



THE UNIVERSITY
of EDINBURGH

Database Tuples Play Cooperative Games

Ester Livshits

Joint work with:

Leopoldo Bertossi, Benny Kimelfeld, Alon Reshef, Moshe Sebag

AUTHOR	
Name	Affiliation
Alice	UCLA
Bob	NYU
Cathy	MIT
David	UCSD
Ellen	NYU

INSTITUTE	
Name	STATE
UCLA	CA
UCSD	CA
NYU	NY
MIT	MA

PUBLICATION	
Author	Paper
Alice	A
Alice	B
Bob	C
Cathy	C
Cathy	D
David	C

CITATIONS	
PAPER	CITS
A	18
B	2
C	8
D	12

$q(z, w): \neg \text{AUTHOR}(x, y), \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, z), \text{CITATIONS}(z, w)$

Why we obtained a particular answer?

PAPER	CITS
A	18
B	2
C	8

Why we did **not obtain some other answer?**

AUTHOR	
Name	Affiliation
Alice	UCLA
Bob	NYU
Cathy	MIT
David	UCSD
Ellen	NYU

INSTITUTE	
Name	STATE
UCLA	CA
UCSD	CA
NYU	NY
MIT	MA

PUBLICATION	
Author	Paper
Alice	A
Alice	B
Bob	C
Cathy	C
Cathy	D
David	C

CITATIONS	
PAPER	CITS
A	18
B	2
C	8
D	12

$q(z, w): \neg \text{AUTHOR}(x, y), \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, z), \text{CITATIONS}(z, w)$

Why we obtained a particular answer?

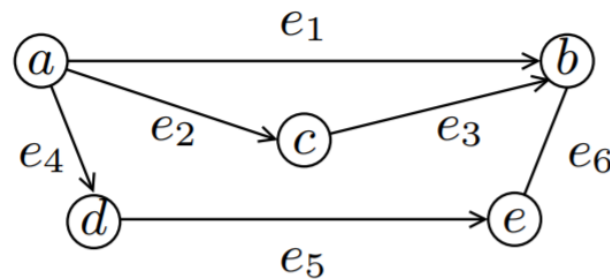
PAPER	CITS
A	18
B	2
C	8

Why we did *not* obtain some other answer?

**Which tuples in the database
explain the query result?**

Measuring Contribution

- Causal responsibility [Meliou et al. 2010]
 - ❖ t is a **counterfactual cause** for q if $D \models q$ and $D \setminus \{t\} \not\models q$
 - ❖ t is an **actual cause** for q if $D \setminus \Gamma \models q$ and $D \setminus \{\Gamma \cup \{t\}\} \not\models q$ for some $\Gamma \subseteq D \setminus \{t\}$
 - ❖ The responsibility of t is $\frac{1}{1+|\Gamma_{\min}|}$
- Not extendable to aggregate queries
- May be counterintuitive



Is there a path from a to b?

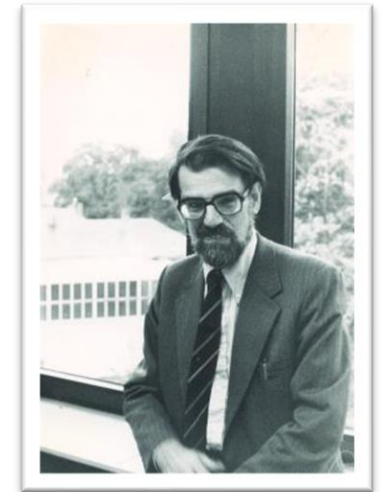
Measuring Contribution

- Causal effect [Salimi et al. 2016]
 - ❖ See the database as a probabilistic database
 - ❖ $CE(t) = E(q \mid t \in D) - E(q \mid t \notin D)$

What makes the choice of a contribution score a good one?

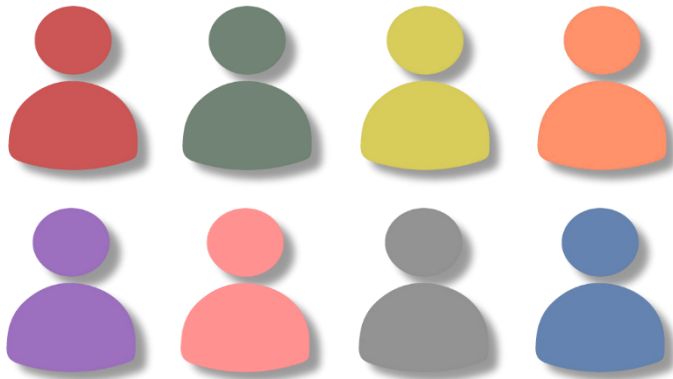
Shapley Value

- A widely known profit-sharing formula in cooperative game theory
- Introduced by Lloyd Shapley in 1953
- Applied in various areas beyond cooperative game theory:
 - ❖ Pollution responsibility in **environmental management**
 - ❖ Influence measurement in **social network analysis**
 - ❖ Identifying candidate autism **genes**
 - ❖ Bargaining foundations in **economics**
 - ❖ Takeover corporate rights in **law**
 - ❖ Explanations in **machine learning**



Shapley Value




Set A of players:





How to distribute the total wealth among the players?

