Efficient Algorithms, Spring 2021

3. Functional Aggregate Queries

Prof. Dan Olteanu

DaST

Data • (Systems+Theory)

University of Zurich UZH

March 8, 2021

https://lms.uzh.ch/url/RepositoryEntry/16979230773

Notation

- $i \in [n] = \{1, \dots, n\}$
- $X_1, \dots, X_n$ are variables
- x_i are values in discrete domain $\mathrm{Dom}(X_i)$
- $\mathbf{x} = (x_1, \dots, x_n) \in \mathrm{Dom}(X_1) \times \cdots \times \mathrm{Dom}(X_n)$
- For any $S \subseteq [n]$,

$$\begin{aligned} \mathbf{x}_S &= (x_i)_{i \in S} \in \prod_{i \in S} \mathrm{Dom}(X_i) \\ \text{e.g., } \mathbf{x}_{\{2,5,8\}} &= (x_2, x_5, x_8) \in \mathrm{Dom}(X_2) \times \mathrm{Dom}(X_5) \times \mathrm{Dom}(X_8) \end{aligned}$$

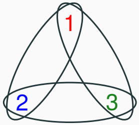
Hypergraphs

$\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is a (multi)hypergraph

- $\mathcal{V}$ is the set of nodes, with one node per variable
- $\mathcal{E}$ is the (multi)set of hyperedges, with each hyperedge $S \in \mathcal{E}$ a subset of $\mathcal{V}$

Examples:

- $\mathcal{V} = \{1, 2, 3\}$
- $\mathcal{E} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$



The hyperedges could be the edge relation of a graph

Hypergraphs

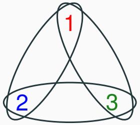
$\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is a (multi)hypergraph

- $\mathcal{V}$ is the set of nodes, with one node per variable
- $\mathcal{E}$ is the (multi)set of hyperedges, with each hyperedge $S \in \mathcal{E}$ a subset of $\mathcal{V}$

Examples:

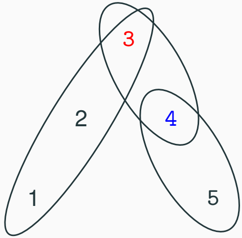
- $\mathcal{V} = \{1, 2, 3\}$
- $\mathcal{E} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$

The hyperedges could be the edge relation of a graph

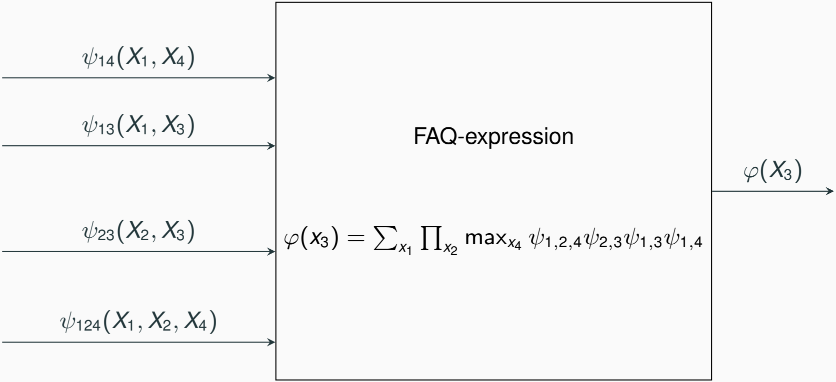


- $\mathcal{V} = \{1, 2, 3, 4, 5\}$
- $\mathcal{E} = \{\{1, 2, 3\}, \{3, 4\}, \{4, 5\}\}$

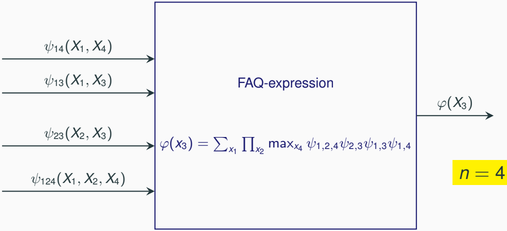
The hyperedges could be: Orders(customer, day, dish), Dishes(dish, item), Items(item, price)



Functional Aggregate Queries: The Problem



Functional Aggregate Queries: The Input



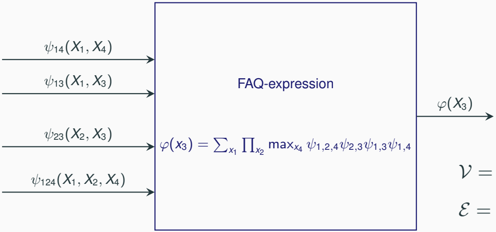
All factors have the same range **D**

$n = 4$

- n variables $X_1, \ldots, X_n$
- a multi-hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
  - Each vertex is a variable: $\mathcal{V} = [n]$
  - To each hyperedge $S \in \mathcal{E}$ there corresponds a *factor* ψ_S

$$\psi_S : \prod_{i \in S} \mathrm{Dom}(X_i) \rightarrow \mathbf{D}$$

Functional Aggregate Queries: The Input



All factors have the same range **D**

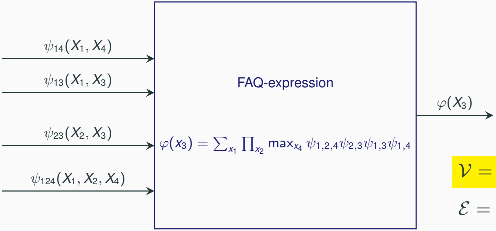
$\mathcal{V} = \{1, 2, 3, 4\}$

$\mathcal{E} = \{\{1, 4\}, \{1, 3\}, \{2, 3\}, \{1, 2, 4\}\}$

- *n* variables $X_1, \ldots, X_n$
- a multi-hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
  - Each vertex is a variable: $\mathcal{V} = [n]$
  - To each hyperedge $S \in \mathcal{E}$ there corresponds a *factor* ψ_S

