FACULTY OF ENGINEERING AND ARCHITECTURE

DEPARTMENT ELECTRONICS AND INFORMATION SYSTEMS (ELIS)
RESEARCH GROUP IDLAB

MINING SUBJECTIVELY INTERESTING PATTERNS IN DATA

PART 2/5: THE FORSIED FRAMEWORK

Jefrey Lijffijt Tijl De Bie
Ghent University

GHENT UNIVERSITY

European Research Council
Established by the European Commission

fwo

GENERIC FRAMEWORK

GHENT UNIVERSITY

De Bie, KDD 2011

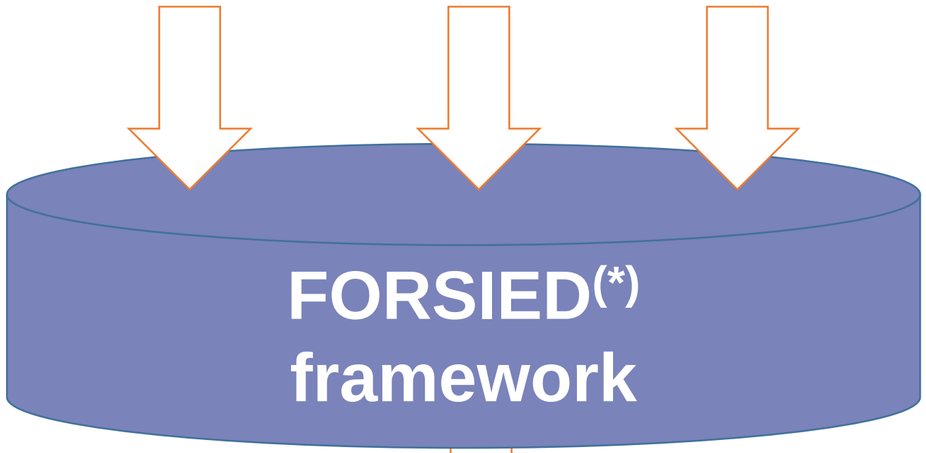
De Bie, DAMI 2011

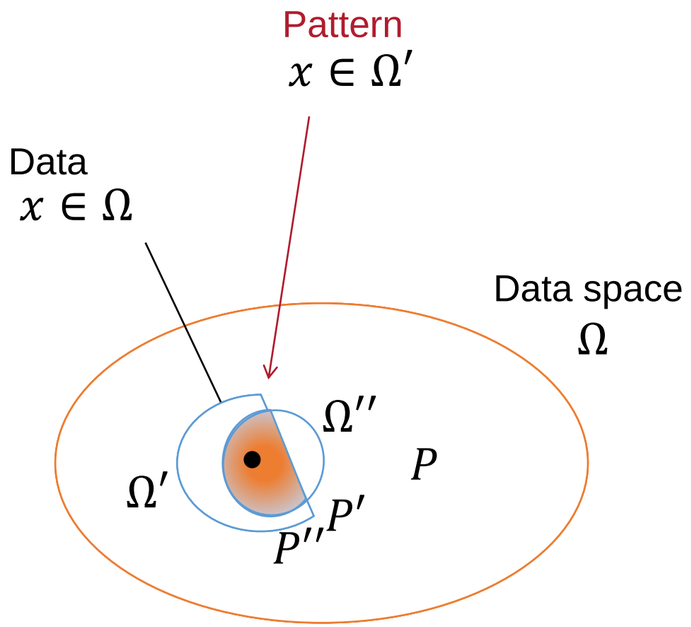
- **Data:** the adjacency matrix of a graph under study
- **Patterns:** the claim that a specified set of nodes are densely connected
- **Prior beliefs:** the degrees of the nodes, a known block structure,...
- **Interestingness:** subjective information density
- Overlapping communities!

GHENT UNIVERSITY

Data Patterns Model of user beliefs ("background distribution")

P evolves!





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

Pattern User

Subjective!

