
2. train roles R on $UPA^{(1)}$ **and $TDI^{(1)}$**
3. assign users from $UPA^{(2)}$ to roles based on x% of their permissions **and TDI**
4. predict remaining (100-x)% of permissions
5. compute prediction error



Closer to RC* ⇨ better prediction error

Discussion

1. Wich problem do we want to solve?
2. How should we validate solutions?

Summary statistics of role mining concepts [FBB10]

	formal definition	solution algorithm	quality measure
size of RBAC configuration (number of roles, no. of assignments, etc.)	3	6	5
linear combination of size measures (*wsc* or "costs" with specified weights)		8	8
comparison with original roles (if known)		3 (some deployed roles are given)	9
likelihood		3 1	
agreement with top-down information	1	3	5
0-consistency with *UPA*: required	3	10	
0-consistency with *UPA*: not required	2	8	

Discussion

1. Wich problem do we want to solve?
    - **Inference RMP**
        - **Assume that there are roles to be found (if not, don‘t use role mining)**
        - **Roles should correspond to business (TDI)**

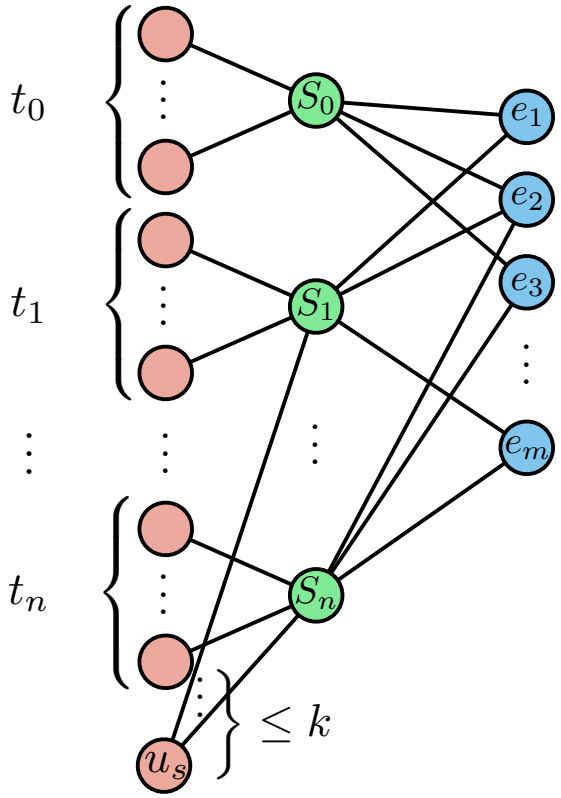
2. How should we validate solutions?
    - **Comparison with true roles (if known)**
    - **Generalization test (if not known)**

Exact Role Mining

WSC Theoretical Results

[MCL+1x]

- **NP**-Complete in general
  - SETCOVER
- Some trivial cases
  - Not interesting
- No polynomial approximation (**P≠NP**)
  - Edge-Concentration

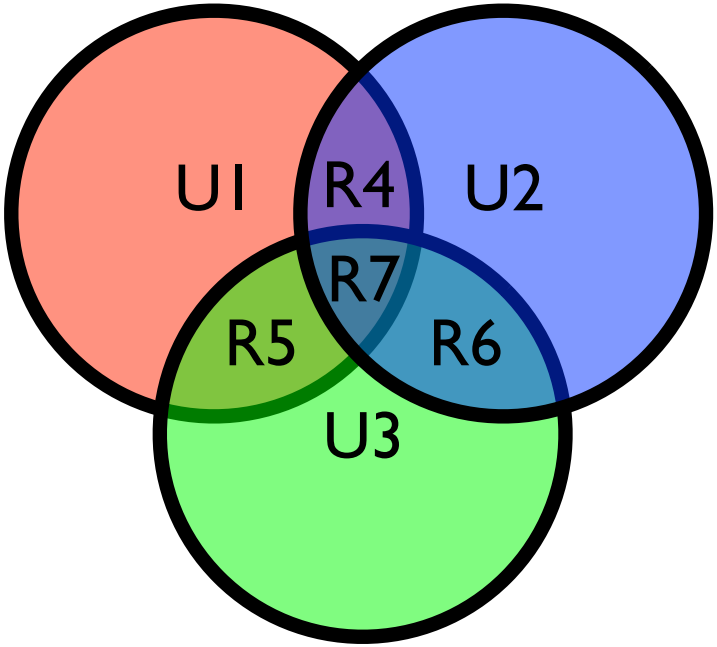


{Fast,Complete}Miner [VAW06]

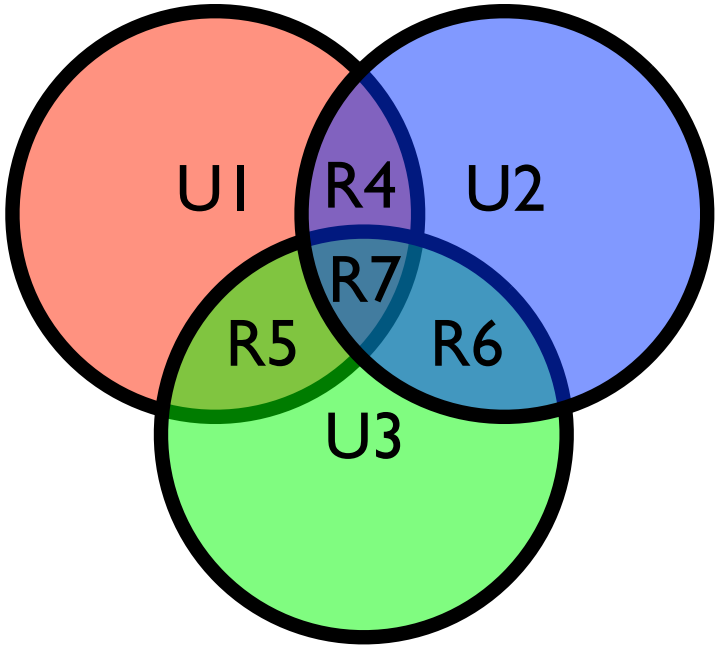
- Each user is an initial role
- Intersect initial roles
- Order roles by

  $\alpha * e(r) + n(r)$

- How many original roles are recovered?



{Fast, Complete}Miner



FastMiner

U1
U2
U3
R4∪R7
R6∪R7
R5∪R7

CompleteMiner

R4
R5
R6
R7

Database Tiling [VAG07]

- Define the role mining problem (RMP)
  - Minimize the number of roles for UP
- Show RMP is NP-Complete
  - Reduce to database tiling

Database Tiling [VAG07]

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	1	1	0	0	1	1	1
u_2	0	0	0	1	1	1	1
u_3	1	1	0	1	1	0	0
u_4	1	1	0	0	0	0	0

R1

R2

R3

Database Tiling [VAG07]

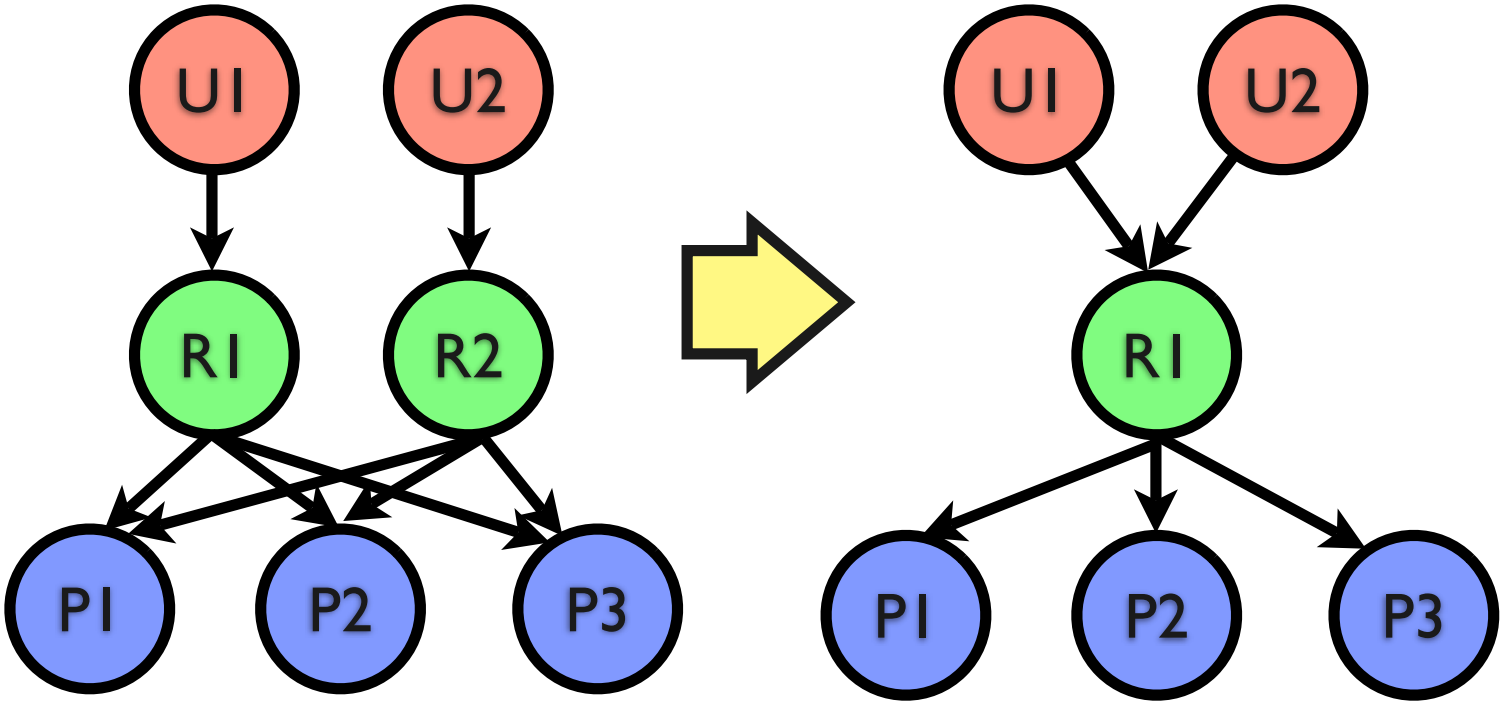
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	1	1	0	0	1	1	1
u_2	0	0	0	1	1	1	1
u_3	1	1	0	1	1	0	0
u_4	1	1	0	0	0	0	0

- Greedy Solution
- Tile that covers largest uncovered permissions
- Subproblem is NP-hard

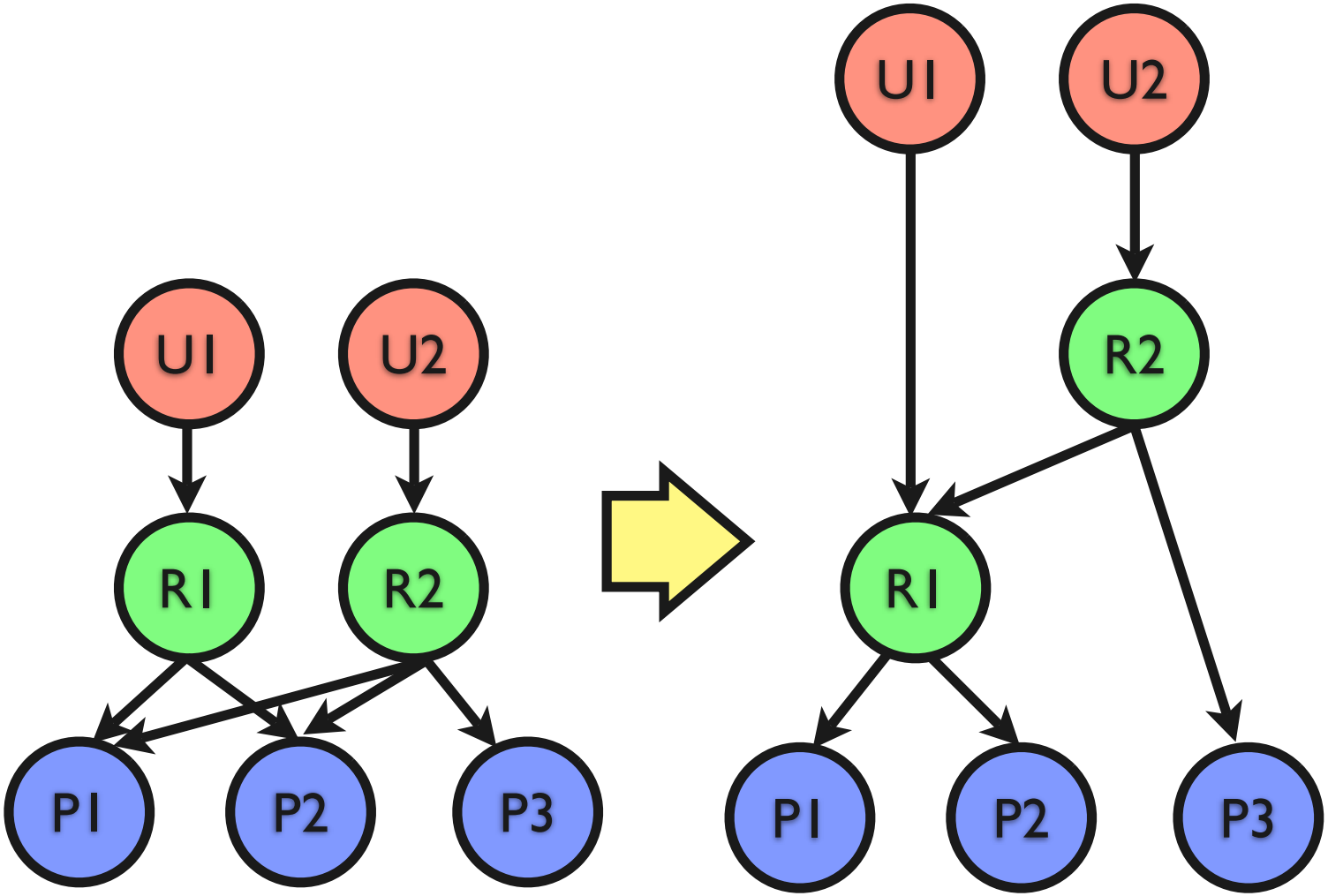
Graph Optimization

Graph Optimization [ZRE07]