$$\psi_S : \prod_{i \in S} \mathrm{Dom}(X_i) \to \mathbf{D}$$

Functional Aggregate Queries: The Input



All factors have the same range **D**

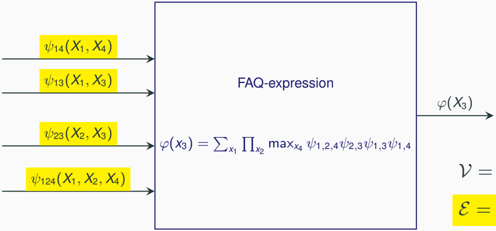
$\mathcal{V} = \{1, 2, 3, 4\}$

$\mathcal{E} = \{\{1, 4\}, \{1, 3\}, \{2, 3\}, \{1, 2, 4\}\}$

- *n* variables $X_1, \ldots, X_n$
- a multi-hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
  - Each vertex is a variable: $\mathcal{V} = [n]$
  - To each hyperedge $S \in \mathcal{E}$ there corresponds a *factor* ψ_S

$$\psi_S : \prod_{i \in S} \mathrm{Dom}(X_i) \to \mathbf{D}$$

Functional Aggregate Queries: The Input



All factors have the same range **D**

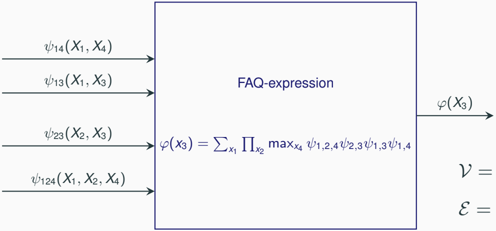
$\mathcal{V} = \{1, 2, 3, 4\}$

$\mathcal{E} = \{\{1, 4\}, \{1, 3\}, \{2, 3\}, \{1, 2, 4\}\}$

- n variables $X_1, \ldots, X_n$
- a multi-hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
  - Each vertex is a variable: $\mathcal{V} = [n]$
  - To each hyperedge $S \in \mathcal{E}$ there corresponds a *factor* ψ_S

$$\psi_S : \prod_{i \in S} \text{Dom}(X_i) \to \mathbf{D}$$

Functional Aggregate Queries: The Input



All factors have the same range **D**

$$\mathcal{V} = \{1, 2, 3, 4\}$$

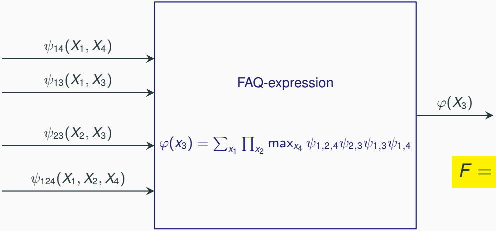
$$\mathcal{E} = \{\{1, 4\}, \{1, 3\}, \{2, 3\}, \{1, 2, 4\}\}$$

- n variables $X_1, \dots, X_n$
- a multi-hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
  - Each vertex is a variable: $\mathcal{V} = [n]$
  - To each hyperedge $S \in \mathcal{E}$ there corresponds a *factor* ψ_S

$$\psi_S : \prod_{i \in S} \text{Dom}(X_i) \to \mathbf{D}$$

↑ $\mathbb{R}$, $\{\text{true}, \text{false}\}$, $\{0, 1\}$, $2^{\mathcal{U}}$, etc.

Functional Aggregate Queries: The Input



All factors have the same range **D**

$F = \{3\}$

- n variables $X_1, \dots, X_n$
- a multi-hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
  - Each vertex is a variable: $\mathcal{V} = [n]$
  - To each hyperedge $S \in \mathcal{E}$ there corresponds a *factor* ψ_S

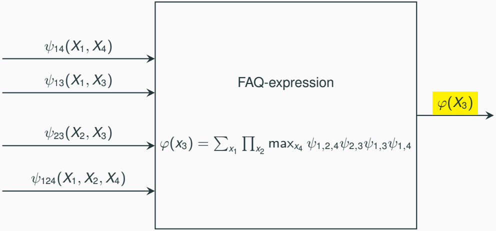
$$\psi_S : \prod_{i \in S} \text{Dom}(X_i) \to \mathbf{D}$$

$\uparrow$

$\mathbb{R}$, $\{\text{true}, \text{false}\}$, $\{0, 1\}$, $2^{\mathcal{U}}$, etc.

- a set $F \subseteq \mathcal{V}$ of free variables; w.l.o.g., $F = [f] = \{1, \dots, f\}$

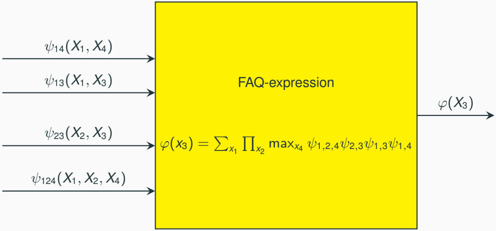
Functional Aggregate Queries: The Output



All functions have the same range **D**

- Compute the function $\varphi : \prod_{i \in F} \text{Dom}(X_i) \to \mathbf{D}$.

