FACULTY OF ENGINEERING
AND ARCHITECTURE

DEPARTMENT ELECTRONICS AND INFORMATION SYSTEMS (ELIS)
RESEARCH GROUP IDLAB

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)

GHENT UNIVERSITY

European Research Council
Established by the European Commission

fwo

SUBJECTIVITY = KEY

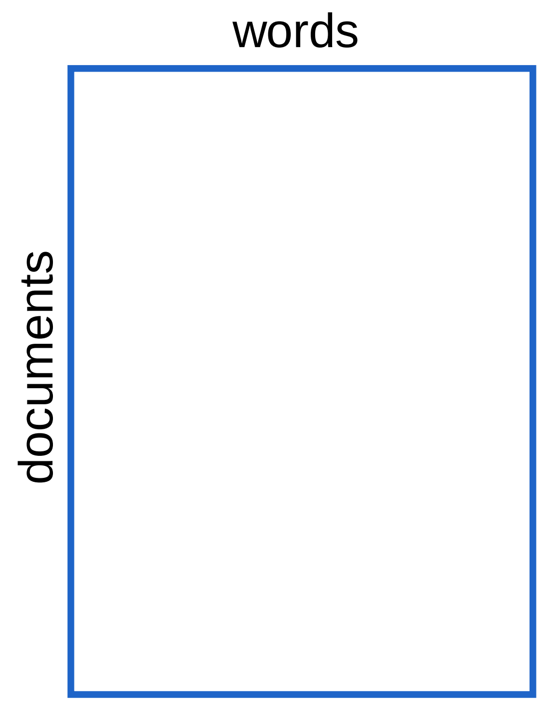
Three motivating examples:

1. **Frequent itemset mining**
   – Support = bad
   – Support x itemset size (= ‘surface area’) = a bit better
   – ...
2. **Graph embedding**
   – Where do high degree nodes go?
3. **Privacy-preserving data publishing (& mining)**
   – The trouble of background knowledge attacks

ASSOCIATION ANALYSIS / ITEMSET MINING

Subjective interestingness ranking Prior info on: Row & column sums	#docs	Support x size (area)	#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

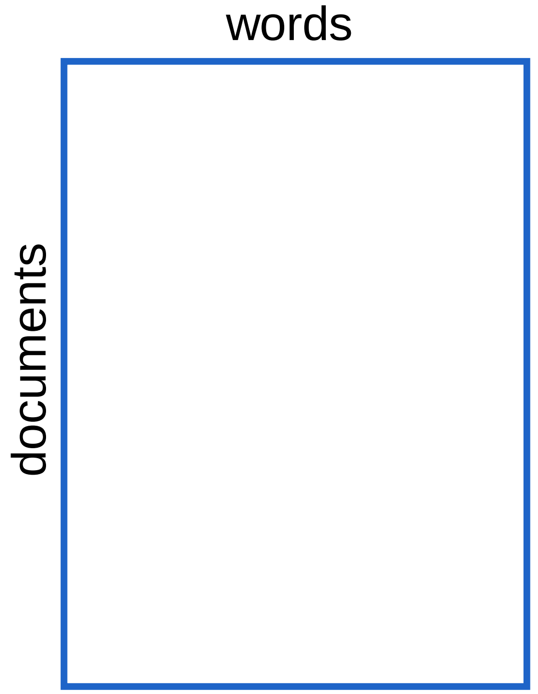
NIPS abstracts dataset:



GHENT UNIVERSITY

ASSOCIATION ANALYSIS / ITEMSET MINING

NIPS abstracts dataset:

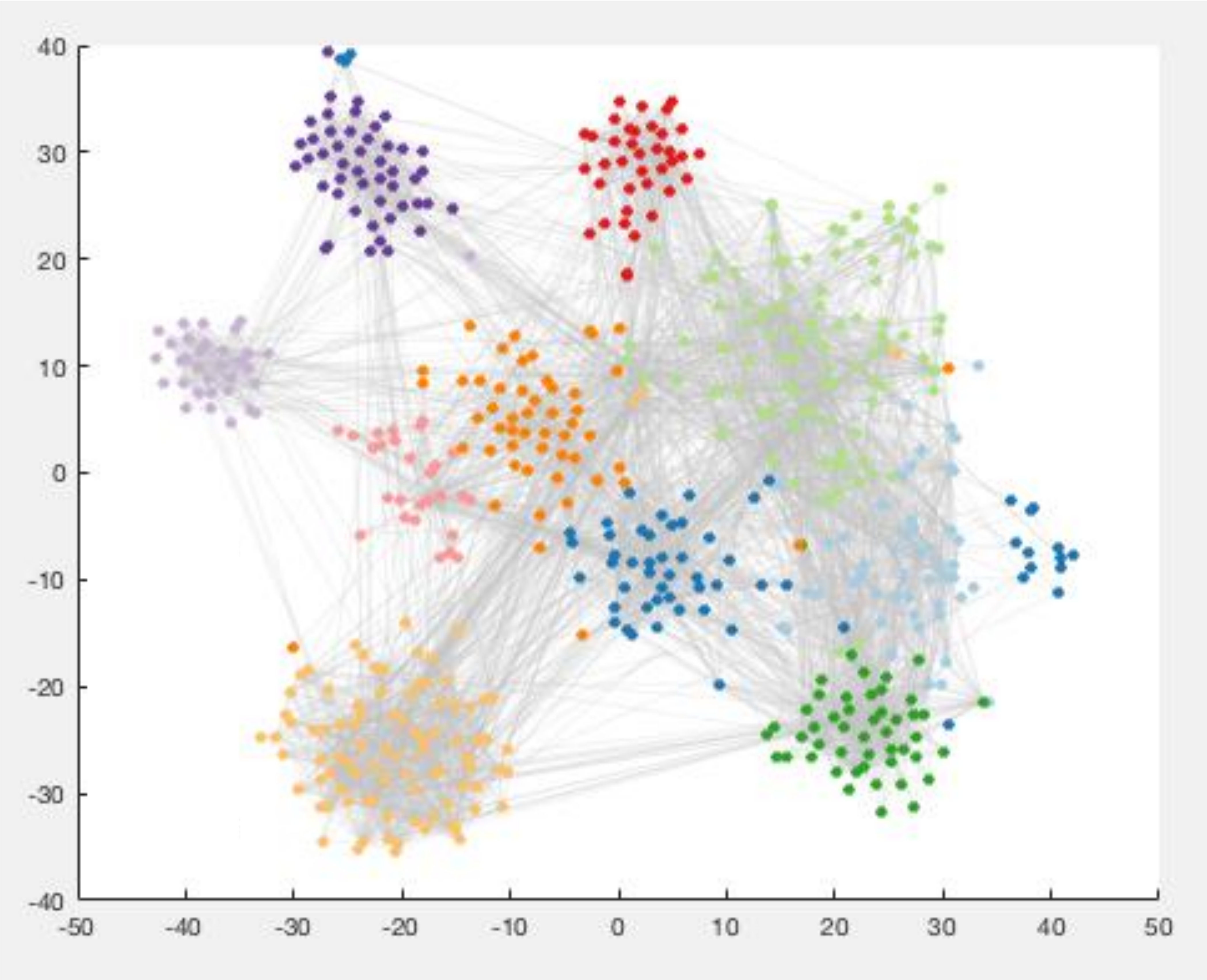


Subjective interestingness ranking **Prior info on:** **Row & column sums**	**#docs**	**Support x size (area)**	**#docs**
svm**,** support, machin, vector	25	data, paper	389
state, art	39	algorithm, propose	246
unlabelled, labelled, supervised, learn	10	data, mine	312
associ, rule, mine	36	base, method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

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

GHENT UNIVERSITY

CONDITIONAL GRAPH EMBEDDINGS



GHENT UNIVERSITY

PRIVACY-PRESERVING DATA PUBLISHING

– Prevent linking identity with values of a sensitive attribute

Quasi-identifiers

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

– Prevent linking identity with values of a sensitive attribute

Quasi-identifiers

Anonymized patient database

ZIP	D.O.B.	Sex	Diagnosis
94701	‘51-’01	F	Healthy
94701	‘51-’01	F	Obesitas
94702-5	‘51-’01	M	Healthy
94702-5	‘51-’01	M	Prostate cancer
94702-5	‘51-’01	F	Healthy
94702-5	‘51-’01	M	Obesitas
94702-5	‘51-’01	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

EXPLORING DATA

- The search for interesting *patterns* in *data*
  - 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,…
  - Association analysis — Frequency, lift, confidence, leverage, coverage,...
  - Graph embedding — Node2Vec, Path2Vec, MetaPath2Vec,...
  - Sanitized data publishing — Discernibility, generalization height, average group size,...
  - …
- Zillions of
  - Objective functions
  - Quality functions
  - Utility functions
  - Cost functions
  - …

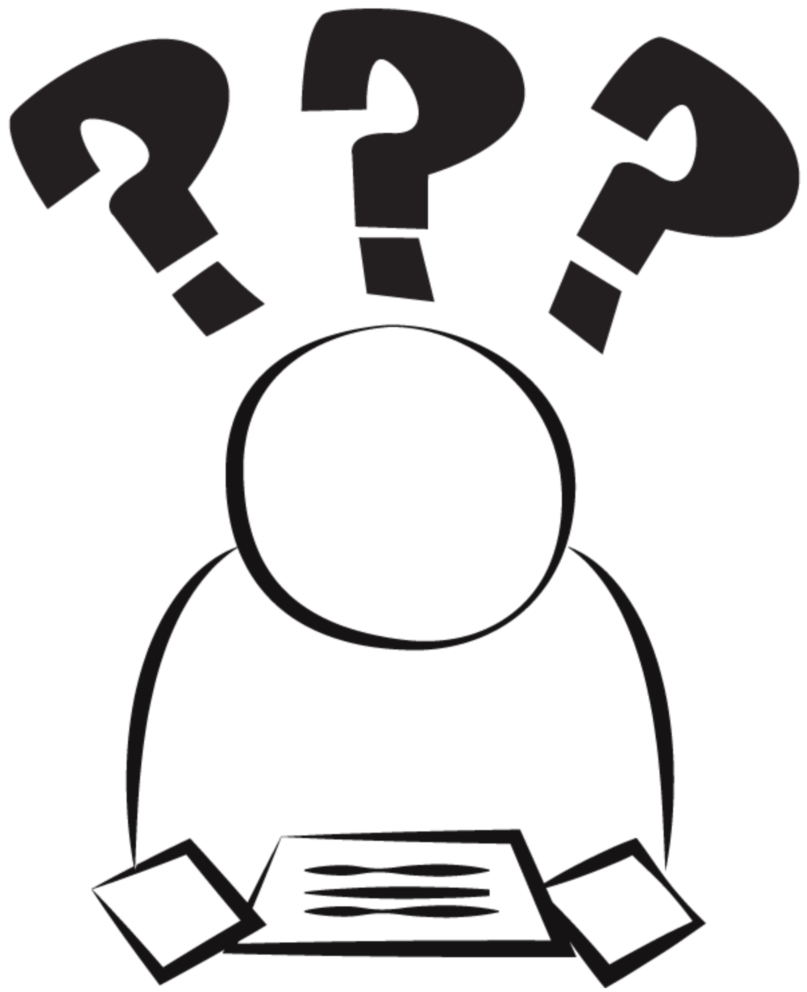
‘Interestingness measures’

GHENT UNIVERSITY

THE CHALLENGE

- Zillions of interestingness measures = good & bad
  - Good: more options!
  - Bad: the trees & the forest…

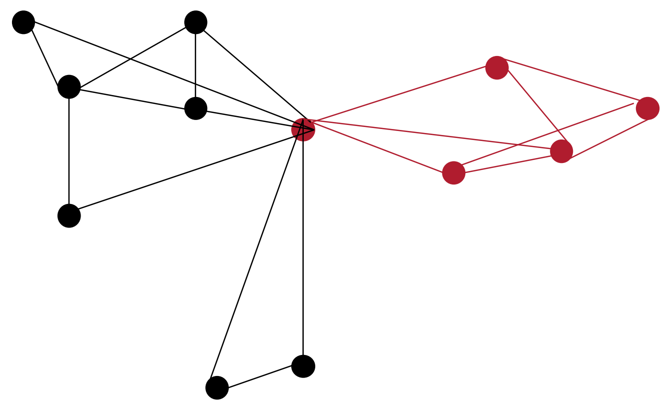
- Challenge:
  - **Formalise *true* interestingness!**
    - With minimal user interaction
    - Without requiring user expertise



GHENT UNIVERSITY

MOTIVATING EXAMPLE

- Community detection:



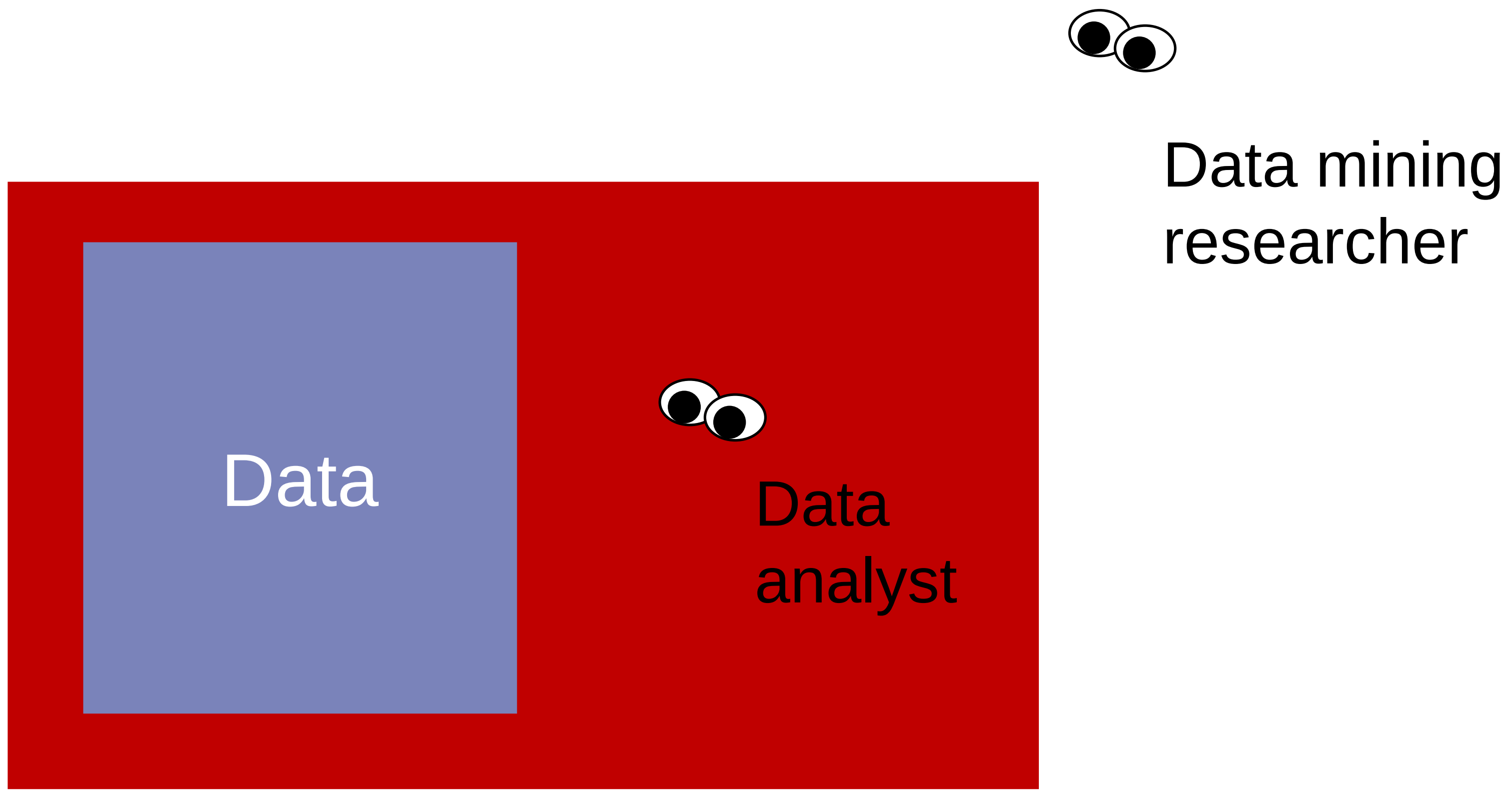
- What makes for an interesting community?
  - Densely connected?
  - Large?
  - Few neighbours outside community?
  - Unrelated to certain known ‘affiliations’?
  - …

GHENT UNIVERSITY

THE FORSIED APPROACH

Interestingness(pattern)

THE FORSIED APPROACH



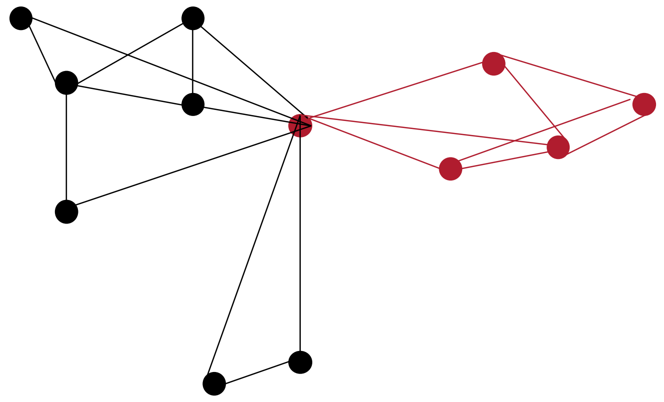
Interestingness(pattern) → Interestingness(pattern, analyst)

Interestingness = subjective

GHENT
UNIVERSITY

MOTIVATING EXAMPLE

- Community detection:

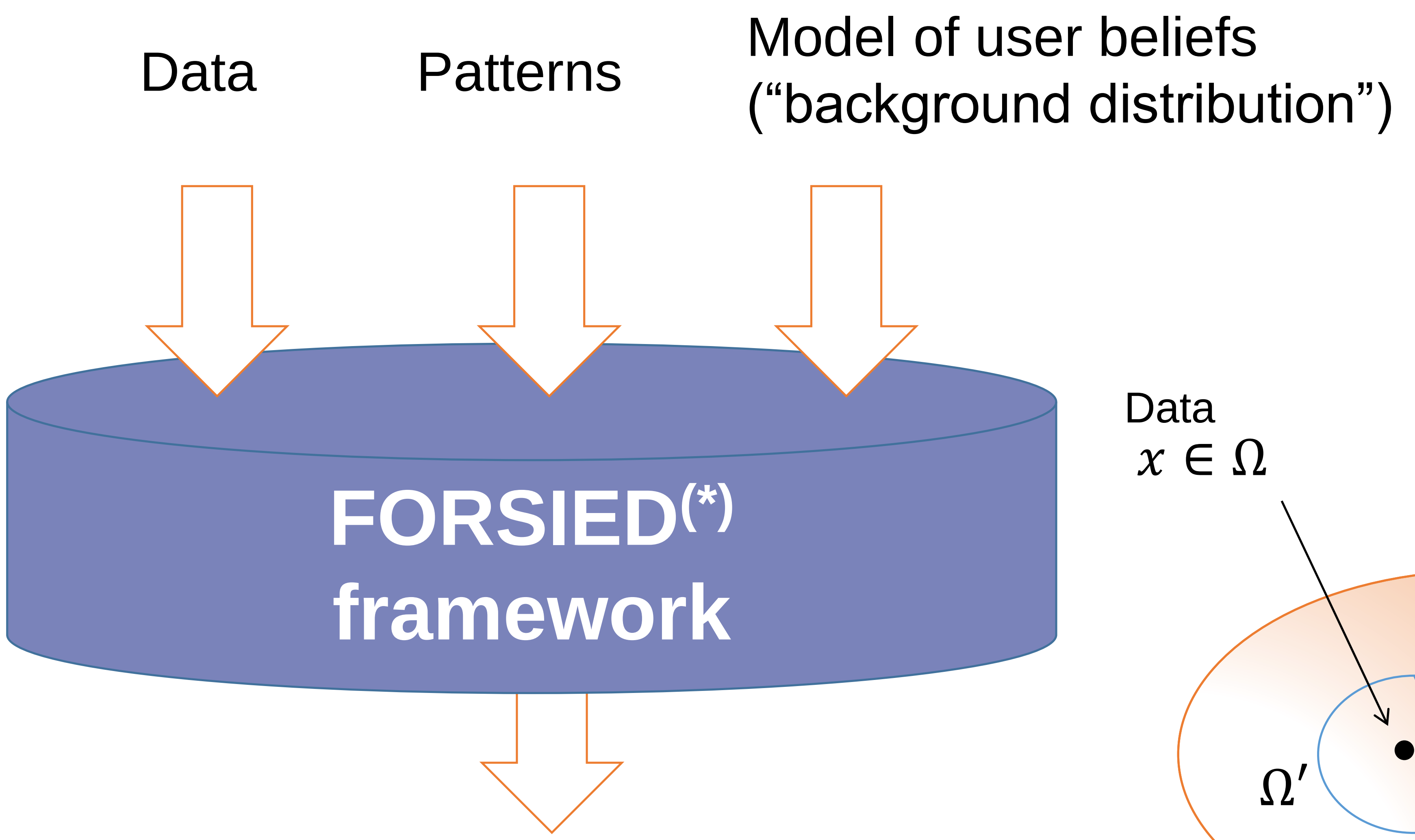


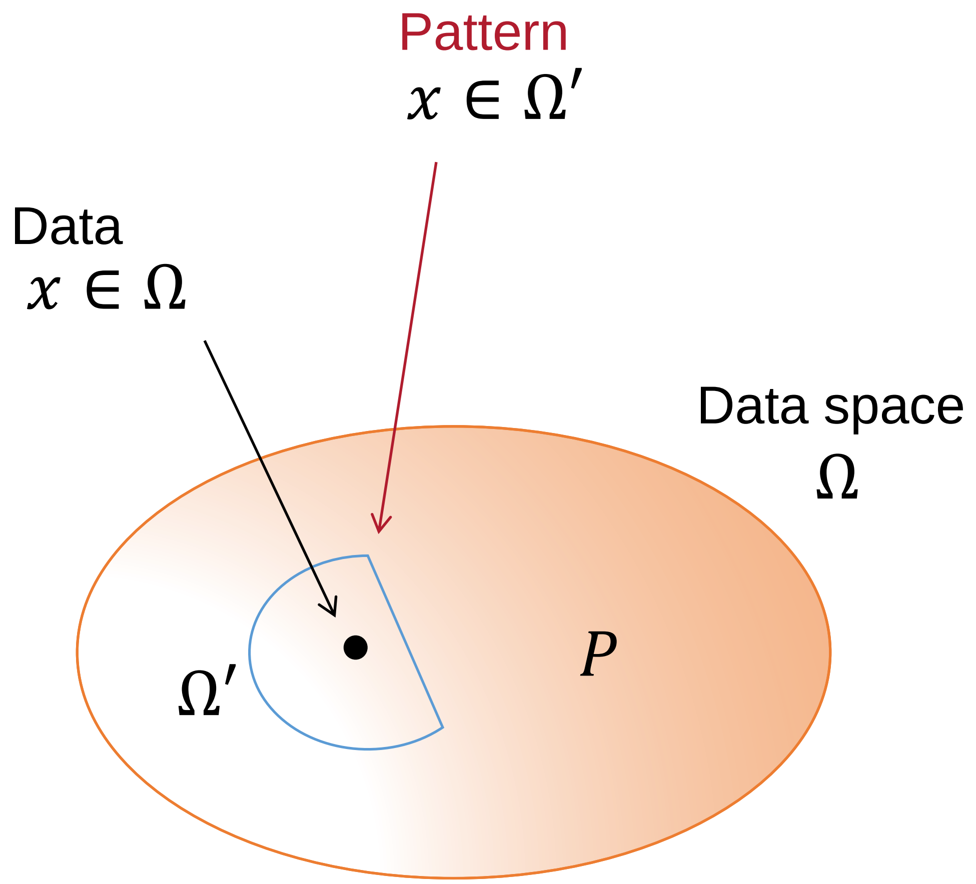
- User states expectations / beliefs
  - Formalized as a ‘background distribution’
- Any ‘pattern’ that contrasts with this and is easy to describe

  = *subjectively interesting*

GHENT UNIVERSITY

AN INFORMATION THEORETIC APPROACH





$$\boxed{\text{InformationContent}(\Omega', P) = -\log(P(\Omega'))}$$

Pattern User

Subjective!

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

GHENT UNIVERSITY

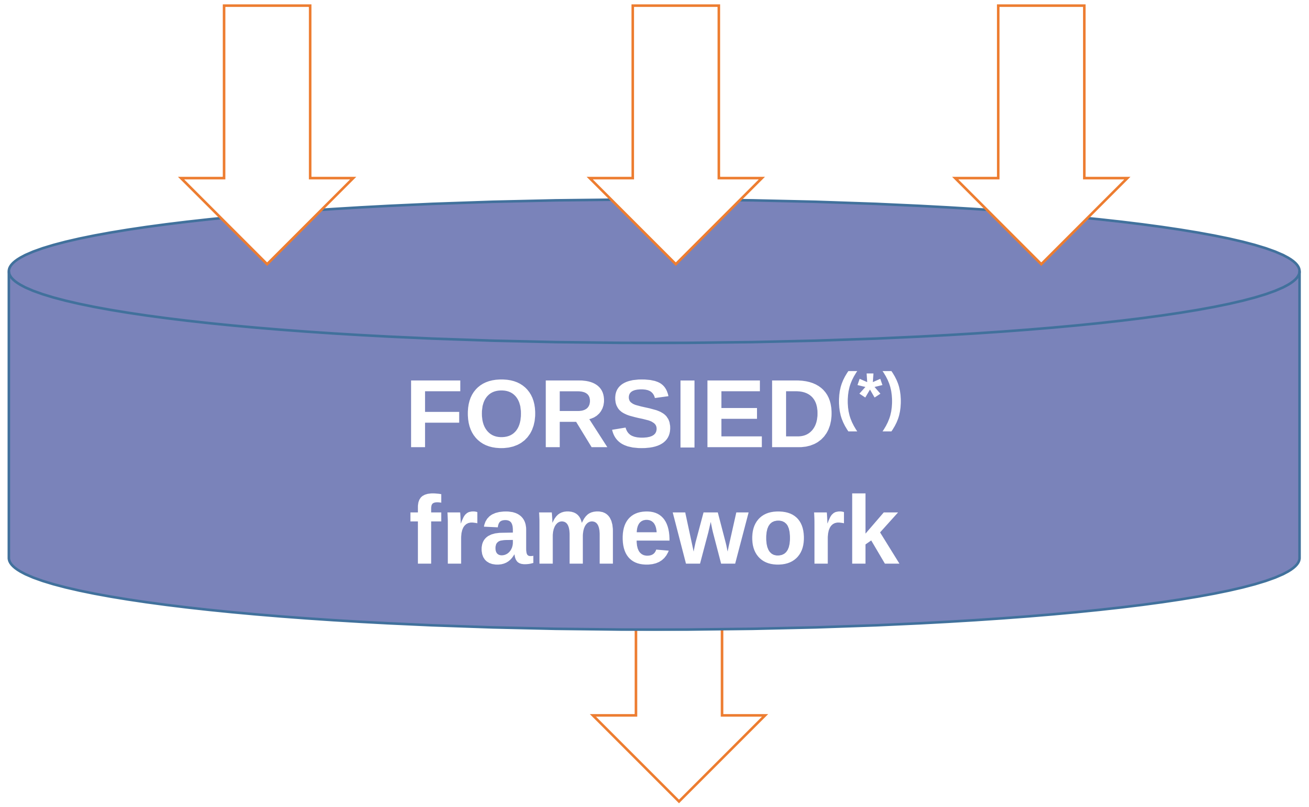
(*)Formalizing Subjective Interestingness in Exploratory Data mining

P evolves!

