Rekonstukce databázového modelu na základě nepřesných dat

Martin Řimnáč

ITAT 2004

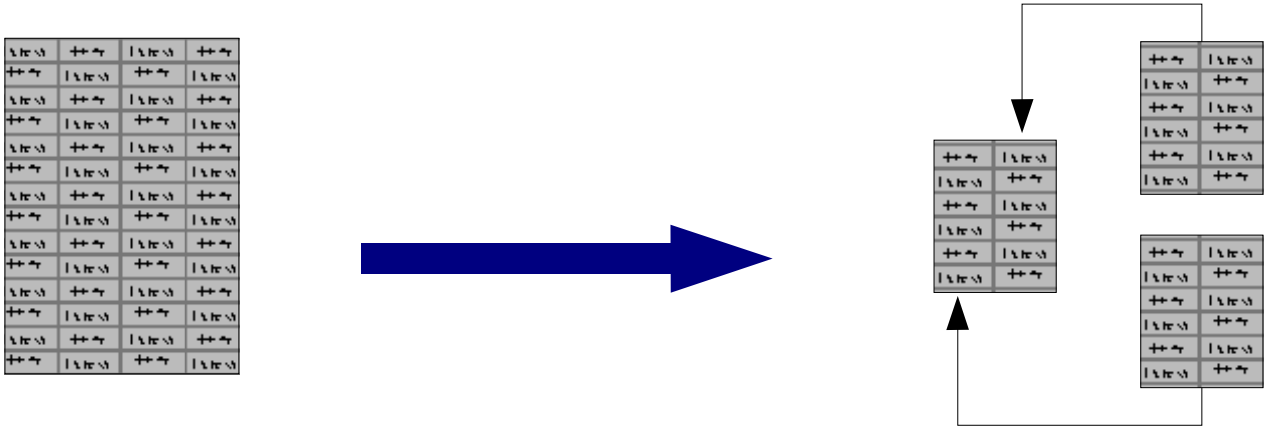
Ústav informatiky
Akademie věd ČR

Katedra kybernetiky
České vysoké učení technické

Dekompozice modelu

Nalezení takového databázového modelu, který odpovídá

- Požadované normální formě
- Vstupním datům (složená relace)

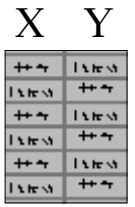


Metoda zahrnuje:

- Detekci funkčních závislostí
- Odstranění redundantních závislostí
- Generování databázového modelu

Funkční závislosti

Atribut Y je funkčně závislý na atributu X (X->Y), pokud vzoru (X) odpovídá právě jeden obraz (Y).



Vlastnosti:

- Transitivita — X->Y a Y->Z, pak X->Z
- Hierarchie — X->Y, pak i {X,Z}->Y

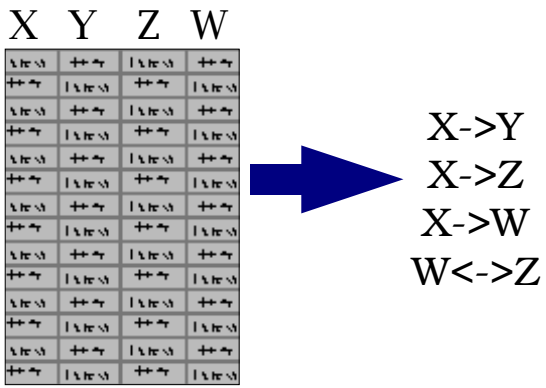
Značení:

- Není funkčně závislý — X-/>Y
- Vzájemně závislý — X<->Y — X->Y a Y->X

Detekce funkční závislosti

Detekce funkční závislosti X->Y:

- Seřazení záznamů podle X, sekundárně Y
- Testování všech záznamů podle definice



Složitost:

- Rekonstukce tabulky — *o(1)* až *o(NC)*
- Seřazení záznamů — *o(N log(N))*
- Testování rovnosti — *o(N)*

