

Logical and Statistical Constraints

Michel de Rougemont

University Paris II & IRIF CNRS

(joint work with R. Lassaigne)



Constraints in Data

1. Statistical constraints

- Statistical queries
- Relation on Distributions
- Definability
- Declarative queries for Prediction:

2. CQA with both constraints

- FD as logical constraints
- Statistical constraints
- Repairs have natural probabilities

1. Statistical constraints

1. Statistical query

$$\delta(x) = y \quad \sum_x y = 1$$

$$\delta(x_1, x_2) = y \quad \sum_{x_1, x_2} y = 1$$

1. Statistical constraint

Relation between 2 distributions

- Equality: $\delta_1 = \delta_2$
- $\text{dist}(\delta_1, \delta_2) < \varepsilon$

Statistical query

1. Words on {a,b}

- Distribution of letters
- Distribution of k-factors $ustat_k$
- $Next_k$
 - Alphabet of Tokens (3.10^4 for Chatgpt)

2. Graphs

- Degree distribution

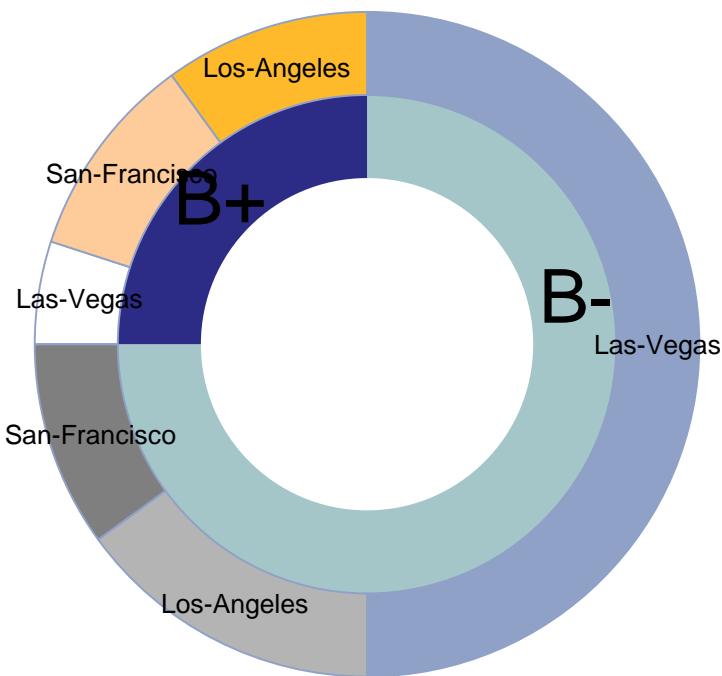
3. Relational Data

- OLAP

2-dimensional distribution

0.5	0.15	0.1
0.1	0.1	0.05

0.75
0.25



2-dimensional distribution

