

October 28, 2024



Inter-University Research Institute Corporation /
Research Organization of Information and Systems

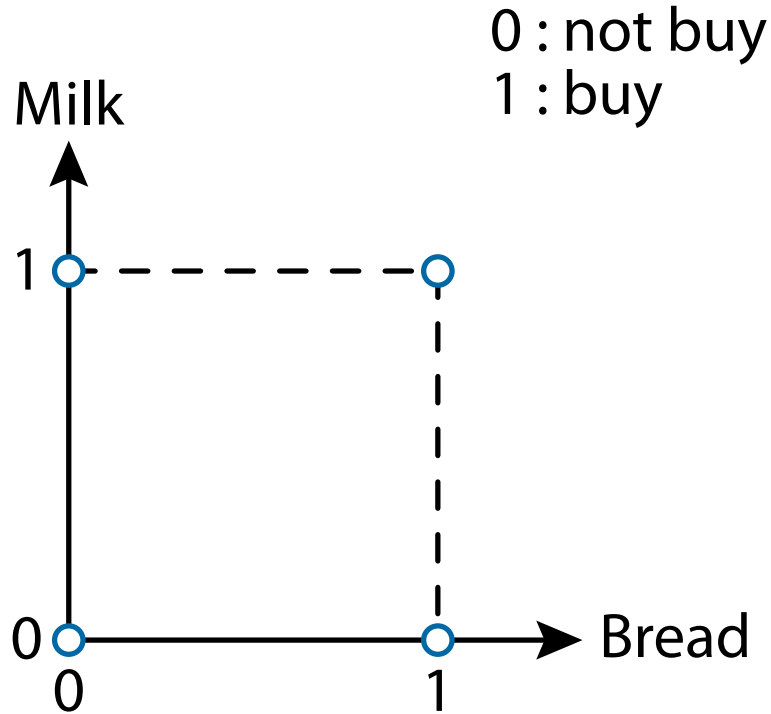
National Institute of Informatics

Pattern Discovery and Generative Models

Introduction to Intelligent Systems Science II

Mahito Sugiyama

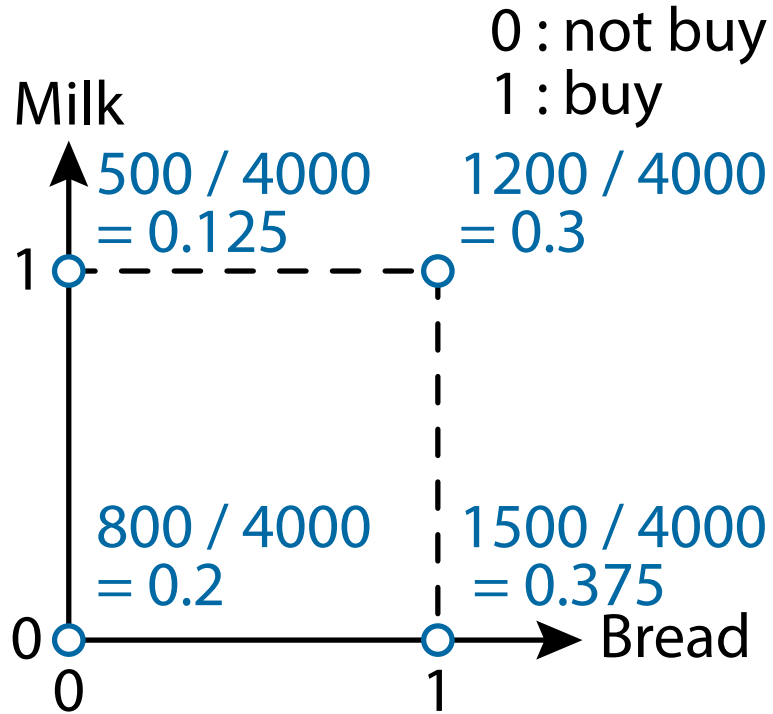
Binary Data → Pattern Mining



- Dataset:

	Bread	Milk
1	0	1
2	1	1
3	1	1
4	1	0
⋮	⋮	⋮
4000	1	0

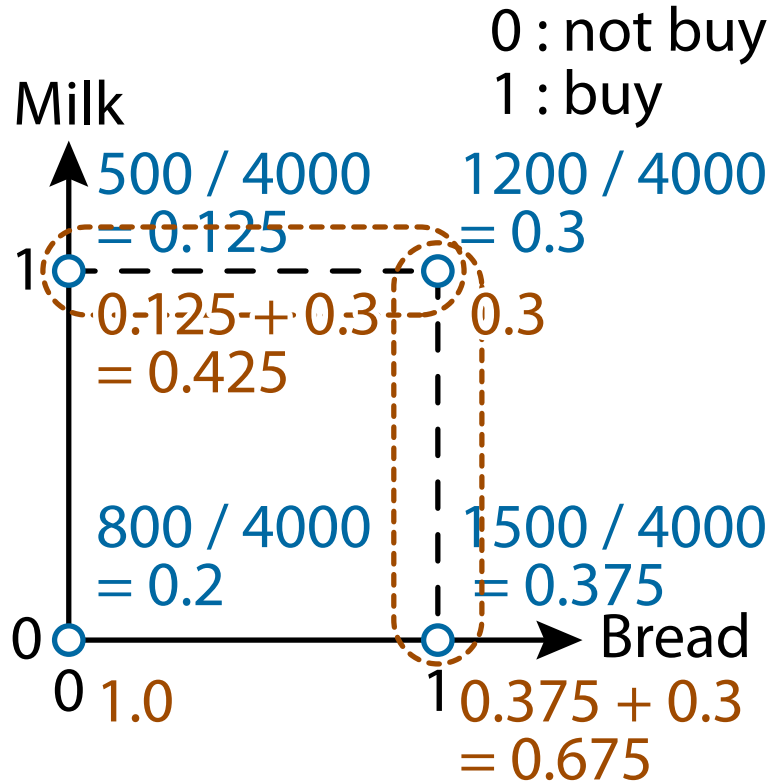
Find Frequent Patterns (Probability)



- Dataset:

