ALGODAN Algorithmic Data Analysis

Overview of Algodan 2010-2012

Esko Ukkonen
Director of CoE

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

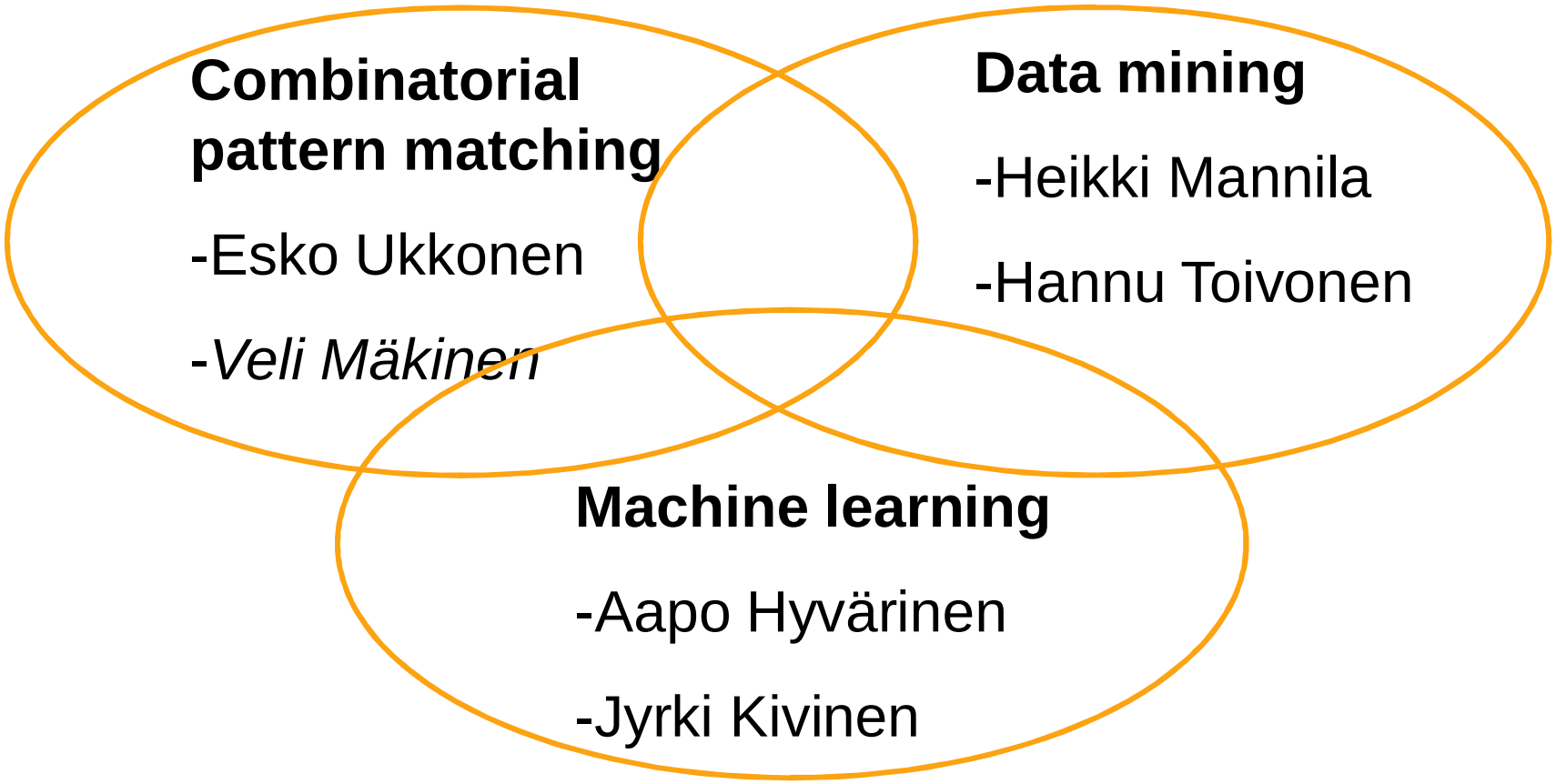
Scientific goals of the centre

- The Algorithmic Data Analysis CoE develops new concepts, algorithms, principles, and frameworks for data analysis.
- The work combines strong basic research in computer science with interdisciplinary work in a variety of scientific disciplines and industrial problems.
- Theory <=> Applications

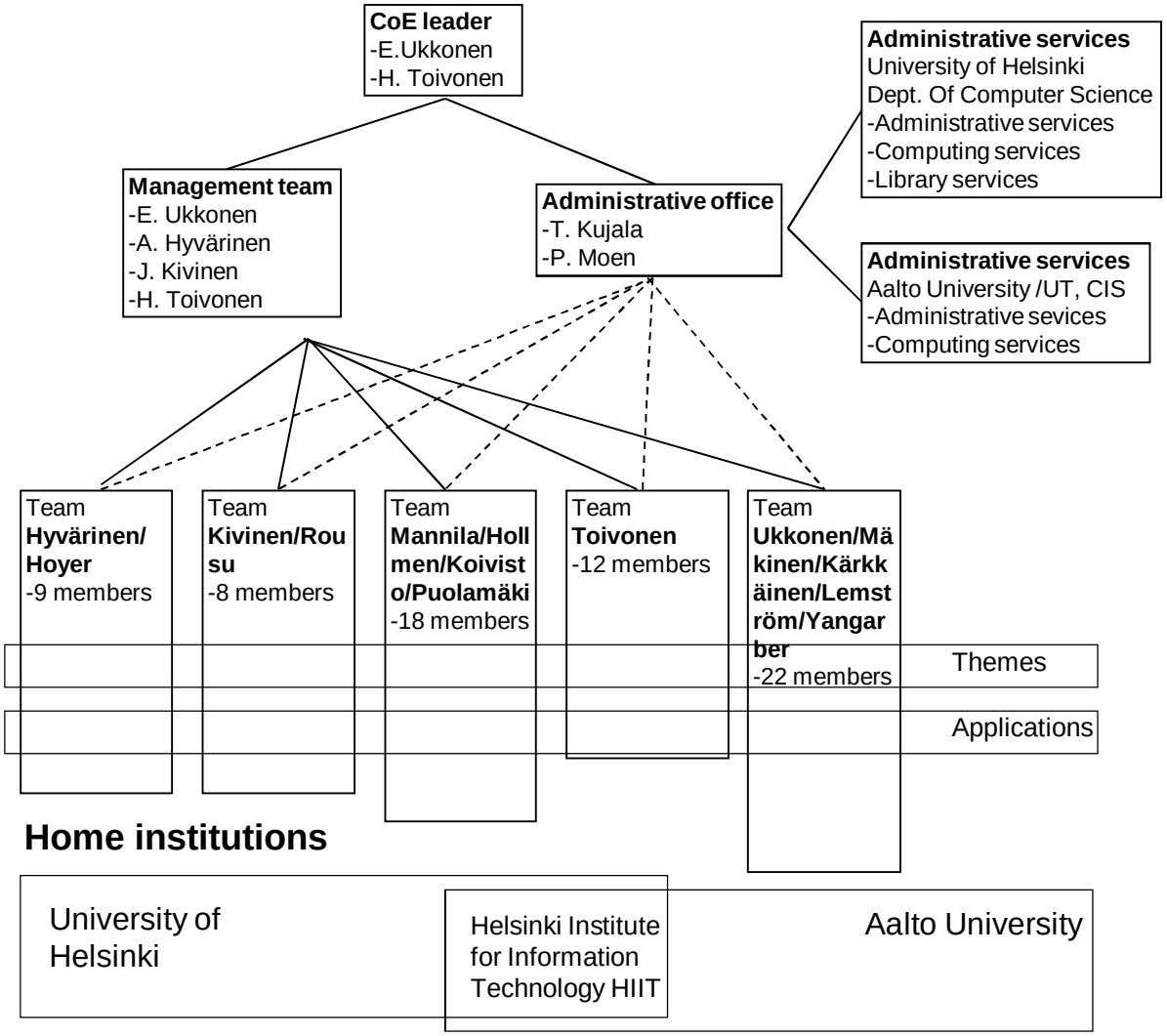
Five research teams (state of 2012)

- **Combinatorial pattern matching**
  - Ukkonen, Mäkinen (-12/2011), Kärkkäinen, Lemström, Yangarber, 4 postdocs, 8 PhD students
- **Data mining: theory and applications**
  - Mannila (- 2/2012), Hollmen, Koivisto, Kaski, Puolamäki, 2 postdocs, 6 PhD students
- **Pattern and link discovery**
  - Toivonen, 1 postdoc , 7 PhD students
- **Machine learning**
  - Kivinen, Rousu, 1 postdocs, 3 PhD students
- **Neuroinformatics**
  - Hyvärinen, Hoyer, 4 postdocs, 3 PhD students
- about 70 persons in total: 58 Univ Helsinki, 12 Aalto Univ

Profile in Computer Science



Organization chart of Algodan



ALGODAN Algorithmic Data Analysis

Funding

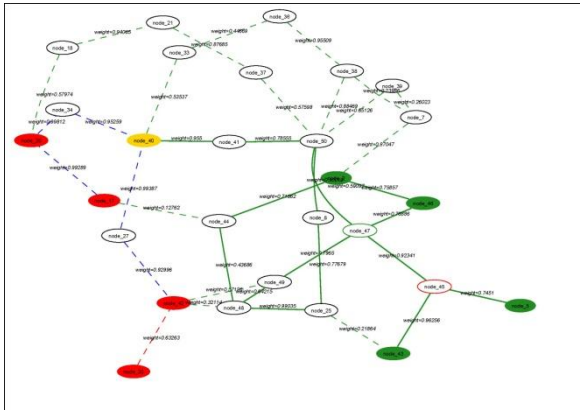
- Basic funding from the Academy of Finland (2010-2013): 520 k€ / year
- Basic funding from the home universities: 300 k€ / year
- Home universities: infrastucture, salaries
- Academy: researcher positions
- Project funding: Academy; TEKES; EU; NIH; private foundations; industry; ...

Four main research themes

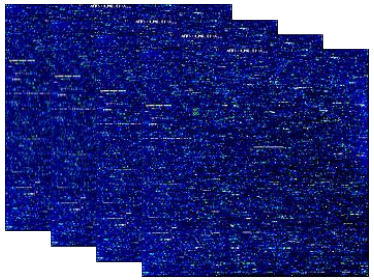
- Sequence analysis (**S**)

cgccgagtgacagagacgctaatcaggctgt
gttctcaggatgcgtac…

- Learning from and mining structured and heterogeneous data (**L**)



- Discovery of hidden structure in high-dimensional data (**D**)



- Foundations of algorithmic data analysis (**F**)

Scientific activity & progress: indicators

	2008	2009	2010	2011	2012	Total
Journal + conf publications + books	40+68	34+46+1	34+75	31+52	13+23	152+264+1
PhD degrees	7	7	5	3	4	26
External funding (incl. Academy) k€	2 038	2 160	2 033	1 691		7 922
Software	?	14				
Foreign personnel	9	19			21	

Algodan seminars

Jan 2008 (kick-off)

Jan 2010

Oct 2011

International visibility

- International conference organization
  - ICML 2008, COLT 2008, UAI 2008, IDA 2010, DS 2011, ALT 2011, SWAT 2012, CPM 2012, IDA 2012, …
- International calls for PhD students and postdocs
- Hirings from abroad: 21 / 70 (~ 30 %)

Researcher career development

- Former Algodan postdocs/PhDs now elsewhere
  - Academia: G Garriga, E Terzi, C Pizzi, F Nicolas, U Köster, E Pitkänen, P Rastas, W Hämäläinen, …
  - Industry: T Mielikäinen, M Kääriäinen, A Gionis, P Sevon, A Rantanen, S Hyvönen, I Autio, J Makkonen, J Lindgren, S Ruosaari, J Tikka, ,J Seppänen, K Laasonen, J Kollin, E Junttila, P Hintsanen, M Lukk,…
- New professors: Veli Mäkinen (2010), Juho Rousu (2012), Petteri Kaski (2012)

	2008	2009	2012
Prof & Senior researcher	13	15 (0 females)	13 (0)
PostDoc	16	19 (2)	12 (2)
PhD student	26	32 (7)	27 (2)
Student	15	20 (5)	18 (3)

Collaboration in computer science

- Longer visits in and out in 2011: about 40 person months
- European Union research consortia:
  - Pattern Analysis, Statistical Modeling and Computational Learning (J Shawe-Taylor, London)
- Numerous collaborations with individual researchers (D Gunopulos (Athens), F Fomin (Bergen), T Husfeldt (Lund), J Nederlöf, S. Szeider, A Apostolico (Padua & Georgia Tech), G Navarro (Chile), P Ferragina (Pisa), G Wiggins (London), C Iliopoulos (Kings C), N Lavrac (Ljubjana), L De Raedt (Leuven), J Shawe-Taylor (UCL), S M Smith (Oxford), F Eberhardt (Canegie Mellon), D Janzig (Max Planck Instit), S Ishii (Kyoto), …)

Collaboration in applications: Bioinformatics, neuroinformatics

- International & European Union projects
  - EU-Project: Systems biology of colorectal cancer (J Taipale)
  - European Bioinformatics Institute, UK: Dr Alvis Brazma
  - Center for Neurobehavioral Genetics at the University of California Los Angeles (UCLA)
  - S Luyssaert & I Janssens, Univ Antwerp (carbon balance)
- University of Helsinki:
  - CoE on Translational Genome-Scale Biology: J Taipale, L Aaltonen
  - CoE in Microbial Food Safety (A Palva)
  - prof Sakari Knuutila (genetics), prof Liisa Holm (bioinformatics), prof. A Urtti (pharmacology), P Hari & E Nikinmaa (forestry)
  - Institute for Molecular Medicine in Finland (FIMM) and National Institute of Health and Welfare (THL)
- Aalto University
  - CoE on systems neuroscience and neuroimaging (Riitta Hari, S Vanni)
- VTT Biotechnology:
  - prof H Söderlund, prof M Penttilä (CoE)

Collaboration in applications: Environmental research

- University of Helsinki:
  - CoE on Metapopulation research: prof I Hanski
  - CoE on Physics, Chemistry and Biology of Atmospheric Composition and Climate Change: prof M Kulmala
  - CoE on Developmental Biology: prof. M Fortelius, prof. J Jernvall
  - ESO project with astronomers: prof. K Mattila

Collaboration in applications: Linguistics and language technology

- University of Helsinki
  - CoE on Language Variation and Changes: prof T Nevalainen
  - Univ Helsinki: prof. K Koskenniemi (computer linguistics), L Carlson (computer linguistics)

- Research Institute for the Languages of Finland:
  - prof R-L Pitkänen

- European Commission's Joint Research Centre (JRC, Ispra), EC Frontex Agency , Global Health Security Initiative (GHSI), European Center for Disease Control (ECDC) , Russian Academy of Sciences

ALGODAN Algorithmic Data Analysis

Influence Attribution in Citation Networks

Panagiotis Papapetrou

Postdoctoral Researcher

Department of Information and Computer Science

Aalto University

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Problem description

- People always intrigued by characterizing influential ideas, books, scientists, politicians, etc
- Main question: who is influential?
- Examples:

  who are the most influential scientists?

  which actors influence a movie rating the most?

Our setting

- Individuals accomplish tasks in a collaborative manner

- **Influence attribution:** each individual is assigned with a score based on performance

Instantiation: author-publication

- Individual == author
- Task == publication
- Impact score:

  CC: citation count of the publication

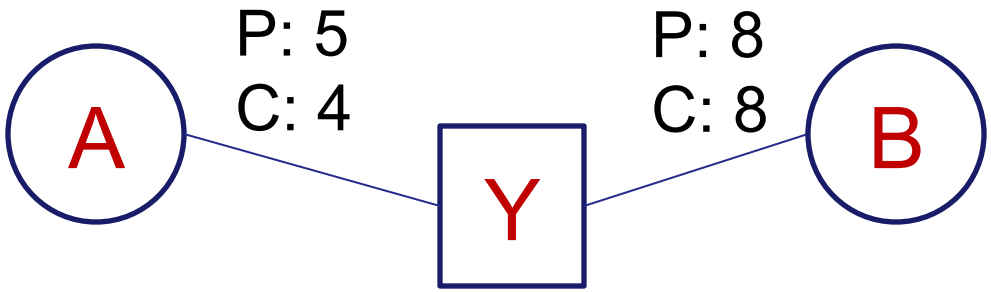
  PR: PageRank score of the publication

Example: author-publication

- Two researchers A and B
- Question: who is more influential?

Example: author-publication

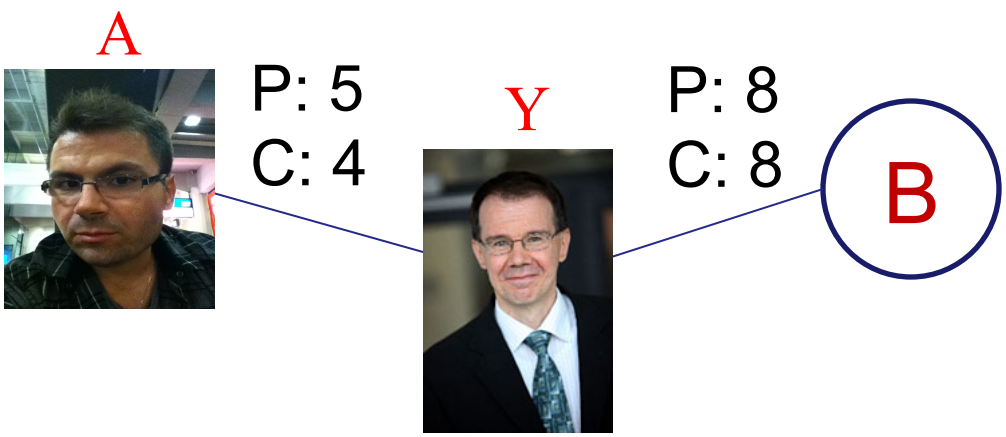
- One common collaborator: Y



P: number of papers
C: number of citations per paper

Example: author-publication

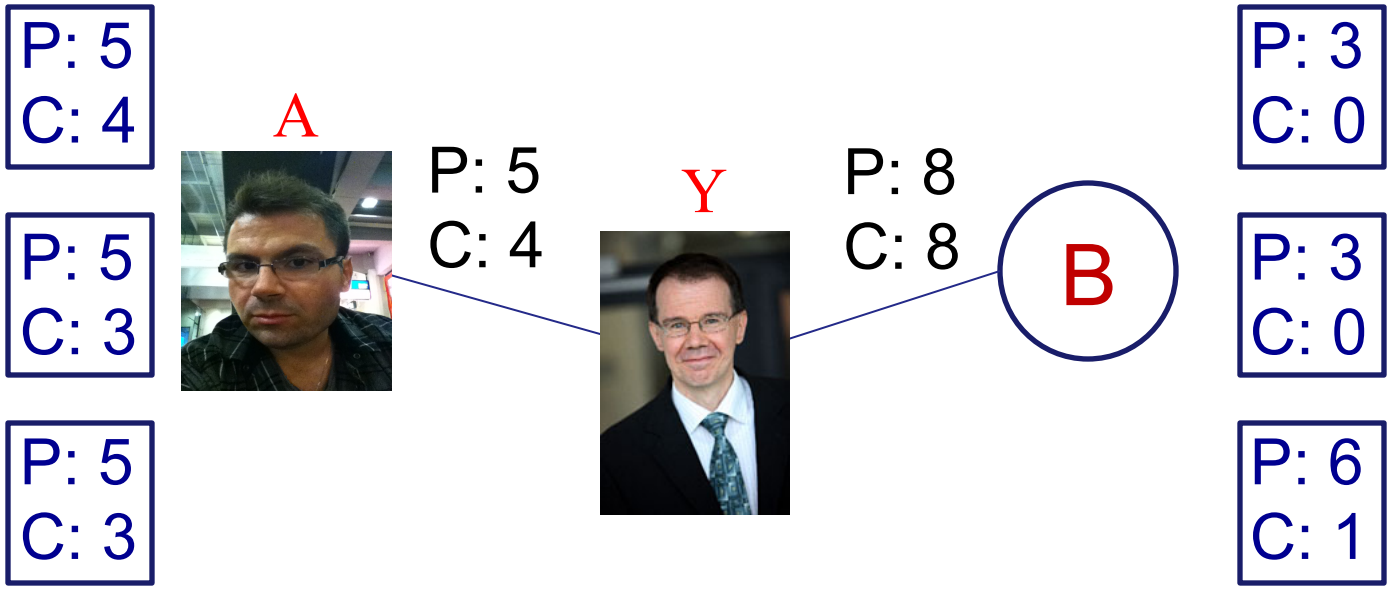
- One common collaborator: Y



P: number of papers
C: number of citations per paper

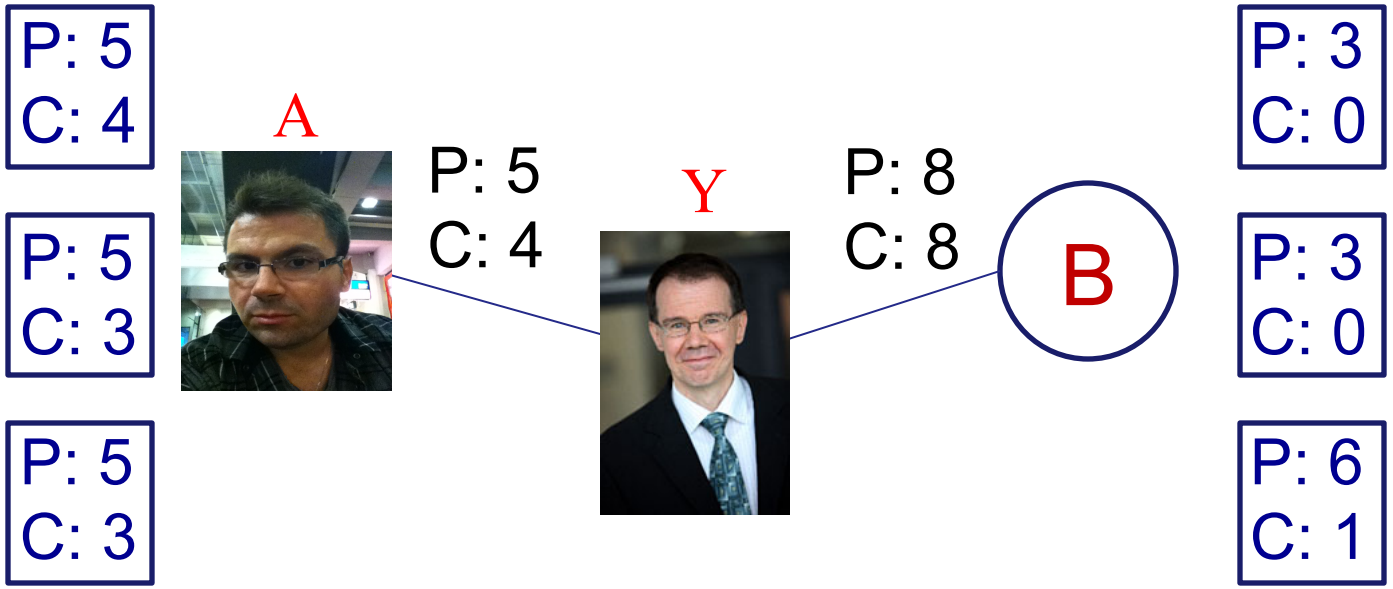
Example: author-publication

- Three additional collaborators for A and B



Example: author-publication

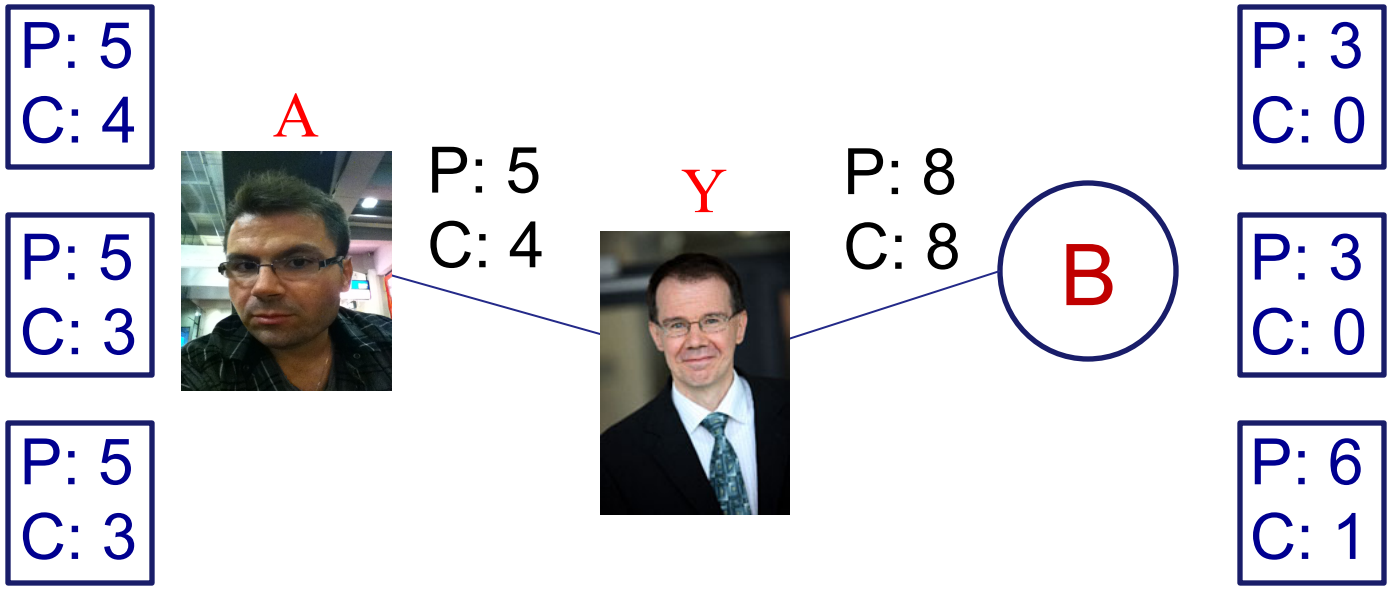
- Three additional collaborators for A and B



Researcher	Papers	Citations	H-index
A	20	70	4
B	20	70	8

Example: author-publication

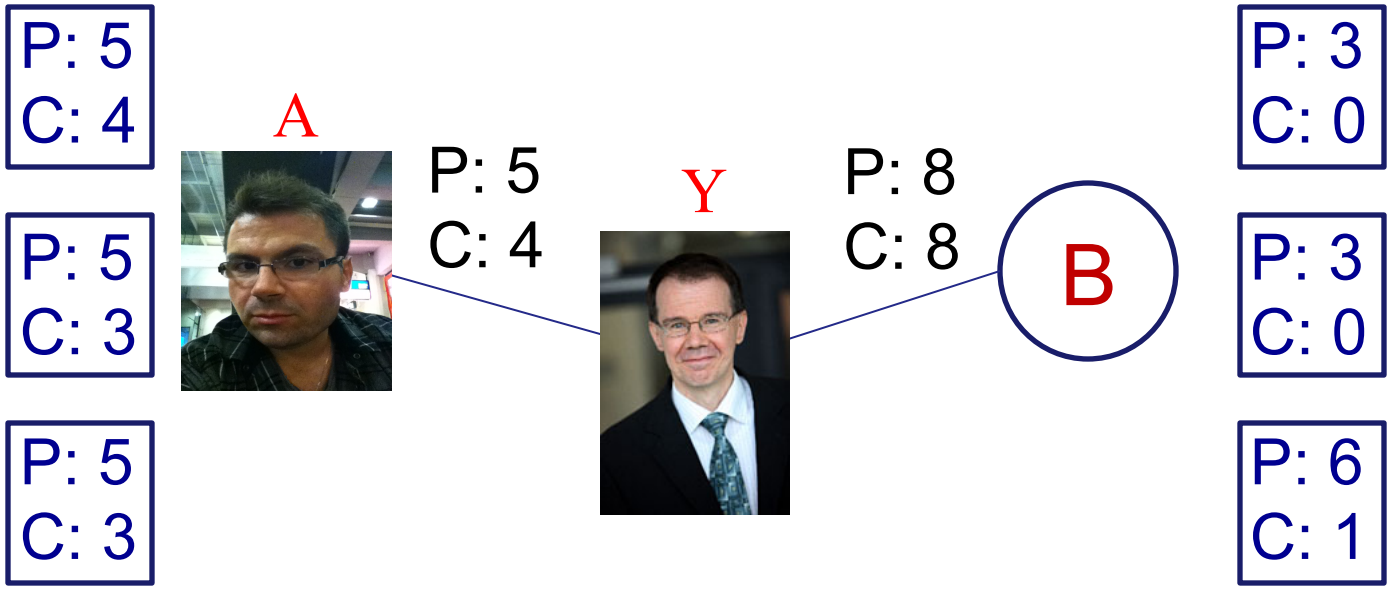
- Three additional collaborators for A and B



Researcher	Papers	Citations	H-index
A	20	70	4
B	20	70	8

Example: author-publication

- Three additional collaborators for A and B



But is B indeed that influential?