X	Y
1	1
2	1
2	1
3	2
4	2

X->Y

X	Y
1	1
2	1
2	1
2	2
3	2

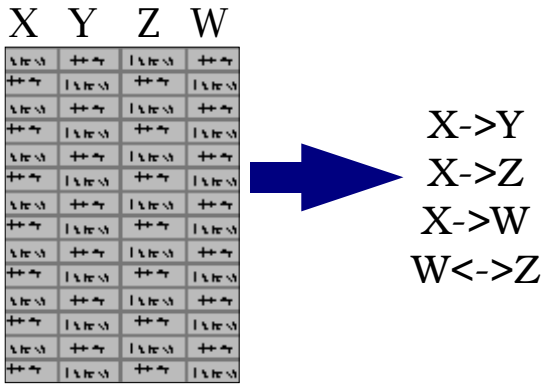
X-/>Y

Jednoatributové funkční závislosti: *C(C-1)* testů.

Detekce funkční závislosti

Detekce funkční závislosti X->Y:

- Seřazení záznamů podle X, sekundárně Y
- Testování všech záznamů podle definice



Složitost:

- Rekonstukce tabulky — *o(1)* až *o(NC)*
- Seřazení záznamů — *o(N log(N))*
- Testování rovnosti — *o(N)*

X	Y
1	1
2	1
2	1
3	2
4	2

X->Y

X	Y
1	1
2	1
2	1
2	2
3	2

X-/>Y

Jediný chybný záznam rozhodne o neplatnosti funkční závislosti

Jednoatributové funkční závislosti: *C(C-1)* testů.

Matice závislostí

Popis funkčních závislostí

- Pomocí matice závislostí M
- Omezení pouze na jednoatributové závislosti

Matice závislostí

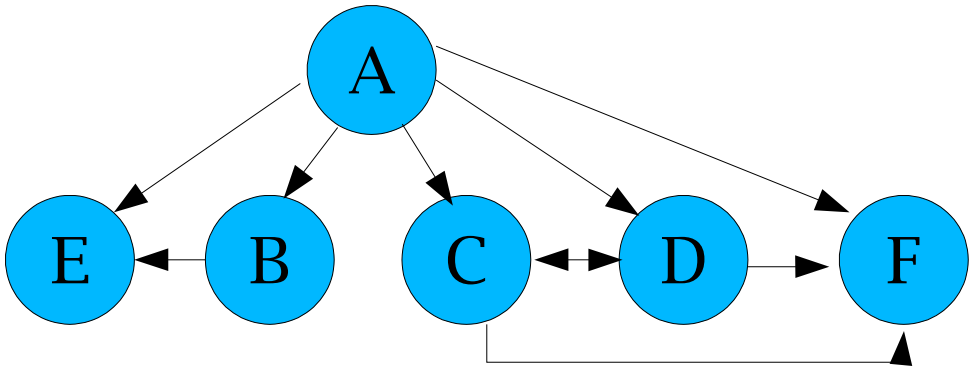
- $M_{ij} = +1$ pokud $A_i \rightarrow A_j$ a $A_j \not\rightarrow A_i$
- $M_{ij} = -1$ pokud $A_i \not\rightarrow A_j$ a $A_j \rightarrow A_i$
- $M_{ij} = 0$ jinak

Ilustrační příklad

A	B	E	C	D	F
Výrobek	Podkategorie	Kategorie	Cena	Cena+DPH	Cena-Akce
…	…	…	…	…	…
…	…	…	…	…	…

Detekované funkční závislosti:

- A->B, A->E, A->C, A->D, A->F, B->E, C<->D, C->F



	A	B	E	C	D	F
A	0	1	1	1	1	1
B	-1	0	1	0	0	0
E	-1	-1	0	0	0	0
C	-1	0	0	0	0	1
D	-1	0	0	0	0	1
F	-1	0	0	-1	-1	0

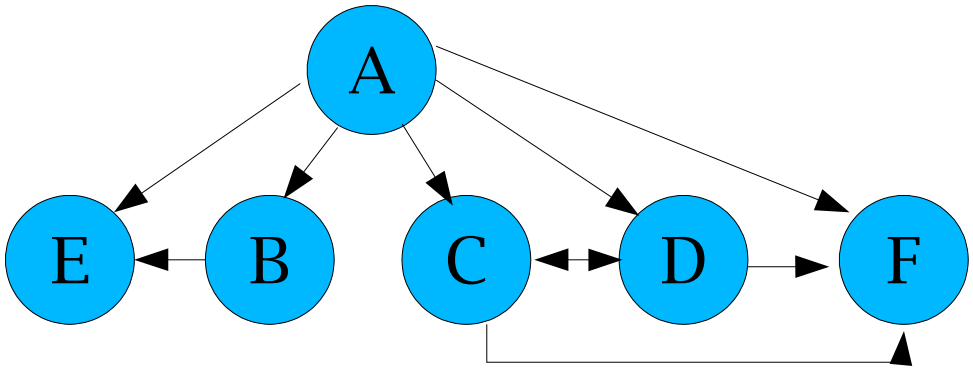
Uspořádání na matici závislostí

Definujme uspořádání atributů na základě tak, že

$\Sigma_k m_{ik} > \Sigma_k m_{jk}$, pak $i<j$

Vlastnost (díky tranzitivitě):

Pokud A_i -> A_j , pak $i<j$.



	A	B	E	C	D	F
A	0	1	1	1	1	1
B	-1	0	1	0	0	0
E	-1	-1	0	0	0	0
C	-1	0	0	0	0	1
D	-1	0	0	0	0	1
F	-1	0	0	-1	-1	0
Σ	-5	0	2	1	1	3

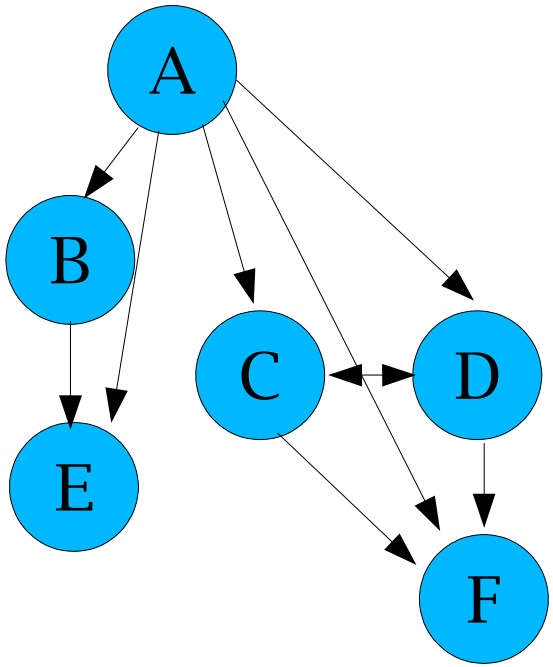
Uspořádání na matici závislostí

Definujme uspořádání atributů na základě tak, že

$\Sigma_k m_{ik} > \Sigma_k m_{jk}$, pak $i<j$

Vlastnost (díky tranzitivitě):

Pokud A_i -> A_j , pak $i<j$.



