

# AN INFORMATION-THEORETIC FRAMEWORK FOR DATA EXPLORATION

FROM ITEMSETS TO EMBEDDINGS, FROM INTERESTINGNESS TO PRIVACY

Tijl De Bie  
Ghent University

(In collaboration with many others)

# SUBJECTIVITY = KEY

Three motivating examples:

- 1. Frequent itemset mining**

- Individually frequent items = probably frequent together

- 2. Graph embedding**

- High degree nodes = probably embedded centrally

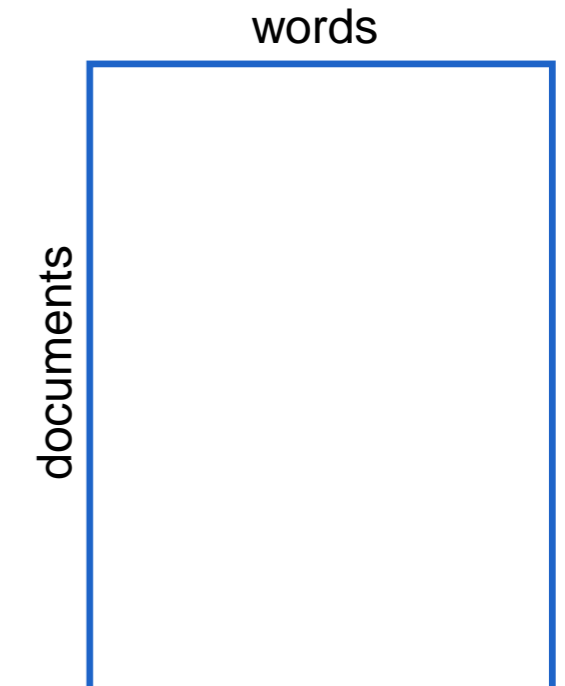
- 3. Privacy-preserving data publishing**

- Accounting for background knowledge is key

# ASSOCIATION ANALYSIS / ITEMSET MINING

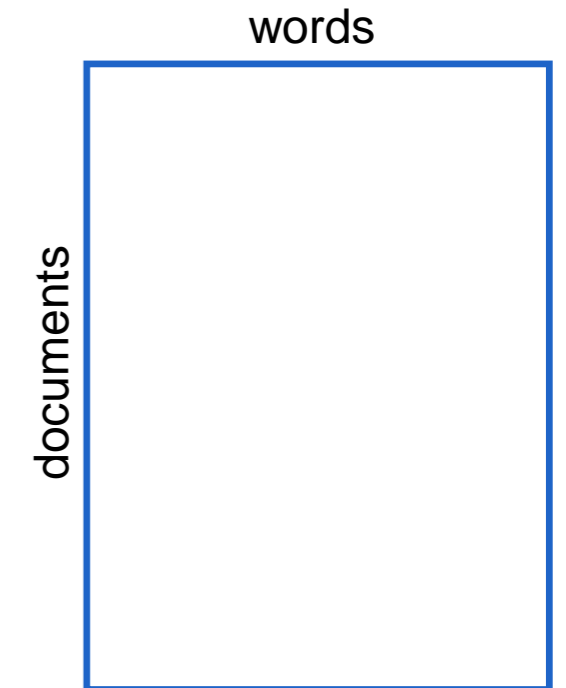
Subjective interestingness ranking		Support x size (area)	
Prior info on: Row & column sums	#docs		#docs
svm, support, machin, vector	25	data, paper	389
state, art	39	algorithm, propose	246
unlabelled, labelled, supervised, learn	10	data, mine	312
associ, rule, mine	36	base, method	202
gene, express	25	result, show	196
frequent, itemset	28	problem	373
large, social, network, graph	15	data, set	279
column, row	13	approach	330
algorithm, order, magnitud, faster	12	model	301
paper, propos, algorithm, real, synthetic, data	27	present	296

NIPS abstracts dataset:



# ASSOCIATION ANALYSIS / ITEMSET MINING

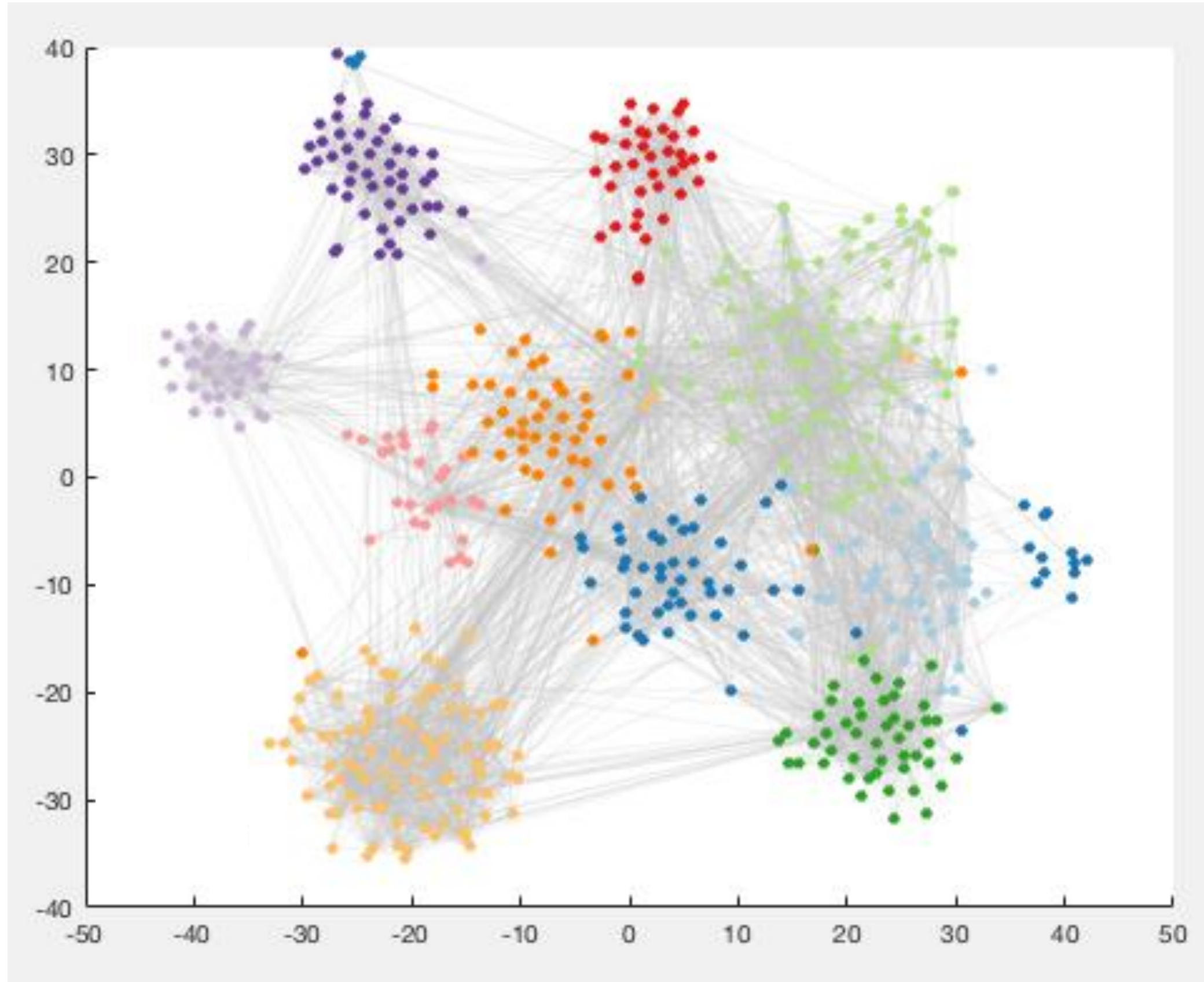
NIPS abstracts dataset:



Subjective interestingness ranking	#docs	Support x size (area)	#docs
Prior info on: Row & column sums			
svm, support, machin, vector	25	data, paper	389
state, art	39	algorithm, propose	246
unlabelled, labelled, supervised, learn	10	data, mine	312
associ, rule, mine	36	base, method	202
gene, express	25	result, show	196
frequent, itemset	28	problem	373
large, social, network, graph	15		
column, row	13		
algorithm, order, magnitud, faster	12		
paper, propos, algorithm, real, synthetic, data	27		

Subjective interestingness ranking	#docs	Subjective interestingness ranking	#docs
Prior info on: Row & column sums		Additionally prior info on: Keyword tiles	
svm, <b>support</b> , machin, vector	25	art, state	39
state, art	39	row, column, algorithm	12
unlabelled, labelled, <b>supervised</b> , learn	10	unlabelled, labelled, data	14
<b>associ</b> , rule, mine	36	answer, question	18
<b>gene</b> , express	25	Precis, recal	14

# CONDITIONAL GRAPH EMBEDDINGS



# PRIVACY-PRESERVING DATA PUBLISHING

- **Anonymization** insufficient to protect sensitive attributes (**linkage attack**)

Anonymized patient database

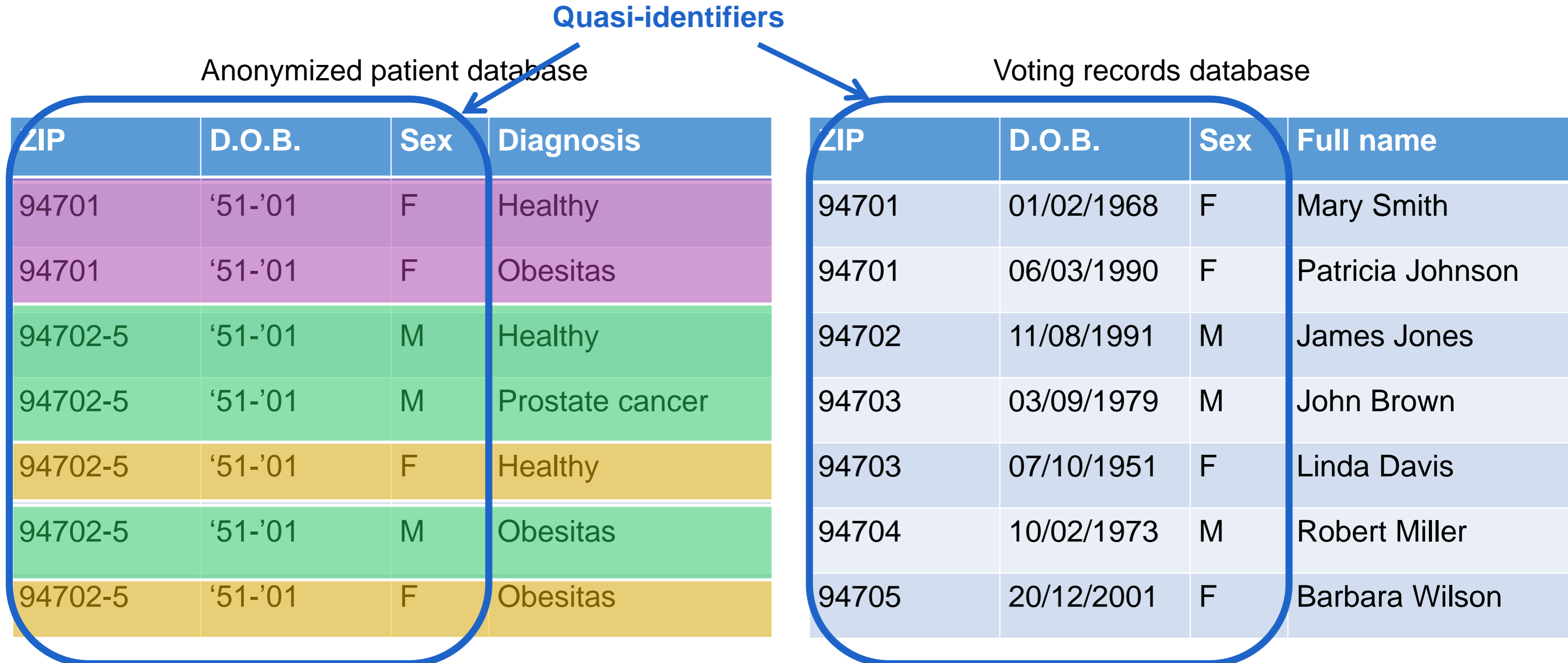
ZIP	D.O.B.	Sex	Diagnosis
94701	01/02/1968	F	Healthy
94701	06/03/1990	F	Obesitas
94702	11/08/1991	M	Healthy
94703	03/09/1979	M	Prostate cancer
94703	07/10/1951	F	Healthy
94704	10/02/1973	M	Obesitas
94705	20/12/2001	F	Obesitas

Voting records database

ZIP	D.O.B.	Sex	Full name
94701	01/02/1968	F	Mary Smith
94701	06/03/1990	F	Patricia Johnson
94702	11/08/1991	M	James Jones
94703	03/09/1979	M	John Brown
94703	07/10/1951	F	Linda Davis
94704	10/02/1973	M	Robert Miller
94705	20/12/2001	F	Barbara Wilson

# PRIVACY-PRESERVING DATA PUBLISHING

- **Anonymization** insufficient to protect sensitive attributes (**linkage attack**)
- **Generalization!**



# EXPLORING DATA

– The search for interesting *patterns* in *data*

- Association analysis

Frequency, lift, confidence, leverage, coverage,...

- Graph embedding

Node2Vec, Path2Vec, MetaPath2Vec,...

- Privacy-preserving data publishing

Discernibility, generalization height, average group size,...

- Dimensionality reduction

PCA, ICA, projection pursuit, Laplacian Eigenmaps, tSNE, LLE,...

- Clustering

K-means clustering, hierarchical clustering, Mixture of Gaussians, spectral clustering,...

- Community detection

Stochastic block modelling, modularity, k-cores, quasi-cliques, dense subgraphs,...

- ...

– Zillions of

Objective functions

Quality functions

Utility functions

Cost functions

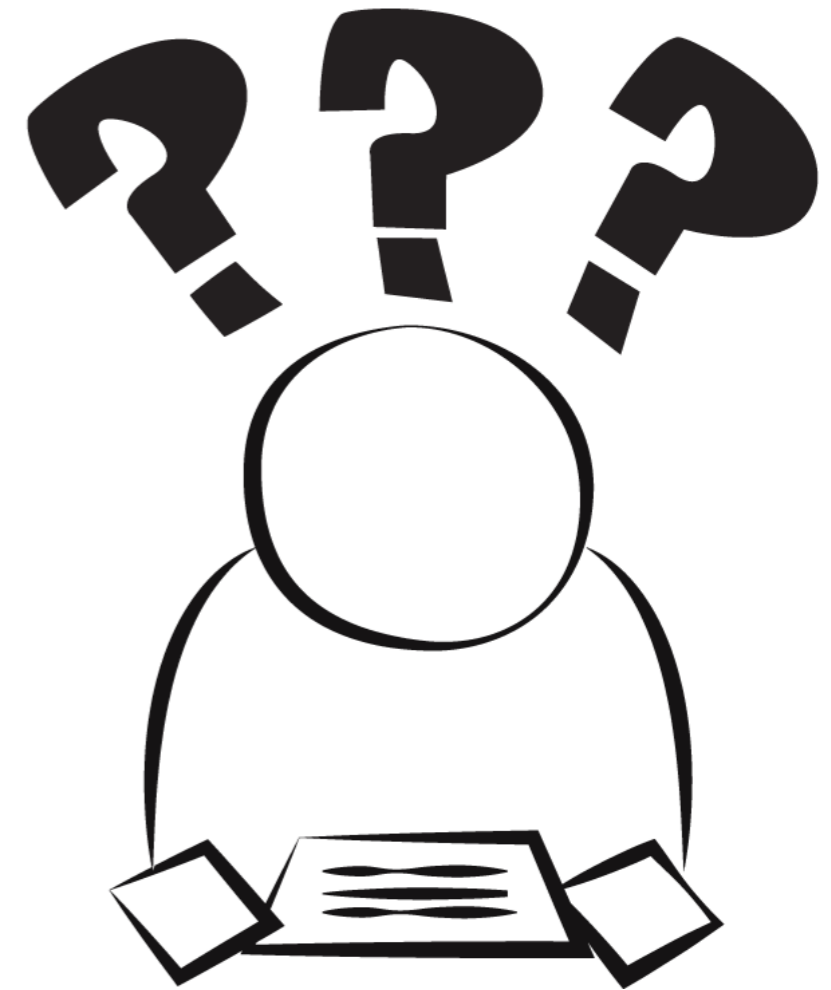
...