$\delta_1(x, y)$

0.5	0.15	0.1
0.1	0.1	0.05

0.75
0.25

$\delta_2(x, y)$

0.6	0.15	0.1
0.05	0.05	0.05

0.85
0.15

$$\text{Dist}(\delta_1, \delta_2) = 0.2$$

$\delta_1(y, x)$

0.5	0.1
0.15	0.1
0.1	0.05

0.6
0.25
0.15

Distances

1. L_1

- Sum of the Δ : représentation matricielle
- Generalization to L_p

2. EMD

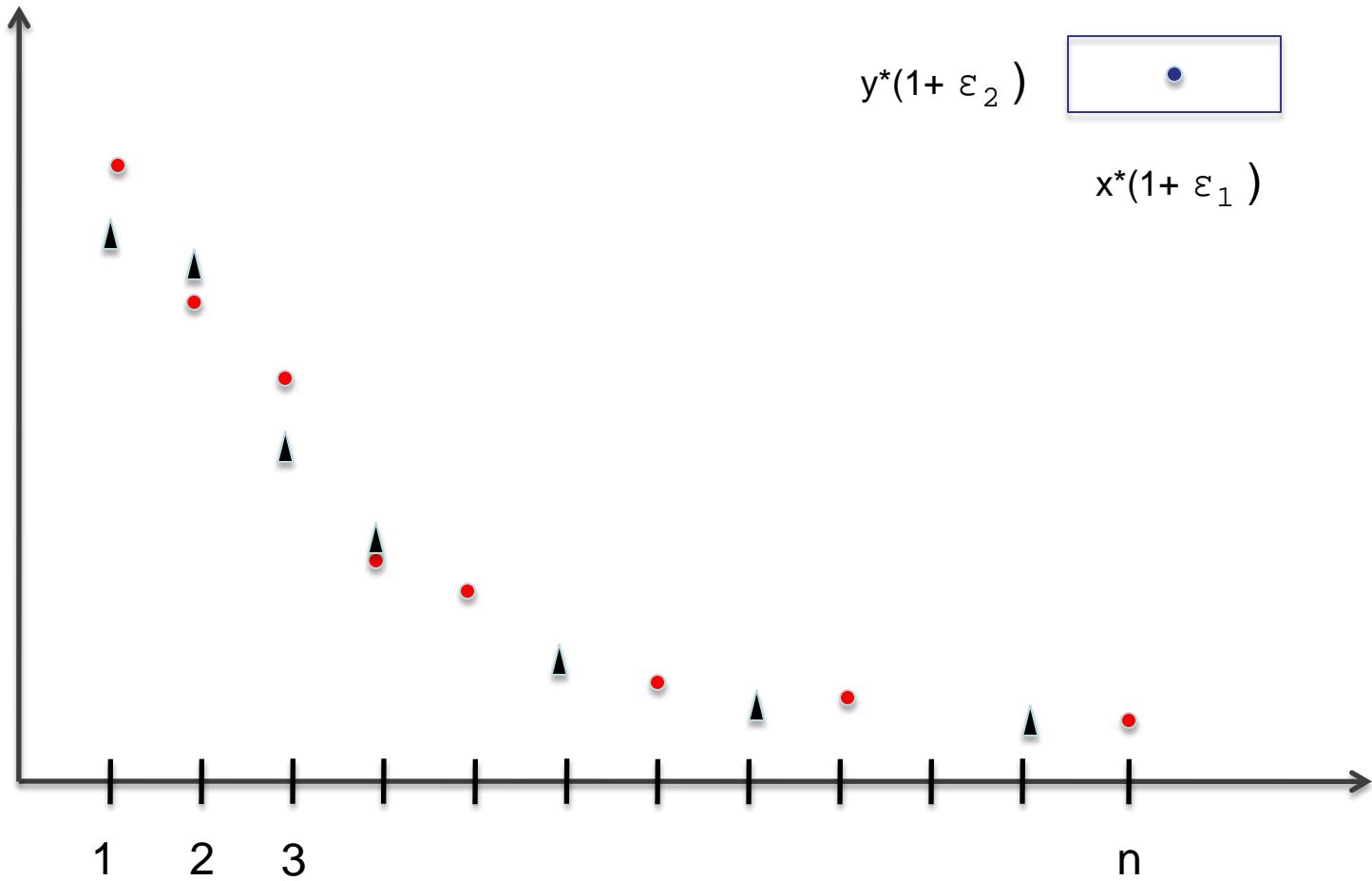
- Earth-moving distance

3. Fréchet

- Absolute/relative

4. Kullback-Liebler Divergence

Fréchet relative distance



Two degree distributions

Definability of Counting Queries

1. Words $W_n = (\{1, 2, \dots, n\}, \Sigma_k, P, <; Q, +, ^*, /)$

$Q(x): \exists y P(x, y)$

$Q(x): \#y P(x, y)$

$Q(x): \%y P(x, y)$

$Q(x, y_1, y_2): \exists x' P(x, y_1) \cap y_1 < y_2 \cap P(x', y_2)$

$Q(x): \exists x', y_1, y_2 P(x, y_1) \cap y_1 < y_2 \cap P(x', y_2)$

$Q(x): \%(y_1, y_2) \exists x' P(x, y_1) \cap y_1 < y_2 \cap P(x', y_2)$

$Q(x): \%(y_1, y_2) \exists x' P(x, y_1) \cap y_2 = y_1 + 1 \cap P(x', y_2)$

Definability of Counting Queries

Graphs $G_n = (\{1, 2, \dots, n\}, E; Q, +, ^*, /)$