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

SI = IC/DL

(*)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 $x \in \Omega'$

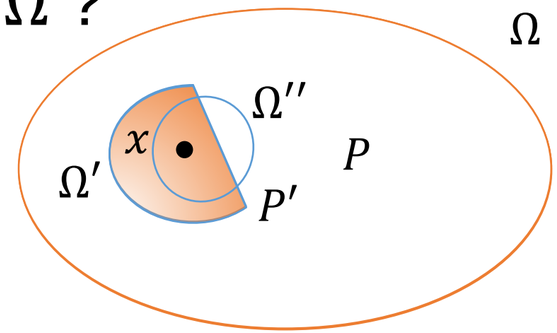
$$P'(\Omega'') = \frac{P(\Omega'' \cap \Omega')}{P(\Omega')}$$

$$\Rightarrow \quad -\log\big(P'(\boldsymbol{x})\big) = -\log\big(P(\boldsymbol{x})\big) + \log\big(P(\Omega')\big)$$

Information content in data after pattern

Information content in data before pattern

Minus information content of pattern under P



- Description length?
  - Essentially problem-dependent

GHENT UNIVERSITY

WHY MAXIMUM ENTROPY / CONDITIONING?

- Most **unbiased estimate**
  - Informal... no bias other than the constraints
- Assume **cautious / pessimistic user**
  - A user who expects to be very surprised
- Leads to **most robust estimate** of true subjective information content
  - Information content estimated with maxent P will never differ much from information content w.r.t. true prior belief of user

GHENT UNIVERSITY

A FIRST INSTANTIATION: COMMUNITY DETECTION

van Leeuwen, De Bie, Spyropoulou, Mesnage, MLj, 2016

GHENT UNIVERSITY

COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

Prior beliefs:

1. Overall density
2. or: Vertex degrees

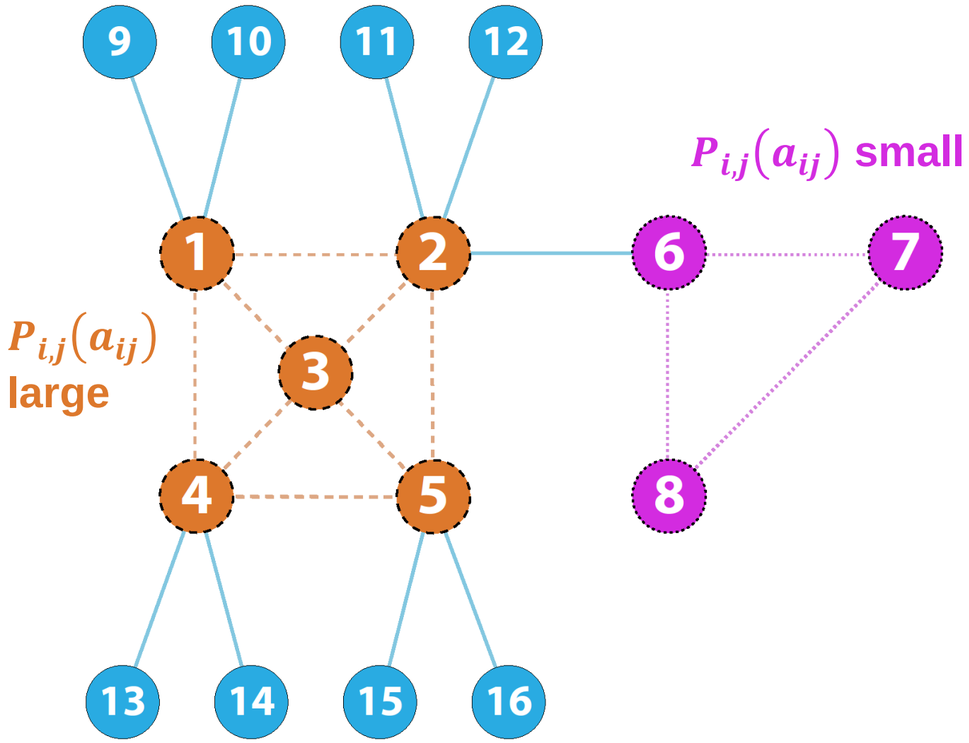
➔ MaxEnt distribution:

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

Adjacency matrix

Edge indicator variables

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



GHENT UNIVERSITY

COMMUNITY DETECTION IN NETWORKS

Data:

- Graph

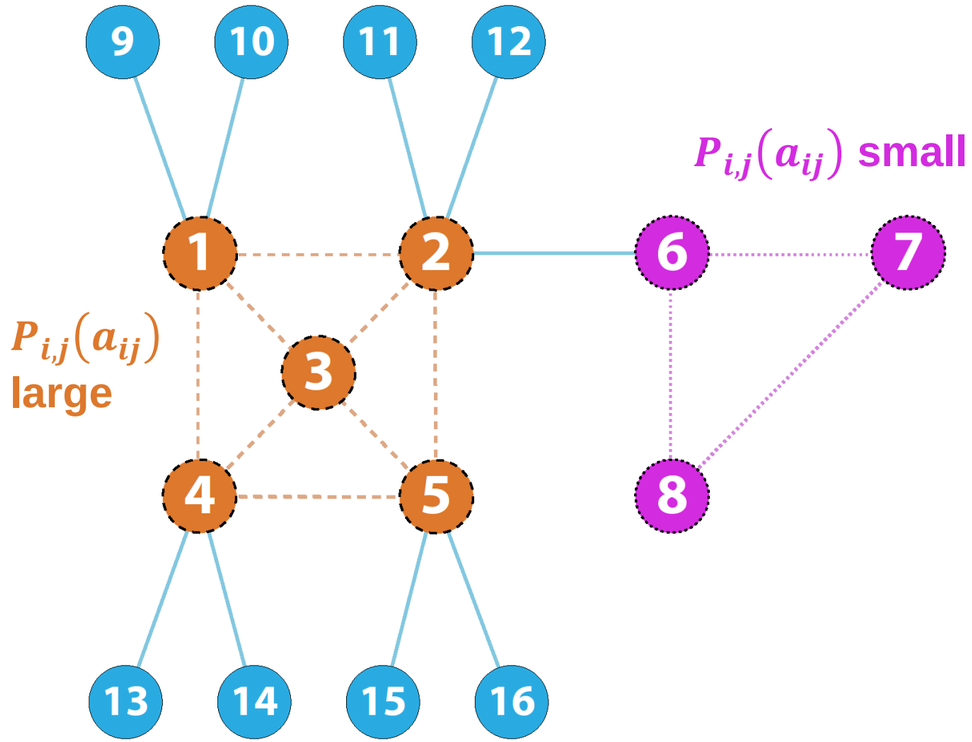
Prior beliefs:

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:

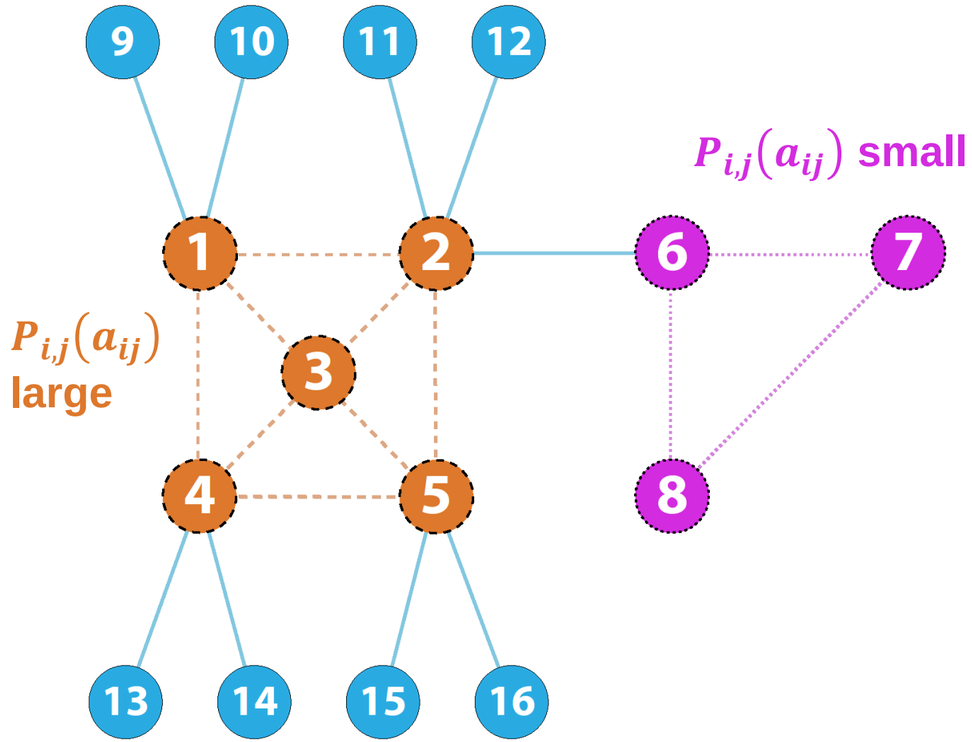
1. Overall density
2. or: Vertex degrees