← **‘Interestingness measures’**



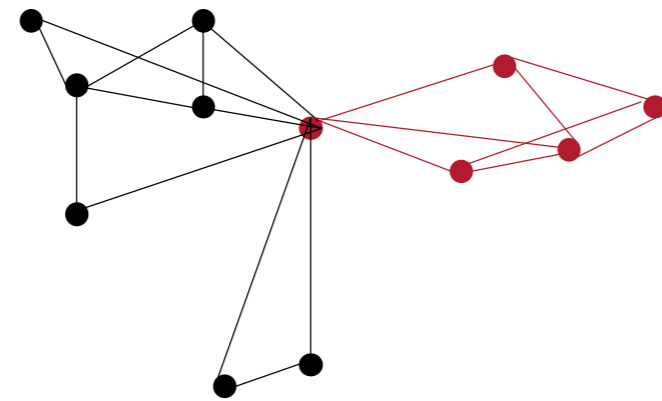
# THE CHALLENGE

- Zillions of interestingness measures = good & bad
  - Good: more options!
  - Bad: the trees & the forest...
- Challenge:
  - **Formalise *true* interestingness!**
    - With minimal user interaction
    - Without requiring user expertise



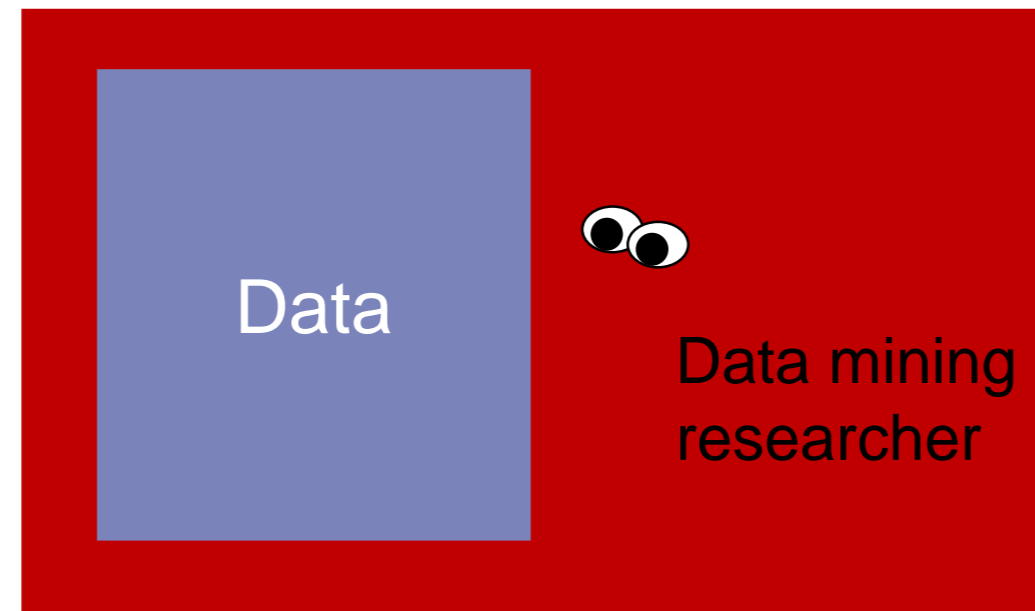
# MOTIVATING EXAMPLE

- Community detection:



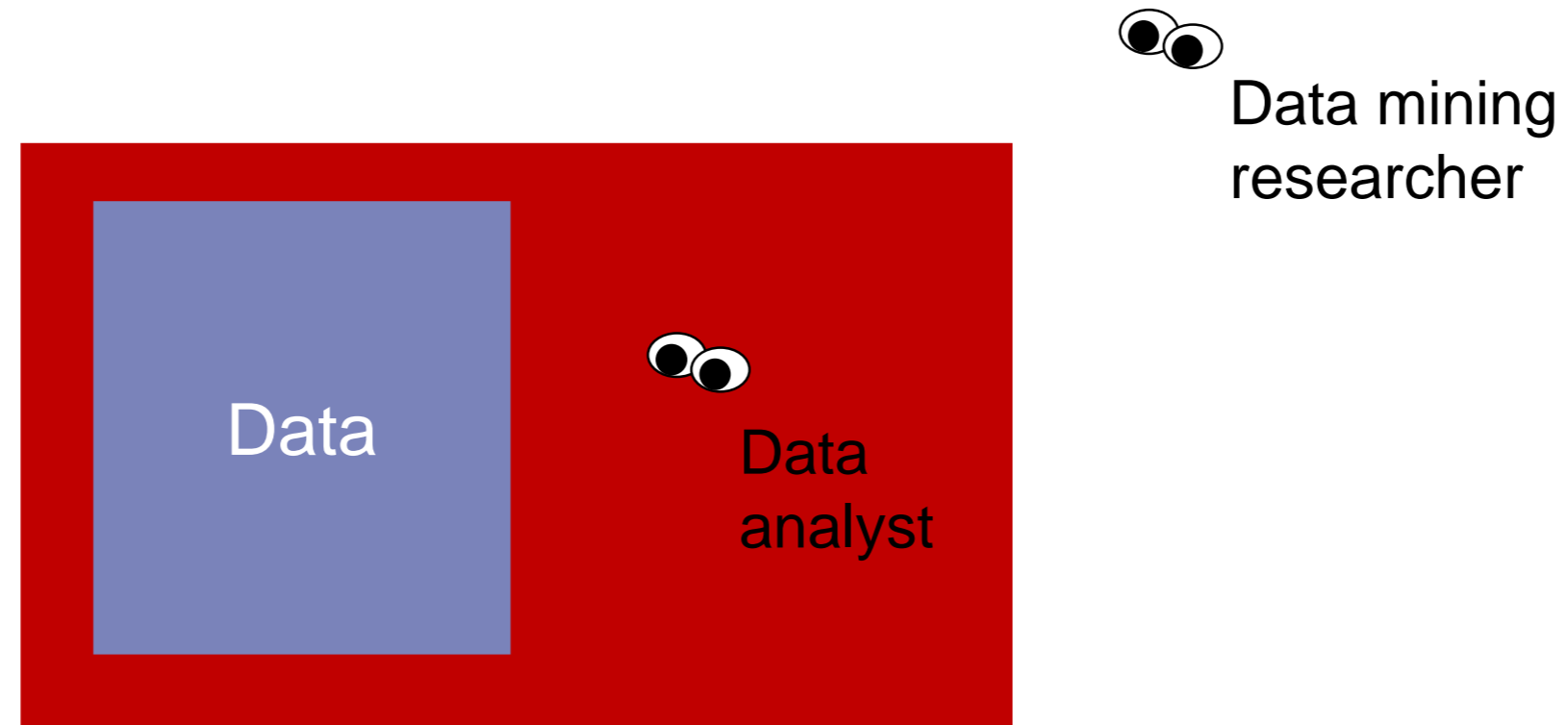
- What makes for an interesting community?
  - Densely connected?
  - Large?
  - Few neighbours outside community?
  - Unrelated to certain known ‘affiliations’?
  - ...

# THE FORSIED APPROACH



Interestingness(pattern)

# THE FORSIED APPROACH

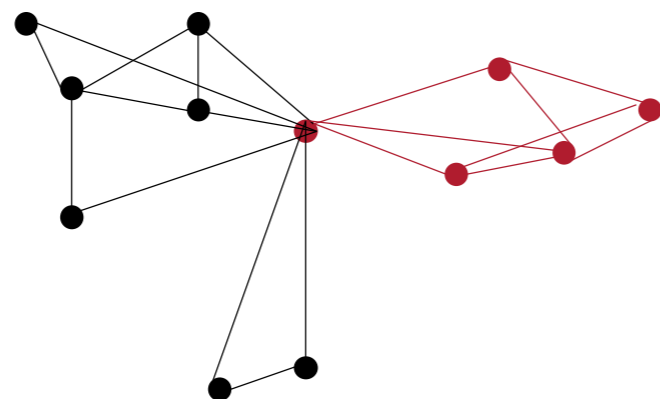


Interestingness(pattern) → Interestingness(pattern, analyst)

Interestingness = **subjective**

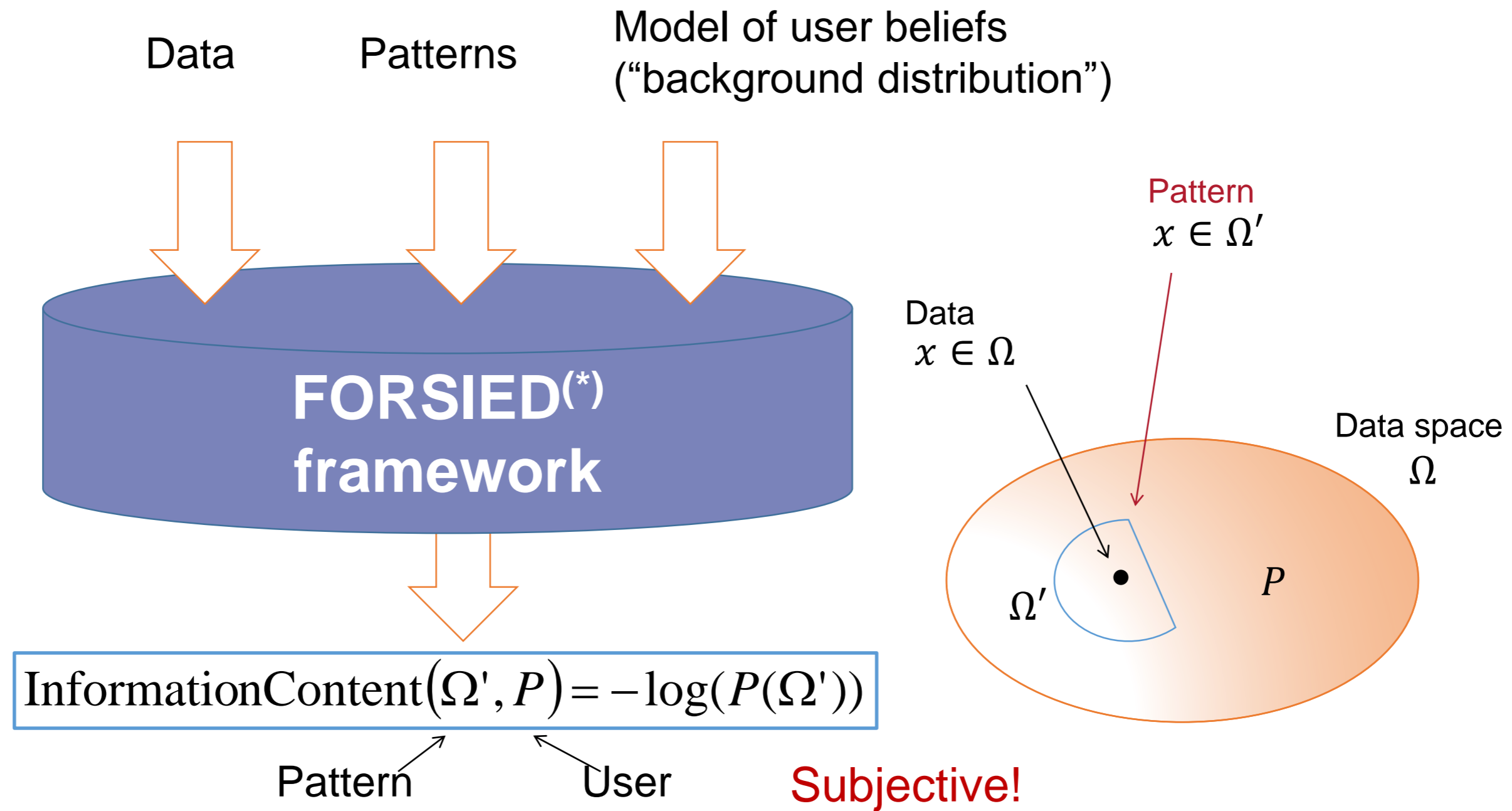
# MOTIVATING EXAMPLE

- Community detection:

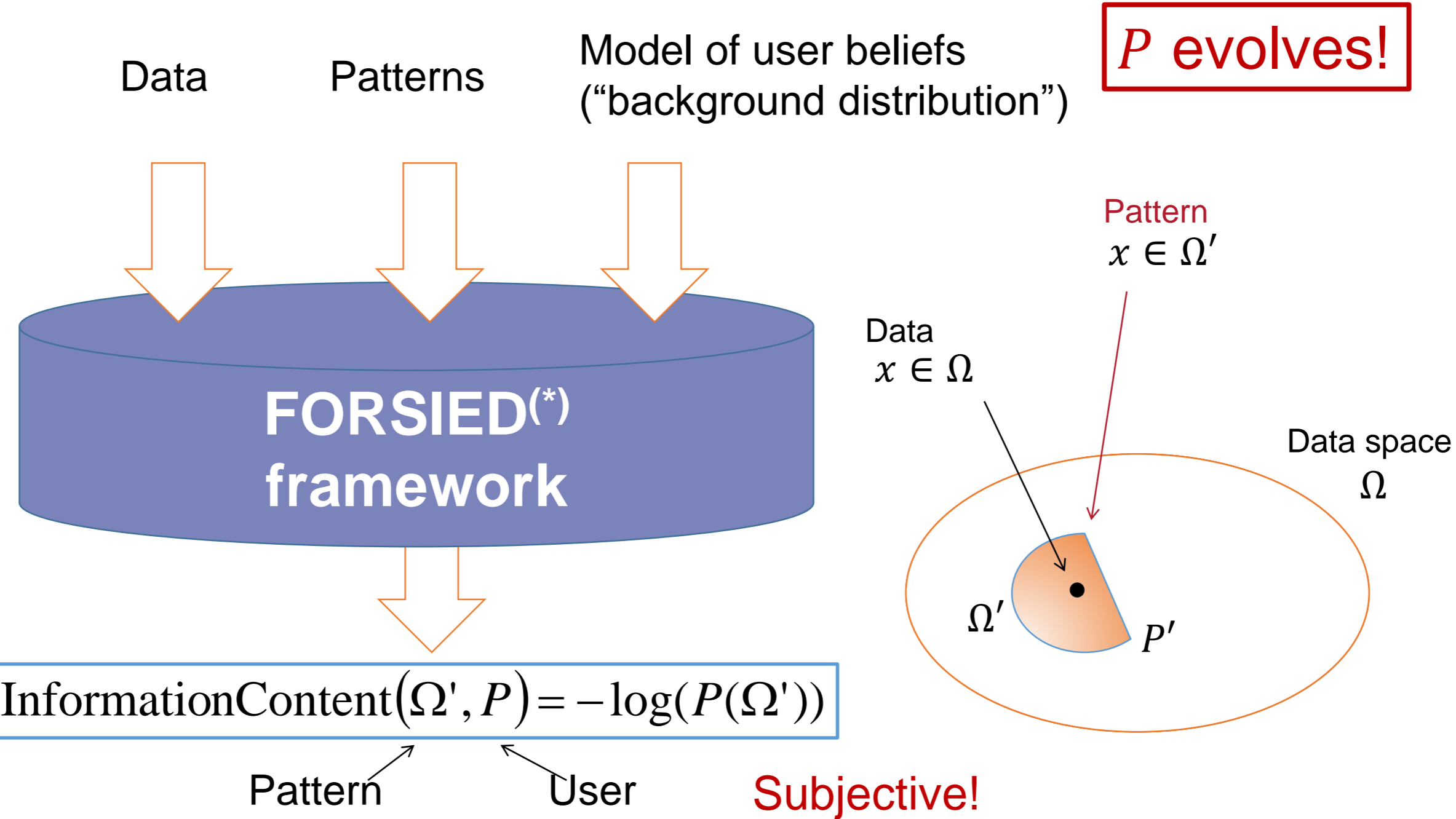


- User states expectations / beliefs
  - Formalized as a ‘background distribution’
- Any ‘pattern’ that contrasts with this and is easy to describe
  - = *subjectively interesting*