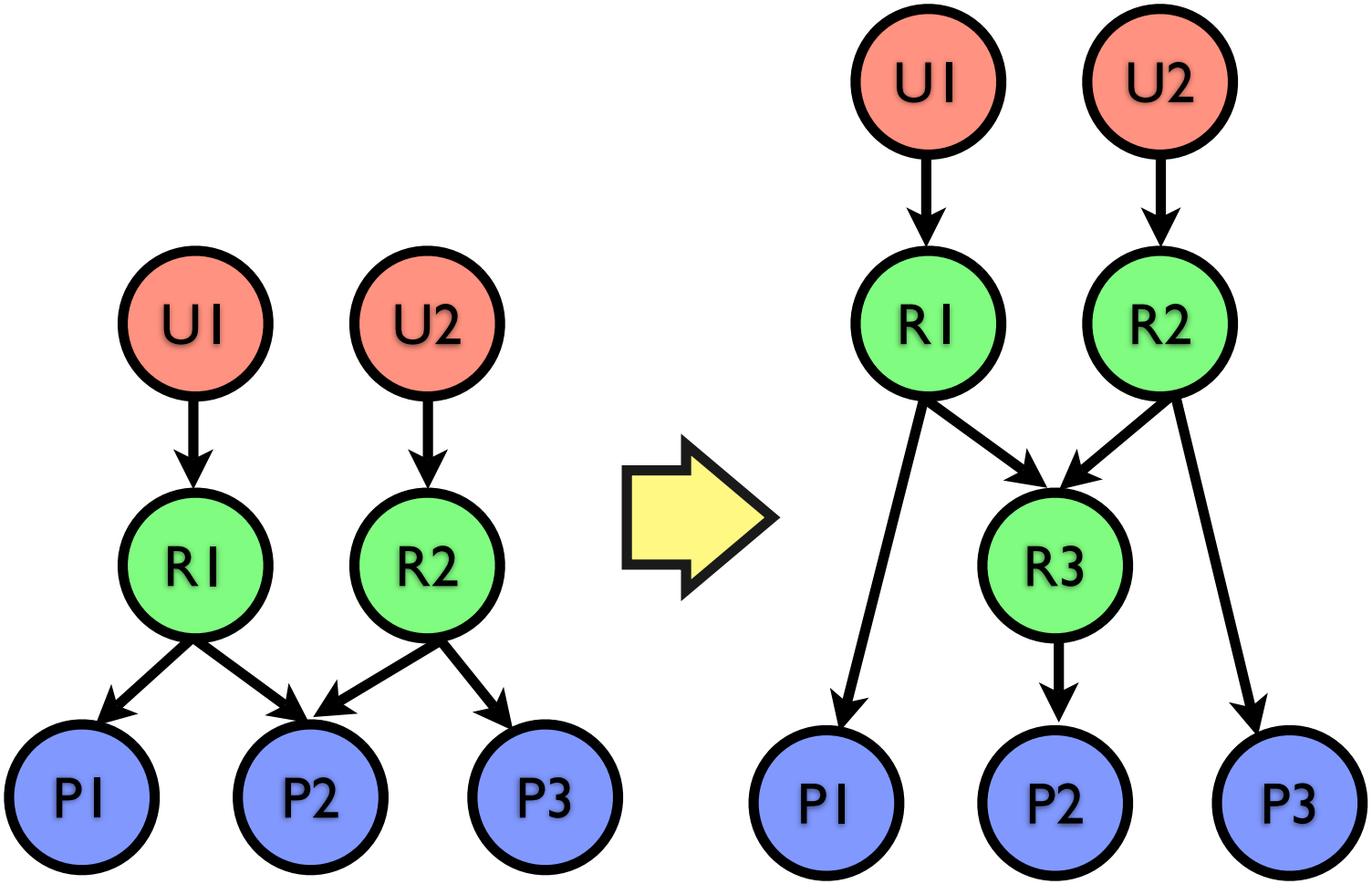
- Each user defines an initial role
- Role, user, permission node on a graph
- Perform pair-wise optimizations
- Minimize:
  - $|UA| + |PA| + |RH|$
  - $|R| + |UA| + |PA| + |RH|$



