
On Generating Near-Optimal Tableaux for Conditional Functional Dependencies

Lukasz Golab, Howard Karloff, Flip Korn,
Divesh Srivastava, Bei Yu

Key Takeaways

- Philosophy: constraints used to represent database semantics
 - Conditional constraints allow more expressive power
 - Data quality problem: a violation of database semantics
- Contributions: discovery of conditional functional dependencies
 - What is an optimal CFD?
 - How difficult is it to find an optimal CFD?
 - How practical and scalable is discovery of a good CFD?

Outline

- Motivation: examples of FDs, CFDs
- Problem statement: optimal hold and fail tableaux generation
- Complexity and efficient approximation algorithms
- Experimental evaluation

Example: Sales Table, FD

<u>Tid</u>	<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>Price</u>	<u>Tax</u>
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Example: Sales Table, CFD

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau [BFG+07]

Name	Type	Country	Price	Tax
-	Clothing	-	-	-
-	Book	France	-	0
-	-	UK	-	-

Example: Sales Table, CFD

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]
Confidence = 0.70

Hold Tableau [BFG+07]

Name	Type	Country	Price	Tax
-	Clothing	-	-	-
-	Book	France	-	0
-	-	UK	-	-

Example: Sales Table, CFD

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau [BFG+07]

Name	Type	Country	Price	Tax
-	Clothing	-	-	-
-	Book	France	-	0
-	-	UK	-	-

Example: Sales Table, Fail Tableau

<u>Tid</u>	<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>Price</u>	<u>Tax</u>
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]
Confidence = 0.70

Hold Tableau [BFG+07]

<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>Price</u>	<u>Tax</u>
-	Clothing	-	-	-
-	Book	France	-	0
-	-	UK	-	-

Fail Tableau [GKK+08]

<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>Price</u>	<u>Tax</u>
-	-	USA	-	-

Outline

- Motivation: examples of FDs, CFDs
- Problem statement: optimal hold and fail tableaux generation
- Complexity and efficient approximation algorithms
- Experimental evaluation

Problem Statement

- Given a table R and an FD, generate hold and fail tableaux
 - Philosophy: discover hidden semantics from data
- What makes for a good hold tableau?
 - High support, high confidence, parsimony
- Given a hold tableau, what makes for a good fail tableau?
 - High (residual) support, low (residual) confidence, parsimony

Example: Good Hold, Fail Tableaux

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: Support = 0.75, Confidence = 0.8

Name	Type	Country	Price	Tax
-	Clothing	-	-	-
-	Book	France	-	0
-	-	UK	-	-

Fail Tableau: Support = 0.25, Confidence = 0.4

Name	Type	Country	Price	Tax
-	-	USA	-	-

Example: Bad Hold Tableau

<u>Tid</u>	<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>Price</u>	<u>Tax</u>
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: Not parsimonious

<u>Name</u>	<u>Type</u>	<u>Country</u>		<u>Price</u>	<u>Tax</u>
HP	Book	France		-	-
TLotR	Book	France		-	-
AS	Clothing	UK		-	-
ASI	Clothing	UK		-	-
PS	Clothing	France		-	-
...

Metrics: Local Support, Confidence

- Given a CFD $\Phi = (R: X \rightarrow Y, T_p)$ and table instance $\text{adom}(R)$
 - $\text{Cover}(t_p) = \{t \mid (t \in \text{adom}(R)) \ \& \ (t[X] \text{ matches } t_p[X])\}$
- Local support of a pattern t_p in a tableau T_p
 - $\text{LS}(t_p) = |\text{Cover}(t_p)|/|\text{adom}(R)|$
- Local confidence of a pattern t_p in a tableau T_p
 - Let $\text{Keepers}(t_p)$ denote records in $\text{Cover}(t_p)$ after removing fewest records needed to eliminate all disagreements
 - $\text{LC}(t_p) = |\text{Keepers}(t_p)|/|\text{Cover}(t_p)|$