Wealth function $v: \mathcal{P}(A) \rightarrow \mathbb{R}$:



Machine learning
[Lundberg, Lee 2017]

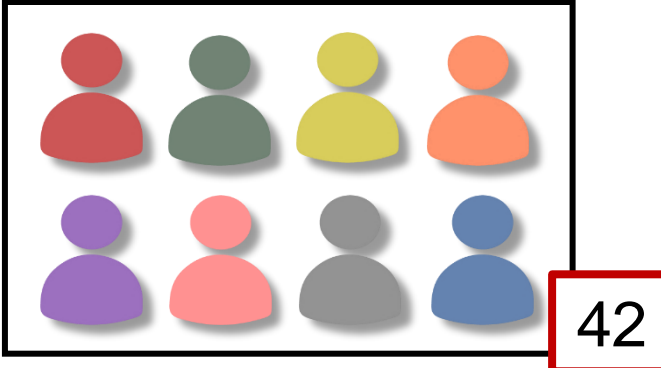
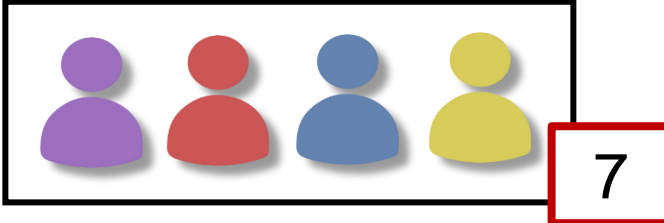
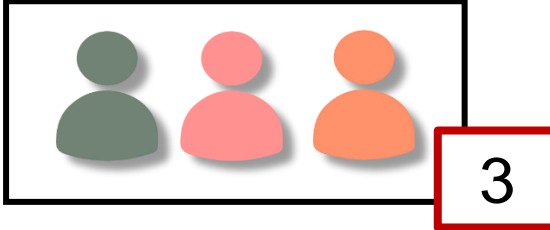
Features  Prediction  

Query answering
[L et al. 2020]

Tuples  Answer 

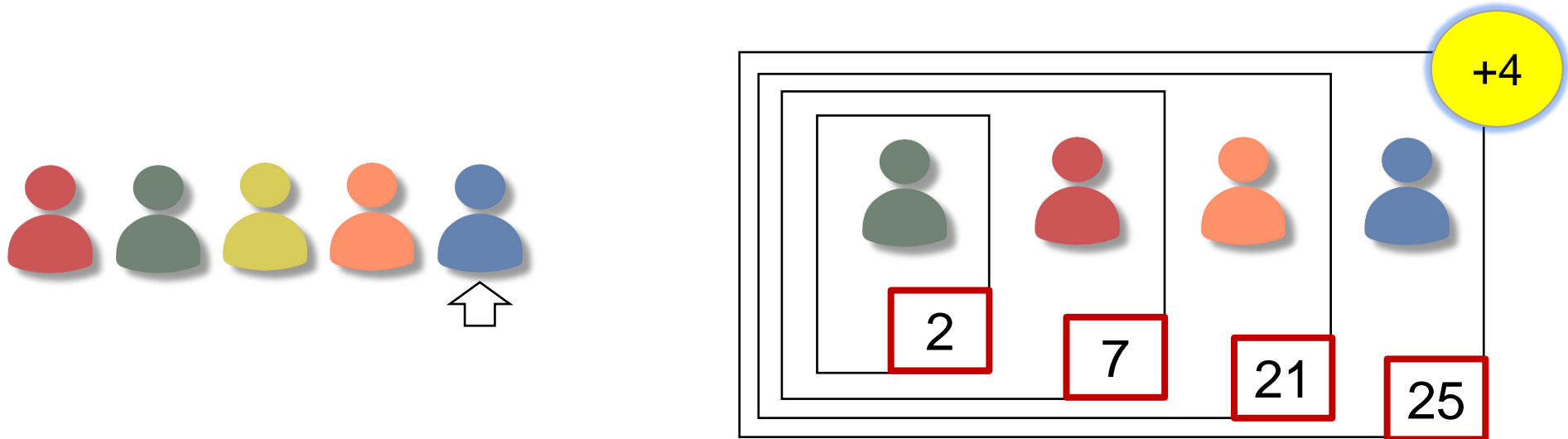
Inconsistency
[L, Kimelfeld 2021]

Tuples  Measure 



Shapley Value

$$\text{Shapley}(A, v, a) = \sum_{B \subseteq A \setminus \{a\}} \frac{|B|! (|A| - |B| - 1)!}{|A|!} (v(B \cup \{a\}) - v(B))$$



The Shapley value is the expected delta due to the addition in a random permutation

Shapley Value for Database Queries

➤ Which tuples in the database explain the query result?