Identical Roles

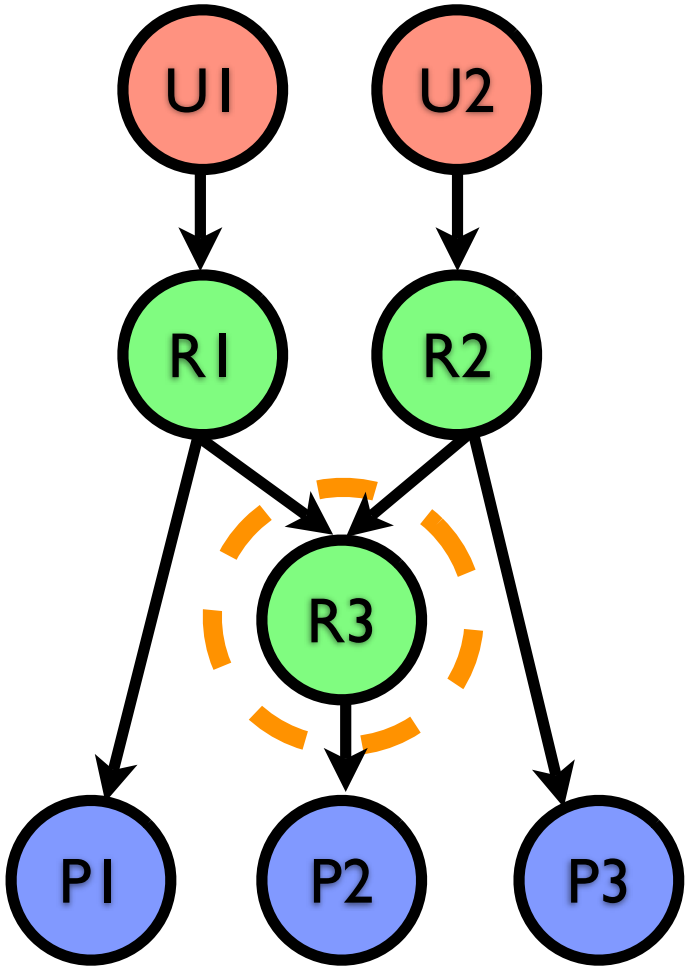


Subset Relation



Sufficient Overlap

Repeat by pairing R3
with all current roles
or end after *k* rounds

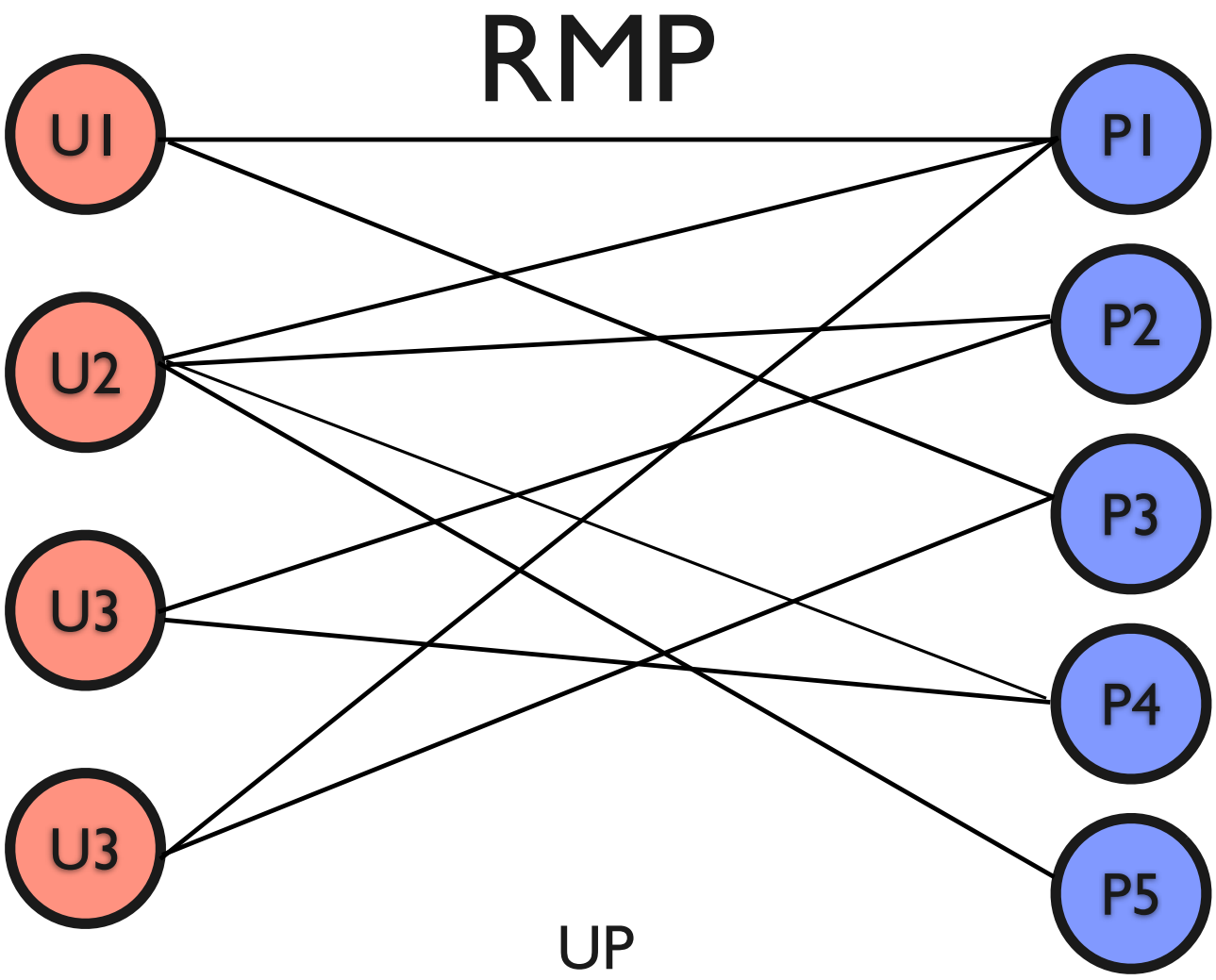


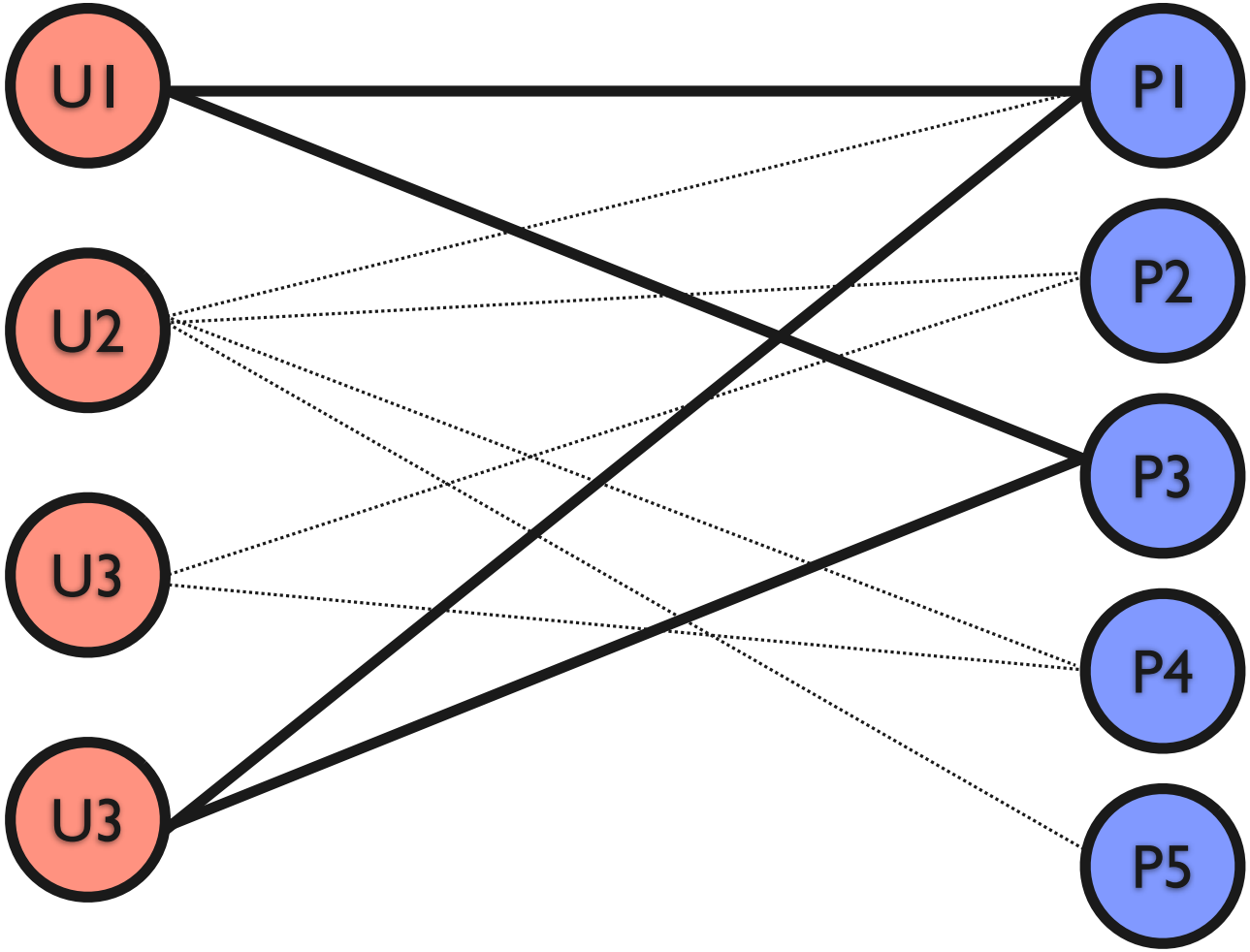
Repeat as Necessary

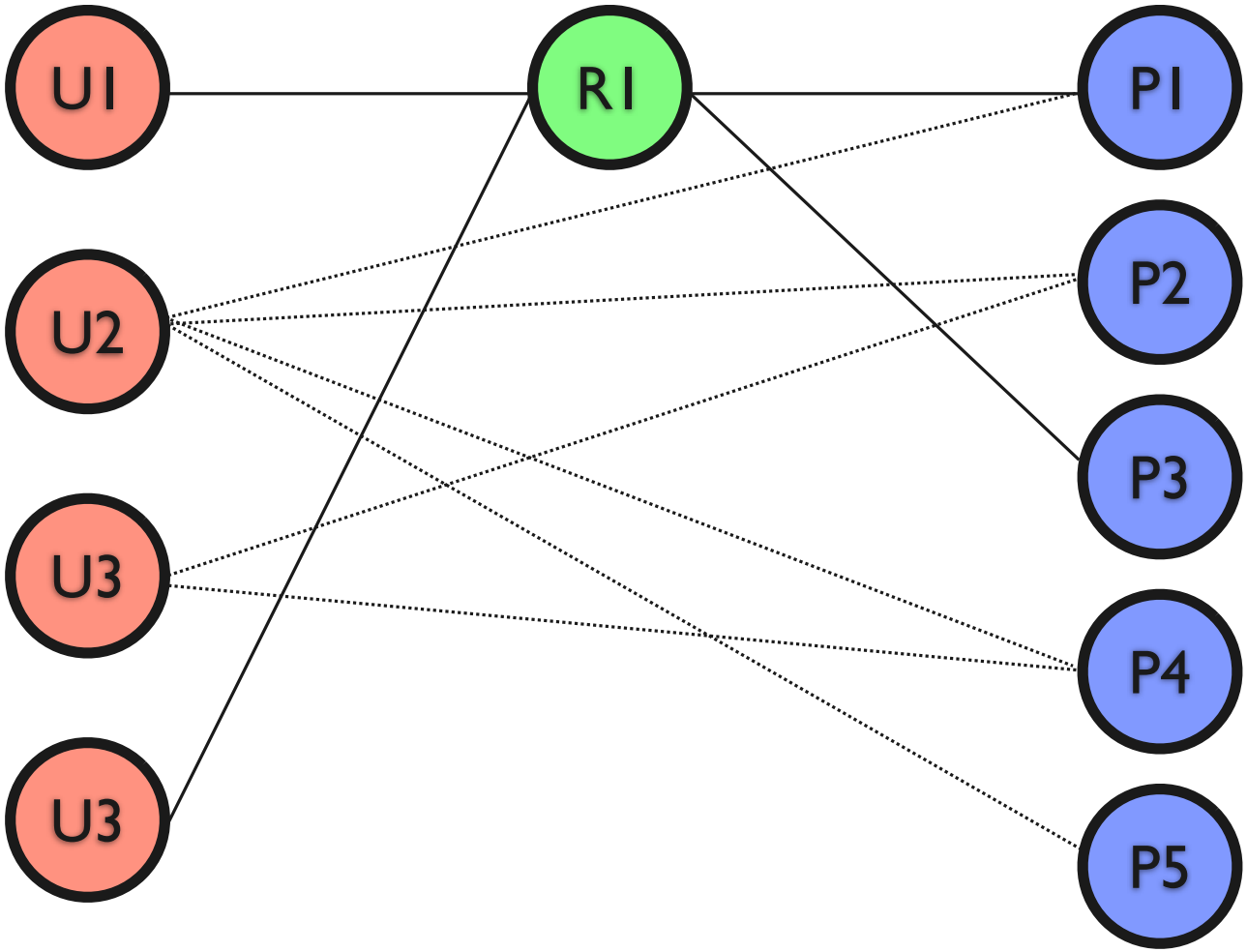
Maximal Bicliques

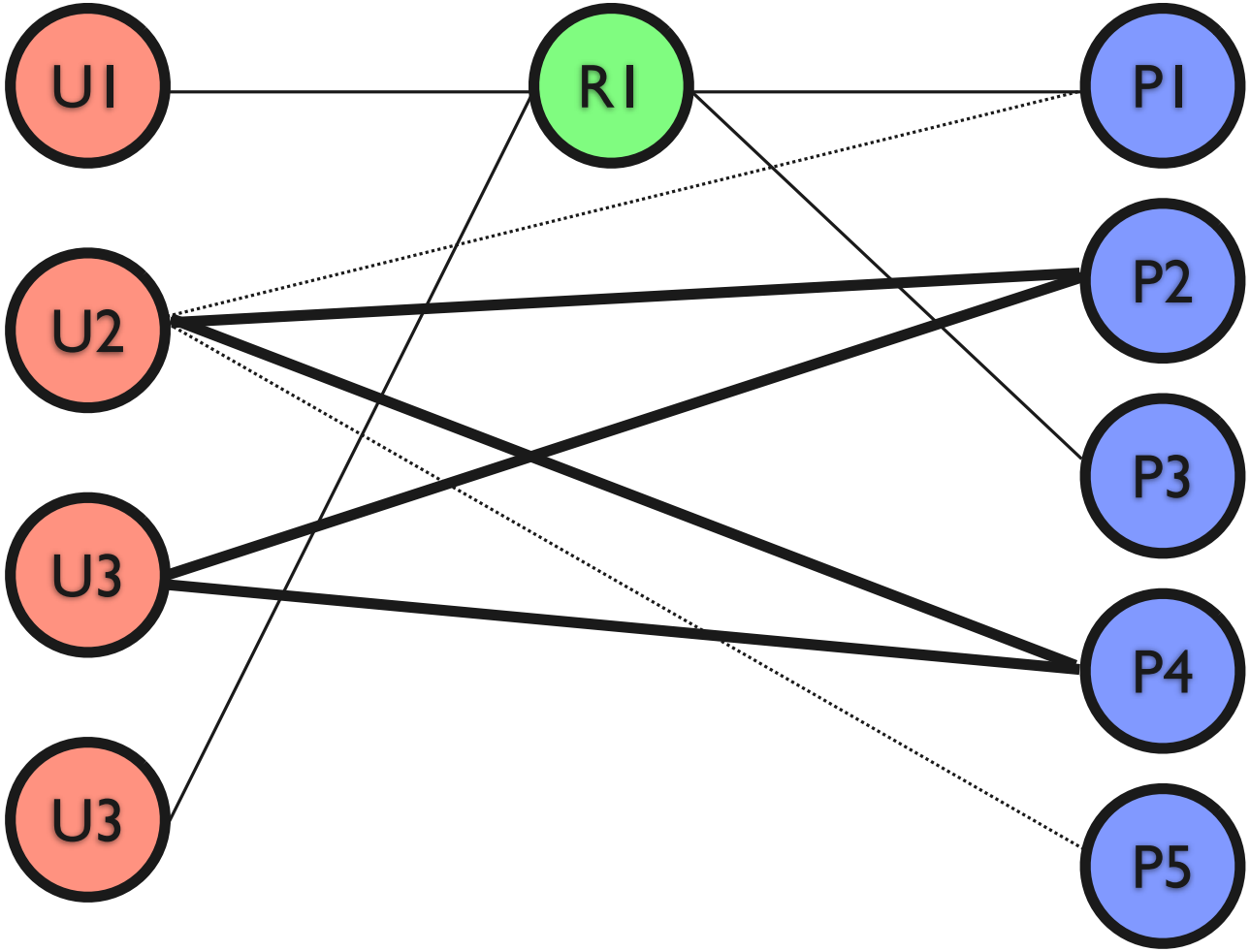
Biclique Cover [EHM+08]

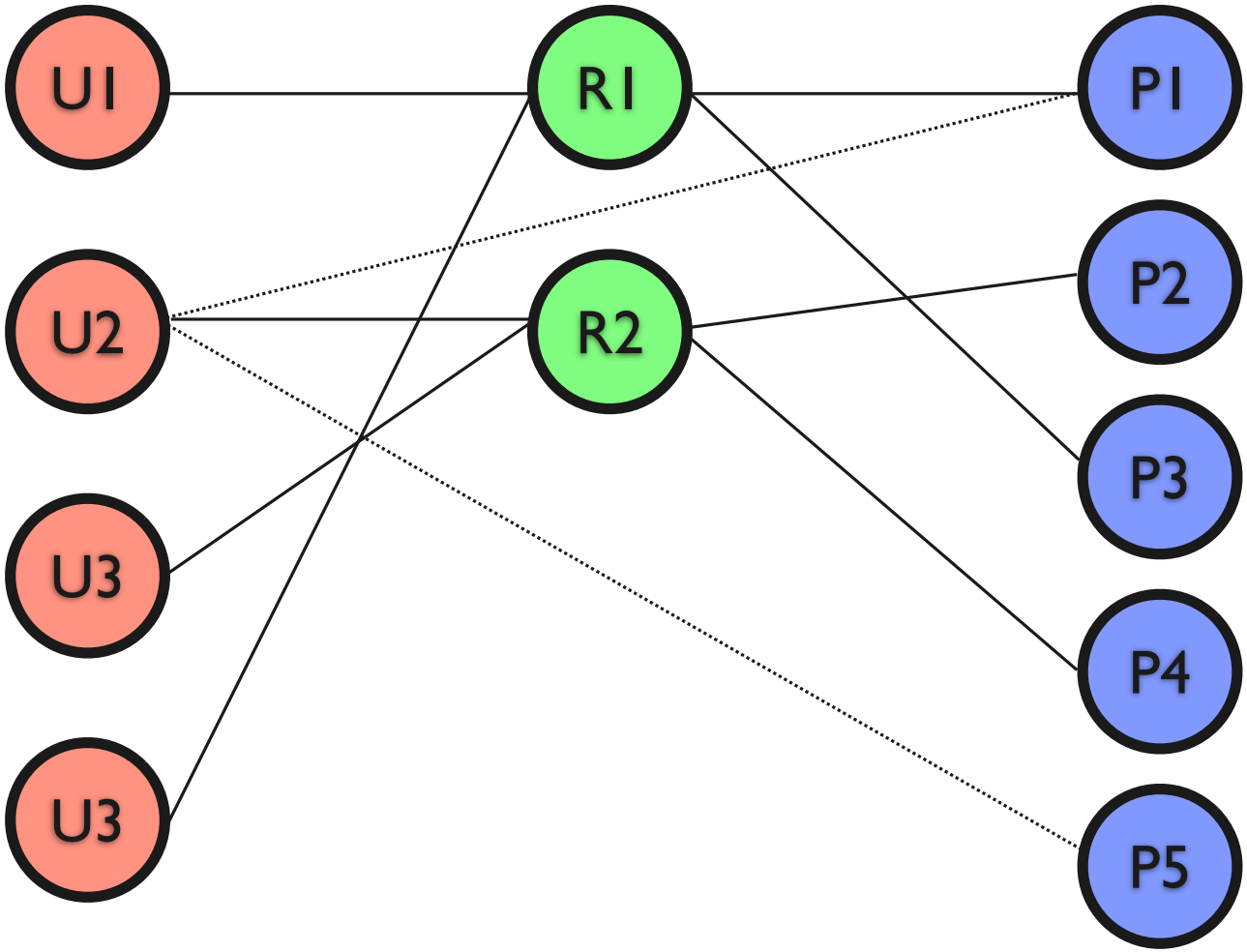
- Users and permissions are vertexes
- Permission assignments are edges
- UP is a bipartite graph
- [Flat] RBAC is a tripartite graph
- Minimum biclique cover is RMP