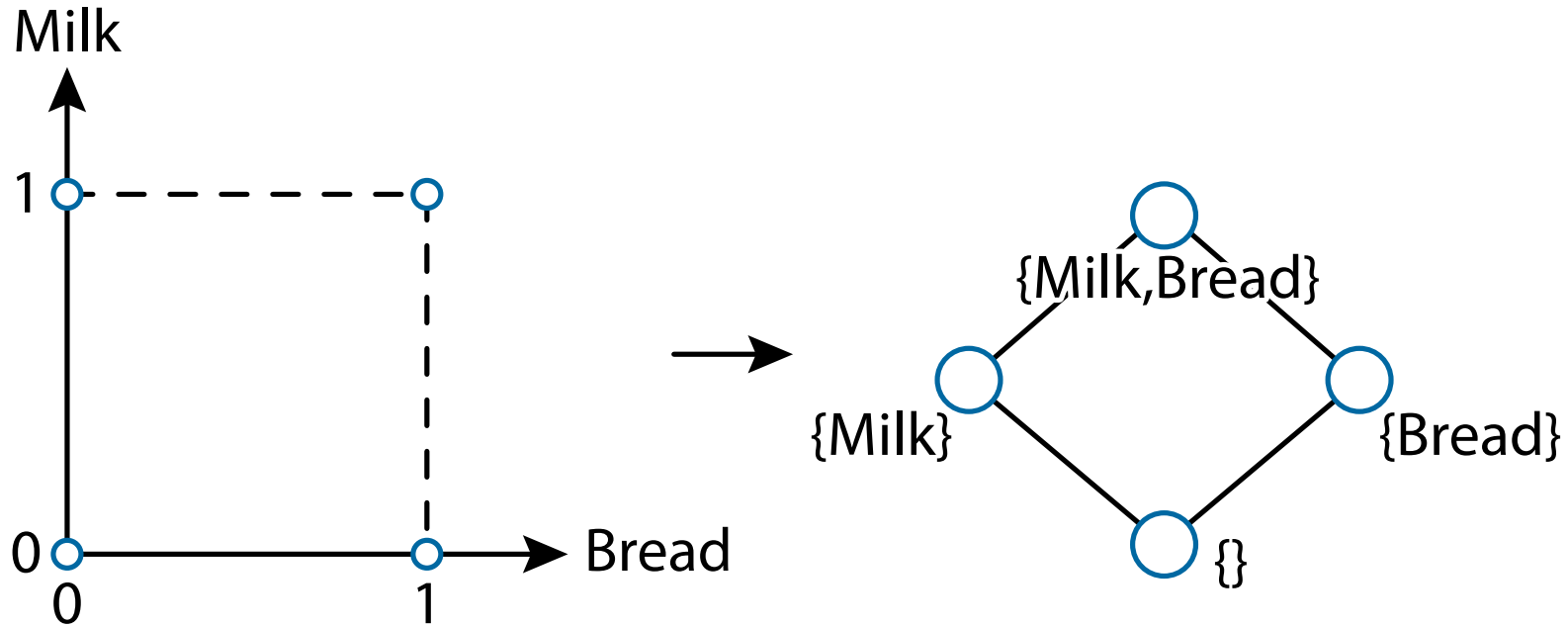
	Bread	Milk
1	0	1
2	1	1
3	1	1
4	1	0
⋮	⋮	⋮
4000	1	0

Find Frequent Patterns (Frequency)

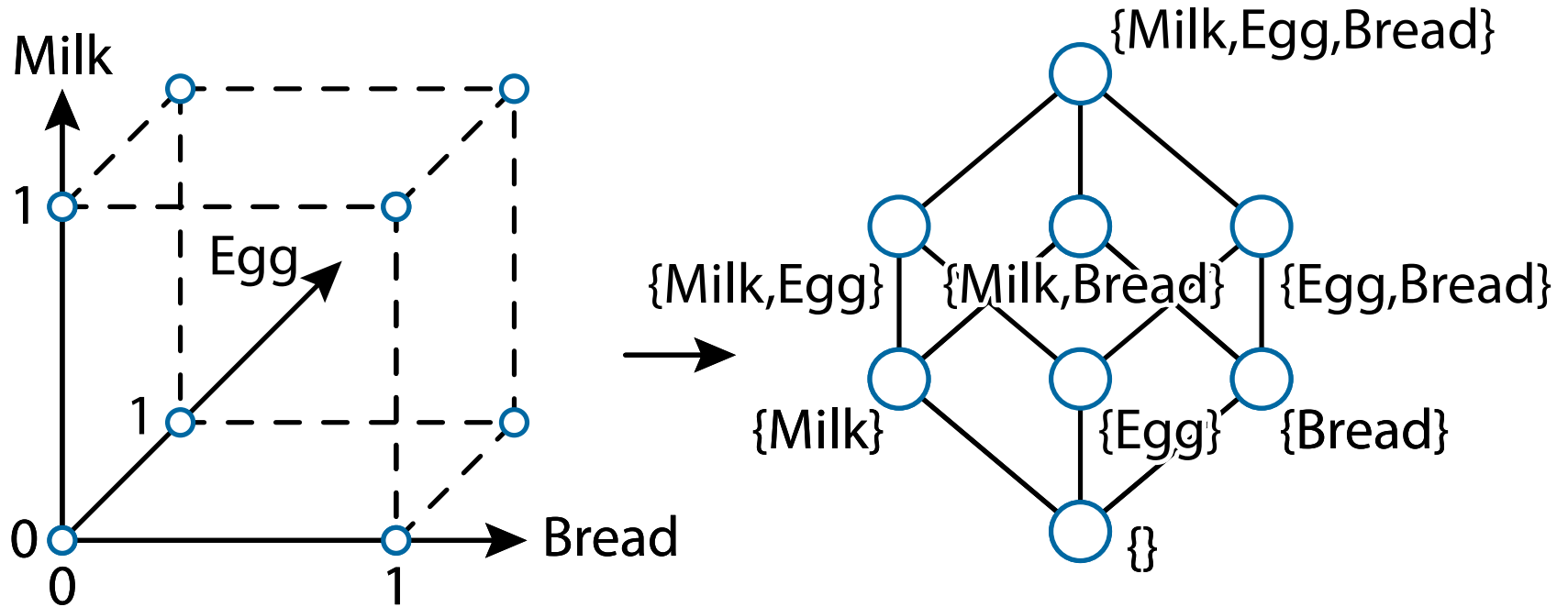


- 67.5% customers bought {bread}
- 42.5% customers bought {milk}
- 30.0% customers bought {bread, milk}
- An item combination is called a **pattern**

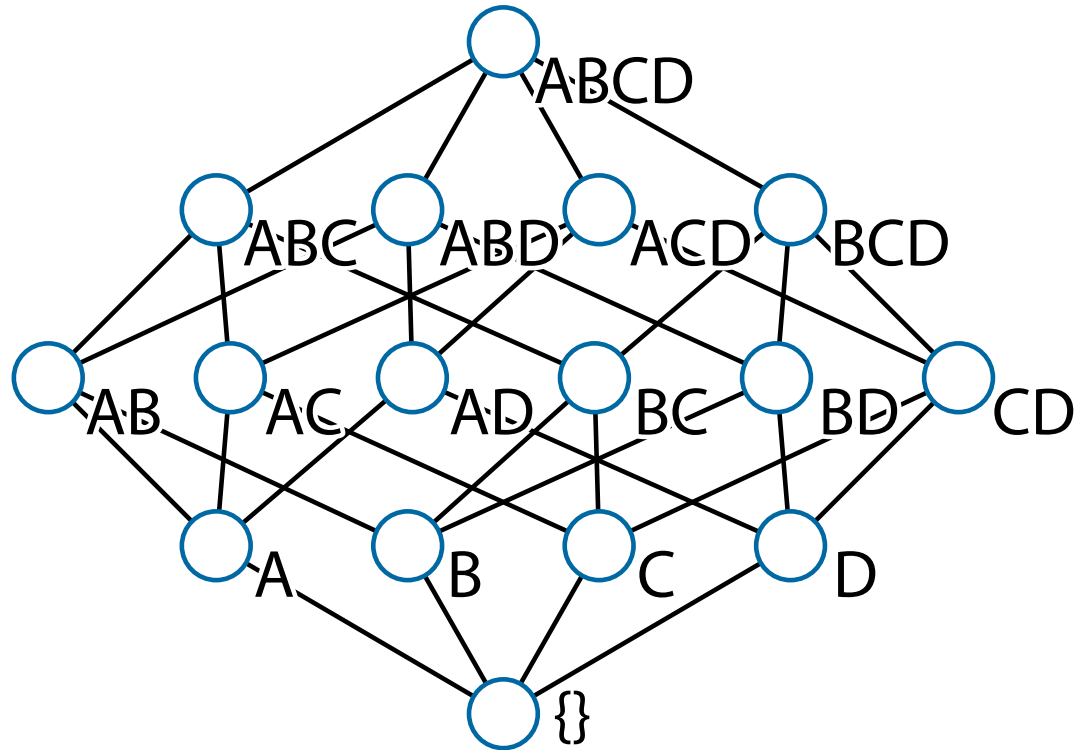
Lattice Representation (2 Variables)