AUTHOR	
Name	Affiliation
Alice	UCLA
Bob	NYU
Cathy	MIT
David	UCSD
Ellen	NYU

INSTITUTE	
Name	STATE
UCLA	CA
UCSD	CA
NYU	NY
MIT	MA

PUBLICATION	
Author	Paper
Alice	A
Alice	B
Bob	C
Cathy	C
Cathy	D
David	C

CITATIONS	
PAPER	CITS
A	18
B	2
C	8
D	12

Players

$q(z, w): -\text{AUTHOR}(x, y), \text{PUBLICATION}(x, z), \text{CITATIONS}(z, w)$

$\text{SUM}_w \langle q(z, w) \rangle$

← Wealth function

$SV(\text{Alice}) = 20$
 $SV(\text{Cathy}) = 14.67$
 $SV(\text{Bob}) = 2.67$
 $SV(\text{David}) = 2.67$
 $SV(\text{Ellen}) = 0$

AUTHOR	
Name	Affiliation
Alice	UCLA
Bob	NYU
Cathy	MIT
David	UCSD
Ellen	NYU

INSTITUTE	
Name	STATE
UCLA	CA
UCSD	CA
NYU	NY
MIT	MA

PUBLICATION	
Author	Paper
Alice	A
Alice	B
Bob	C
Cathy	C
Cathy	D
David	C

CITATIONS	
PAPER	CITS
A	18
B	2
C	8
D	12

$q(z, w): \neg \text{AUTHOR}(x, y), \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, z), \text{CITATIONS}(z, w)$

PAPER	CITS
A	18
B	2
C	8

AUTHOR	
Name	Affiliation
Alice	UCLA
Bob	NYU
Cathy	MIT
David	UCSD
Ellen	NYU

INSTITUTE	
Name	STATE
UCLA	CA
UCSD	CA
NYU	NY
MIT	MA

PUBLICATION	
Author	Paper
Alice	A
Alice	B
Bob	C
Cathy	C
Cathy	D
David	C

CITATIONS	
PAPER	CITS
A	18
B	2
C	8
D	12

$q(z, w): \neg \text{AUTHOR}(x, y), \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, z), \text{CITATIONS}(z, w)$

PAPER	CITS
A	18
B	2
C	8

$q(): \neg \text{AUTHOR}(x, y), \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, 'A'), \text{CITATIONS}('A', 18)$

Outline

- Explaining Query Answers
- **Computational Complexity**
- Responsibility to Inconsistency

$q(): -\text{AUTHOR}(x, y), \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, z)$

Computational Complexity

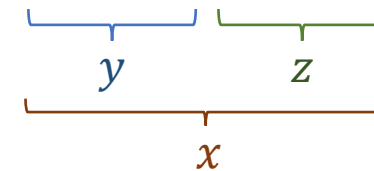
Query	Hierarchical	Non-hierarchical
SJFCQ	PTIME	$\text{FP}^{\#P}$ -complete
SJFCQ with negations	PTIME	$\text{FP}^{\#P}$ -complete
sum \ count	PTIME	$\text{FP}^{\#P}$ -complete

[L et al.
ICDT 2020]

[Reshef et al.
PODS 2020]

- A CQ q is **hierarchical** if for every two existential variables x and y :
 - ❖ $\text{Atoms}(x) \subseteq \text{Atoms}(y)$ or
 - ❖ $\text{Atoms}(y) \subseteq \text{Atoms}(x)$ or
 - ❖ $\text{Atoms}(x) \cap \text{Atoms}(y) = \emptyset$

$q_1(): -R(x, y), S(x, z)$



$q(): -\text{AUTHOR}(x, y), \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, z)$

Computational Complexity

Query	Hierarchical	Non-hierarchical
SJFCQ	PTIME	FP ^{#P} -complete
SJFCQ with negations	PTIME	FP ^{#P} -complete
sum \ count	PTIME	FP ^{#P} -complete

[L et al.
ICDT 2020]

[Reshef et al.
PODS 2020]

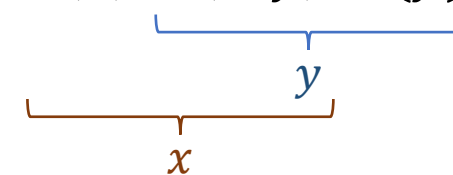
➤ A CQ q is **hierarchical** if for every two existential variables x and y :