	A	B	C	D	E	F
A	0	1	1	1	1	1
B	-1	0	0	0	1	0
C	-1	0	0	0	0	1
D	-1	0	0	0	0	1
E	-1	-1	0	0	0	0
F	-1	0	-1	-1	0	0
Σ	-5	0	1	1	2	3

Vzájemné závislostí

Vzájemná závislost atributů A_i <-> A_j

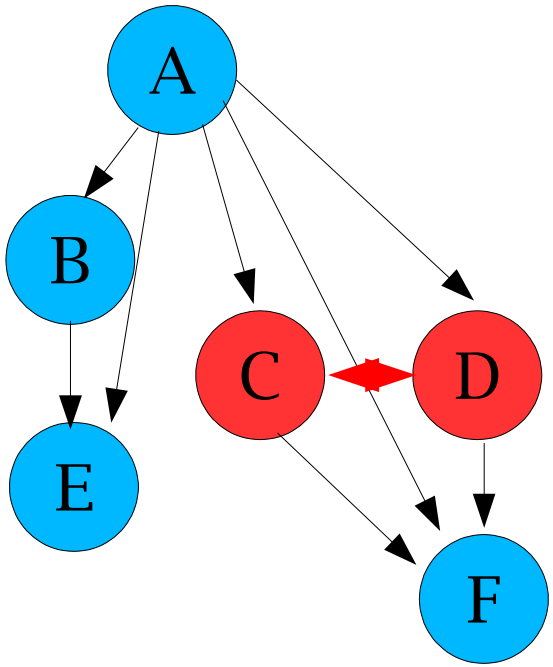
A_i -> A_j a zároveň A_j ->A_i

Platí:

$$\Sigma_k m_{ik} = \Sigma_k m_{jk}$$

$$m_{ik} = m_{jk}$$

$$m_{ki} = m_{kj}$$



	A	B	C	D	E	F
A	0	1	1	1	1	1
B	-1	0	0	0	1	0
C	-1	0	0	0	0	1
D	-1	0	0	0	0	1
E	-1	-1	0	0	0	0
F	-1	0	-1	-1	0	0

Vzájemné závislosti

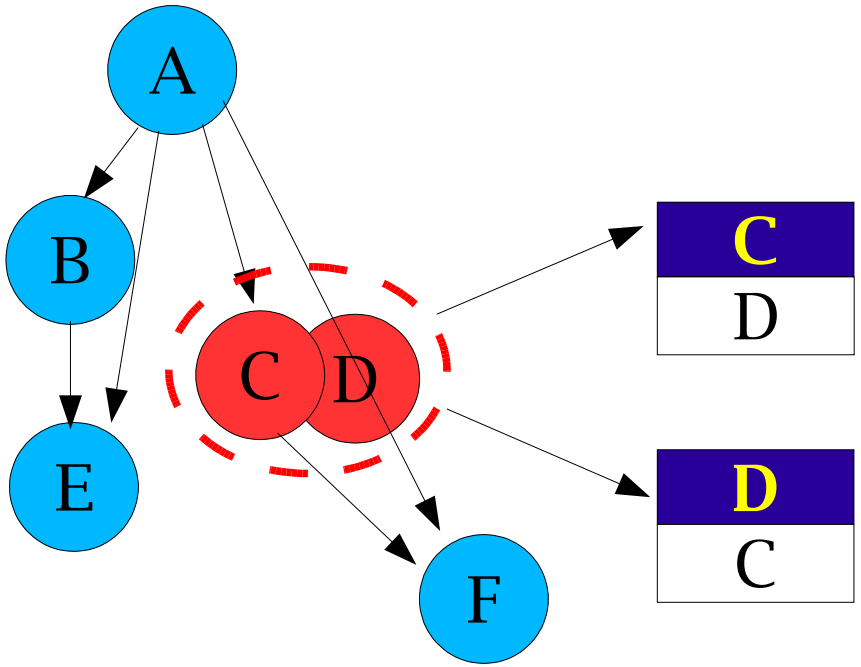
Vzájemná závislost atributů A_i <-> A_j

Vytvoření vzájemné relace

Neurčitost primárního klíče – výběr libovolného atributu

Redukce matice (omezení stavového prostoru)

Ve schématu zůstane pouze atribut představující primární klíč



	A	B	C D	E	F
A	0	1	1	1	1
B	-1	0	0	1	0
C D	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Vzájemné závislosti