# AN INFORMATION THEORETIC APPROACH

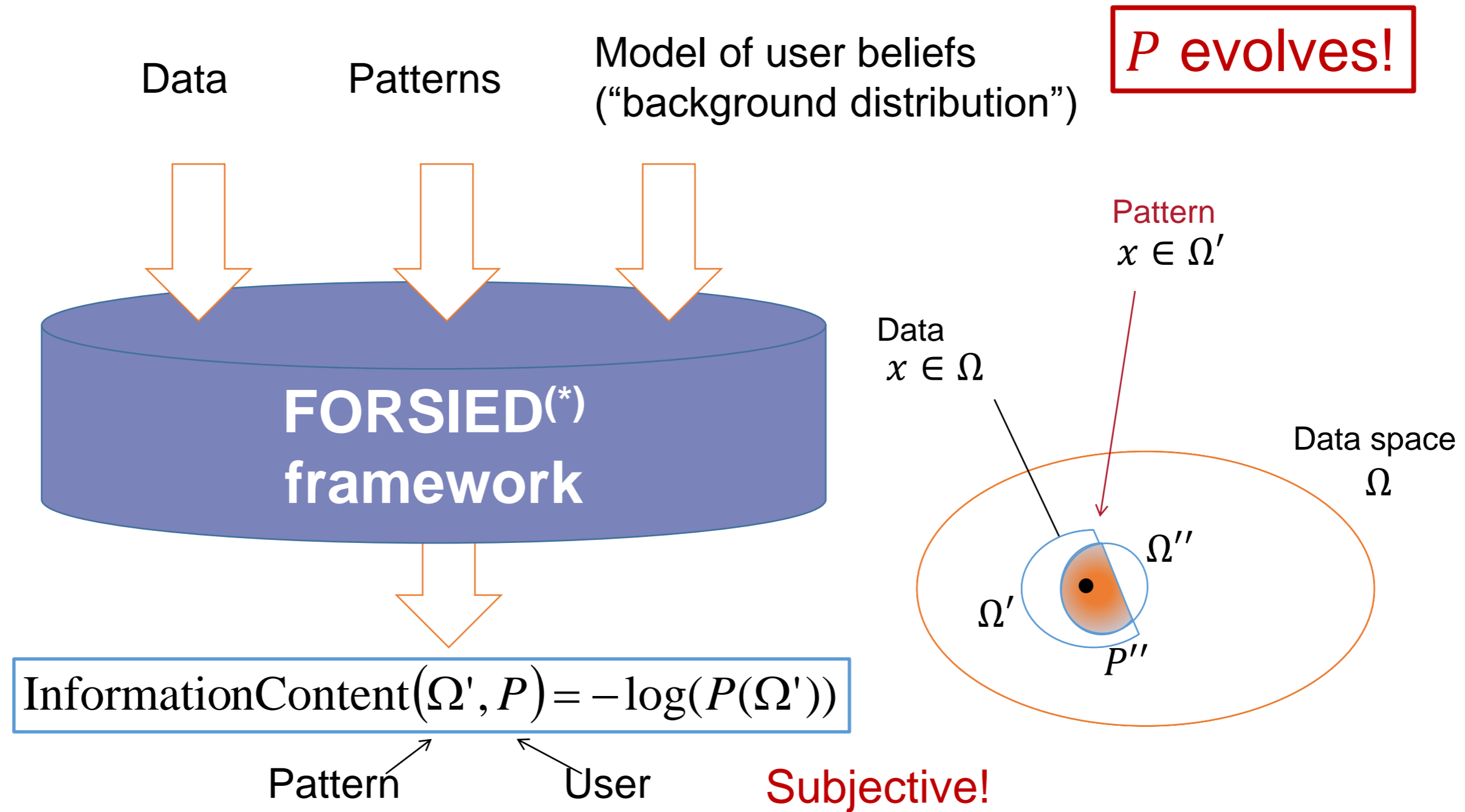


$$\text{Interestingness}(\Omega', P) = \frac{\text{InformationContent}(\Omega', P)}{\text{DescriptonalComplexity}(\Omega')}$$



$$\text{Interestingness}(\Omega', P) = \frac{\text{InformationContent}(\Omega', P)}{\text{DescriptiveComplexity}(\Omega')}$$





$$\text{Interestingness}(\Omega', P) = \frac{\text{InformationContent}(\Omega', P)}{\text{DescriptiveComplexity}(\Omega')}$$

# THE FINE PRINT

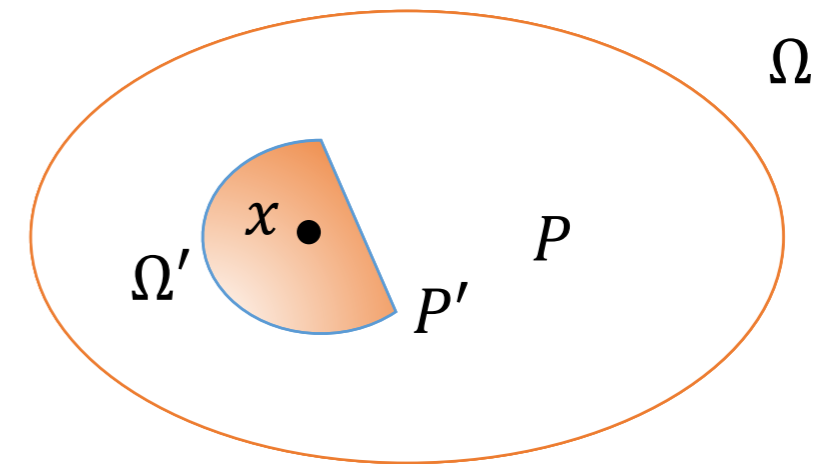
- Initial background distribution  $P$ ?
- **Maximum entropy** distribution

$$\max_P E_{X \sim P} \{-\log P(X)\}$$

- Updated background distribution  $P'$  given pattern  $x \in \Omega'$ ?
- $P$  **conditioned** onto event  $x \in \Omega'$

$$P'(\Omega'') = \frac{P(\Omega'' \cap \Omega')}{P(\Omega')}$$

- Descriptive complexity?
- Essentially problem-dependent



# FORSIED INSTANTIATIONS

# COMMUNITY DETECTION IN NETWORKS

with Matthijs van Leeuwen, Eirini Spyropoulou

## Data:

– Graph

## Prior beliefs:

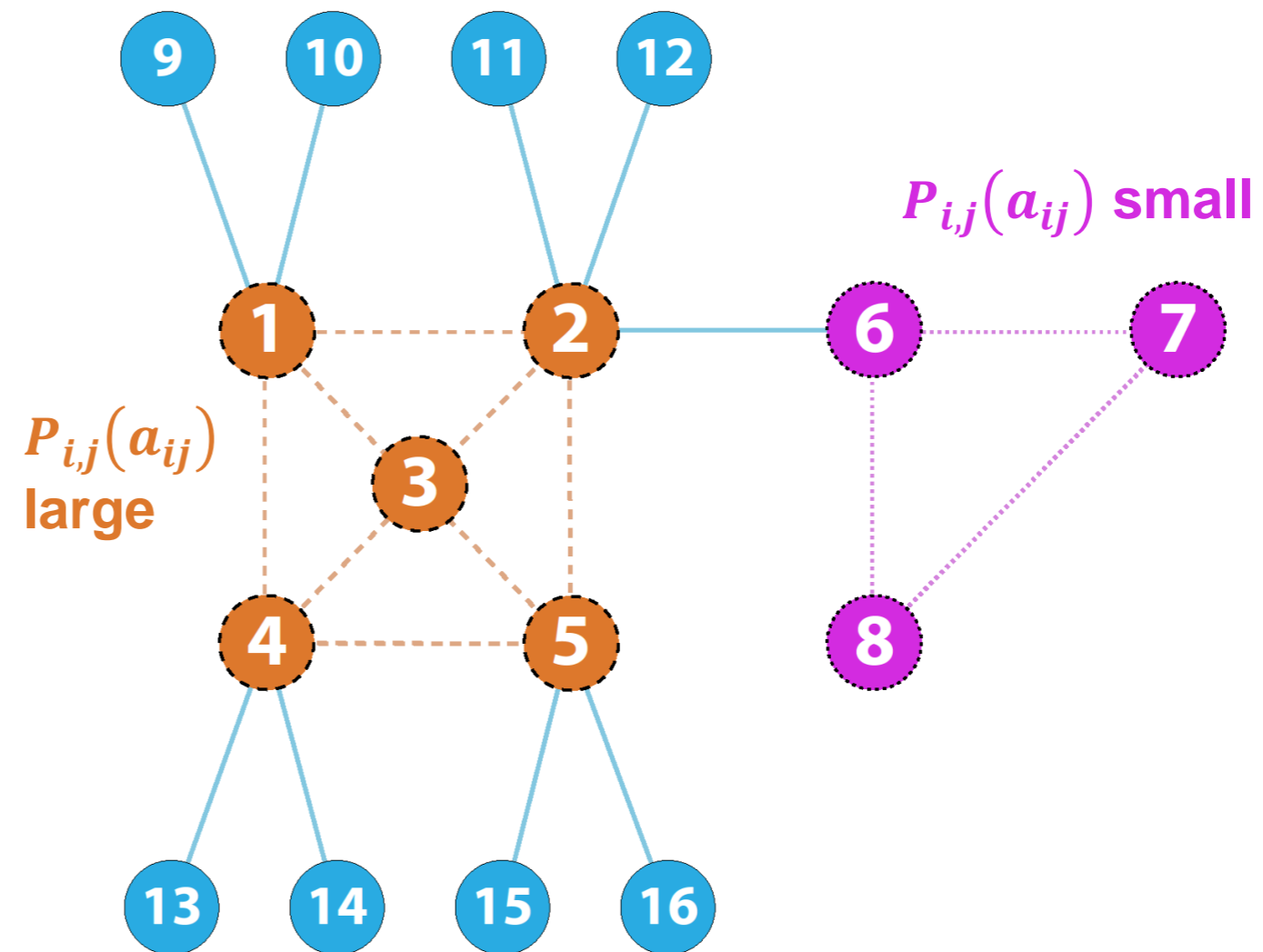
1. Overall density
2. or: Vertex degrees

→ MaxEnt distribution:

$$P(A) = \prod_{i>j} P_{i,j}(a_{ij})$$

Adjacency matrix  $\uparrow$   $P_{i,j}(a_{ij})$   $\uparrow$  Edge indicator variables

$$P_{i,j}(a_{ij}) = \frac{\exp(a_{ij} \cdot (\lambda_i + \lambda_j))}{1 + \exp(\lambda_i + \lambda_j)}$$



# COMMUNITY DETECTION IN NETWORKS

## Data:

- Graph

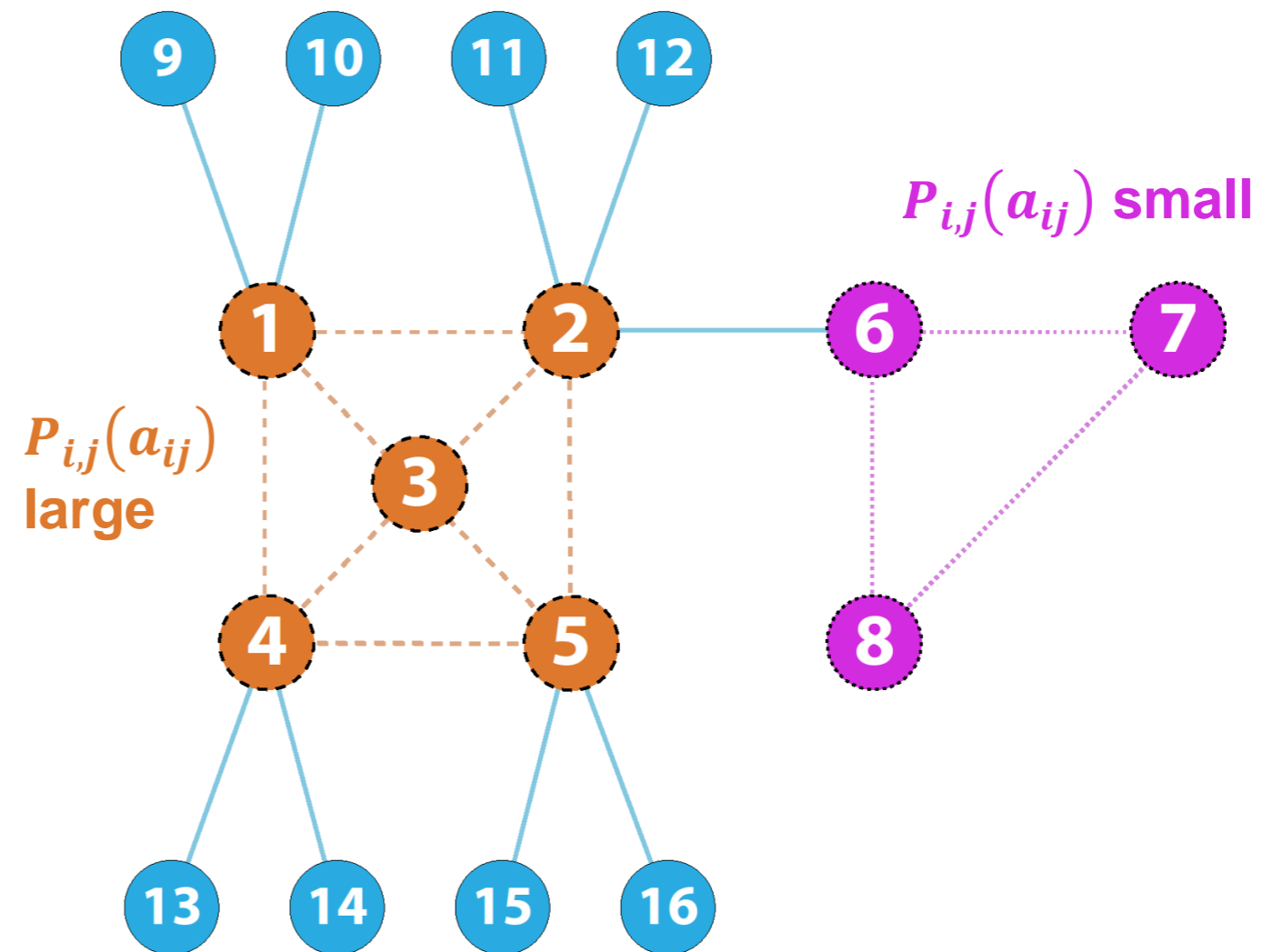
## Prior beliefs:

1. Overall density
2. or: Vertex degrees

## Pattern:

- Dense subgraphs

$$\sum_{i,j \in \text{subgraph}} a_{ij} \geq k$$



# COMMUNITY DETECTION IN NETWORKS

## Data:

- Graph

## Prior beliefs:

1. Overall density
2. or: Vertex degrees

## Pattern:

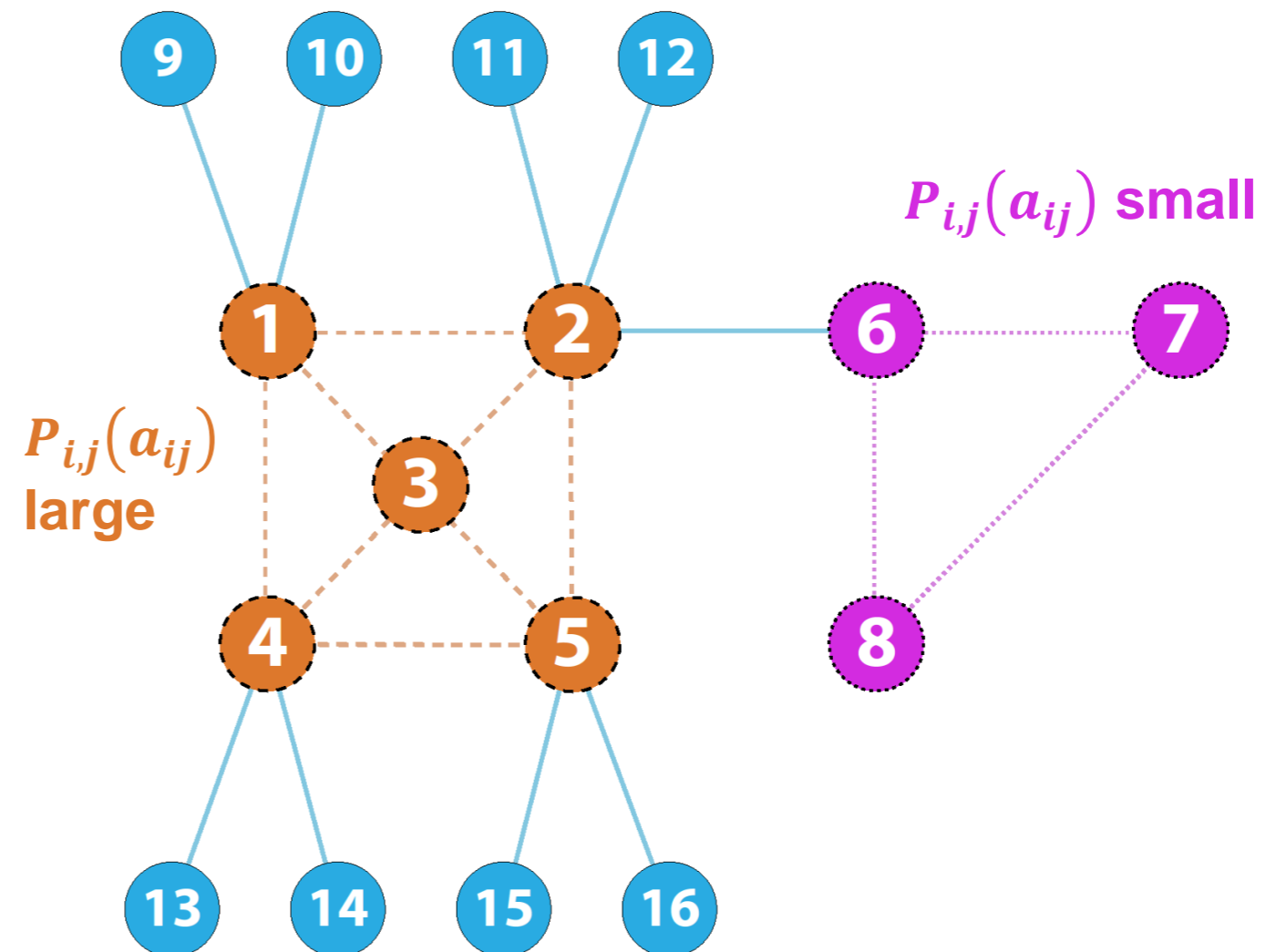
- Dense subgraphs

## Interestingness:

$$-\log P(\text{pattern})$$

---

$$\text{DescriptiveComplexity}(\text{pattern})$$



# COMMUNITY DETECTION IN NETWORKS

## Data:

- Graph

## Prior beliefs:

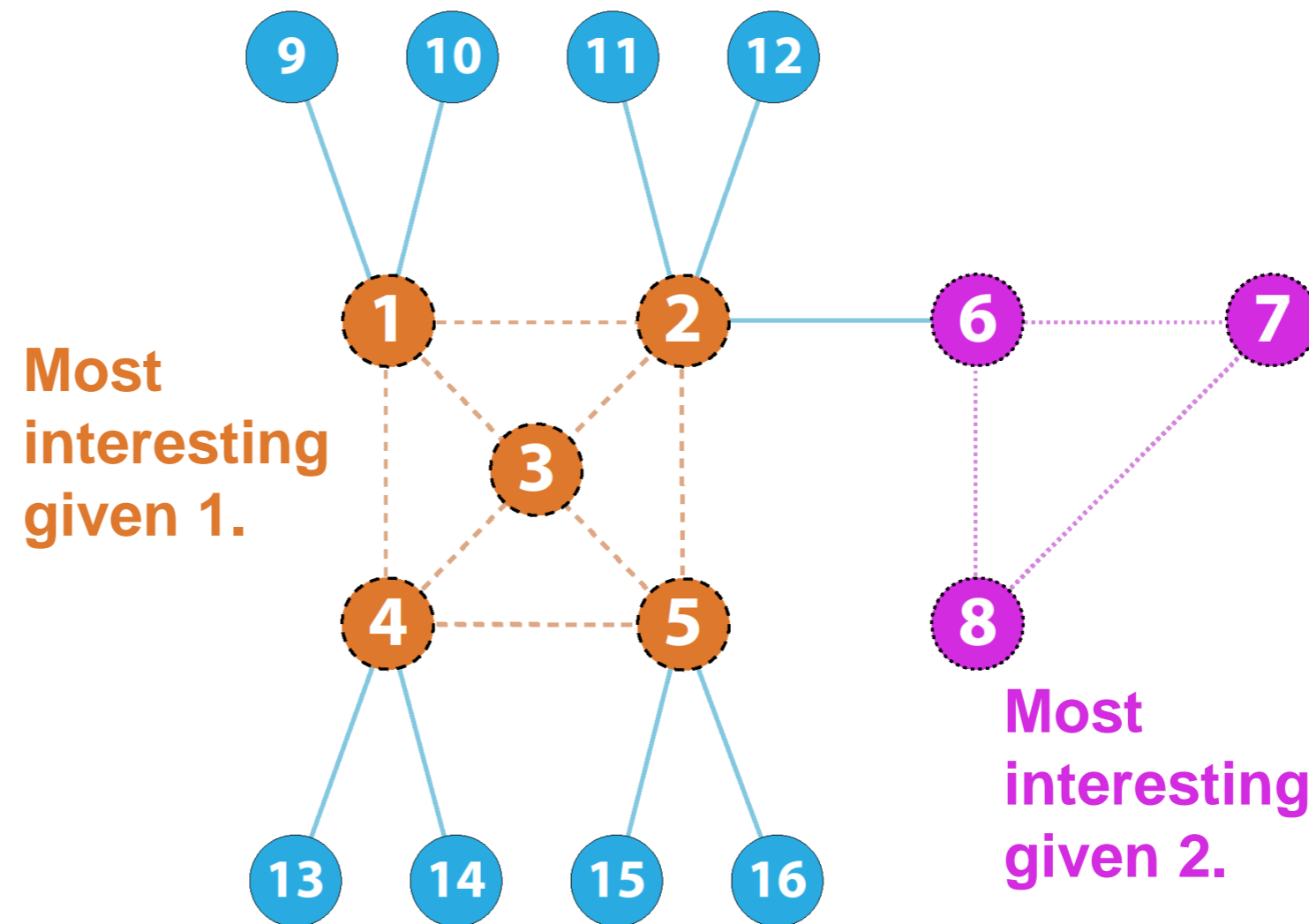
1. Overall density
2. or: Vertex degrees

## Pattern:

- Dense subgraphs

## Interestingness:

- Density vs. size
- 2. → preferably low degree nodes



# COMMUNITY DETECTION IN NETWORKS

## Data:

- Graph

## Prior beliefs:

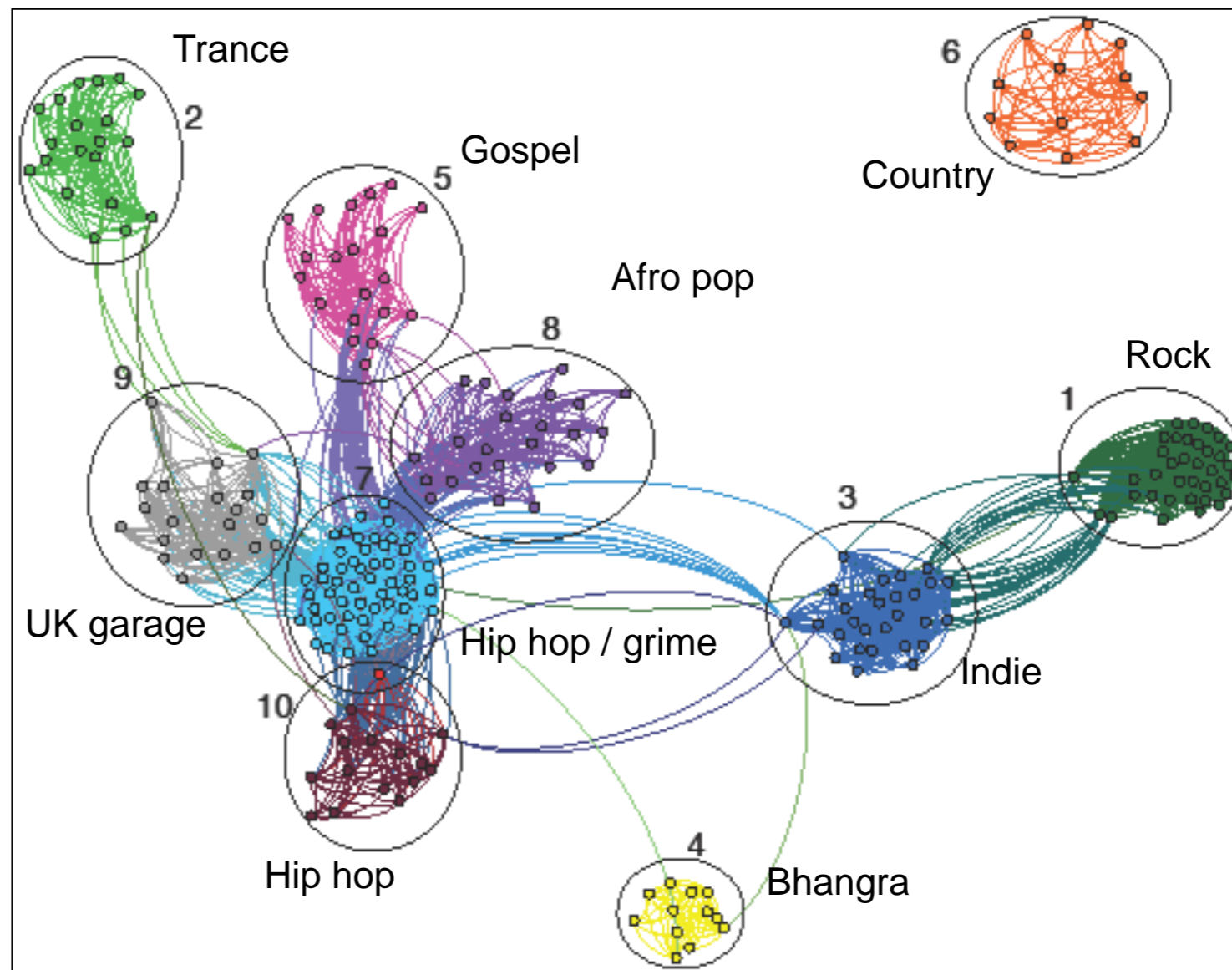
1. Overall density
2. or: Vertex degrees

## Pattern:

- Dense subgraphs

## Interestingness:

- Density vs. size
- 2. → preferably low degree nodes





# ASSOCIATION ANALYSIS

with Akis Kontonasis

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$

	Beer	Diapers	Lipstick	Carrier
Alice	1	1		1
Bob	1		1	1
Charlie	1	1		
Denise	1			1
Eve			1	1
Frankie		1		1

# ASSOCIATION ANALYSIS

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$
- **Prior beliefs:** row and column sums  
 → Background distribution  $P$

$$P(\mathbf{X}) = \prod_{i,j} P_{i,j}(x_{ij}) \qquad P_{i,j}(x_{ij}) = \frac{\exp(x_{ij} \cdot (\mu_i + \lambda_j))}{1 + \exp(\mu_i + \lambda_j)}$$

	Beer	Diapers	Lipstick	Carrier	SUM
Alice	1	1		1	3
Bob	1		1	1	3
Charlie	1	1			2
Denise	1			1	2
Eve			1	1	2
Frankie		1		1	2
SUM	4	3	2	5	

# ASSOCIATION ANALYSIS

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$
- **Prior beliefs:** row and column sums  
→ Background distribution  $P$
- **Patterns:** *tiles*

	Beer	Diapers	Lipstick	Carrier	SUM
Alice	1	1		1	3
Bob	1		1	1	3
Charlie	1	1			2
Denise	1			1	2
Eve			1	1	2
Frankie		1		1	2
SUM	4	3	2	5	

# ASSOCIATION ANALYSIS

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$
- **Prior beliefs:** row and column sums  
→ Background distribution  $P$
- **Patterns:** *tiles*

	Beer	Diapers	Lipstick	Carrier	SUM
Alice	1	1		1	3
Bob	1		1	1	3
Charlie	1	1			2
Denise	1			1	2
Eve			1	1	2
Frankie		1		1	2
SUM	4	3	2	5	

# ASSOCIATION ANALYSIS

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$
- **Prior beliefs:** row and column sums  
→ Background distribution  $P$
- **Patterns:** *tiles*

	Beer	Diapers	Lipstick	Carrier	SUM
Alice	1	1		1	3
Bob	1		1	1	3
Charlie	1	1			2
Denise	1			1	2
Eve			1	1	2
Frankie		1		1	2
SUM	4	3	2	5	

# ASSOCIATION ANALYSIS

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$
- **Prior beliefs:** row and column sums  
→ Background distribution  $P$
- **Patterns:** *tiles*

	Beer	Diapers	Lipstick	Carrier	SUM
Alice	1	1		1	3
Bob	1		1	1	3
Charlie	1	1			2
Denise	1			1	2
Eve			1	1	2
Frankie		1		1	2
SUM	4	3	2	5	

# ASSOCIATION ANALYSIS

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$
- **Prior beliefs:** row and column sums  
→ Background distribution  $P$
- **Patterns:** *tiles*