❖ $\text{Atoms}(x) \subseteq \text{Atoms}(y)$ or

❖ $\text{Atoms}(y) \subseteq \text{Atoms}(x)$ or

❖ $\text{Atoms}(x) \cap \text{Atoms}(y) = \emptyset$

$q_2(): -R(x), S(x, y), T(y)$

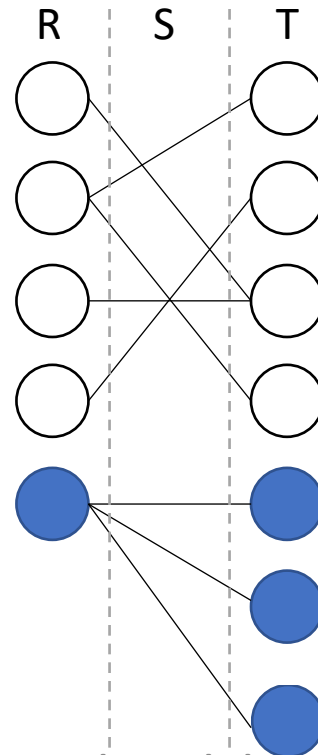


Conjunctive Queries

- To prove hardness, we consider the simplest non-hierarchical query

$$q_{RST}(): -R(x), S(x, y), T(y)$$

- Reduction from **counting independent sets in a bipartite graph**



Conjunctive Queries

- Each instance provides us with an equation over $|IS(g, k)|$
- $|IS(g, k)|$ - number of independent sets of size k in g

$$\begin{pmatrix} 0!(n+1)! & 1!n! & \dots & n!1! \\ 0!(n+2)! & 1!(n+1)! & \dots & n!2! \\ \vdots & \vdots & \vdots & \vdots \\ 0!(2n+1)! & 1!(2n)! & \dots & n!(n+1)! \end{pmatrix} \begin{pmatrix} |IS(g, 0)| \\ |IS(g, 1)| \\ \vdots \\ |IS(g, n)| \end{pmatrix} = \begin{pmatrix} (n+2)!S_1 - c_1v_0 \\ (n+3)!S_2 - c_2v_0 \\ \vdots \\ (2n+2)!S_{n+1} - c_{n+1}v_0 \end{pmatrix}$$

$q(): \neg \text{AUTHOR}(x, y), \neg \text{INSTITUTE}(y, 'CA'), \text{PUBLICATION}(x, z)$

Computational Complexity

Query	Hierarchical	Non-hierarchical
SJFCQ	PTIME	FP ^{#P} -complete
SJFCQ with negations	PTIME	FP ^{#P} -complete
sum \ count	PTIME	FP ^{#P} -complete

[L et al.
ICDT 2020]

[Reshef et al.
PODS 2020]

- A CQ q is **hierarchical** if for every two existential variables x and y :
 - ❖ $\text{Atoms}(x) \subseteq \text{Atoms}(y)$ or
 - ❖ $\text{Atoms}(y) \subseteq \text{Atoms}(x)$ or
 - ❖ $\text{Atoms}(x) \cap \text{Atoms}(y) = \emptyset$

$q(z, w): \neg \text{AUTHOR}(x, y), \text{PUBLICATION}(x, z), \text{CITATIONS}(z, w)$

$\text{SUM}_w \langle q(z, w) \rangle$

Computational Complexity

Query	Hierarchical	Non-hierarchical
SJFCQ	PTIME	FP ^{#P} -complete
SJFCQ with negations	PTIME	FP ^{#P} -complete
sum \ count	PTIME	FP ^{#P} -complete

[L et al.
ICDT 2020]

[Reshef et al.
PODS 2020]

- A CQ q is **hierarchical** if for every two existential variables x and y :
 - ❖ $\text{Atoms}(x) \subseteq \text{Atoms}(y)$ or
 - ❖ $\text{Atoms}(y) \subseteq \text{Atoms}(x)$ or
 - ❖ $\text{Atoms}(x) \cap \text{Atoms}(y) = \emptyset$

$q(z, w): \neg \text{AUTHOR}(x, y), \text{PUBLICATION}(x, z), \text{CITATIONS}(z, w)$

$\text{MAX}_w \langle q(z, w) \rangle, \text{MIN}_w \langle q(z, w) \rangle, \text{AVERAGE}_w \langle q(z, w) \rangle$

Computational Complexity

Query	Hierarchical	Non-hierarchical
SJFCQ	PTIME	FP ^{#P} -complete
SJFCQ with negations	PTIME	FP ^{#P} -complete
sum \ count	PTIME	FP ^{#P} -complete

[L et al.
ICDT 2020]

[Reshef et al.
PODS 2020]

Hardness can be extended to
general numerical queries

- A CQ q is **hierarchical** if for every two existential variables x and y :
 - ❖ $\text{Atoms}(x) \subseteq \text{Atoms}(y)$ or
 - ❖ $\text{Atoms}(y) \subseteq \text{Atoms}(x)$ or
 - ❖ $\text{Atoms}(x) \cap \text{Atoms}(y) = \emptyset$

Approximation Complexity

- Computing the Shapley value is often hard
- The picture is more positive when allowing approximation