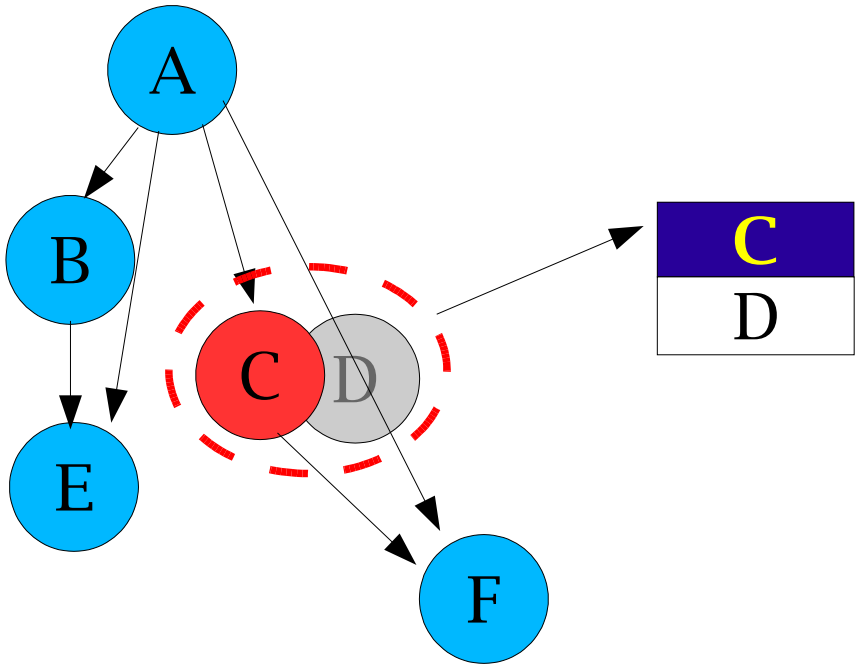
Vzájemná závislost atributů A_i <-> A_j

Vytvoření vzájemné relace

Neurčitost primárního klíče – výběr libovolného atributu

Redukce matice (omezení stavového prostoru)

Ve schématu zůstane pouze atribut představující primární klíč



	A	B	C	E	F
A	0	1	1	1	1
B	-1	0	0	1	0
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Antisymetrie matice

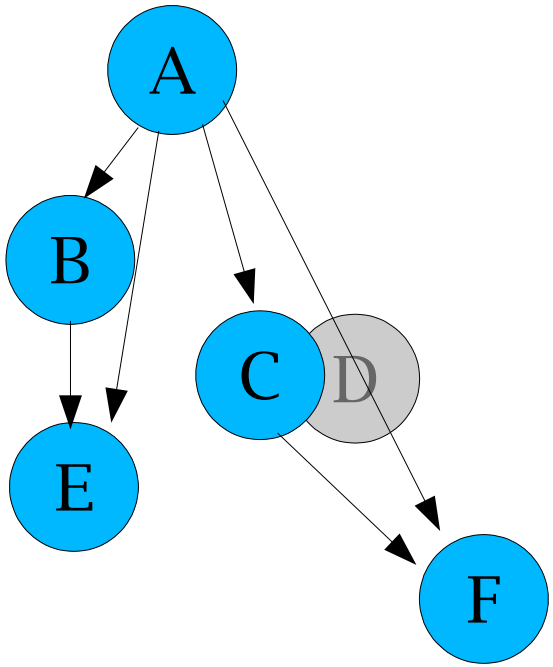
Matice obsahuje pouze jednosměrné funkční závislosti