$$Q(x, i): \exists^i y \ E(x, y)$$

$$Q_1(i): \%x \ \exists^i y \ E(x, y)$$

Zipf degree distribution:

$$Q(): \%x \ \exists^i y \ E(x, y) = c/i^2$$

OLAP Queries

CUST		
CID	CNAME	City
C1	John	L.A. <i>f₁</i>
C2	Mary	L.V. <i>f₂</i>
C2	Mary	S.F. <i>f₃</i>
C3	Don	L.V. <i>f₄</i>
C4	Jen	L.A. <i>f₅</i>

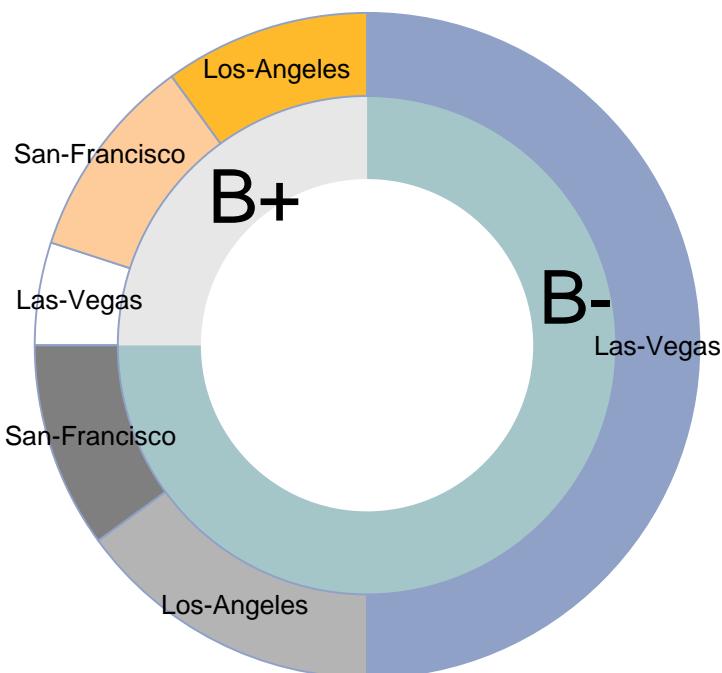
ACCOUNTS			
ACCID	Type	BAL	
A1	Checking	900	<i>f₆</i>
A2	Checking	1000	<i>f₇</i>
A3	Saving	1200	<i>f₈</i>
A3	Saving	-100	<i>f₉</i>
A4	Saving	-300	<i>f₁₀</i>

CUSTACC		
CID	ACCID	
C1	A1 <i>f₁₁</i>	
C2	A2 <i>f₁₂</i>	
C2	A3 <i>f₁₃</i>	
C3	A4 <i>f₁₄</i>	

VISITS				
CID	ACCID	DATE	DURATION	
C1	A1	Jan 1.	15mn	<i>f₁₅</i>
C2	A2	Jan 2.	30 mn	<i>f₁₆</i>
C2	A2	Jan 2.	20 mn	<i>f₁₇</i>
C3	A4	Jan 3.	45mn	<i>f₁₈</i>
C2	A3	Jan 5.	30 mn	<i>f₁₉</i>
C2	A3	Jan 6.	10 mn	<i>f₂₀</i>

Statistical constraint: 2-dimensional distribution

0.5	0.15	0.1	0.75
0.1	0.1	0.05	0.25



Definability of OLAP Queries

OLAP schema: FDs

Dimensions: variables

Aggregation operators: Count, Sum(Duration)