Query	Hierarchical	Non-hierarchical
SJFCQ	PTIME	FPRAS
sum \ count	PTIME	FPRAS

$$\Pr \left[\frac{f(x)}{1 + \epsilon} \leq A(x, \epsilon, \delta) \leq (1 + \epsilon)f(x) \right] \geq 1 - \delta$$

- Generalizes to unions of CQs

Approximation Complexity

- Additive approximation via Monte Carlo sampling

$$\Pr[f(x) - \epsilon \leq A(x, \epsilon, \delta) \leq f(x) + \epsilon] \geq 1 - \delta$$

- Also a multiplicative approximation due to the “gap property”

For every tuple t in the database D :

$$\text{Shapley}(t)=0 \text{ or } \text{Shapley}(t) \geq \frac{1}{p(|D|)}$$

- Does not hold when allowing negation
- Negation **fundamentally changes** the complexity picture!

Approximation Complexity

- With negation, the contribution can be negative

Student	TA	Register	
Name	Name	Student	Course
Alice	Alice	Alice	OS
Bob	Bob	Alice	AI
Cathy	David	Bob	OS
David		Cathy	DB
		Cathy	IC

$q(): \neg\text{Student}(x), \neg\text{TA}(x), \text{Register}(x, y)$

In some cases, deciding whether $\text{Shapley}(t) \neq 0$ is hard

Banzhaf Power Index

- Causal effect [Salimi et al. 2016]
 - ❖ See the database as a probabilistic database
 - ❖ $CE(t) = E(q \mid t \in D) - E(q \mid t \notin D)$
- Coincides with the Banzhaf Power Index [Banzhaf 1965]
- Our complexity results extend to this measure

Outline

- Explaining Query Answers
- Computational Complexity
- **Responsibility to Inconsistency**

Inconsistent Databases

- A database is **inconsistent** if it violates integrity constraints



Inconsistent Databases

- Imprecise **data sources**
 - ❖ Crowd, Web pages, social encyclopedias, sensors, ...
- Imprecise **data generation**
 - ❖ natural-language processing, sensor/signal processing, image recognition, ...
- Conflicts in **data integration**
 - ❖ Crowd + enterprise data + KB + Web + ...
- Data **staleness**
 - ❖ Entities change address, status, ...
- And so on...



Idea:

Quantify the extent to which integrity constraints are violated

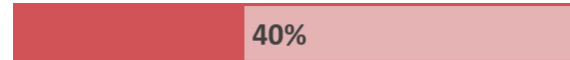
Reliability estimation

How reliable is a new data source?