Lattice Representation (3 Variables)



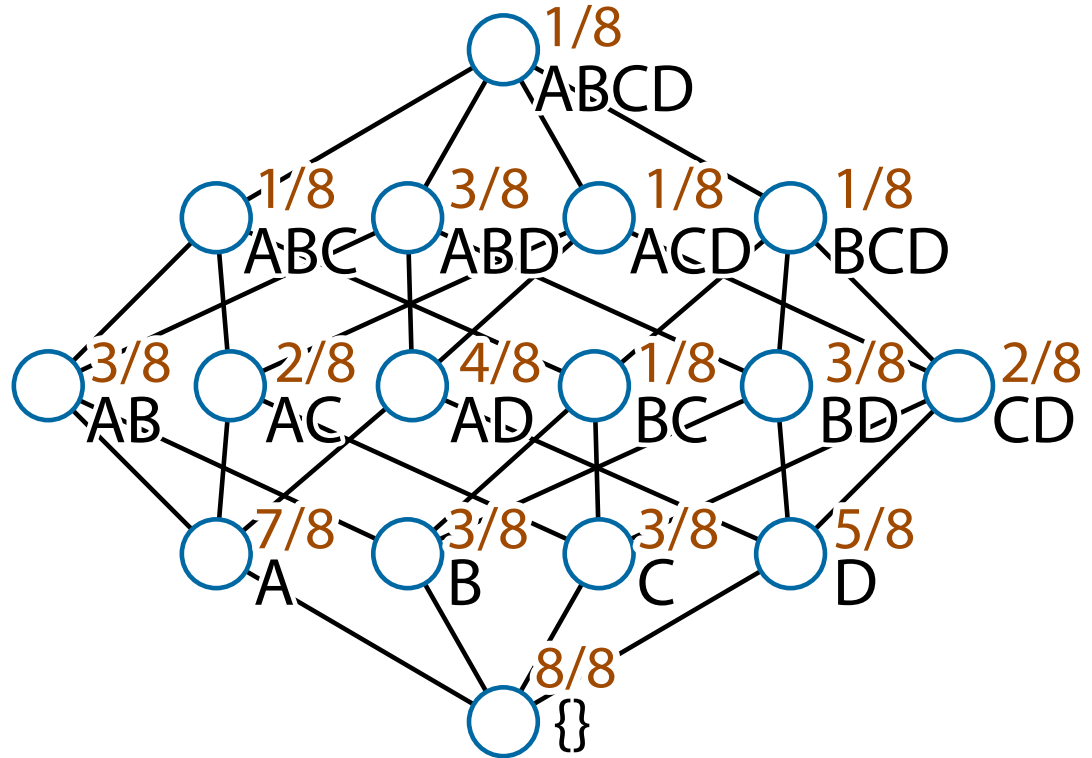
Lattice for 4 Variables



Frequency of Patterns

Dataset:

	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1



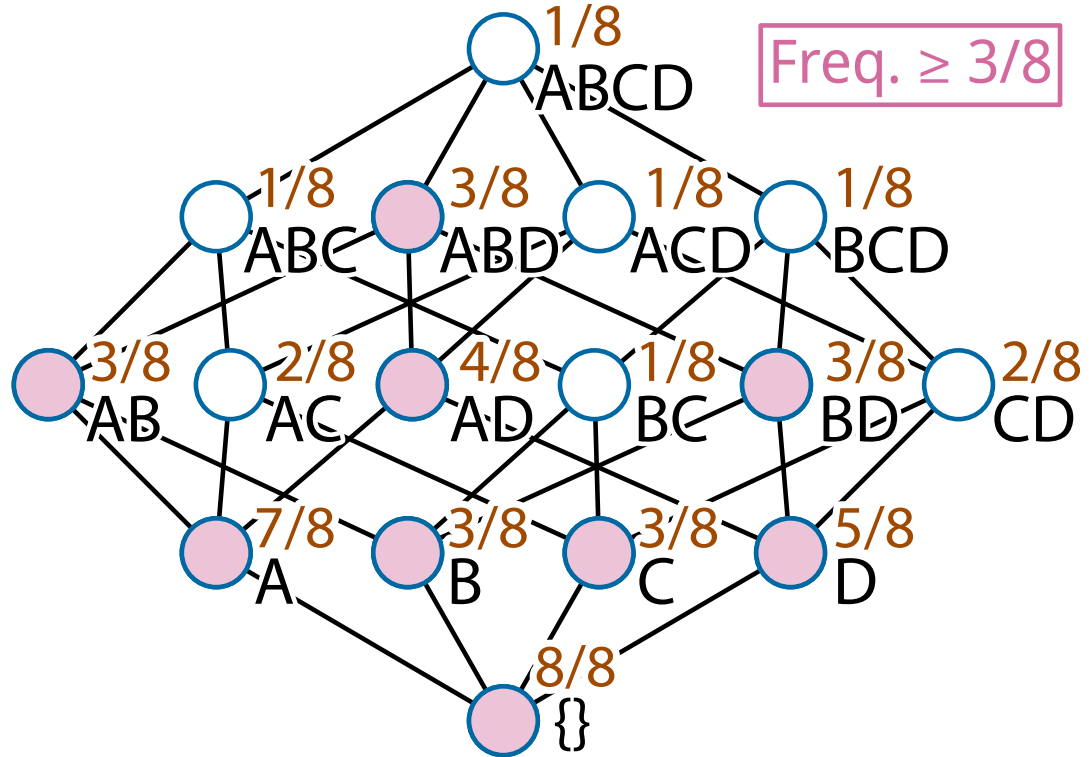
Pattern Mining

- Find **frequent patterns** (variable combinations)
- Representative methods:
 - Apriori, FP-growth, LCM, ...
- Input: Set of binary vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \{0, 1\}^d$
 - Multiple occurrence is allowed in X
- Output: Set of patterns (itemsets):
 $\{s \subseteq \{1, 2, \dots, d\} \mid \eta(s) > \sigma\}$
 - $\eta(s)$ is frequency of s , σ is a threshold given by the user

Search for Frequent Patterns

Dataset:

	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1



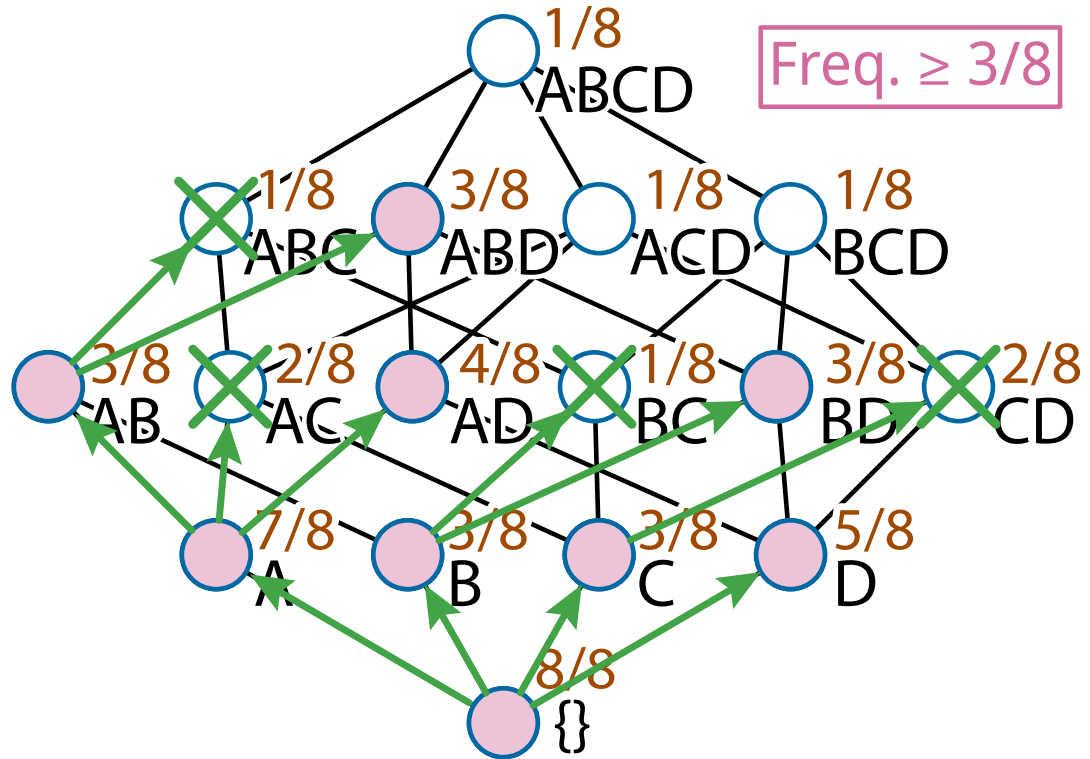
Combinatorial Explosion!! [\[YouTube\]](#)

# var.	# patterns	Runtime in naïve enum.
20	2^{20}	0.00059 sec.
30	2^{30}	0.6 sec.
40	2^{40}	10.2 min.
50	2^{50}	174 hours.
70	2^{70}	7 million days
100	2^{100}	8 thousand billion days

Apriori [Agrawal & Srikant, 1994]

Dataset:

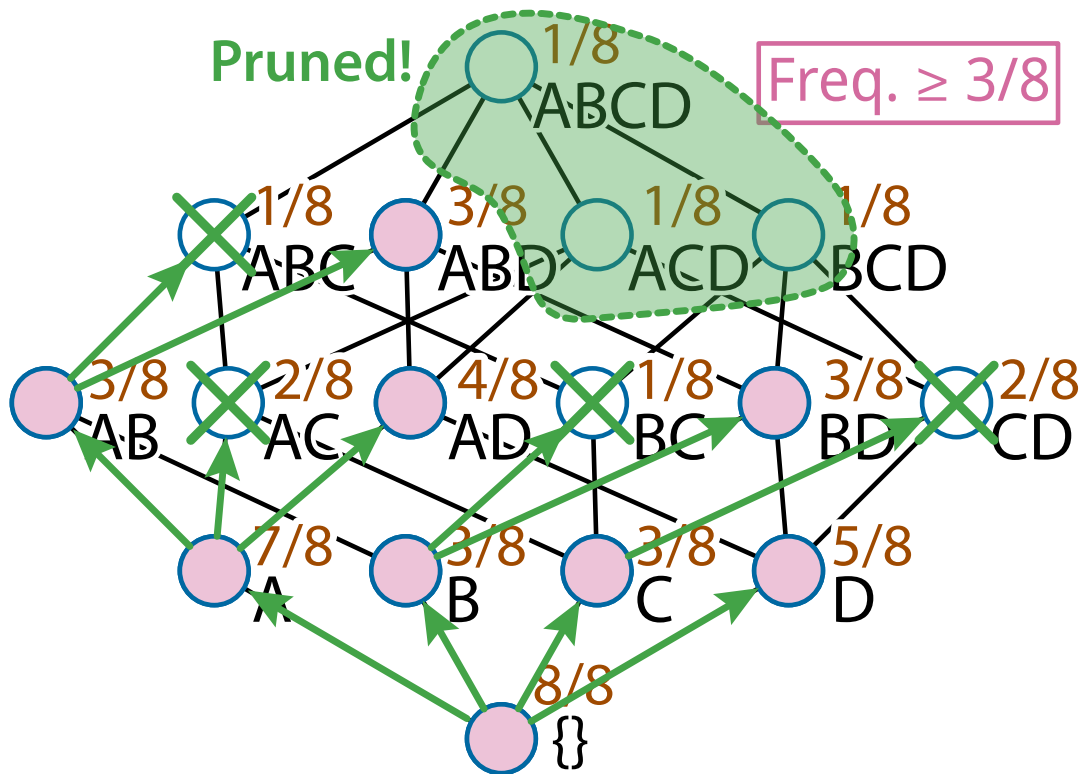
	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1



Apriori [Agrawal & Srikant, 1994]

Dataset:

	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1



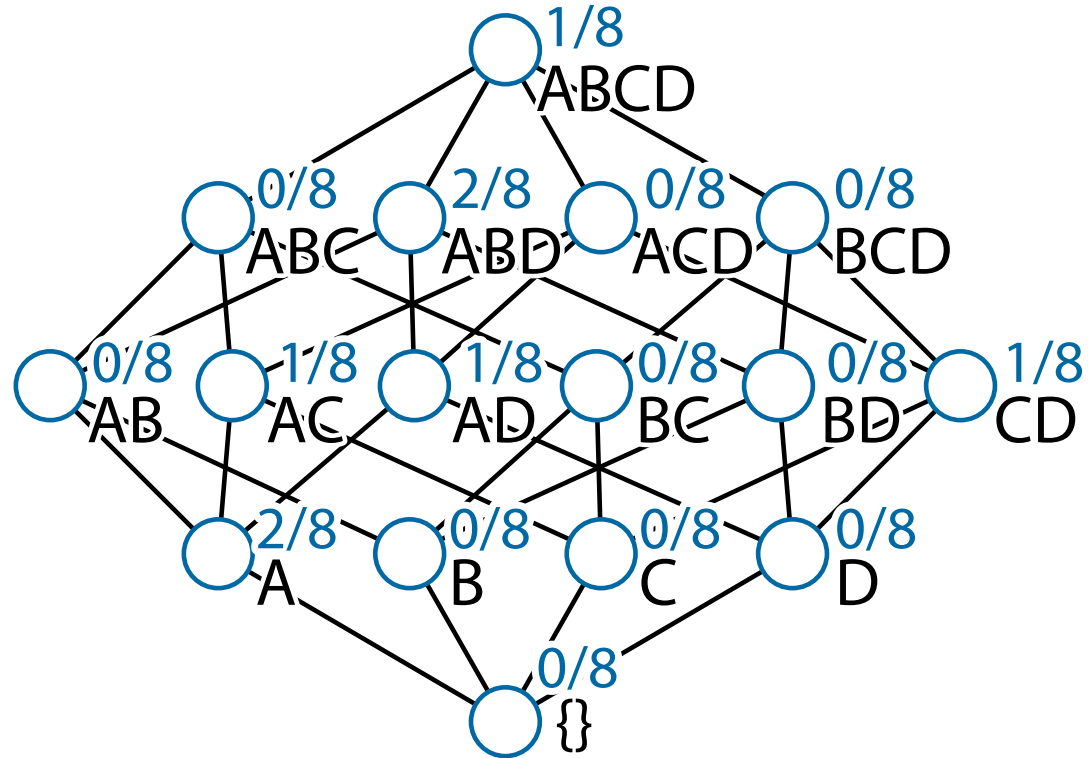
Algorithm 1: Apriori algorithm

```
1 PATTERNMINING( $\sigma$ )
2   | PATTERNENUMERATION ( $\perp$ ,  $\sigma$ )
3 PATTERNENUMERATION( $x$ ,  $\sigma$ )
4   | foreach  $s \supset x$  and  $|s| = |x| + 1$  do
5     |   | if  $\eta(s) \geq \sigma$  then
6       |   |   | Output  $s$ 
7       |   |   | PATTERNENUMERATION ( $s$ ,  $\sigma$ )
```