$A_i \rightarrow A_j$ a $A_j \not\rightarrow A_i$

Antisymetrická matice $m_{ij} = -m_{ji}$

Kladné prvky nad diagonálou, nulová diagonála

Redukce

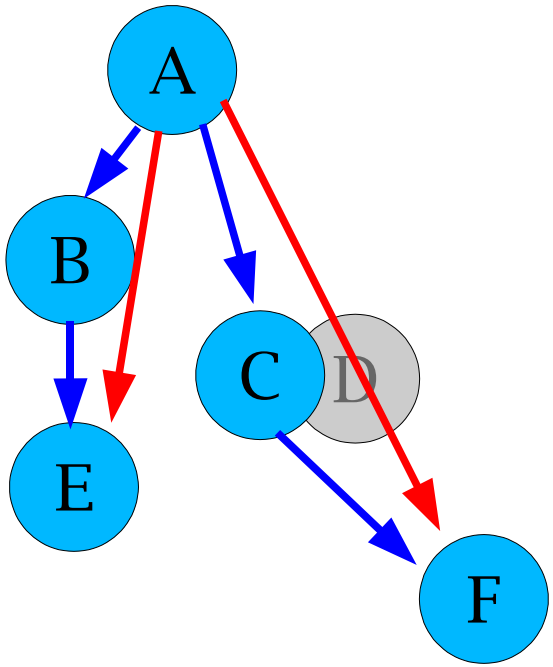


C
D

	A	B	C	E	F
A	0	1	1	1	1
B	-1	0	0	1	0
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Tranzitivita

Redundance díky tranzitivitě
X->Y a Y->Z, pak X->Z



C
D

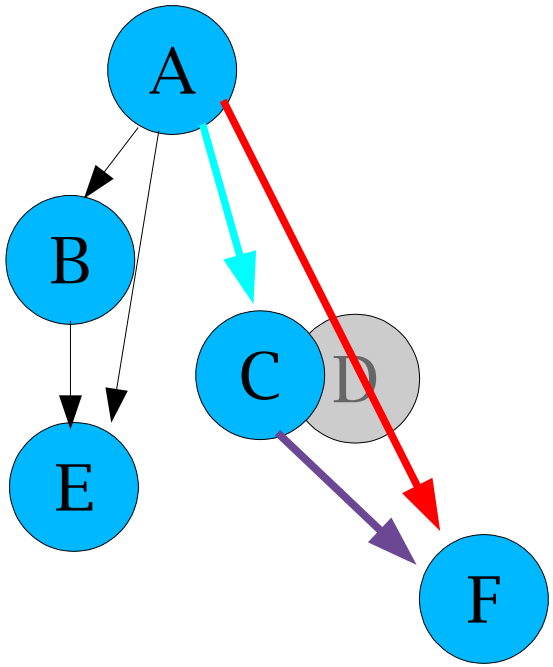
	A	B	C	E	F
A	0	1	1	1	1
B	-1	0	0	1	0
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Odstranění redundance

Pokud m_{ij} = +1,

pak v *i*-tém řádku nuluj všechny +1,

pokud ve stejném sloupci je v *j*-tém řádku +1.



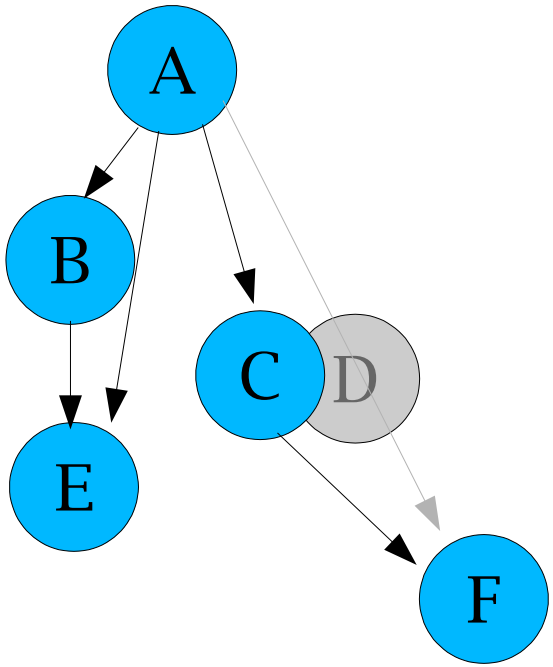
C
D

	A	B	C	E	F
A	0	1	1	1	1
B	-1	0	0	1	1
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Odstranění redundance

Pokud m_{ij} = +1,

pak v *i*-tém řádku nuluj všechny +1,
pokud ve stejném sloupci je v *j*-tém řádku +1.