Progress indication

Progress bar for data repairing

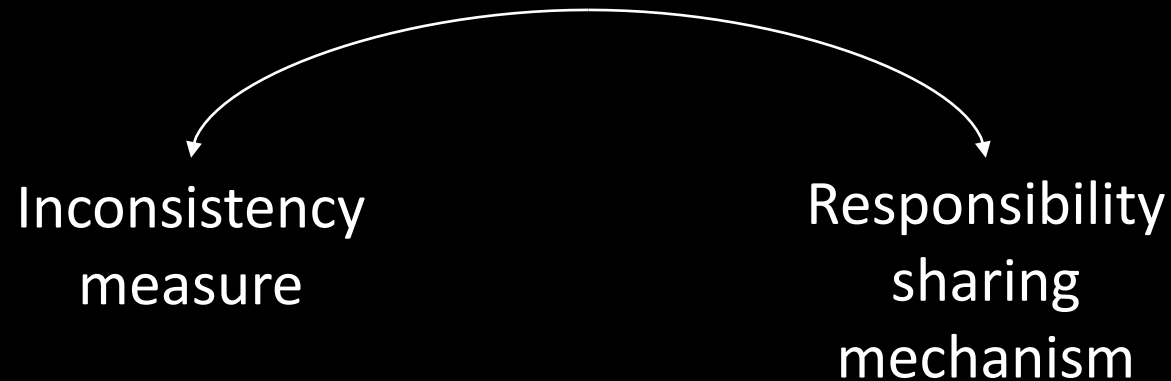


Action prioritization

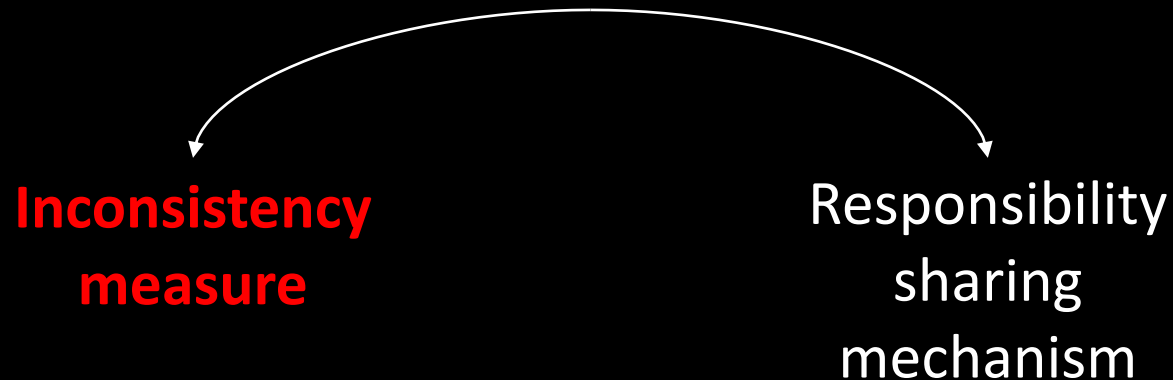
Which tuples are mostly responsible for inconsistency?



How can we quantify the responsibility of individual tuples to inconsistency?

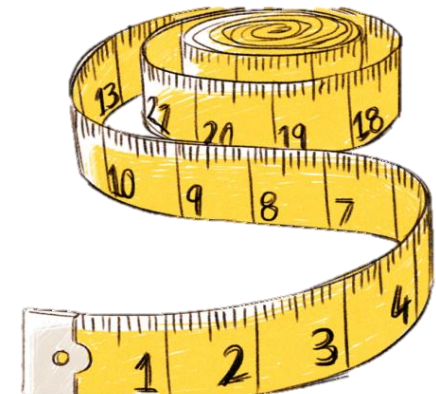


How can we quantify the responsibility of individual tuples to inconsistency?

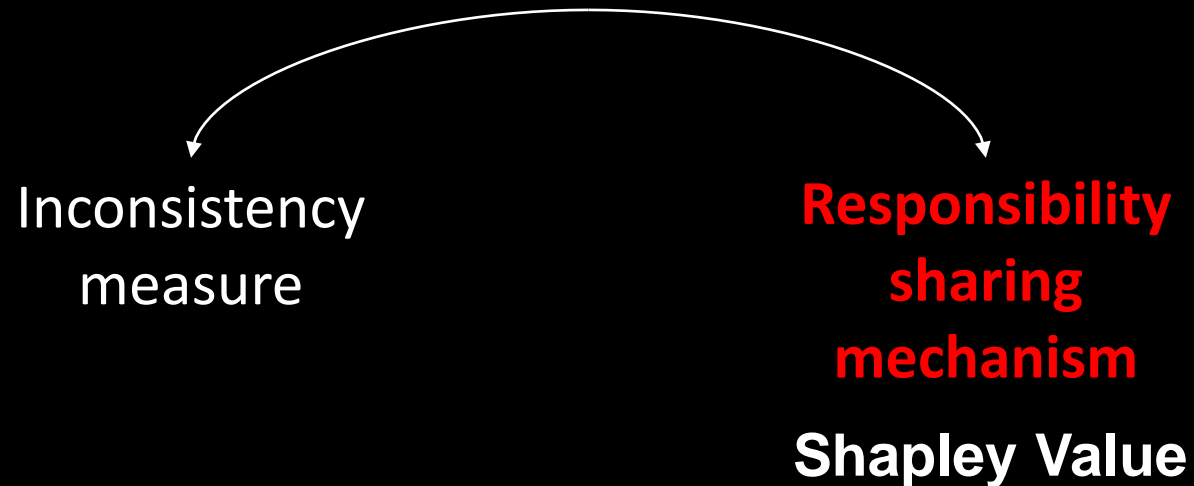


How to Measure Inconsistency?

- Several measures proposed by the KR and DB communities
 - ❖ The drastic measure – 1 if inconsistent, 0 otherwise [Thimm 2017]
 - ❖ #minimal inconsistent subsets [Hunter and Konieczny 2008]
 - ❖ #problematic tuples [Grant and Hunter 2011]
 - ❖ Minimal #tuples to remove to satisfy the constraints [Grant and Hunter 2013], [Bertossi 2018]
 - ❖ #maximal consistent subsets [Grant and Hunter 2011]
- What makes a measure a good one? [L et al. SIGMOD 2021]



How can we quantify the responsibility of individual tuples to inconsistency?



FD: birthCity → birthState

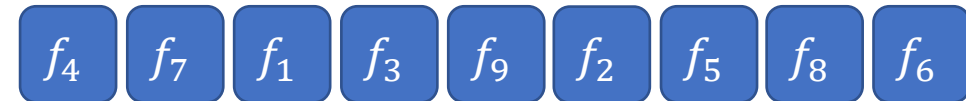
Computational Complexity

Measure	lhs chain	No lhs chain, tractable c-repair	other
drastic	PTIME	FP ^{#P} -complete	
#min- inconsistent	PTIME		
#problematic tuples	PTIME		
cardinality repair	PTIME	Open	NP-hard
#repairs	PTIME	FP ^{#P} -complete	