Generalize the quantifier $\%x$ to $\%Y \quad Y = \sum y \quad y = \text{Duration}$

$Q(z): \% (x, u, v, w) \exists \quad y \text{Cust}(x, y, z) \cap \text{Visits}(x, u, v, w)$

$Q(z): \% t \quad t = (x, u, v, w) \cap \exists \quad y$
 $\text{Cust}(x, y, z) \cap \text{Visits}(x, u, v, w)$

$Q(\textcolor{red}{z}): \% W \quad W = \sum t.w \cap t = (x, u, v, w) \cap \exists \quad y, t, u, v, w$
 $\text{Cust}(x, y, \textcolor{red}{z}) \cap \text{Visits}(x, u, v, w)$

Analysis of Sum of durations by Cities

Prediction: Graph Learning, GNN

Predict Sales for next week, for active buyers

Selection: Active buyers, (customers who bought something last month)

Dimension: customer

Measure: Price

Aggregation: Sum

Other solution: Count

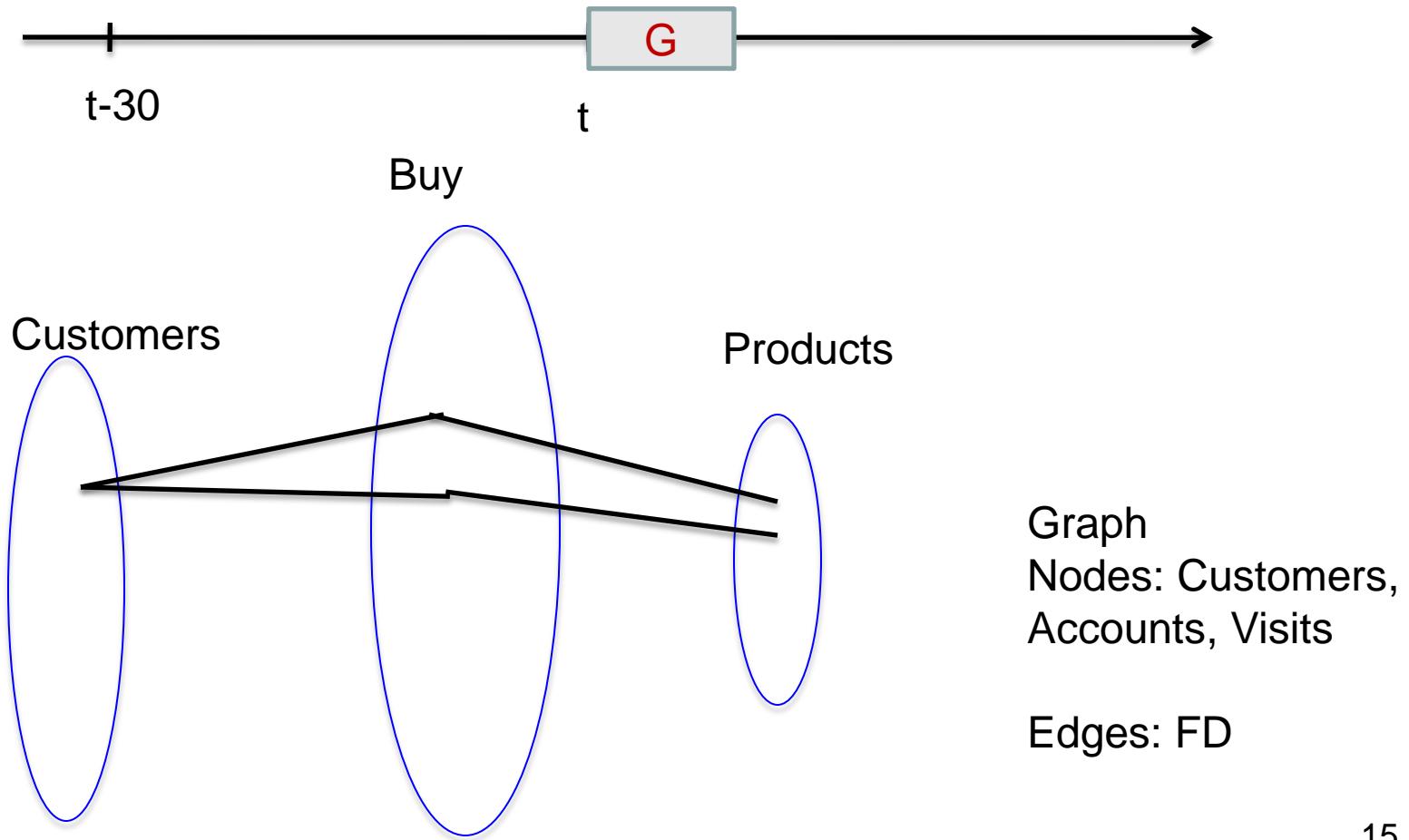
Extension of SQL with a **Predict** operator (J. Leskovec)