Functional Aggregate Queries: The Output

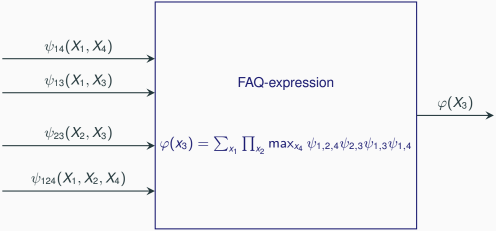


All functions have the same range **D**

- Compute the function $\varphi : \prod_{i \in F} \text{Dom}(X_i) \to \mathbf{D}$.
- φ defined by the FAQ-expression

$$\varphi(\mathbf{x}_{[f]}) = \bigoplus_{x_{f+1} \in \text{Dom}(X_{f+1})}^{(f+1)} \cdots \bigoplus_{x_{n-1} \in \text{Dom}(X_{n-1})}^{(n-1)} \bigoplus_{x_n \in \text{Dom}(X_n)}^{(n)} \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

Functional Aggregate Queries: The Output



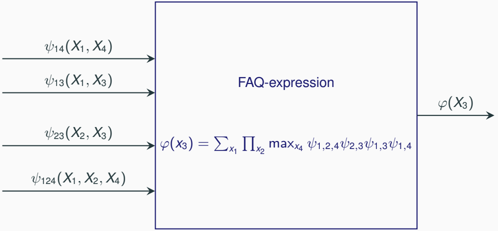
All functions have the same range **D**

- Compute the function $\varphi : \prod_{i \in F} \text{Dom}(X_i) \to \mathbf{D}$.
- φ defined by the FAQ-expression

$$\varphi(\mathbf{x}_{[f]}) = \bigoplus_{x_{f+1} \in \text{Dom}(X_{f+1})}^{(f+1)} \cdots \bigoplus_{x_{n-1} \in \text{Dom}(X_{n-1})}^{(n-1)} \bigoplus_{x_n \in \text{Dom}(X_n)}^{(n)} \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

- For each $\bigoplus^{(i)}$

Functional Aggregate Queries: The Output



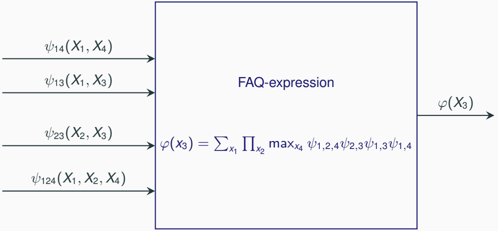
All functions have the same range **D**

- Compute the function $\varphi : \prod_{i \in F} \mathrm{Dom}(X_i) \to \mathbf{D}$.
- φ defined by the FAQ-expression

$$\varphi(\mathbf{x}_{[f]}) = \bigoplus_{x_{f+1} \in \mathrm{Dom}(X_{f+1})}^{(f+1)} \cdots \bigoplus_{x_{n-1} \in \mathrm{Dom}(X_{n-1})}^{(n-1)} \bigoplus_{x_n \in \mathrm{Dom}(X_n)}^{(n)} \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

- For each $\bigoplus^{(i)}$
  - Either $\left(\mathbf{D}, \bigoplus^{(i)}, \bigotimes, \mathbf{0}, \mathbf{1}\right)$ is a commutative semiring

Functional Aggregate Queries: The Output



All functions have the same range **D**

- Compute the function $\varphi : \prod_{i \in F} \mathrm{Dom}(X_i) \to \mathbf{D}$.
- φ defined by the FAQ-expression

$$\varphi(\mathbf{x}_{[f]}) = \bigoplus^{(f+1)}_{x_{f+1} \in \mathrm{Dom}(X_{f+1})} \cdots \bigoplus^{(n-1)}_{x_{n-1} \in \mathrm{Dom}(X_{n-1})} \bigoplus^{(n)}_{x_n \in \mathrm{Dom}(X_n)} \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

- For each $\bigoplus^{(i)}$
  - Either $\left(\mathbf{D}, \bigoplus^{(i)}, \bigotimes, \mathbf{0}, \mathbf{1}\right)$ is a commutative semiring
  - Or $\bigoplus^{(i)} = \bigotimes$

Functional Aggregate Queries: Putting it Together

An FAQ expression has the form

$$\varphi(\mathbf{x}_{[f]}) = \bigoplus_{x_{f+1} \in \mathrm{Dom}(X_{f+1})}^{(f+1)} \cdots \bigoplus_{x_n \in \mathrm{Dom}(X_n)}^{(n)} \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

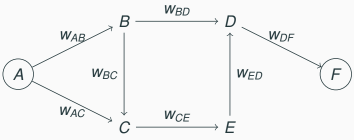
where

- $X_1, \ldots, X_n$ are variables with domains $\mathrm{Dom}(X_1), \ldots, \mathrm{Dom}(X_n)$
  - $X_1, \ldots, X_f$ are the free variables, $X_{f+1}, \ldots, X_n$ are the bound variables
- $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is a multi-hypergraph, where
  - $\mathcal{V} = [n]$ is the index set of variables
  - $\mathcal{E}$ is a set of subsets S of $\mathcal{V}$
- Each ψ_S is an input function or factor over variables with index set $S \in \mathcal{E}$
  - ψ_S maps tuples over S to elements in a finite set $\mathbf{D}$
  - Semirings $(\mathbf{D}, \oplus^{(i)}, \otimes, \mathbf{0}, \mathbf{1})$ have the same support $\mathbf{D}$