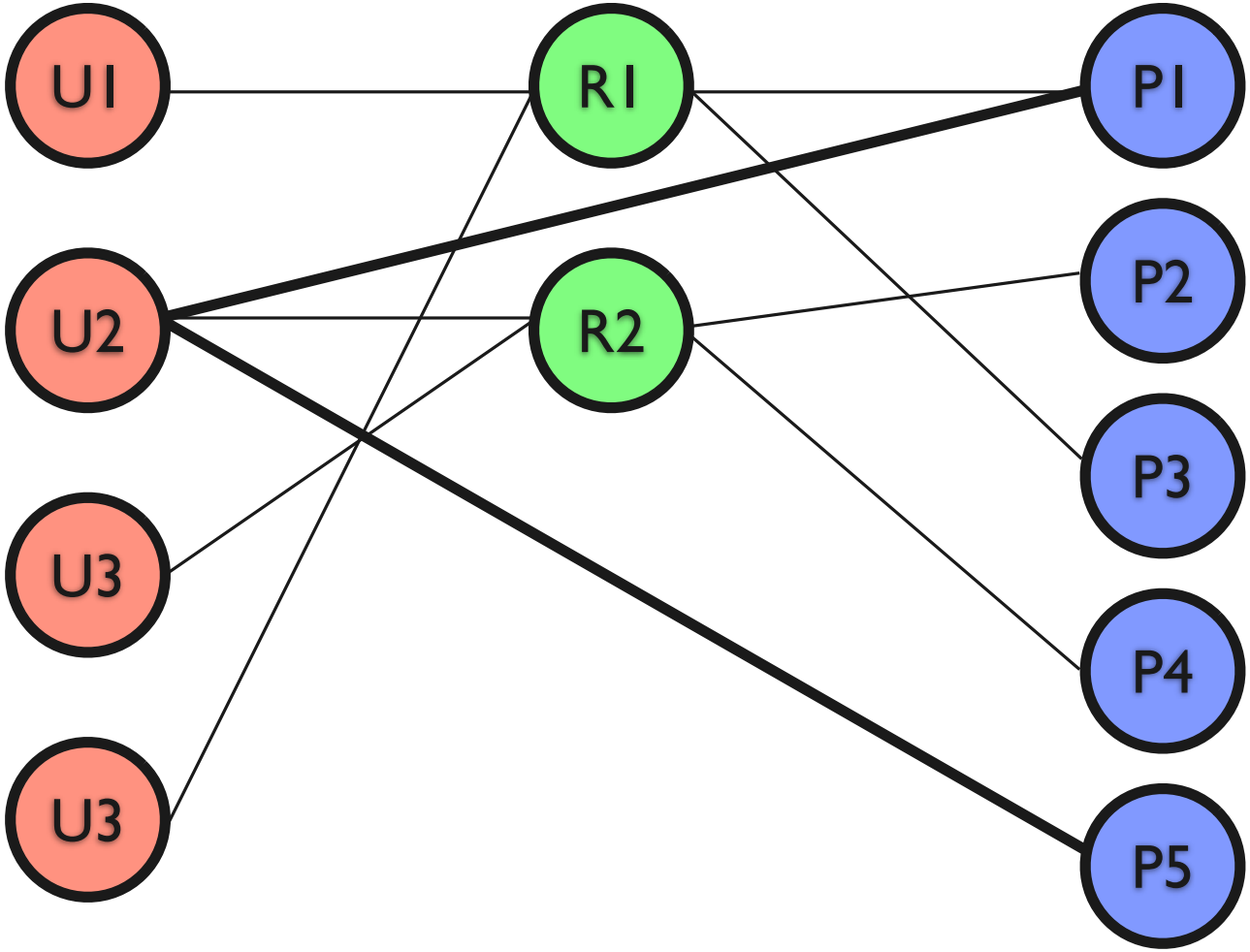

RMP
U1
U2
U3
U3
P1
P2
P3
P4
P5
UP

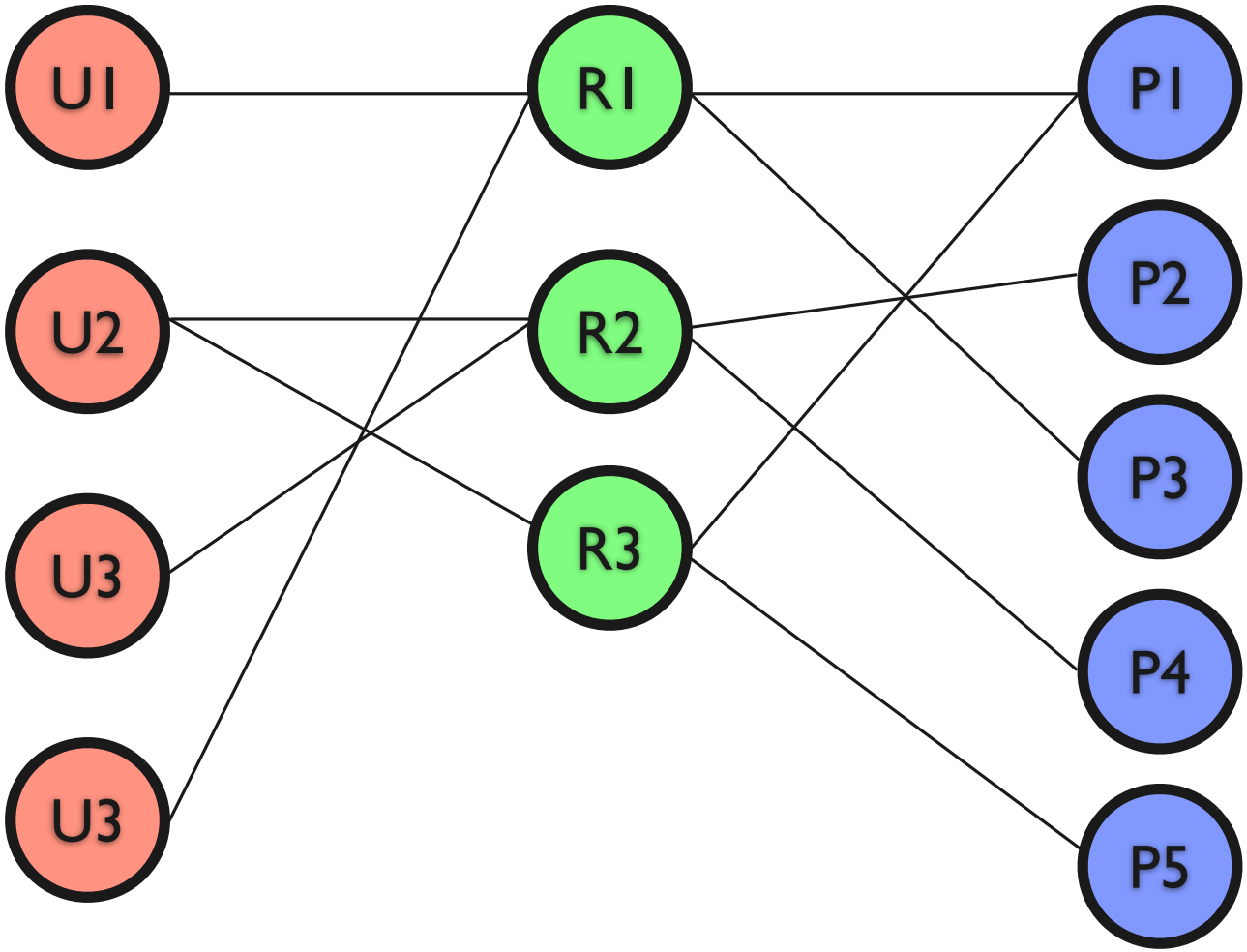

U1
U2
U3
U3
P1
P2
P3
P4
P5


U1
U2
U3
U3
R1
P1
P2
P3
P4
P5


U1
R1
P1
U2
P2
U3
P3
U3
P4
P5


U1
U2
U3
U3
R1
R2
P1
P2
P3
P4
P5


U1
U2
U3
U3
R1
R2
P1
P2
P3
P4
P5


U1
U2
U3
U3
R1
R2
R3
P1
P2
P3
P4
P5

Formal Concept Analysis and Lattices

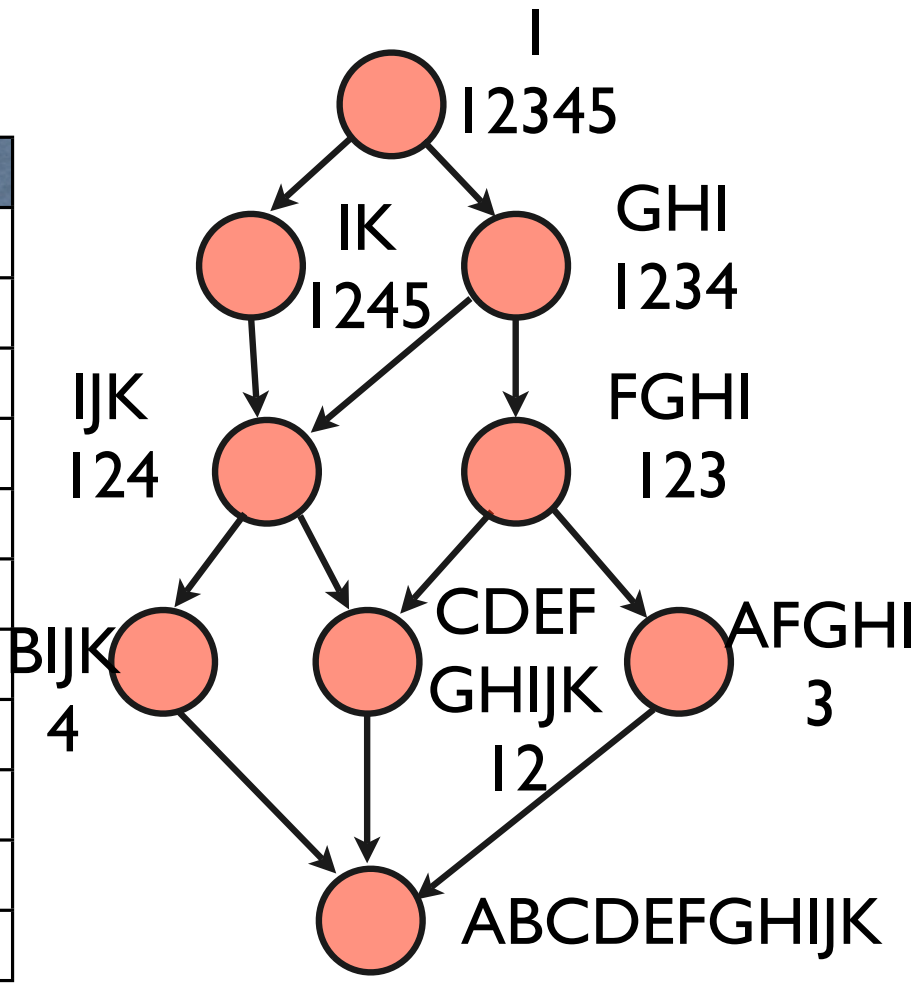
Formal Concept Analysis

- Context triple (G,M,I)
  - $I \subseteq G \times M$
- *G* users, *M* permissions, $I = UP$
- Concept *(X,Y)*
  - $X \subseteq G, Y \subseteq M$
  - *X* and *Y* maximal bicliques
  - Arrange on a full lattice

Permissions

Users

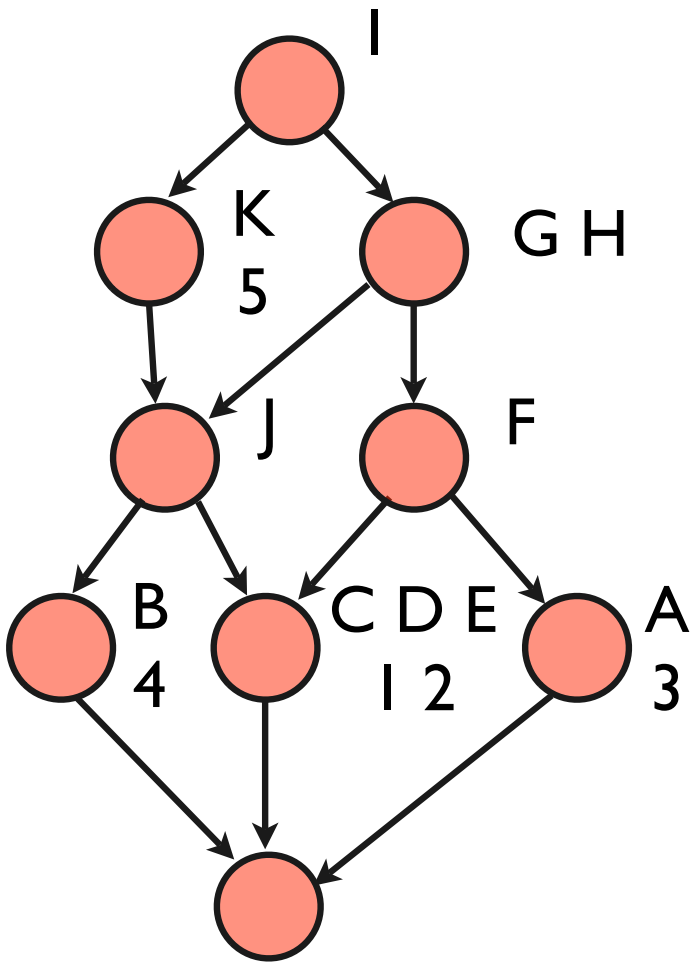
	1	2	3	4	5
A			✓		
B				✓	
C	✓	✓			
D	✓	✓			
E	✓	✓			
F	✓	✓	✓		
G	✓	✓	✓	✓	
H	✓	✓	✓	✓	
I	✓	✓	✓	✓	✓
J	✓	✓		✓	
K	✓	✓		✓	✓