	Beer	Diapers	Lipstick	Carrier	SUM
Alice	1	1		1	3
Bob	1		1	1	3
Charlie	1	1			2
Denise	1			1	2
Eve			1	1	2
Frankie		1		1	2
SUM	4	3	2	5	

# ASSOCIATION ANALYSIS

- **Data:** binary matrix:  $X \in \{0,1\}^{m \times n}$
- **Prior beliefs:** row and column sums  
→ Background distribution  $P$
- **Patterns:** *tiles*

	Beer	Diapers	Lipstick	Carrier	SUM
Alice	1	1		1	3
Bob	1		1	1	3
Charlie	1	1			2
Denise	1			1	2
Eve			1	1	2
Frankie		1		1	2
SUM	4	3	2	5	



# ASSOCIATION ANALYSIS

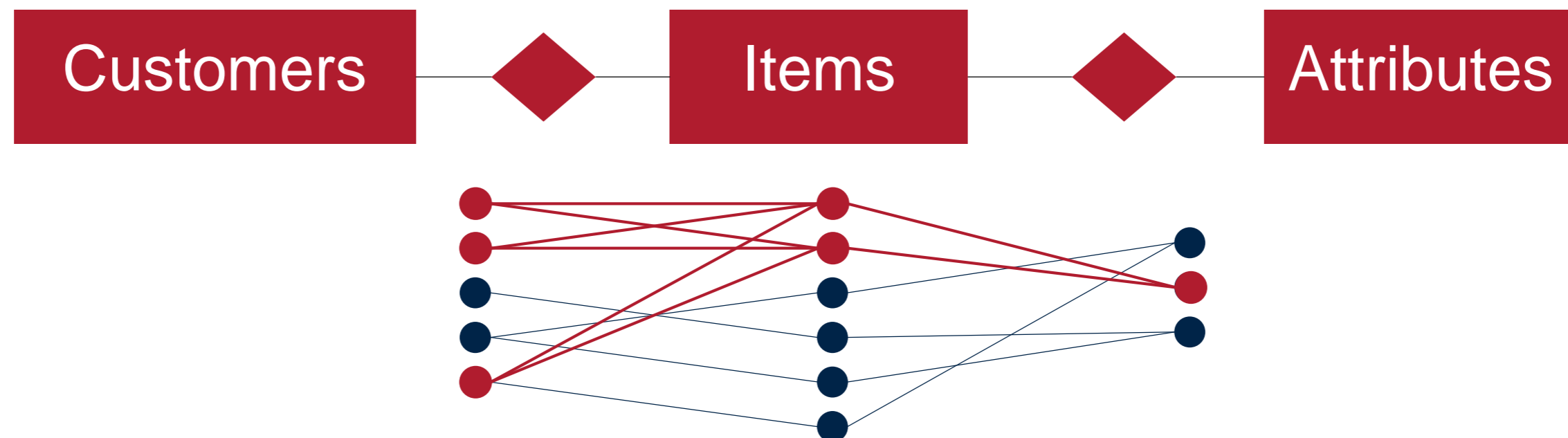
Subjective interestingness ranking		Support x size (area)	
Prior info on: Row & column sums	#docs		#docs
svm, support, machin, vector	25	data, paper	389
state, art	39	algorithm, propose	246
unlabelled, labelled, supervised, learn	10	data, mine	312
associ, rule, mine	36	base, method	202
gene, express	25	result, show	196
frequent, itemset	28	problem	373
large, social, network, graph	15		
column, row	13		
algorithm, order, magnitud, faster	12		
paper, propos, algorithm, real, synthetic, data	27		

Subjective interestingness ranking		Subjective interestingness ranking	
Prior info on: Row & column sums	#docs	Additionally prior info on: Keyword tiles	#docs
svm, <b>support, machin, vector</b>	25	art, state	39
state, art	39	row, column, algorithm	12
unlabelled, labelled, <b>supervised, learn</b>	10	unlabelled, labelled, data	14
<b>associ, rule, mine</b>	36	answer, question	18
<b>gene, express</b>	25	Precis, recal	14

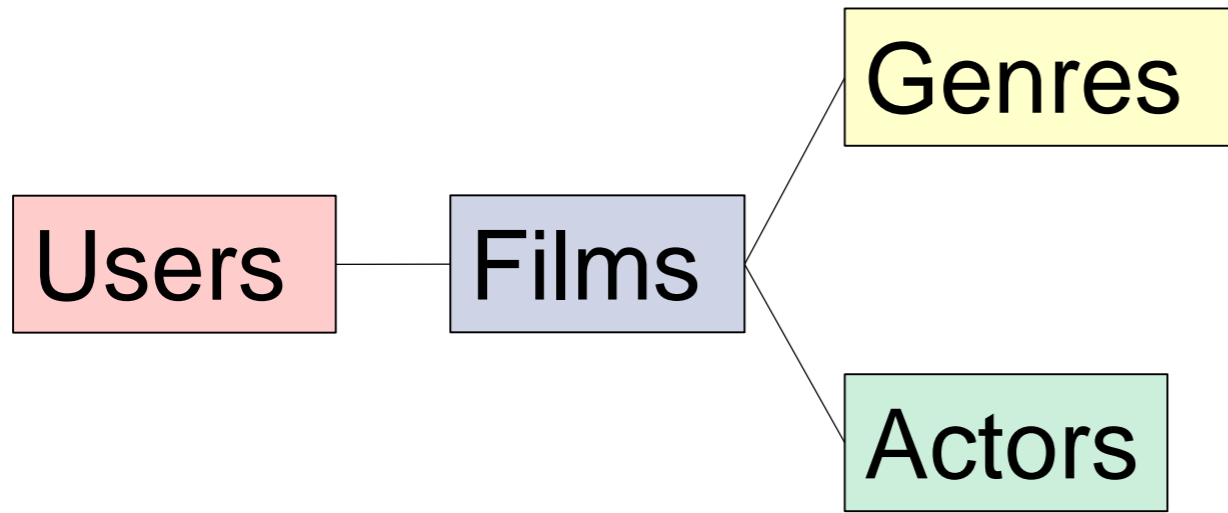
# RELATIONAL PATTERN MINING

with Eirini Spyropoulou, Jeffrey Lijffijt, Mario Boley, Bo Kang, Tias Guns, Achille Aknin

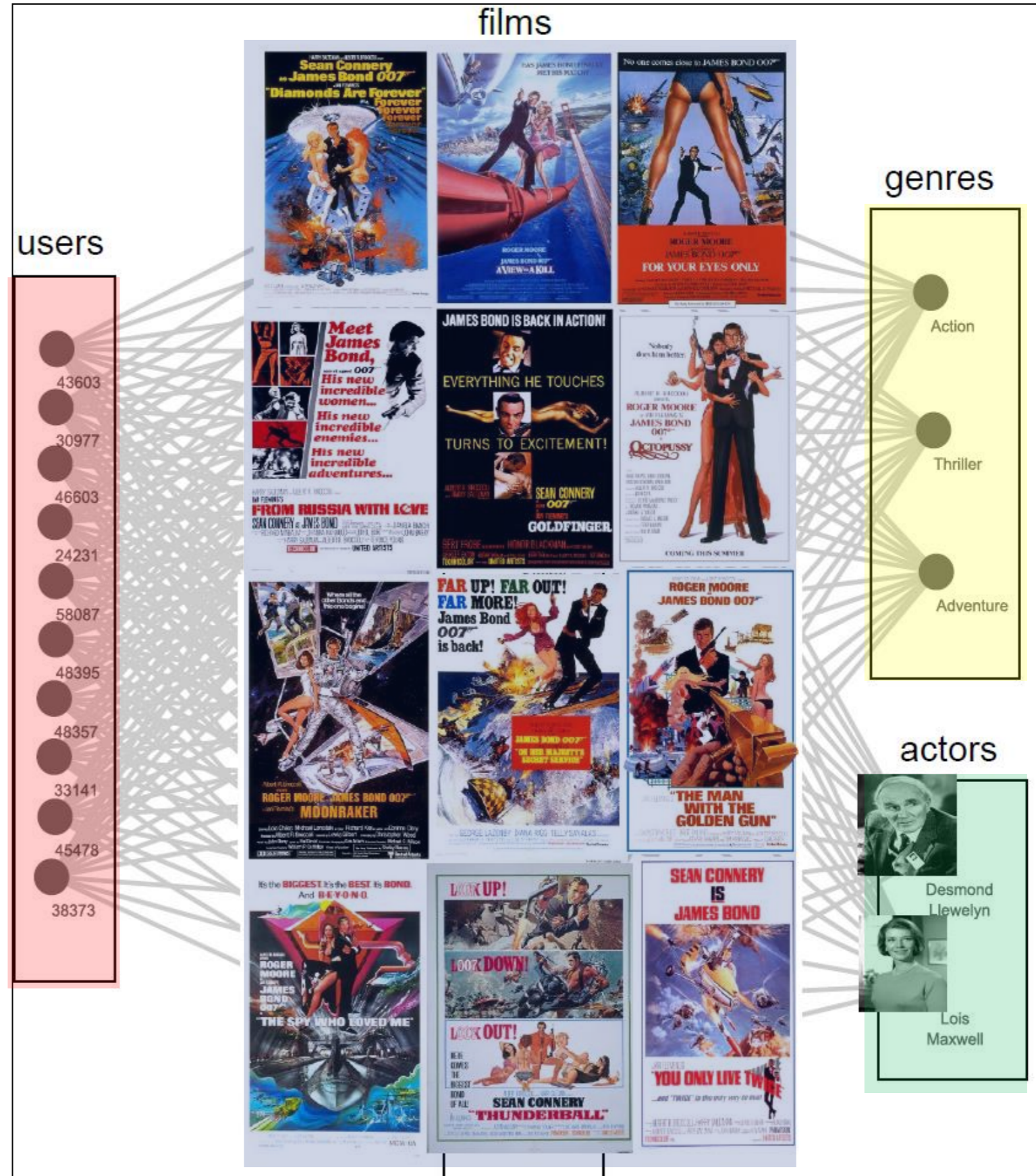
- **Data:** relational database
- **Pattern:** connected complete subgraphs
- **Prior beliefs:** degree of each node in each relationship



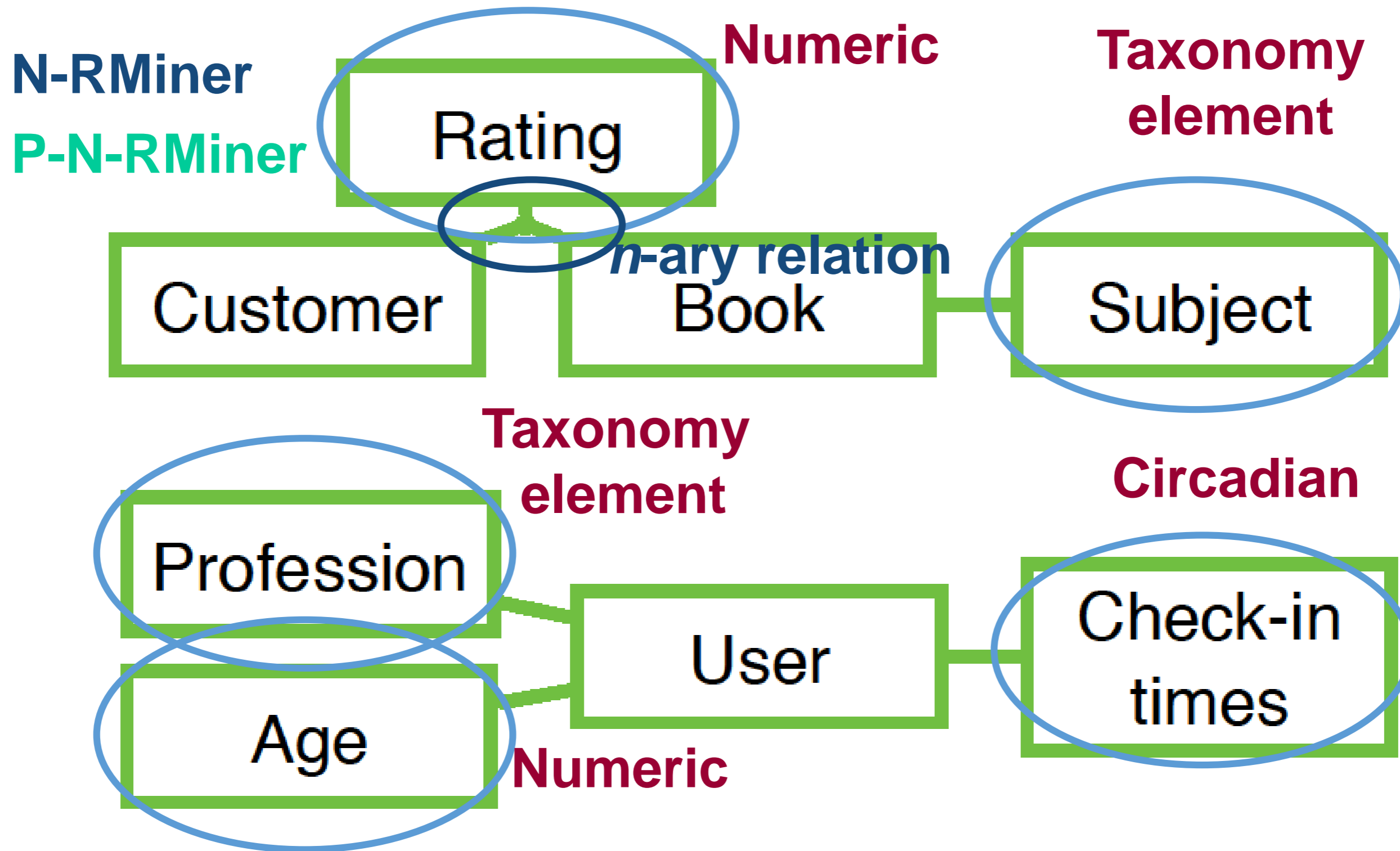
# RELATIONAL PATTERN MINING



RMiner



# RELATIONAL PATTERN MINING



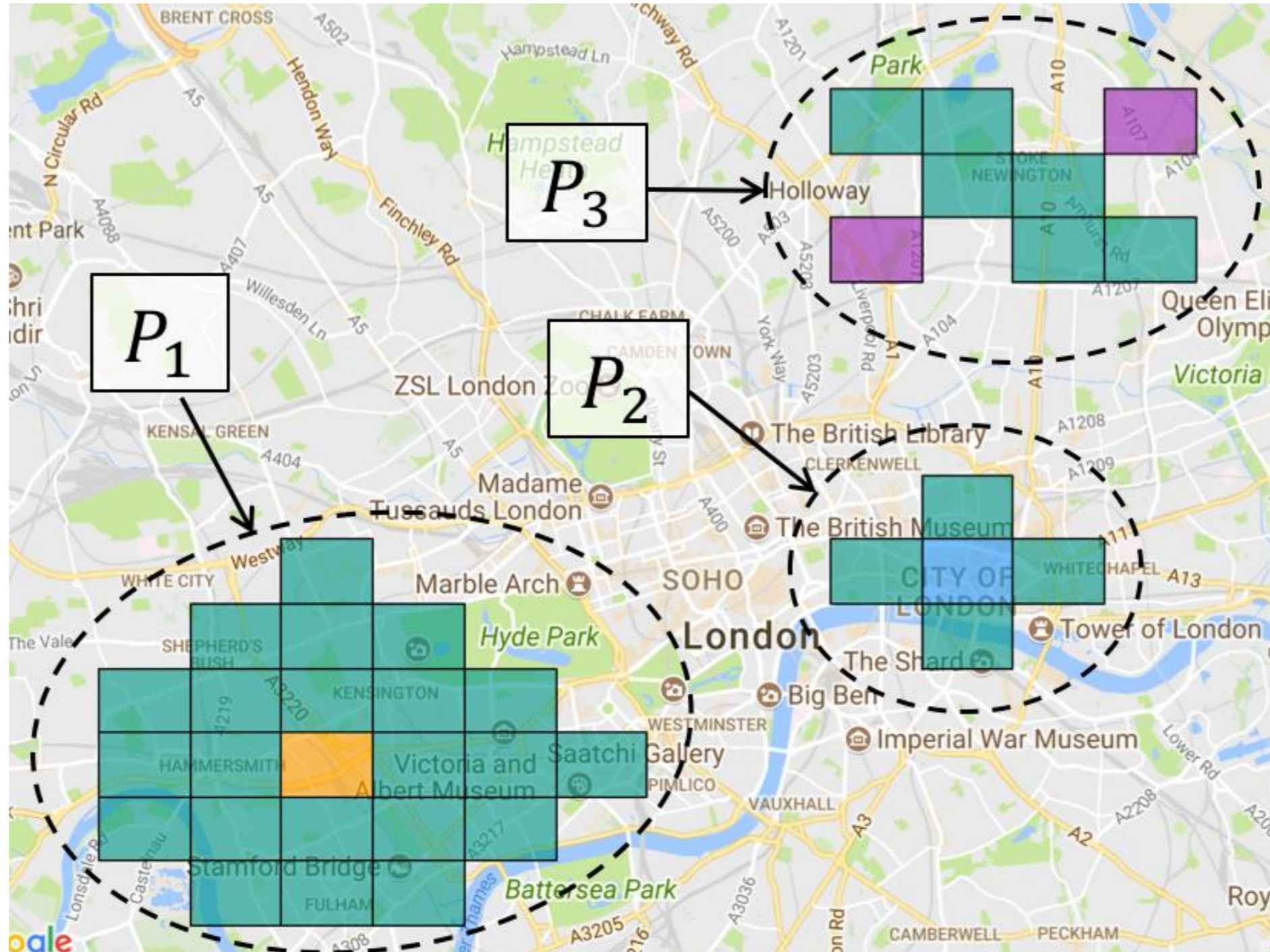
# COHESIVE SUBGRAPHS IN ATTRIBUTED GRAPHS