<https://kumo.ai/freetrial>

Prediction: Graph Learning, GNN

Predict Sales for next week, for active buyers

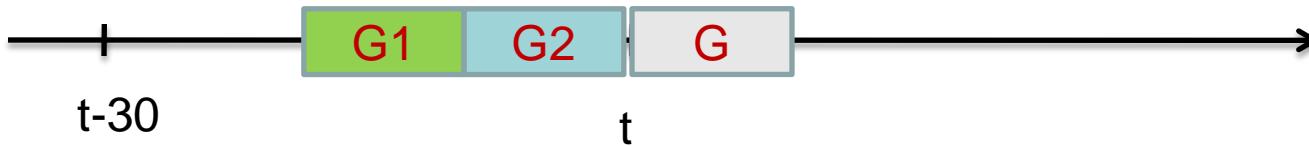
Time is a key attribute:



Prediction: Graph Learning, GNN

Predict Sales for next week, for active buyers

Time is a key attribute:



Analyze OLAP query on G1, G2: δ_1, δ_2

Output distribution: Interpolation: $\delta_1 + \delta_2$

Output graph $G = « G1+G2 »$

2. CQA with logical and statistical constraints

Source has errors:

Repair: minimal structure which satisfies logical constraints

$$r = (ab)^*c^*$$

w: abababb**b**abcc**a**c

R1:ababab**a**babcccc

R2:ababababcccc

R3:abababcccc not a repair

Relational Data FDs $CID \rightarrow CNAME, City$ $ACCID \rightarrow Type, BAL$

CUST

CID	CNAME	City	
C1	John	L.A.	f_1
C2	Mary	L.V.	f_2
C2	Mary	S.F.	f_3
C3	Don	L.V.	f_4
C4	Jen	L.A.	f_5

ACCOUNTS

ACCID	Type	BAL	
A1	Checking	900	f_6
A2	Checking	1000	f_7
A3	Saving	1200	f_8
A3	Saving	-100	f_9
A4	Saving	-300	f_{10}