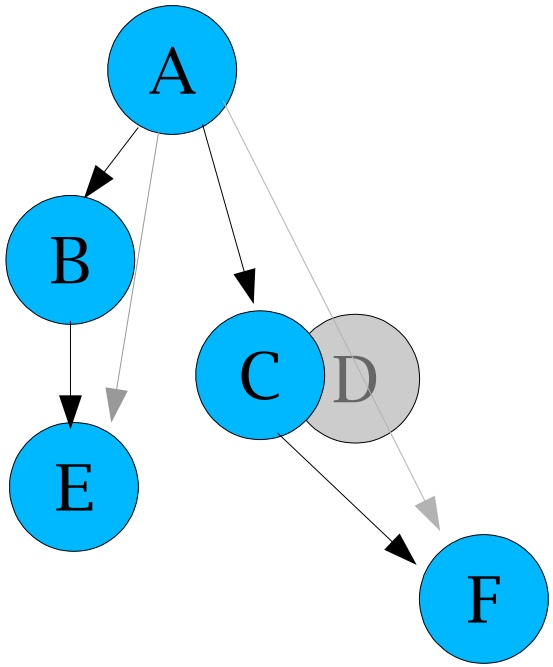
C
D

	A	B	C	E	F
A	0	1	1	1	0
B	-1	0	0	1	0
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Odstranění redundance

Pokud m_{ij} = +1,

pak v *i*-tém řádku nuluj všechny +1,
pokud ve stejném sloupci je v *j*-tém řádku +1.



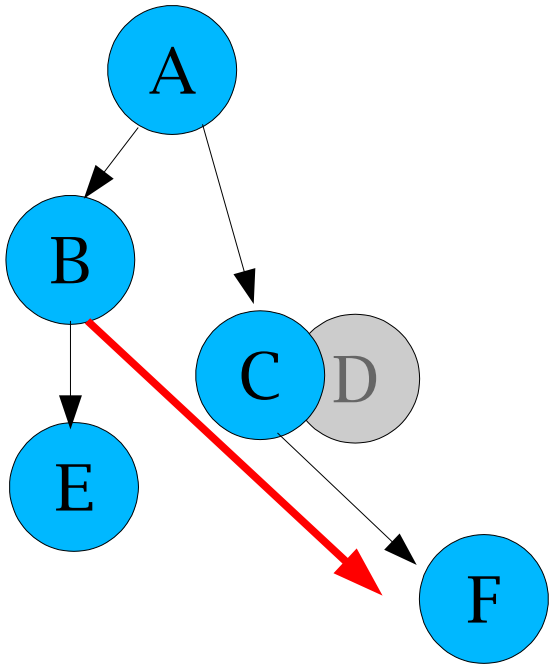
C
D

	A	B	C	E	F
A	0	1	1	0	0
B	-1	0	0	1	0
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Odstranění redundance

Můžeme považovat za redundantní všechny závislosti,
mezi nimiž a diagonálou je alespoň jedna +1.

Pokud do uzlu vstupuje více hran,
může se jednat o speciální případ vícehodnotové závislosti.