Example: Local Metrics

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau

Name	Type	Country	LS	LC
-	Clothing	-	0.35	6/7
-	-	UK	0.35	6/7

Metrics: Global Support, Confidence

- Given a CFD $\Phi = (R: X \rightarrow Y, T_p)$ and table instance $\text{adom}(R)$
- Global support of a tableau T_p
 - $\text{GS}(T_p) = |\mathbf{U} \text{Cover}(t_p)|/|\text{adom}(R)|$
 - $\text{Max}(\text{LS}(t_p)) \leq \text{GS}(T_p) \leq \sum(\text{LS}(t_p))$
- Global confidence of a tableau T_p
 - $\text{GC}(T_p) = |\mathbf{U} \text{Keepers}(t_p)|/|\mathbf{U} \text{Cover}(t_p)|$
 - Possible that $\text{GC}(T_p) \leq \min(\text{LC}(t_p))$
 - Possible that $\text{GC}(T_p) \geq \max(\text{LC}(t_p))$

Example: Global Metrics

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.5, GC = 0.8

Name	Type	Country	LS	LC
-	Clothing	-	0.35	6/7
-	-	UK	0.35	6/7

Tableau Generation Problem 1

- Given a CFD $\Phi = (R: X \rightarrow Y, T_p)$, table instance $\text{adom}(R)$ is
 - $(s,c)_{\text{gg}}$ -satisfied by Φ iff $\text{GS}(T_p) \geq s$ and $\text{GC}(T_p) \geq c$
- Tableau generation problem with GS and GC
 - Given FD $R: X \rightarrow Y$, $\text{adom}(R)$, and (s,c) find T_p of smallest size such that $\text{adom}(R)$ is $(s,c)_{\text{gg}}$ -satisfied by $(R: X \rightarrow Y, T_p)$
- Complexity of problem
 - NP-complete
 - Provably hard to approximate within $|\text{adom}(R)|^{1/2 - \epsilon}$, $\epsilon > 0$

Tableau Generation Problem 2

- Given a CFD $\Phi = (R: X \rightarrow Y, T_p)$, table instance $\text{adom}(R)$ is
 - $(s,c)_{gl}$ -satisfied by Φ iff $GS(T_p) \geq s$ and for all $t_p \in T_p$, $LC(t_p) \geq c$
- Tableau generation problem with GS and LC
 - Given FD $R: X \rightarrow Y$, $\text{adom}(R)$, and (s,c) find T_p of smallest size such that $\text{adom}(R)$ is $(s,c)_{gl}$ -satisfied by $(R: X \rightarrow Y, T_p)$
- Complexity of problem
 - NP-complete: reduction from vertex cover in tripartite graphs
 - Provably hard to approximate to within $34/33$

Problem 2: Greedy Approximation

- Problem with GS and LC can be reduced to Partial Set Cover
 - Partial Set Cover admits a greedy approximation algorithm
- Reduction
 - Generate all candidate patterns (data cube) from $\text{adom}(R)$
 - Iteratively choose pattern with highest marginal local support, satisfying $\text{LC}(t) \geq c$
 - Stop when (hold) tableau's $\text{GS} \geq s$
- Approximation guarantee, complexity of greedy approximation
 - $|T|/|T^*| \leq 1 + \ln(s \cdot |\text{adom}(R)|)$, compared to optimal T^*
 - Complexity is $O(2^K \cdot |\text{adom}(R)|)$, $K = \text{antecedents}(\text{FD})$

Example: Greedy Algorithm

<u>Tid</u>	<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>Price</u>	<u>Tax</u>
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8

<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>MLS</u>	<u>LC</u>
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.35	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	Clothing	UK	0.2	1.0

Example: Greedy Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.0

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.35	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	Clothing	UK	0.2	1.0

Example: Greedy Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.35

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.15	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	Clothing	UK	0.0	1.0

Example: Greedy Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.6

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.15	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	Clothing	UK	0.0	1.0

Example: Greedy Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.75

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.15	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	Clothing	UK	0.0	1.0

Issues with Greedy Algorithm

- Very high initial cost
 - Generate **all** candidate patterns (data cube) from $\text{adom}(R)$
 - Example: 48 candidate patterns generated, most useless
- Very high incremental cost
 - Iteratively maintain all (even unused) marginal local supports
 - Example: 3 iterations
- Good news
 - Initial and incremental costs can be substantially reduced