Permissions

Users

	1	2	3	4	5
A			✓		
B				✓	
C	✓	✓			
D	✓	✓			
E	✓	✓			
F	✓	✓	✓		
G	✓	✓	✓	✓	
H	✓	✓	✓	✓	
I	✓	✓	✓	✓	✓
J	✓	✓		✓	
K	✓	✓		✓	✓



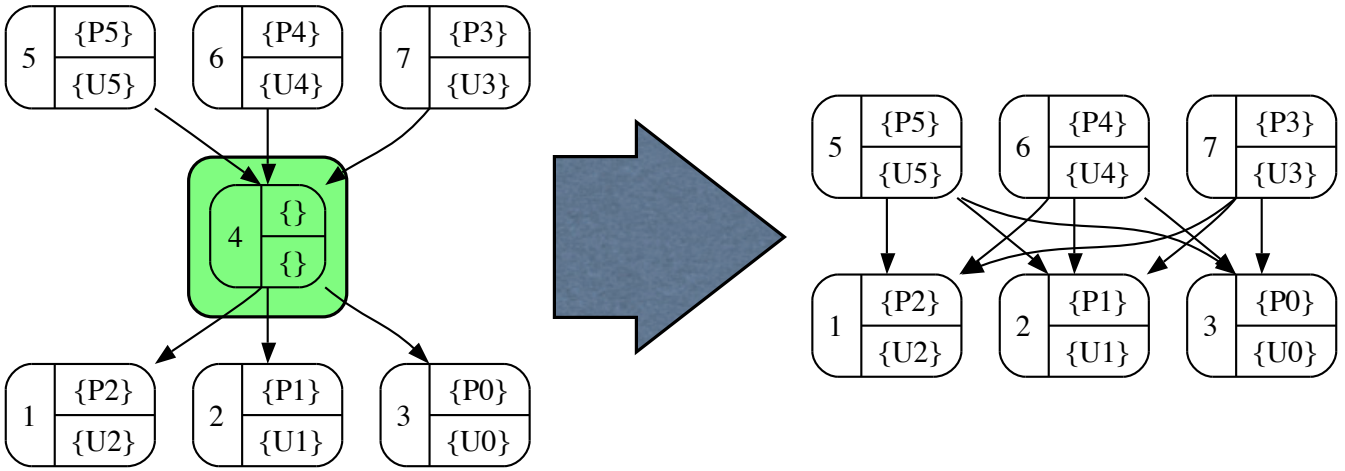
Hierarchical Miner [MCL+08]

- Greedy algorithm to heuristically optimize
- Local restructuring when WSC beneficial
- Four rules to prune roles—Can be extended
- Stops when no restructures decrease WSC

7 Roles
6 RH

VS.
(Based on WSC)

6 Roles
9 RH

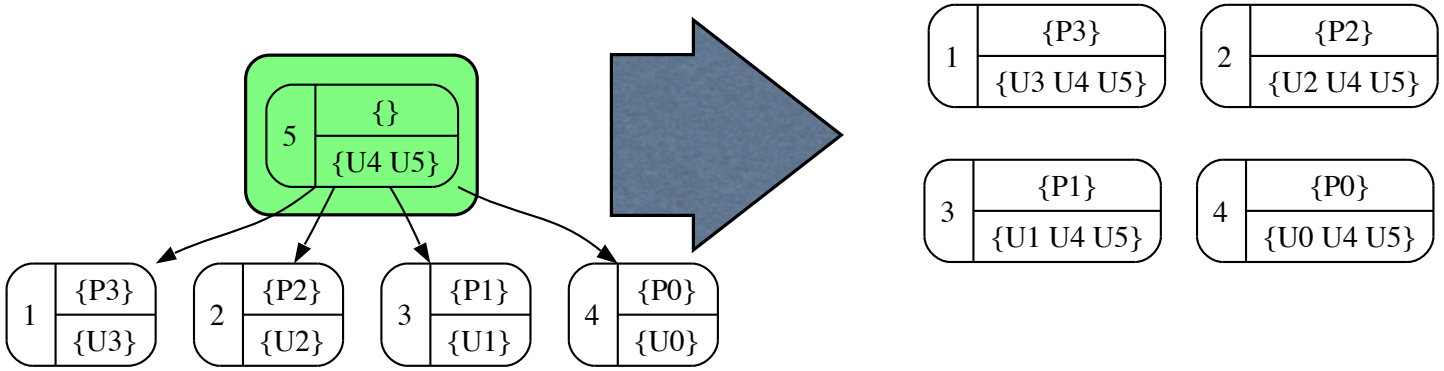


Case 1: No Users or Permissions

5 Roles
4 RH
6 UA

VS.
(Based on WSC)

4 Roles
0 RH
12 UA

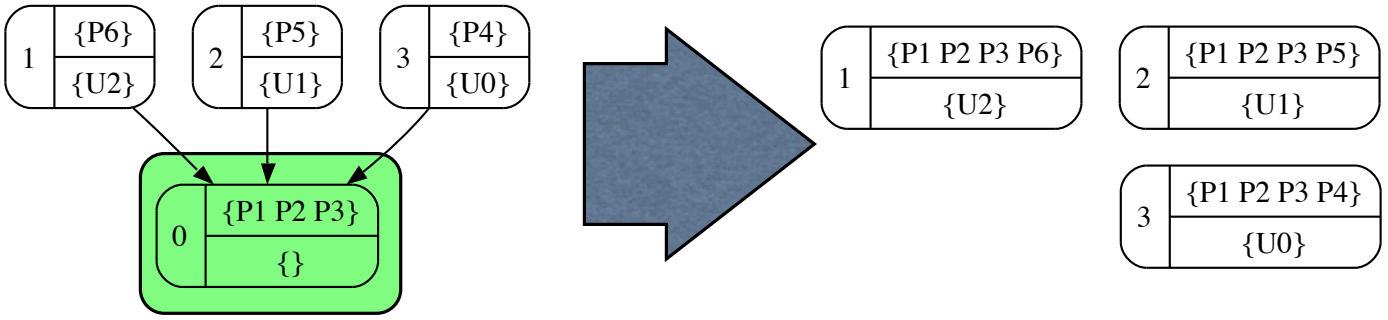


Case 2: No Permissions

4 Roles
3 RH
6 PA

VS.
(Based on WSC)

3 Roles
0 RH
12 PA

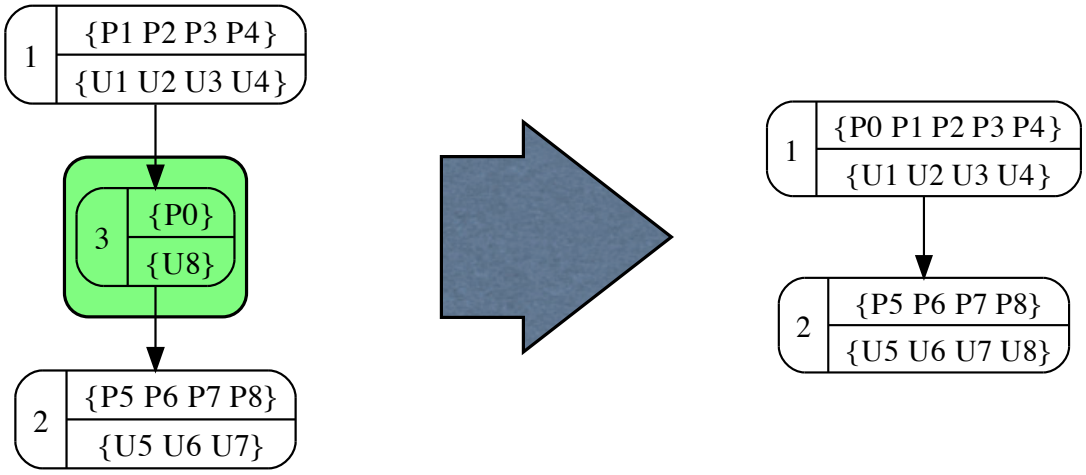


Case 3: No Users

3 Roles
2 RH
0 DUPA

VS.
(Based on WSC)

2 Roles
1 RH
1 DUPA



Case 4: Users and Permissions

Noisy Data
&
Policy Errors

Noise in Access Control

1. Errors or Correctness
   - Type I - Over assigned *(not revoked)*
   - Type II - Under assigned *(never assigned)*
2. Applicability
   - RBAC is a compression
   - *80–20* Rule

Other "Noise"

- Missing and unknown values
- Redundant Attributes
  - e.g., US, USA, United States
- Multiple Accounts
  - e.g., imolloy, immolloy, molloyim
- Artificial Users
  - e.g., www-data, root

Approaches

- Rank-Reduced Matrix Factorization
- Detect noise (errors) and anomalies
- Perform prediction of unknown values
- Leverage attributes and additional relations

Matrix Decomposition

- $UP \in \{0,1\}^{n \times m}$
- $A \in \mathbb{R}^{n \times k},\ B \in \mathbb{R}^{m \times k}$
- $UP \approx AB^T$

$$g : \mathbb{R}^{n \times m} \rightarrow \{0,1\}^{n \times m}$$

	P1	P2	P3
U1	1	1	0
U2	0	1	1
U3	1	1	1
U4	1	1	1
U5	1	1	1
U6	1	1	1

=

	R1	R2
U1	1	0
U2	0	1
U3	1	1
U4	1	1
U5	1	1
U6	1	1

x

	P1	P2	P3
R1	1	1	0
R2	0	1	1

UP = UA x PA

Decomposition Models

- Singular Value Decomposition *(SVD)*
- Non-Negative Matrix Factorization *(NMF)*
- Logistic PCA *(LPCA)*

- Disjoint Decomposition Model *(DDM)*
- Multi-Assignment Clustering *(MAC)*

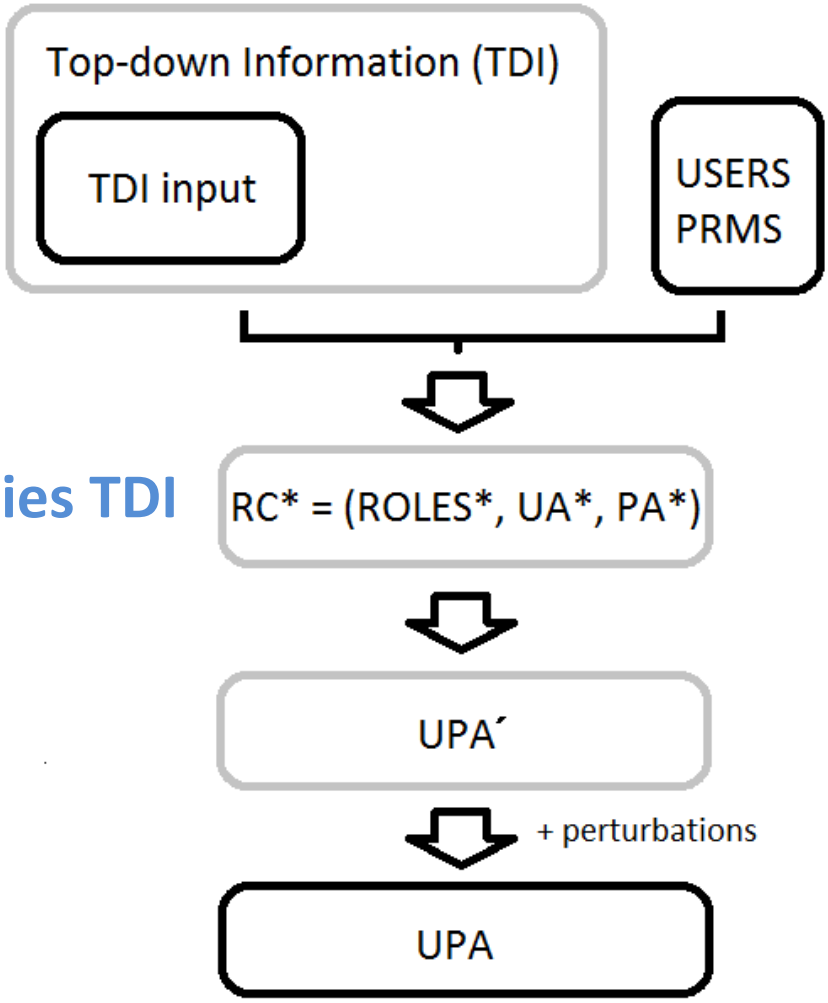
Probabilistic Role Mining

Goal of Probabilistic Role Mining

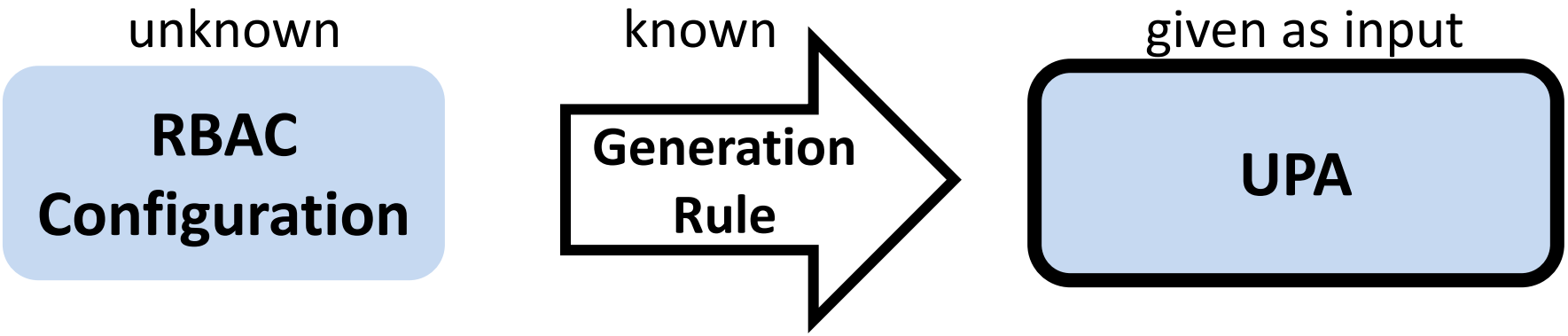
Solve the **Inference RMP**

Assumptions:

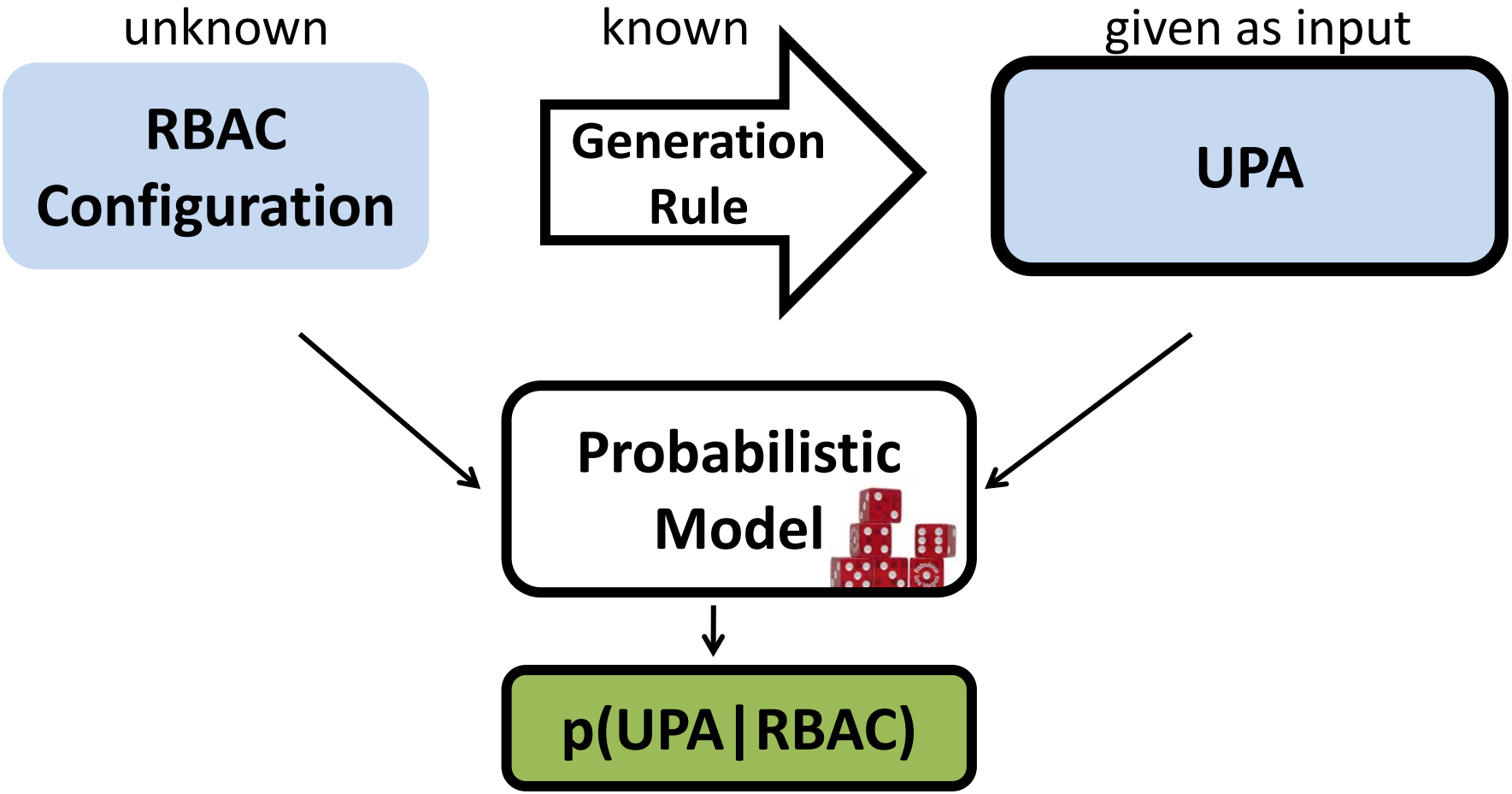
- Underlying structure RC* in UPA
- **Structure reflects business properties TDI**
- Exceptions in UPA



From roles to permissions

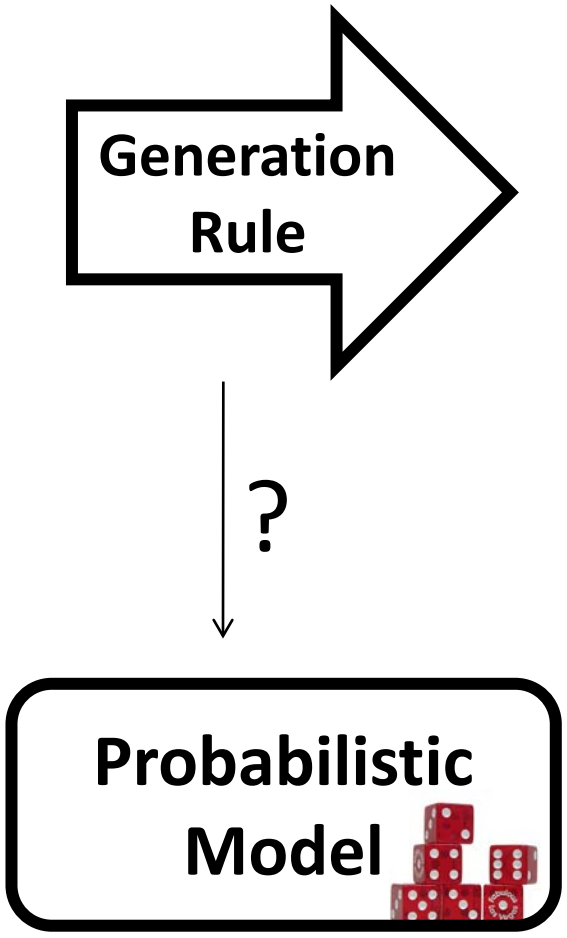


Maximum-likelihood principle



Select the RBAC configuration that **maximizes** p(UPA|RBAC).

From Generation Process to Model

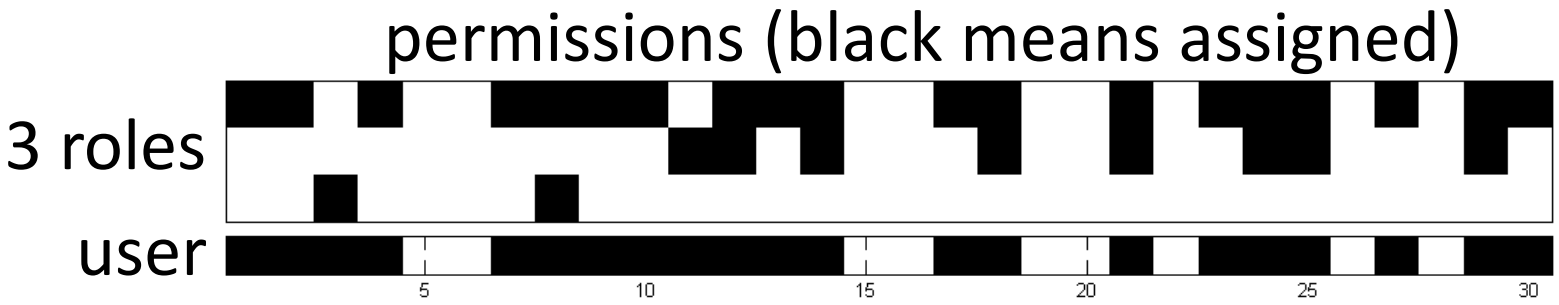


From Generation Process to Model

Generation process:
Boolean disjunction of all roles a user is assigned to.

$$x_{ij} = \bigvee_k [z_{ik} \wedge u_{kj}]$$

Example for a user with 3 roles:

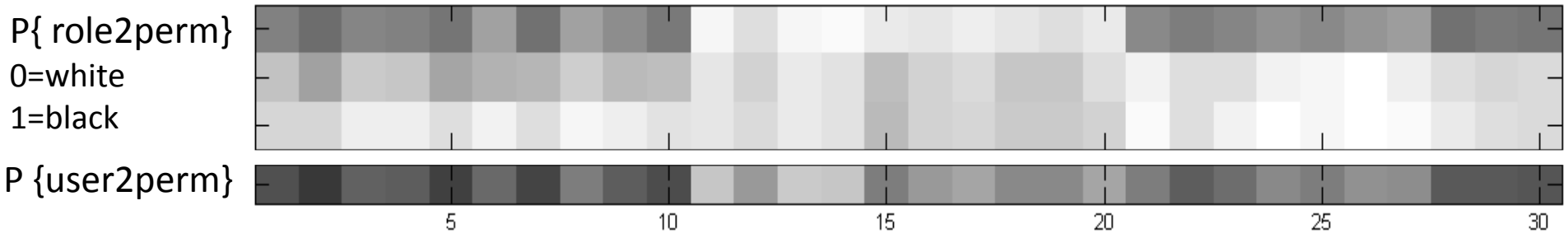


From Generation Process to Model

Replace **binary** permissions by **probabilities** [FBB08] :

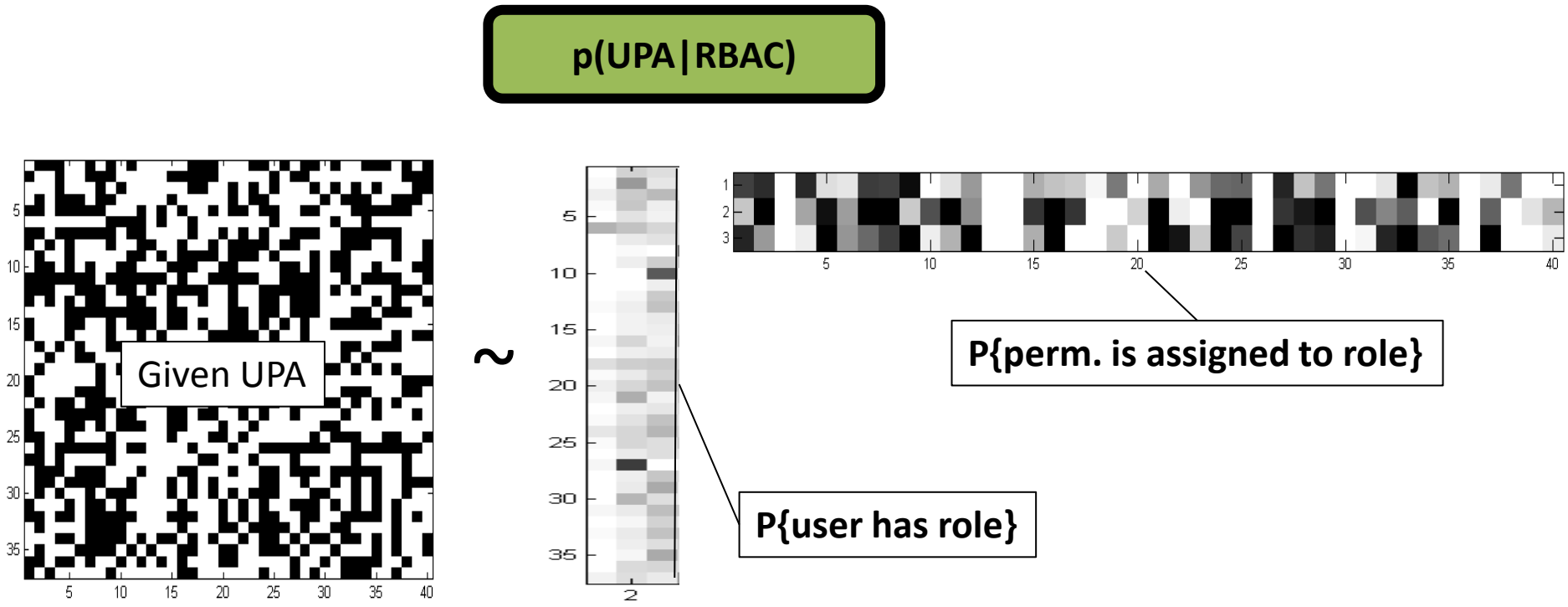
role-permission assignment ⇨ P{perm. is assigned to role}

Example for a user with 3 roles:

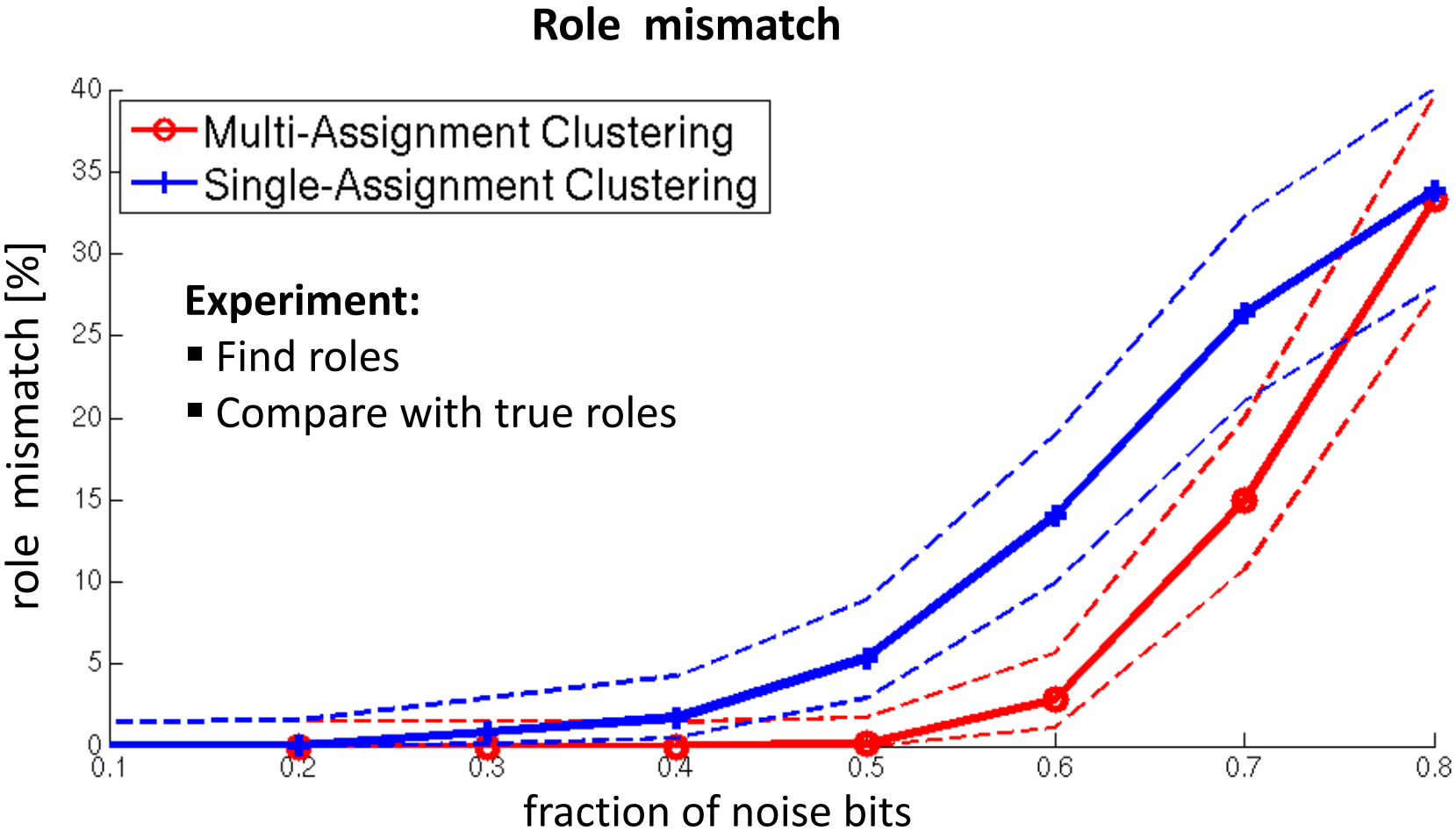


The Model

- Describe the problem with a **probabilistic model.** [SFB+09]
- Infer the model parameters that make UPA **most likely**.

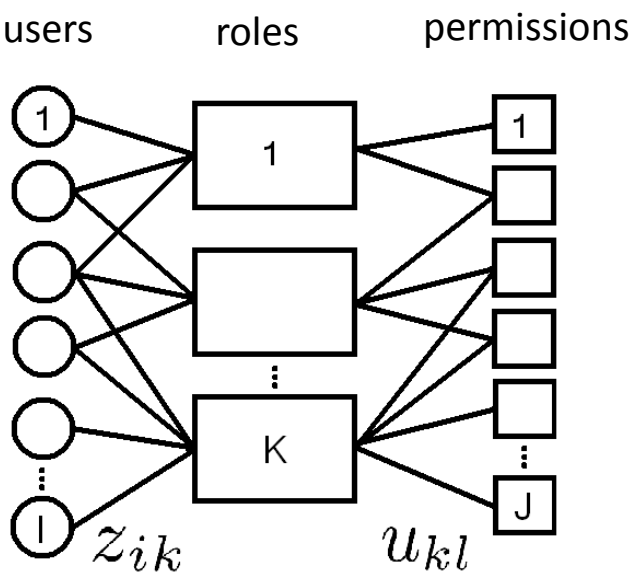


Great for Infering Underlying Roles

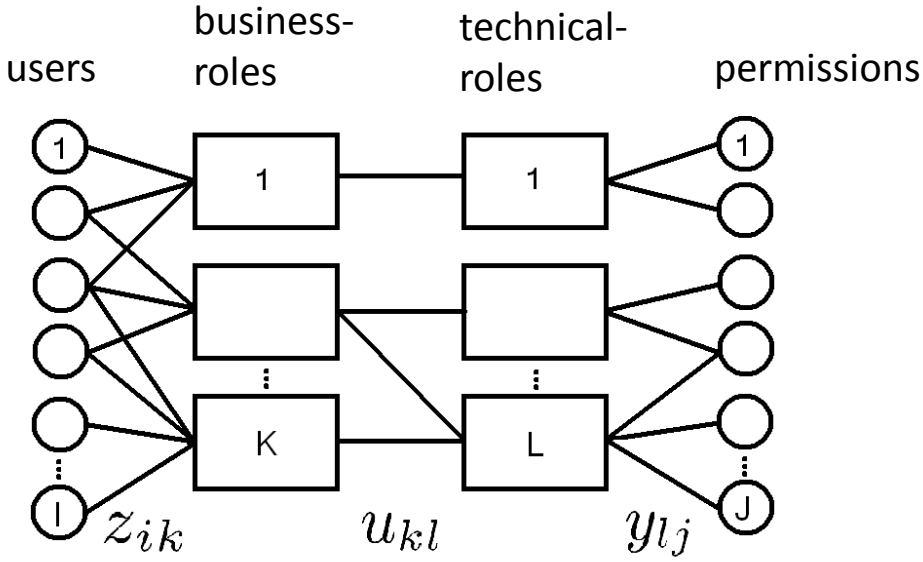


Model variants

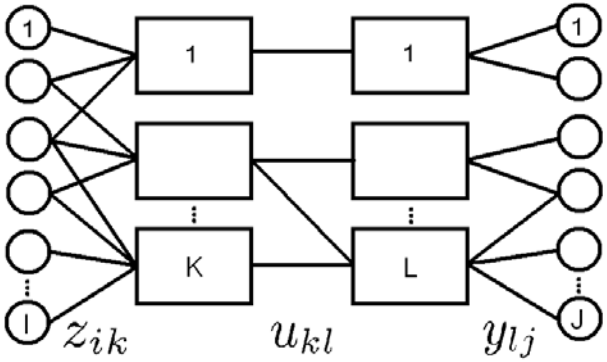
$$\mathbf{x} = \mathbf{z} \otimes \mathbf{u}$$



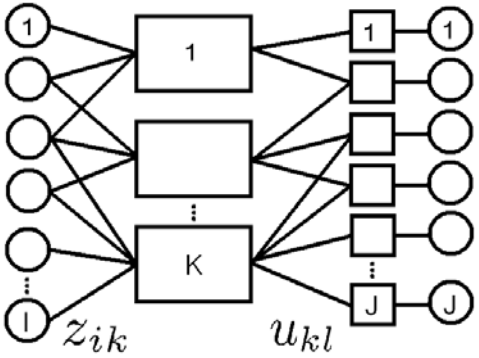
$$\mathbf{x} = \mathbf{z} \otimes \mathbf{u} \otimes \mathbf{y}$$



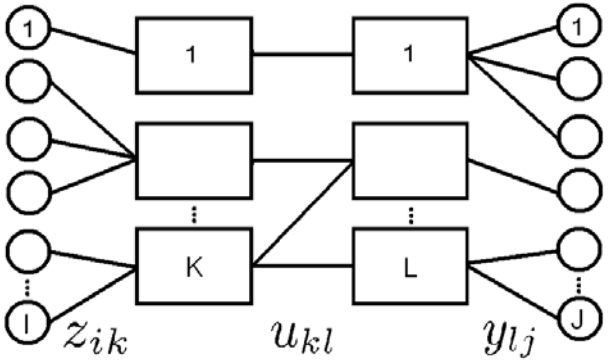
Generic class of models [FBB08]



General



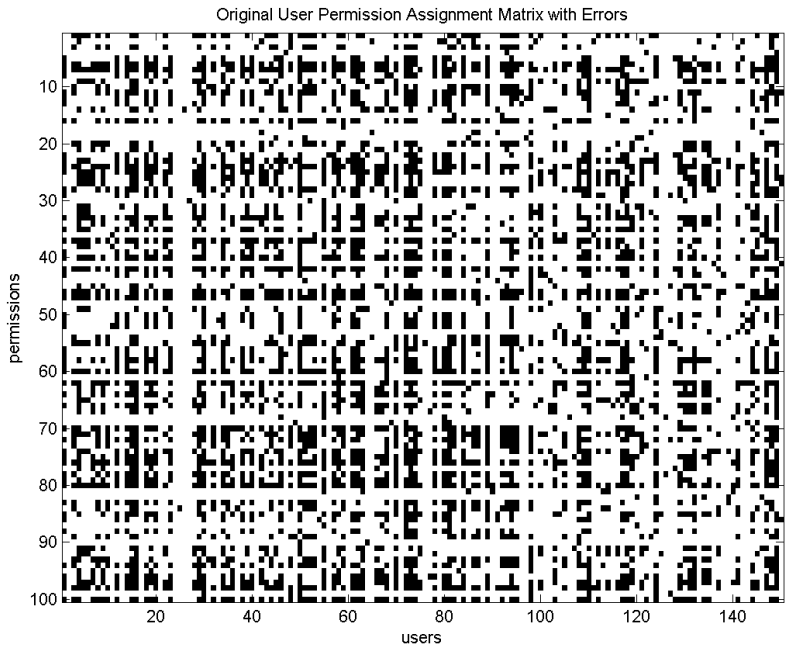
Plain RBAC

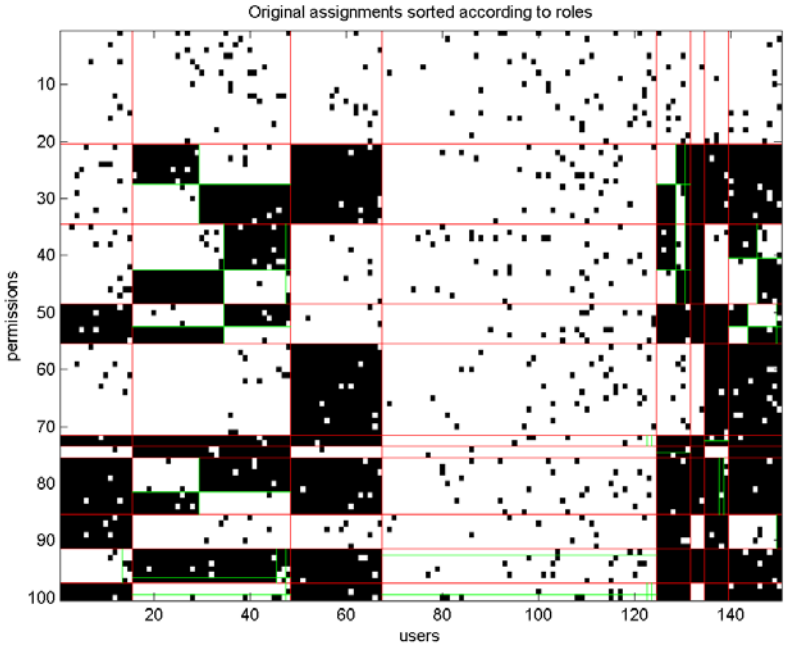


Disjoint Decomposition

$$p\left(\mathbf{x} \mid \mathbf{z}, \mathbf{y}\right) = \prod_{i,j} \left[1 - \prod_{k,l} p(\overline{u_{kl}})^{y_{lj} z_{ik}}\right]^{x_{ij}} \left[\prod_{k,l} p(\overline{u_{kl}})^{y_{lj} z_{ik}}\right]^{1-x_{ij}}$$

Disjoint Decomposition Model (DDM)

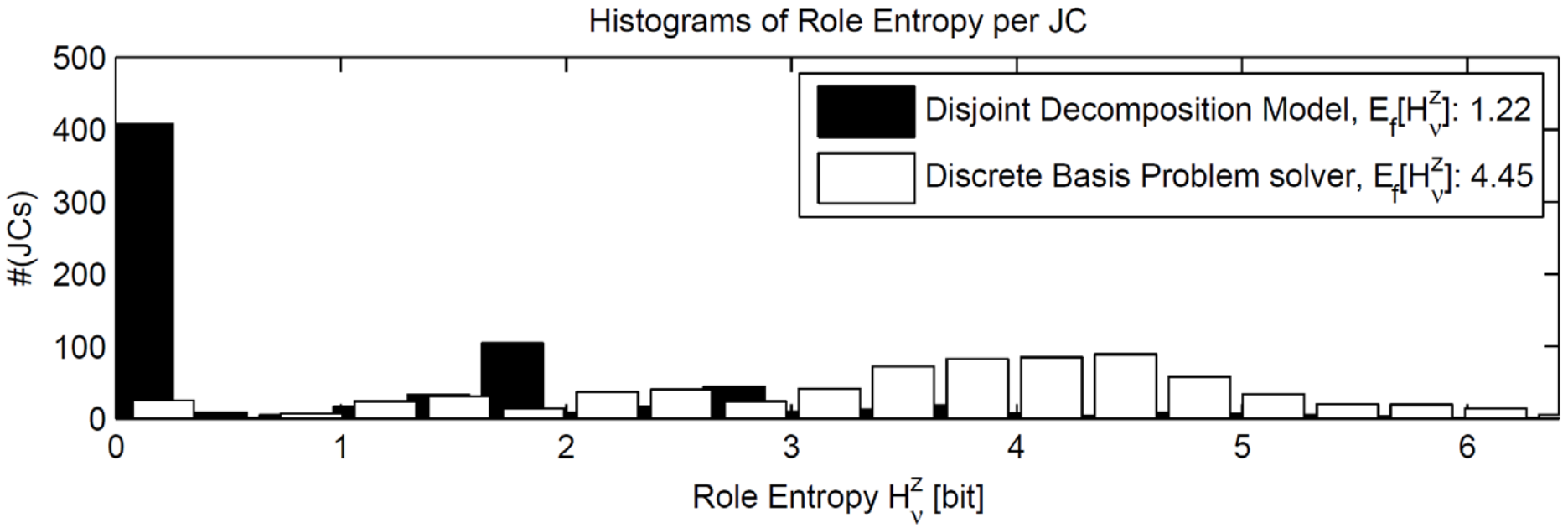




DDM @ work

5000 users on 1323 permissions +740 job-functions

Assessment via **job-code entropy** (business meaning):



