FACULTY OF ENGINEERING AND ARCHITECTURE

DEPARTMENT ELECTRONICS AND INFORMATION SYSTEMS (ELIS)
RESEARCH GROUP IDLAB

MINING SUBJECTIVELY INTERESTING PATTERNS IN DATA

PART 1/5: MOTIVATION, BACKGROUND & OUTLINE

Jefrey Lijffijt Tijl De Bie
Ghent University

GHENT UNIVERSITY

European Research Council
Established by the European Commission

fwo

MOTIVATION

GHENT
UNIVERSITY

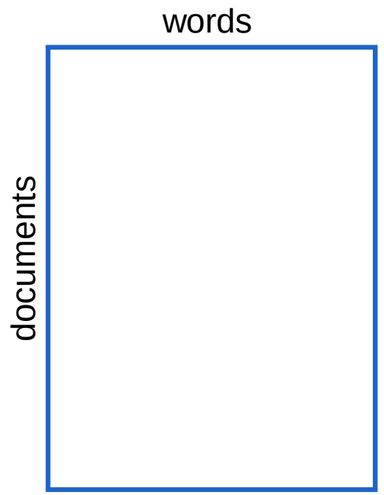
SUBJECTIVITY = KEY

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

KDD abstracts dataset:

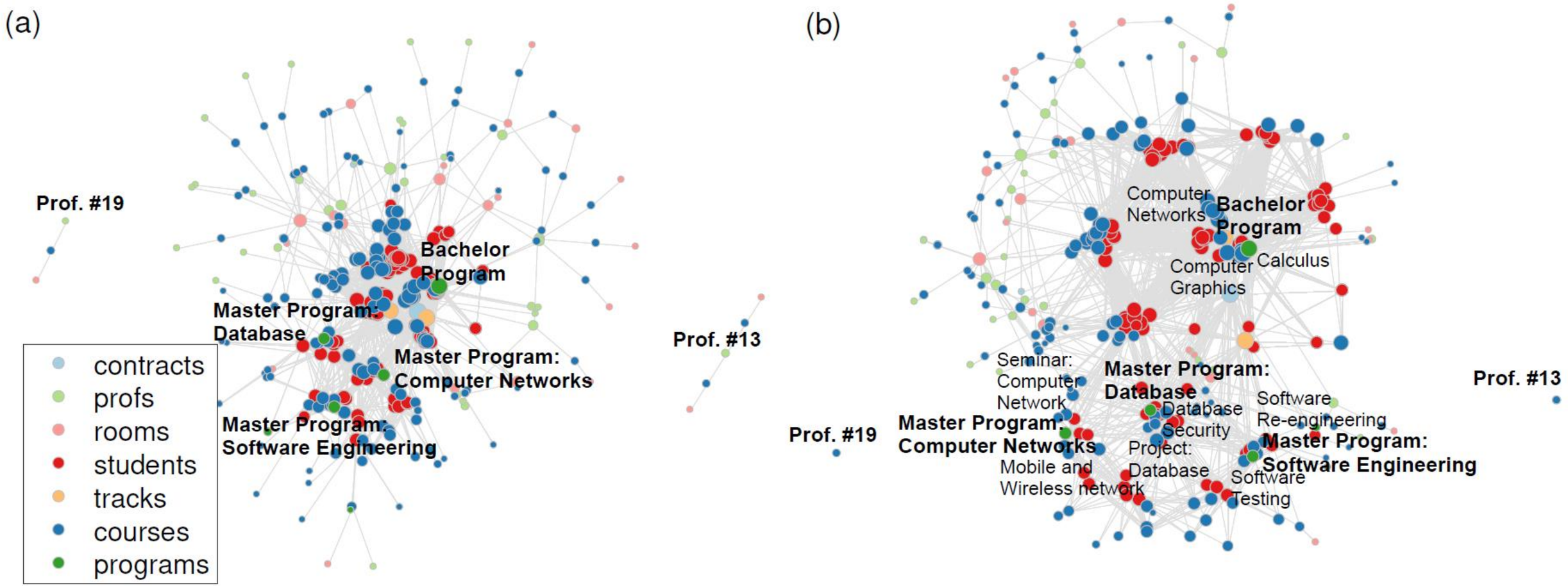


Support x size (area)	#docs
data, paper	389
algorithm, propose	246
data, mine	312
base, method	202
result, show	196
problem	373