Pattern:

– Dense subgraphs

Interestingness:

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



GHENT UNIVERSITY

COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

Prior beliefs:

1. Overall density
2. or: Vertex degrees

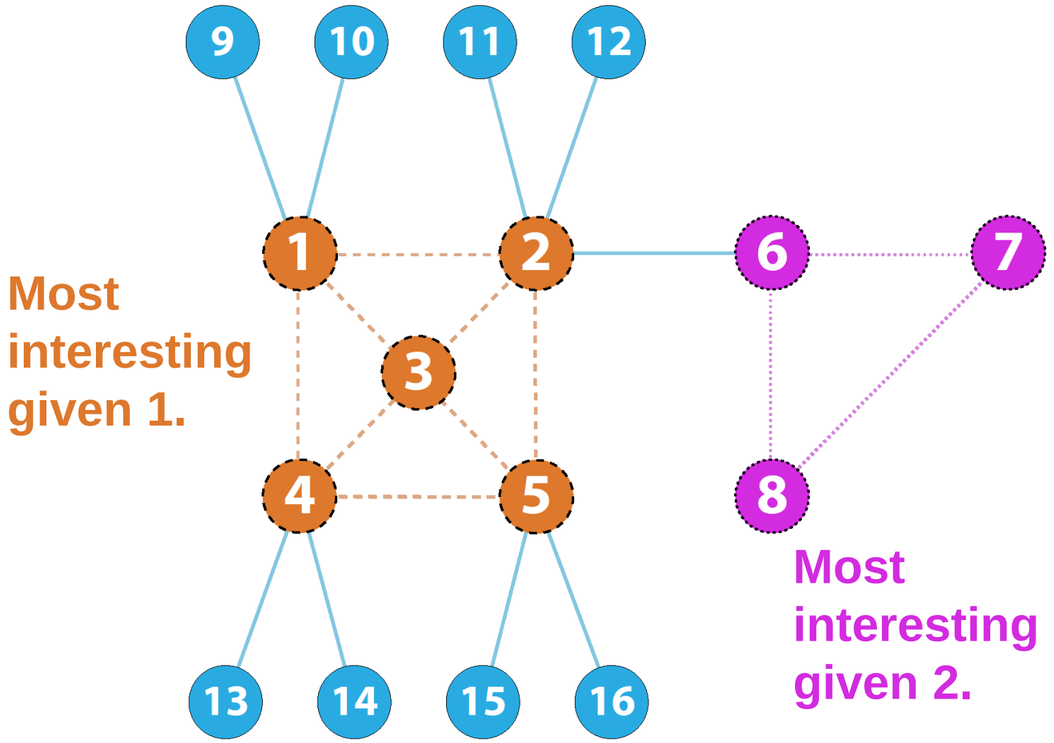
Pattern:

– Dense subgraphs

Interestingness:

– Density vs. size

– 2. ➔ preferably low degree nodes



GHENT UNIVERSITY

COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

Prior beliefs:

1. Overall density
2. or: Vertex degrees

Pattern:

– Dense subgraphs

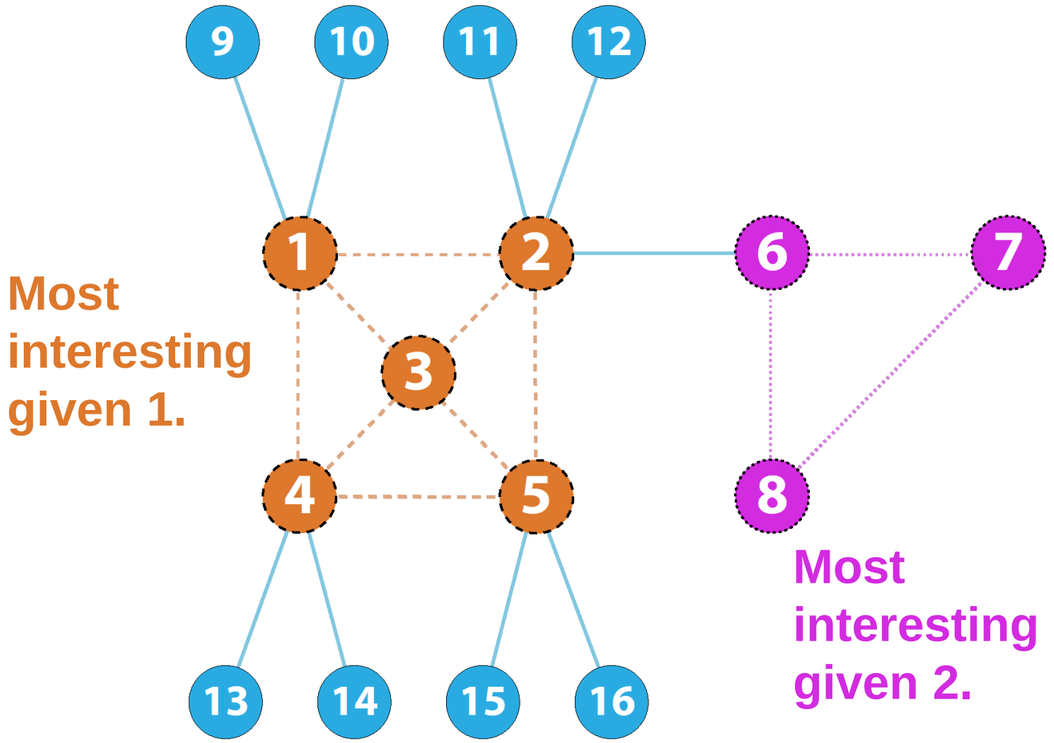
Interestingness:

– Density vs. size

– 2. ➔ preferably low degree nodes

Hill-climbing for search

Update P after each pattern



COMMUNITY DETECTION IN NETWORKS

Data:

– Graph

Prior beliefs:

1. Overall density
2. or: Vertex degrees

Pattern:

– Dense subgraphs

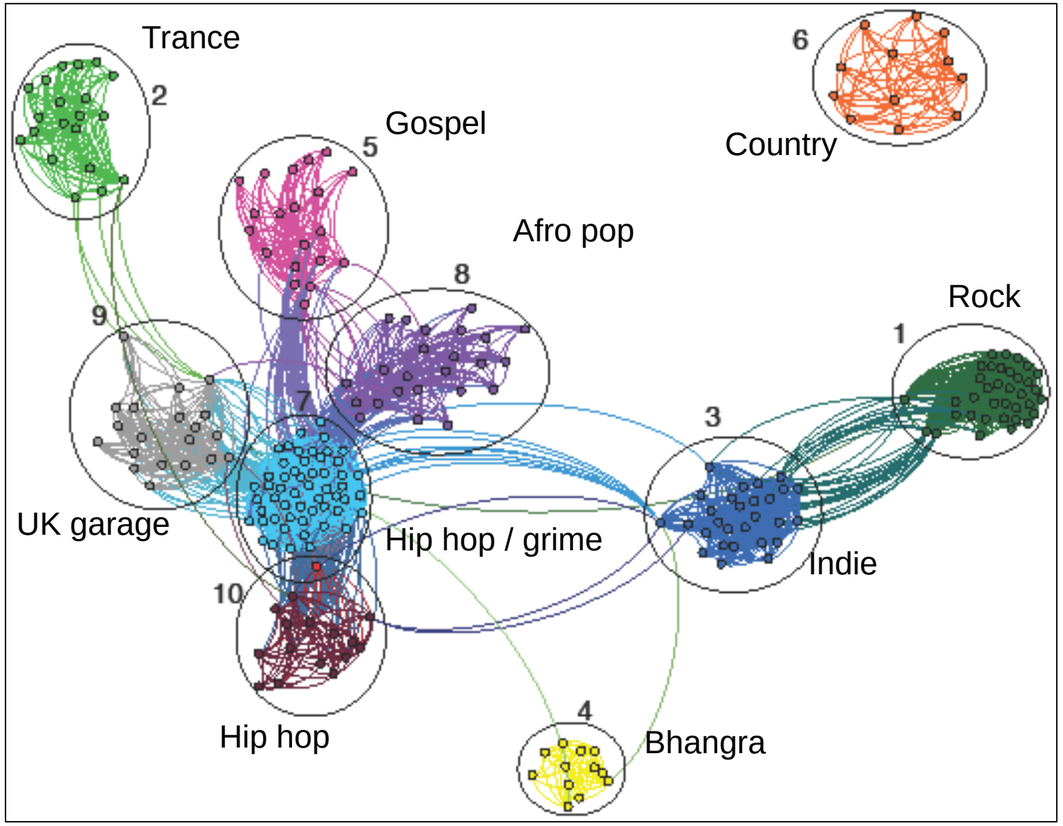
Interestingness:

– Density vs. size

– 2. ➔ preferably low degree nodes

Hill-climbing for search

Update P after each pattern



TAKE-AWAYS

1. What is the **data**?
2. Determine suitable **pattern** syntax
3. What are the **prior beliefs**? (= what is irrelevant to user?)
   - Compute **background distribution P** using maximum entropy
4. Formulate **subjective interestingness**:

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

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

5. Design an **algorithm to optimize** it
6. Find out how to **condition background distribution** on a pattern

GHENT UNIVERSITY

THE BACKGROUND DISTRIBUTION: MAXENT

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$\max_P \sum_{\boldsymbol{A}} -P(\boldsymbol{A}) \cdot \log\big(P(\boldsymbol{A})\big)$$

Entropy

$$\text{s.t.} \sum_{\boldsymbol{A}} P(\boldsymbol{A}) \cdot \left(\sum_{j=1:n} a_{ij} \right) = d_i, \quad \forall i = 1:n$$

Degree

Expected degree constraint

$$\sum_{\boldsymbol{A}} P(\boldsymbol{A}) = 1$$

Normalization

Convex!

Lagrangian:

$$L(P, \boldsymbol{\lambda}, \mu) = \sum_{\boldsymbol{A}} -P(\boldsymbol{A}) \cdot \log\big(P(\boldsymbol{A})\big) + \sum_{i=1:n} \lambda_i \left(\sum_{\boldsymbol{A}} P(\boldsymbol{A}) \cdot \left(\sum_{j=1:n} a_{ij} \right) - d_i \right) + \mu \left(\sum_{\boldsymbol{A}} P(\boldsymbol{A}) - 1 \right)$$

Lagrange multipliers

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$L(P, \boldsymbol{\lambda}, \mu) = \sum_{\boldsymbol{A}} -P(\boldsymbol{A}) \cdot \log\big(P(\boldsymbol{A})\big) + \sum_{i=1:n} \lambda_i \left(\sum_{\boldsymbol{A}} P(\boldsymbol{A}) \cdot \left(\sum_{j=1:n} a_{ij} \right) - d_i \right) + \mu \left(\sum_{\boldsymbol{A}} P(\boldsymbol{A}) - 1 \right)$$

GHENT
UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$L(P, \boldsymbol{\lambda}, \mu) = \sum_{\boldsymbol{A}} -P(\boldsymbol{A}) \cdot \log\big(P(\boldsymbol{A})\big) + \sum_{i=1:n} \lambda_i \left(\sum_{\boldsymbol{A}} P(\boldsymbol{A}) \cdot \left(\sum_{j=1:n} a_{ij} \right) - d_i \right) + \mu \left(\sum_{\boldsymbol{A}} P(\boldsymbol{A}) - 1 \right)$$

Optimality condition:

$$\frac{\partial}{\partial P(\boldsymbol{A})} L(P, \boldsymbol{\lambda}, \mu) = 0$$

$$\frac{\partial}{\partial P(\boldsymbol{A})} L(P, \boldsymbol{\lambda}, \mu) = -\log\big(P(\boldsymbol{A})\big) - 1 + \left(\sum_{i,j=1:n} \lambda_i a_{ij} \right) + \mu = 0$$

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$\frac{\partial}{\partial P(\boldsymbol{A})} L(P, \boldsymbol{\lambda}, \mu) = -\log\big(P(\boldsymbol{A})\big) - 1 + \left(\sum_{i,j=1:n} \lambda_i a_{ij} \right) + \mu = 0$$

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$\frac{\partial}{\partial P(\boldsymbol{A})} L(P, \boldsymbol{\lambda}, \mu) = -\log\big(P(\boldsymbol{A})\big) - 1 + \left(\sum_{i,j=1:n} \lambda_i a_{ij} \right) + \mu = 0$$

So:

$$P(\boldsymbol{A}) = \exp(\mu - 1) \cdot \exp\left(\sum_{i,j=1:n} \lambda_i a_{ij} \right)$$