Probability Distribution on Lattice

Dataset:

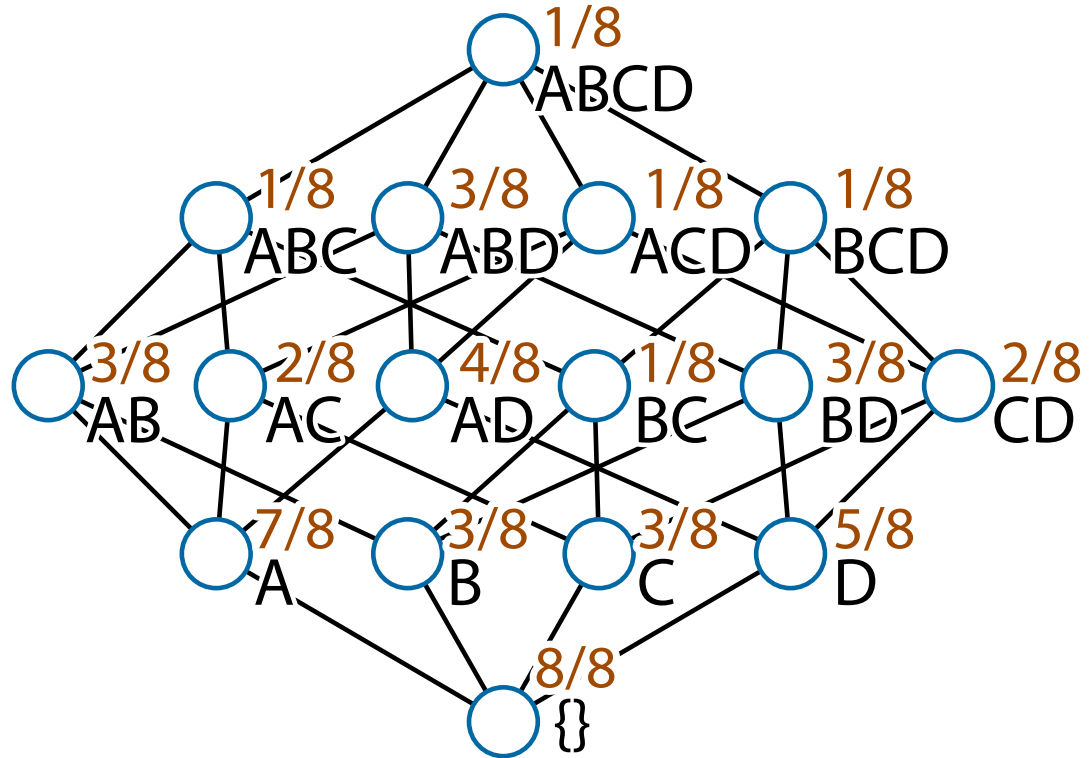
	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1



Probability And Frequency

Dataset:

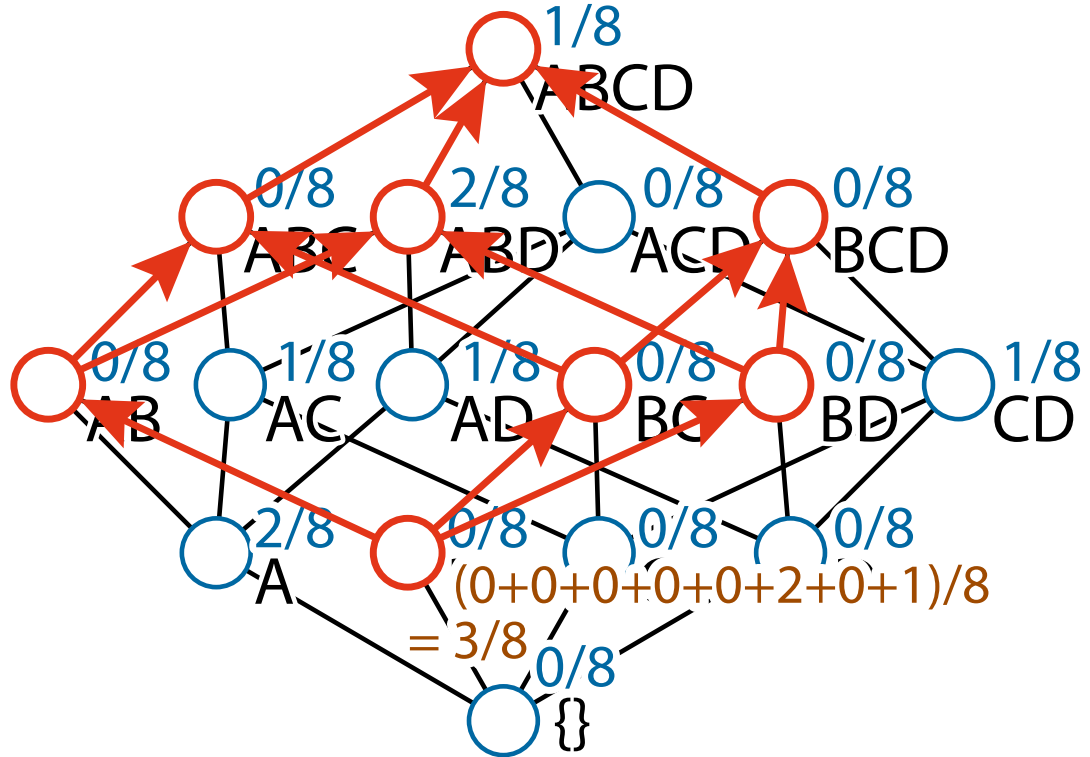
	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1



Sum of prob. in Upper Set = Frequency

Dataset:

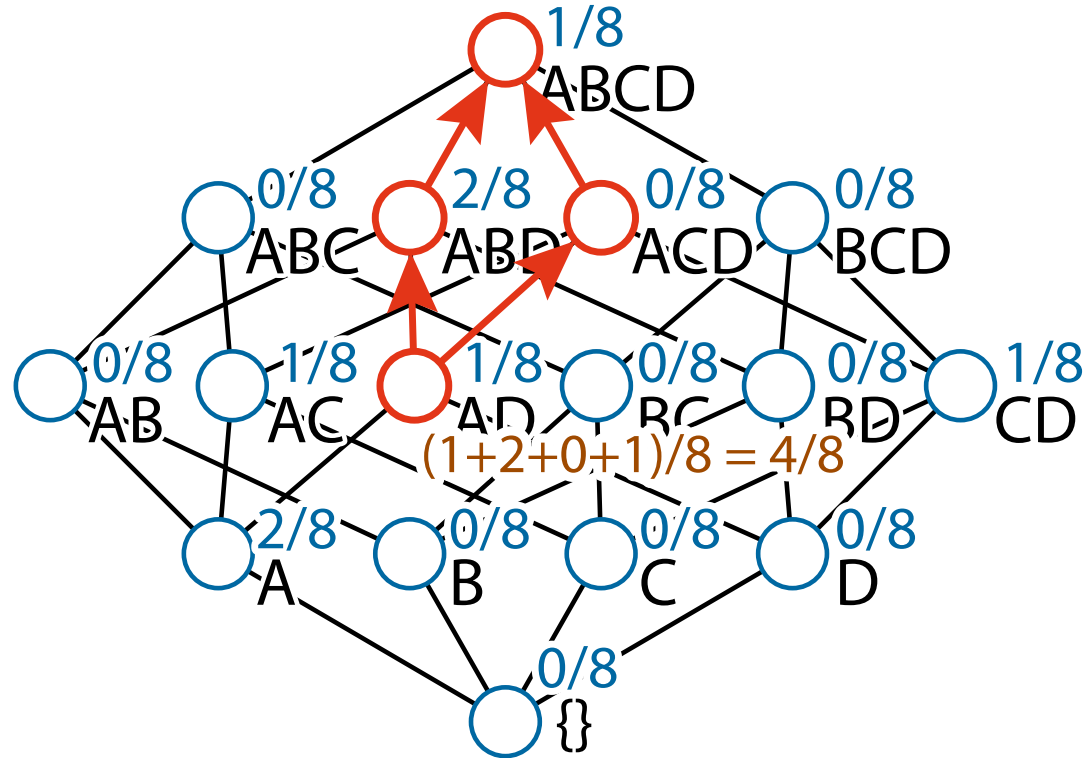
	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1



Sum of prob. in Upper Set = Frequency

Dataset:

	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1

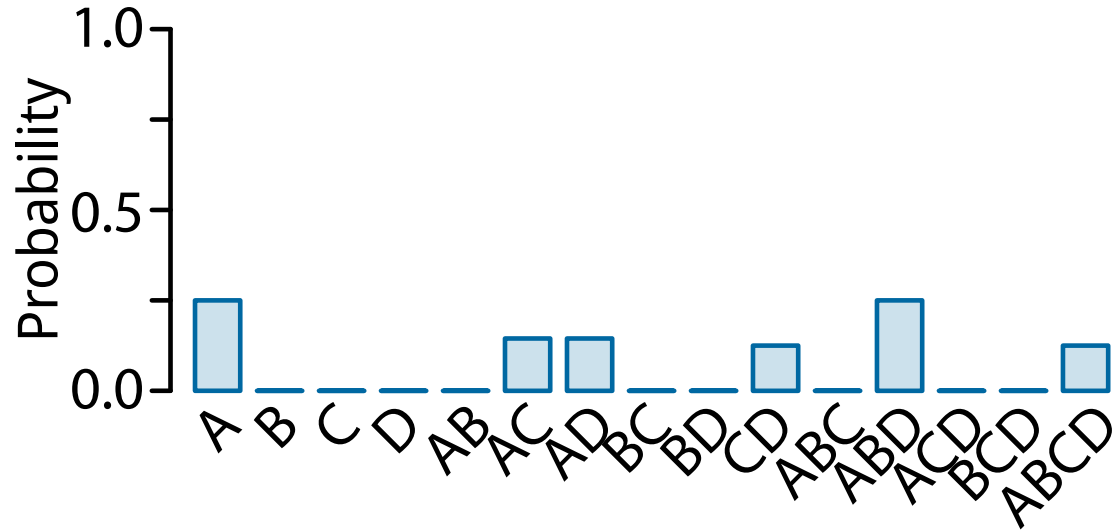


Estimation of Probability Distribution

Dataset:

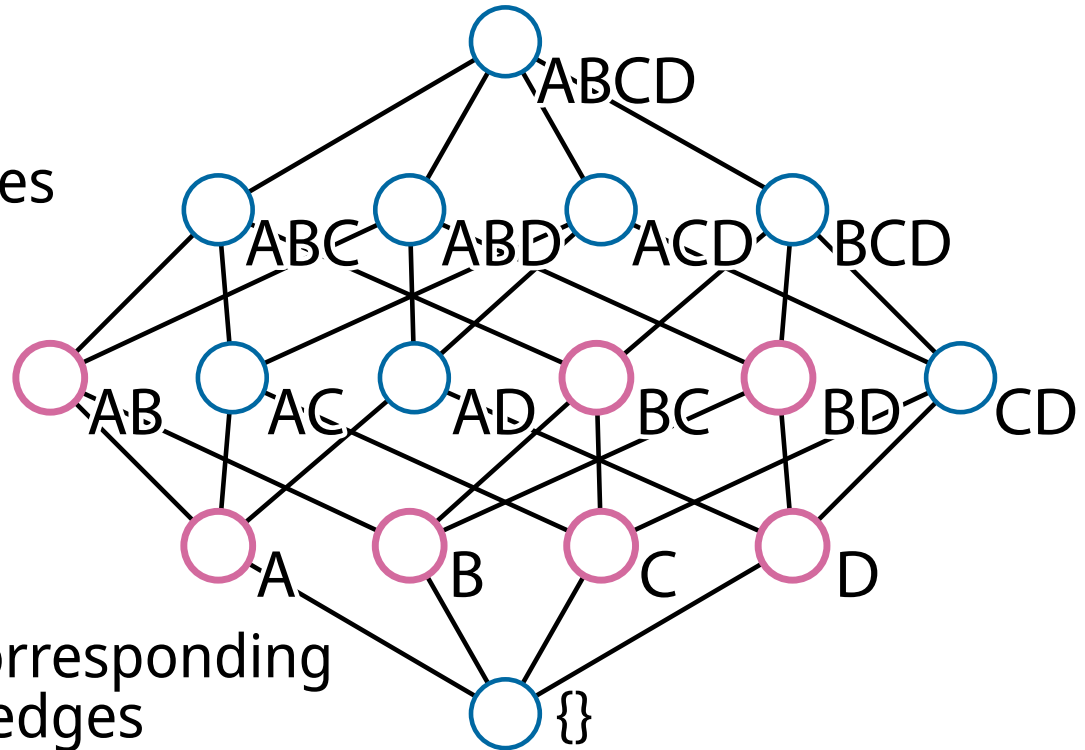
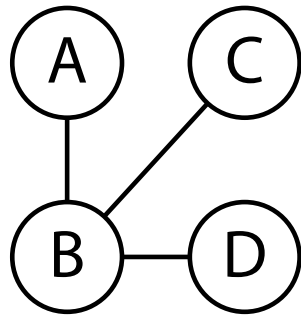
	A	B	C	D
ID 1:	1	1	0	1
ID 2:	1	0	1	0
ID 3:	1	0	0	0
ID 4:	0	0	1	1
ID 5:	1	1	0	1
ID 6:	1	0	0	0
ID 7:	1	0	0	1
ID 8:	1	1	1	1

Empirical distribution
→ What is true distribution?