Tractable Measures

➤ I_{MI} - Number of minimal inconsistent subsets

	Train	Departs	Arrives	Time	Duration
f_1	16	NYP	BBY	1030	315
f_2	16	NYP	PVD	1030	250
f_3	16	PHL	WIL	1030	20
f_4	16	PHL	BAL	1030	70
f_5	16	PHL	WAS	1030	120
f_6	16	BBY	PHL	1030	260
f_7	16	BBY	NYP	1030	260
f_8	16	BBY	WAS	1030	420
f_9	16	WAS	PVD	1030	390



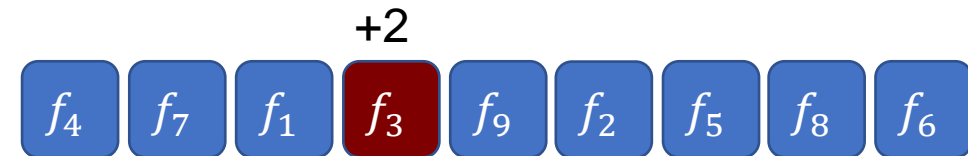
Train Time → Departs
Train Time Duration → Arrives

Tractable Measures

➤ I_{MI} - Number of minimal inconsistent subsets

	Train	Departs	Arrives	Time	Duration
f_1	16	NYP	BBY	1030	315
f_2	16	NYP	PVD	1030	250
f_3	16	PHL	WIL	1030	20
f_4	16	PHL	BAL	1030	70
f_5	16	PHL	WAS	1030	120
f_6	16	BBY	PHL	1030	260
f_7	16	BBY	NYP	1030	260
f_8	16	BBY	WAS	1030	420
f_9	16	WAS	PVD	1030	390

Train Time → Departs
 Train Time Duration → Arrives

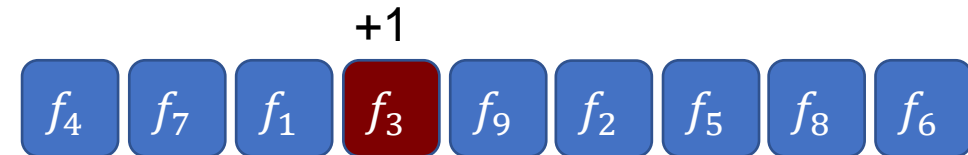


f increases the value of I_{MI} by k if k of the previous tuples conflict with it

Tractable Measures

➤ I_P - Number of problematic tuples

	Train	Departs	Arrives	Time	Duration
f_1	16	NYP	BBY	1030	315
f_2	16	NYP	PVD	1030	250
f_3	16	PHL	WIL	1030	20
f_4	16	PHL	BAL	1030	70
f_5	16	PHL	WAS	1030	120
f_6	16	BBY	PHL	1030	260
f_7	16	BBY	NYP	1030	260
f_8	16	BBY	WAS	1030	420
f_9	16	WAS	PVD	1030	390



f increases the value of I_p by k if $(k - 1)$ of the previous tuples:

- (1) conflict with f ,
- (2) do not conflict with other tuples that occur before f .

Train Time → Departs
 Train Time Duration → Arrives

Computational Complexity

Measure	lhs chain	No lhs chain, tractable c-repair	other
drastic	PTIME	FP ^{#P} -complete	
#min- inconsistent	PTIME		
#problematic tuples	PTIME		
cardinality repair	PTIME	Open	NP-hard
#repairs	PTIME	FP ^{#P} -complete	