with Anes Bendimerad, Ahmad Mel, Marc Plantevit, Celine Robardet, Jeffrey Lijffijt

$P_1$ : +food

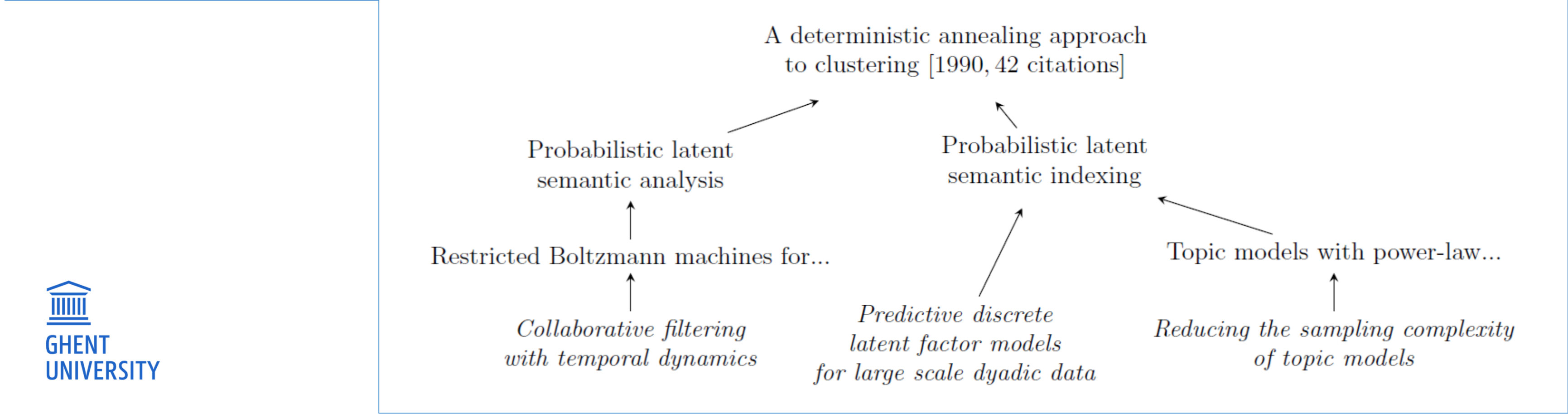
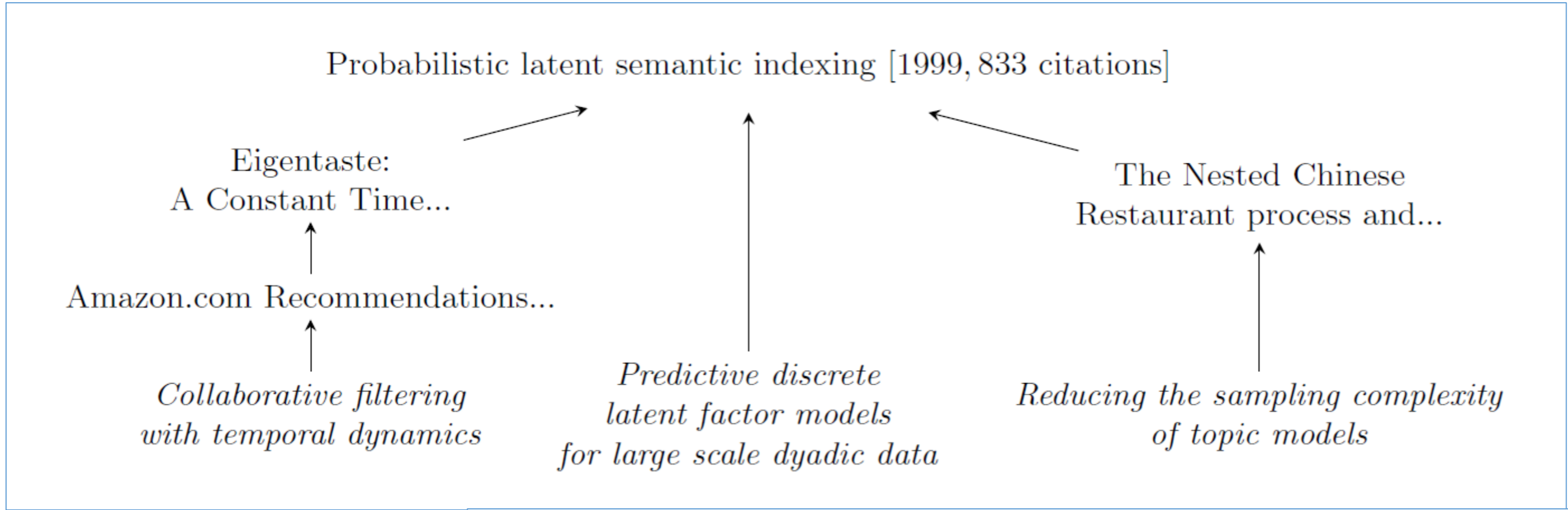
$P_2$ : +professional, +nightlife,  
+outdoors, +college

$P_3$ : +nightlife, +food, -college



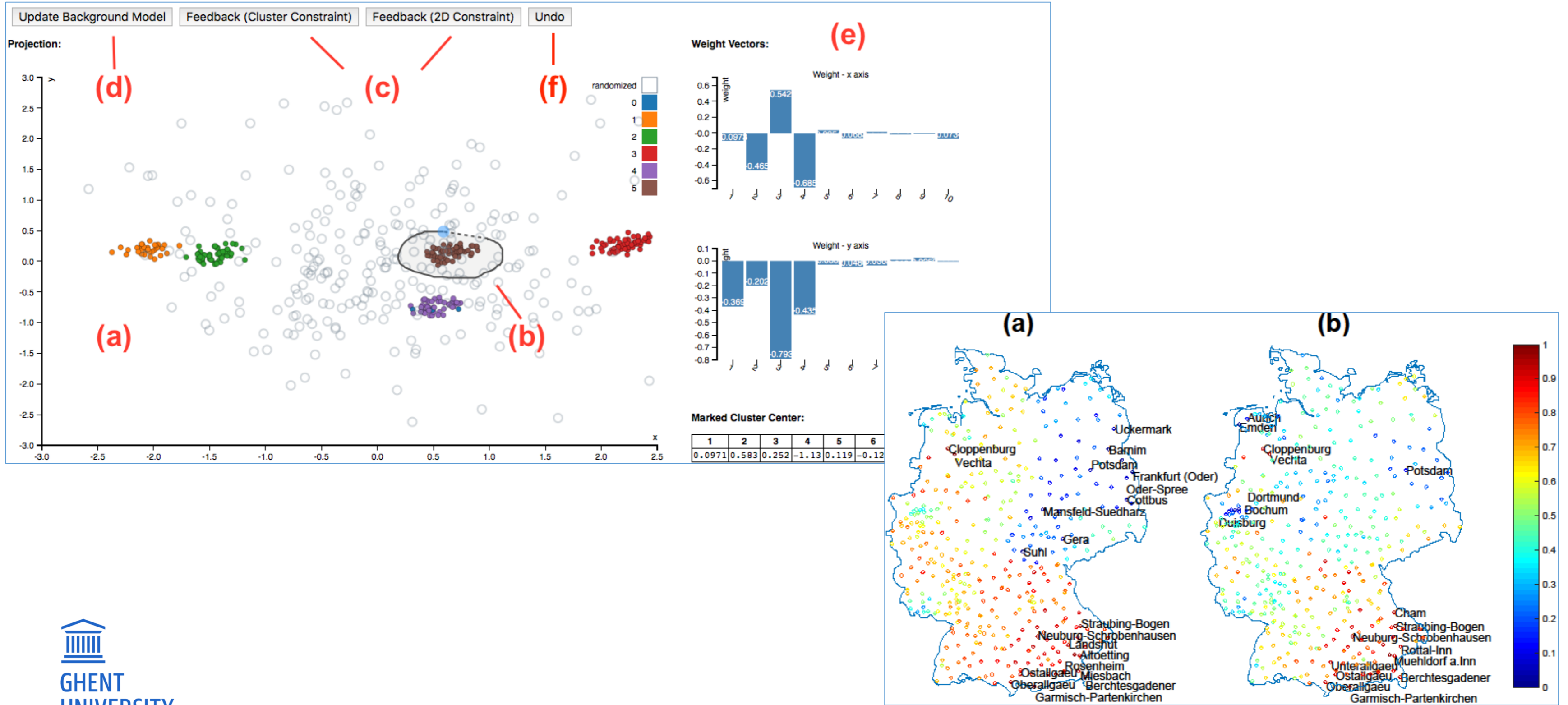
# INTERESTING CONNECTING TREES

with Florian Adriaens, Jeffrey Lijffijt



# DATA PROJECTIONS

with Bo Kang, Jefrey Lijffijt, Kai Puolamaki, Emilia Oikarinen, Raul Santos-Rodriguez

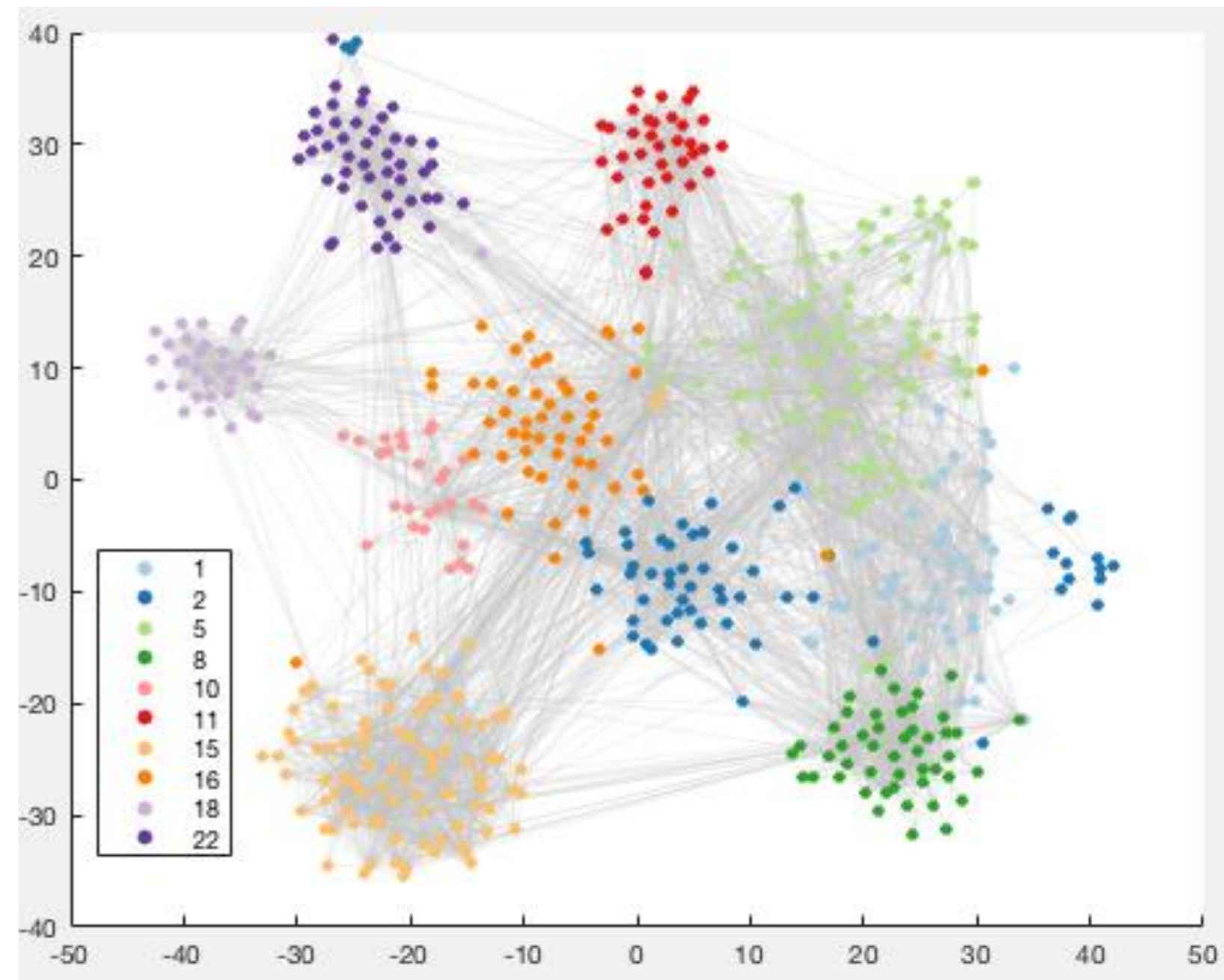


# CONDITIONAL GRAPH EMBEDDINGS

with Bo Kang, Jeffrey Lijffijt

- **Data:** a graph  $G$  w. adj. matrix  $A$
- **Pattern:** a metric embedding  $X$ 
  - Probabilistic info about the graph
  - $P(\|x_i - x_j\| | a_{ij}) = \text{Half-Normal}$
- **Prior beliefs:**  $P_{i,j}(a_{ij})$ 
  - overall density
  - degrees
  - block structure
  - assortativity
  - ...
- **Find ML embedding:**

$$\max_X P(G|X)$$





# AND MORE

with all past and current members of the FORSIED team and Jilles Vreeken, Antonis Matakos, Dario Garcia-Garcia, Siegfried Nijssen,...