Or is B just being favored due to the fame of Y?

Example: author-publication

- Drop Y out of the picture

The performance of A remains quite high

The performance of B is weakened a lot

Example: author-publication

- Drop Y out of the picture

Researcher	Papers	Citations	H-index
A	15	50	4
B	12	6	1

Observation

- Most existing bibliometrics indicators are
  - based on the *publication* or *citation* count
  - simple to compute
- Ignore the underlying structure of
  - the citation graph
  - the co-authorship graph

Solution

- For each individual compute:

  what difference does an individual make to the coalition if dropped from it

- Individuals who form many strong coalitions are favored against those who form weak ones

Shapley value

- V: set of individuals
- u: gain function
- Φ: Shapley value

  sum of marginal gains contributed by each individual to a coalition (subsets of V)

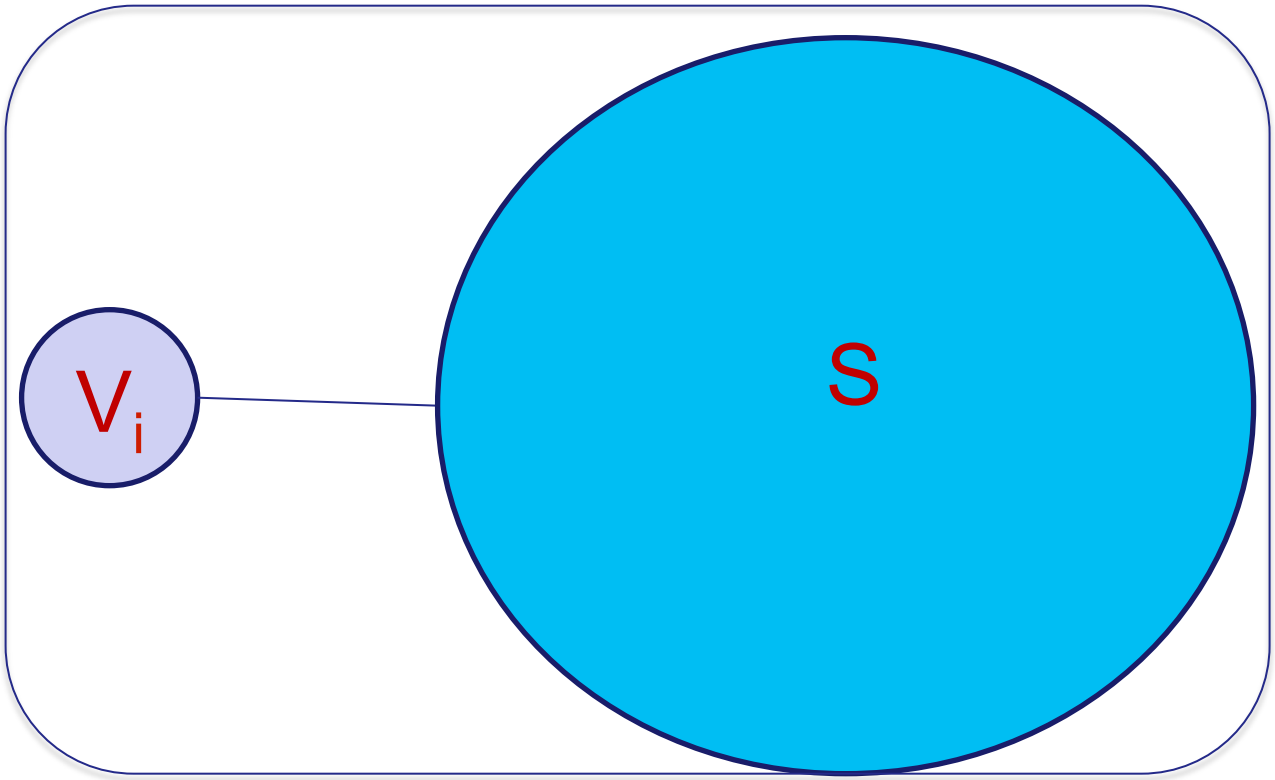
$$\phi_i(v) = \sum_{\mathcal{S} \subseteq \mathcal{V}} \frac{|\mathcal{S}|!(|\mathcal{V}| - |\mathcal{S}| - 1)!}{|\mathcal{V}|!} (v(\mathcal{S} \cup \{V_i\}) - v(\mathcal{S}))$$

Our approach

- Not all coalitions may be available or defined
- We compute the marginal gains *u(S)* by averaging only over **coalitions for which impact scores are available**
- For the author-publication case: iterate over all papers
- We approximate the rest

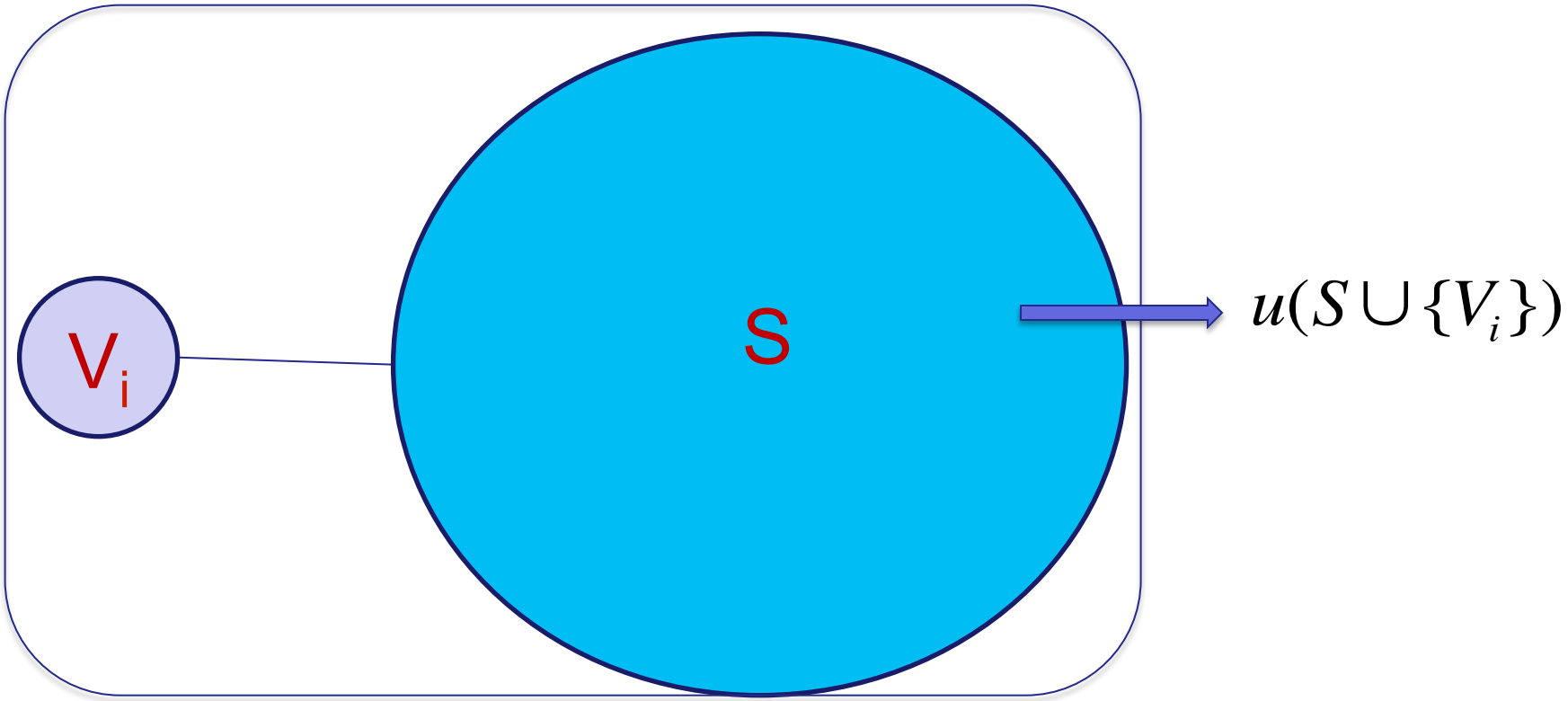
Iterative method

- We choose to take into account all cases for which $S \cup \{V_i\}$ is available



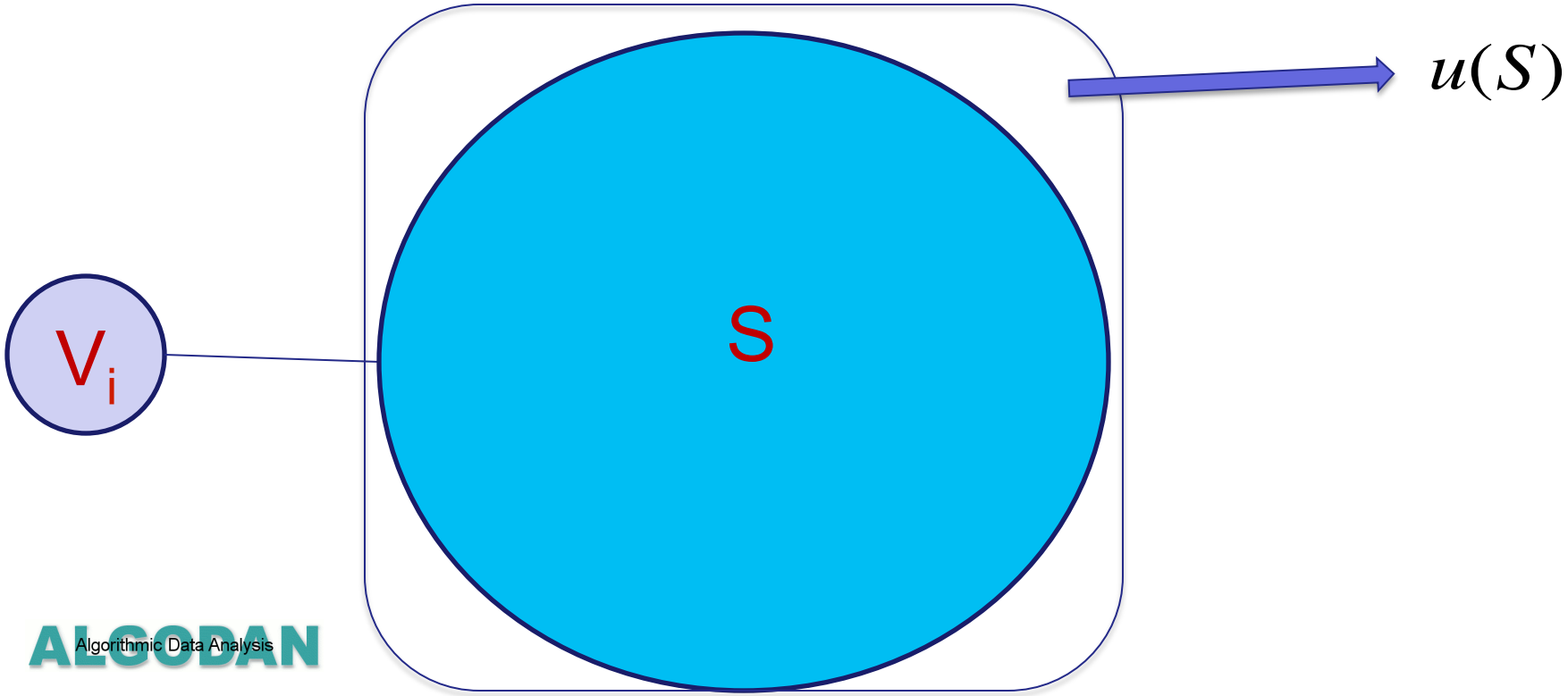
Iterative method

- We choose to take into account all cases for which $S \cup \{V_i\}$ is available



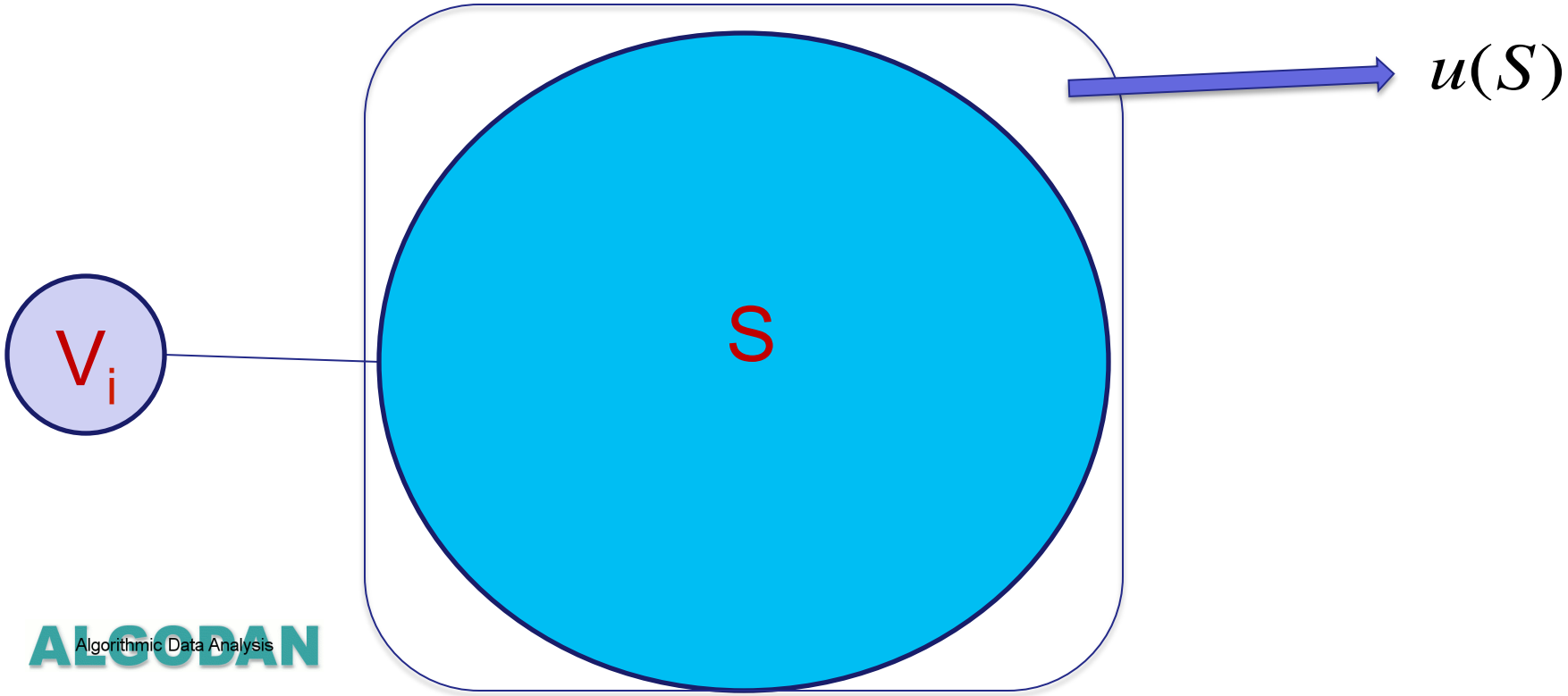
Iterative method

- Then compute the gain of S



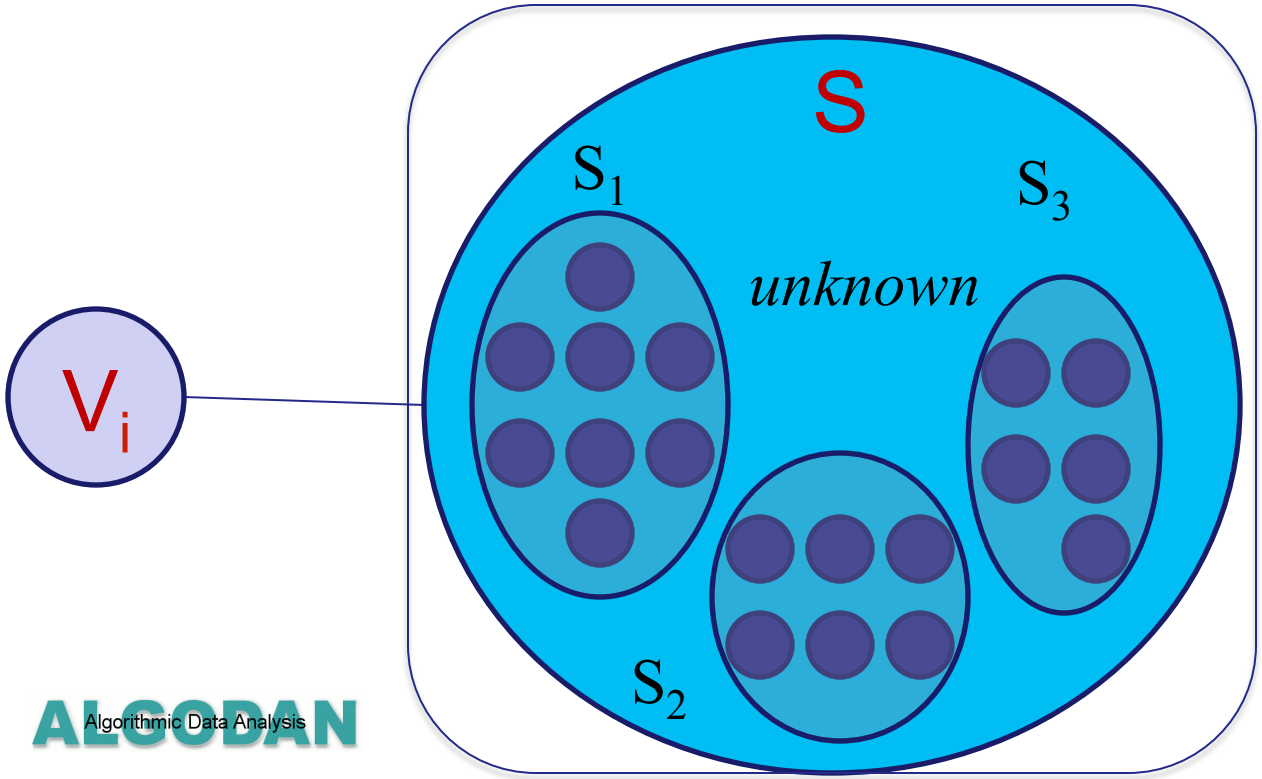
Iterative method

- What if for some set S we have no complete information about the coalitions?



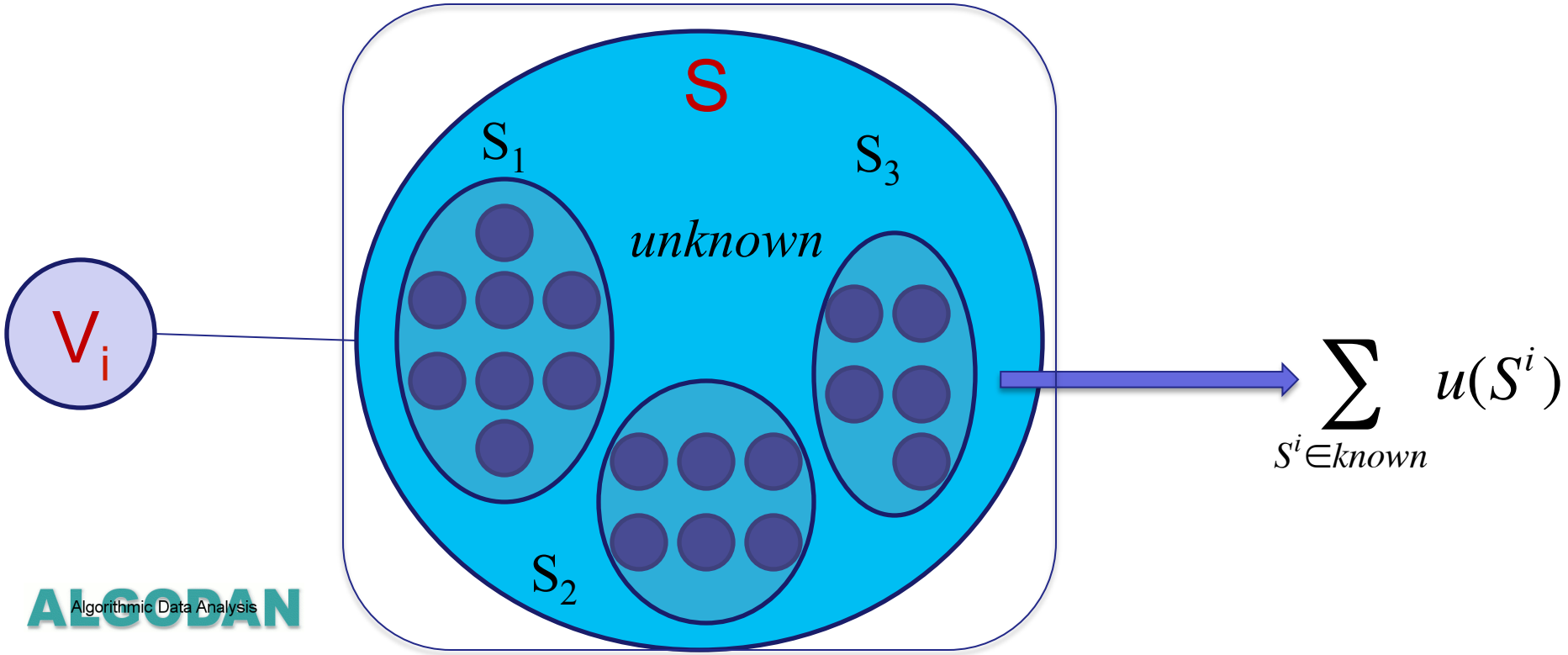
Iterative method

- There should be some “known” coalitions inside S
- The rest are considered “unknown”



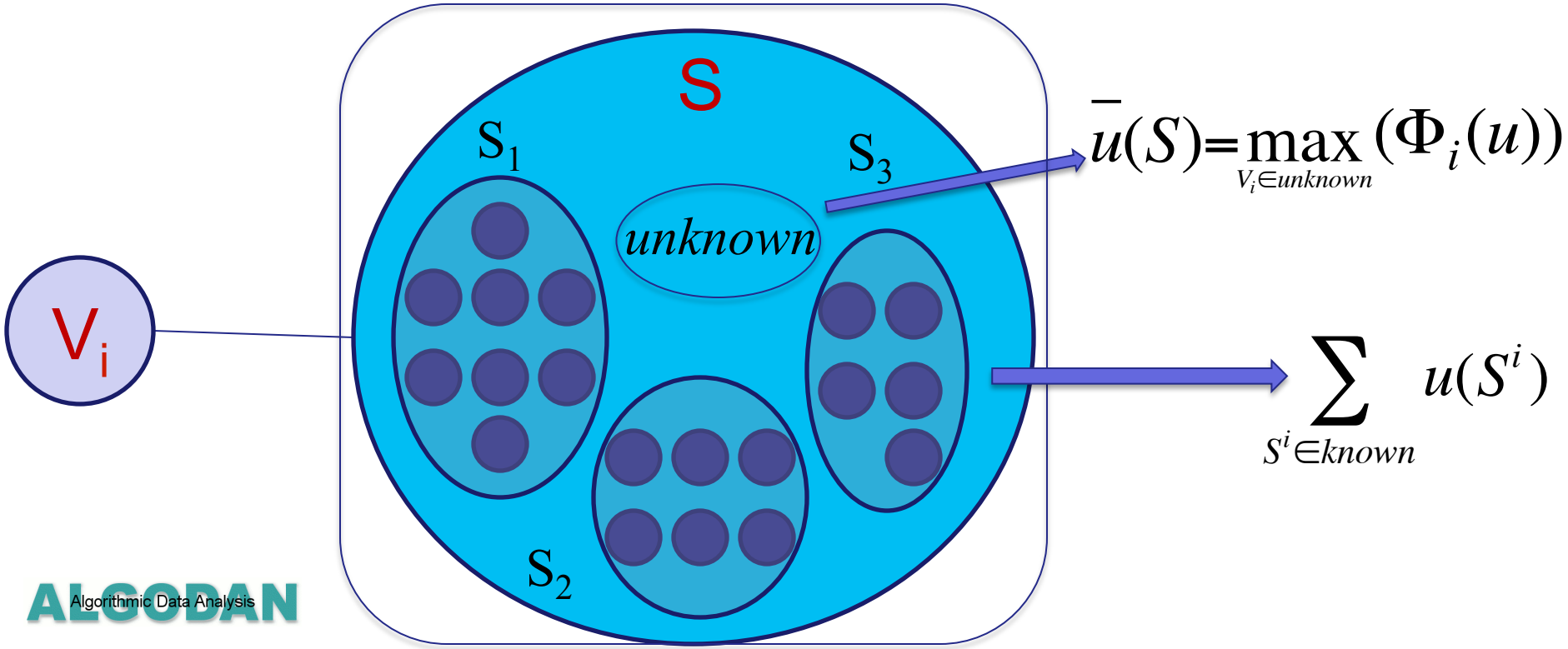
Iterative method

- Compute the gain function for the “known” coalitions



Iterative method

- Approximate the “unknown” coalitions



Monotonicity requirement

- Monotonicity of the gain function
  - bigger coalitions should have higher impacts
  - not always the case: e.g., author-publications
- We impose it using a heuristic

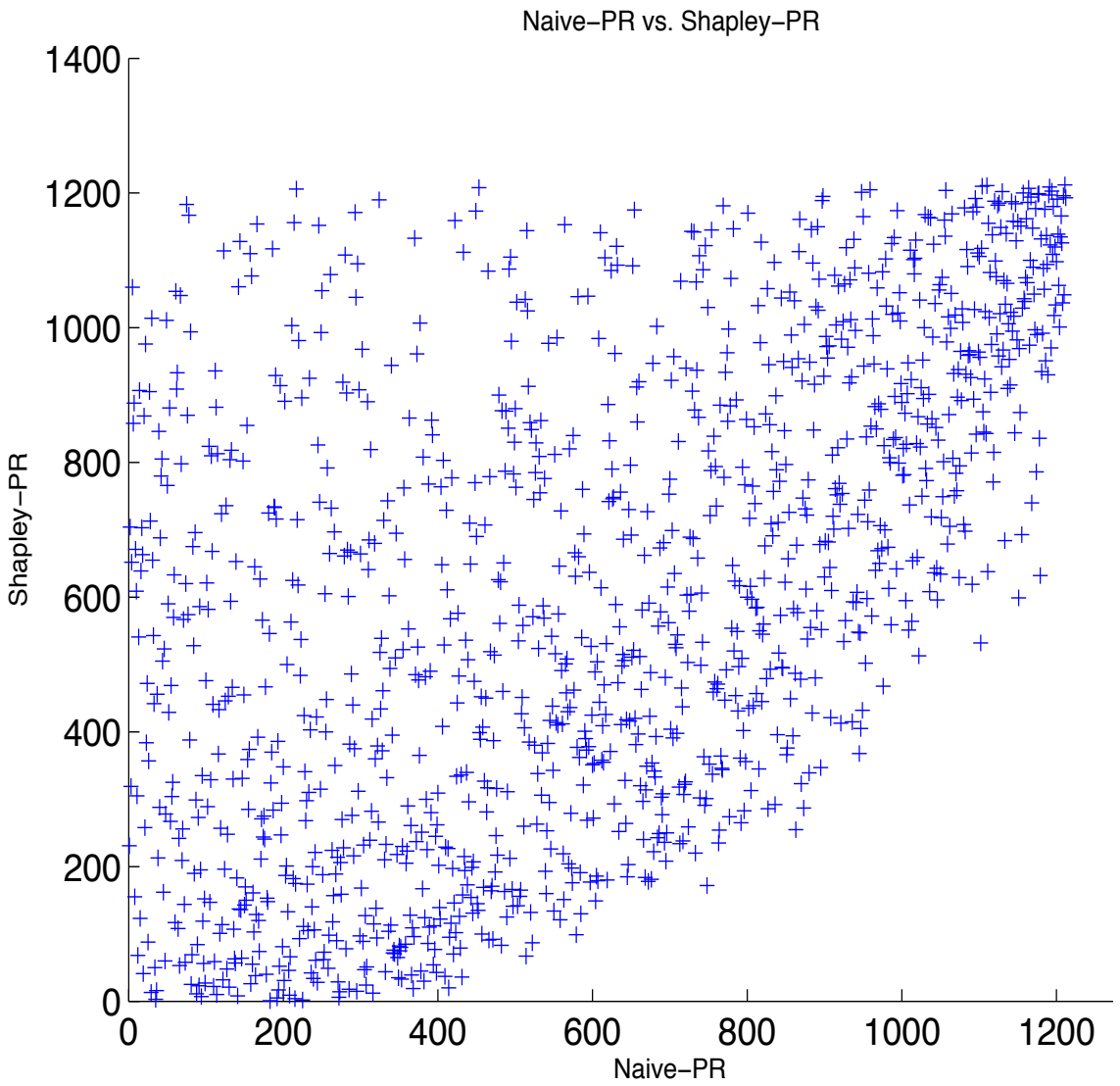
Datasets

- ISI Web of Science:
  - Publication years 2003 and 2009
  - ICS Aalto University & Yahoo! Research, Barcelona
  - 1212 authors and 4506 publications
- Internet Movie Database:
  - 2000 male actors and 4560 comedy/action movies