Expressing Problems in FAQ

PATH: Example

Graph



Edge relation E

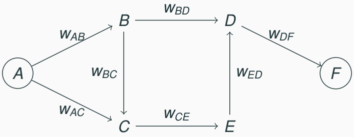
start	end	weight
A	B	w_{AB}
A	C	w_{AC}
B	D	w_{BD}
B	C	w_{BC}
C	E	w_{CE}
E	D	w_{ED}
D	F	w_{DF}

Vertices V

Node
A
B
C
D
E
F

PATH: Example

Graph



Edge relation E

start	end	weight
A	B	w_{AB}
A	C	w_{AC}
B	D	w_{BD}
B	C	w_{BC}
C	E	w_{CE}
E	D	w_{ED}
D	F	w_{DF}

Vertices V

Node
A
B
C
D
E
F

- The factor ψ maps each edge $(x_i, x_j) \in E$ to w_{ij}
- FAQ expresses graph traversal from 1 hop to 5 hops (length of longest path)

$$\psi(A, F) \oplus \left(\bigoplus_{x_1 \in V} \psi(A, x_1) \otimes \psi(x_1, F) \right) \oplus \left(\bigoplus_{x_1, x_2 \in V} \psi(A, x_1) \otimes \psi(x_1, x_2) \otimes \psi(x_2, F) \right)$$

$$\oplus \left(\bigoplus_{x_1, x_2, x_3 \in V} \psi(A, x_1) \otimes \psi(x_1, x_2) \otimes \psi(x_2, x_3) \otimes \psi(x_3, F) \right)$$

$$\oplus \left(\bigoplus_{x_1, x_2, x_3, x_4 \in V} \psi(A, x_1) \otimes \psi(x_1, x_2) \otimes \psi(x_2, x_3) \otimes \psi(x_3, x_4) \otimes \psi(x_4, F) \right)$$

PATH: Algebraic Path Problems

Compute path problem over vertices V and weighted edge relation E

FAQ encoding over the semiring $(\mathbf{D}, \oplus, \otimes, \mathbf{0}, \mathbf{1})$:

$$\Phi() = \left(\bigoplus_{i \in [2] : x_i \in V} \psi_{12}(x_1, x_2) \right) \oplus \qquad \text{// 1 hop}$$

$$\left(\bigoplus_{i \in [3] : x_i \in V} \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \right) \oplus \cdots \oplus \qquad \text{// 2 hops}$$

$$\left(\bigoplus_{i \in [n+1] : x_i \in V} \psi_{12}(x_1, x_2) \otimes \cdots \otimes \psi_{n,n+1}(x_n, x_{n+1}) \right) \qquad \text{// } n \text{ hops}$$

$\psi_{i,i+1}(x_i, x_{i+1})$ maps each edge $(x_i, x_{i+1}) \in E$ to a semiring-dependent weight

- min-sum over reals for shortest distance
- max-min over $\{0, 1\}$ for connectivity and over reals for largest capacity
- max-product over $[0, 1]$ for maximum reliability
- concatenate-union over strings for language accepted by automaton

SAT: Example

$(R_{CH} \lor G_{CH} \lor B_{CH})$	"Switzerland has *at least* one colour."
$\bigwedge$	
$(\neg R_{CH} \lor \neg G_{CH}) \land (\neg R_{CH} \lor \neg B_{CH}) \land (\neg G_{CH} \lor \neg B_{CH})$	"Switzerland has *at most* one colour."
$\bigwedge$	
$(\neg R_{CH} \lor \neg R_{DE}) \land (\neg G_{CH} \lor \neg G_{DE}) \land (\neg B_{CH} \lor \neg B_{DE})$	"Switzerland and Germany have different colours."
$\ldots$	$\ldots$

SAT: Example

$(R_{CH} \vee G_{CH} \vee B_{CH})$ — "Switzerland has *at least* one colour."

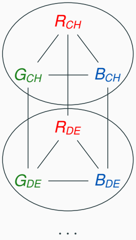
$\wedge$

$(\neg R_{CH} \vee \neg G_{CH}) \wedge (\neg R_{CH} \vee \neg B_{CH}) \wedge (\neg G_{CH} \vee \neg B_{CH})$ — "Switzerland has *at most* one colour."

$\wedge$

$(\neg R_{CH} \vee \neg R_{DE}) \wedge (\neg G_{CH} \vee \neg G_{DE}) \wedge (\neg B_{CH} \vee \neg B_{DE})$ — "Switzerland and Germany have different colours."

$\dots$ — $\dots$



SAT: Example

$(R_{CH} \vee G_{CH} \vee B_{CH})$ — "Switzerland has *at least* one colour."

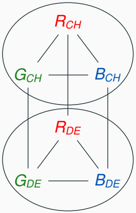
$$\bigwedge$$

$(\neg R_{CH} \vee \neg G_{CH}) \wedge (\neg R_{CH} \vee \neg B_{CH}) \wedge (\neg G_{CH} \vee \neg B_{CH})$ — "Switzerland has *at most* one colour."

$$\bigwedge$$

$(\neg R_{CH} \vee \neg R_{DE}) \wedge (\neg G_{CH} \vee \neg G_{DE}) \wedge (\neg B_{CH} \vee \neg B_{DE})$ — "Switzerland and Germany have different colours."