Data

Patterns

Model of user beliefs
("background distribution")

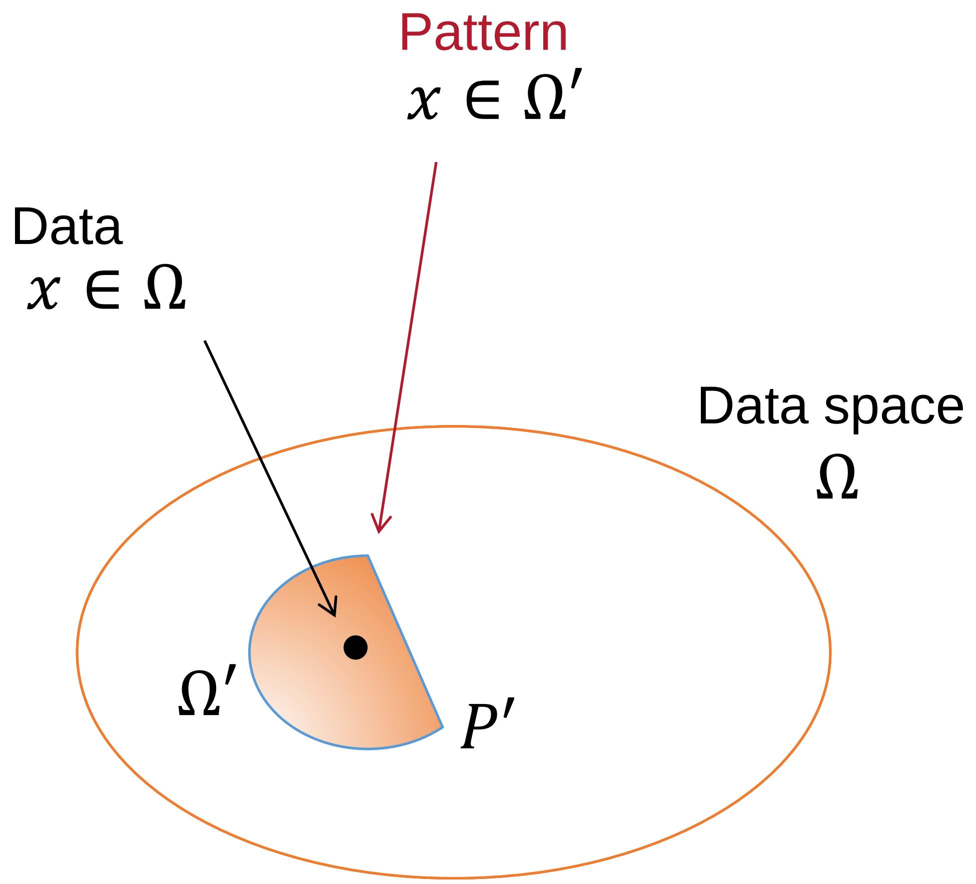


$$\text{InformationContent}(\Omega', P) = -\log(P(\Omega'))$$

Pattern

User

Subjective!

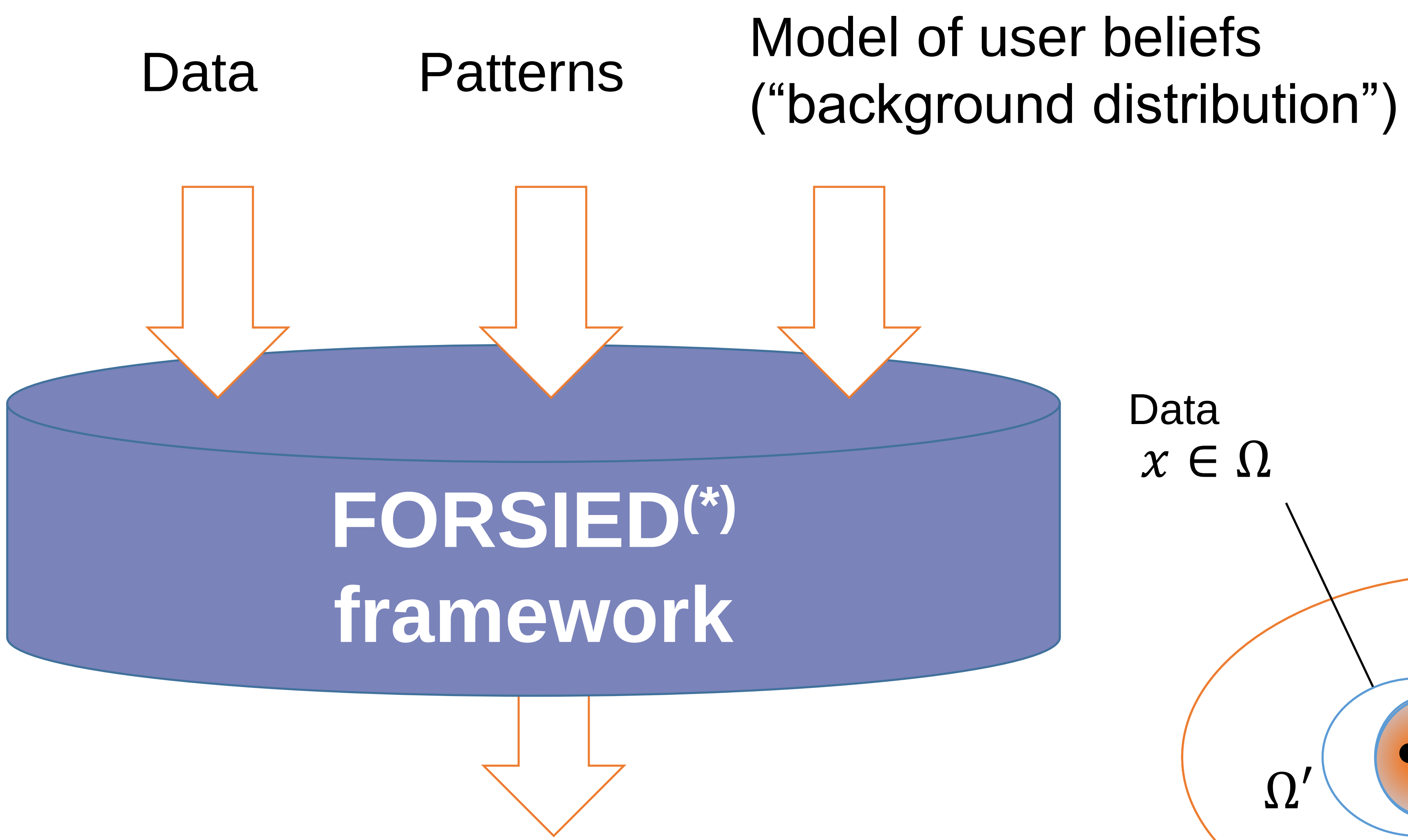


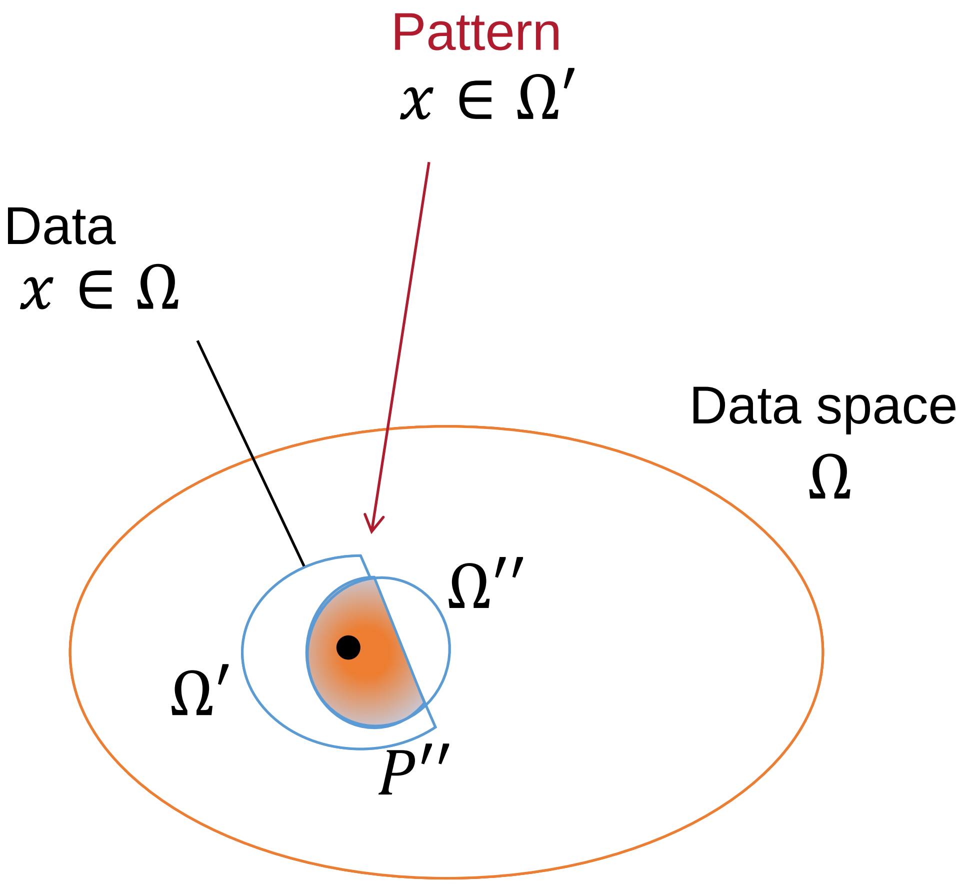
$$\text{Interestingness}(\Omega', P) = \frac{\text{InformationContent}(\Omega', P)}{\text{DescriptionalComplexity}(\Omega')}$$

GHENT UNIVERSITY

(*)Formalizing Subjective Interestingness in Exploratory Data mining

P evolves!





$$\text{InformationContent}(\Omega', P) = -\log(P(\Omega'))$$

Pattern   User

Subjective!

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

GHENT UNIVERSITY

(*)Formalizing Subjective Interestingness in Exploratory Data mining

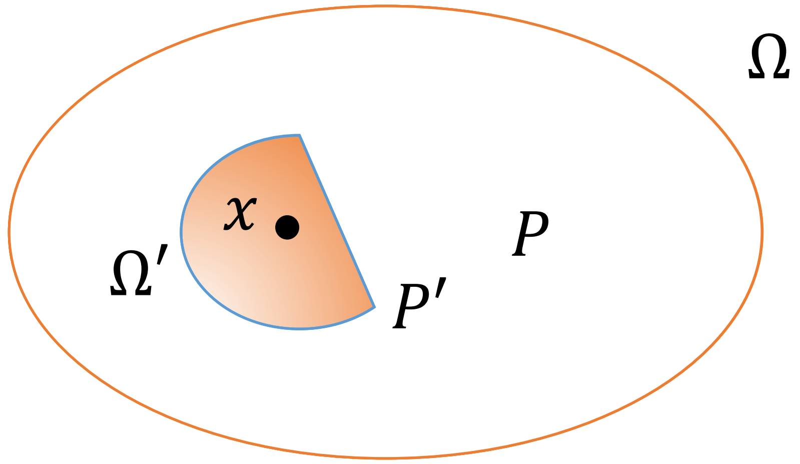
THE FINE PRINT

- 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 $x \in \Omega'$

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



- Descriptional complexity?
  - Essentially problem-dependent

GHENT UNIVERSITY

FORSIED INSTANTIATIONS

COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

Prior beliefs:

1. Overall density
2. or: Vertex degrees

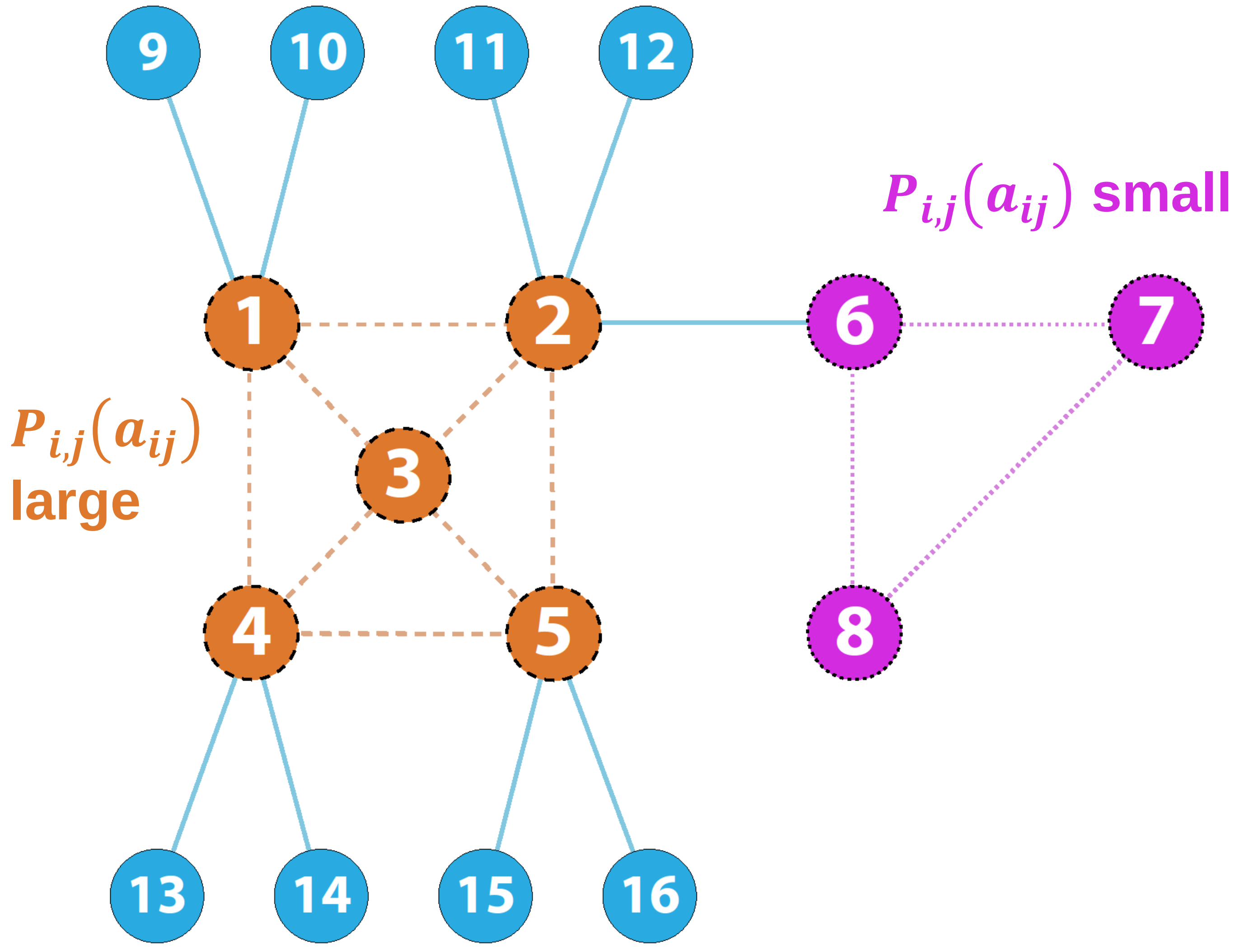
➔ MaxEnt distribution:

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

Adjacency matrix

Edge indicator variables

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



GHENT UNIVERSITY

COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

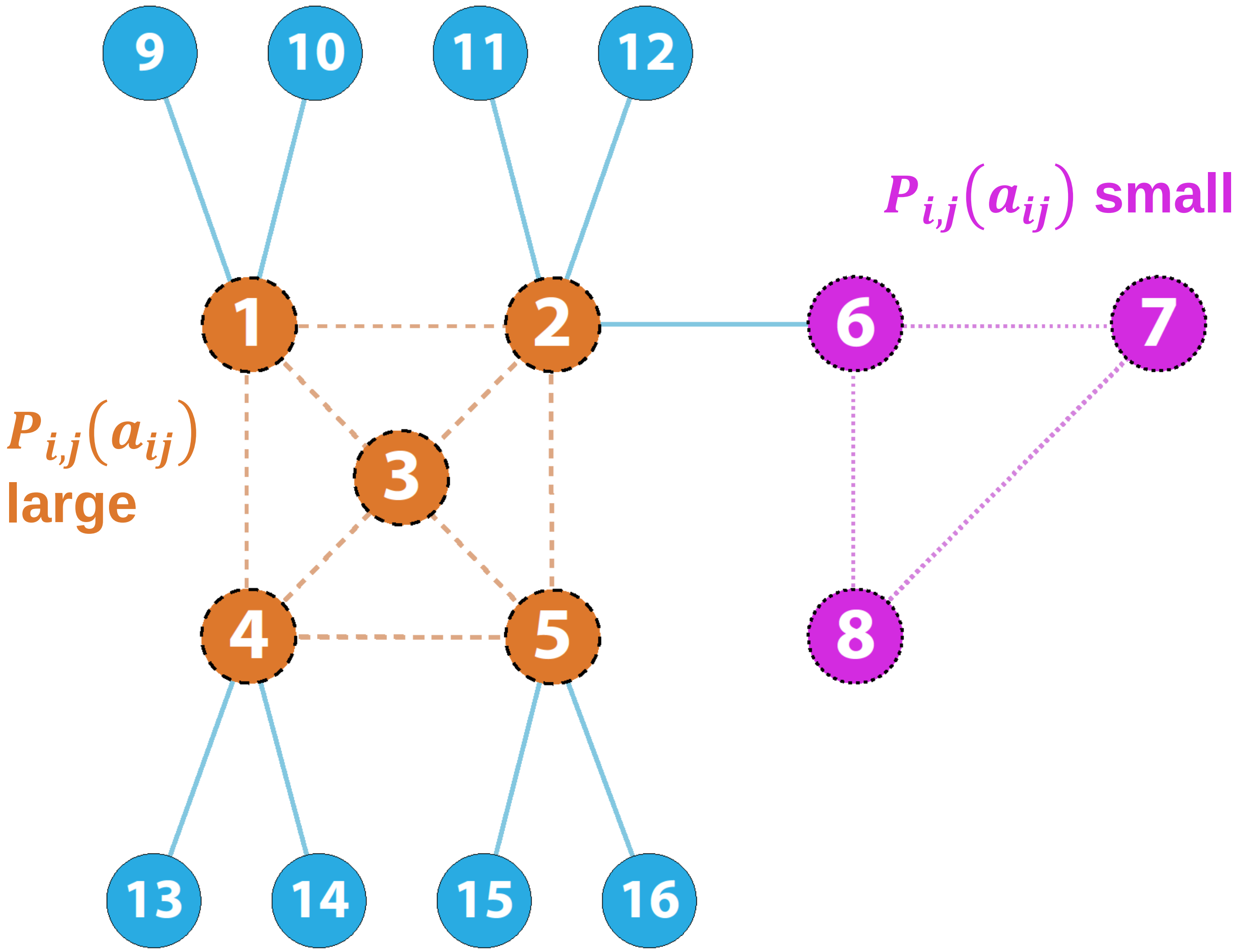
Prior beliefs:

1. Overall density
2. or: Vertex degrees

Pattern:

– Dense subgraphs

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



GHENT UNIVERSITY

COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

Prior beliefs:

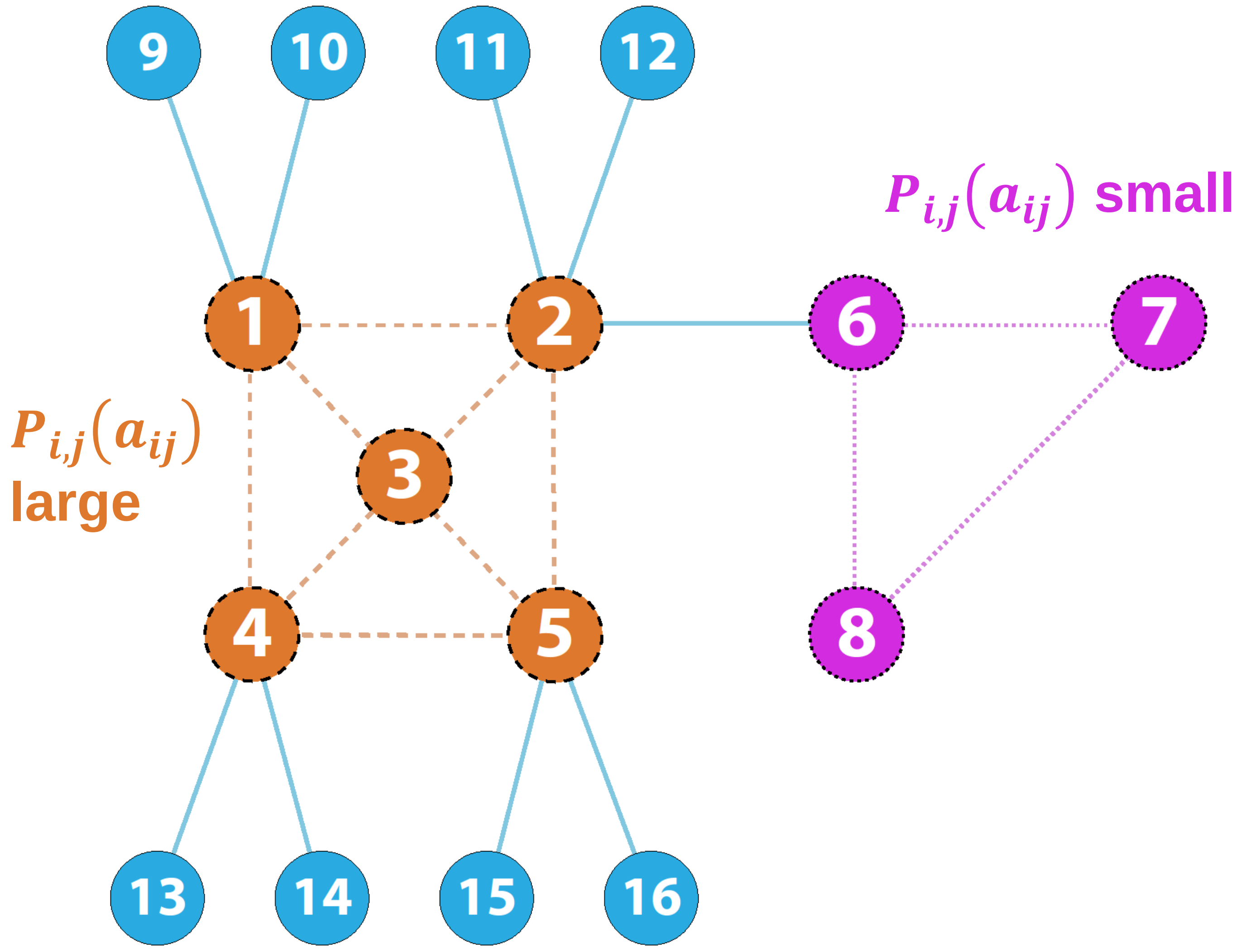
1. Overall density
2. or: Vertex degrees

Pattern:

– Dense subgraphs

Interestingness:

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



GHENT UNIVERSITY

COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

Prior beliefs:

1. Overall density
2. or: Vertex degrees

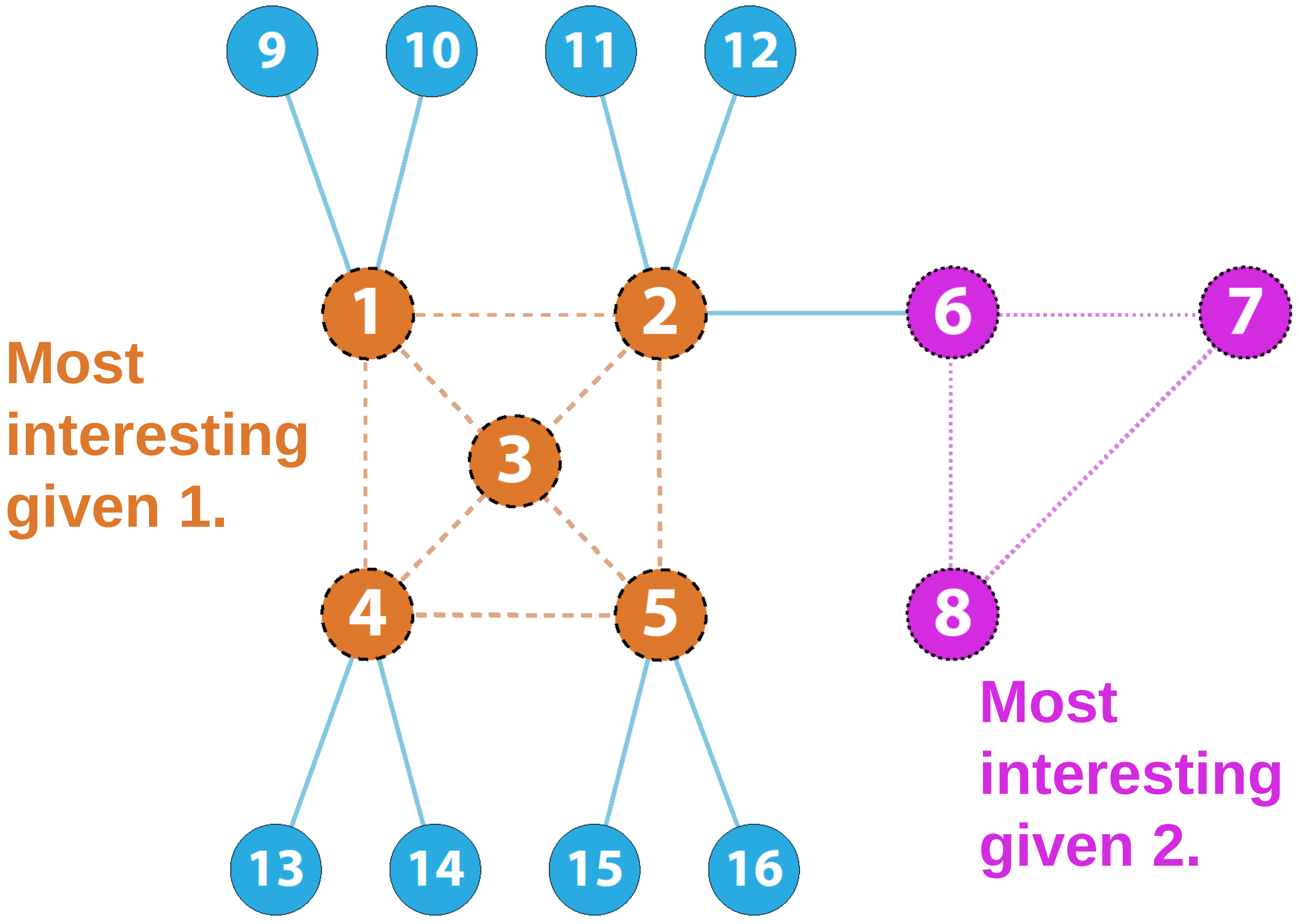
Pattern:

– Dense subgraphs

Interestingness:

– Density vs. size

– 2. ➔ preferably low degrees nodes



GHENT UNIVERSITY

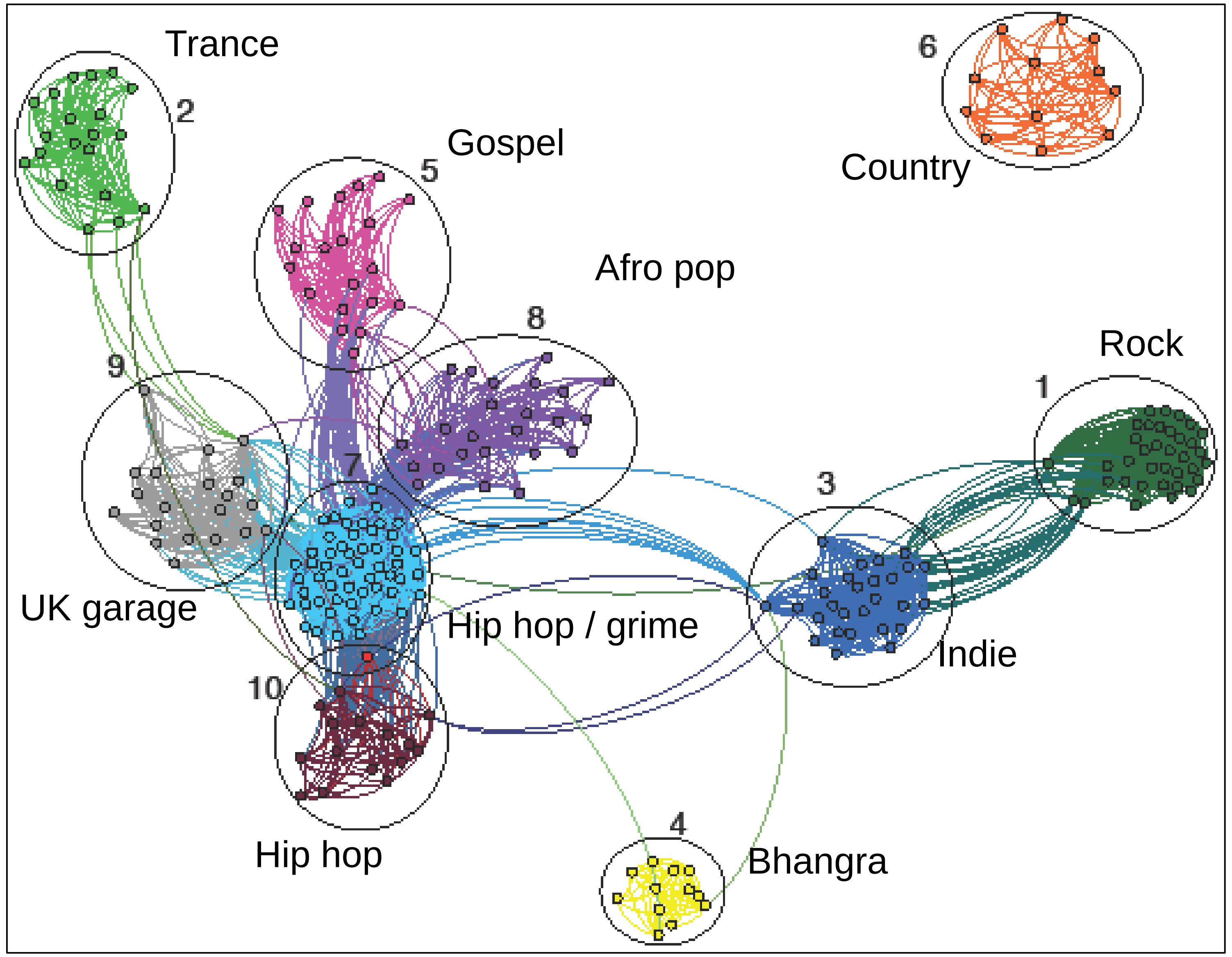
COMMUNITY DETECTION IN NETWORKS

Data:
- Graph

Prior beliefs:
1. Overall density
2. or: Vertex degrees

Pattern:
- Dense subgraphs

Interestingness:
- Density vs. size
- 2. ➔ preferably low degrees nodes



GHENT UNIVERSITY

ASSOCIATION ANALYSIS

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

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

GHENT UNIVERSITY

ASSOCIATION ANALYSIS

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

  → Background distribution P

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

	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	

GHENT UNIVERSITY

ASSOCIATION ANALYSIS

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

GHENT UNIVERSITY

ASSOCIATION ANALYSIS

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

GHENT UNIVERSITY

ASSOCIATION ANALYSIS

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

GHENT UNIVERSITY

ASSOCIATION ANALYSIS

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

GHENT UNIVERSITY

ASSOCIATION ANALYSIS

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

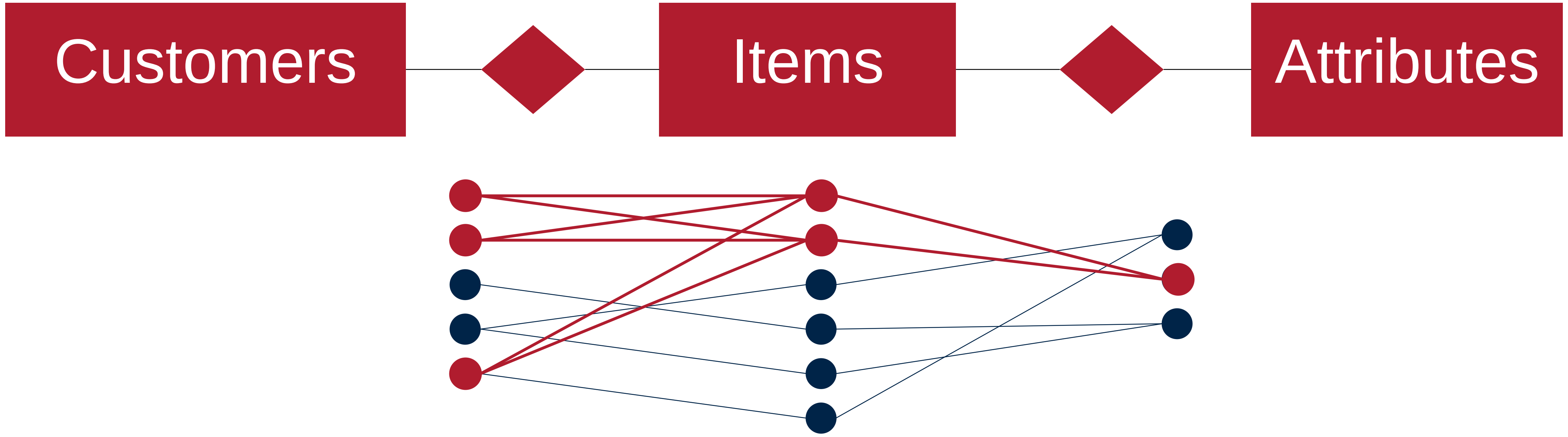
Subjective interestingness ranking **Prior info on:** **Row & column sums**	**#docs**	**Support x size (area)**	**#docs**
svm**,** support, machin, vector	25	data, paper	389
state, art	39	algorithm, propose	246
unlabelled, labelled, supervised, learn	10	data, mine	312
associ, rule, mine	36	base, method	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 **Prior info on:** **Row & column sums**	**#docs**	**Subjective interestingness ranking** **Additionally prior info on:** **Keyword tiles**	**#docs**
svm**, support**, **machin, vector**	25	art, state	39
state, art	39	row, column, algorithm	12
unlabelled, labelled, **supervised, learn**	10	unlabelled, labelled, data	14
associ, rule, mine	36	answer, question	18
gene, express	25	Precis, recal	14

GHENT
UNIVERSITY

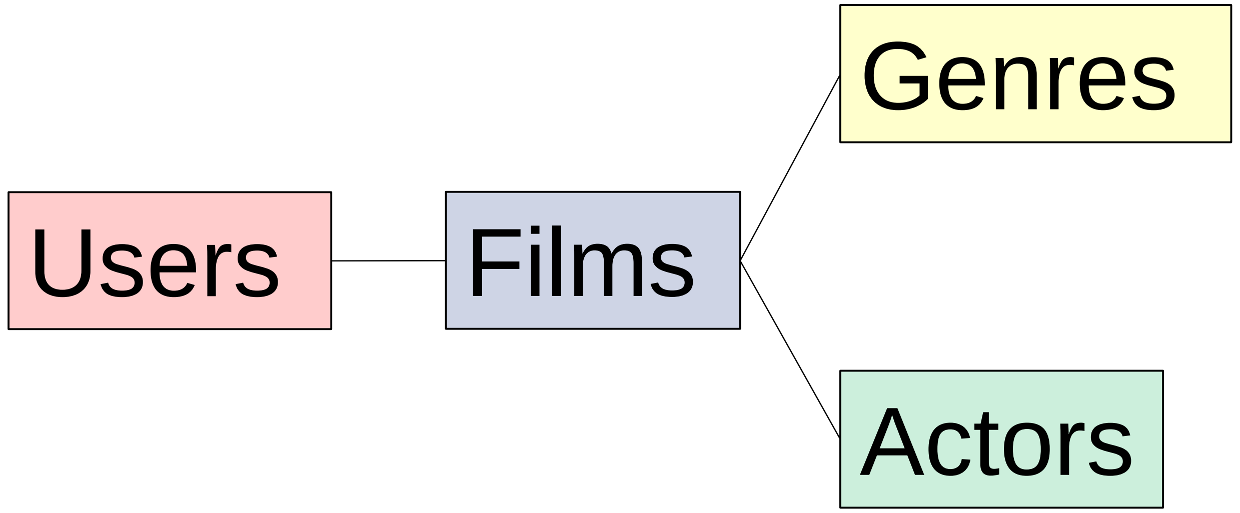
RELATIONAL PATTERN MINING

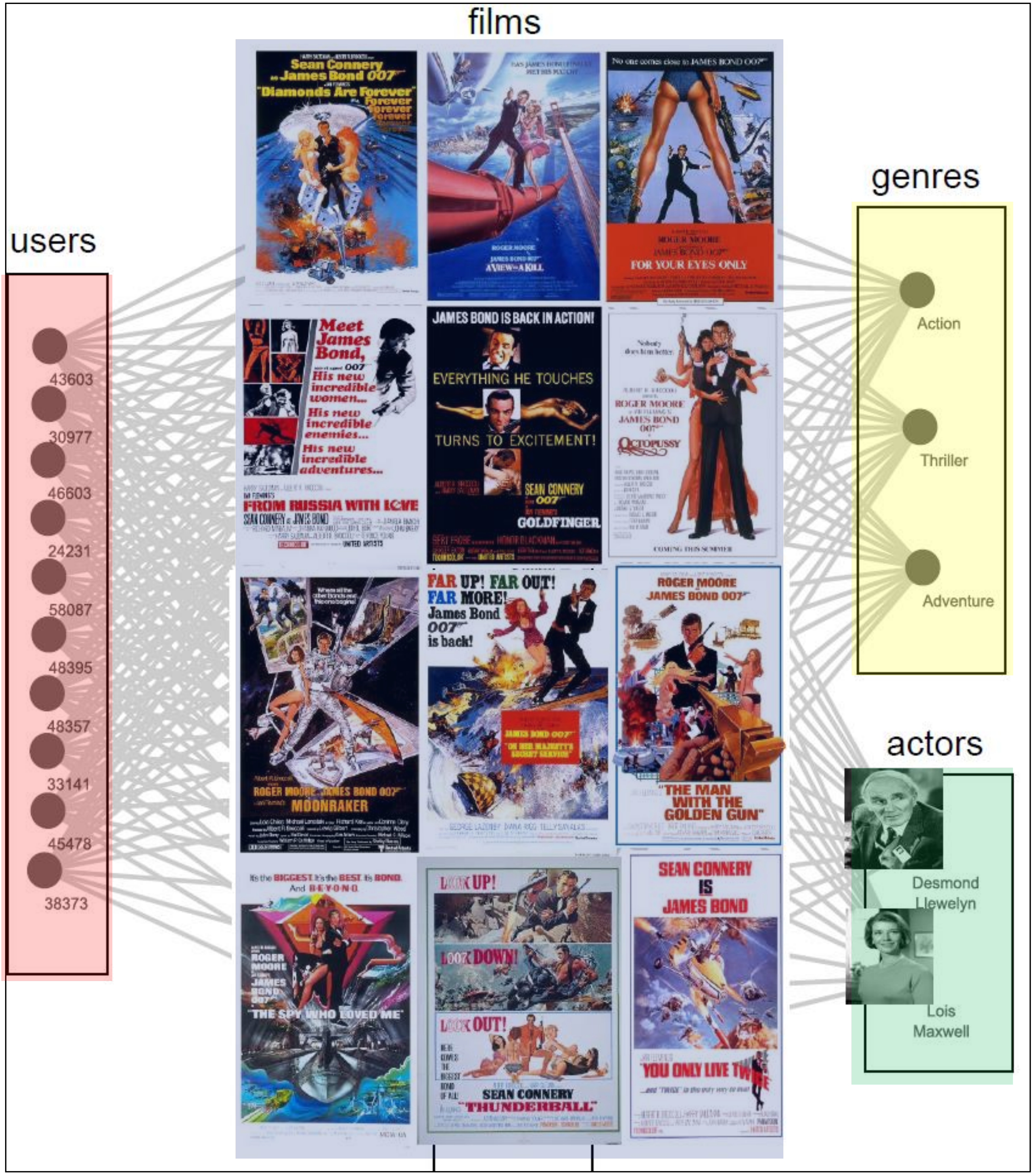
- **Data**: relational database
- **Pattern**: connected complete subgraphs
- **Prior beliefs**: degree of each node in each relationship