ISI Web of Science

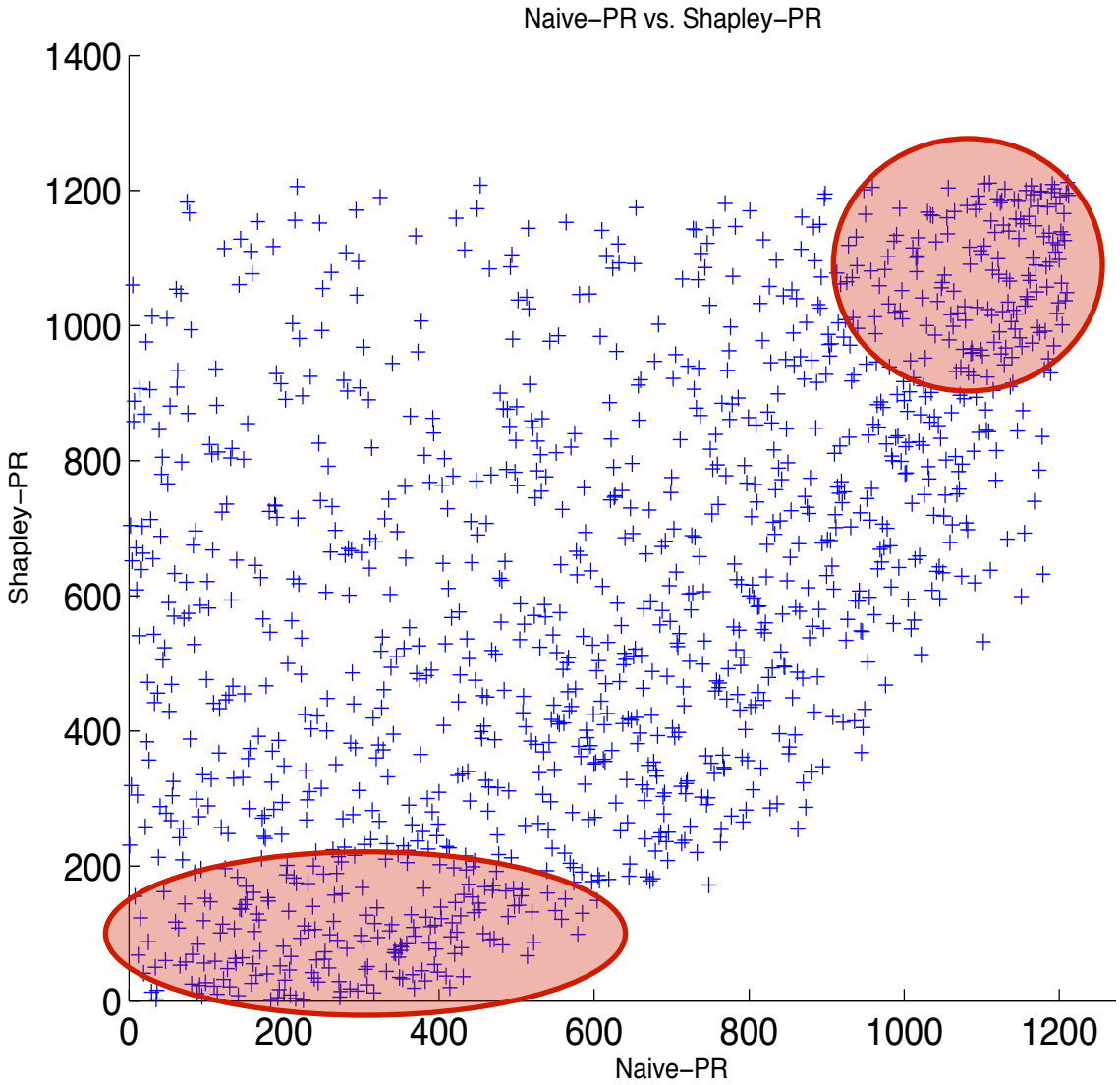
Naïve-PR

Distribute citation scores **equally** to all co-authors



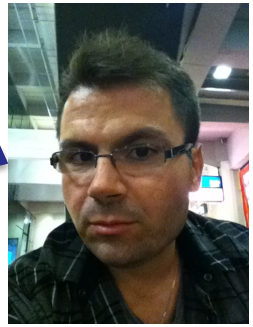
ISI Web of Science

Shapley-PR	Naïve-PR
1	183
2	225
3	35
4	215
5	192
6	272
7	94
8	141
9	208
10	114



Proof of concept

Shapley-PR	Naïve-PR
1	183
2	225
3	35
4	215
5	192
6	272
7	94
8	141
9	208
10	114



Panagiotis Papapetrou, Aris Gionis, and Heikki Mannila, "A Shapley value Approach for Influence Attribution" *ECML-PKDD 2011*

Future work

- Monotonicity property
- Investigation of other domains such as:
  - —user-blogs
  - —social media sites
- How additional information about the individuals can affect/be taken into account
- Further evaluate the quality of the obtained rankings by performing user studies

ALGODAN Algorithmic Data Analysis

A statistical significance testing approach to mining the most informative set of patterns

Jefrey Lijffijt

Doctoral Student @ ICS Department, Aalto University

Joint work with: Panagiotis Papapetrou, Kai Puolamäki

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Introduction

- Early pattern mining: focus on efficient enumeration
  - Apriori / Level-wise algorithm (Agrawal et al. 1994, Mannila et al. 1994), ECLAT (Zaki et al. 1997), FP-GROWTH (Han et al. 2000), Etc.
  - Result: zillions of patterns
- Next approach: condensed representations
  - Maximal patterns (Bayardo Jr. 1998), Closed patterns (Pasquier et al. 1999), Non-derivable patterns (Calders et al. 2002), Significant patterns (Webb 2007, Gionis et al. 2007), Etc.
  - Still too many and redundancy in pattern set

Introduction

- AB, AC, AD, AE, BC, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE
- These could all be significant and non-derivable
- Can all these subsets be explained by ABCDE?

Non-Redundant Sets of Patterns

- Approach 1: Local objective function / heuristics
- Mining *top-k patterns* using ranking / iterative mining (Mielikäinen & Mannila 2003, Bringmann & Zimmermann 2007, Gallo et al. 2007, Hanhijärvi et al. 2009)
- Approach 2: Global objective function
- 'Intuitive' quality measures (Knobbe & Ho 2006), Krimp (Siebes et al. 2006), MaxEnt (Kontonasios & De Bie 2010)

Our Approach (Lijffijt et al. *submitted*)

1. Define null hypothesis
   - What we currently know about the data
2. Choose test statistic
   - Patterns that we want to explain
3. Choose possible constraints
   - Patterns that we want to find

- Algorithmic challenge: find *k* constraints that maximize the p-value of the data
  - P-value $p = \Pr(T_{null} \geq T_{data})$

1. Null Hypothesis

- What you currently know about the data
  - I.e., what is not interesting to find
- Examples of used null hypothesis
  - Data matrix: retain marginal distributions
    - Binary: exact, expectation
    - Real-valued: limit on Kolmogorov-Smirnov statistic
  - Graph data: retain degree distributions, connected components
  - Time series: retain power spectrum
- More advances hypothesis: cluster structure, etc.

Citations omitted here for brevity

3. Constraints

- The patterns that we allow as output
- Itemsets (see example later)
- Clusters / Segments / Tiles
  - Group of objects that are similar (tiles: for restricted set of vars)
- Constraint correspond to enforcing some statistic over these objects
  - Frequency of an itemset
  - Distance between a set of objects

2. Test Statistic

- The patterns that we want to explain
- Itemsets
  - Sum over statistic of some chosen set of itemsets
    - See example later
- Clusters / Segments / Tiles
  - Total clustering/segmentation cost
  - Description length of the data
- Often directly related to the constraints

Our Approach (Reminder)

1. Define null hypothesis
2. Choose test statistic

- 1 and 2 give a p-value for the data $p = \Pr(T_{null} \geq T_{data})$

3. Choose possible constraints

- Algorithmic challenge: find *k* constraints that maximize the p-value of the data

Complexity

- NP-Hard in general (as are *Krimp* and *MaxEnt*)
- There can exist no general fixed-ratio approximation scheme
- Greedy algorithm:
  1. Select constraint that maximizes the p-value of the data
  2. In case of ties, prefer higher test statistic
- Greedy algorithm is optimal if constraints are *independent*
- Greedy algorithm has approximation ratio if constraints are *approximately independent*

Application Example: Mining Itemsets with High Lift (1/3)

- *Paleo* data set (Fortelius 2005, Puolamäki et al. 2006)
  - Genus of fossils in Europe and Asia
  - 124 sites, 139 genus
- Null model: uniform distribution over all binary matrices with same row and column margins
- Constraints: itemsets with $support \geq 0.1$ and $lift \geq 1$
  - 118 possible constraints
- Test statistic $T(\omega) = \sum_{i=1}^{118} lift(X_i, \omega)$

Application Example: Mining Itemsets with High Lift (2/3)

- NB. We should be able to compute a p-value for the data under any combination of constraints
  - Preferably analytically
  - If not → randomization
  - 'Only' a constant factor slower

- Here: randomize data with *itemset-swap* algorithm (Hanhijärvi et al. 2009)

Application Example: Mining Itemsets with High Lift (3/3)

- Result: initial $\hat{p} = 10^{-299.94}$
- P-value rapidly increases with first 4 itemsets $\hat{p} = 10^{-42.75}$
- P-value is maximal after 17 itemsets $\hat{p} = 10^{-27.98}$
  - Low p-value due to maintaining frequency approximately
- Set of patterns does not simply contain the patterns with highest lift

➢ Redundancy is well accounted for

Conclusions

- Other application examples in the paper
  - Clustering, segmentation
- Novel general approach to mining a small set of *interesting* and *non-redundant* patterns
- Open questions include:
  - Null models that lead to analytical p-values (such as *MaxEnt*)
  - What if we mix various types of constraints (itemsets, clusters)
    - Should we weight constraints according to complexity?

ALGODAN Algorithmic Data Analysis

On Analyzing Environment: Natural and Man-Made

Jaakko Hollmén
Parsimonious Modelling
Aalto University School of Science

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Parsimonious Modelling

- Monitoring and diagnosis of man-made **engineered structures**, such as bridges
- Analysis and monitoring of the **natural environment**: climate studies in the context of forest and tree growth
- Parsimonious modeling aims at learning *simple, compact, or sparse models* from data
- Showcases: learn parsimonious models from data in environmental informatics area

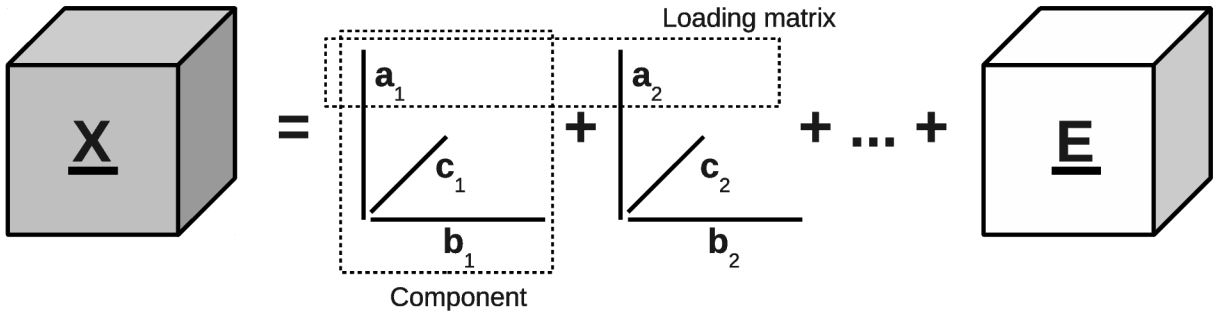
Recent Research and Publications

- **Three-way analysis of structural health monitoring data.** Miguel A. Prada, Janne Toivola, Jyrki Kullaa, Jaakko Hollmén. Neurocomputing, April 2012.
- **Collaborative filtering for coordinated monitoring in wireless sensor networks.** Janne Toivola and Jaakko Hollmén. In Proceedings of International Conference on Data Mining Workshops, December 2011. poster!
- **Environmental proxy selection in temperature reconstruction.** Mikko Korpela et al., Manuscript in preparation. poster!

Three-way analysis of structural health monitoring data

- Monitoring and diagnosis of man-made structures by modeling the vibration profile
- Invariant features: ratios of frequency amplitudes of sensors s_i and s_j
- Represent data as tensors: <sensors, frequencies, time>
- Matrix factorization, or signal decomposition methods
- Model: $\mathbf{X} = \mathbf{\Sigma}_{f=1..F}\ \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f + \mathbf{E}$

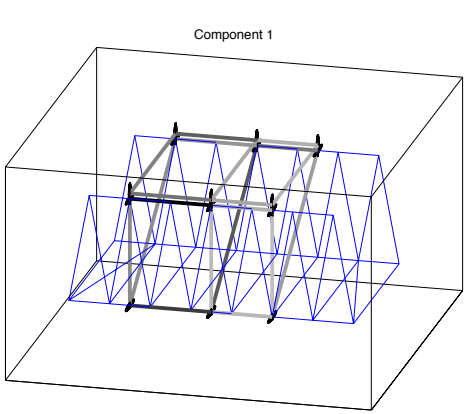
Three-way analysis of structural health monitoring data

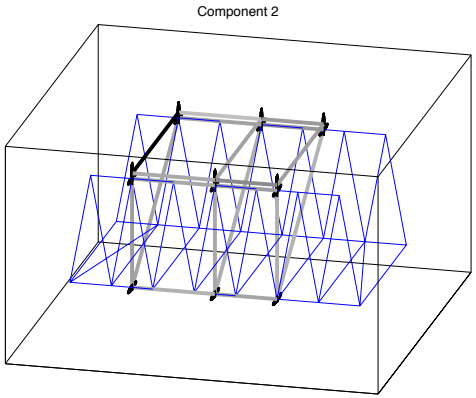


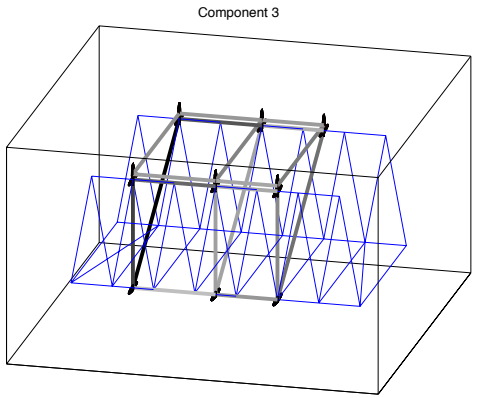
- Model: $\mathbf{X} = \Sigma_{f=1..F}\ \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f + \mathbf{E}$

- Alternating least squares estimation (PARAFAC)
- Projection of the data onto the factors, here, time

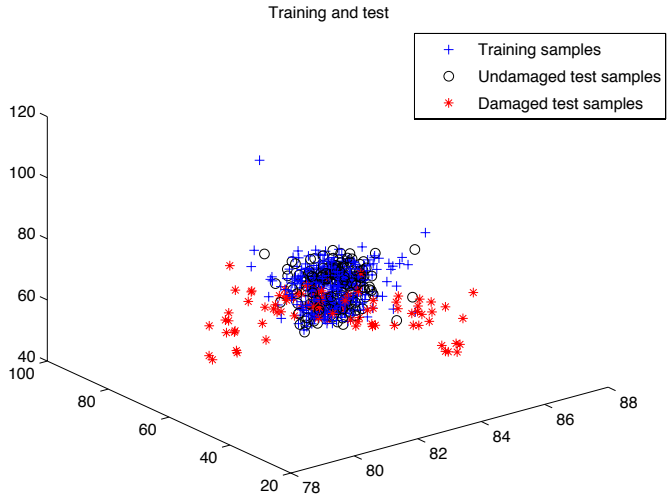
Interpretability of the results





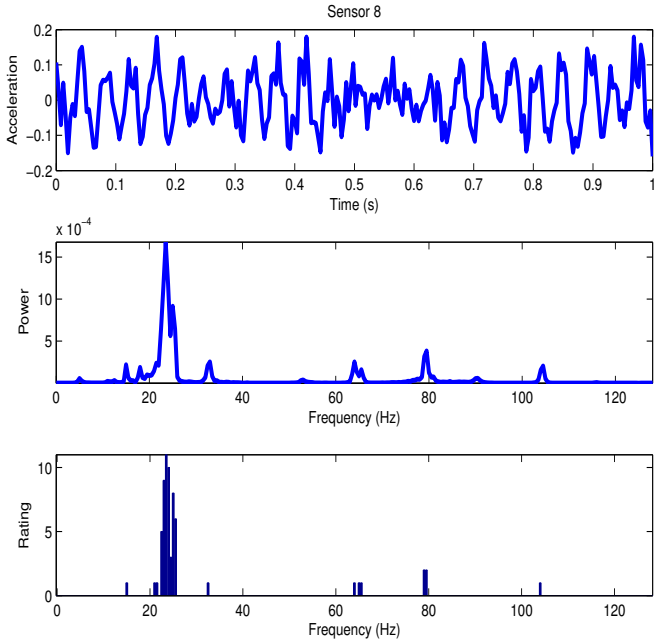


- Resulting three factors in terms of sensor pairs (up)
- Novelty detection (right)

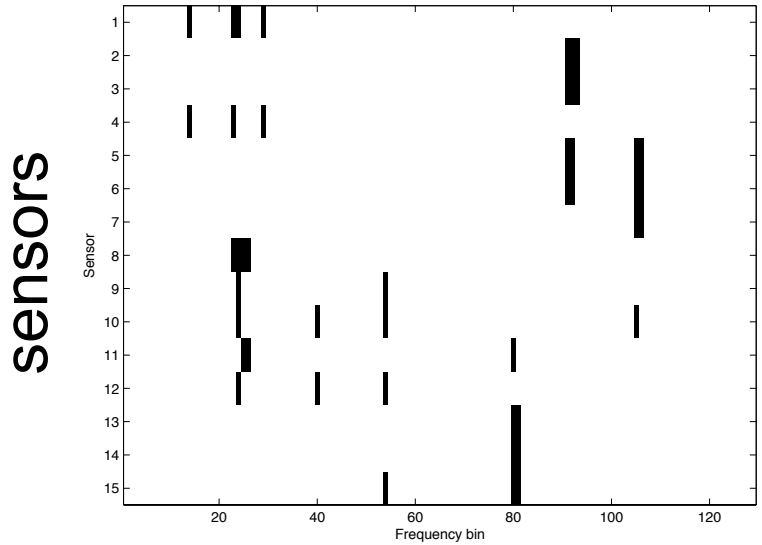


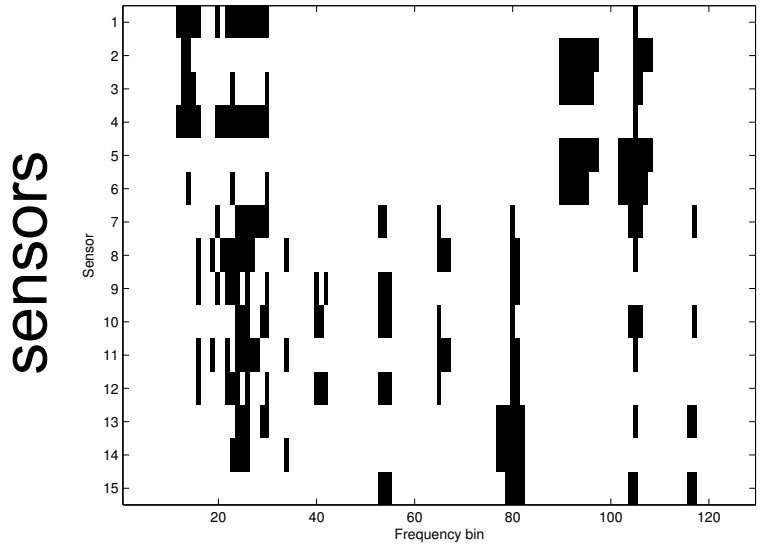
Coordinated monitoring in sensor networks

- Wireless sensor network environment: constrained communication and computation
- Coordinated monitoring: sensors recommend features, features recommend sensors
- Rating algorithm
- Unsupervised feature selection



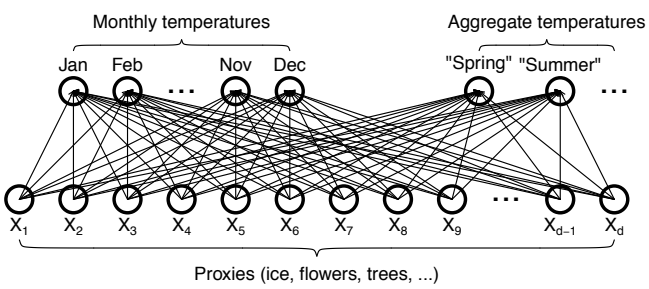
Coordinated monitoring in sensor networks: Results

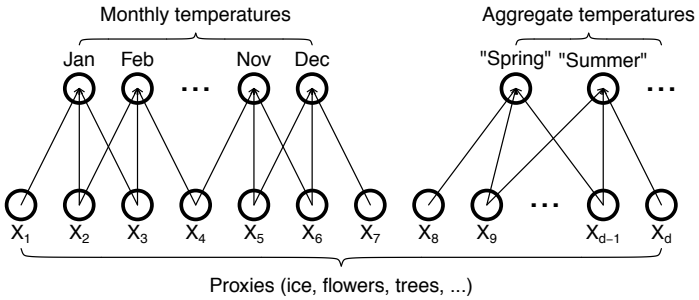




- Resulting selection of sensor-feature pairs
- Controlled sparsity of the solution (left vs. right)
- Reduction of communication, measurement, and computation poster!

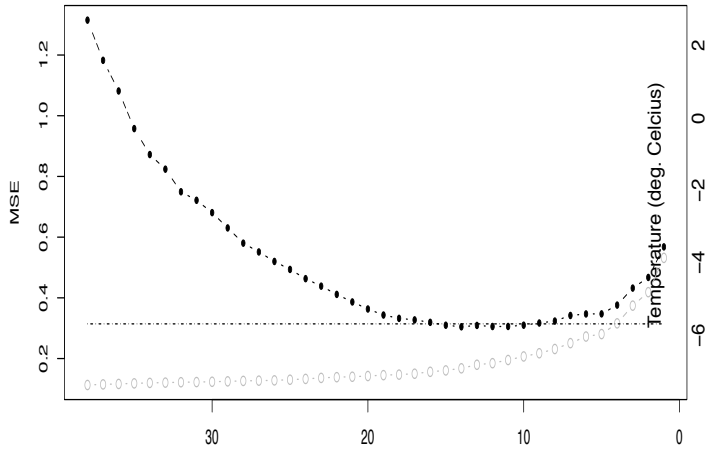
Environmental proxy variable selection



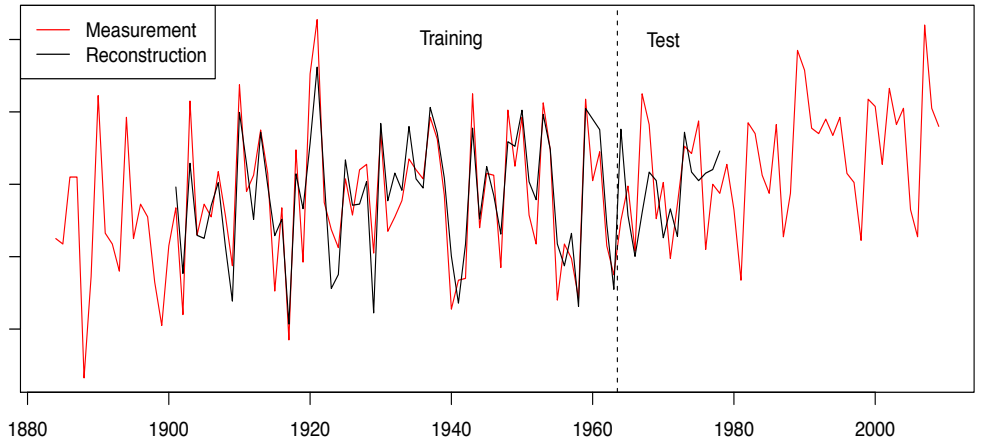


- Learn prediction models for temperature reconstruction
- Proxy variables store temperature information indirectly
- Explosion of the space of models, aim at compact models
- Search-based feature selection: include suboptimal models in the family of solutions poster!

Environmental proxy variable selection



model selection



Reconstruction of temperature

- Example of temperature reconstruction with a prediction model identified with a search procedure in model-space.
- poster!

Summary and Conclusions

- Environmental data analysis problems
- Parsimonious modelling
  - Signal decomposition by three-way analysis
  - Collaborative filtering for sensor-frequency selection poster!
  - Search-based feature selection in model space poster!
- Further exploration of the parsimonious modelling in application areas of cancer genomics and environmental informatics

ALGODAN Algorithmic Data Analysis

Relevant and Non-redundant Object Retrieval

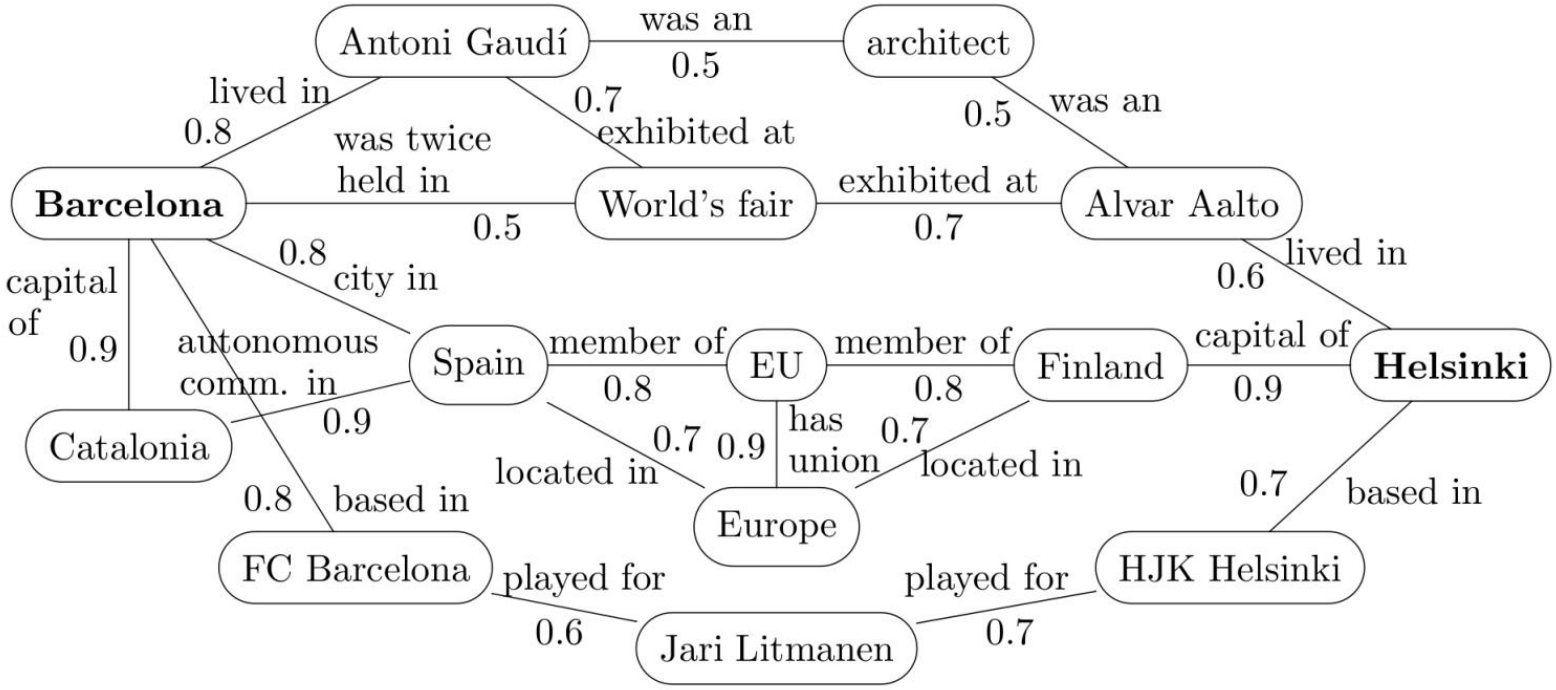
Laura Langohr and Hannu Toivonen

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Motivation

- How to identify objects that are relevant w.r.t. *Barcelona* and *Helsinki*, but non-redundant w.r.t. each other?