$\ldots$ — $\ldots$



$\ldots$

There is a factor encoding each clause.

r_{ch} is a value of R_{CH} (true or false).

$$\bigvee_{\substack{r_{ch}, g_{ch}, b_{ch} \\ r_{de}, g_{de}, b_{de}}} \psi_{R_{CH}, G_{CH}, B_{CH}}(r_{ch}, g_{ch}, b_{ch}) \wedge$$

$$\psi_{R_{CH}, B_{CH}}(r_{ch}, b_{ch}) \wedge \psi_{G_{CH}, B_{CH}}(g_{ch}, b_{ch}) \wedge \psi_{R_{CH}, G_{CH}}(r_{ch}, g_{ch}) \wedge$$

$$\psi_{R_{CH}, R_{DE}}(r_{ch}, r_{de}) \wedge \psi_{G_{CH}, G_{DE}}(g_{ch}, g_{de}) \wedge \psi_{B_{CH}, B_{DE}}(b_{ch}, b_{de}) \wedge$$

$$\ldots$$

SAT: Satisfiability

Check satisfiability of CNF formula $\bigwedge_{i\in[m]} c_i$ with hypergraph $\mathcal{H}$

FAQ encoding over the Boolean semiring $(\{\text{true}, \text{false}\}, \vee, \wedge, \text{false}, \text{true})$:

$$\Phi() = \bigvee_{\mathbf{x}} \bigwedge_{S\in\mathcal{E}} \psi_S(\mathbf{x}_S)$$

where Φ has the same hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ as the CNF formula

- $\mathcal{V} = [n]$ is the set of indices of variables $X_1, \ldots, X_n$ in the CNF formula
- Each clause c_i over variables with indices in S defines a factor ψ_S
    - For variable assignment $\mathbf{x}_S$, $\psi_S(\mathbf{x}_S)$ returns the truth of c_i
    - Naïve $O(2^{|S|})$ representation of ψ_S is the truth table of c_i over the variables in S
    - Alternative $O(|S|)$ representation is just the clause

#SAT: Counting the Number of Satisfying Assignments

Count the satisfying assignments of CNF formula $\bigwedge_{i \in [m]} c_i$ with hypergraph $\mathcal{H}$

FAQ encoding over the sum-product semiring $(\mathbb{N}, +, *, 0, 1)$:

$$\Phi() = \sum_{\mathbf{x}} \prod_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

where Φ has the same hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ as the CNF formula

- $\mathcal{V} = [n]$ is the set of indices of variables $X_1, \ldots, X_n$ in the CNF formula
- Each clause c_i over variables with indices in S defines a factor ψ_S

$$\psi_S(\mathbf{x}_S) = \begin{cases} 1 & \text{if } \mathbf{x}_S \text{ satisfies } c_i \\ 0 & \text{otherwise.} \end{cases}$$

3-Colorability: A Different Approach

Recall again the 3-colorability problem instance.

We use the map graph as the hypergraph of the problem:

- $\mathcal{V} = \{CH, DE, FR, IT, AT, LI, \ldots\}$
- $\mathcal{E} = \{(CH, DE), (CH, FR), (CH, IT), (CH, AT), (CH, LI), \ldots\}$

For each edge (u, v) there is a factor $\psi_{uv}(c_1, c_2) = (c_1 \neq c_2)$

- Each variable can take value $1, 2$, or 3 representing one of the three colours
- For a pair of colours (c_1, c_2) for nodes (u, v), ψ_{uv} is true if $c_1 \neq c_2$

The FAQ is (the part for CH):

$$\Phi() = \bigvee_{\substack{c_{ch}, c_{de}, c_{fr}, \\ c_{it}, c_{at}, c_{li}, \ldots}} \psi_{CH,DE}(c_{ch}, c_{de}) \wedge \psi_{CH,FR}(c_{ch}, c_{fr}) \wedge \psi_{CH,IT}(c_{ch}, c_{it}) \wedge \psi_{CH,AT}(c_{ch}, c_{at}) \wedge \psi_{CH,LI}(c_{ch}, c_{li}) \wedge \ldots$$

k-Colorability

Check k-colorability for a graph $G = (V, E)$

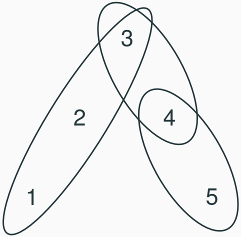
FAQ encoding over the Boolean semiring $(\{\text{true}, \text{false}\}, \vee, \wedge, \text{false}, \text{true})$:

$$\Phi() = \bigvee_{\mathbf{x}} \bigwedge_{(u,v)\in E} \psi_{uv}(x_u, x_v), \text{ where}$$

- Every edge $(u, v) \in E$ defines a factor $\psi_{uv}(c_1, c_2) = (c_1 \neq c_2)$
- Every node $v \in V$ defines a variable X_v with domain $\text{Dom}(X_v) = [k]$
- Φ has the hypergraph with vertices $\mathcal{V} = \{X_v | v \in V\}$ and edges $\mathcal{E} = E$

DB: Example (1/4)

We map our previous DB example with customers ordering dishes to FAQ:



$\underbrace{\texttt{Orders(customer, day, dish)}}_{\psi_{123}(x_1, x_2, x_3)}$

$\underbrace{\texttt{Dishes(dish, item)}}_{\psi_{34}(x_3, x_4)}$

$\underbrace{\texttt{Items (item, price)}}_{\psi_{45}(x_4, x_5)}$

The FAQ over the union-intersection semiring to capture the join:

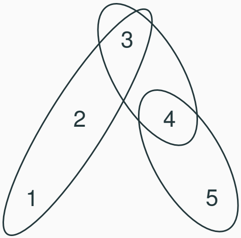
$$\Phi() = \bigcup_{x_1, x_2, x_3, x_4, x_5} \psi_{123}(x_1, x_2, x_3) \cap \psi_{34}(x_3, x_4) \cap \psi_{45}(x_4, x_5)$$

Φ maps the empty tuple to the join result.