GHENT UNIVERSITY

RELATIONAL PATTERN MINING

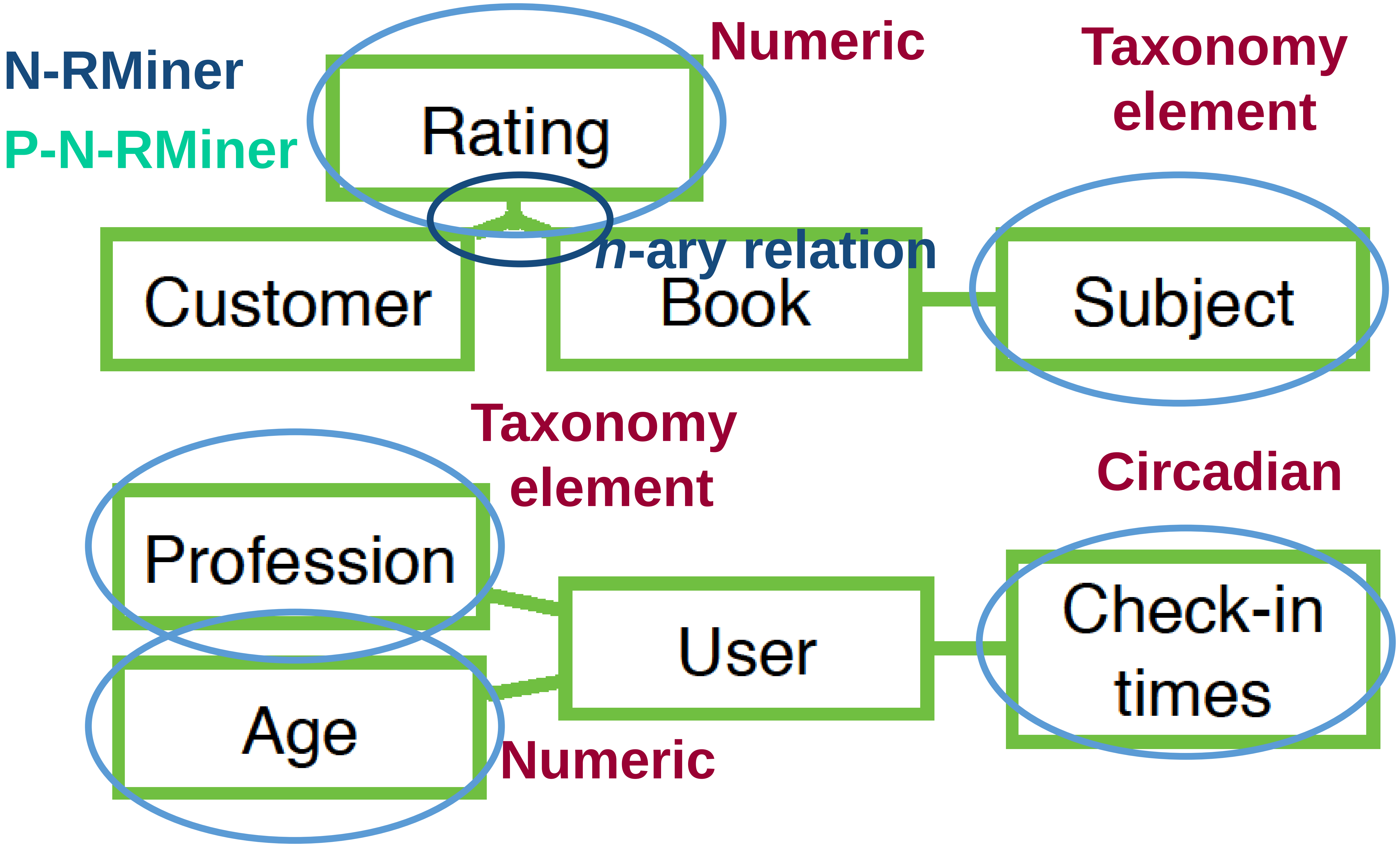




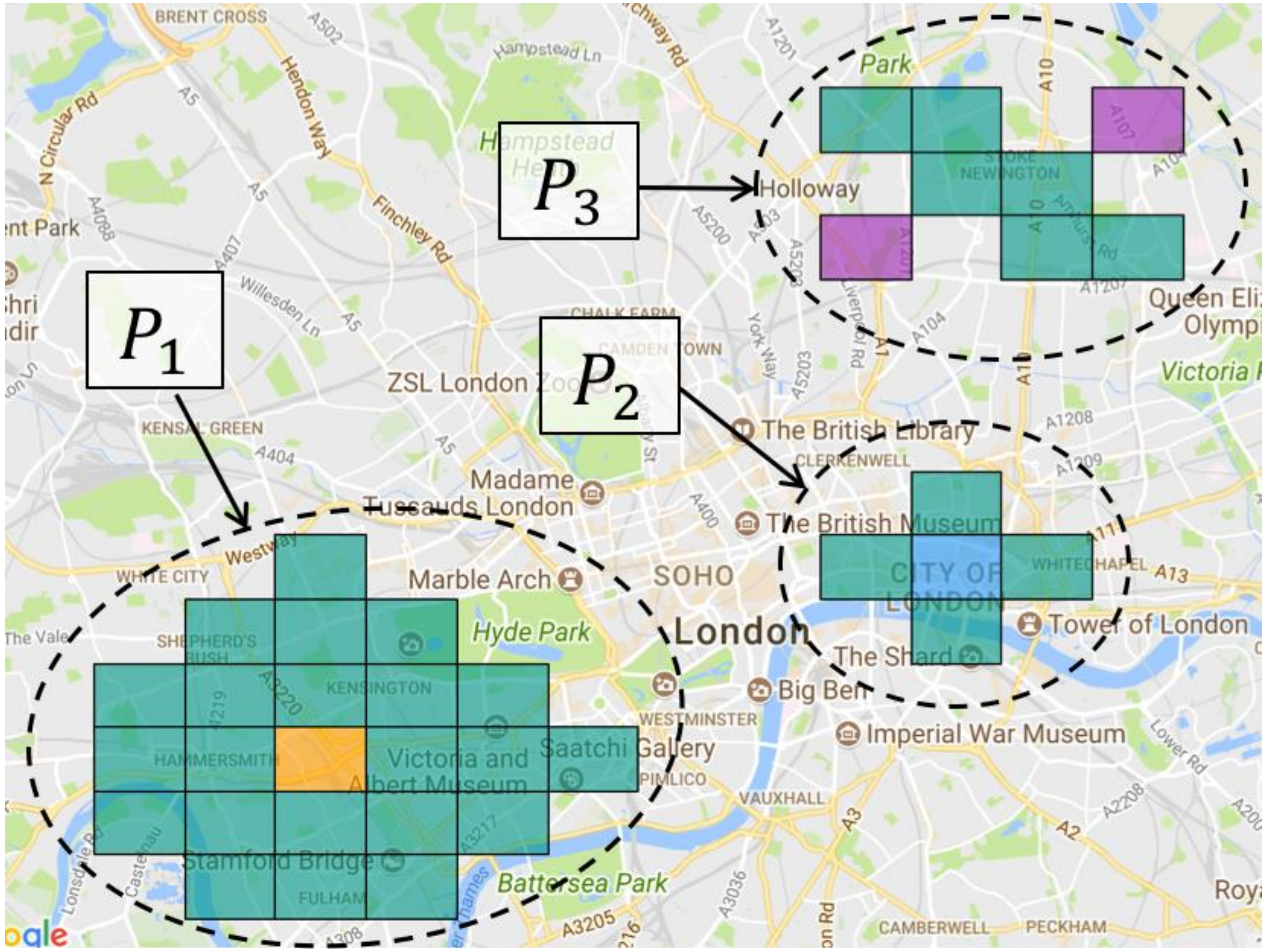
RMiner

GHENT UNIVERSITY

RELATIONAL PATTERN MINING

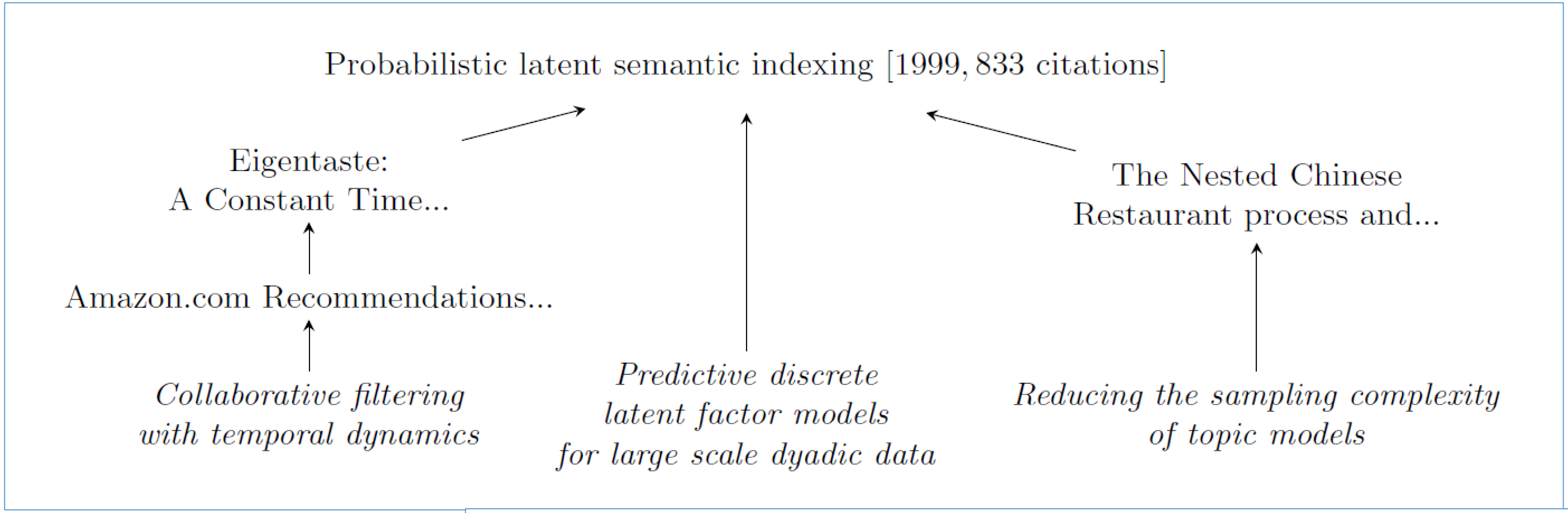


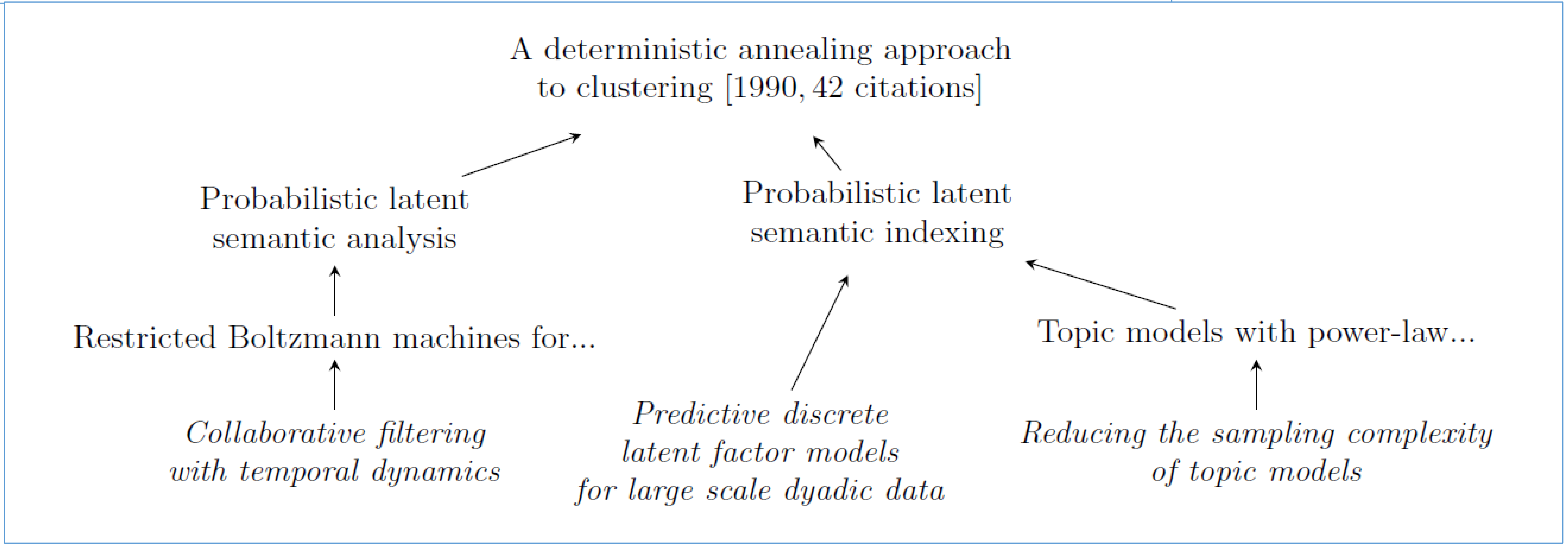
COHESIVE SUBGRAPHS IN ATTRIBUTED GRAPHS



GHENT UNIVERSITY

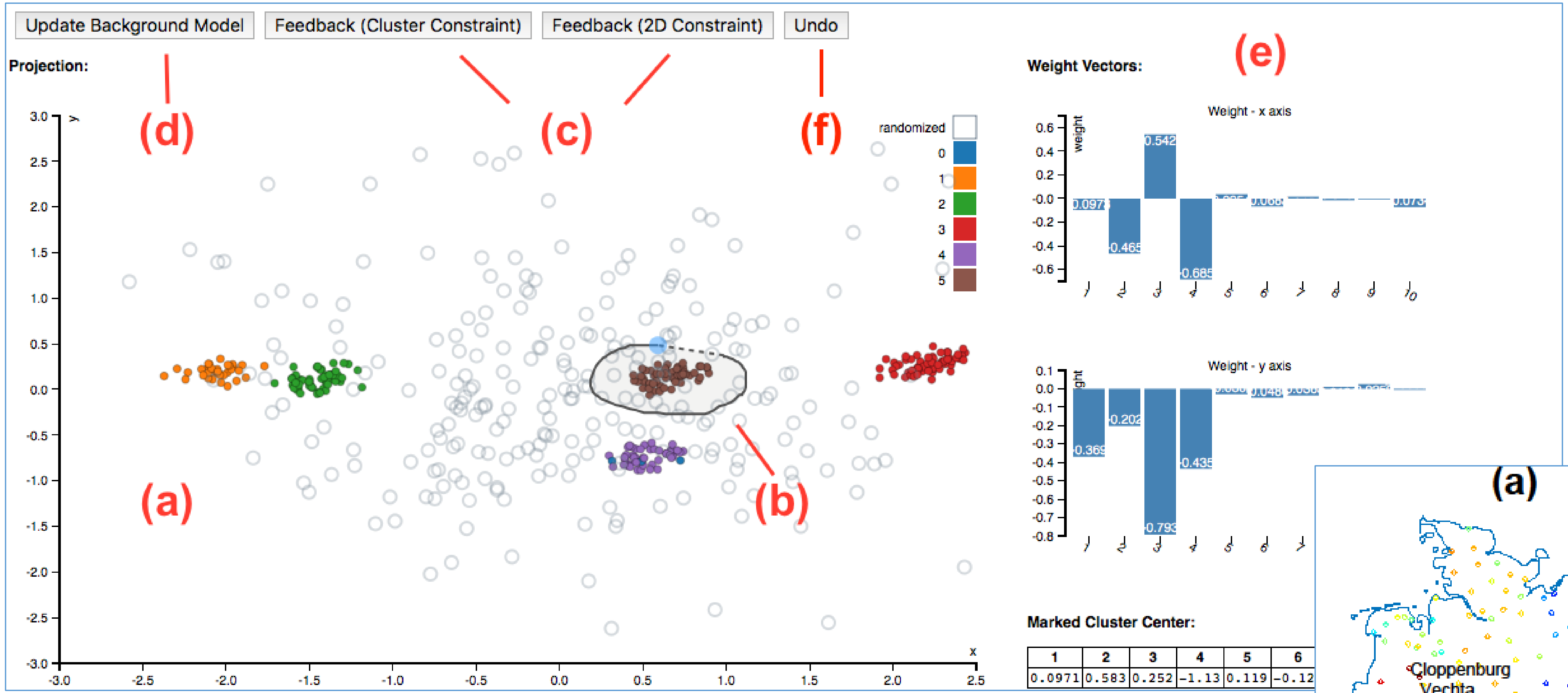
INTERESTING CONNECTING TREES

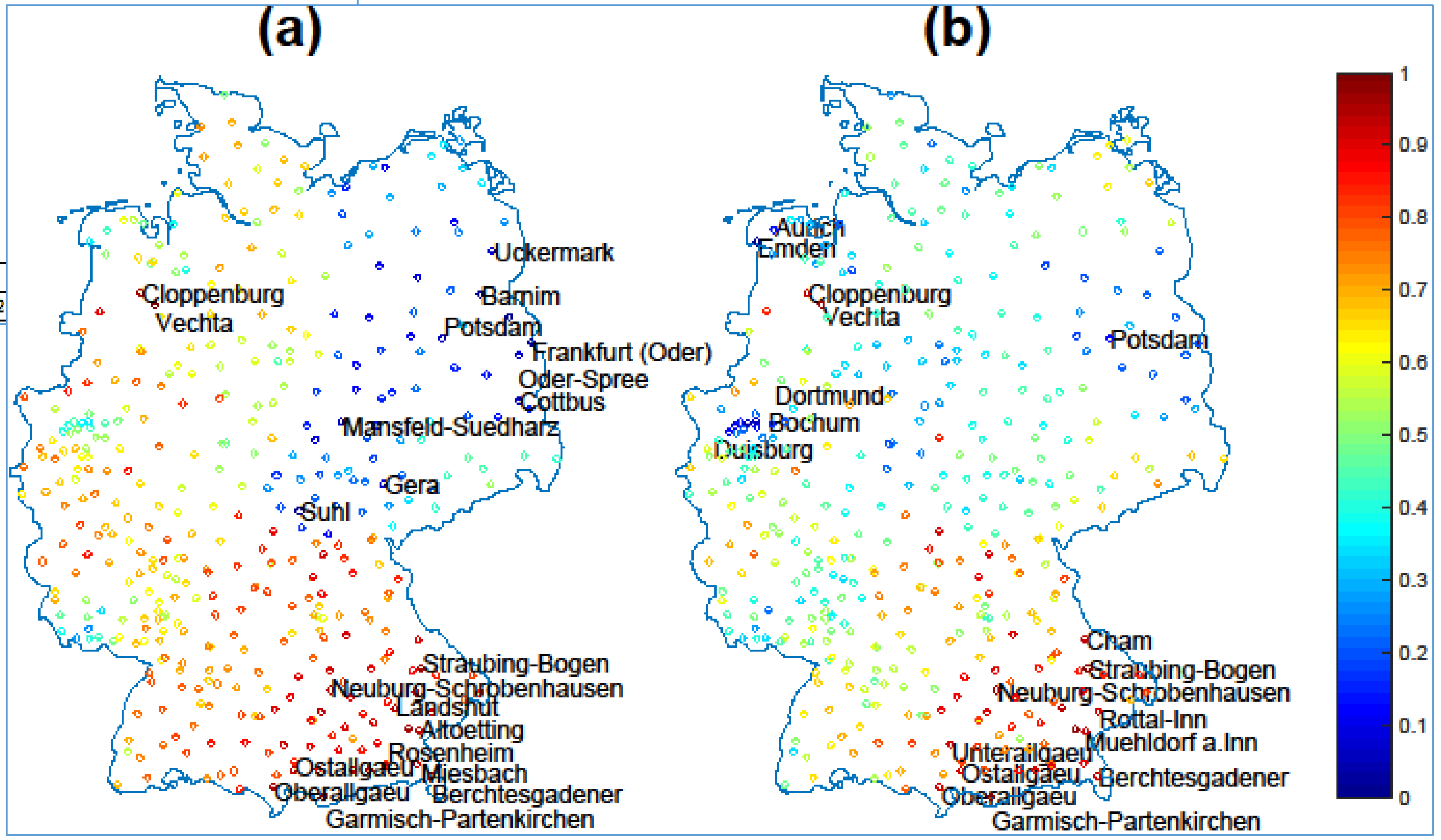




GHENT UNIVERSITY

DATA PROJECTIONS



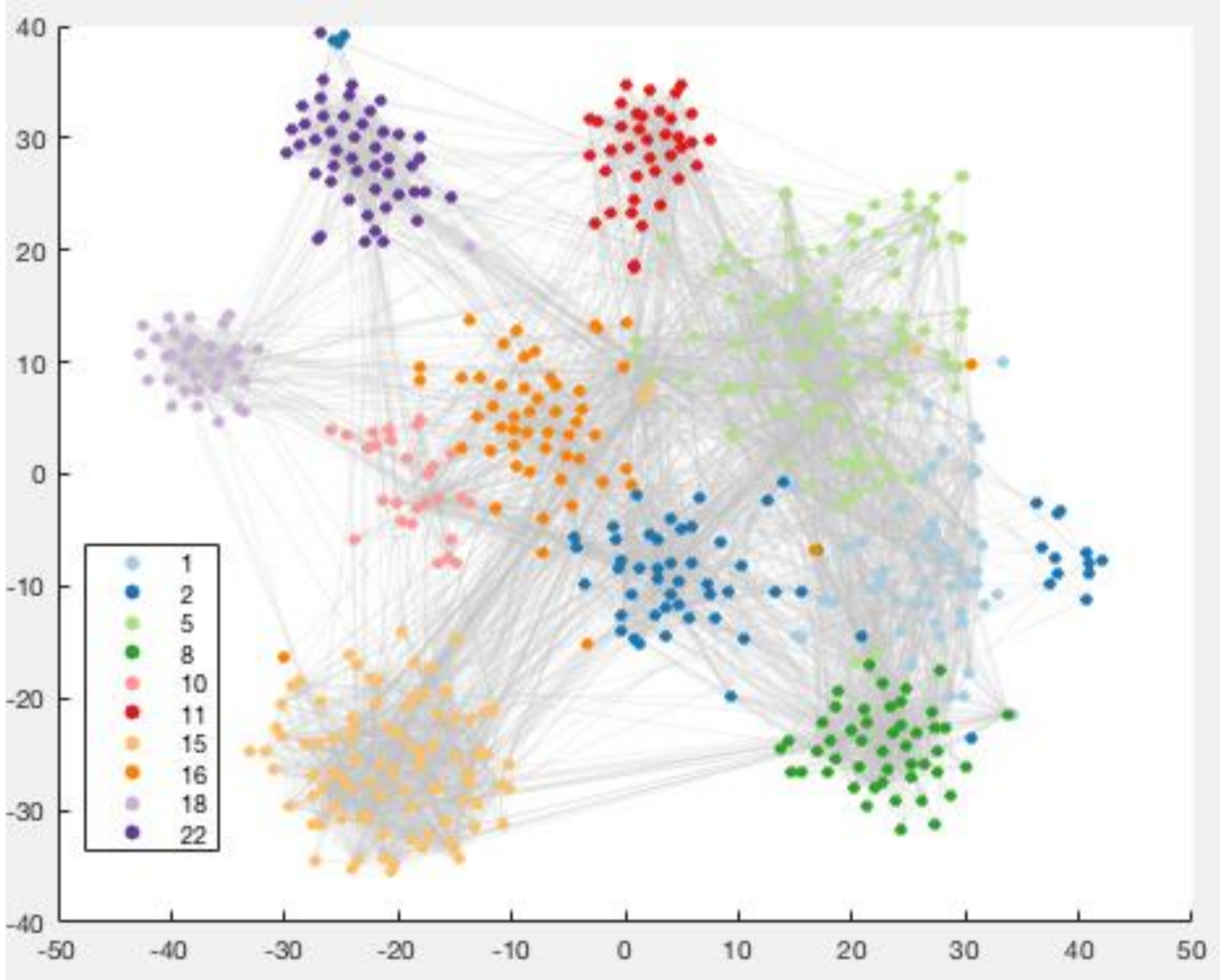


GHENT UNIVERSITY

CONDITIONAL GRAPH EMBEDDINGS

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

$$\max_{\boldsymbol{X}} P(G|\boldsymbol{X})$$



GHENT UNIVERSITY

AND MORE

- **Past**
  - Data clustering
  - Exceptional model mining
  - 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

GHENT UNIVERSITY

PRIVACY...

- **Privacy-preserving data publishing**
  - Prevent linking identity with values of a sensitive attribute

Anonymized patient database

Quasi-identifiers

Voting records 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

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...

- **Privacy-preserving data publishing:**
  - Prevent linking identity with values of a sensitive attribute
  - Mere anonymization of records is not enough

PRIVACY...

– **Generalize** quasi-identifiers to limit privacy loss

Patient database

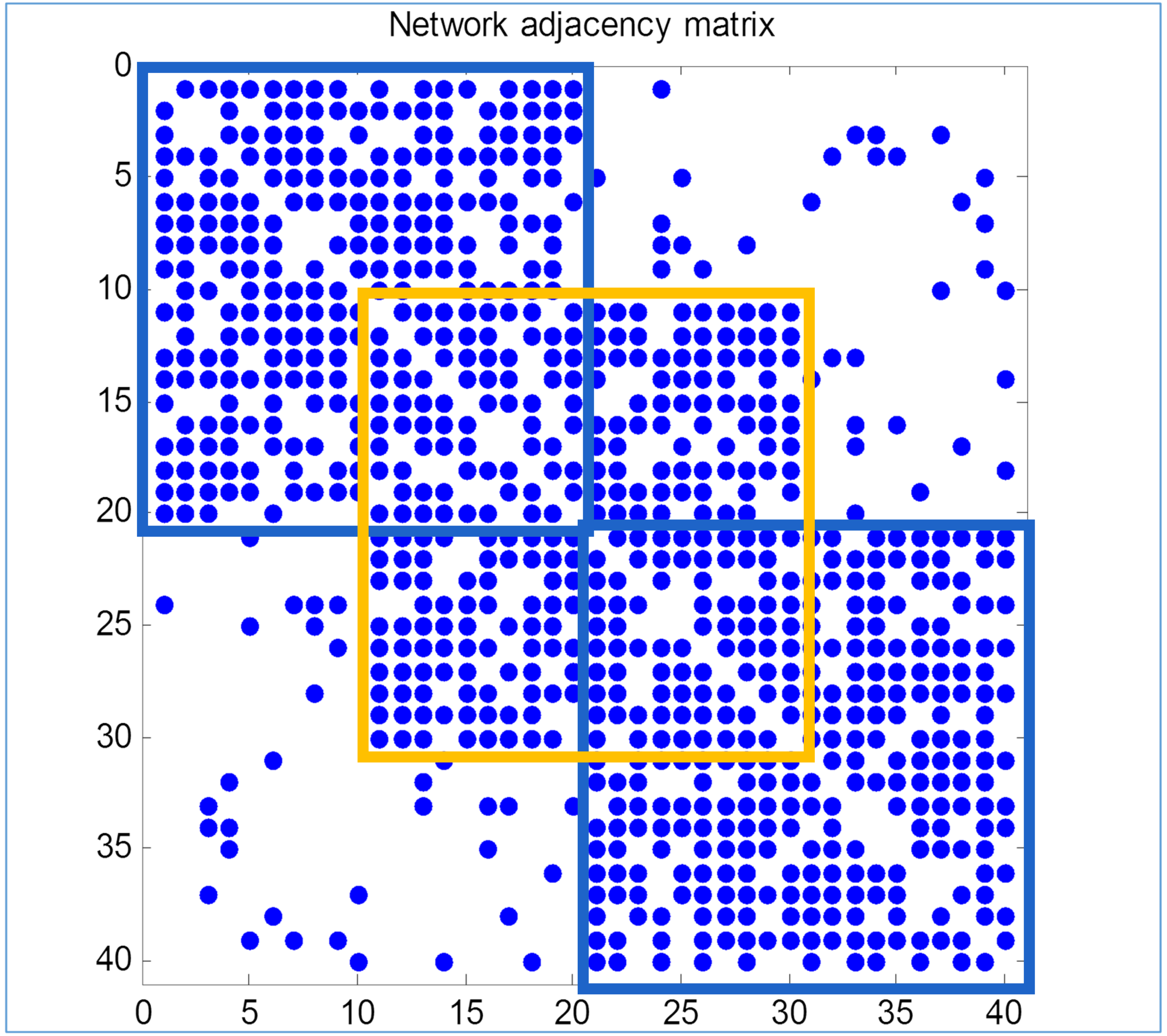
ZIP	D.O.B.	Sex	Diagnosis
94701	01/02/1968	F	Healthy
94701	06/03/1990	F	Obesitas
94702	11/08/1991	M	Healthy
94703	03/09/1979	M	Cervical cancer
94703	07/10/1951	F	Healthy
94704	10/02/1973	M	Obesitas
94705	20/12/2001	F	Obesitas

Generalized patient database

ZIP	D.O.B.	Sex	Diagnosis
94701	‘51-’01	F	Healthy
94701	‘51-’01	F	Obesitas
94702-5	‘51-’01	M	Healthy
94702-5	‘51-’01	M	Prostate cancer
94702-5	‘51-’01	F	Healthy
94702-5	‘51-’01	M	Obesitas
94702-5	‘51-’01	F	Obesitas

... AND MORE

– Existence of a tight community in a network



GHENT UNIVERSITY

... 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 **sanitized version of database**,