Other application domains include …

- Obtaining an overview of different term contexts
  - Query term: *root*
  - Result: *plant*, *equation*, *word* - a representative set of terms representing different contexts (botany, mathematics, linguistics)
- Understanding co-authorship relations
  - Identify authors that are relevant w.r.t. query authors, but non-redundant w.r.t. each other
- Biomedicine
  - Identify phenotypes that are relevant w.r.t. query genes, but non-redundant w.r.t. each other

Related Work

- Identifying relevant objects (typically documents) is a classical problem in information retrieval (IR).
- Our problem differs from these as in our work as:
  1. Objects are not assumed to have any attributes
  2. Relevance is based solely on a proximity function
  3. Queries are specified by objects, not by keywords
- A variant of random walk with restart addresses non-redundancy (Tong et al. 2011)
  - We could alternatively use these relevance and non-redundancy measures
  - However, negative query nodes are not considered

Some Concepts

- **Relevance:** A node is relevant if it is closely related to **all** positive query nodes.
- **Irrelevance:** A node is irrelevant if it is closely related to **any** negativ query node.
- **Redundancy:** A node is redundant if it is closely related to **any** other selected node.

Relevance

- V, a set of objects
- $q \in V$, a positive query object
- $d : V \times V \rightarrow \mathrm{R}^{+}$, a distance measure for objects in V
- Alternatively: $s : V \times V \rightarrow \mathrm{R}^{+}$, a proximity function

- Relevance of an object $u \in V$ w.r.t. the query object $q \in V$

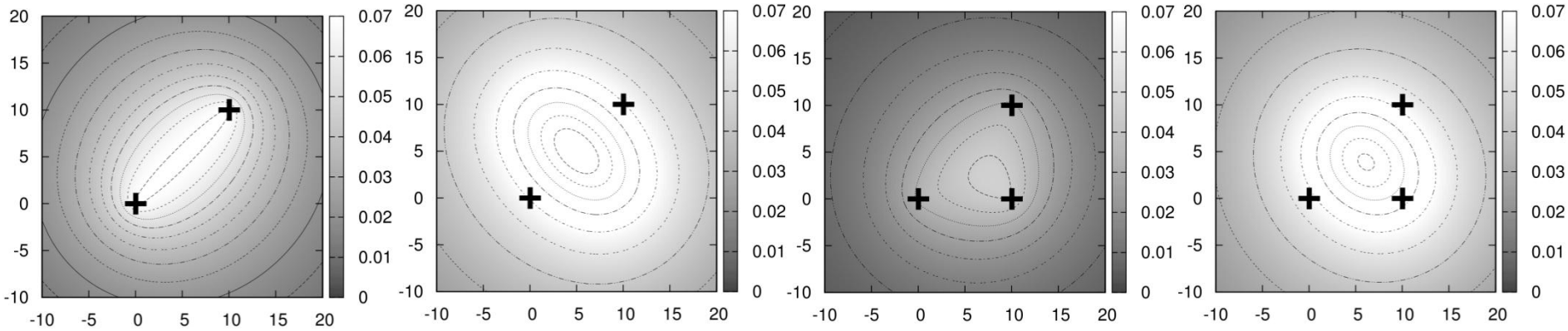
$$rel_P(u, q) = s(u, q) = 1 / d(u, q)$$

Relevance

- Relevance of an object u w.r.t. a set $Q_P \subset V$ of query objects

$$rel_P(u, Q_P) = \left(\sum_{q \in Q_P} d(u,q)^{\alpha} \right)^{-1/\alpha}$$

- With larger values of $\alpha \geq 1$, larger distances dominate the function:



Irrelevance

- Negative query objects to specify subjective irrlevance (uninterestingness) of objects

- Irrelevance (negative relevance) of an object u w.r.t. a negative query object $\bar{q} \in V$

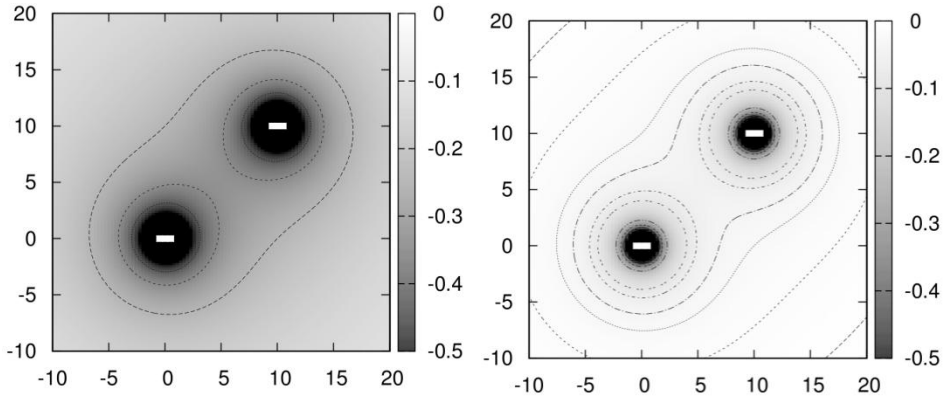
$$rel_N(u, \bar{q}) = s(u, \bar{q}) = 1 / d(u, \bar{q})$$

Irrelevance

- Irrelevance of object u
  w.r.t. a set $Q_N \subset V$ of negative query objects

$$rel_N(u, Q_N) = \sum_{\bar{q} \in Q_N} d(u, \bar{q})^{-\beta} = \sum_{\bar{q} \in Q_N} s(u, \bar{q})^{\beta}$$

- Increasing $\beta \geq 1$, increases the local concentration around negative query nodes:



$\beta = 1$ $\beta = 2$

Non-redundancy

- We want to retrieve a list of relevant objects, but we also want them to be mutually non-redundant.

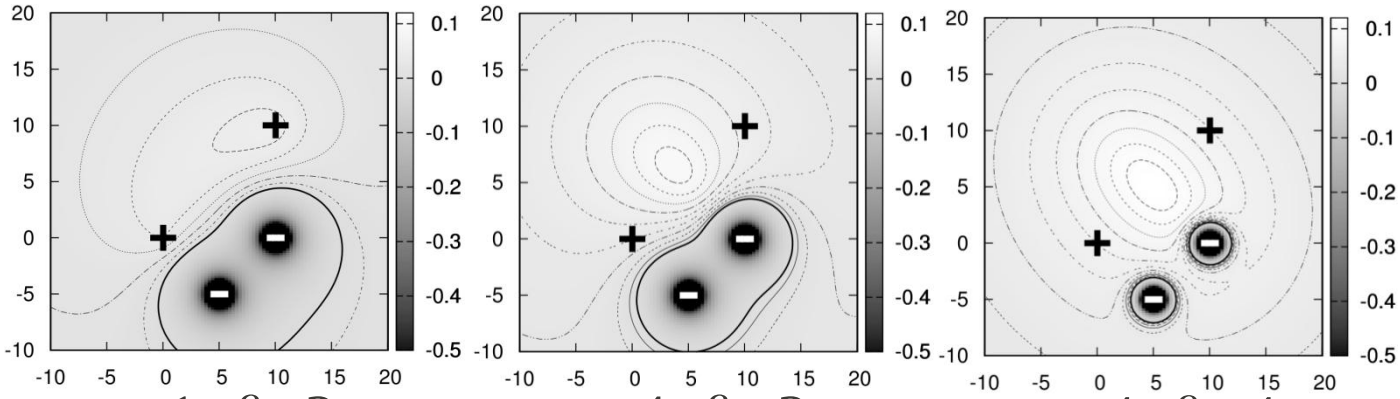
- This is similar to the effect of negative query objects. Hence, we define redundancy of a set $R \subset V$ of objects similarly.

$$red(R) = \sum_{\substack{u,v \in R, \\ u \neq v}} d(u,v)^{-\beta} = \sum_{\substack{u,v \in R, \\ u \neq v}} s(u,v)^{\beta}$$

Relevance and Non-redundancy

- Overall relevance and non-redundancy of a set of objects $R \subset V$

$$REL\left(R, Q_P, Q_N\right) = \sum_{u \in R} rel_P\left(u, Q_P\right) - \sum_{u \in R} rel_N\left(u, Q_N\right) - red\left(R\right)$$



Probabilistic Relations

- We proposed alternative measures for relevance and non-redundancy with a probabilistic interpretation.

  (Langohr and Toivonen 2012)

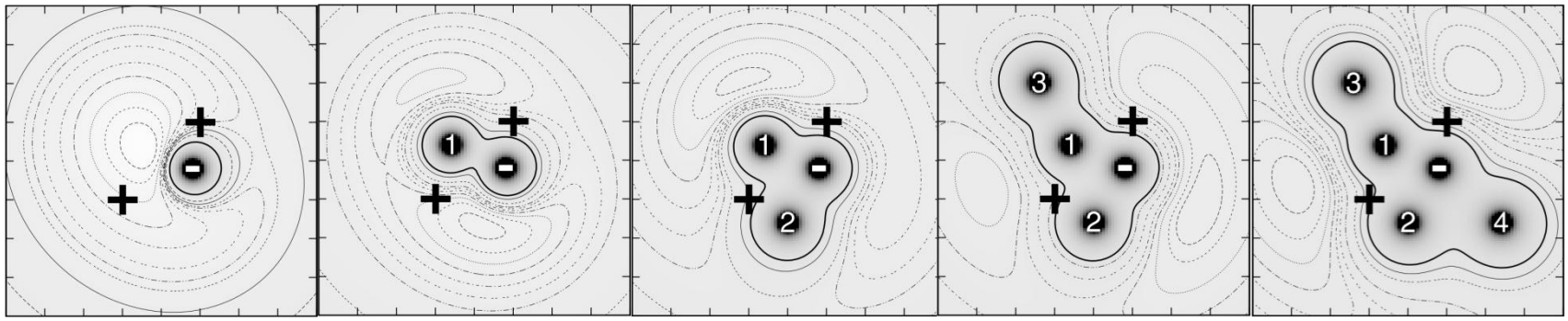
- This is especially interesting in a setting of uncertain networks, where edge weights describe probabilistic relations between nodes.

Greedy Algorithm

- The overall relevance $REL(\cdot)$ is submodular
- Hence, a greedy algorithm is guaranteed to find a set which achieves at least $1/k$ of the optimal score

(Nemhauser 1978)

1. Repeat until k representatives has been retrieved:
   a. Find the most relevant object r w.r.t. Q_P and Q_N
   b. Output r and add it to Q_N



Iterative Algorithm

1. Get an initial solution R of k objects (e.g. random)
2. Repeat while R changes:
    a. Find the optimal swap of any object r in R to any object not in
    b. If the swap improves the result, implement it

Word Relations and Senses

- **Problem:** Identify an overview of different senses or contexts of words
- **Proximity:** word co-occurrence within sentences

word(s)	*bank*	*star*	*root*	*branch, root*
contexts	reserve	planet	plant	tree
	river	trek	equation	indo
	gaza	cluster	word	mathematics
	credit	sirius	irrationality	line
	international	movie	unity	equation

Co-authorship Relations

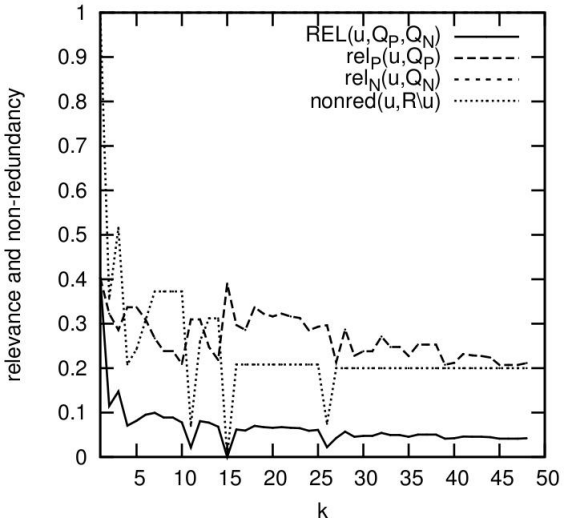
- **Problem:** Identify authors that are relevant w.r.t *J. Han* and *C. Faloutsos*, but non-redundant w.r.t. each other.
- **Proximity:** proportional to the number of co-authorships

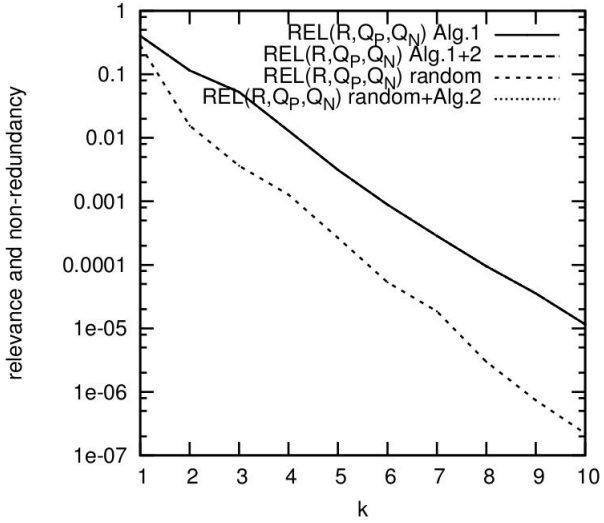
(Potamias et al. 2012)

Rel. and non-redundancy		**Relevance only**	
P.S. Yu	IL, USA	P.S. Yu	IL, USA
D. Srivastava	NJ, USA	R.T. Ng	Canada
H.J. Zhang	China	S. Papadimitriou	NY, USA
Y. Tao	Hong Kong	L.V.S. Lakshmanan	Canada
C. Liu	WA, USA	H.V. Jagadish	MI, USA
B. Chin Ooi	Singapore	X. Yan	CA, USA
T.K. Sellis	Greece	J. Yang	OH, USA
J. Gao	IL, USA	W. Fan	NY, USA

Co-authorship relations

- **Problem:** as before …
- **Proximity:** as before ...



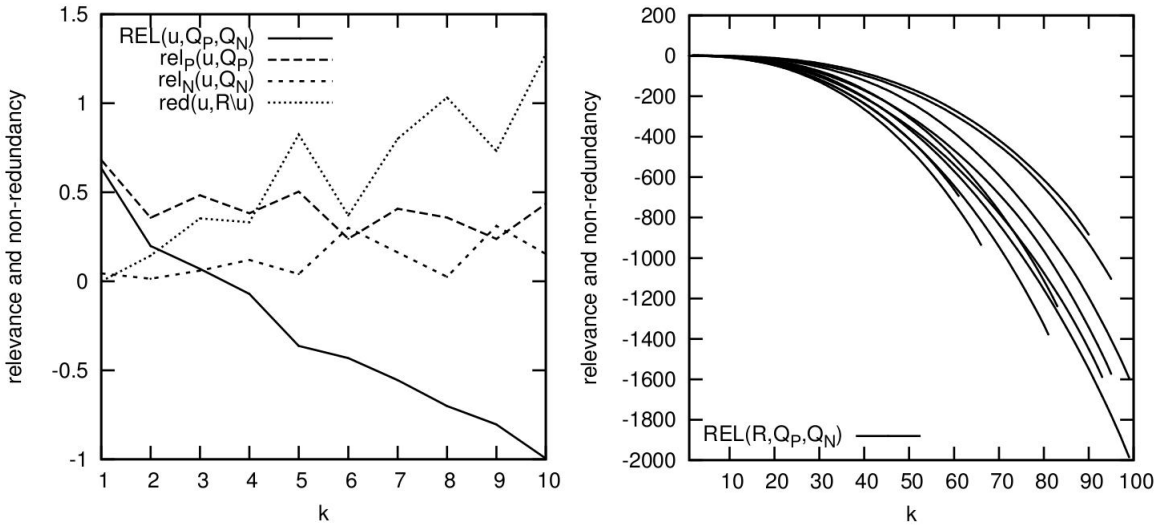


=> The iterative algorithm produces only marginally better results than the greedy ranking.

Biomedicine

- **Problem:** Identify phenotypes that are relevant w.r.t. query genes, but non-redundant w.r.t. each other.
- **Proximity:** probabilistic proximities from Biomine

(Sevon et al. 2006)



=> The quality decreases in quite a similar manner for different cases

Conclusion

- Problem: Find a non-redundant set of relevant objects, given positive and negative query objects
- Relevance, irrelevance, and non-redundancy definitions are based on object distance/proximity.
- Greedy and iterative algorithms produce a good set of objects, with high relevance and low redundancy.
- The iterative algorithm produces only marginally better results than the greedy ranking.

Future Work

- A deeper analysis of the problem and its properties
- The proposed algorithms are simple but efficient if the proximities are given. For more complex and larger cases faster algorithms are needed.
- More extensive experiments to understand the practical behaviour of the methods and parameters
- Adaption of the approach to different applications

ALGODAN Algorithmic Data Analysis

Discovering Knowledge about the Evolution of Bacterial Metabolism: Weighted Graphs and Compression

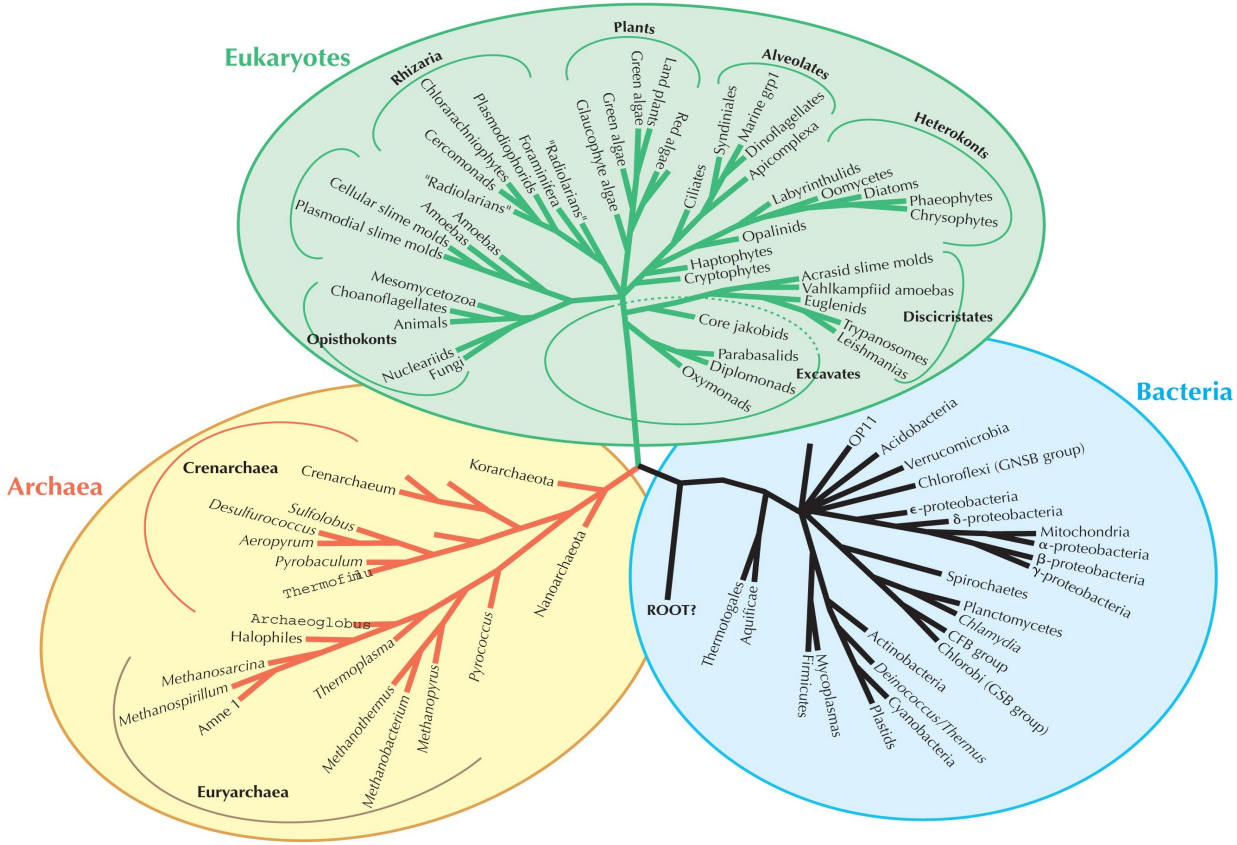
Fang Zhou

PhD student, Discovery group

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tree of life



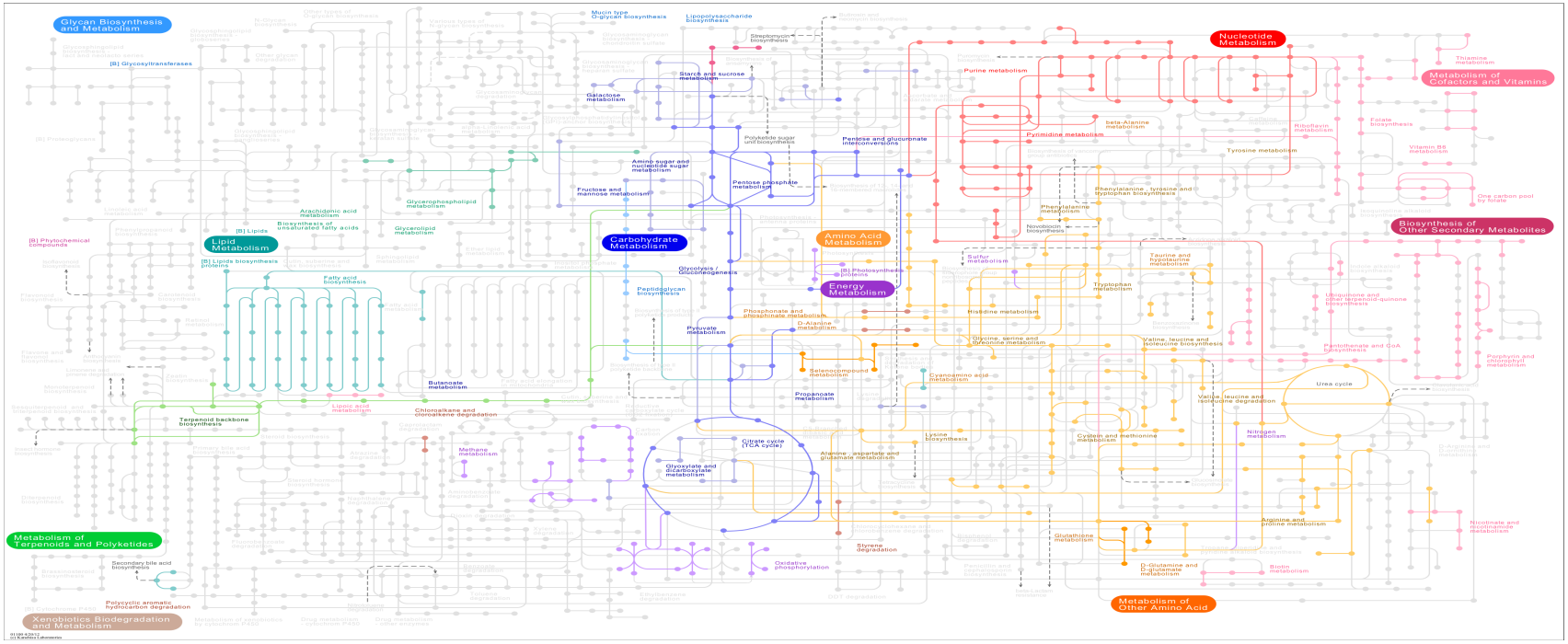
We are interested in the biodiversity within metabolism in *Archaea* and *Eubacteria*, two main branches of life.

Problem

- **Question:** whether the evolution process is constrained by the environment and biochemistry or it is a stochastic process?
  - The central problem of this question is that it is not possible to repeat the experiment – evolution.
  - However, it is possible to get some understandings of the stochasticity of the problem by looking at cases where evolution started from similar starting points.
- **Goal:** understand the conservative of metabolisms.
- **Idea:** compare the essential part of metabolisms.

Metabolism

- The metabolism of one species.



- The metabolism contains thousands of enzymes. The network is too complex to easily analyzed.

Metabolism

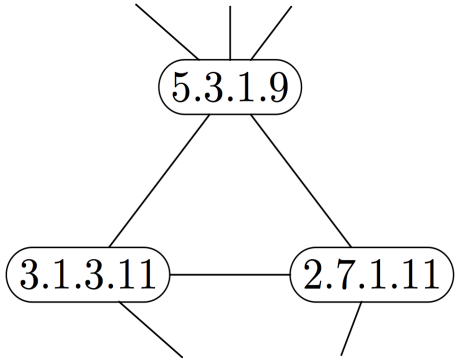
- Different species have different metabolisms.
- Given a large number of species, how can we compare their essential parts?
- **Our method:**
  - Integrate metabolisms of species into one graph.
  - Extract essential parts by using graph compression.
  - Compare essential parts of graphs.

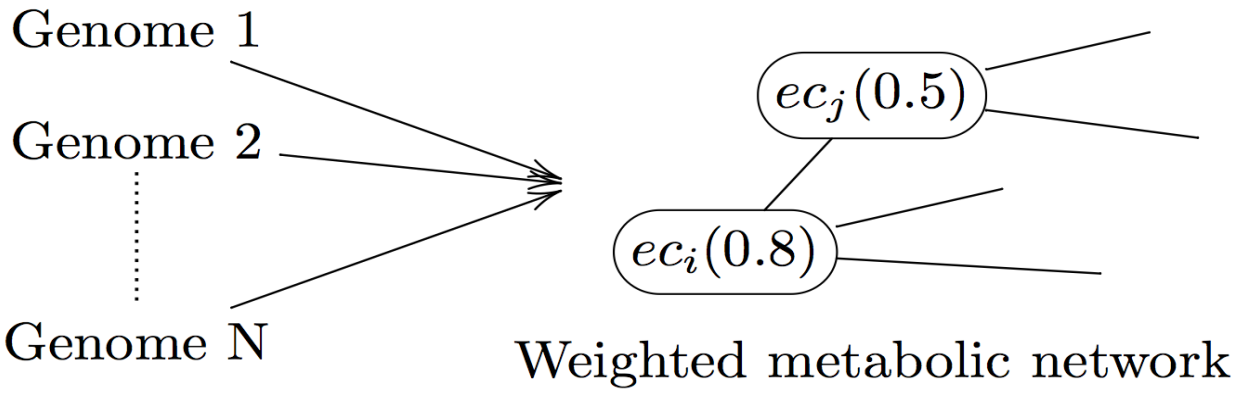
Outline

- Representing metabolism with graphs
- Weighted graph compression
- Application to metabolic networks
- Conclusion

Weighted metabolic networks

- How to integrate metabolic networks into one graph?
  - Represent the meta-metabolic network as a graph with enzymes as nodes. Two enzymes are connected if they catalyze reactions that share metabolites.
  - Assign weights to enzymes based on how frequent they are in the species.



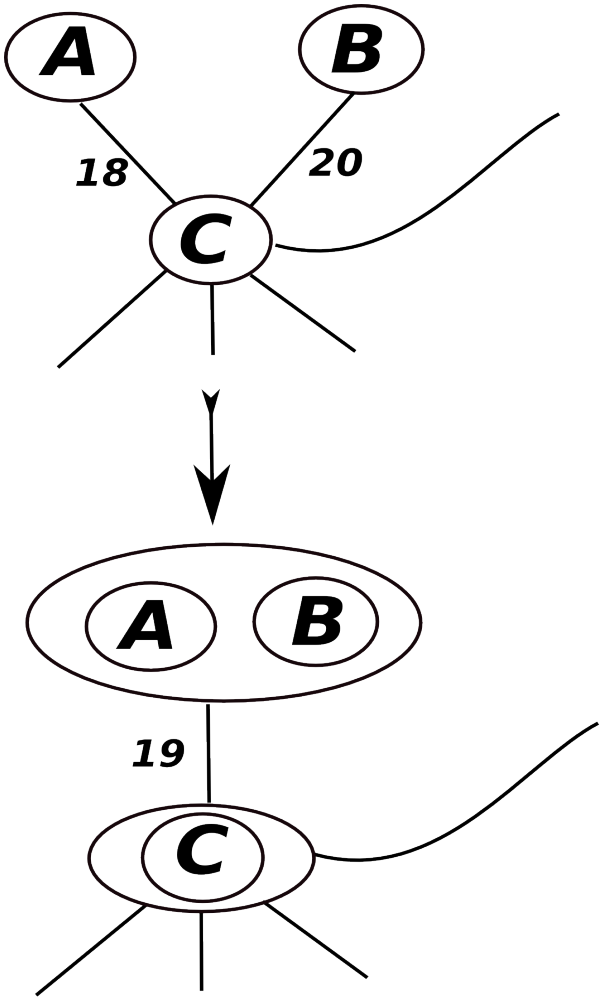


Outline

- Representing metabolism with graphs
- Weighted graph compression
- Application to metabolic networks
- Conclusion

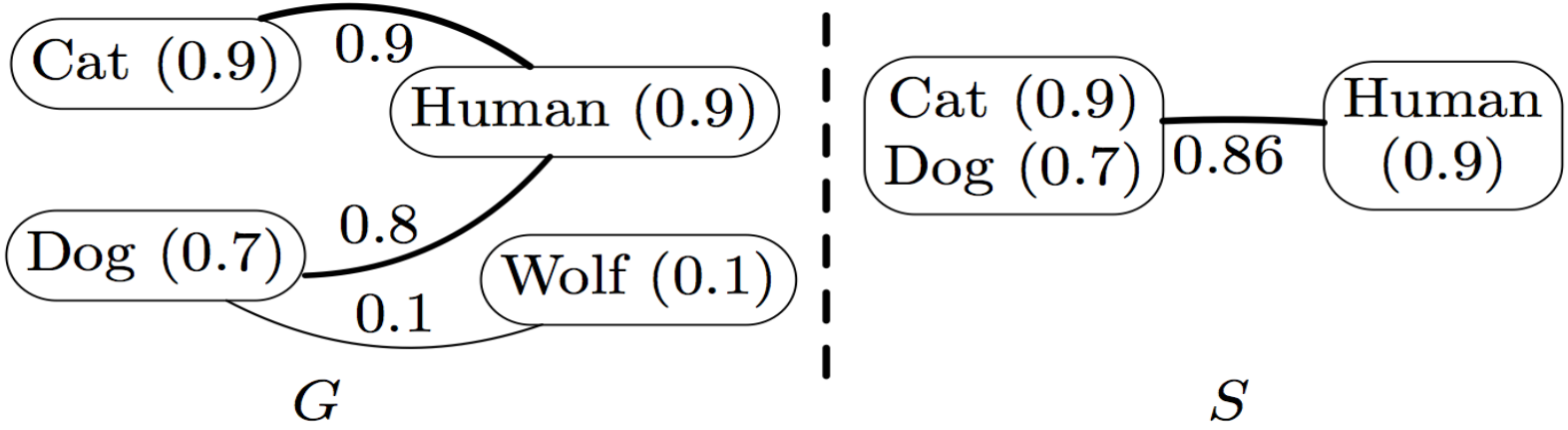
Weighted graph compression

- Weighted graph compression= grouping nodes that have similar link structure. (KDD2011)
  - A supernode represent all original nodes within it.
  - A superedge represent all possible edges between the corresponding original nodes.
  - The superedge weight is the mean of the original edge weights.
- Differs from clustering or community detection.