In SQL, this join is expressed as follows:

```
SELECT * FROM Orders NATURAL JOIN Dishes NATURAL JOIN Items
```

DB: Example (2/4)



$$\underbrace{\texttt{Orders(customer, day, dish)}}_{\psi_{123}(x_1,x_2,x_3)}$$

$$\underbrace{\texttt{Dishes(dish, item)}}_{\psi_{34}(x_3,x_4)}$$

$$\underbrace{\texttt{Items (item, price)}}_{\psi_{45}(x_4,x_5)}$$

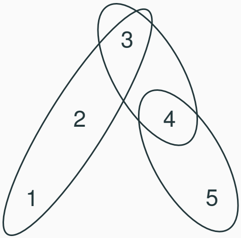
The FAQ over the Boolean semiring to capture the Boolean conjunctive query:

$$\Phi() = \bigvee_{x_1,x_2,x_3,x_4,x_5} \psi_{123}(x_1, x_2, x_3) \wedge \psi_{34}(x_3, x_4) \wedge \psi_{45}(x_4, x_5)$$

In SQL, this is expressed as follows:

```
SELECT true FROM Orders NATURAL JOIN Dishes NATURAL JOIN Items
```

DB: Example (3/4)



$$\underbrace{\texttt{Orders(customer, day, dish)}}_{\psi_{123}(x_1,x_2,x_3)}$$

$$\underbrace{\texttt{Dishes(dish, item)}}_{\psi_{34}(x_3,x_4)}$$

$$\underbrace{\texttt{Items (item, price)}}_{\psi_{45}(x_4,x_5)}$$

FAQ over the sum-product semiring to express a COUNT query:

$$\Phi(x_1) = \sum_{x_2,x_3,x_4,x_5} \psi_{123}(x_1, x_2, x_3) \cdot \psi_{34}(x_3, x_4) \cdot \psi_{45}(x_4, x_5)$$

Each factor maps tuples from corresponding relation to 1.

In SQL, this FAQ is expressed as follows:

```
SELECT customer, COUNT(*)
FROM Orders NATURAL JOIN Dishes NATURAL JOIN Items
GROUP BY customer
```

DB: Example (4/4)

More interesting aggregates captured by appropriately defining the factors

Query: Total price per customer and day

In FAQ:

- Let ψ_{45} map (x_4, x_5) to x_5 (price), all other factors map tuples to 1

$$\Phi(x_1, x_2) = \sum_{x_3, x_4, x_5} \psi_{123}(x_1, x_2, x_3) \cdot \psi_{34}(x_3, x_4) \cdot \psi_{45}(x_4, x_5)$$

In SQL:

```
SELECT customer, day, SUM(price)
FROM Orders NATURAL JOIN Dishes NATURAL JOIN Items
GROUP BY customer, day
```

BCQ: Boolean Conjunctive Queries

Compute the Boolean query $\exists X_1 \ldots \exists X_n : \bigwedge_{R \in \text{atoms}} R(\text{vars}(R))$

- atoms is the set of relation symbols in the query, e.g., Dishes(dish, item)
- Each relation symbol $R \in$ atoms has variables (attributes) vars(R)

FAQ encoding over the Boolean semiring $(\{\text{true}, \text{false}\}, \vee, \wedge, \text{false}, \text{true})$:

$$\Phi() = \bigvee_{\mathbf{x}} \bigwedge_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

- Φ has the hypergraph $(\mathcal{V}, \mathcal{E})$
- $\mathcal{V} = \bigcup_{R \in \text{atoms}(\Phi)} \text{vars}(R)$ and $\mathcal{E} = \{\text{vars}(R) \mid R \in \text{atoms}(\Phi)\}$
- For $S \in \mathcal{E}$ corresponding to relation R, there is a factor ψ_S such that

$$\psi_S(\mathbf{x}_S) = (\mathbf{x}_S \in R)$$

Join: Natural Join Queries

Compute the natural join query $\bowtie_{R \in \text{atoms}} R$

FAQ encoding over the set semiring $(2^{\mathcal{U}}, \cup, \cap, \emptyset, \mathcal{U})$:

$$\Phi() = \bigcup_{\mathbf{x}} \bigcap_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

- Φ has the hypergraph $(\mathcal{V}, \mathcal{E})$ and maps the empty tuple to the join result
- $\mathcal{U} = \prod_{i=1}^{n} \text{Dom}(X_i)$ is the set of all possible tuples
- $2^{\mathcal{U}}$ is the powerset of $\mathcal{U}$, i.e., the set of all possible subsets of $\mathcal{U}$
- $\mathcal{V}$ is the set of variables (attributes) in the atoms of the join query
- For $S \in \mathcal{E}$ corresponding to relation R, there is a factor ψ_S such that

$$\psi_S(\mathbf{x}_S) = \begin{cases} \{\mathbf{t} | \pi_S(\mathbf{t}) = \mathbf{x}_S\} & \text{if } \mathbf{x}_S \in R \\ \emptyset & \text{if } \mathbf{x}_S \notin R \end{cases}$$

MCM: Matrix Chain Multiplication

Compute the matrix product $\mathbf{A} = \mathbf{A}_1 \cdots \mathbf{A}_n$, where $\forall i \in [n] : \mathbf{A}_i \in \mathbb{R}^{p_i \times p_{i+1}}$

FAQ encoding over the real sum-product semiring $(\mathbb{R}, +, *, 0, 1)$:

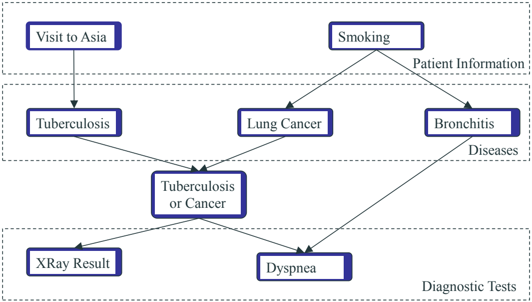
$$\Phi(x_1, x_{n+1}) = \sum_{x_2 \in \text{Dom}(X_2)} \cdots \sum_{x_n \in \text{Dom}(X_n)} \prod_{i \in [n]} \psi_{i,i+1}(x_i, x_{i+1}).$$

- We use $n + 1$ variables $X_1, \ldots, X_{n+1}$ with domains $\text{Dom}(X_i) = [p_i]$
- Each matrix $\mathbf{A}_i$ can be viewed as a function of two variables:

  $$\psi_{i,i+1} : \text{Dom}(X_i) \times \text{Dom}(X_{i+1}) \to \mathbb{R}, \text{ where } \psi_{i,i+1}(x_i, x_{i+1}) = (\mathbf{A}_i)_{x_i x_{i+1}}$$

  One variable is the row index, the other variable is the column index

PGM: Example



Network represents a knowledge structure that models the relationship between diseases, their causes and effects, patient information and diagnostic tests

A	$P(A)$
T	.01
F	.99

S	$P(S)$
T	.4
F	.6

A T	$P(T\|A)$
T T	.05
T F	.95
F T	.01
F F	.99

S B	$P(B\|S)$
T T	.6
T F	.4
F T	.3
F F	.7

S L	$P(L\|S)$
T T	.1
T F	.9
F T	.01
F F	.99

T L O	$P(O\|T, L)$
T T T	1
T T F	0
T F T	1
T F F	0
F T T	1
F T F	0
F F T	0
F F F	1

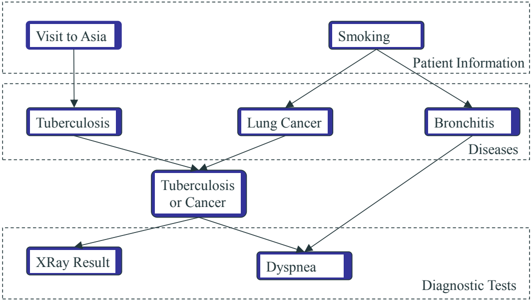
O B D	$P(D\|O, B)$
T T T	.9
T T F	.1
T F T	.7
T F F	.3
F T T	.8
F T F	.2
F F T	.1
F F F	.9

O X	$P(X\|O)$
T T	.98
T F	.02
F T	.05
F F	.95

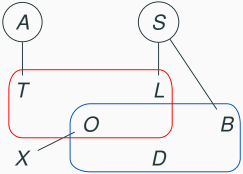
Variable O:
Tuberculosis or Cancer

PGM: Example

Hypergraph: one node per random variable and one hyperedge per conditional probability table



Network represents a knowledge structure that models the relationship between diseases, their causes and effects, patient information and diagnostic tests



One factor ψ_S for each conditional probability table over variables S

- Maps tuple x_S of values of variables S to the associated probability
- Factors: ψ_A, $\psi_{A,T}$, ψ_S, $\psi_{S,L}$, $\psi_{S,B}$, $\psi_{T,L,O}$, $\psi_{O,X}$, $\psi_{O,B,D}$

PGM: Example

The original joint probability distribution

$$P(A, T, S, L, B, O, X, D) = P(A) \cdot P(T|A) \cdot P(S) \cdot P(L|S) \cdot P(B|S) \cdot P(O|T, L) \cdot P(X|O) \cdot P(D|O, B)$$

PGM: Example

The original joint probability distribution and the corresponding FAQ encoding

$$\underbrace{P(A, T, S, L, B, O, X, D)}_{\Phi} = \underbrace{P(A)}_{\psi_A} \cdot \underbrace{P(T|A)}_{\psi_{A,T}} \cdot \underbrace{P(S)}_{\psi_S} \cdot \underbrace{P(L|S)}_{\psi_{S,L}} \cdot \underbrace{P(B|S)}_{\psi_{S,B}} \cdot \underbrace{P(O|T, L)}_{\psi_{T,L,O}} \cdot \underbrace{P(X|O)}_{\psi_{O,X}} \cdot \underbrace{P(D|O, B)}_{\psi_{O,B,D}}$$

PGM: Example

The original joint probability distribution and the corresponding FAQ encoding

$$\underbrace{P(A,T,S,L,B,O,X,D)}_{\Phi} = \underbrace{P(A)}_{\psi_A} \cdot \underbrace{P(T|A)}_{\psi_{A,T}} \cdot \underbrace{P(S)}_{\psi_S} \cdot \underbrace{P(L|S)}_{\psi_{S,L}} \cdot \underbrace{P(B|S)}_{\psi_{S,B}} \cdot \underbrace{P(O|T,L)}_{\psi_{T,L,O}} \cdot \underbrace{P(X|O)}_{\psi_{O,X}} \cdot \underbrace{P(D|O,B)}_{\psi_{O,B,D}}$$