- identifying **dense subgraphs**,
- finding **clusters**,
- mining **frequent itemsets**, etc

Data mining patterns

GENERAL STRATEGY

- **Data:** $\boldsymbol{x}$
  - Data mining goal: **reveal as much as possible about $\boldsymbol{x}$**
- **Sensitive aspects:** $f(\boldsymbol{x}) \in \Phi$
  - the sensitive attributes' values
  - density of a specified subgraph
  - existence of a tight cluster
  - frequency of an item
  - Goal: **reveal as little as possible about $f(\boldsymbol{x})$**

(data) $\boldsymbol{x} \rightarrow f(\boldsymbol{x})$ (sensitive aspects)

- **Current/future work:**
  - Quantify the information revealed by a pattern about $\boldsymbol{x}$ and $f(\boldsymbol{x})$
  - So they can be traded-off with each other

Considering the user's prior beliefs!
➔ 'Subjective' measures

GHENT UNIVERSITY

QUANTIFYING SENSITIVE INFORMATION LOSS

- **How much does knowing that $\boldsymbol{x} \in \Omega'$ reveal about $f(\boldsymbol{x}) \in \Phi$?**
- **Background distribution** P_f for sensitive aspects $f(\boldsymbol{x})$:

$$P_f(\Phi') \triangleq P(\Omega')$$

  where $\Omega' = f^{-1}(\Phi')$.
- $P_f(\Phi')$ represents degree of belief analyst attaches to the statement that the sensitive aspects $f(\boldsymbol{x})$ belong to $\Phi' \in \Phi$

- **Updating** $P \to P'$ results in updating $P_f \to P'_f$.
  - More complex than conditioning!
  - $P_f\big(f(\boldsymbol{x})\big)$ can be larger or smaller than $P'_f\big(f(\boldsymbol{x})\big)$

GHENT UNIVERSITY

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):
     $$-\log\left(P'_f\left(f(\boldsymbol{x})\right)\right)$$
   - Entropy of P_f (**uncertainty about sensitive attributes**):
     $$-E_{x\sim P_f}\left\{\log\left(P'_f\left(f(\boldsymbol{x})\right)\right)\right\}$$
   - **Knowledge gained** about actual value of the sensitive aspects:
     $$-\log\left(\frac{P_f\left(f(\boldsymbol{x})\right)}{P'_f\left(f(\boldsymbol{x})\right)}\right)$$
   - Degree of belief that the sensitive aspects are within a specified set $\Phi^* \subseteq \Phi$:
     $$P'_f(\Phi^*)$$
   - ...