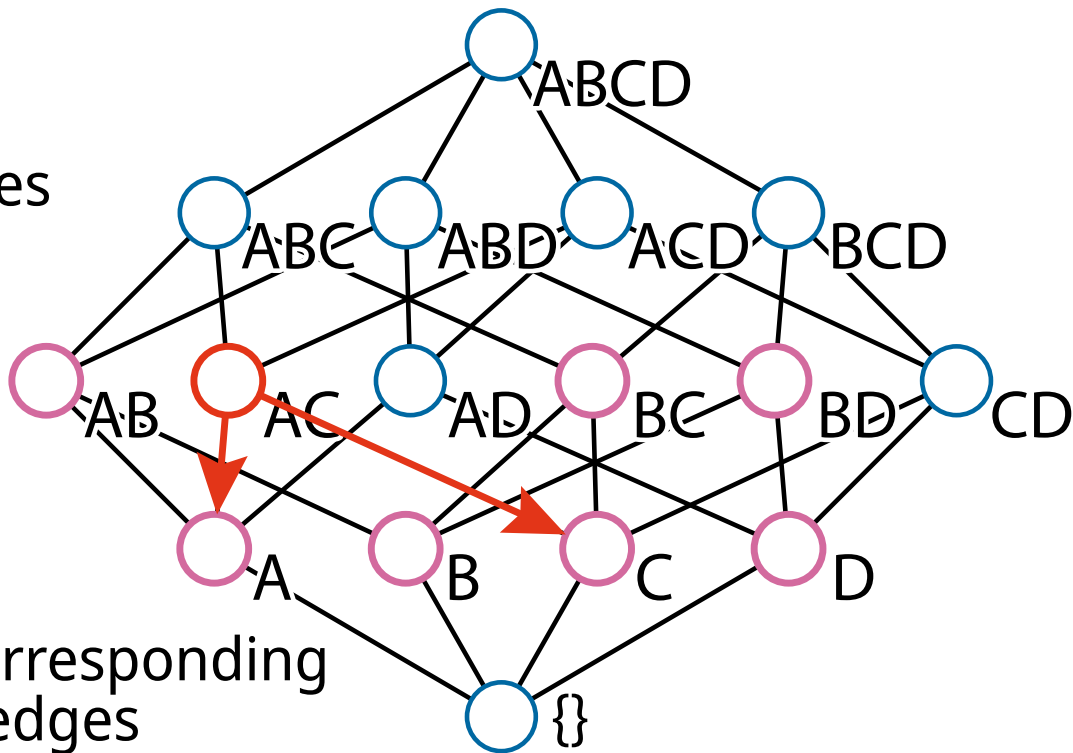
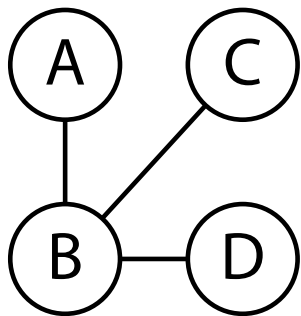
Boltzmann Machine [Ackley, Hinton & Sejnowski, 1985]

Boltzmann machines



Σ Param. in Lower Set = Probability

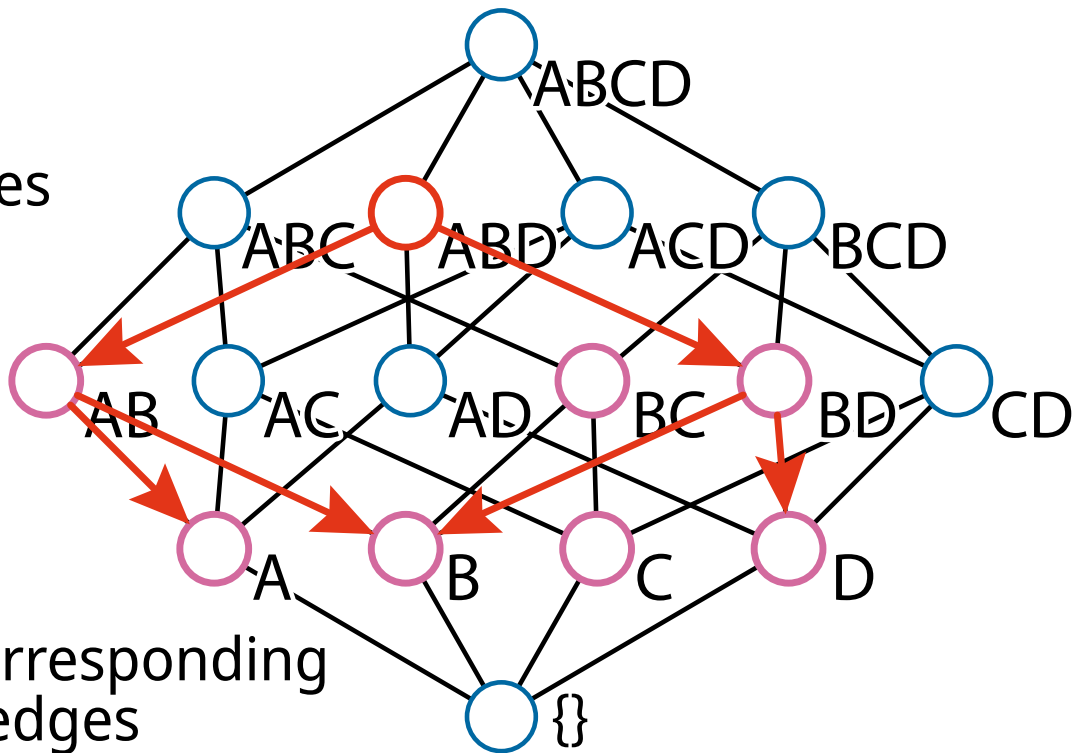
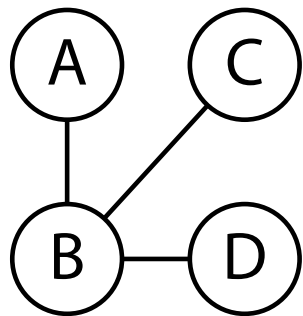
Boltzmann machines



: Parameters corresponding to nodes and edges

Σ Param. in Lower Set = Probability

Boltzmann machines



: Parameters corresponding to nodes and edges

Probability Computation in BM

- Boltzmann Machine: Undirected graph $G = (V, E)$
 - $V = \{A, B, C, D\}$, $E = \{(A, B), (B, C), (B, D)\}$
- Parameter: $\theta = (\theta_A, \theta_B, \theta_C, \theta_D, \theta_{AB}, \theta_{BC}, \theta_{BD})$
- Probability of **model distribution**:

$$p(AC; \theta) = \exp(\theta_A + \theta_C) / Z$$

$$p(ABD; \theta) = \exp(\theta_A + \theta_B + \theta_D + \theta_{AB} + \theta_{BD}) / Z$$

$$Z = \exp(-\theta_{\emptyset}) \quad (\text{normalizing constant})$$

Learning of Parameter θ : MLE

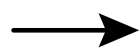
- For a dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$,
find a parameter vector θ that maximizes likelihood L

$$L_X(\theta) = p(\mathbf{x}_1; \theta) \cdot p(\mathbf{x}_2; \theta) \cdot \dots \cdot p(\mathbf{x}_n; \theta)$$