Subjective interestingness ranking Prior info on: Row & column sums	#docs
svm**, support**, **machin, vector**	25
state, art	39
unlabelled, labelled, **supervised, learn**	10
associ, rule, mine	36
gene, express	25

GHENT UNIVERSITY

CONDITIONAL NETWORK EMBEDDINGS



GHENT UNIVERSITY

PRIVACY-PRESERVING DATA PUBLISHING

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

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*

- 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
- …

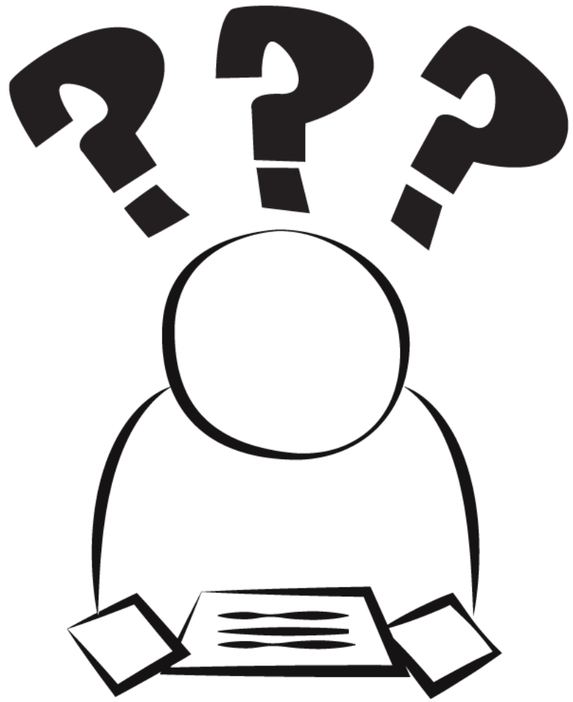
‘Interestingness measures’

GHENT UNIVERSITY

THE CHALLENGE

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

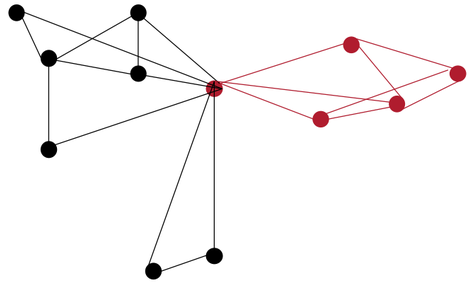
- Challenge:
  - **Formalise *true* interestingness!**
    - With minimal user interaction
    - Without requiring user expertise



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

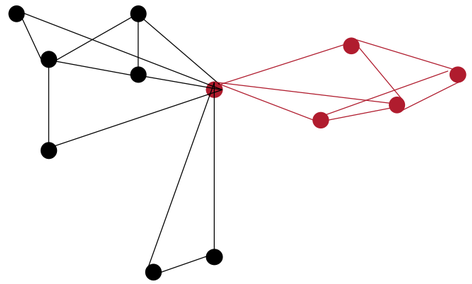
Interestingness(pattern) → Interestingness(pattern, analyst)

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

OUTLINE

GHENT
UNIVERSITY

OUTLINE

Part 1: Introduction and motivation

15mins

Part 2: The FORSIED framework

40mins

Part 3: Binary matrices, graphs, and relational data

45mins

COFFEE BREAK

Part 4: Numeric and mixed data

55mins

Part 5: Advanced topics, outlook & conclusions

30mins

Q&A

15mins

GHENT UNIVERSITY

Feel free to interrupt for questions anytime

GHENT UNIVERSITY