Marginal Distribution $P(A, B, D)$ using the sum-product semiring:

$$\Phi(a,b,d) = \sum_{t,s,l,o,x} \psi_A(a) \cdot \psi_{A,T}(a,t) \cdot \psi_S(s) \cdot \psi_{S,L}(s,l) \cdot \psi_{S,B}(s,b) \cdot \psi_{T,L,O}(t,l,o) \cdot \psi_{O,X}(o,x) \cdot \psi_{O,B,D}(o,b,d)$$

Maximum A-Posteriori $P(A, B, D)$ using the max-product semiring:

$$\Phi(a,b,d) = \max_{t,s,l,o,x} \psi_A(a) \cdot \psi_{A,T}(a,t) \cdot \psi_S(s) \cdot \psi_{S,L}(s,l) \cdot \psi_{S,B}(s,b) \cdot \psi_{T,L,O}(t,l,o) \cdot \psi_{O,X}(o,x) \cdot \psi_{O,B,D}(o,b,d)$$

MAP-PGM: Maximum A-Posteriori in Probabilistic Graphical Models

Compute the MAP estimate over variables $X_1, \ldots, X_f$

FAQ encoding over the max-product semiring $([0, \infty), \max, *, 0, 1)$

$$\Phi(x_1, \ldots, x_f) = \max_{x_{f+1} \in \mathrm{Dom}(X_{f+1})} \cdots \max_{x_n \in \mathrm{Dom}(X_n)} \prod_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

where $(\mathcal{V}, \mathcal{E})$ is the hypergraph of the (undirected) probabilistic graphical model

- $\mathcal{V} = [n]$: indices of n discrete random variables $X_1, \ldots, X_n$
- There is a factor $\psi_S : \prod_{i \in S} \mathrm{Dom}(X_i) \rightarrow [0, \infty)$ for each edge $S \in \mathcal{E}$

MD-PGM: Marginal Distribution in Probabilistic Graphical Models

Compute the marginal distribution of the set of variables $X_1, \ldots, X_f$

FAQ encoding over the sum-product semiring $(\mathbb{R}_+, +, *, 0, 1)$

$$\Phi(x_1, \ldots, x_f) = \sum_{x_{f+1} \in \mathrm{Dom}(X_{f+1})} \cdots \sum_{x_n \in \mathrm{Dom}(X_n)} \prod_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

where $(\mathcal{V}, \mathcal{E})$ is the hypergraph of the (undirected) probabilistic graphical model

- $\mathcal{V} = [n]$: indices of n discrete random variables $X_1, \ldots, X_n$
- There is a factor $\psi_S : \prod_{i \in S} \mathrm{Dom}(X_i) \to \mathbb{R}_+$ for each edge $S \in \mathcal{E}$

For conditional distributions $P(\mathbf{X}_A \mid \mathbf{X}_B = \mathbf{x}_B)$, we set $\mathbf{X}_B$ to $\mathbf{x}_B$.

Sample of Problems Expressible in FAQ

Boolean Semiring

$(\{\text{true}, \text{false}\}, \vee, \wedge, \text{false}, \text{true})$

- Constraint satisfaction problems (CSP) [FAQ]
- Boolean conjunctive query evaluation (BCQ) [FAQ]
- Conjunctive query evaluation (CQ)* [FAQ]
- Join evaluation [FAQ]
- Satisfiability (SAT) [FAQ]
- *k*-colorability [FAQ]
- List recovery problem (coding theory) [FAQ]

(*) also expressible using the set semiring

Set and Natural Sum-Product Semirings

$(2^{\mathcal{U}}, \cup, \cap, \emptyset, \mathcal{U})$

- Conjunctive query evaluation (CQ)* [FAQ]
- Join evaluation [FAQ]

$(\mathbb{N}, +, *, 0, 1)$

- Complex network analysis [FAQ]
- Count constraint satisfaction problems (#CSP) [FAQ]
- Count satisfiability (#SAT) [FAQ]

(*) also expressible using the Boolean semiring

Real Sum-Product Semiring

$(\mathbb{R}, +, *, 0, 1)$

- Permanent [FAQ]
- Discrete Fourier transform [FAQ,AjiMcEl]
- Hadamard transform [AjiMcEl]
- Inference in probabilistic graphical models [FAQ]
- Probability propagation in AI [AjiMcEl]
- Matrix chain multiplication [FAQ,AjiMcEl]
- Graph homomorphism [FAQ]
- BCJR decoding (Bahl, Cocke, Jelinek, Raviv) [AjiMcEl]
- Holant problem [FAQ]

Max-Product and Min-Sum Semirings

$([0,\infty), \max, *, 0, 1)$

- MAP queries in probabilistic graphical models [FAQ]
- Quantified conjunctive query evaluation (QCQ)* [FAQ]

$((-\infty,\infty], \min, +, \infty, 0)$

- Gallager-Tanner-Wiberg decoding [AjiMcEl]
- Viterbi decoding [AjiMcEl]
- Trellis path problem [AjiMcEl]
- Graph optimization [KohlWils]
- Queuing systems [KohlWils]
- Discrete event systems [KohlWils]
- Optimization for weighted CSPs [KohlWils]

(*) also expressible using the max-product, min-product semirings

Problems Expressible With Two Semirings

$([0, \infty), \max, *, 0, 1)$, $((0, \infty], \min, *, \infty, 1)$

- Quantified conjunctive query evaluation (QCQ)* [FAQ]

$(\mathbb{N}, \max, *, 0, 1)$, $(\mathbb{N}, +, *, 0, 1)$

- Count conjunctive query evaluation (#CQ) [FAQ]
- Count quantified conjunctive query evaluation (#QCQ) [FAQ]

(*) also expressible using the max-product semiring