Compression based on node importance

- We extend the definition of graph compression to also consider node importance.
  - The goal is to produce a smaller graph with less error related to more important nodes.
  - Nodes are merged to supernodes, and edges to superedges.
  - Low-weight edges and unimportant nodes are removed.



Compression error

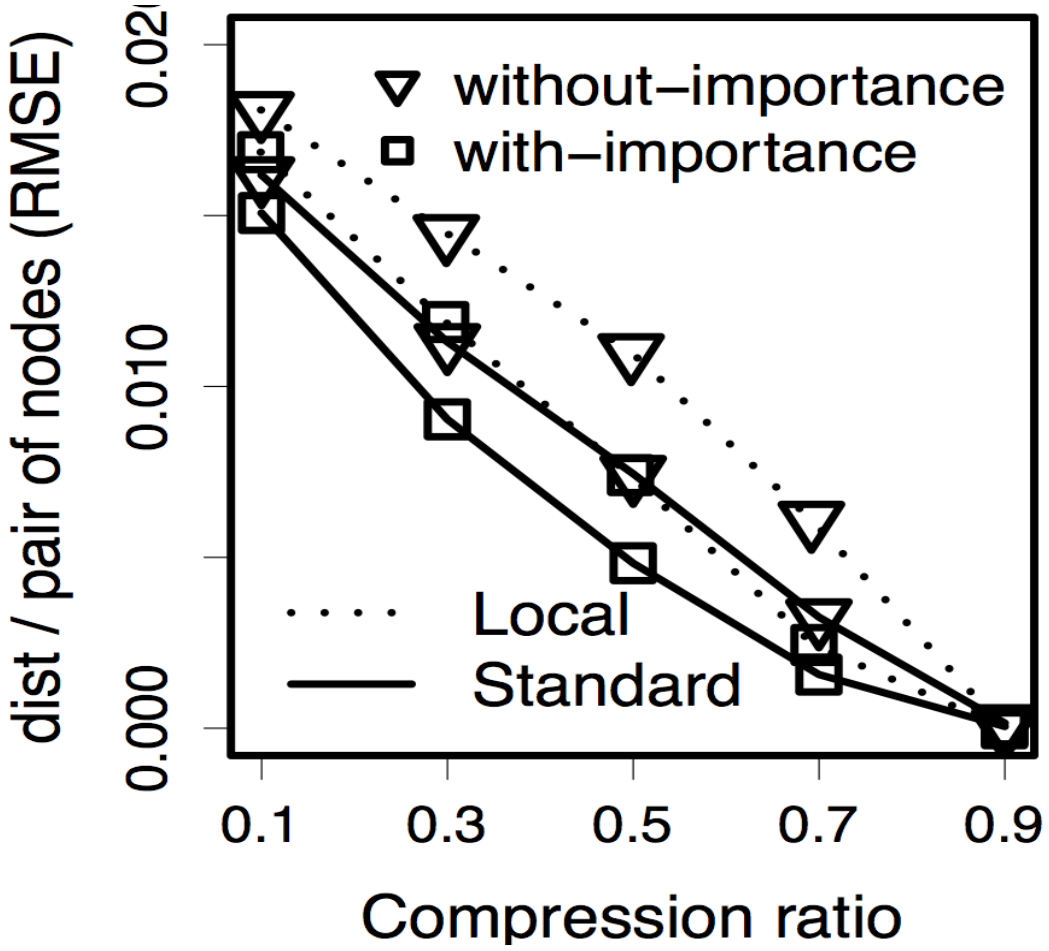
- Possible compression error
  - Missing edges and nodes;
  - Extra edges;
  - Inaccurate edge weights.
- Compression error = Euclidean distance between the original and the compressed graphs weighted by node importance

Algorithms

- Agglomerative process: execute one operation per time until the specific compression ratio is reached.
  - Operation 1: node-pair merger (with possible omission of (super)edges and (super)nodes)
  - Operation 2: individual (super)edge deletion.
- Two basic algorithms
  - Brute-force executes the best possible operation in each iteration.
  - Randomized algorithm randomly picks a node u, and then chooses v whose combination with u gives the best possible result.

Results on synthetic datasets

- Node importances can guide the process to a better compression.

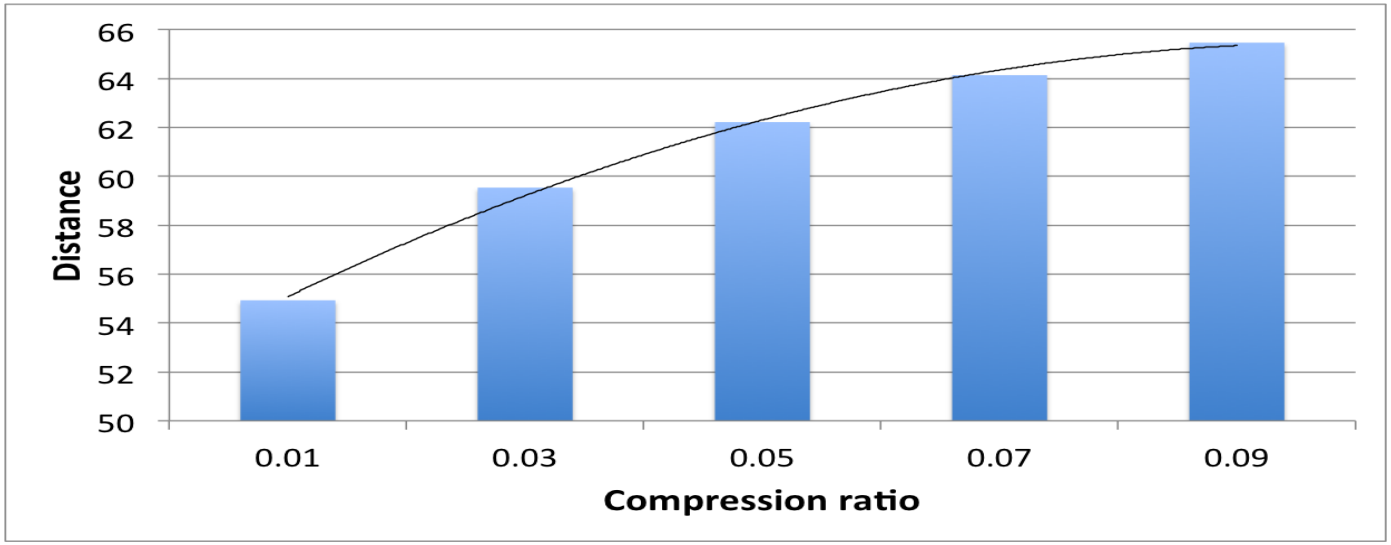


Outline

- Representing metabolism with graphs
- Weighted graph compression
- Application to metabolic networks
- Conclusion

Correlation between two kingdoms

- **Question:** whether the evolution process is constrained by the environment and biochemistry?



- The compressed graph gives the essential part of metabolism.
- **Results show:** more compression actually gives a smaller distance. The evolution process is constrained.

Outline

- Representing metabolism with graphs
- Weighted graph compression
- Application to metabolic networks
- Conclusion

Main contributions

- We presented the weighted graph compression problem, and also extended and generalized the problem definition to also consider node importances.
- Two compression operations and four algorithms are proposed, and are evaluated on real datasets.
- The use of weighted graphs and compression provides a framework to investigate the existence of constraints in the evolution of metabolism.

Future work

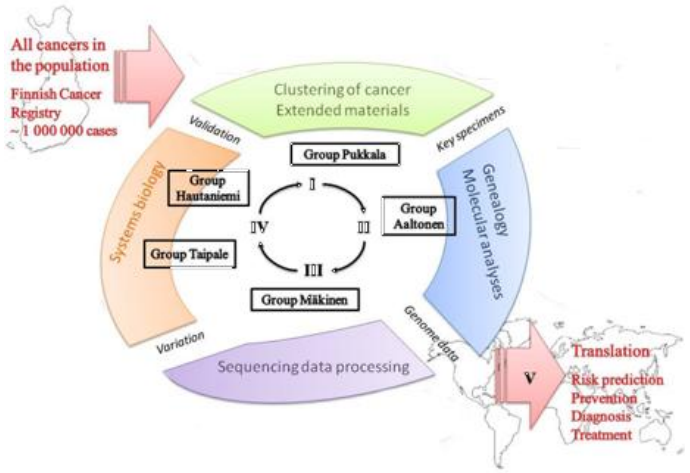
- Develop more effective compression methods.
- Apply the compression method to compare the importance of pathways in the different kingdoms.
- Extract an approximate ancestral metabolism, which is a connected subgraph with enzymes that are common to both kingdoms.

Acknowledgement

- Hannu Toivonen (UH, Finland)
- Ross D. King (University of Manchester, UK)
- Aleksi Hartikainen (UH, Finland)
- Atte Hinkkka (UH, Finland)
- Jonathan H. Badger (J. Craig Venter Institute, US)

Thanks for your time!

ALGODAN Algorithmic Data Analysis



Enhanced Variation Calling

Veli Mäkinen

Genome-scale algorithmics group / CoE in Cancer Genetics Research (2012-)

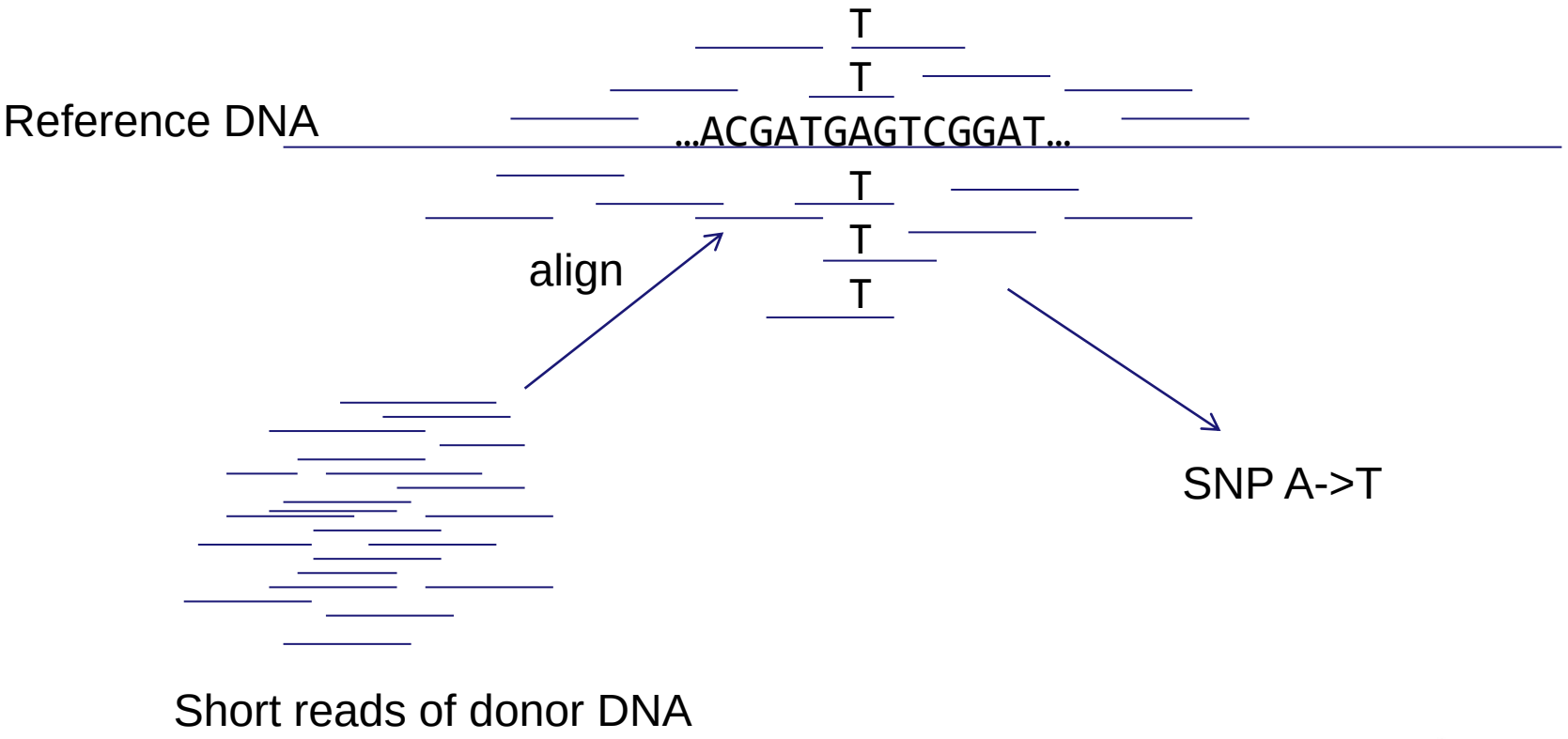
Succinct Data Structures group / ALGODAN (-2011)

Joint work with

Jouni Sirén, Niko Välimäki, Serikzhan Kazi, Esa Pitkänen, Riku Katainen, Eevi Kaasinen

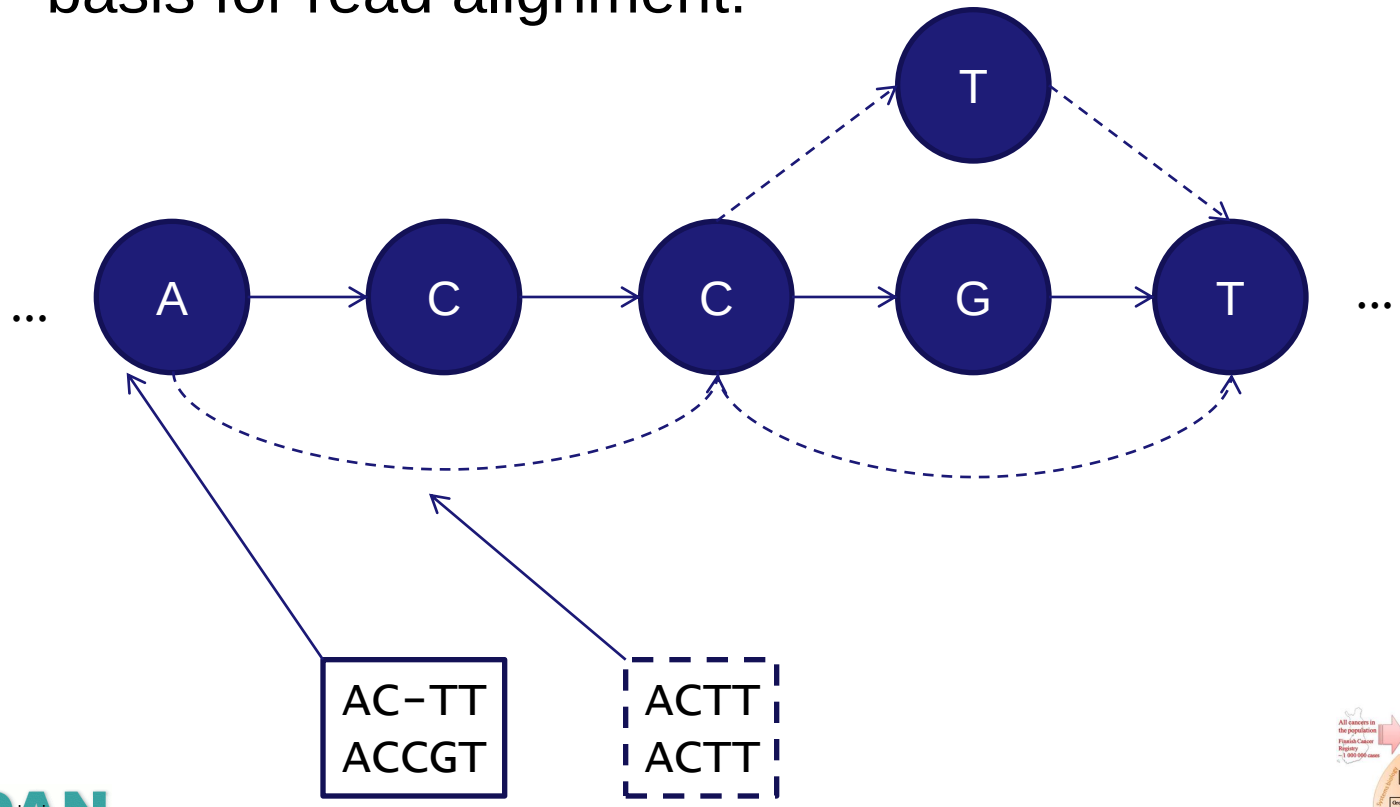
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Variation calling

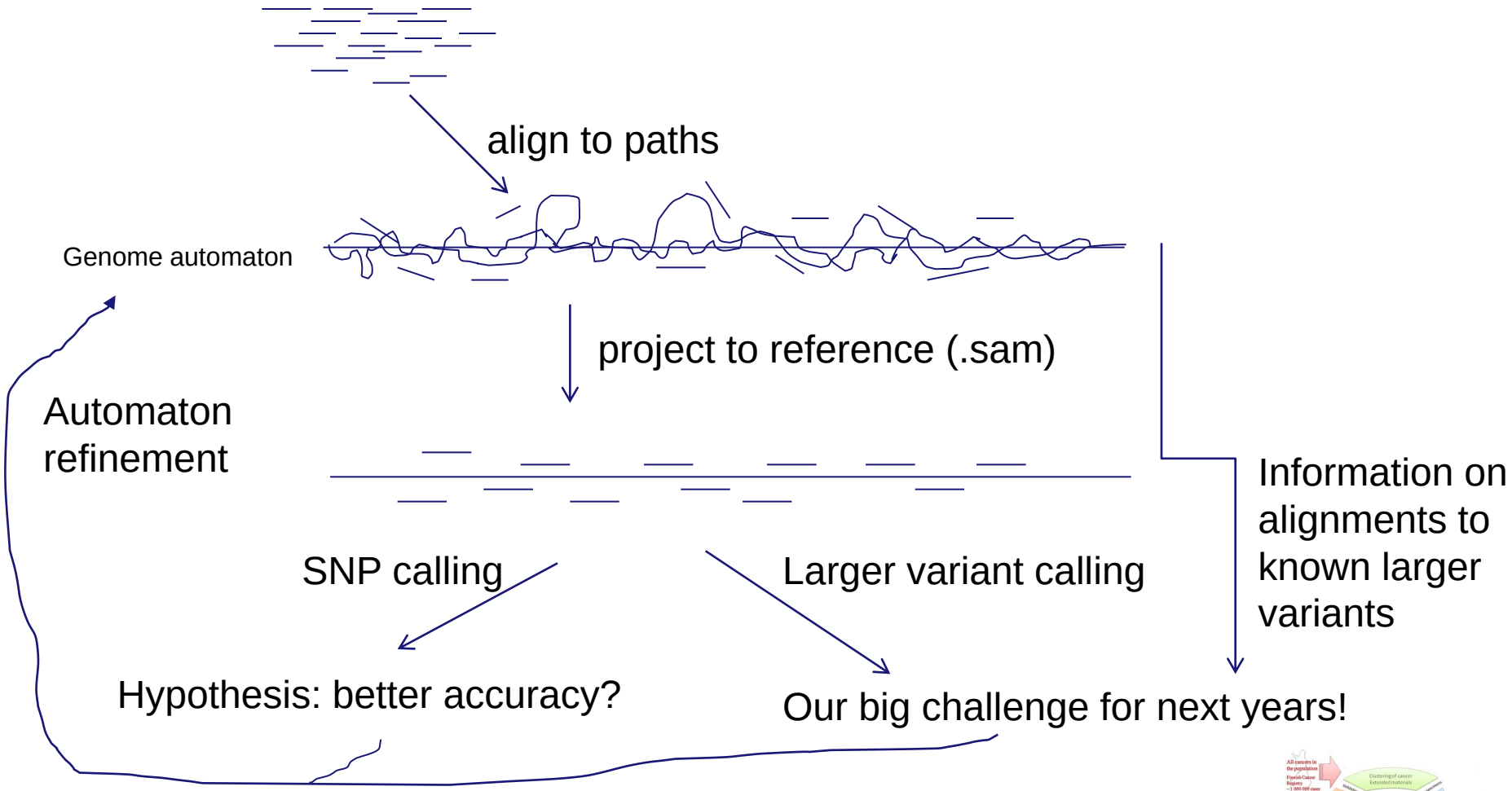


Enhanced variation calling

- Why always only one reference is used?
- We propose to use reference + known variations as the basis for read alignment:

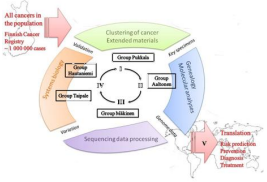


Enhanced variation calling pipeline



Feasibility of the approach?

- Sirén, Välimäki, Mäkinen. Indexing Finite Language Representation of Population Genotypes. *WABI 2011*.
  - Generalization of Burrows-Wheeler transform for acyclic finite automata -> Generalized compressed suffix array (GCSA)
    - Based on our work in *RECOMB 2009* for multiple genomes.
  - Supports alignment of reads alike the other read aligners
    - Given a read P of length m, one can count the paths starting with P in O(m) time
    - Extends to approximate search with the general backtracking & branch-and-bound mechanism.
  - Similar space usage as for other read aligners
    - 3.3 GB for Human Genome + Finnish subpopulation of 1000 genomes data!
      - Construction requires 173 GB and takes 18 hours….



Alignment experiment

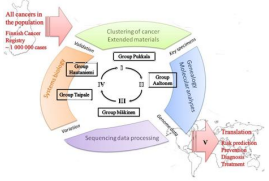
% reads mapped

Errors	GCSA	Time	RLCSA	Time
0	86,3%	1,5 h	82,7%	0,5 h
1	91,9%	1,9 h	91,4%	0,8 h
2	94,1%	7,0 h	93,9%	5,4 h

BWT index on automaton,
Supports k-errors search faithfully.

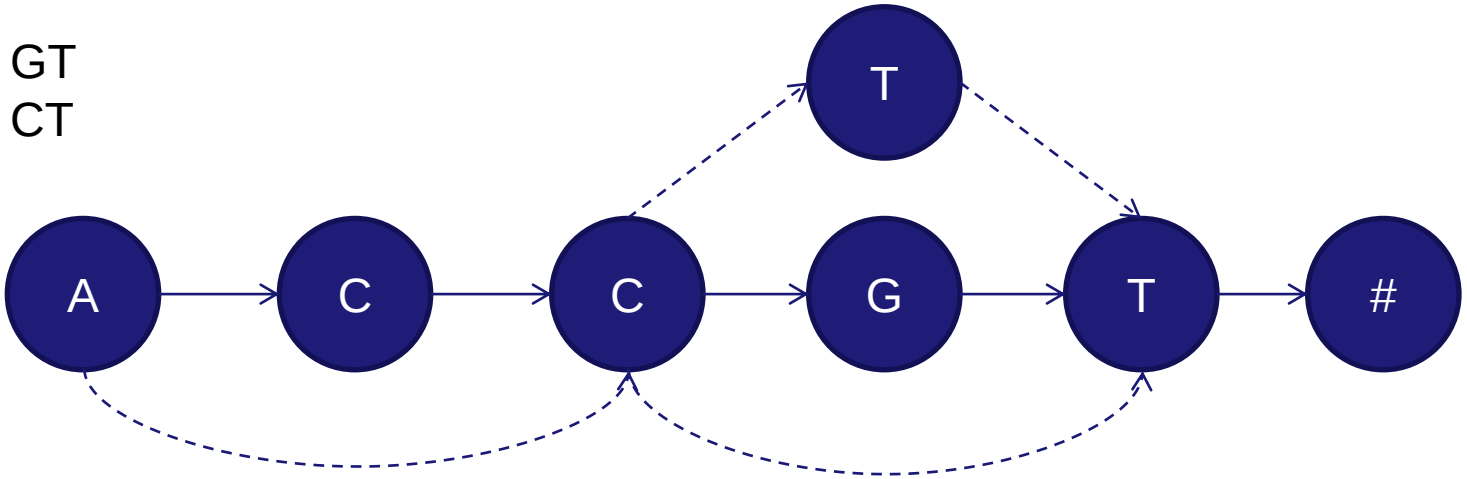
BWT index on reference genome,
alike Bowtie, BWA, SOAP2.
Supports k-errors search faithfully
-> better yardstick for GCSA.

10 million reads of length 56 using one thread



Some insights

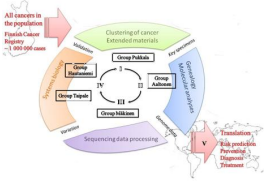
Search GT
Search CT



	GBWT	Distinguishing prefixes	Edges
1 G, 4 C	T	#	1
1 G, 4 C	#	A	2
1 G, 4 C, 0 G, 2 C	A	CC	1
1 G, 4 C, 0 G, 2 C	AC	CG	1
1 G, 4 C, 0 G, 2 C	AC	CT	2
1 G, 4 C, 0 G, 2 C	C	G	1
1 G, 4 C	CGT	T#	1
1 G, 4 C	C	TT	1

All paths in lex order

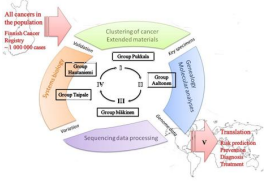
#
ACCGT#
ACCTT#
ACGT#
ACTT#
CCGT#
CCTT#
CGT#
CT#
CTT#
GT#
T#
TT#



ALGODAN
Algorithmic Data Analysis

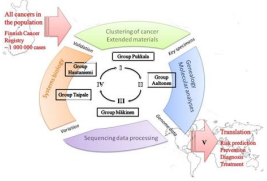
Summary

- Make finite automaton from reference + SNP data or from multiple alignment.
- Make it reverse deterministic (skipped details).
- Sort distinguishing prefixes (prefix doubling, bucket sort, others?)
- Output GBWT.
- Read alignment almost identical to normal BWT read aligners (like bowtie, bwa, SOAP2).