MAXENT MODEL S.T. DEGREE BELIEFS

$$\frac{\partial}{\partial P(\boldsymbol{A})} L(P, \boldsymbol{\lambda}, \mu) = -\log\big(P(\boldsymbol{A})\big) - 1 + \left(\sum_{i,j=1:n} \lambda_i a_{ij} \right) + \mu = 0$$

So:

$$P(\boldsymbol{A}) = \exp(\mu - 1) \cdot \exp\left(\sum_{i,j=1:n} \lambda_i a_{ij} \right)$$

$$= \frac{1}{Z(\boldsymbol{\lambda})} \cdot \exp\left(\sum_{i>j} (\lambda_i + \lambda_j) a_{ij} \right)$$

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$\frac{\partial}{\partial P(\boldsymbol{A})} L(P, \boldsymbol{\lambda}, \mu) = -\log\big(P(\boldsymbol{A})\big) - 1 + \left(\sum_{i,j=1:n} \lambda_i a_{ij} \right) + \mu = 0$$

So:

$$P(\boldsymbol{A}) = \exp(\mu - 1) \cdot \exp\left(\sum_{i,j=1:n} \lambda_i a_{ij} \right)$$

$$= \frac{1}{Z(\boldsymbol{\lambda})} \cdot \exp\left(\sum_{i>j} (\lambda_i + \lambda_j) a_{ij} \right) = \frac{1}{Z(\boldsymbol{\lambda})} \cdot \prod_{i>j} \exp\big((\lambda_i + \lambda_j) \cdot a_{ij} \big)$$

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$\frac{\partial}{\partial P(\boldsymbol{A})} L(P, \boldsymbol{\lambda}, \mu) = -\log\big(P(\boldsymbol{A})\big) - 1 + \left(\sum_{i,j=1:n} \lambda_i a_{ij} \right) + \mu = 0$$

So:

$$P(\boldsymbol{A}) = \exp(\mu - 1) \cdot \exp\left(\sum_{i,j=1:n} \lambda_i a_{ij} \right)$$

$$= \frac{1}{Z(\boldsymbol{\lambda})} \cdot \exp\left(\sum_{i>j} (\lambda_i + \lambda_j) a_{ij} \right) = \frac{1}{Z(\boldsymbol{\lambda})} \cdot \prod_{i>j} \exp\Big((\lambda_i + \lambda_j) \cdot a_{ij} \Big)$$

$$= \prod_{i>j} \frac{\exp\Big((\lambda_i + \lambda_j) \cdot a_{ij} \Big)}{1 + \exp(\lambda_i + \lambda_j)}$$

$P_{i,j}(a_{ij})$

Product of independent Bernoulli distributions!

Thanks to fact that prior belief constraint is on a (weighted) sum of the a_{ij}

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

To **find optimal values of Lagrange multipliers**, solve the dual:

$$\min_{\boldsymbol{\lambda}} L(P, \boldsymbol{\lambda})$$

where P is given as:

$$P(\boldsymbol{A}) = \prod_{i>j} \frac{\exp\left(\left(\lambda_i + \lambda_j\right) \cdot a_{ij}\right)}{1+\exp\left(\lambda_i + \lambda_j\right)}$$

After some calculations:

$$\min_{\boldsymbol{\lambda}} \sum_{i>j} \log\left(1 + \exp\left(\lambda_i + \lambda_j\right)\right) - \sum_{i=1:n} \lambda_i d_i$$

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$\min_{\boldsymbol{\lambda}} \sum_{i>j} \log\left(1+\exp\left(\lambda_i+\lambda_j\right)\right) - \sum_{i=1:n} \lambda_i d_i$$

GHENT UNIVERSITY

MAXENT MODEL S.T. DEGREE BELIEFS

$$\min_{\lambda} \sum_{i>j} \log\left(1+\exp\left(\lambda_i+\lambda_j\right)\right) - \sum_{i=1:n} \lambda_i d_i$$

Can be solved using **gradient** descent:

$$\frac{\partial}{\partial \lambda_k}\left(\sum_{i>j} \log\left(1+\exp\left(\lambda_i+\lambda_j\right)\right) - \sum_{i=1:n} \lambda_i d_i\right)$$

$$= \sum_{i=1:n} \frac{\exp(\lambda_i+\lambda_k)}{1+\exp(\lambda_i+\lambda_k)} - d_k$$

Expected degree of node k

Required expected degree of node k

Lots of computational speed-ups possible...