Problem 2: On-demand Algorithm

- Key observation: generate candidate patterns only as needed
- Algorithm
 - Start with the “all wildcards” candidate pattern in frontier
 - Iteratively visit frontier patterns in decreasing MLS order
 - If candidate pattern meets LC threshold, include in tableau else consider adding its “children” patterns to frontier
 - Important: generate a pattern only when all “parents” visited
 - Stop when (hold) tableau’s $GS \geq s$
- Result: correspondence with (off-demand) greedy algorithm
 - Same patterns chosen for tableau in same order

Example: On-demand Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.35	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	Clothing	UK	0.2	1.0

Example: On-demand Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.0

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.35	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8

Example: On-demand Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.0

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.35	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8

Example: On-demand Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.35

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.15	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8

Example: On-demand Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.35

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.15	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	DVD	USA	0.15	1/3

Example: On-demand Algorithm

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: GS = 0.75, LC = 0.8, CGS = 0.75

Name	Type	Country	MLS	LC
-	Book	-	0.35	5/7
-	Clothing	-	0.35	6/7
-	-	UK	0.15	6/7
-	-	USA	0.25	0.4
-	DVD	-	0.3	0.5
-	-	France	0.4	6/8
-	Book	France	0.25	0.8
-	Book	USA	0.1	0.5
-	DVD	USA	0.15	1/3

Binding the Consequent

- So far: tableau patterns only bound attributes in the antecedent
 - Specificity in the consequent → stronger assertions
 - But specificity in consequent does not lead to parsimony
- Our approach
 - First, generate tableau without constants in consequent
 - For each tableau pattern, assign most number of constants to consequent while ensuring that $LC \geq c$
 - Idea: compute keeper counts for all bindings of consequent

Example: Binding the Consequent

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau

Name	Type	Country	Price	Tax
-	Book	France	-	-

Price	Tax	Count(*)
all	all	5
10	all	3
25	all	2
all	0	4
all	0.05	1
10	0	2
10	0.05	1
25	0	2

Example: Binding the Consequent

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau

Name	Type	Country	Price	Tax
-	Book	France	-	0

Price	Tax	Count(*)
all	all	5
10	all	3
25	all	2
all	0	4
all	0.05	1
10	0	2
10	0.05	1
25	0	2

Fail Tableau Generation

- Given FD $R: X \rightarrow Y$, a table instance $\text{adom}(R)$, a hold tableau T_p , and (s_f, c_f) , find T_f of smallest size such that
 - (Marginal) global support of $T_f \geq s_f$
 - For all patterns $t_f \in T_f$, (marginal) local confidence $\leq c_f$
- Can reuse on-demand algorithm with two minor change
 - Recompute counts using records not covered by T_p
 - If candidate pattern below c_f threshold, include in tableau T_f else consider adding its “children” patterns to frontier

Example: Fail Tableau Generation

Tid	Name	Type	Country	Price	Tax
1	Harry Potter	Book	France	10	0
2	Harry Potter	Book	France	10	0
3	Harry Potter	Book	France	10	0.05
4	The Lord of the Rings	Book	France	25	0
5	The Lord of the Rings	Book	France	25	0
6	Algorithms	Book	USA	30	0.04
7	Algorithms	Book	USA	40	0.04
8	Armani suit	Clothing	UK	500	0.05
9	Armani suit	Clothing	UK	500	0.05
10	Armani slacks	Clothing	UK	250	0
11	Armani slacks	Clothing	UK	250	0
12	Prada shoes	Clothing	France	200	0.05
13	Prada shoes	Clothing	France	200	0.05
14	Prada shoes	Clothing	France	500	0.05
15	Spiderman	DVD	UK	19	0
16	Star Wars	DVD	UK	29	0
17	Star Wars	DVD	UK	25	0
18	Terminator	DVD	USA	25	0.08
19	Terminator	DVD	USA	25	0
20	Terminator	DVD	USA	20	0