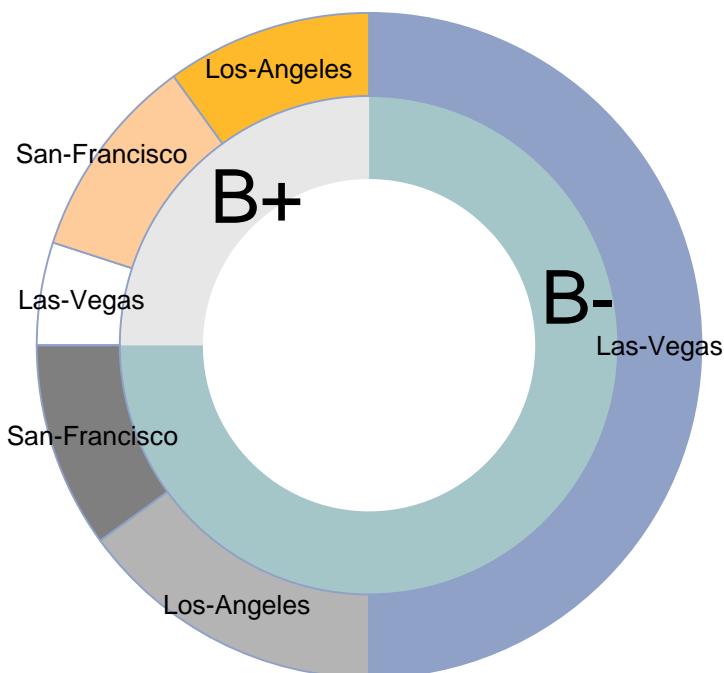
VISITS

CUSTACC	
CID	ACCID
C1	A1
C2	A2
C2	A3
C3	A4

CID	ACCID	DATE	DURATION	
C1	A1	Jan 1.	15mn	f_{15}
C2	A2	Jan 2.	30 mn	f_{16}
C2	A2	Jan 2.	20 mn	f_{17}
C3	A4	Jan 3.	45mn	f_{18}
C2	A3	Jan 5.	30 mn	f_{19}
C2	A3	Jan 6.	10 mn	f_{20}

Statistical constraint: 2-dimensional distribution

0.5	0.15	0.1	0.75
0.1	0.1	0.05	0.25



Probabilistic Repairs

Repairs: choose R_i with probability p_i

The probability is related to the distance to the statistical constraint.

If distance is small, probability is high:

$$p = c \cdot e^{-dist(\delta_1, \delta_2)}$$

Probabilistic space: probabilistic partition

Modèle partitionnel

Source

ID	A	B	C
1	a	b	c
2	b	a	c'
3	a	b'	c

$I(a)=\{(1,1),(2,2),(3,1)\}$ notation matricielle

$I(b)=\{(1,2),(2,1)\}$

$I(ab)=\{(1,1,2)\}$

$I(ba)=\{(2,1,2)\}$

Modèle probabiliste 1:

Î sous-ens aléatoire de I : {1,2,3} p1, p2, p3
Proba {1,2} p1.p2. (1-p3)

$\hat{I}(a)=\{(1,1),(2,2)\}$

Lien avec les BD relationnelles probabilistes

Modèle relationnel probabiliste

S1

ID	A	B	C	
1	a	b	c	0.8
2	b	a	c'	0.4
3	a	b'	c	0.2

$$I(a) = \{(1,1), (2,2), (3,1)\} \text{ notation matricielle}$$
$$I(b) = \{(1,2), (2,1)\}$$

$$I(ab) = \{(1,1,2)\}$$
$$I(ba) = \{(2,1,2)\}$$

Modèle probabiliste 1:

\hat{I} sous-ensemble aléatoire de $I = \{1,2,3\}$ $p_1=0.8$, $p_2=0.4$, $p_3=0.2$
Proba [$\hat{I} = \{1,2\}$] = $p_1.p_2. (1-p_3)$

$$\hat{I}(a) = \{(1,1), (2,2)\}$$

$$\hat{I}(b) = \{(1,2), (2,1)\}$$

Les BD relationnelles probabilistes spécifient p_1, p_2, \dots, p_n

Modèle relationnel probabiliste

S1

ID	A	B	C	
1	a	b	c	0.8
2	b	a	c'	0.4
3	a	b'	c	0.2

$$I(a) = \{(1,1), (2,2), (3,1)\} \text{ notation matricielle}$$
$$I(b) = \{(1,2), (2,1)\}$$

$$I(ab) = \{(1,1,2)\}$$
$$I(ba) = \{(2,1,2)\}$$

Modèle relationnel probabiliste : \hat{U}

$$\text{Proba } [\hat{U} \models Q] = \frac{\#\hat{U} \text{ satisfont } Q}{2^n} :$$

Modèle d'apprentissage des probabilités à partir d'exemples. [V. d. Broeck 2020]