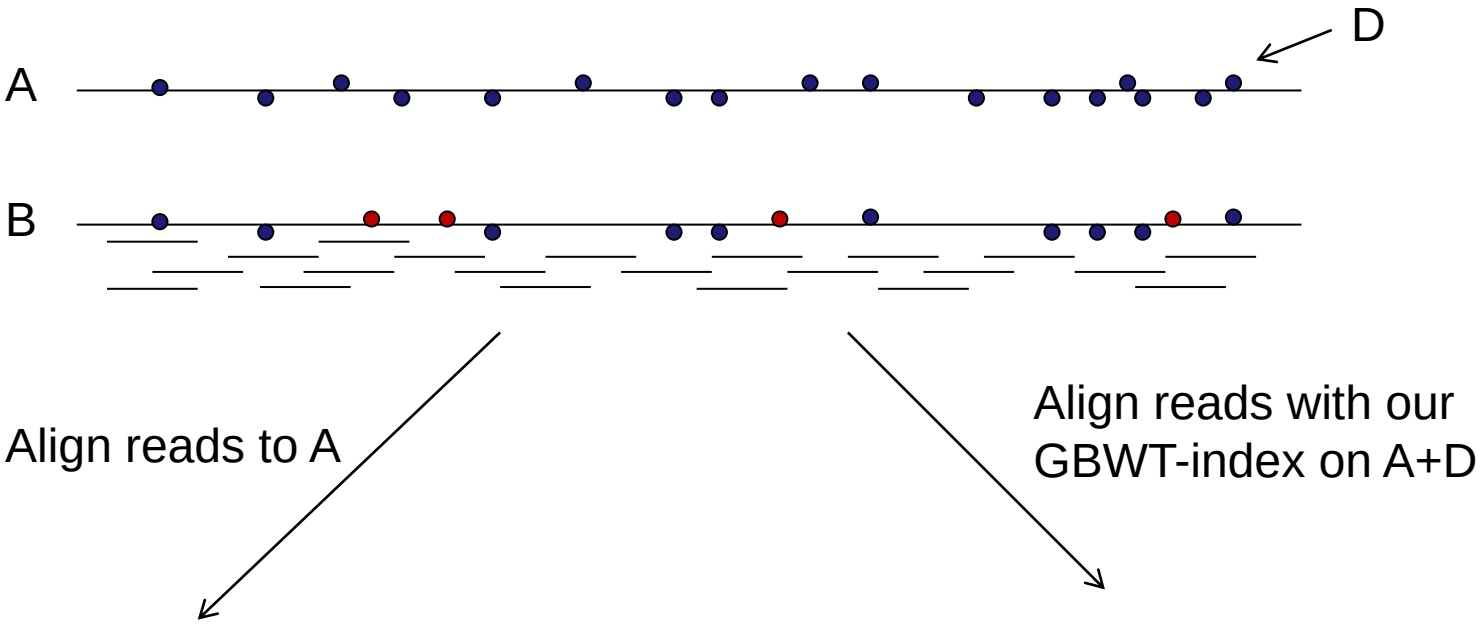


What now?

- Distributed construction algorithm in progress -> no need for massive main memory.
- **Validation test** -> can we get better accuracy for small variant calling?
  - Anchored alignment -> projection to reference (done, just running tests)
- Large variant calling directly from paths? -> Less data for *de novo* calling -> should improve accuracy.
- Other applications: primer/adapter design



Validation test



- Do variant calling identically with both approaches.
- Calculate precision/recall for both approaches.

ALGODAN Algorithmic Data Analysis

Improving Burrows-Wheeler Compression

Juha Kärkkäinen

Simon J. Puglisi, Dominik Kempa, Pekka Mikkola

Practical Algorithms on Strings Group

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Burrows-Wheeler Compression

- In 1994, the Burrows-Wheeler transform (BWT) introduced a completely new way to compress text
  - Used in data compressors such as `bzip2`
- In 2000, the FM-index based on BWT created a new type of text index: the compressed full-text self-index
  - Store text in compressed form
  - Support fast pattern matching queries (index)
  - Used in bioinformatics tools such as `bowtie`

Burrows-Wheeler Transform (BWT)

- Invertible permutation of text

ABRACADABRA → RDARCAAAABB

- Easier to compress than text
- Indexing is easier too
  - Pattern matching queries on text can be implemented using simpler queries on BWT

rank(A,7) = 3 = number of A's in first 7 characters

Our Recent Improvements

- Grammar precompression
  - Faster compression and decompression

POSTER

- Faster algorithms for inverse BWT
  - Faster decompression

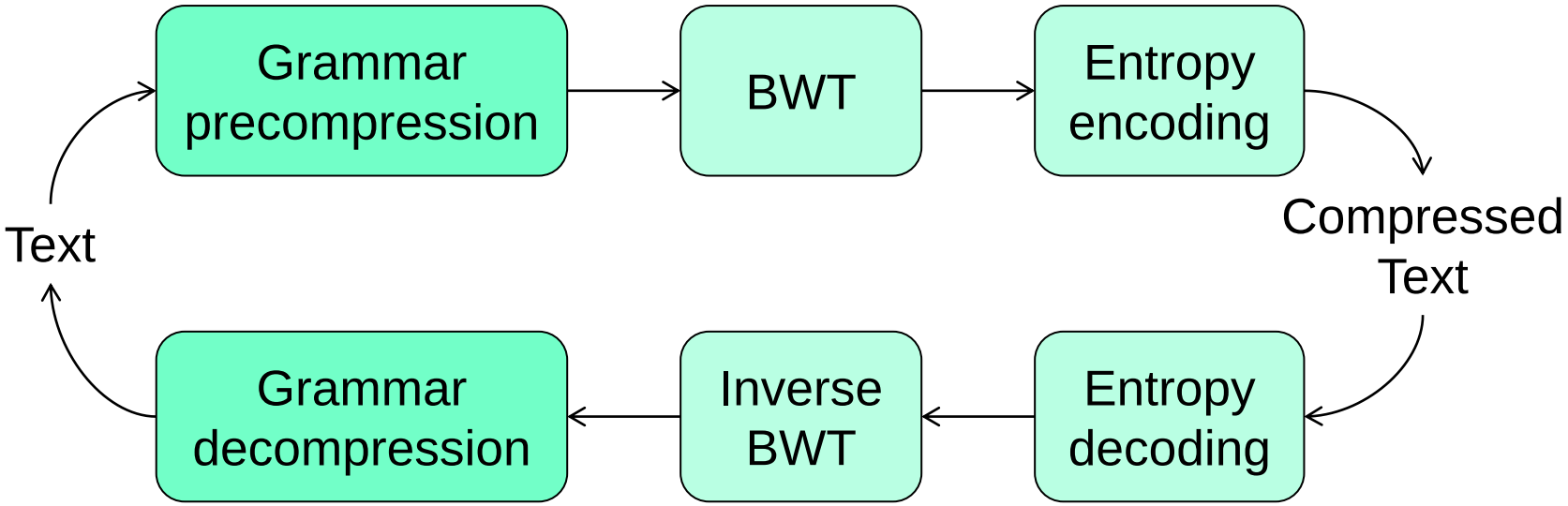
POSTER

- Fixed-block compression boosting
  - Simpler, faster, smaller FM-index

Grammar Precompression

POSTER

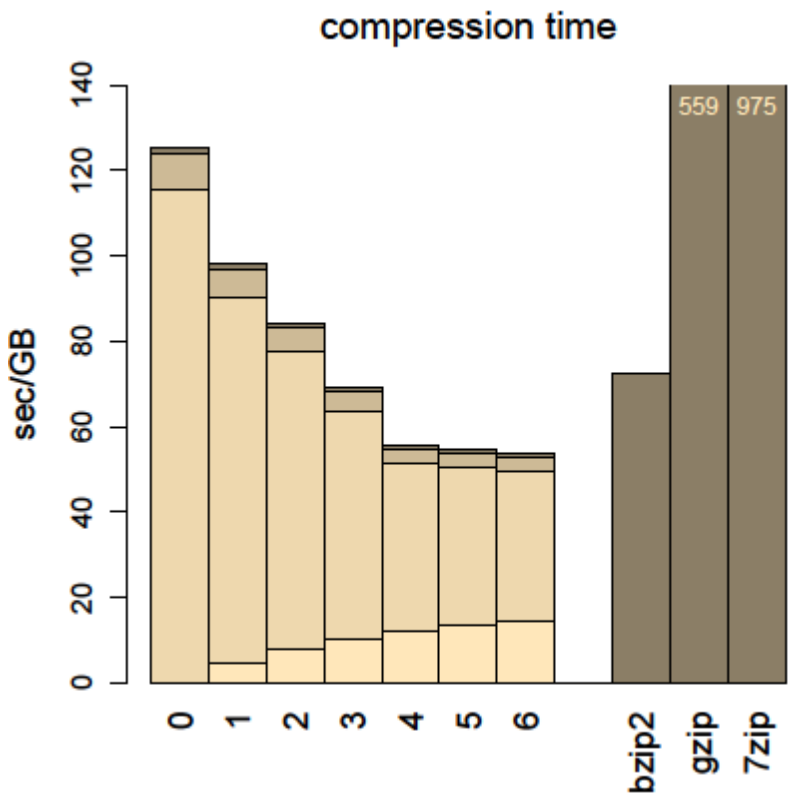
- Initial compression before BWT

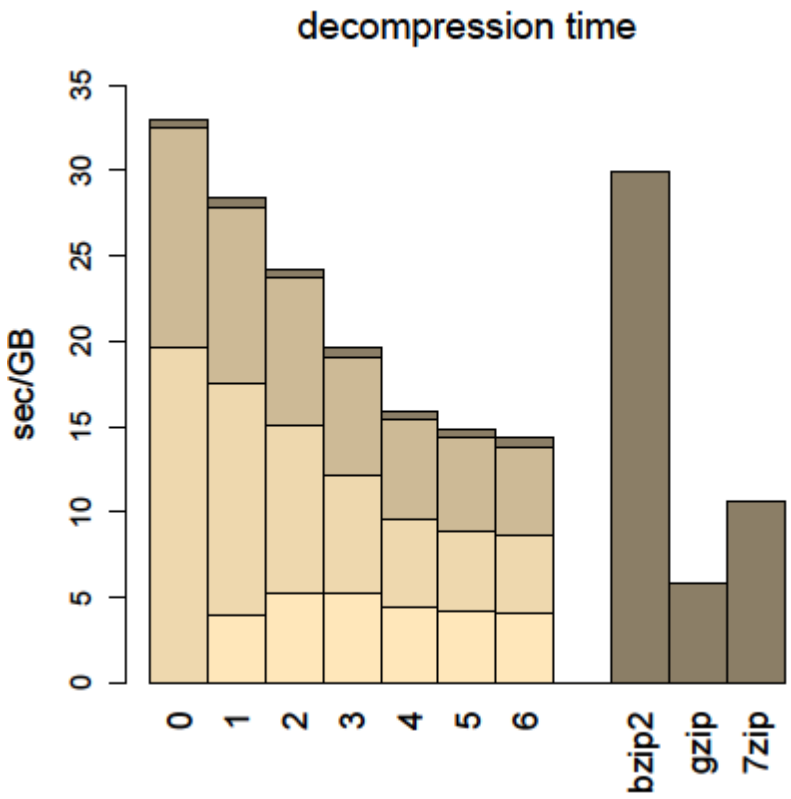


- Less data to process by BWT and its inverse

Grammar Precompression

POSTER





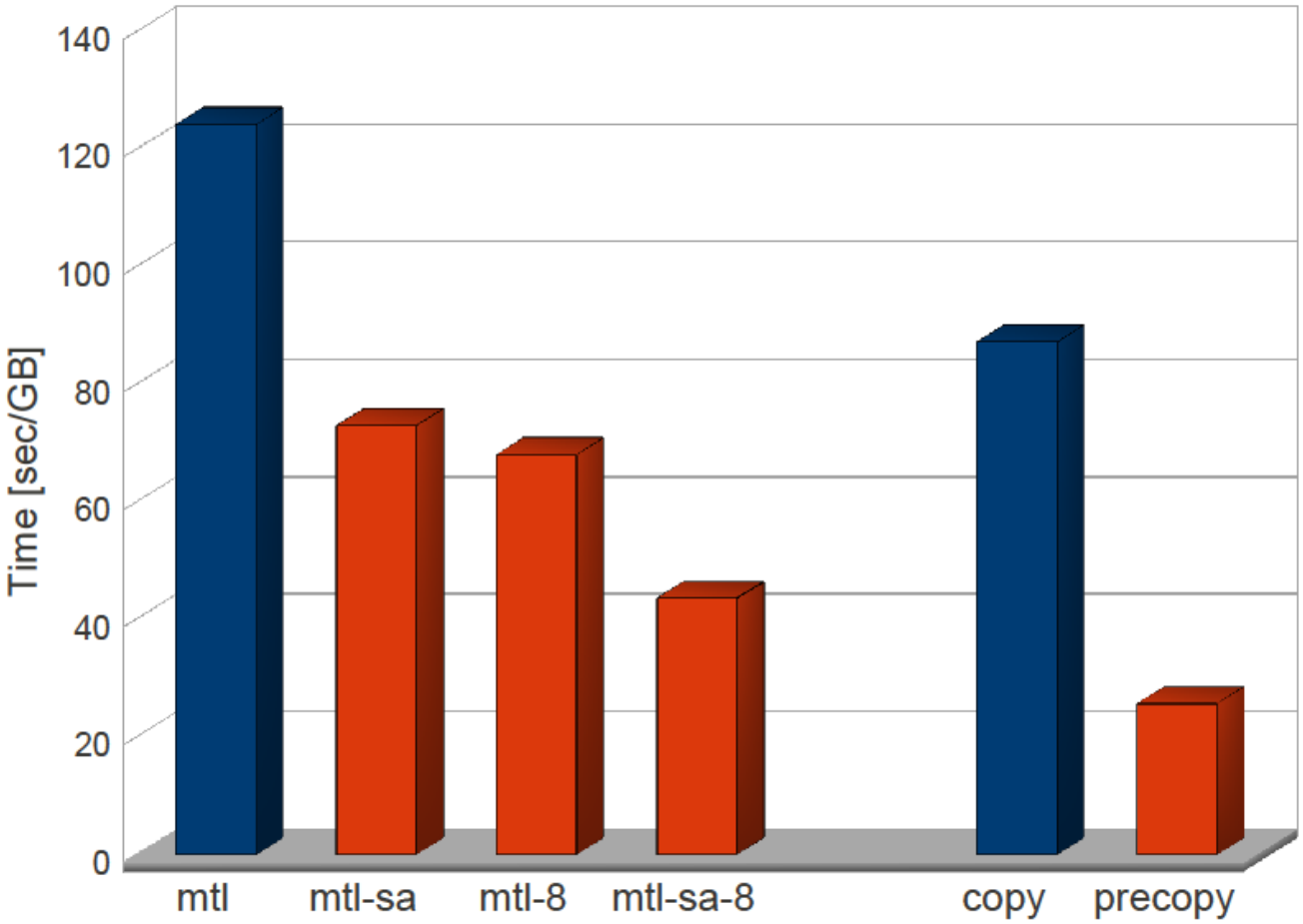
Faster Inverse BWT

POSTER

- Inverse BWT is bottleneck in decompression
  - Slow because of CPU cache misses
- Our new algorithms
  - Algorithm that halves the number of cache misses
  - Algorithm with asymptotic reduction in cache misses for highly repetitive data
  - Technique that reduces the cost of cache misses
- Based on
  - Combinatorics of BWT
  - Cache complexity analysis
  - Out-of-order execution

Faster Inverse BWT

POSTER

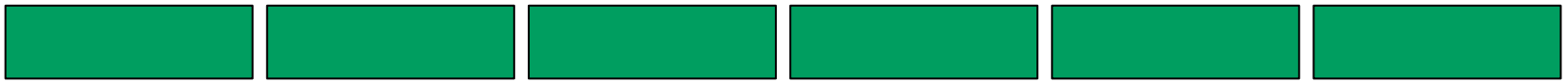


Fixed-Block Compression Boosting

- Compression boosting improves FM-index compression
  - Divide BWT into blocks in a specific way

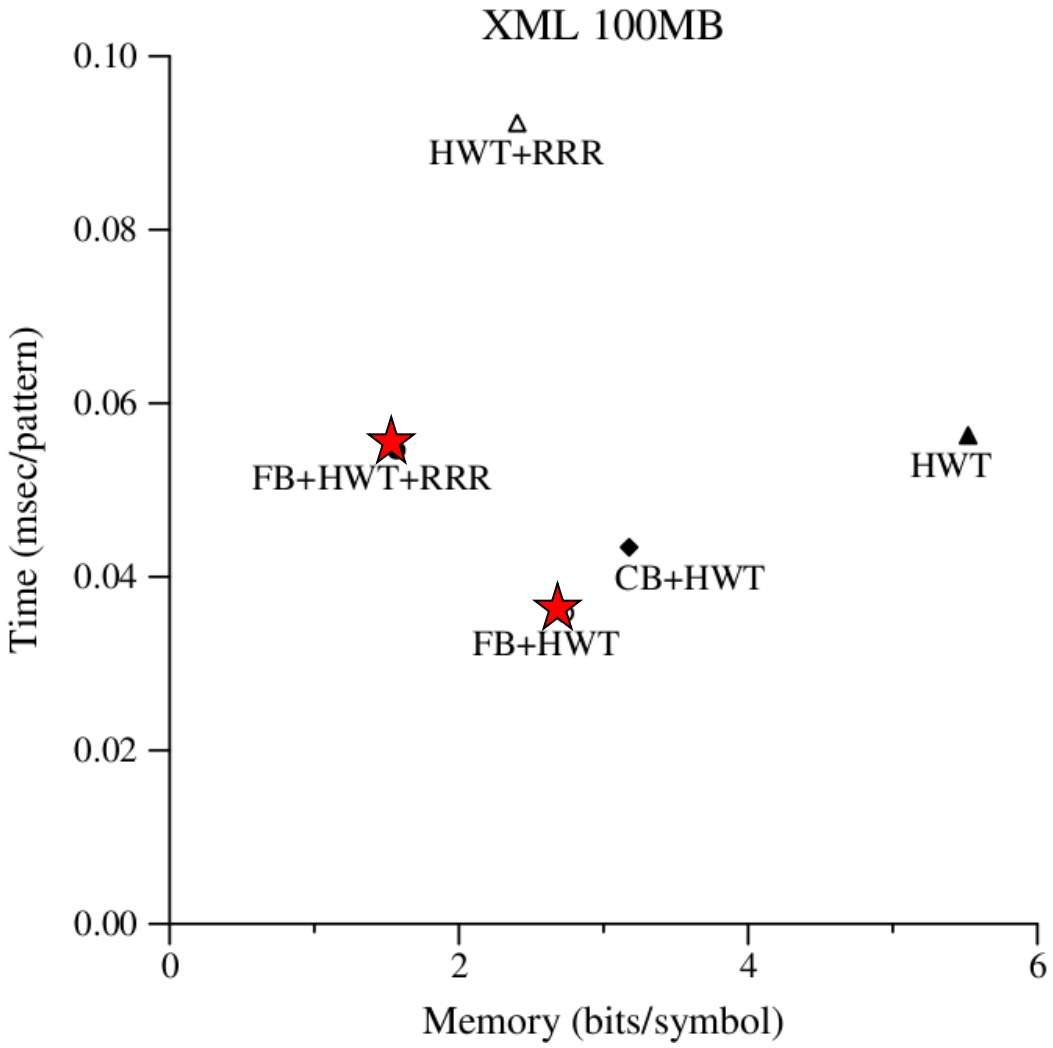
  - Compress each block separately

- We show that blocks of fixed size work as well



  - Simpler to implement, faster to construct
  - Better compression, faster queries

Fixed-Block Compression Boosting



Publications and Code

- Juha Kärkkäinen, Pekka Mikkola, Dominik Kempa. **Grammar Precompression Speeds Up Burrows-Wheeler Compression**. Submitted, 2012.
- Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi. **Slashing the Time for BWT Inversion**. 2012 Data Compression Conference.
- Juha Kärkkäinen, Simon J. Puglisi. **Fixed Block Compression Boosting in FM-indexes**. SPIRE 2011.
- https://github.com/pjmikkol/bwtc
  - Experimental, open source Burrows-Wheeler compressor

Future work

- Improve entropy coding in compressor
  - We have a good coder and a fast coder
  - We want a coder that is both good and fast
  - Completes the Burrows-Wheeler compressor
- Improve FM index entropy coding
  - Entropy coding techniques are completely different because of the need to support rank queries
  - But the same properties of BWT are exploited
  - Study both types of techniques together

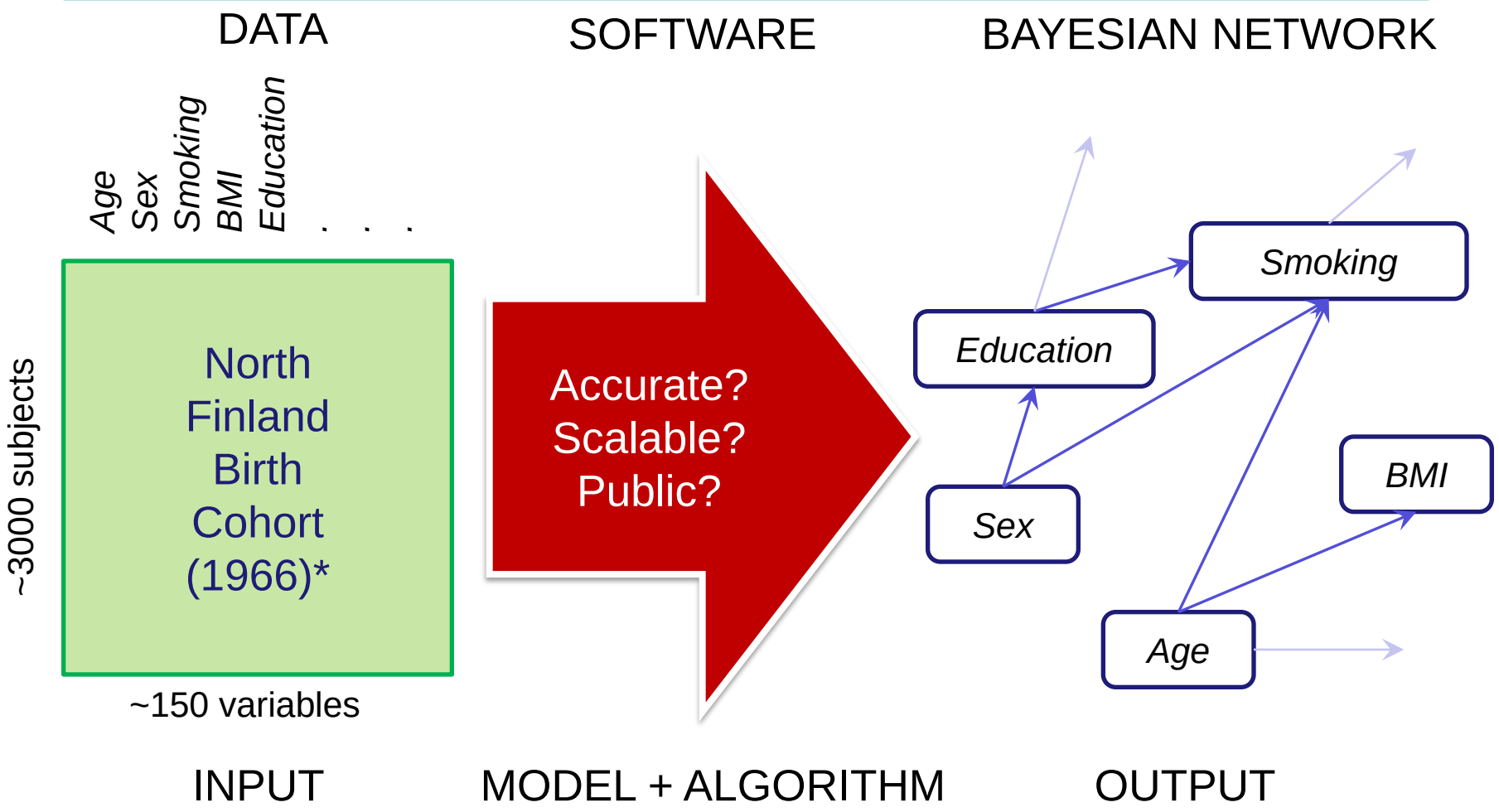
ALGODAN Algorithmic Data Analysis

Advanced Algorithms for Bayesian Network Discovery

Mikko Koivisto
CO-ALCO & Phenomics Group

Aalto University
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Objective

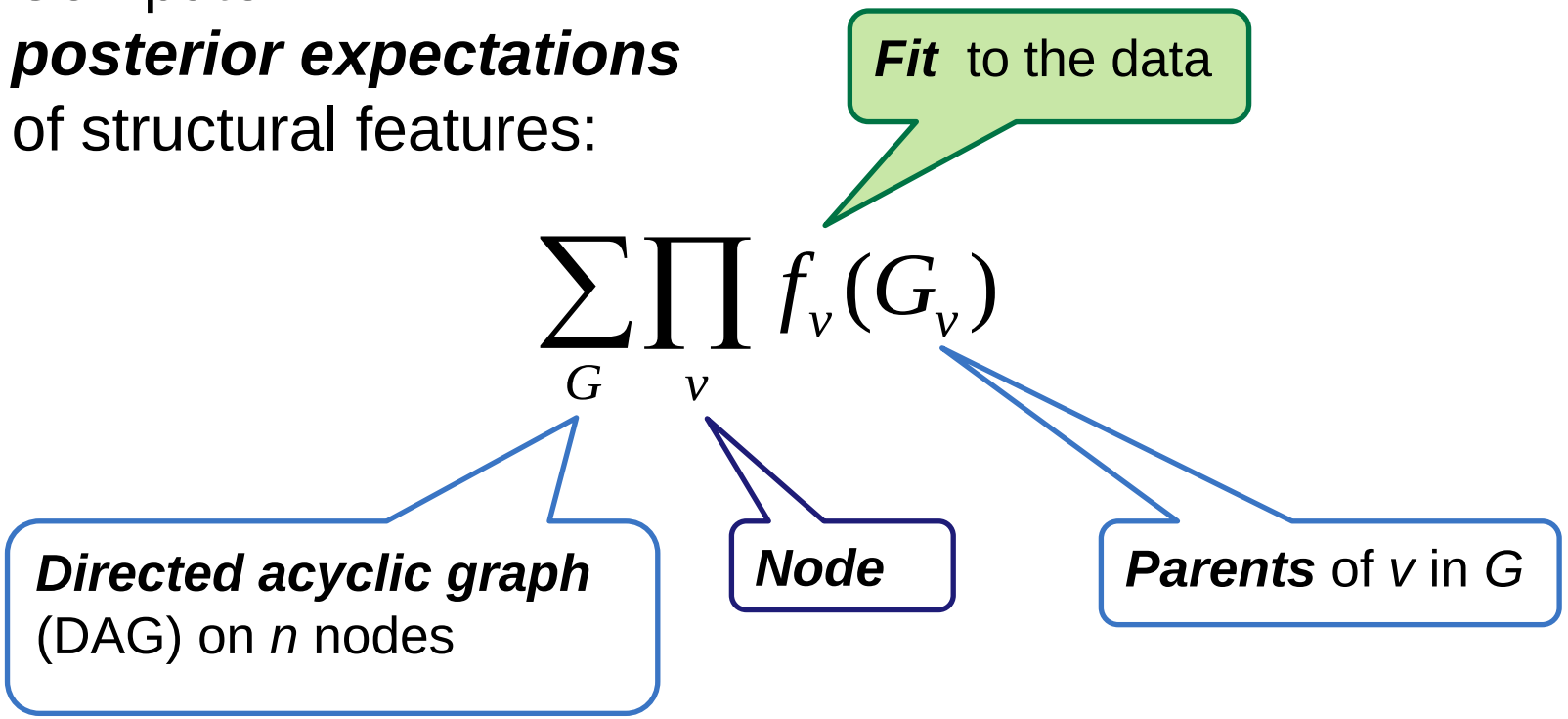


ALGODAN Algorithmic Data Analysis

*In collaboration with Dr. Tiina Paunio, Institute for Molecular Medicine in Finland & National Institute for Health and Welfare

Problem

Compute
posterior expectations
of structural features:



Idea

(Parviainen & Koivisto, AISTATS'10; Niinimäki, Parviainen, Koivisto, UAI'11)

$$\sum_{G} \prod_{v} f_v(G_v)$$

Summation splitting for tradeoffs

smaller, smoother

$$\sum_{X}$$

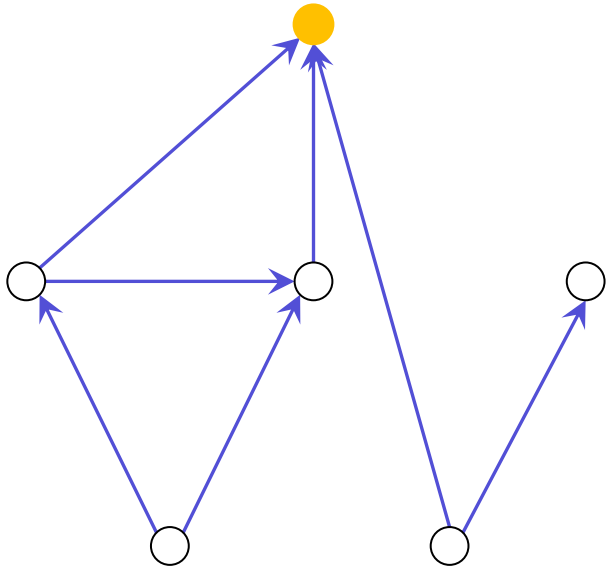
$$\sum_{G:X} \prod_{v} \tilde{f}_v(G_v)$$

faster

DAG	$O(n)$ linear
linear order	$O(n^{k+1})$ polynomial
partial order	$O(n^{k+1}+n^2 I)$ *polynomial*
void	$O(2^n n)$ exponential