Hybrid Role Mining

- User's department, location, title, etc.
- Permission's object, right, granularity, etc.
- Give roles semantic meaning
- Correct recurring errors

Model-Based Hybrid Role Mining [FSB+09]

Combine **two objectives**:

1) The negative **log-likelihood**

$$R_{i,\mathcal{L}}^{(ll)} = -\log\left(\prod_j p_M\left(x_{ij} \mid z_{i\cdot}, \beta, r, \epsilon\right)\right)$$

2) **Business properties objective function**

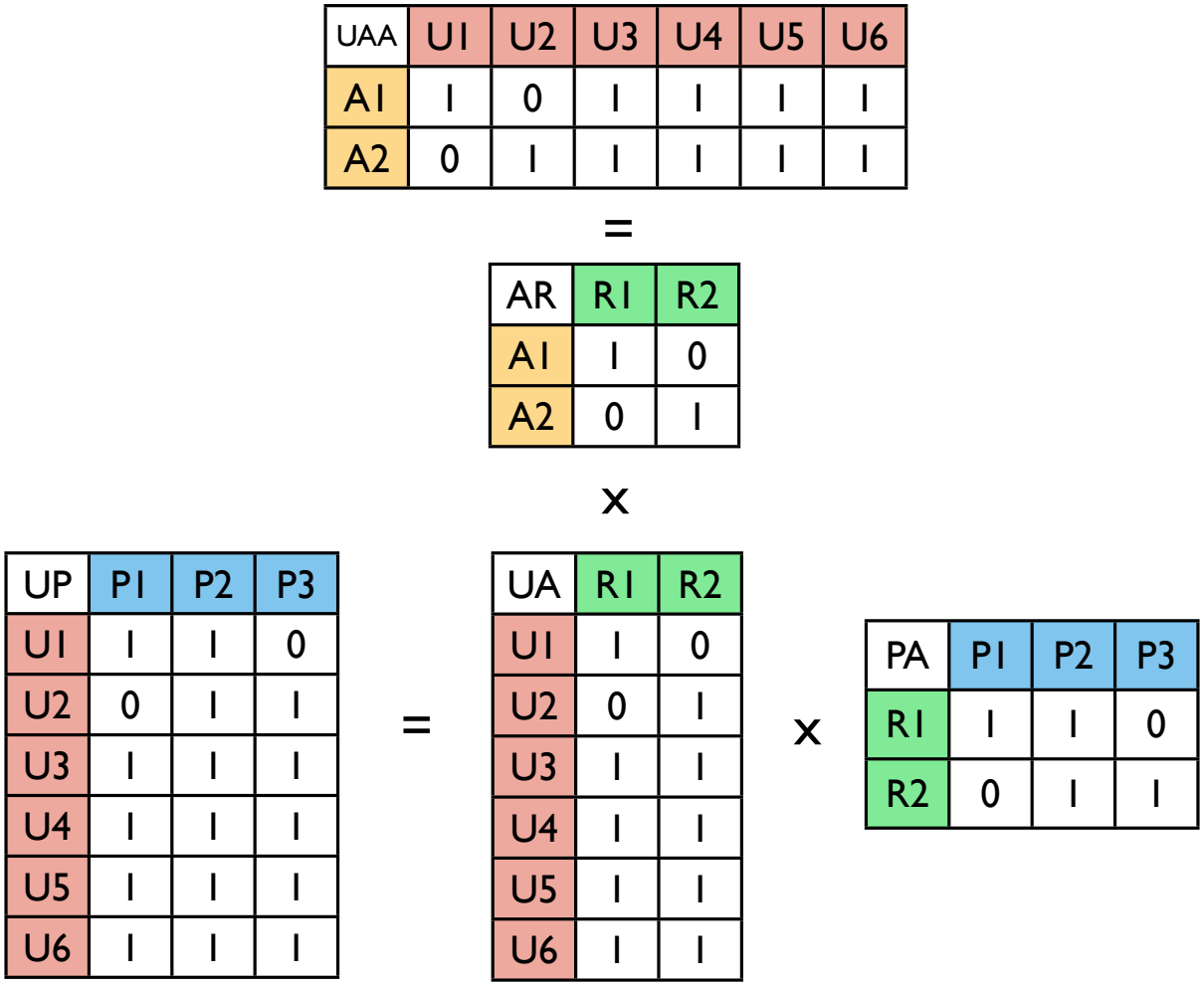
Example: pairwise costs

$$R^{(S)} = \frac{1}{N}\sum_s \sum_{i,i'} w_{is} w_{i's} \sum_k z_{i'k}\left(1 - 2 z_{i'k} z_{ik}\right)$$

Combined objective function: $R = R^{(ll)} + \lambda R^{(S)}$

Collective Matrix Factorization [SG08]

- $UAA \in \mathbb{R}^{\ell \times n} \quad UP \in \{0,1\}^{n \times m}$
- $A \in \mathbb{R}^{n \times k},\ B \in \mathbb{R}^{m \times k},\ C \in \mathbb{R}^{\ell \times k}$
- $UP \approx AB^T \quad UAA \approx CA^T$ Share Matrix A
- $\alpha D(UP \parallel AB^T) + (1-\alpha) D(UAA \parallel CA^T)$


UAA U1 U2 U3 U4 U5 U6
A1 1 0 1 1 1 1
A2 0 1 1 1 1 1
=
AR R1 R2
A1 1 0
A2 0 1
x
UP P1 P2 P3
U1 1 1 0
U2 0 1 1
U3 1 1 1
U4 1 1 1
U5 1 1 1
U6 1 1 1
=
UA R1 R2
U1 1 0
U2 0 1
U3 1 1
U4 1 1
U5 1 1
U6 1 1
x
PA P1 P2 P3
R1 1 1 0
R2 0 1 1

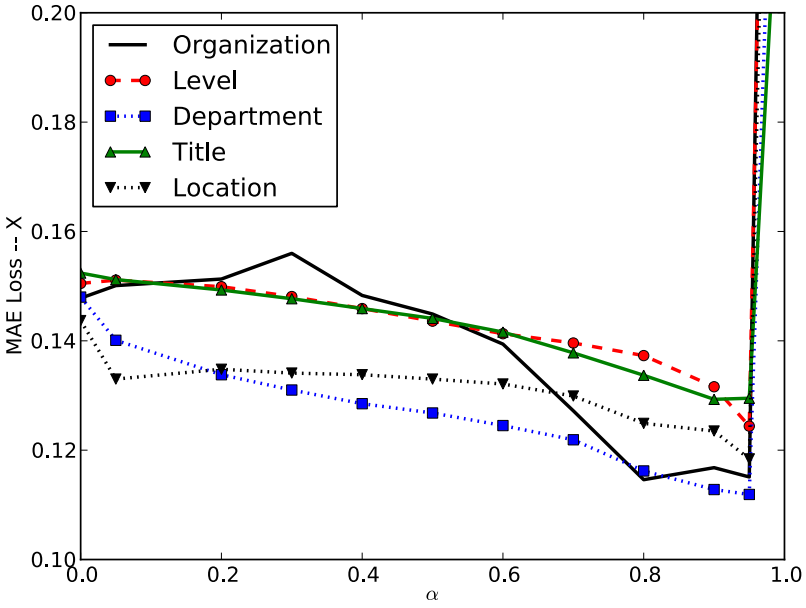
Which Attributes?

- All [MCL+08]
- Entropy Reduction [FSB+09]
  - $\frac{h(p_i) - h(p_i \mid A)}{h(p_i)}$
  - h Shannon Entropy
  - Select the greatest entropy reduction
  - Balance attribute granularity

Prediction with Attributes

[MLL+10]

- Attributes improve predictive performance
- Clusters have more semantic meaning
- Organization outperforms Level



Attributes [MLL+10]

Attribute	Order	Uncert.	Pred. Improv.
Manager	298	2186.03	17.5%
Department	192	1931.95	24.4%
Title	527	1878.51	15.2%
Location	53	1316.92	17.6%
Organization	***12***	***789.46***	***22.5%***
Level	***17***	***170.34***	***17.3%***
Contractor	2	78.44	12.0%
	*		

What is Next?

Technical Challenges

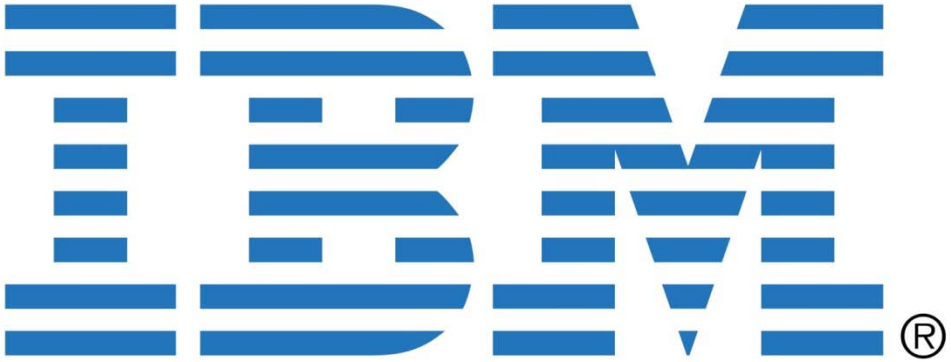
- Data mapping and enforcement
  - Definition of a Permission
- Generating **Role Names**
- Certification and recertification of roles
- Avoid **model-mismatch** (Probabilistic Role Mining)
  - Model selection
- Deal with **structured errors**
  - Add feedback loop
- **Errors** vs. **intended exceptions**
  - Add plausibility analysis
- Dataset **Size**

Sizes

- Size of the datasets
  - Millions of users and permissions
  - Running times, memory, etc.
- Partitioning the data
- Conceptually limited by administrator
  - Visualization

Example Dataset Sizes

	Users	Perms	|UP|
Anon.	3,068	3,133	71,596
Customer	854	885	6,753
Swiss Bank	22,353	1,786	
HP	3,485	10,127	185,294



380,000+ Users?

Handle 500,000,000+ Users?

Future Research

- Dynamic Data
  - Historical data, user and **role evolution**
- Compliance, PCI, SOX, HIPPA, etc.
- **Provision new users**, applications
- Permission granularity and **paramaterized roles**

References

- [CDO+09] A. Colantonio, R. Di Pietro, A. Ocello, and N.V. Verde. **A Formal Framework to Elicit Roles with Business Meaning in RBAC Systems**. SACMAT, 2009.
- [EHM+08] A. Ene, W. Horne, N. Milosavljevic, P. Rao, R. Schreiber, and R. Targan. **Fast Exact and Heuristic Methods for Role Minimization Problems**. SACMAT, 2008.
- [FBB08] M. Frank, D. Basin, and J. Buhmann. **A Class of Probabilistic Models for Role Engineering**. CCS, 2008.
- [FBB10] M. Frank, J. Buhmann, and D. Basin. **On the Definition of Role Mining**. SACMAT, 2010.
- [FSB+09] M. Frank, A. Streich, D. Basin, and J. Buhmann. **A Probabilistic Approach to Hybrid Role Mining**. CCS, 2009.
- [KSS03] M. Kuhlmann, D. Shohat, and G. Schimpf. **Role mining — Revealing Business Roles for Security Administration using Data Mining Technology**. SACMAT, 2003.
- [LMQ+07] N. Li, T. Li, I. Molloy, Q. Wang, E. Bertino, S. Calo, and J. Lobo. **Role mining for engineering and optimizing role based access control systems**. Technical report, 2007.
- [LVA08] H. Lu, V. Vaidya, and V. Atluri. **Optimal Boolean Matrix Decomposition: Application to Role Engineering**. ICDE, 2008.
- [MCL+08] I. Molloy, H. Chen, T. Li, Q. Wang, N. Li, E. Bertino, S. Calo, and J. Lobo. **Mining Roles with Semantic Meanings**. SACMAT, 2008
- [MCL+1x] I. Molloy, H. Chen, T. Li, Q. Wang, N. Li, E. Bertino, S. Calo, and J. Lobo. **Mining Roles with Multiple Objectives**. TISSEC, accepted 26 April, 2010.
- [MLL+09] I. Molloy, N. Li, T. Li, Z. Mao, Q. Wang, and J. Lobo. **Evaluating Role Mining Algorithms**. SACMAT, 2009.
- [MLL+10] I. Molloy, N. Li, J. Lobo, Y. Qi, and L. Dickens. **Mining Roles with Noisy Data**. SACMAT, 2010
- [SFB+09] A. Streich, M. Frank. D. Basin, and J. Buhmann. **Multi-Assignment Clustering for Boolean Data**. ICML, 2009.
- [SG08] A. Singh, and G. Gordon. **Relational Learning via Collective Matric Factorization**. KDD, 2008.
- [VAG07] J. Vaidya, V. Atluri, and Q. Guo. **The Role Mining Problem: Finding a Minimal Descriptive Set of Roles**. SACMAT, 2007.
- [VAW06] J. Vaidya, V. Atluri, and J. Warner. **RoleMiner: Mining Roles using Subset Enumeration**. CCS, 2006
- [ZRE07] D. Zhang, K. Ramamahanarao, and T. Ebringer. **Role Engineering using Graph Optimization**. SACMAT, 2007.