Modèle de corrélations

ID	A	B	C
1	a	b	c
2	b	a	c'
3	a	b'	c
4	a	a	a
5	a	b	c'

$$I(a)=\{(1,1),(2,2),(3,1),(4,1), (4,2), (4,3), (5,1),\}$$

$$I(b)=\{(1,2),(2,1),(5,2)\}$$

$$I(ab)=\{ (1,1,2),(5,1,2) \}$$

$$I(ba)=\{ (2,1,2) \}$$

Modèle probabiliste non indépendant:

† sous-ensemble aléatoire de I : {1,2,3,4,5}

Proba {1,2, 3} p=0.8 sinon {1}, {2}, {3} (1-p)/3

Proba {4,5} q =0.4 sinon {4}, {5} (1-q)/2

†(a)={(1,1),(2,2),(3,1)} avec proba 0.8. (1-0.4)=0.48

Modèle partitionnel non indépendant

ID	A	B	C
1	a	b	c
2	b	a	c'
3	a	b'	c
4	a	a	a
5	a	b	c'

$$I(a)=\{(1,1),(2,2),(3,1),(4,1), (4,2), (4,3), (5,1),\}$$

$$I(b)=\{(1,2),(2,1),(5,2)\}$$

$$I(ab)=\{ (1,1,2),(5,1,2) \}$$

$$I(ba)=\{ (2,1,2) \}$$

Modèle probabiliste non indépendant:

Q: A=a and C=c

$$\text{Proba} [\hat{U} \models Q] = 0.8 + 0.133 = 0.933$$

Events: {1,2,3}, {1}, {3}

Modèle partionnel probabiliste 2 tables

S1

ID	A	B	C
1	a	b	c
2	b	a	c'
3	a	b'	c

ID	A	D	E
10	a	d	e
20	b	d'	e'
30	a	d'	e

Modèle probabiliste non indépendant:

l' sous-ensemble aléatoire de l : $\{1,2,3,4,5\}^*\{10,20,30\}$

Proba $\{1,2\}^*\{10,20\}$ p=0.8 sinon $\{1\}^*\{20\}$ (1-p)

Proba $\{3\}^*\{30\}$ q =0.4 sinon $\{3\}^*\{\}$ (1-q)

$\hat{l}=\{1,2,3\}^*\{10,20\}$ avec proba 0.8. (1-0.4)=0.48

Modèle partionnel pour CUST

CUST				
CID	CNAME	City		
1	C1	John	L.A.	f_1
2	C2	Mary	L.V.	f_2
3	C2	Mary	S.F.	f_3
4	C3	Don	L.V.	f_4
5	C4	Jen	L.A.	f_5

$$R1 = \{1, 2, 4, 5\}$$

avec proba p et

R2={1,3,4,5}

1-p

Distance to SC: 0.7

0,5

p= 0.4

$$1-p=0.6$$

Modèle partionnel pour CUST et VISITS

Las-Vegas

VISITS				
CID	ACCID	DATE	DURATION	
C1	A1	Jan 1.	15mn	f_{15}
C2	A2	Jan 2.	30 mn	f_{16}
C2	A2	Jan 2.	20 mn	f_{17}
C3	A4	Jan 3.	45mn	f_{18}
C2	A3	Jan 5.	30 mn	f_{19}
C2	A3	Jan 6.	10 mn	f_{20}

$$R1 = \{1, 2, 4, 5\}$$

avec proba p et $R2 = \{1, 3, 4, 5\}$ $1-p$

$$\text{Dist}(\delta_1, \delta_0) = 0.2$$

$$p = 0.7$$

$$\text{Dist}(\delta_2, \delta_0) = 0.5$$

$$1-p = 0.3$$

Link to MaxSAT

Consider Q1, Q2

- Stat. Const: reference distribution delta
- Errors on CUST table (m errors)
- Probabilistic corrections
- Evaluate Q1 and Q2