## – **Past**

- Data clustering
- Biclustering
- Exceptional model mining / subgroup discovery
- Time series segments

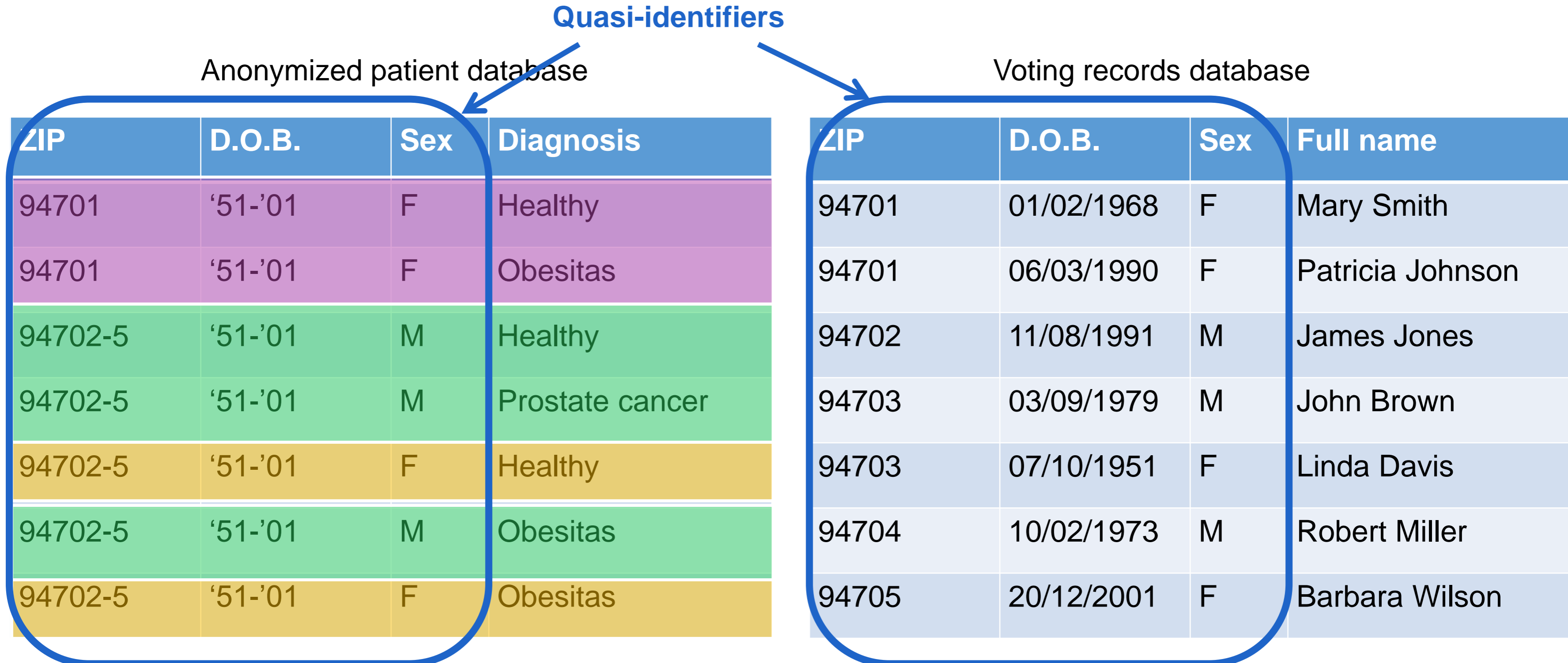
## – **Ongoing / future**

- Backbone of a network
- Insightful ‘generalizations’ of an attributed network
- Conditional t-SNE (a.o. non-linear dimred methods)
- ...

# DATA MINING WITHOUT SPILLING THE BEANS

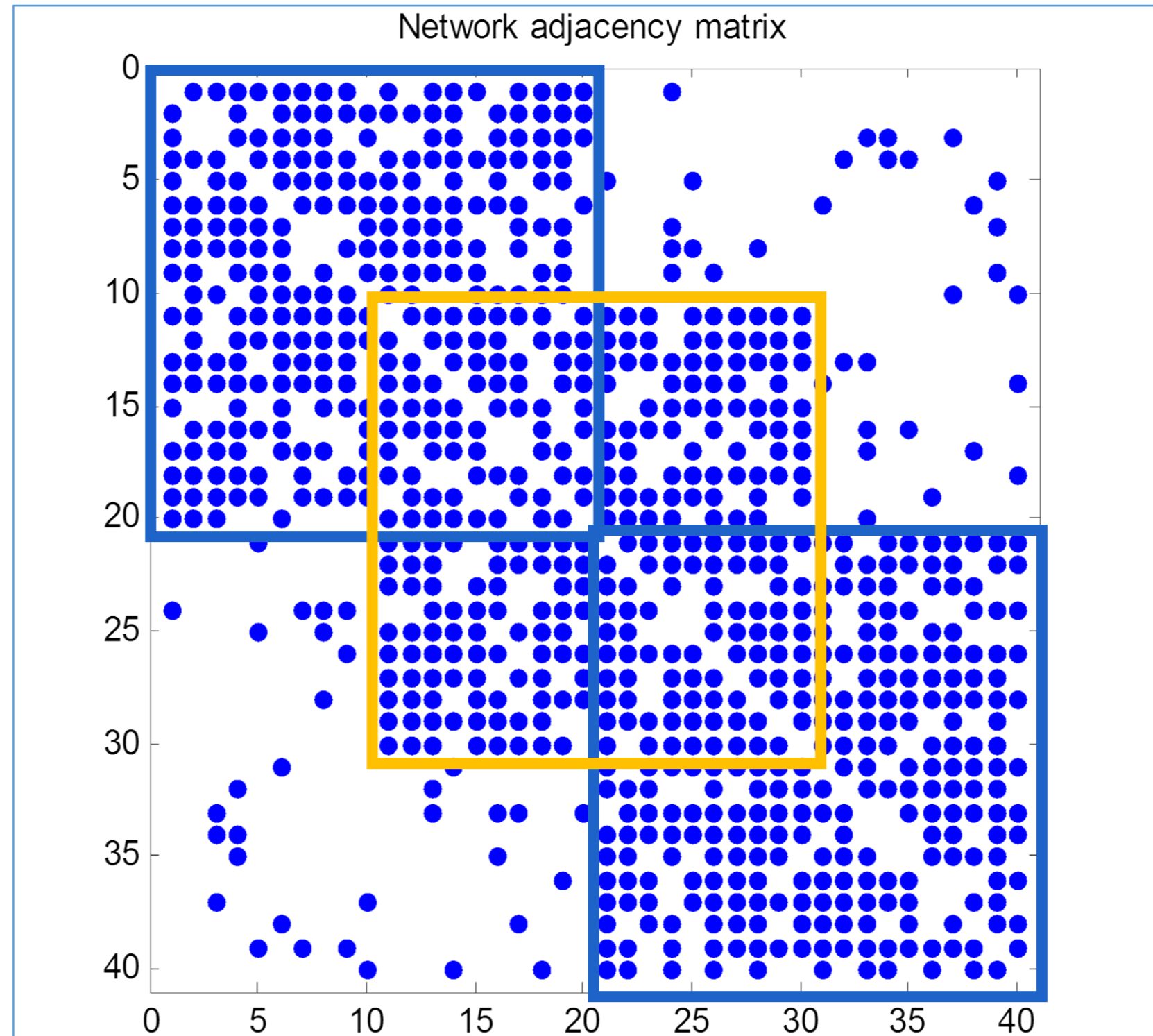
# PRIVACY-PRESERVING DATA PUBLISHING

- **Anonymization** insufficient to protect sensitive attributes (**linkage attack**)
- **Generalization!**



# ... AND MORE

- Existence of a tight community in a network



## ... AND MORE

- Existence of a tight community in a network
- Existence of a cluster in data
- Frequency of particular items / size of particular transactions in a database of purchases

Preserve this while:

- publishing **generalized version of database,**
- identifying **dense subgraphs,**
- finding **clusters,**
- mining **frequent itemsets,** etc

**Data mining patterns**

# GENERAL STRATEGY

– **Data:**  $x$

– Data mining goal: **reveal as much as possible about  $x$**

– **Sensitive aspects:**  $f(x) \in \Phi$

– the sensitive attributes' values

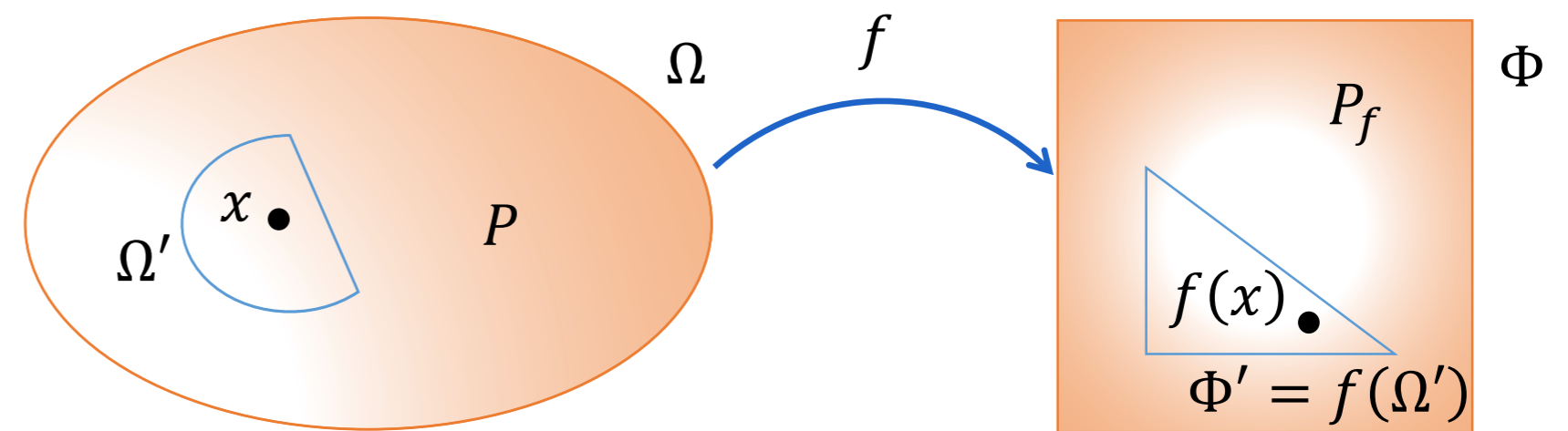
– density of a specified subgraph

– existence of a tight cluster

– frequencies of all items

– Goal: **reveal as little as possible about  $f(x)$**

(data)  $x \rightarrow f(x)$  (sensitive aspects)



– **Updating**  $P \rightarrow P'$  results in updating  $P_f \rightarrow P'_f$

– More complex than conditioning!

–  $P_f(f(x))$  can be larger or smaller than  $P'_f(f(x))$

# TRADING-OFF TWO THINGS

1. **Subjective information content of a pattern**
2. **A criterion on the background distribution about sensitive aspects:**
  - **Information content left in sensitive aspects**  
(surprise in actual value of the sensitive attributes):
$$-\log \left( P'_f(f(\mathbf{x})) \right)$$
  - Entropy of  $P'_f$  (**uncertainty about sensitive attributes**):
$$-E_{\mathbf{x} \sim P'_f} \left\{ \log \left( P'_f(f(\mathbf{x})) \right) \right\}$$
  - **Knowledge gained** about actual value of the sensitive aspects:
$$-\log \left( \frac{P_f(f(\mathbf{x}))}{P'_f(f(\mathbf{x}))} \right)$$
  - Degree of belief that the sensitive aspects are within a specified set  $\Phi^* \subseteq \Phi$ :
$$P'_f(\Phi^*)$$

# EXAMPLES

---



# PRIVACY-PRESERVING DATA PUBLISHING

- **Random synthetic dataset:**

- 5 real-valued quasi-identifiers, generalization through intervals
- 1 sensitive attribute, 3 possible values
- 1 other attribute, 3 possible values
- 100 data records

E.g. • Zip code  
• DOB

E.g. • Sexual orientation  
• Ethnicity

E.g. • Sense of well-being  
• Productivity

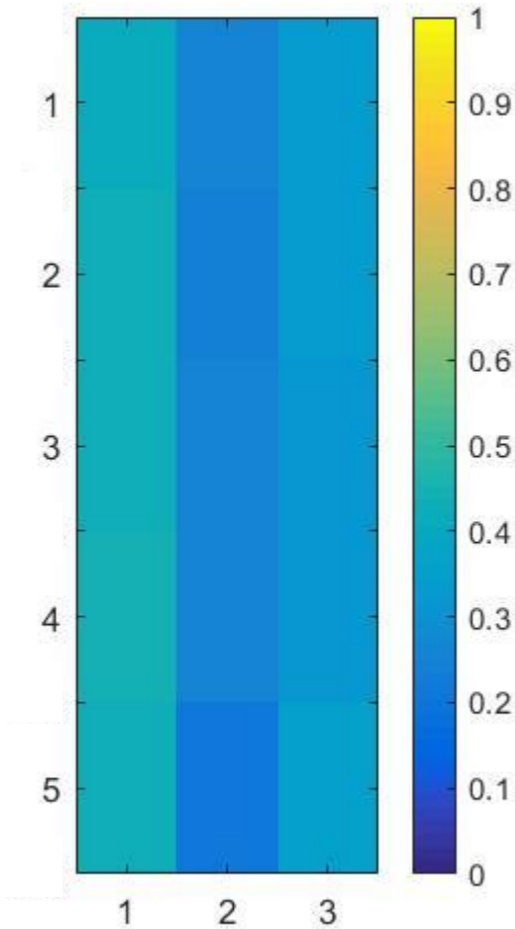
- **Trade-off:**

- information about **data** (other & sensitive attributes)
- knowledge gained about **sensitive attribute**

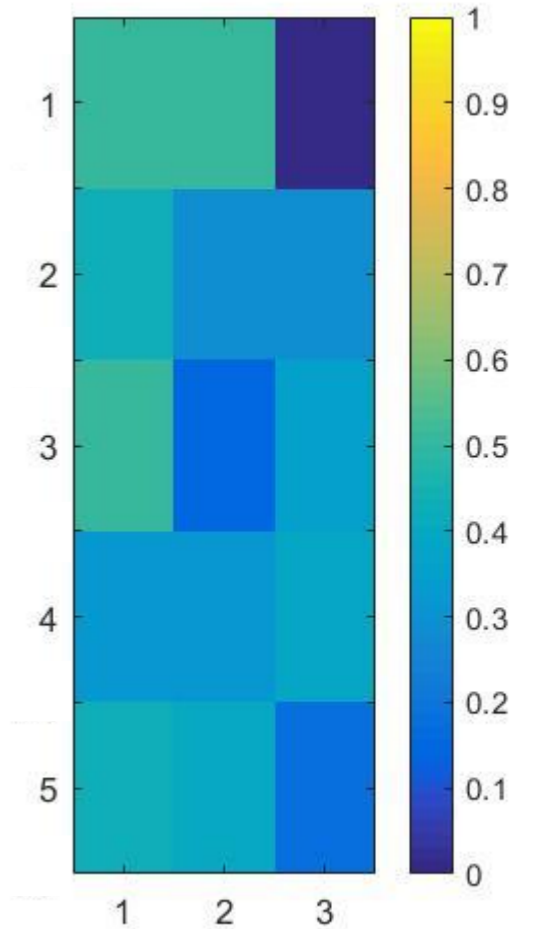
- **Generalize quasi-attributes** → 5 equivalence classes