GHENT UNIVERSITY

EXAMPLES

GHENT UNIVERSITY

PRIVACY-PRESERVING DATA PUBLISHING

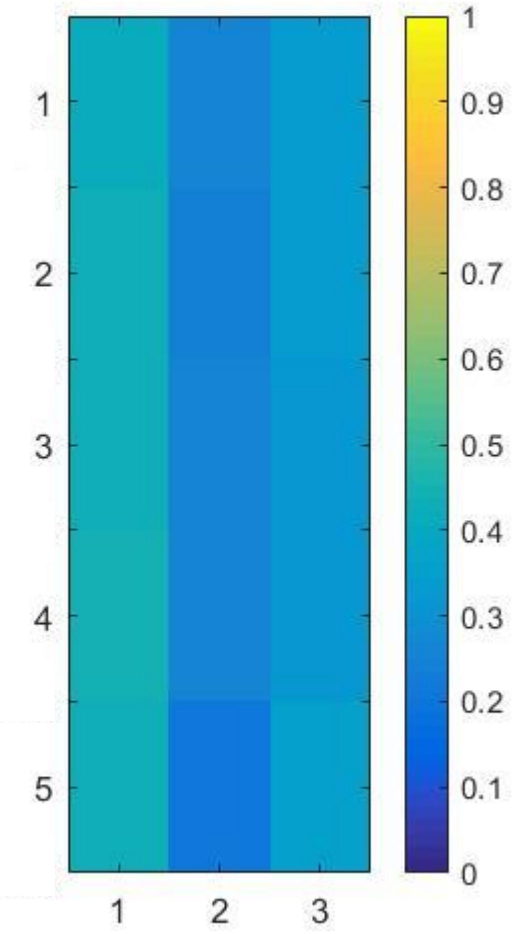
- **Random synthetic dataset:**
  - 5 real-valued quasi-identifiers, generalization through intervals
  - 1 sensitive attribute, 3 possible values
  - 1 other attribute, 3 possible values
  - 100 data records
- **Trade-off:**
  - information about **data** (other **&** sensitive attributes)
  - knowledge gained about **sensitive attribute**
- **Generalize quasi-attributes** → 5 equivalence classes
  - Ensure the maximum information content about any sensitive attribute value is small

E.g.
- Zip code
- DOB

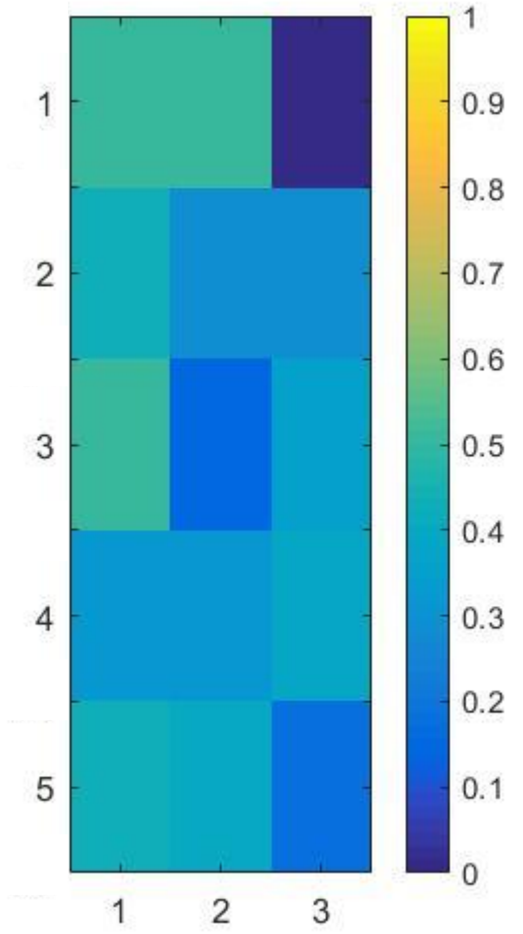
E.g.
- Sexual orientation
- Ethnicity

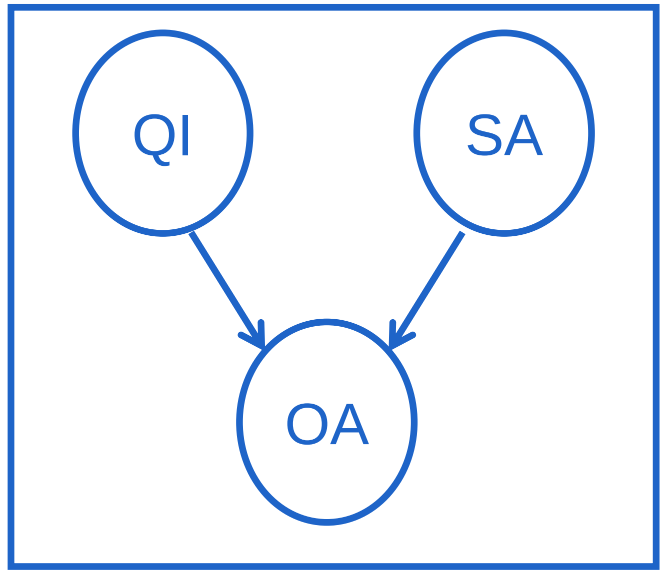
E.g.
- Sense of well-being
- Productivity

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

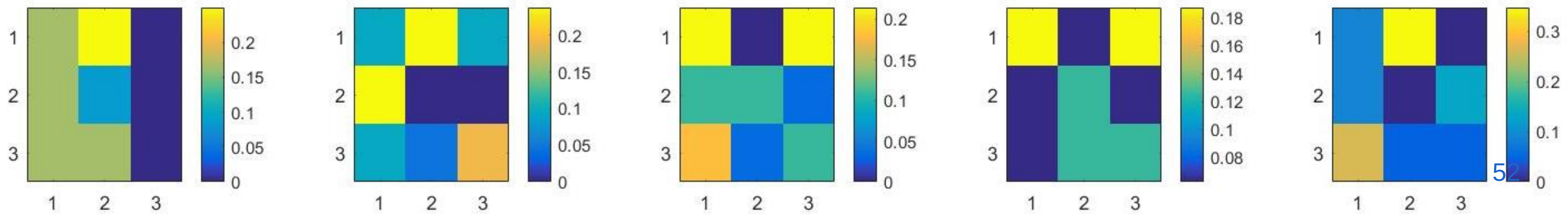


Conditional distributions within the 5 equivalence classes over the 3 **other** attribute values



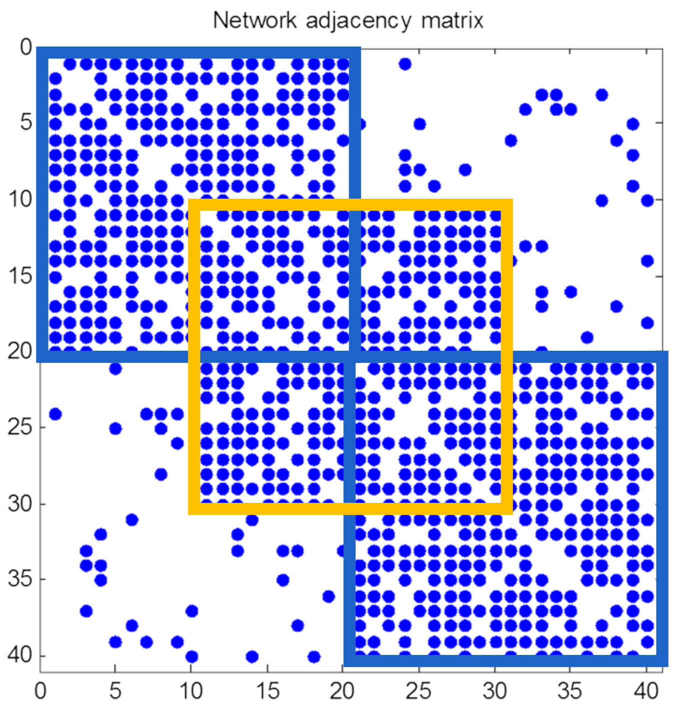


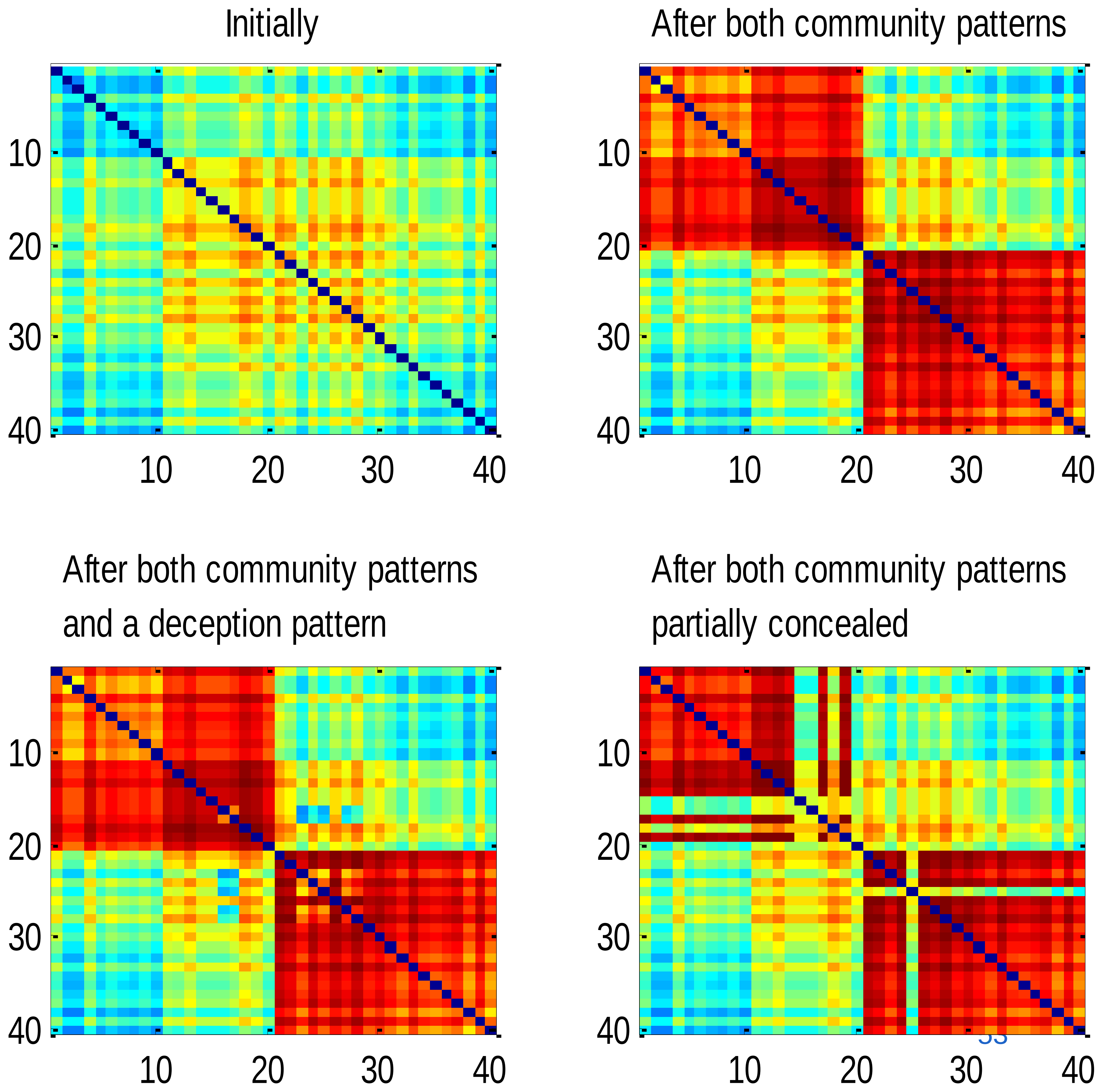
Joint conditional distribution of the **sensitive** (rows) and **other** (columns) attributes, within 5 equivalence classes



DENSE SUBGRAPHS WITHOUT SPILLING BEANS

- **Random network**:
  - 2 non-overlapping communities
  - A 3rd community overlapping both
- The 3rd is **sensitive**
  - Analyst should remain surprised by its presence
- **Task**:
  - Identify (non-)dense subgraphs
  - Without spilling the beans on the 3rd community
- **Approaches** (result from general strategy):
  - Deceive
  - Conceal





“Data Mining without Spilling the Beans: Preserving more than Privacy alone”
Project funded by the FWO (with Jefrey 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

European Research Council
Established by the European Commission

www.forsied.net

GHENT UNIVERSITY

Thanks...