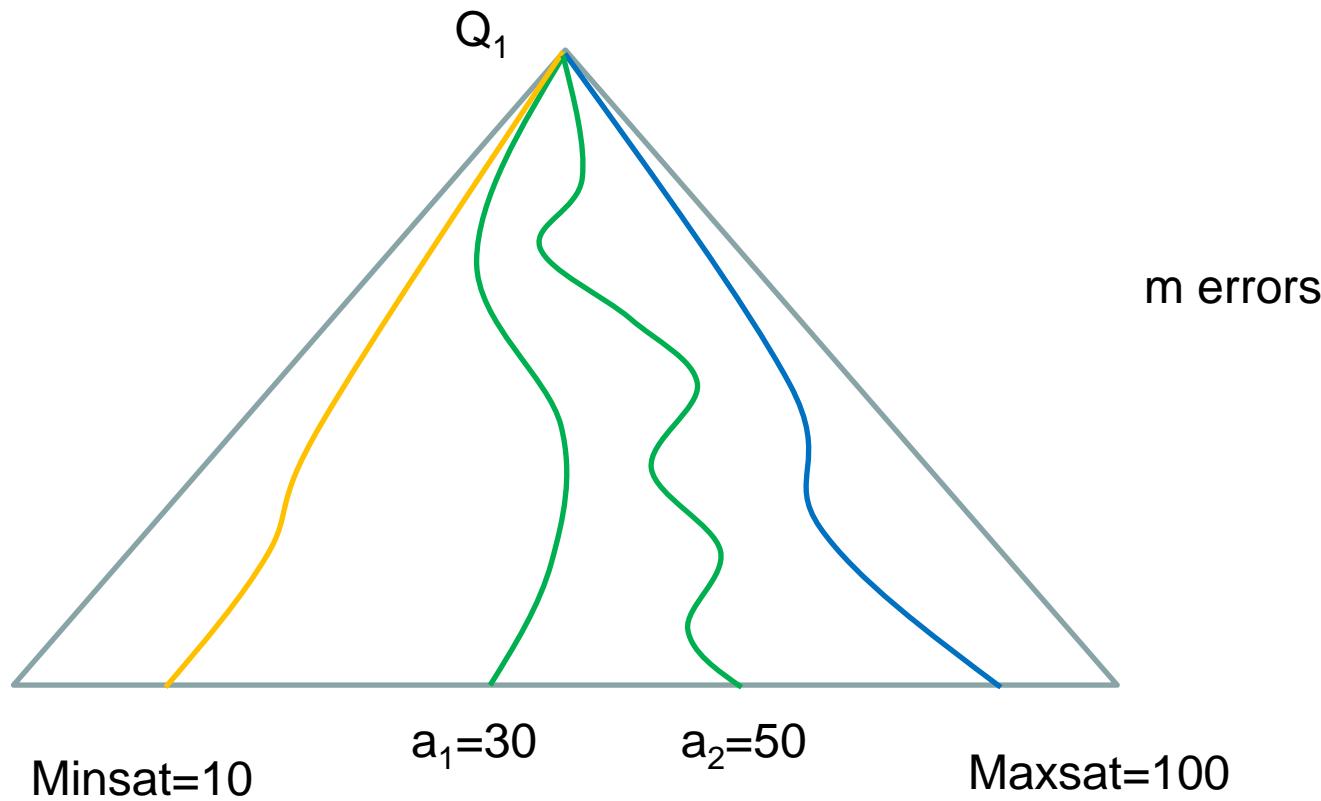
Monte-Carlo algorithm:

For each error, sample the repairs

Evaluate the counting query: $a_1, a_2 \dots a_k$

Output the interval [min, max]

Link to MaxSAT: probabilistic space



DK answer: [10, 100]

Our answer: [30, 50]

Conclusion

1. Statistical constraints
 - Statistical query
 - Relation between statistical queries
2. CQA with logical and statistical constraints
 - Repair: probabilistic R_i with the partition model
 - Monte Carlo algorithm