- Ensure the maximum information content about any sensitive attribute value is small

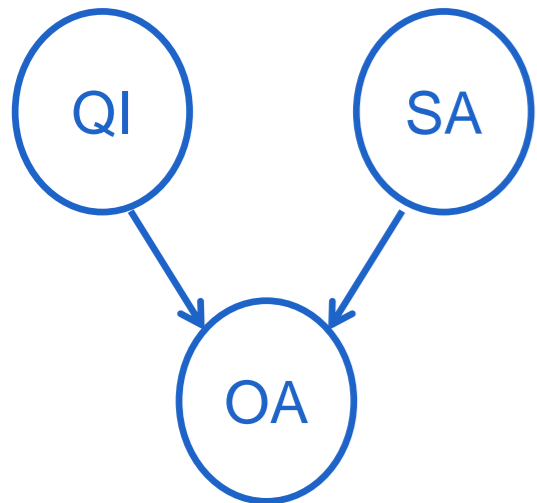
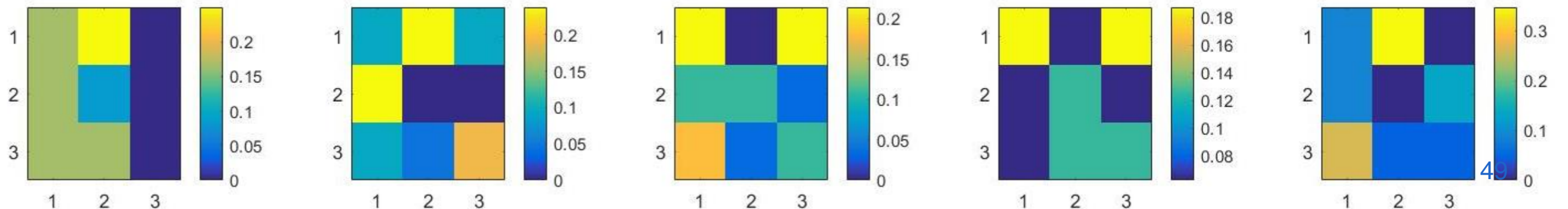
Conditional distributions within the 5 equivalence classes over the 3 **sensitive** attribute values



Conditional distributions within the 5 equivalence classes over the 3 **other** attribute values



Joint conditional distribution of the **sensitive** (rows) and **other** (columns) attributes, within 5 equivalence classes



# DENSE SUBGRAPHS WITHOUT SPILLING BEANS

- **Random network:**

- 2 non-overlapping communities
- A 3<sup>rd</sup> community overlapping both

- The 3<sup>rd</sup> is **sensitive**

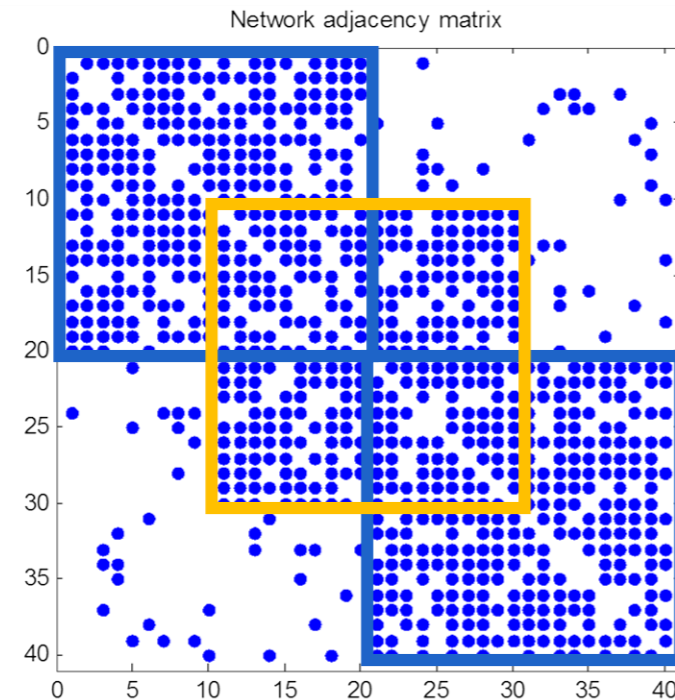
- Analyst should remain surprised by its presence

- **Task:**

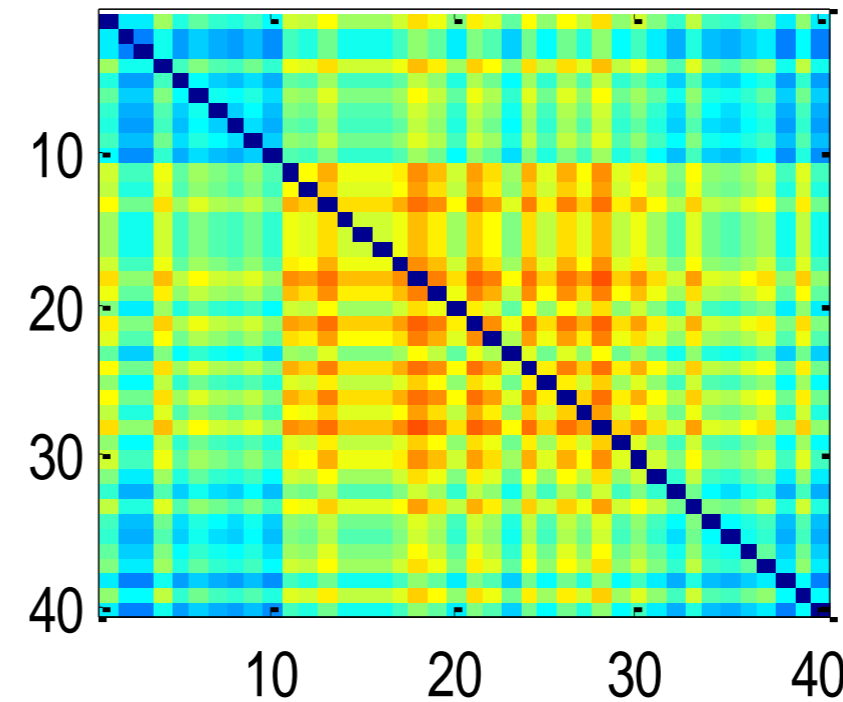
- Identify (non-)dense subgraphs
- Without spilling the beans on the 3<sup>rd</sup> community

- **Approaches** (result from general strategy):

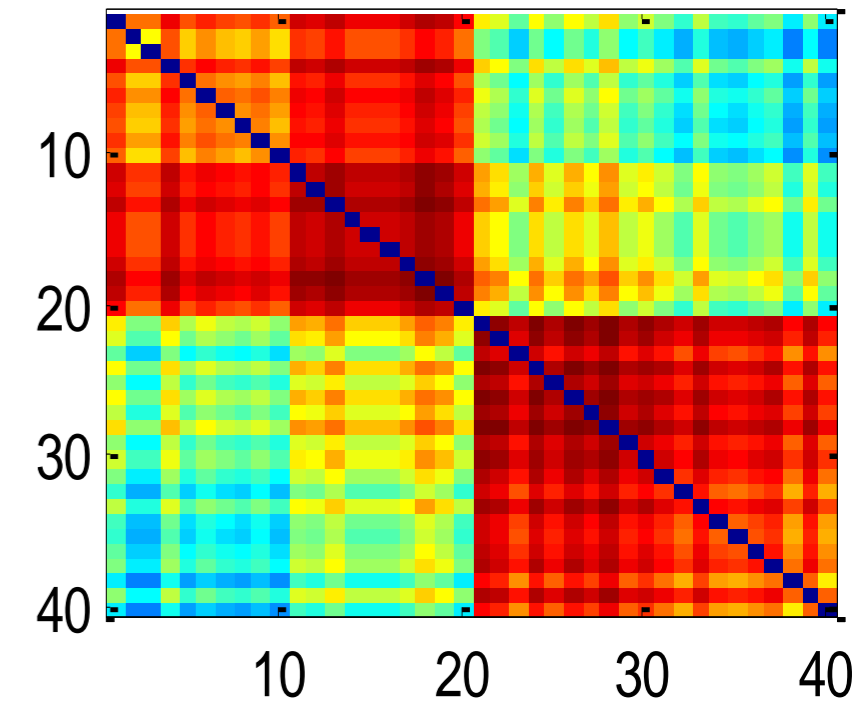
- Deceive
- Conceal



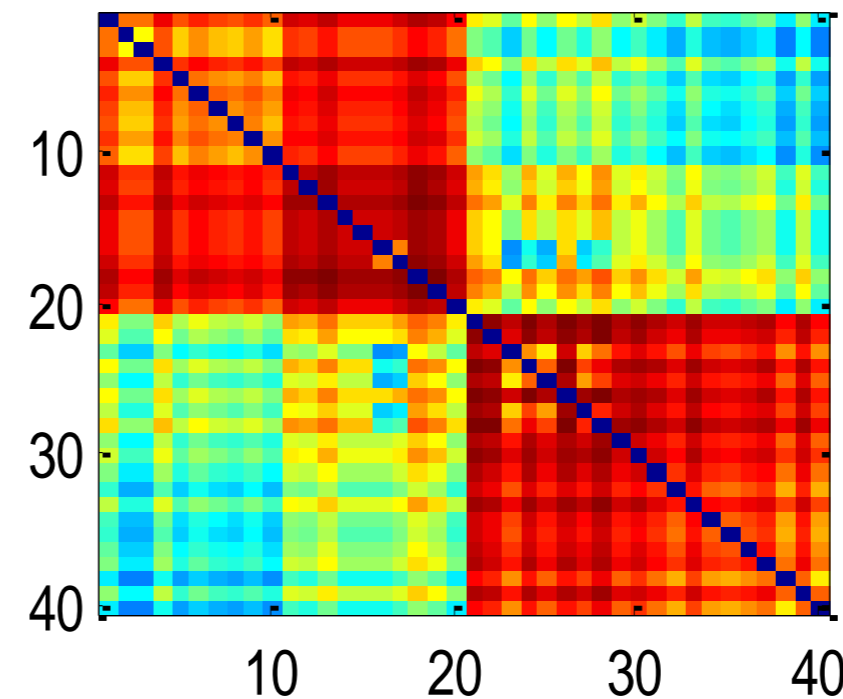
Initially



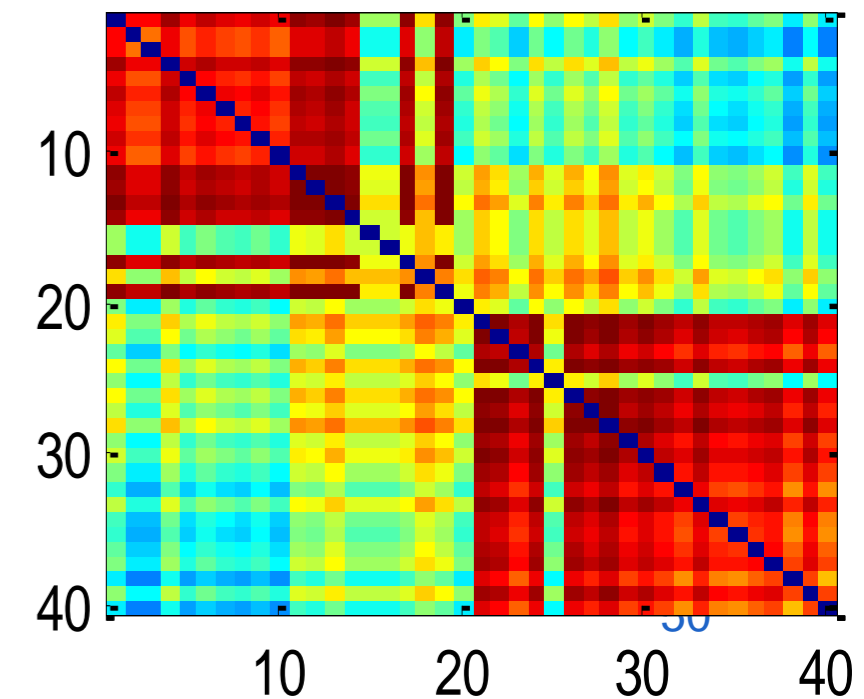
After both community patterns



After both community patterns and a deception pattern



After both community patterns partially concealed



# OUTLOOK

# FURTHER WORK

- **Improve theoretical understanding**
  - Estimating the background distribution (information geometry)
  - Cognitive aspects (cognitive science)
  - User interface (human computer interaction)
  - Visualization (visual analytics)
  - Algorithmic aspects (optimisation theory)
  - Safeguarding sensitive information & fairness
- **More instantiations**
  - Data types (Linked Data!) / pattern types / prior belief types.
- **Applications**
  - Bioinformatics
  - Web and social media mining

**“Data Mining without Spilling the Beans: Preserving more than Privacy alone”**

Project funded by the FWO (with Jeffrey Lijffijt as co-investigator)



**“Exploring Data: Theoretical Foundations and Applications to Web, Multimedia, and Omics Data”**

Odysseus project funded by the FWO

**“Formalizing Subjective Interestingness in Data mining”**

ERC project FORSIED



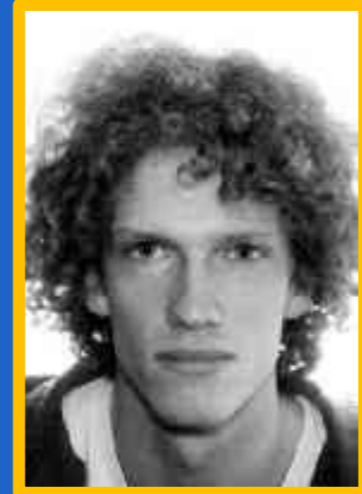
Ahmad Mel



Bo Kang



Florian Adriaens



Paolo Simeone



Alexandru Mara



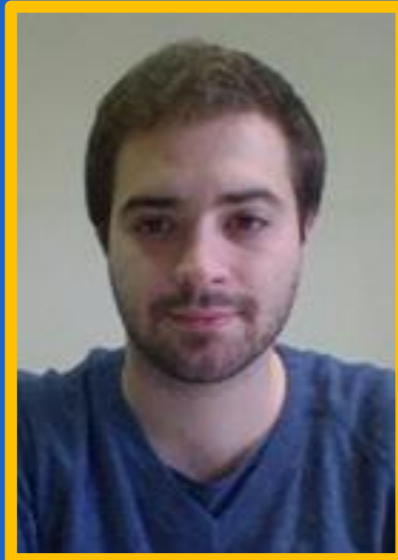
Holly Silk



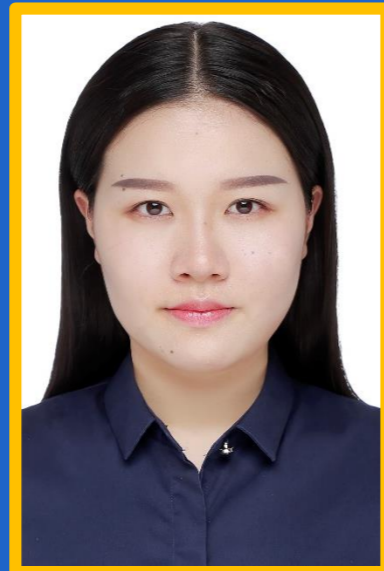
Eirini Spyropoulou



Robin Vandaele



Xi Chen



Jefrey Lijffijt



Junning (Lemon) Deng



Raul Santos-Rodriguez



Akis Kontonasios



Tijl De Bie



Wouter Duivesteijn



Achille Akinin



**We are recruiting!**