GHENT UNIVERSITY

TAKE-AWAYS

- Constraints on the expected value of weighted averages:

$$E_{\boldsymbol{A}\sim P}\left\{\sum_{i,j\in I} f_{ij}a_{ij}\right\} = c$$

  where f_{ij} and c are constants and I is a set of indices,
  lead to convenient **product distributions**

- **Other examples for graphs:**
  - Overall density (trivial)
  - Densities of particular blocks (e.g. block of nodes with same affiliation)
  - Assortativity (approximately)
  - ...

GHENT UNIVERSITY

THE INTERESTINGNESS

GHENT
UNIVERSITY

INFORMATION CONTENT

- **Information content:**

$$\text{InformationContent}(\text{pattern}, P) = -\log\big(P(\text{pattern})\big)$$

- **Pattern:** "number of edges between given set of nodes $W \subseteq V$ is larger than or equal to a specified k_W"
  - A bit tricky...
  - **Cliques as a special case:** "set of nodes $W \subseteq V$ forms a clique". Then:

$$P(\text{pattern}) = \prod_{i>j \in W} P_{i,j}(1)$$

  - So:

$$\text{InformationContent}(\text{pattern}, P) = -\log\big(P(\text{pattern})\big) = -\sum_{i>j \in W} \log\Big(P_{i,j}(1)\Big)$$

  - Larger if $|W|$ is larger and if $P_{i,j}(1)$ for $i, j \in W$ is smaller

GHENT UNIVERSITY

INFORMATION CONTENT

- **Pattern (general case):** “number of edges between given set of nodes $W \subseteq V$ is larger than or equal to a specified k_W”
  - Probability of at least k_W successes in $n_W = \binom{|W|}{2}$ Bernoulli trials?
  - Approximated by:

$$P(\text{pattern}) \approx \exp\left(-n_W \mathbf{KL}\left(\frac{k_W}{n_W} || p_W\right)\right)$$

  - And thus:

$$\text{InformationContent}(\text{pattern}, P) = -\log\big(P(\text{pattern})\big) \approx n_W \mathbf{KL}\left(\frac{k_W}{n_W} || p_W\right)$$

    where p_W is the average probability $P_{i,j}(1)$ for the edges between $i, j \in W$
  - Larger if $|W|$ (and thus n_W) is larger, p_W is smaller, and k_W is larger

DESCRIPTION LENGTH

- For **cliques:**
  - Describe set W
  - A constant (to describe $|W|$)
    plus a linear term in $|W|$ (to describe its elements):

$$\text{DescriptionLength}(\text{pattern}) = \alpha|W| + \beta$$

- For **dense subgraphs:**
  - Constant β also describes k_W

GHENT UNIVERSITY

INTERESTINGNESS

Putting things together:

$$\text{Interestingness}(\text{pattern}, P) = \frac{-\sum_{i>j \in W} \log\left(P_{i,j}(1)\right)}{\alpha|W| + \beta}$$

- A bit more complex for general dense subgraphs:

$$\text{Interestingness}(\text{pattern}, P) \approx \frac{n_W \mathbf{KL}\left(\frac{k_W}{n_W} || p_W\right)}{\alpha|W| + \beta}$$

- Hard to optimize!
  - Exact search for small graphs
  - Effective hill climber for large graphs

GHENT UNIVERSITY

TAKE-AWAYS

- No compromises w.r.t. interestingness
- Often leads to hard search problems
- **Is this intrinsic to subjective interestingness?**

GHENT UNIVERSITY

UPDATING THE BACKGROUND DISTRIBUTION

GHENT UNIVERSITY

UPDATING THE BACKGROUND DISTRIBUTION

Given a pattern,
update the background distribution by conditioning on the pattern

- Easy to do **for cliques W**:
  - Set $P'_{i,j}(a_{ij}) = 1$ for $a_{ij} = 1$ if $i, j \in W$

- Fast to approximate **for (non-clique) dense subgraphs W**:
  - Set $P'_{i,j}(a_{ij}) \propto P_{i,j}(a_{ij}) \cdot \exp(\lambda_W a_{ij})$ if $i, j \in W$ such that the expected density of W is k_W

- Remains a product of Bernoulli's

TAKE-AWAYS

- Updating can be trivial
- Otherwise, often easy to do approximately

GHENT UNIVERSITY