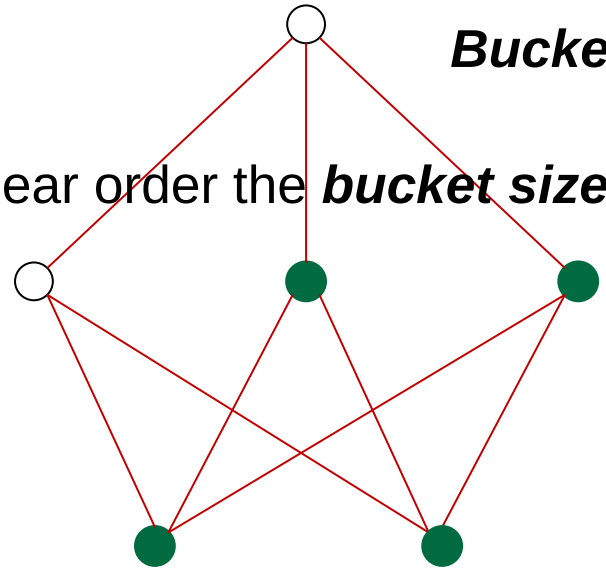
Concepts

A ***DAG*** is ***compatible with*** a ***partial order***
if they have a common linear extension



Bucket order

In a linear order the ***bucket sizes*** are one



A DAG has ***max-indegree*** k

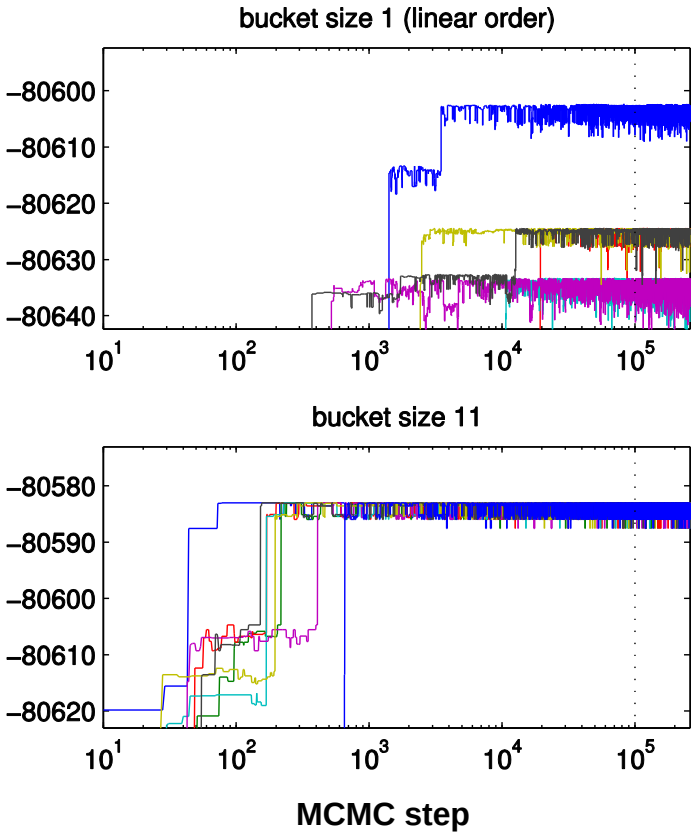
A partial order has I ***ideals***

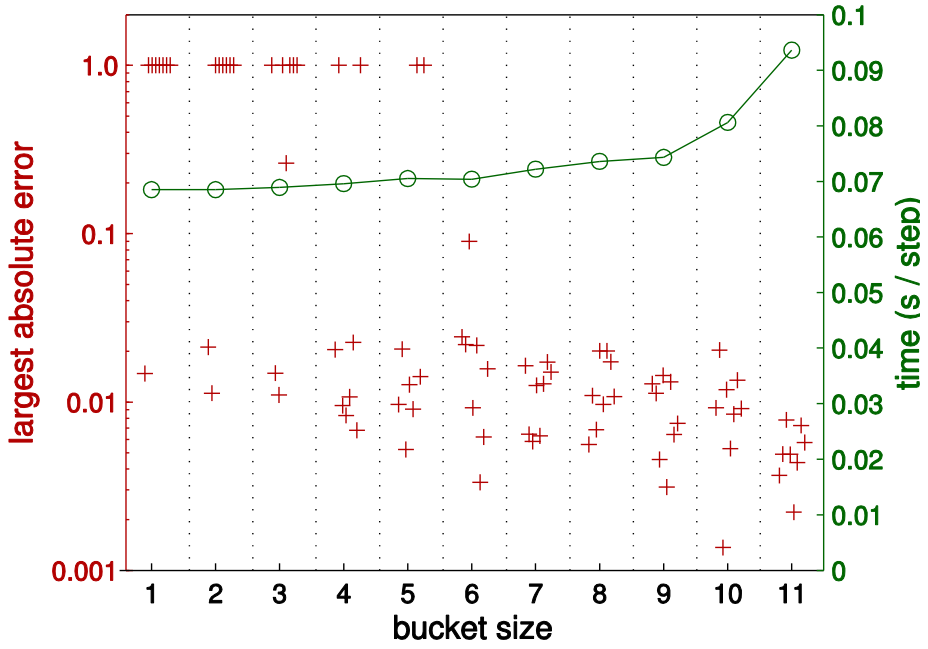
Analytics

- For a fixed partial order, time $O(n^{k+1}+n^2 I)$ suffices thanks to the ***fast sparse zeta transform***.
- ***Single bucket orders with large buckets*** yield a ***better*** tradeoff between sample space size and per-sample runtime than parallel bucket orders with smaller buckets.
- ***Bucket sizes larger than one*** reduce the sample space size substantially. A ***reasonable bucket size*** is $(k\text{-}2) \log n$.

Experiments

”Mushroom” data; n=22; k=5; eight independent MCMC runs





Conclusion

- An important and inspiring sum-product problem.
- Our Partial Order MCMC is the most efficient Bayesian method for structure discovery in Bayesian networks.
- Next: Implement some useful features to our publicly available software (BEANDisco).
- Next: Approximation guarantees?

ALGODAN Algorithmic Data Analysis

Fast Zeta Transforms for Lattices

Petteri Kaski

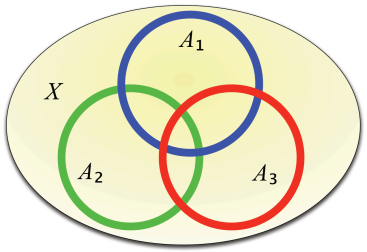
joint work with

Andreas Björklund (Lund), Thore Husfeldt (Copenhagen),
Mikko Koivisto (Helsinki), Jesper Nederlof (Utrecht),
Pekka Parviainen (Helsinki)

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Background

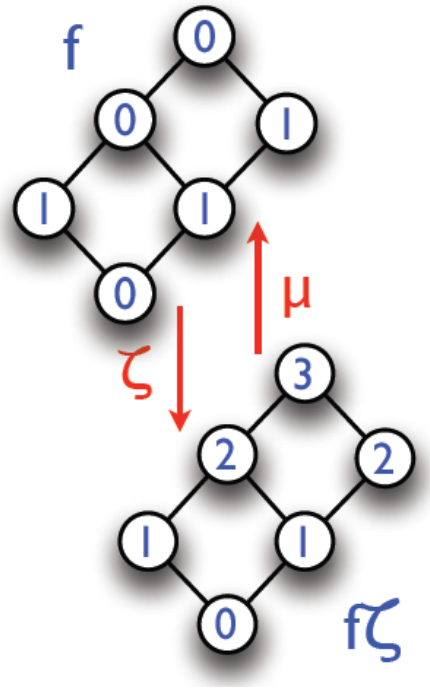


- *Möbius inversion* is a generalization of the principle of inclusion and exclusion to *partially ordered sets*

- Zeta transform ~ Fourier transform
  Möbius transform ~ inv. Fourier transform

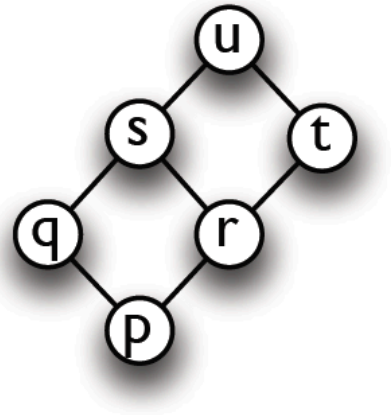
- We know that *fast* Fourier transforms exist — are there *fast* zeta/Möbius transforms?

- For *lattices*, yes (SODA 2012)



(Finite) Lattices

- *Combinatorial definition:*
  A (finite) partially ordered set $(L,\leq)$ such that
  1) there is a minimum element; and
  2) any two elements $x,y \in L$ have a least upper bound (**join**) $x \vee y$

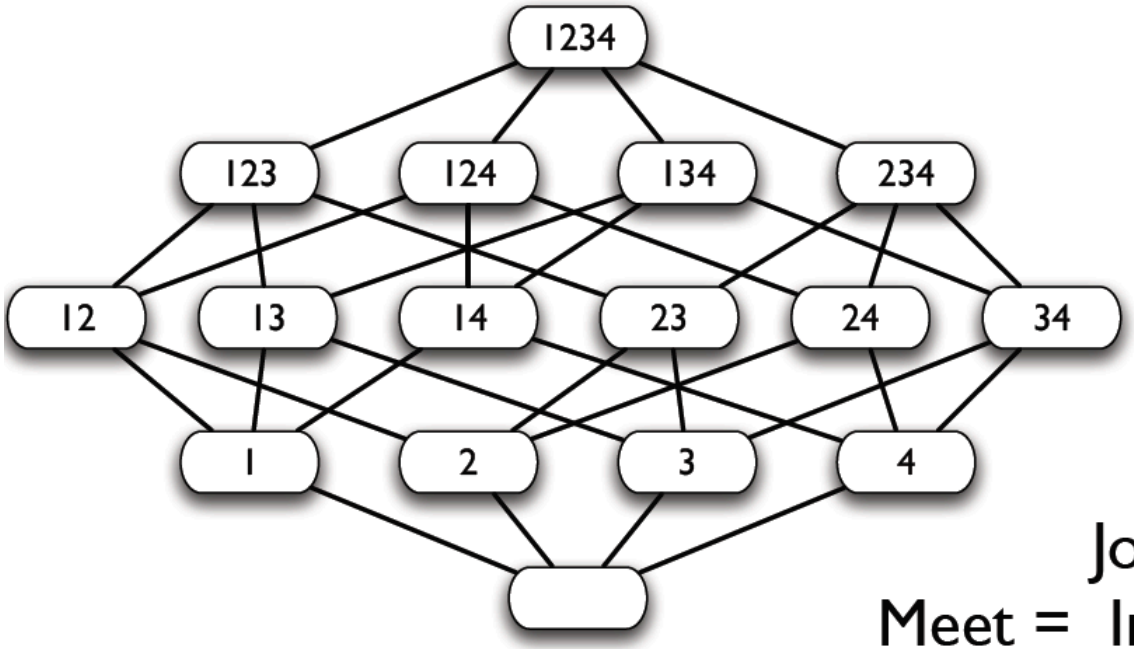


- *Algebraic definition:*
  A (finite) commutative idempotent semigroup $(L,\vee)$ with identity

$\vee$	p	q	r	s	t	u
p	p	q	r	s	t	u
q	q	q	s	s	u	u
r	r	s	r	s	t	u
s	s	s	s	s	u	u
t	t	u	t	u	t	u
u	u	u	u	u	u	u

Example: Subset Lattice

- The set of all 2^n subsets of an n-element set
- Partially ordered by subset inclusion

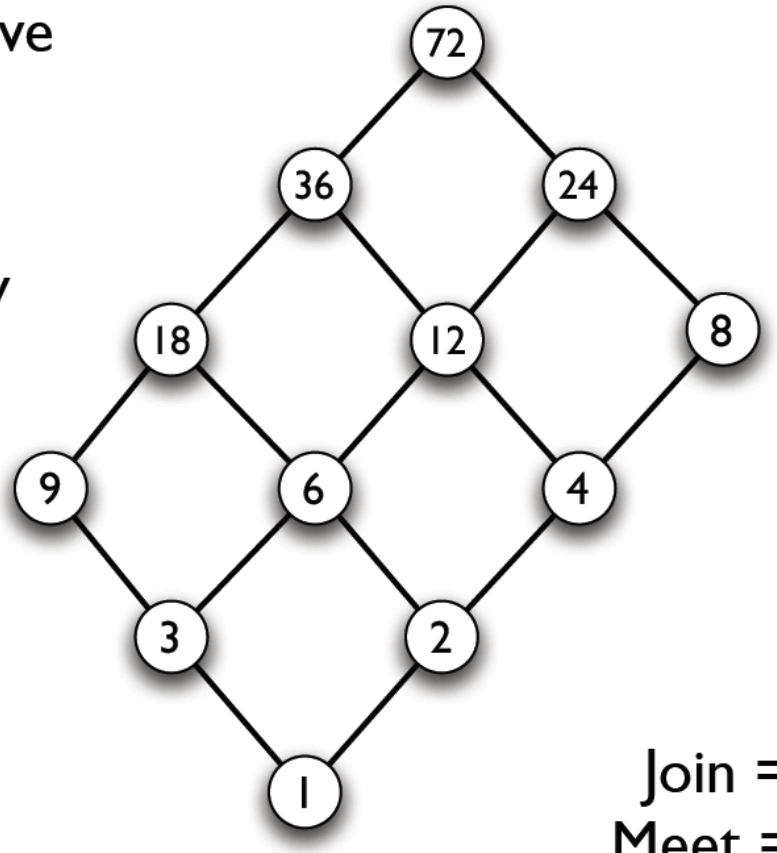


Join = Union

Meet = Intersection

Example: Divisor Lattice

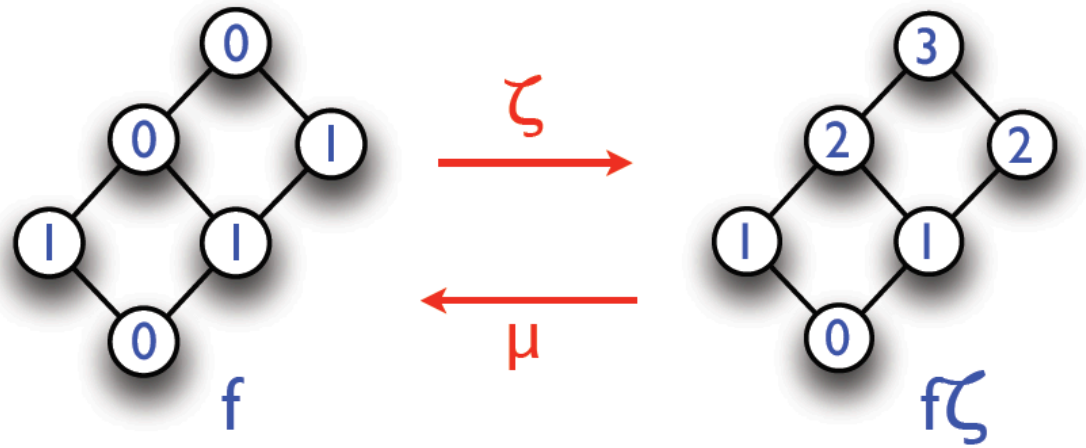
- The set of all positive divisors of a positive integer n
- Partially ordered by divisibility



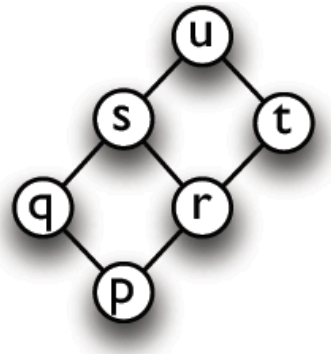
Join = lcm
Meet = gcd

Möbius Inversion [Rota]

- Let $(L, \leq)$ be a lattice
- Let be K a field
- For $f : L \to K$, define the **zeta transform** $f\zeta : L \to K$ for all $y \in L$ by $f\zeta(y) = \sum_{x \in L : x \leq y} f(x)$
- The inverse of ζ is the **Möbius transform** μ



In the Language of Linear Algebra …

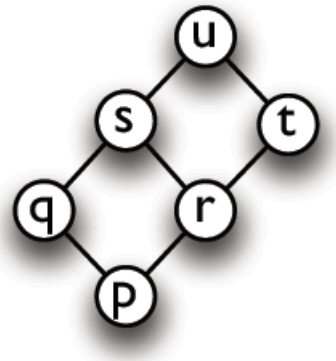


- Suppose L has v elements
- f is a row vector of length v with positions indexed by L
- ζ is a v by v matrix with
  ζ(x,y)=1 if x≤y;
  ζ(x,y)=0 otherwise
- Zeta transform:
  Right-multiply f with ζ

	p	q	r	s	t	u
f	0	0	1	1	1	0
fζ	0	1	1	2	2	3

ζ	p	q	r	s	t	u
p	1	1	1	1	1	1
q	0	1	0	1	0	1
r	0	0	1	1	1	1
s	0	0	0	1	0	1
t	0	0	0	0	1	1
u	0	0	0	0	0	1

Complexity of Evaluation



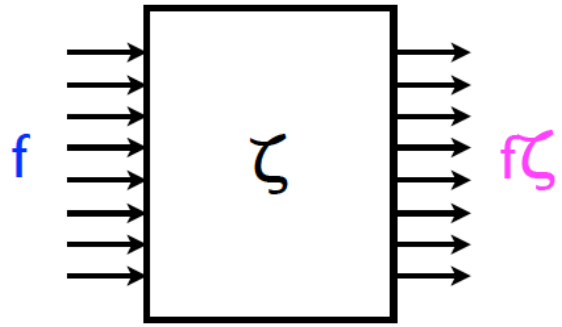
- Assume that L is fixed, $|L| = v$
- Task:
  Given $f : L \to K$ as input,
  compute $f\zeta : L \to K$
- $f\zeta$ can clearly be computed in $O(v^2)$ arithmetic operations in K
- ***But can we go faster?***

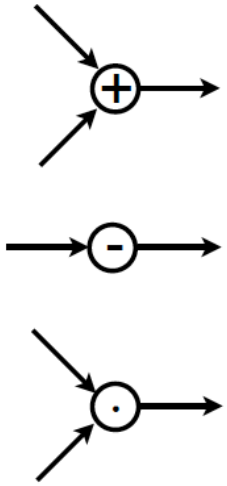
	p	q	r	s	t	u
f	0	0	1	1	1	0
$f\zeta$	?	?	?	?	?	?

ζ	p	q	r	s	t	u
p	1	1	1	1	1	1
q	0	1	0	1	0	1
r	0	0	1	1	1	1
s	0	0	0	1	0	1
t	0	0	0	0	1	1
u	0	0	0	0	0	1

Arithmetic Circuits

- How many gates are sufficient / necessary in an arithmetic circuit that computes $f\zeta$ from f ?
- Trivial circuit has $O(v^2)$ gates
  ——***but do there exist smaller circuits?***





Main Result (Björklund *et al.*, SODA12)

- Let $(L, \leq)$ be a lattice with v elements, n of which are *nonzero* and *join-irreducible*
- Then, there exist arithmetic circuits of size $O(vn)$ both for the zeta transform on L and for the Möbius transform on L
- (The claim holds also if *join-irreducible* is replaced with *meet-irreducible*)

Motivation: Many combinatorially useful lattices have $n = O(\text{polylog } v)$

Why?

- Polynomial multiplication:

  $(1x^0+1x^1+3x^2) \cdot (1x^0+2x^1) = 1x^0+3x^1+5x^2+6x^3$

- ... *fast* multiplication via the fast Fourier transform (FFT)

- "Lattice polynomial" multiplication:

  $(1\{a,b\} + 3\{c,d\}) \cup (1\{b,c\} + 2\{d\}) =$
  $= 1\{a,b,c\} + 3\{b,c,d\} + 2\{a,b,d\} + 6\{c,d\}$

- ... *fast* multiplication via the fast zeta transform & fast Möbius transform (FZT/FMT)

Applications (e.g.)

- (Currently fastest) exact algorithms for many hard problems such as graph colouring
  [Björklund, Husfeldt & Koivisto 2009]
- Constructing FFTs for inverse semigroups
  [Malandro & Rockmore 2010]
- Analysis of Markov chains on semigroups
  [Bidigare, Hanlon & Rockmore 1999;
  Brown 2000; Brown & Diaconis 1998]

Summary & Further Work

- **Main result:**
  There exist arithmetic circuits of size $O(vn)$
  for the zeta & Möbius transforms on $(L, \leq)$
  with v elements and n nonzero join-irreducibles
- Can we go faster?
  —Are there smaller circuits?
  —*Constructing* the circuits efficiently
- Is there a family of lattices L that does not admit (monotone) circuits of size $O(e)$, where e is the number of edges in the diagram of L ?
- Applications in algorithms for hard problems

ALGODAN
Algorithmic Data Analysis

Clustgrams: An Extension to Histogram Densities Based on the MDL Principle

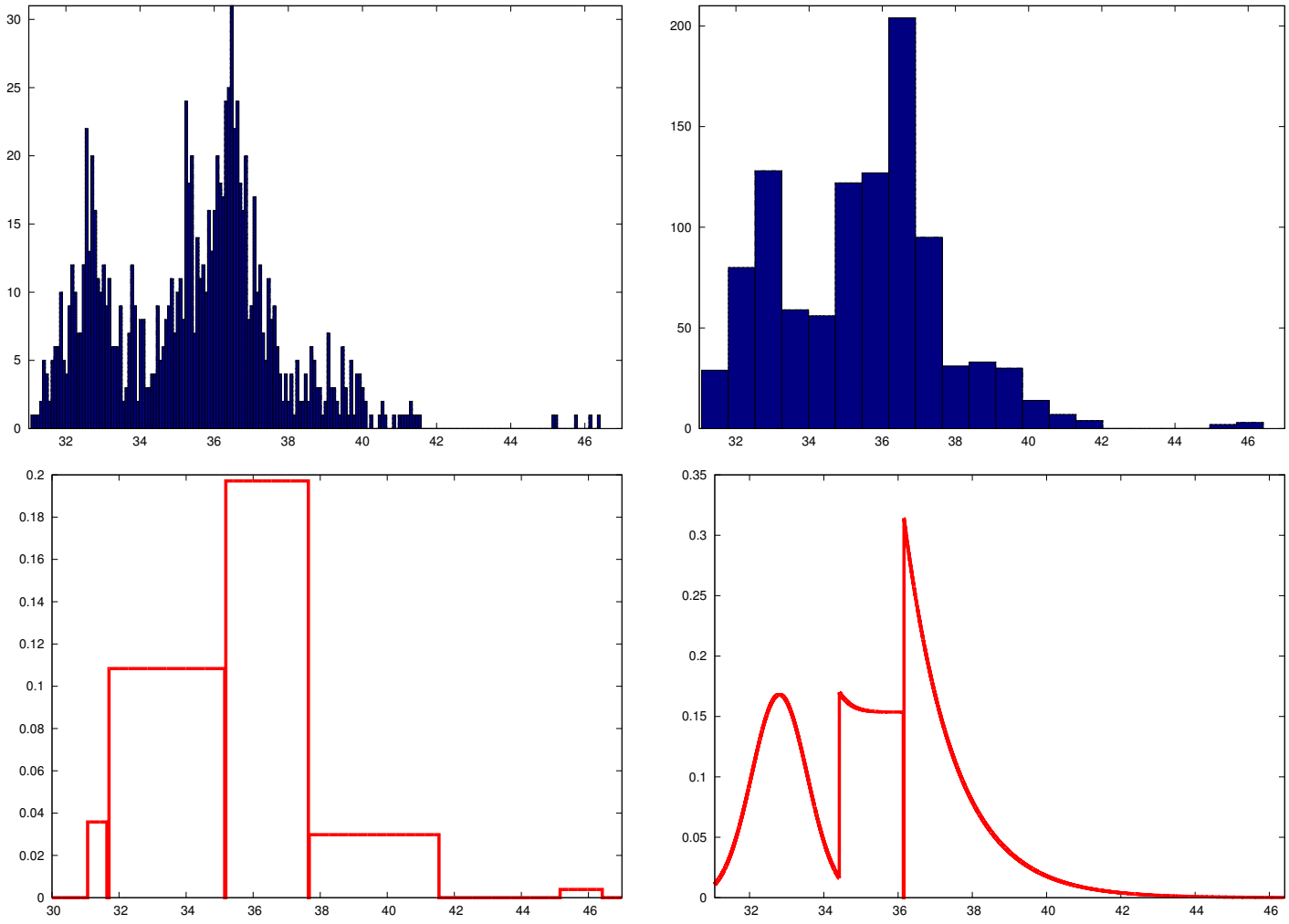
Panu Luosto, Petri Kontkanen and Kerkko Luosto

UNIVERSITY OF HELSINKI

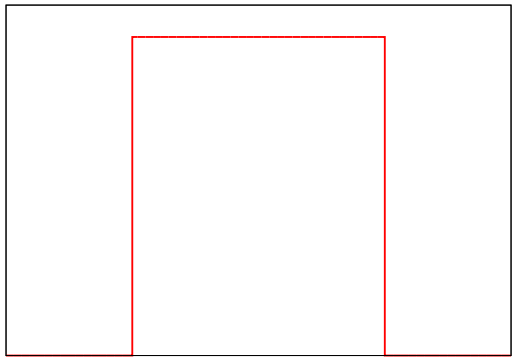
Panu Luosto and Petri Kontkanen:
Clustgrams: An extension to histogram densities based on the MDL principle (*Central European Journal on Computer Science 2011*)

Panu Luosto, Petri Kontkanen and Kerkko Luosto:
The normalized maximum likelihood distribution of the multinomial model class with positive maximum likelihood parameters (*Manuscript*)

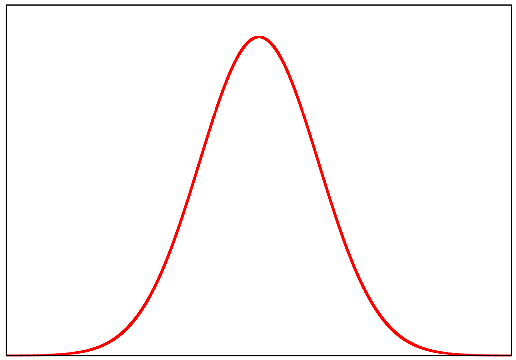
Histograms and a Clustgram



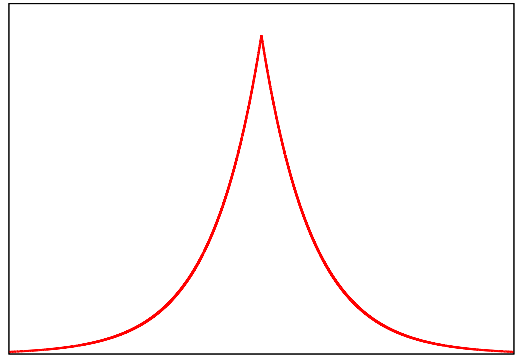
Model Classes in the Location-Scale Family



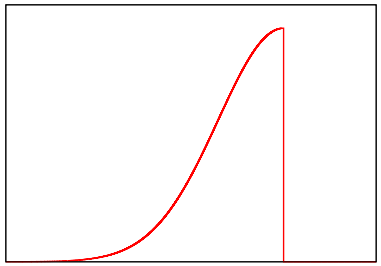
Uniform

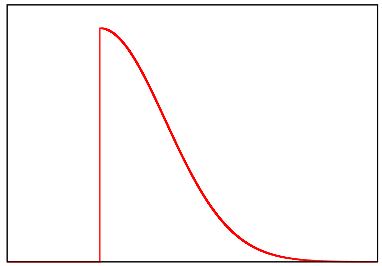


Normal

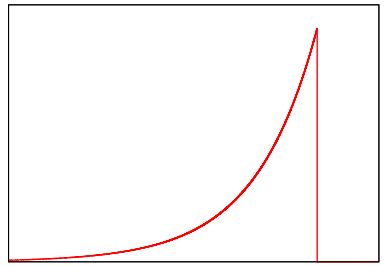


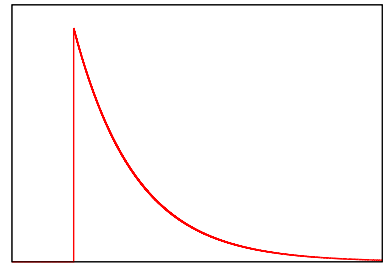
Laplace





Half-normal





Exponential

Main Problem

We would like to use normalized maximum likelihood (NML) for model selection, but...

- How to handle the problem of the infinite parametric complexity of a model class?