$\{B \rightarrow A, BC \rightarrow D, BCF \rightarrow E\}$



$\{B\} \subseteq \{B, C\} \subseteq \{B, C, F\}$

$\{B \rightarrow A, BC \rightarrow D, BCG \rightarrow E, BCF \rightarrow H\}$

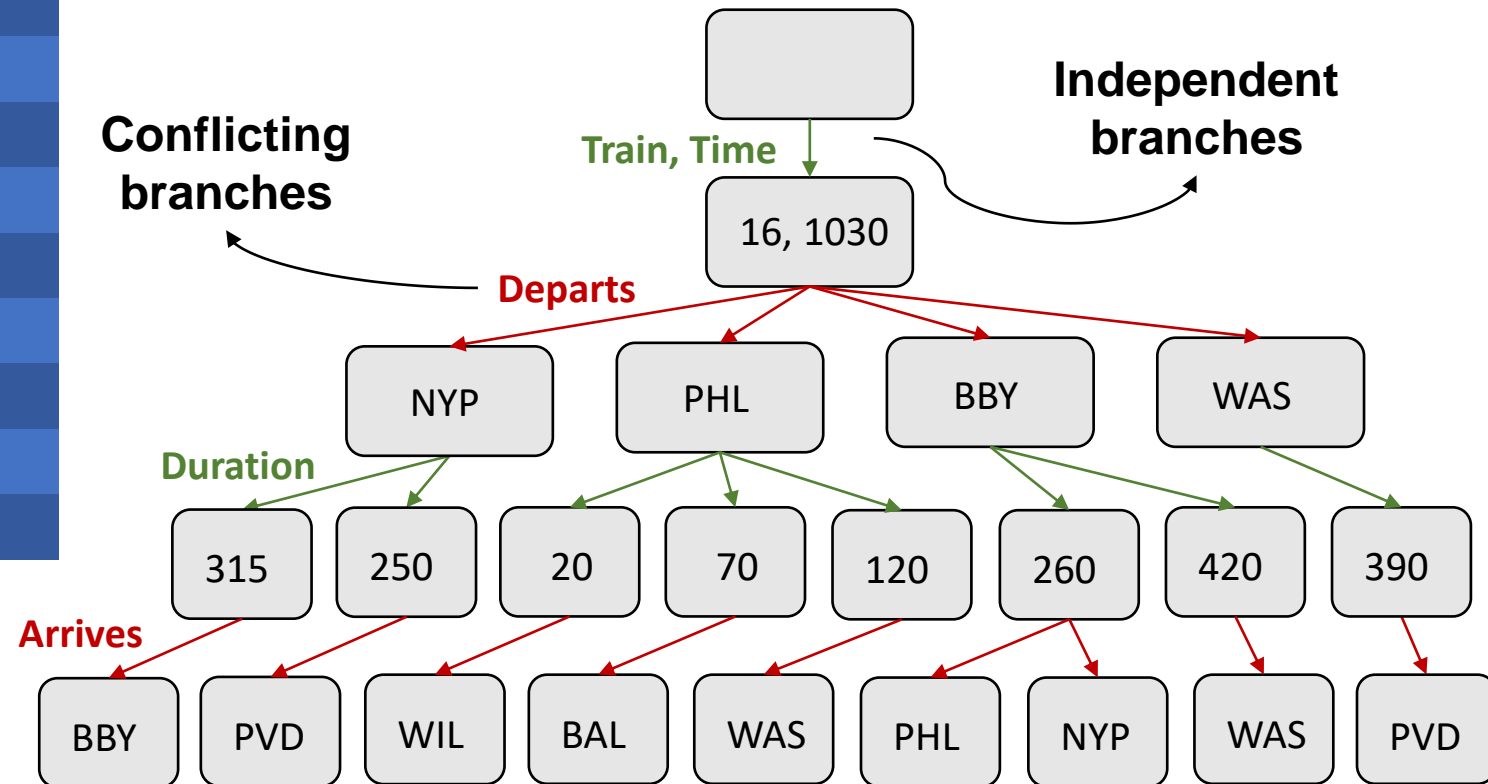


$\{B, C, G\} \not\subseteq \{B, C, F\}, \{B, C, F\} \not\subseteq \{B, C, G\}$

Left-Hand Side Chain

Train Time → Departs
 Train Time Duration → Arrives

	Train	Departs	Arrives	Time	Duration
f_1	16	NYP	BBY	1030	315
f_2	16	NYP	PVD	1030	250
f_3	16	PHL	WIL	1030	20
f_4	16	PHL	BAL	1030	70
f_5	16	PHL	WAS	1030	120
f_6	16	BBY	PHL	1030	260
f_7	16	BBY	NYP	1030	260
f_8	16	BBY	WAS	1030	420
f_9	16	WAS	PVD	1030	390



Efficiency: $\sum_{a \in A} \text{Shapley}(A, v, a) = v(A)$

Computational Complexity

Measure	lhs chain	No lhs chain, tractable c-repair	other
drastic	PTIME	FP ^{#P} -complete	
#min- inconsistent	PTIME		
#problematic tuples	PTIME		
cardinality repair	PTIME	Open	NP-hard
#repairs	PTIME	FP ^{#P} -complete	

$\{B \rightarrow A, BC \rightarrow D, BCF \rightarrow E\}$ ✓

$\{B\} \subseteq \{BC\} \subseteq \{BCF\}$

$\{B \rightarrow A, BC \rightarrow D, BCG \rightarrow E, BCF \rightarrow H\}$ ✗

$\{BCG\} \not\subseteq \{BCF\}, \{BCF\} \not\subseteq \{BCG\}$

Approximation Complexity

Measure	lhs chain	No lhs chain, tractable c-repair	other
drastic	PTIME	FPRAS	
#min- inconsistent	PTIME		
#problematic tuples	PTIME		
cardinality repair	PTIME	FPRAS	No FPRAS
#repairs	PTIME	Open	

Would imply an FPRAS for #MIS in a bipartite graph – long standing open problem

Concluding Remarks

- Two situations where we wish to quantify the responsibility of tuples:
 - ❖ Query answering
 - ❖ Database inconsistency
- We treat the contribution from the viewpoint of game theory
- We investigated the computational complexity

Thank you for listening!

Questions?