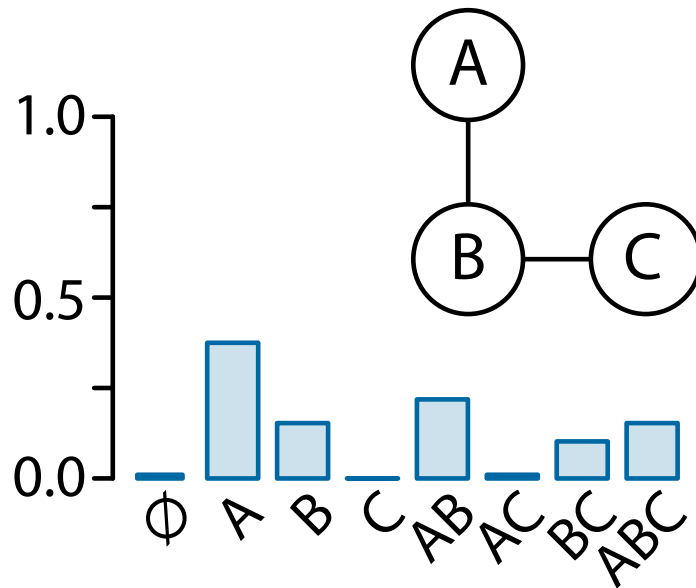
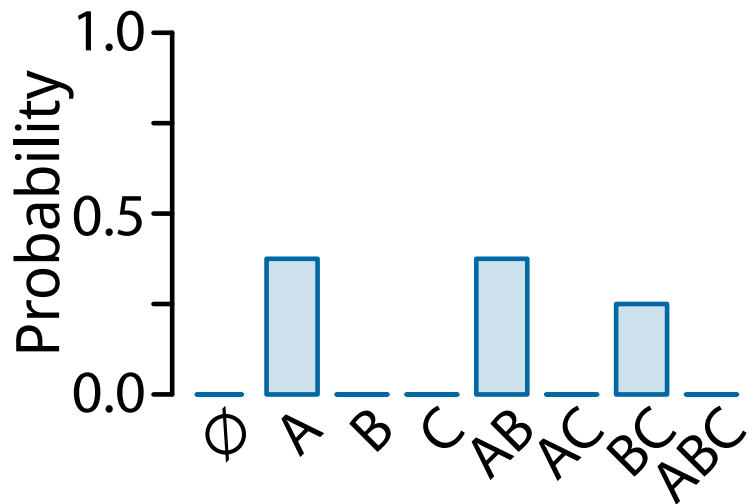
- Basic strategy: **Gradient method** :
 - (i) Start with some θ
 - (ii) Compute the direction (gradient) to the goal
 - Difference b/w freq. of empirical and model dist.
 - (iii) Move θ a bit toward the direction \rightarrow (ii)

Result of BM learning

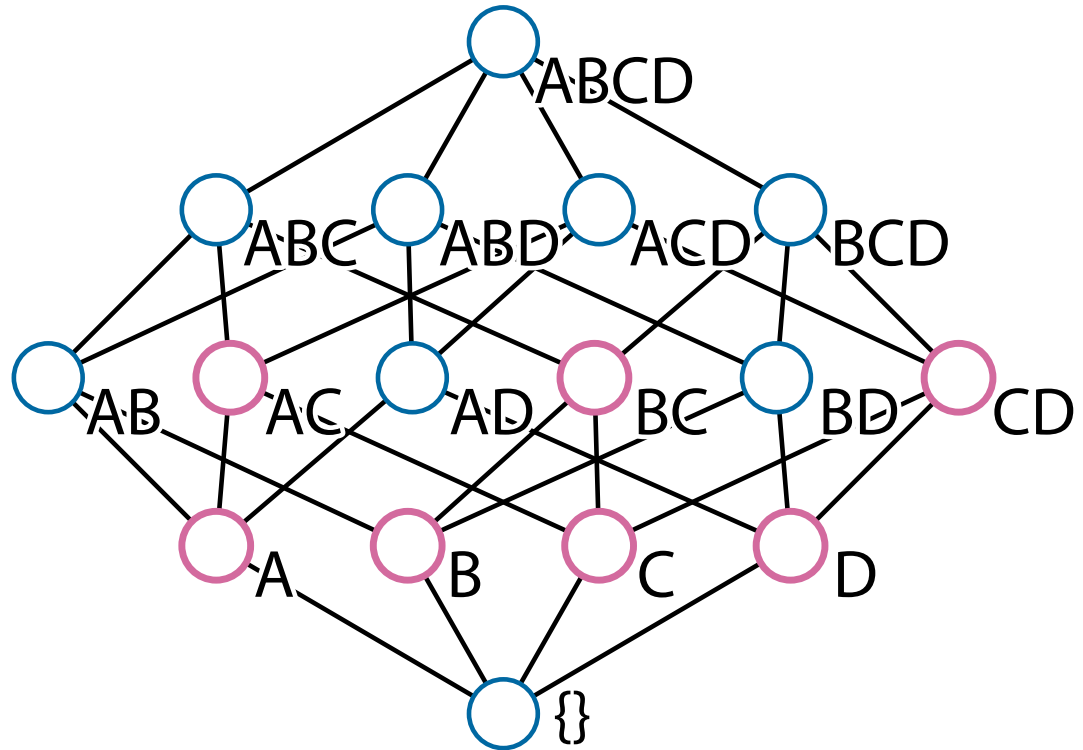
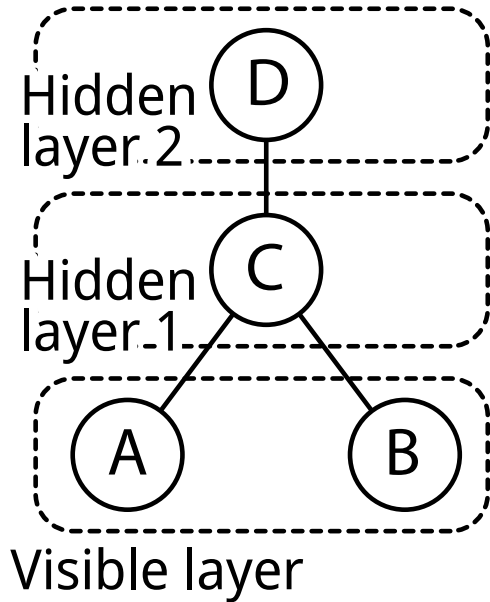
Empirical Distribution



Learned Distribution



Deep BM [Salakhutdinov & Hinton, 2009]



Bias-Variance Tradeoff

- Likelihood increases as # parameters increases, while a model overfits to data
- Extreme cases:
 - Empirical dist. itself is learned if all parameters are used
→ bias = 0, variance is large
 - Uniform dist. is learned if no parameter is used
→ bias is large , variance = 0
- There is a **tradeoff between bias and variance**

Relationship to Info. Geometry

- Consider “the set of probability distributions” (manifold)
- Parameters θ and frequencies η in Boltzmann machines are its coordinate system
- θ and η are orthogonal, leading to dually flat structure
- The MLE formulation can be applied to various tasks if θ and η are well combined
 - Matrix balancing, tensor decomposition, signal separation, ...