FD: [Name, Type, Country] → [Price, Tax]

Confidence = 0.70

Hold Tableau: Support = 0.75, Confidence = 0.8

Name	Type	Country	Price	Tax
-	Clothing	-	-	-
-	Book	France	-	0
-	-	UK	-	-

Fail Tableau: Support = 0.25, Confidence = 0.4

Name	Type	Country	Price	Tax
-	-	USA	-	-

Generating Range Tableaux

- Difference between (ordinary) tableau and range tableau
 - Permit ranges $[a_l, a_r]$ in ordered attributes of patterns
 - Much larger number of candidate patterns
 - Allows for substantially more parsimonious tableaux
- Can reuse on-demand algorithm with suitable changes
 - Children of $[a_l, a_r]$ are $[a_l, a_{r-1}]$ and $[a_{l+1}, a_r]$

Example: Range Tableau

<u>Name</u>	<u>Type</u>	<u>Country</u>	<u>Date</u>		<u>Price</u>	<u>Tax</u>
-	Clothing	-	-		-	-
-	Book	France	[2008/01, 2008/04]		-	0
-	Book	France	[2008/05, 2008/08]		-	0.05
-	-	UK	-		-	-

Outline

- Motivation: examples of FDs, CFDs
- Problem statement: optimal hold and fail tableaux generation
- Complexity and efficient approximation algorithms
- Experimental evaluation

Experiments: Real Data Sets

- 300K sales records from online retailer
 - Sales(tid, itemid, name, type, price, tax, country, city)

FD1	type, name, country → price, tax, itemid
-----	--

- 30-day excerpt of network router configuration table
 - Config(date, router, interface, interface_type, IP_address)

FD2	router, interface → IP_address
FD3	router, interface, interface_type → IP_address
FD4	router, interface, date → IP_address

Experiments: Hold Tableau Sizes

- Summary of hold tableau for FD1, $l_c = 0.88$

support threshold	size	optimal size	global confidence
0.3	1	1	0.908
0.4	2	2	0.916
0.5	2	2	0.916
0.6	2	2	0.916
0.7	3	3	0.922
0.8	41	41	0.924
0.9	1690	1689	0.927

Experiments: Performance

- Comparison of running times, number of patterns considered

<u>GS</u>	<u>Time</u> <u>Off-demand</u>	<u>Time</u> <u>On-demand</u>		<u>Patterns</u> <u>Off-demand</u>	<u>Patterns</u> <u>On-demand</u>
0.5	11.5s	5.2s		610	90
0.7	11.8s	5.4s		610	92
0.8	12.0s	5.9s		610	150
0.85	12.2s	6.0s		610	155
0.9	12.5s	6.5s		610	190

Experiments: Range Tableau Size

- Reducing tableau size for FD4 with attribute ranges

<u>Support Threshold</u>		<u>Tableau Size</u>	<u>Tableau size with ranges</u>
0.5		23	1
0.6		76	1
0.7		328	2
0.8		2634	4
0.9		N/A	320

Experiments: Summary

- Greedy algorithms return near-optimal tableaux
 - Far smaller than upper bound of approximation guarantee
- On-demand algorithm much faster than off-demand algorithm
 - Difference increases with number of candidate patterns
- Range tableaux much smaller than standard tableaux
 - If embedded FD holds for a range of antecedent values

Related Work

- Conditional functional dependencies
 - [BFG+07]: CFD definition, Armstrong axioms extension
 - [CFG+07]: violation detection, table repair to satisfy CFDs
 - [BFM07]: extended CFDs, inspiration for range tableaux
- Approximate FD discovery [HKPT99, KL03]
 - Useful to identify suitable embedded FDs

Summary

- Data quality a serious issue in today's complex databases
 - CFDs help by capturing undocumented semantics
- Contributions: generating good tableaux for CFDs
 - Optimal tableau definition: support, confidence, parsimony
 - Complexity of optimal tableau discovery: NP-hard
 - Efficient approximation algorithms: greedy, on-demand
 - Range tableau: very compact tableau for ordered attributes
 - Experiments: practical and scalable tableau discovery