Contributions

- Code lengths for model classes of the location-scale family in the form

$$-\log f_{\mathrm{ML}}(x^n) + C_\epsilon - \log p(\hat{\alpha}(x^n))$$

- NML code for a clustering sequence

Clustering

In this talk,

- data are a one-dimensional sequence:
  $x^n = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$
- clusters are subsequences
- every point x_i belongs to exactly one cluster

Quality of a Clustering (Clustgram)

Problem: Given a collection of model classes (uniform, normal etc.), find

- the best number of clusters $k \in \{1, 2, \ldots, k_0\}$
- the best clustering and
- the best types of clusters

Criterion: The best clustering enables the best compression of the data and the clustering itself.

Normalized Maximum Likelihood

- Normalized maximum likelihood (NML) given a model class:

$$P_{NML}(x^n) = \frac{P(x^n;\, \hat{\theta}(x^n))}{\sum_{y^n \in X^n} P(y^n;\, \hat{\theta}(y^n))}$$

yields the code length

$$-\log P(x^n;\, \hat{\theta}(x^n)) + \underbrace{\log \sum_{y^n \in X^n} P(y^n;\, \hat{\theta}(y^n))}_{\text{parametric complexity}}$$

Inifinite Parametric Complexity

- The normalizing sum (integral) diverges in many interesting cases:

$$\underbrace{\log \sum_{y \in X} P(y;\, \hat{\theta}(y))}_{\text{parametric complexity}} = \infty$$

- Then, no worst-case optimal solution exists (uniform, normal, half-normal, exponential, Laplace)

NML with Luckiness

- Allow the excess code length to grow as an function of the ML parameters.
- NML (if it exists) has a code length

$$-\log P_{\mathrm{ML}}(x^n) + C$$

- We get

$$-\log P_{\mathrm{ML}}(x^n) + C_\epsilon - \log p(\hat{\alpha}(x^n))$$

where p is an extremely flat density based on Rissanen's prior for integers

NML for a Clustering Sequence

- Given the number fo clusters k, we know that there are *no empty clusters*
- For multinomial NML, calculating

$$\mathcal{C}_1(k,n) = \sum_{\substack{m_1+\cdots+m_k=n,\\ m_1,\ldots,m_k\geq 1}} \frac{n!}{m_1!\ldots m_k!} \prod_{i=1}^{k} \left(\frac{m_i}{n}\right)^{m_i}$$

  fast is untrivial
- Using the technique of generating functions we can prove that

$$\mathcal{C}_1(k+2,n) + 2\mathcal{C}_1(k+1,n) = \left(\frac{n}{k} - 1\right)\mathcal{C}_1(k,n)$$

Two Normalizing Sums

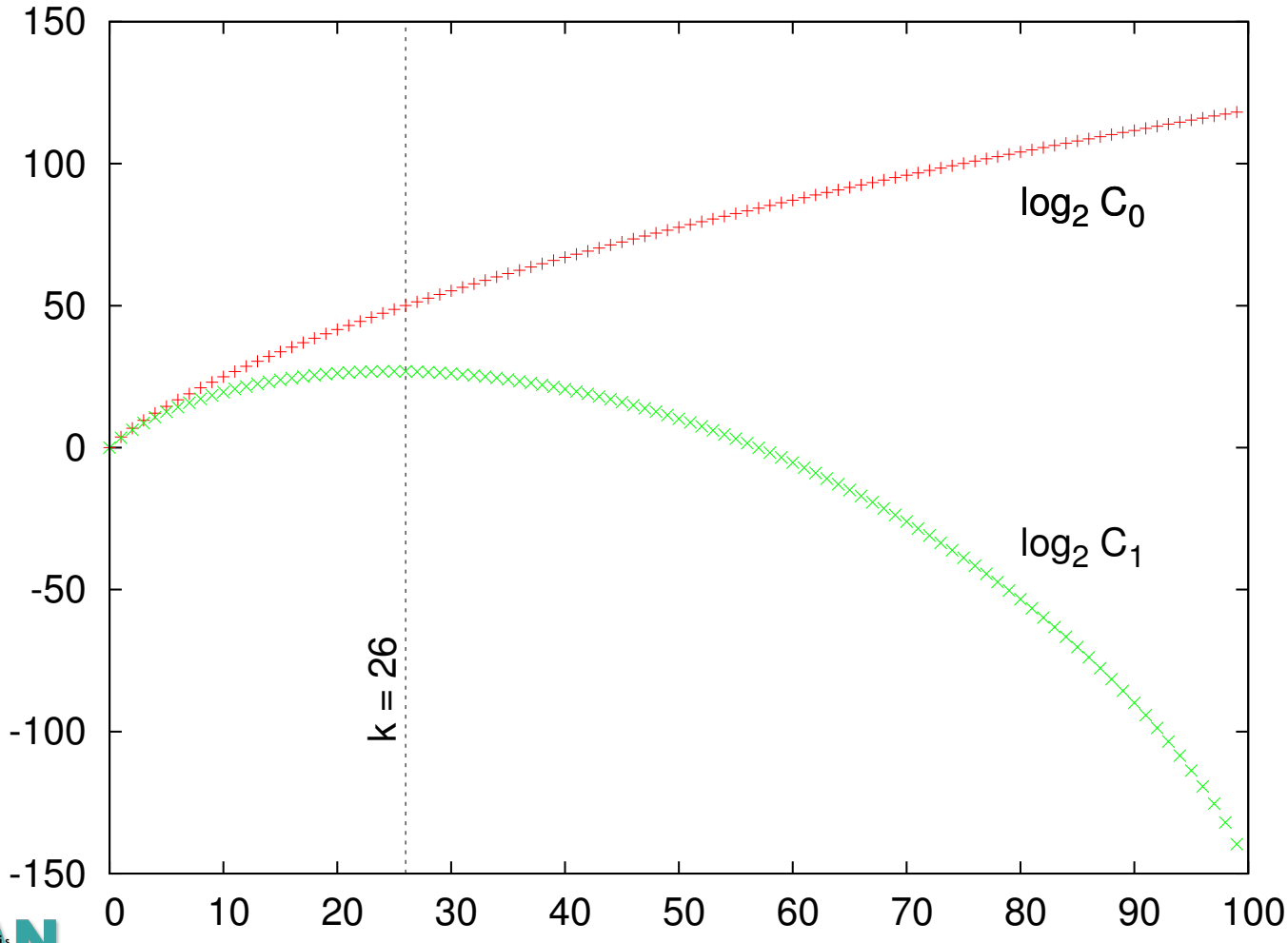
- Is there a difference in practice between

$$C_1(k,n) = \sum_{\substack{m_1+\cdots+m_k=n,\\ m_1,\ldots,m_k\geq 1}} \frac{n!}{m_1!\ldots m_k!} \prod_{i=1}^{k} \left(\frac{m_i}{n}\right)^{m_i}$$

and

$$C_0(k,n) = \sum_{\substack{m_1+\cdots+m_k=n,\\ m_1,\ldots,m_k\geq 0}} \frac{n!}{m_1!\ldots m_k!} \prod_{i=1}^{k} \left(\frac{m_i}{n}\right)^{m_i} \quad ?$$

Parametric Complexity as a Function of k With a Fixed $n = 100$



ALGODAN Algorithmic Data Analysis

Neuroinformatics

Overview, incl. brain imaging

Aapo Hyvärinen

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

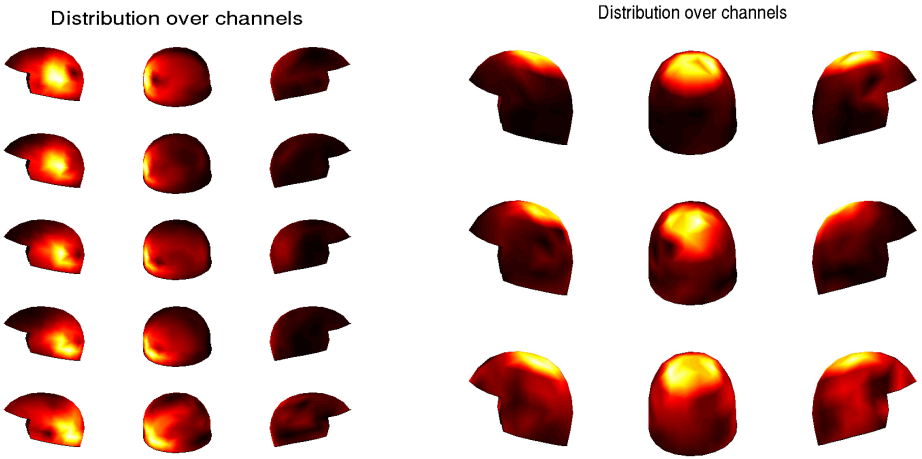
Neuroinformatics Team

- Mission:
  - Develop statistical data analysis methods, with focus on
    - Unsupervised machine learning methods
    - Neuroscience applications
  - Non-Gaussianity a central theoretical framework

- Members:
  - Aapo Hyvärinen, leader
  - Patrik Hoyer, co-leader
  - 4 postdocs, 3 PhD students
  - From 2012, 30% in CoE of Inverse Problems Research

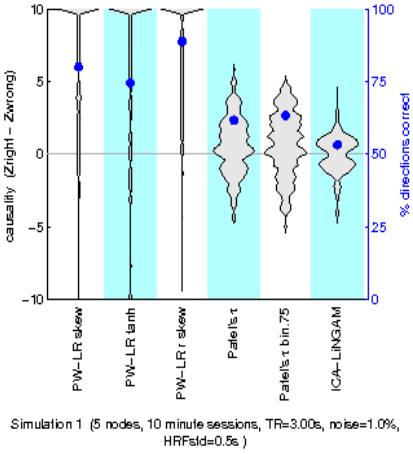
Highlight 1:
Testing independent components

- In independent component analysis, testing almost inexistent
  - Components could be local minima, or random effects
- We developed a method which uses a proper null hypothesis and the theory of classical hypothesis testing (*NeuroImage*, 2011).
  - Do ICA on multiple datasets (e.g. subjects), and see if you get the same component in more than one data set
  - Applications in MEG here:



Highlight 2: Connectivity (causality) in fMRI

- Goal is to tailor our causal analysis framework for fMRI
  - Adapt nonlinearities to the specific distributions in fMRI
  - Develop methods which work with few data points
- Jointly with Stephen Smith
  - Oxford Centre for Functional Imaging of the Human Brain
  - Developer of simulated data for comparing algorithms (*NeuroImage*, 2011)
- Our methods (under revision for JMLR)
  - Have best performance on simulated data
  - Are particularly simple, based on sign of $E\, g(x)\, y - E\, x\, g(y)$ where g is a nonlinearity, such as $g(u)=u^2$



ALGODAN Algorithmic Data Analysis

Learning linear cyclic causal models with latent variables

Patrik Hoyer
Academy Research Fellow
Neuroinformatics group

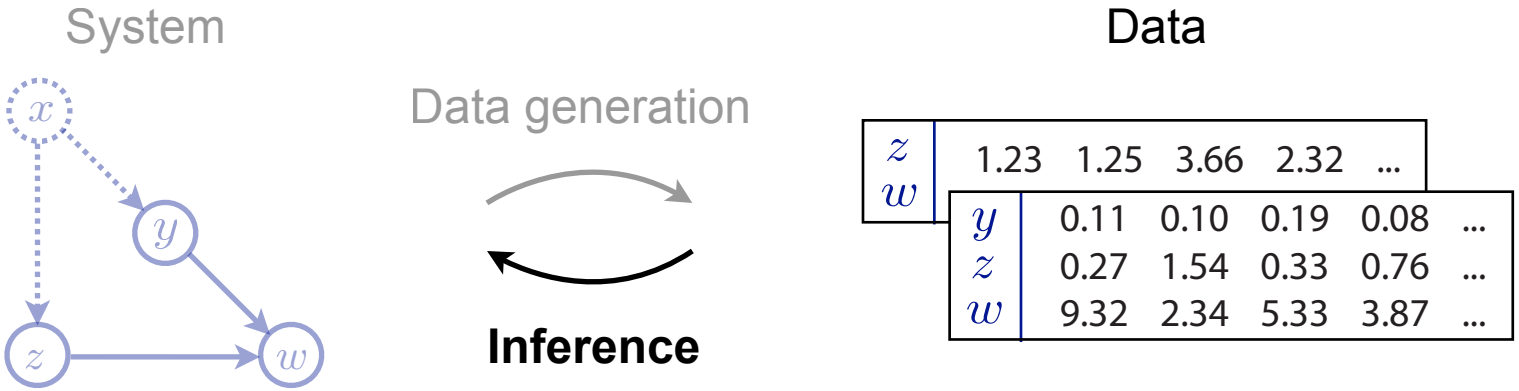
Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Causal discovery

- General problem

(Pearl, 2000; Spirtes et al, 2000)



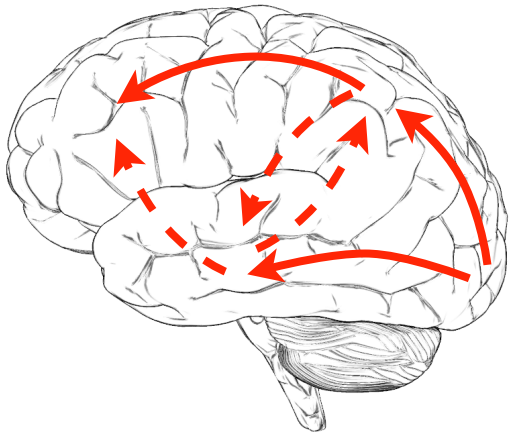
- Using what type of data, what kind of algorithms, and under what assumptions on the underlying system can we recover the data-generating process in the large sample limit? What procedures work well with realistic sample sizes?

Causal discovery

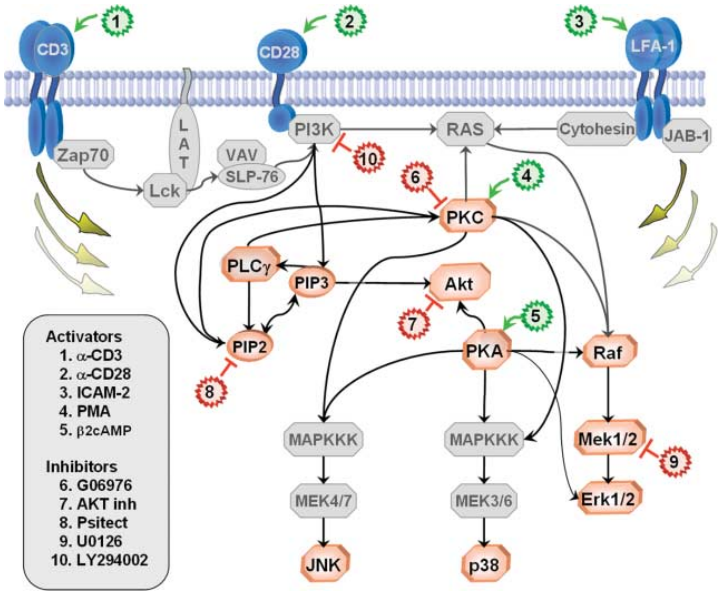
- Examples

(e.g. Ramsey et al, 2011; Sachs et al, 2005)

Neuroinformatics



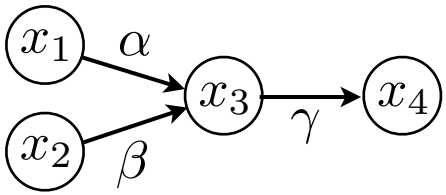
Bioinformatics



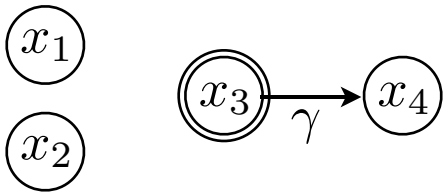
Linear cyclic model with latent variables

(Hyttinen, Eberhardt, and Hoyer, JMLR, minor revisions)

- Model:
  - Observed variables $\mathcal{V}$ grouped into a vector $\mathbf{x}$
  - Exogenous input given by the vector $\mathbf{e}$ (correlations allowed!)
  - Linear generative model $\mathbf{x} := \mathbf{B}\mathbf{x} + \mathbf{e}$
  - Interventions 'cut' all arrows *into* that variable



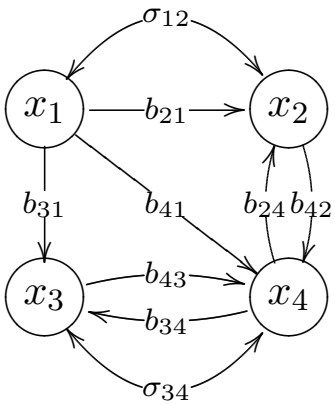
$$
\begin{aligned}
x_1 &= e_1 \\
x_2 &= e_2 \\
x_3 &= \alpha x_1 + \beta x_2 + e_3 \\
x_4 &= \gamma x_3 + e_4
\end{aligned}
$$



$$
\begin{aligned}
x_1 &= e_1 \\
x_2 &= e_2 \\
x_3 &= c_3 \\
x_4 &= \gamma x_3 + e_4
\end{aligned}
$$

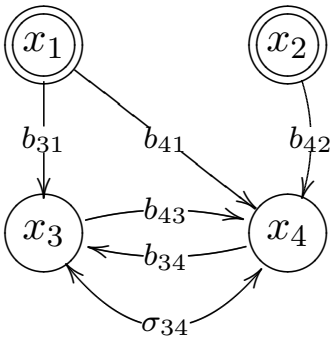
Linear cyclic model with latent variables

- Example:



$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ b_{21} & 0 & 0 & b_{24} \\ b_{31} & 0 & 0 & b_{34} \\ b_{41} & b_{42} & b_{43} & 0 \end{bmatrix} \qquad \mathbf{\Sigma_e} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

- Intervening on x_1 and x_2:



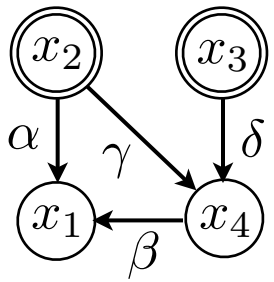
$$\mathbf{U}_k\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ b_{31} & 0 & 0 & b_{34} \\ b_{41} & b_{42} & b_{43} & 0 \end{bmatrix} \qquad \mathbf{U}_k\mathbf{\Sigma_e}\mathbf{U}_k + \mathbf{J}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

Tools for discovery

- 'Passive observational' data
  - Without additional assumptions, the model is not identifiable if we do not intervene on the variables (i.e. if we only 'passively observe' them)
- Randomized controlled experiments
  - 'Gold standard' for learning causal discovery
  - In experiment $\mathcal{E}_k$, randomize the values of one or more variables $\mathcal{J}_k \subseteq \mathcal{V}$ and passively observe the remaining ones $\mathcal{U}_k = \mathcal{V} \setminus \mathcal{J}_k$
  - Allows computing *experimental effects* of each intervened variable onto each passively observed variable (see next slide)

Experimental effects

- 'Experimental effects' measure the influence of each intervened variable on each passively observed variable. Example:



$$t(x_2 \rightsquigarrow x_4 \parallel \{x_2, x_3\}) = \gamma$$
$$t(x_2 \rightsquigarrow x_1 \parallel \{x_2, x_3\}) = \alpha + \gamma\beta$$
$$t(x_3 \rightsquigarrow x_4 \parallel \{x_2, x_3\}) = \delta$$
$$t(x_3 \rightsquigarrow x_1 \parallel \{x_2, x_3\}) = \delta\beta$$

(Note: These are not influenced by latent variables)

- We can derive linear constraints on the direct effects, based on the measured experimental effects:

$$t(x_j \rightsquigarrow x_u \parallel \mathcal{J}_k) = b_{uj} + \sum_{x_i \in \mathcal{U}_k} t(x_j \rightsquigarrow x_i \parallel \mathcal{J}_k)\, b_{ui}$$

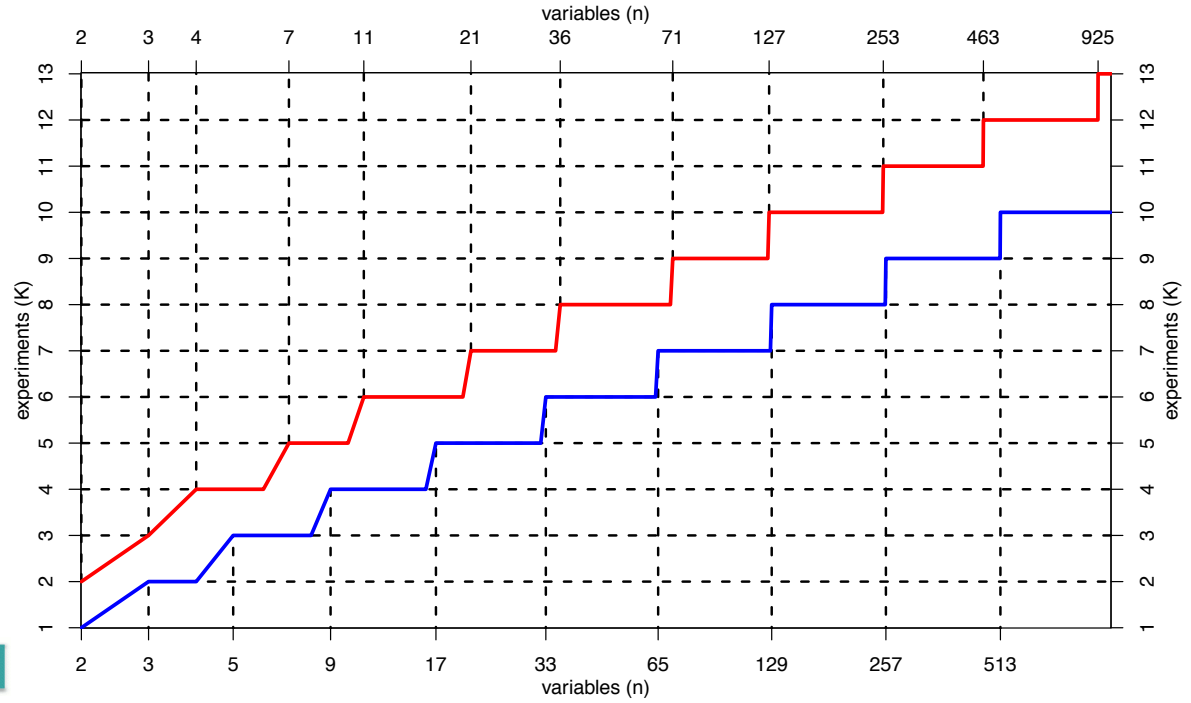
(The constraints are valid for cyclic networks as well, and are not influenced by hidden variables.)

Algorithm & identifiability results

- Algorithm:
  - Collect all constraints on the direct effects from all experiments
  - Solve the resulting linear system to estimate the coefficients
- Theorem 1:
  - The outlined procedure identifies the generating model if and only if for each ordered pair (x_i, x_j) there exists at least one experiment in which x_i is intervened on and x_j is passively observed.
- Theorem 2:
  - No procedure can identify the generating model unless the condition of Theorem 1 is satisfied.

Experiment selection

- What sets of experiments satisfy the condition?
  - n experiments each intervening on a single variable
  - n experiments each intervening on all but a single variable
  - $O(\log n)$ experiments for optimal designs!



Summary and outlook

- Summary (Hyttinen, Eberhardt, and Hoyer, JMLR, minor revisions)
  - Model class: linear cyclic models with latent variables
  - Data needed: experiments intervening on the variables
  - Learning algorithm: sound and complete
  - Identifiability conditions, selecting experiments
  - Application to DREAM challenge data (check the paper!)
- Outlook
  - Extends to 'noisy-or' binary variables (poster!)
  - Background knowledge and faithfulness (poster!)
  - Recent work on 'overlapping' datasets, and on non-parametric approaches

ALGODAN

Algorithmic Data Analysis

Estimation of unnormalized probabilistic models

Michael Gutmann
Postdoctoral researcher
Neuroinformatics group

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Motivation

- Data analysis often requires estimating the parameters θ of a probabilistic model $p(x|\theta)$.
- A popular approach is to choose θ such that the probability of the data is maximized (MLE).
  - MLE is only applicable if $\int p(x|\theta)dx=1$ for all θ.
  - Many models do not integrate to one (e.g. Markov random fields). They are unnormalized.
- Normalizing unnormalized models is computationally expensive in high dimensions (curse of dimensionality).

Goal

Estimation procedure

- applicable to unnormalized models
- with a good trade-off between statistical and computational efficiency

Statistical efficiency: small estimation errors

Computational efficiency: short running times

Proposed procedure (basic idea)

- Intuitively, knowing
  - (1) the properties of a random variable y
  - (2) the differences between x and y

allows you to infer the properties of x.

- More concretely,
  - (1) Choosing a random variable y with known p(y) where sampling is easy
  - (2) Performing logistic regression on the samples from x and y, with $p(x|\theta)/p(y)$ in the regression function

allows you to estimate the parameters θ.

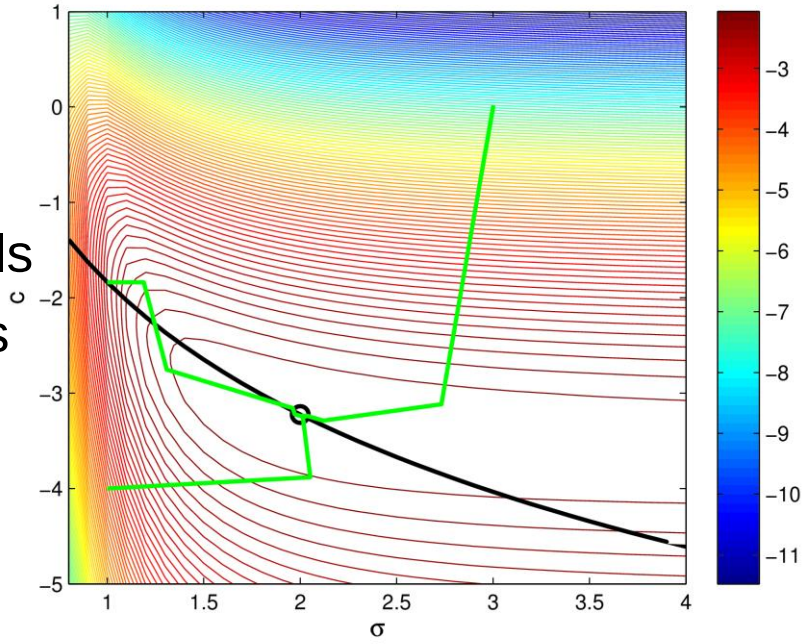
- See next slide: $p(x|\theta)$ can be unnormalized!

(Gutmann & Hyvärinen, JMLR2012; Alternatives to logistic regr.: Pihlaja et al, UAI2010)

Toy example

- Gaussian data with standard deviation σ=2:
  - Unnormalized model: ln $p(x|\sigma,c) = -||x||^2/2\sigma^2 + c$
  - To be estimated: σ, c (normalizing parameter)
  - Auxiliary “noise” distribution p(y): standard Normal

- Contour plot of objective in logistic regression
  - Black: loci of normalized models
  - Green: optimization trajectories
  - Circle: optimum



Properties

- Statistical efficiency:
  - The estimates are consistent even if $p(x|\theta)$ is unnormalized.

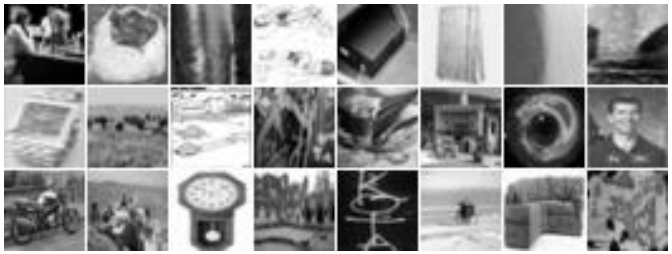
    (There are mild conditions on the noise distribution p(y).)
  - For large ratios ν of noise to data sample size, the estimates become as good as those obtained with MLE.

    (asymptotic Fisher efficiency)
- Computational efficiency:
  - Logistic regression is performed by solving a well defined unconstrained optimization problem. Standard optimization tools are applicable.
  - The ratio ν can be used to control computational complexity.

Application in the modeling of images

- Data: 80 million complete visual scenes (size: 32x32)
- Hierarchical model with three feature extraction layers

  (conference submission)



- After learning:
  - 1st and 2nd layer: “Invariant” detection of edges
  - 3rd layer: Features with enhanced selectivity to orientation/ space; Descriptors of overall image properties?

For three 3rd layer features, images giving maximal (top) and minimal activation (bottom)

Summary

- Problem studied:

  Estimation of unnormalized probabilistic models
- Relevance:
  - Estimation of probabilistic models is ubiquitous in data analysis.
  - Many advanced probabilistic models are unnormalized.
- Our solution: (Gutmann & Hyvärinen, JMLR2012)
  - Based on logistic regression between data and artificial “noise”
  - Good statistical and computational properties
  - Connection between supervised and unsupervised learning
  - Allowed us to formulate and estimate novel models for images
  - Inspired the formulation of a general estimation framework

  (Pihlaja, Gutmann & Hyvärinen, UAI2010; Gutmann & Hirayama, UAI2011 →poster)