Linear Space Direct Pattern Sampling using Coupling From The Past

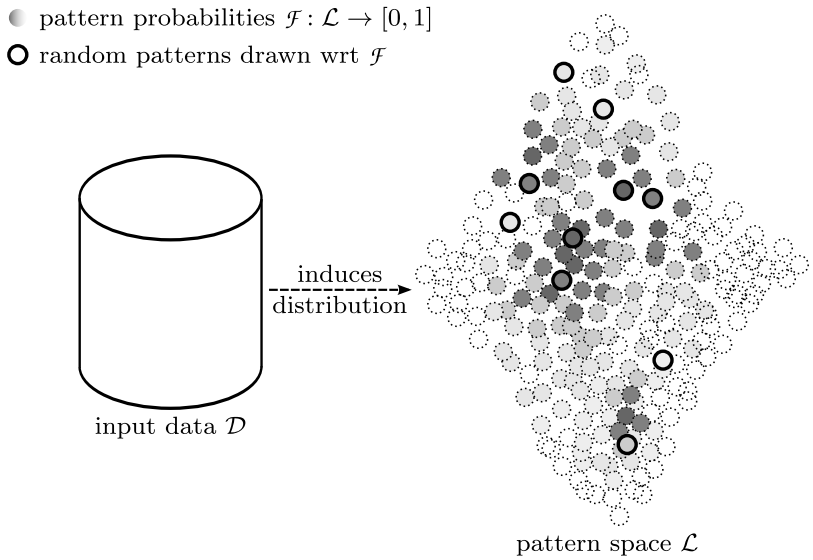
Mario Boley Sandy Moens Thomas Gärtner

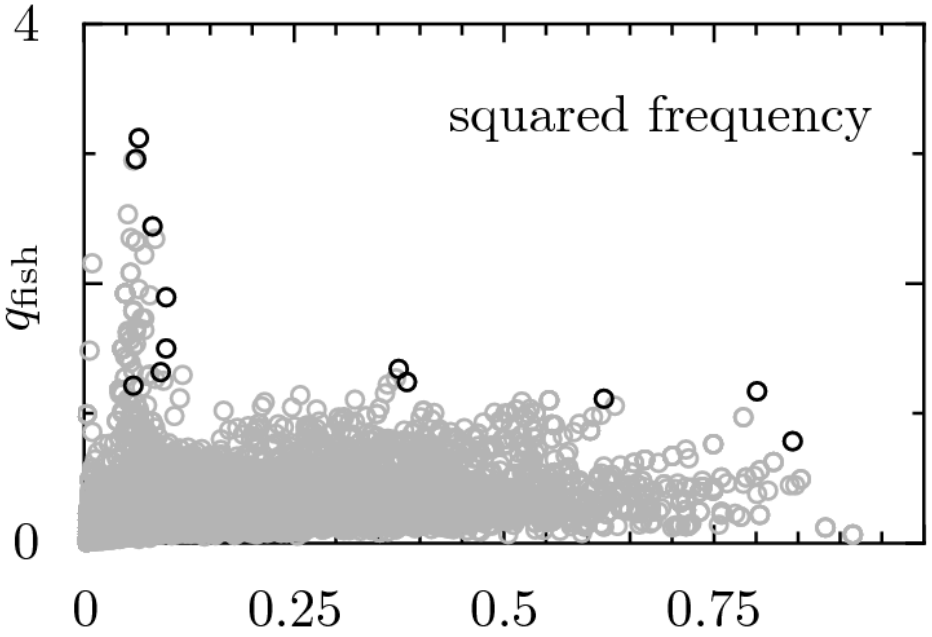
Why Sampling?

- Problems with existing Pattern Mining
  - Finds many-many patterns
  - Finds many redundant patterns
  - Interestingness ≠ support

- Possible solutions:
  - Condensed representations, e.g. closed itemsets
  - Pattern set mining, e.g. KRIMP
  - Pattern sampling, e.g. this work

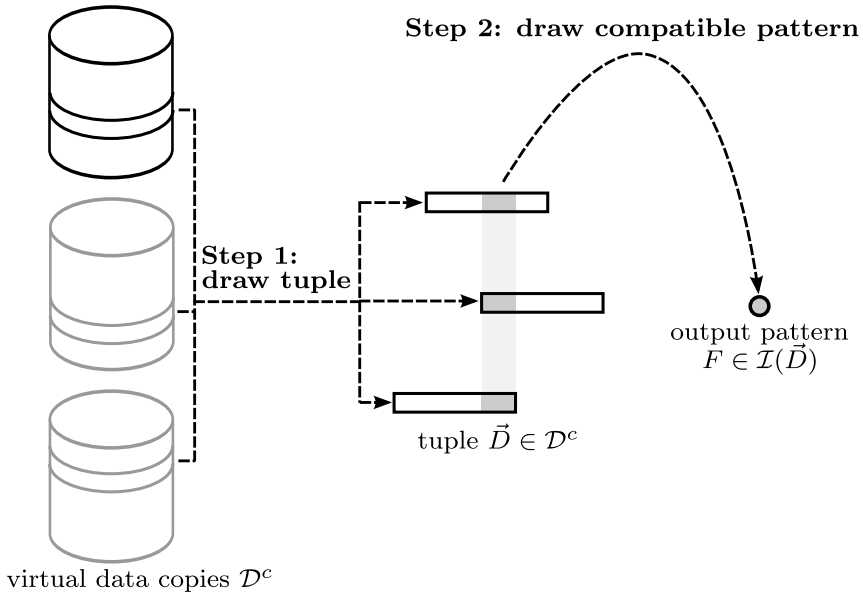
What is Pattern Sampling?





Two-Step Random Procedures

- Step 1:
  - Draw data tuples w.r.t. induced patterns
- Step 2:
  - Generate sample from tuples



Old Approach (Boley et al. 2011)

- Step 1: Logarithmic search through cumulative probability vector
  - Large pre-processing time
  - Super-linear space requirements
- Step 2: Specific sample generation depending on distribution

New General Framework

- Suitably biased patterns: $\mathcal{F}(F) = b_\star(F) \prod_{i=1}^{c} q_i(\mathcal{D}_i, F)/Z$
- Step 1: Coupling From The Past (CFTP)
  - No pre-processing time
  - Linear space requirements
- Step 2: Sequential sampling based on induced patterns

Step 1: Coupling From The Past (Huber, 2004)

- Simulation of Markov Chain
  - Exact sampling from distribution
- Backward simulation using increasing epoch lengths
  - vs. unbounded forward simulation
- Key-issue: detecting coalescence
  - i.e. monitoring for single-state
  - Intuition: Hard-Jupp

Step 2: Generate Sample

- Sequential sampling based on induced patterns
- Idea:
  - For each singleton e evaluate

$$\frac{\#\text{induced patterns including e}}{\#\text{induced patterns by data tuple}}$$

Optimalizations

- Singleton rejection
- Lower bound instead of Hard-Jupp
- Enhanced proposal function
  - Non-uniform

Evaluation

- Running time
- Proposal function
  - Irregularity
- Unsupervised performance
  - KRIMP experiments
- Supervised performance
  - Pattern-based classification

Future work

- Good distributions vs. number of factors
- Stratified techniques
- Large datasets
- …

Questions???