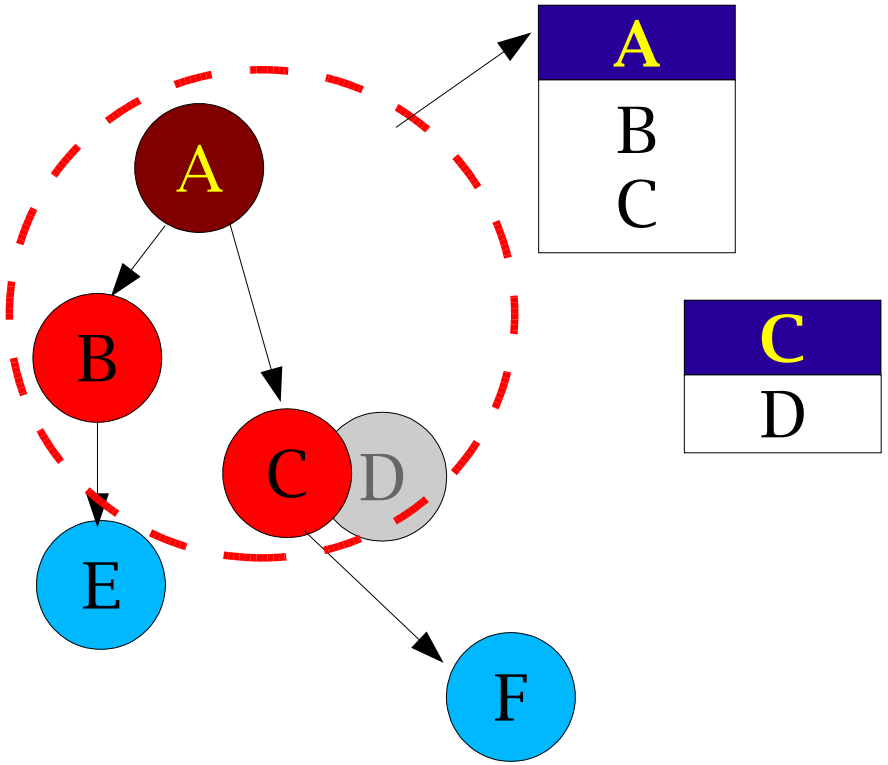


C
D

	A	B	C	E	F
A	0	1	1	0	0
B	-1	0	0	1	1
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Rekontrukce relací

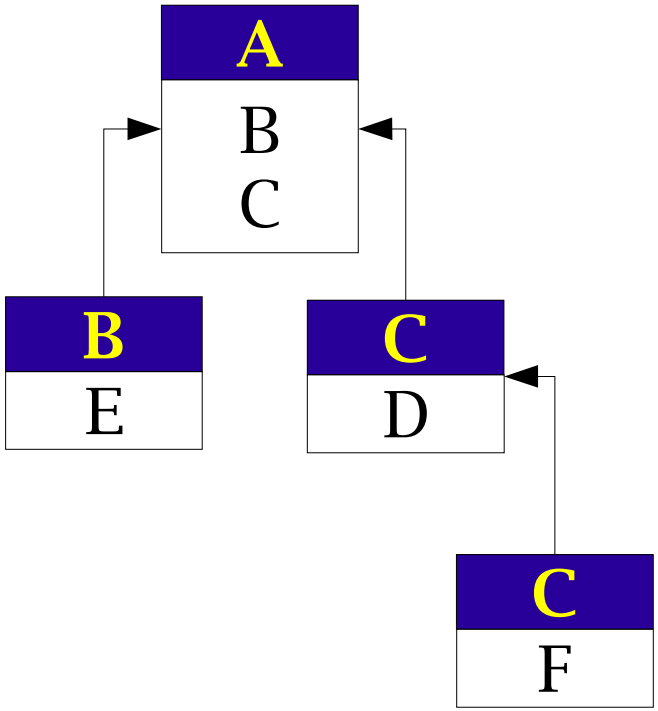
Každý řádek představuje relaci.
Atribut řádku je primárním klíčem relace
Relaci dále tvoří všechny atributy s +1.



	A	B	C	E	F
A	0	1	1	0	0
B	-1	0	0	1	0
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Rekontrukce relací

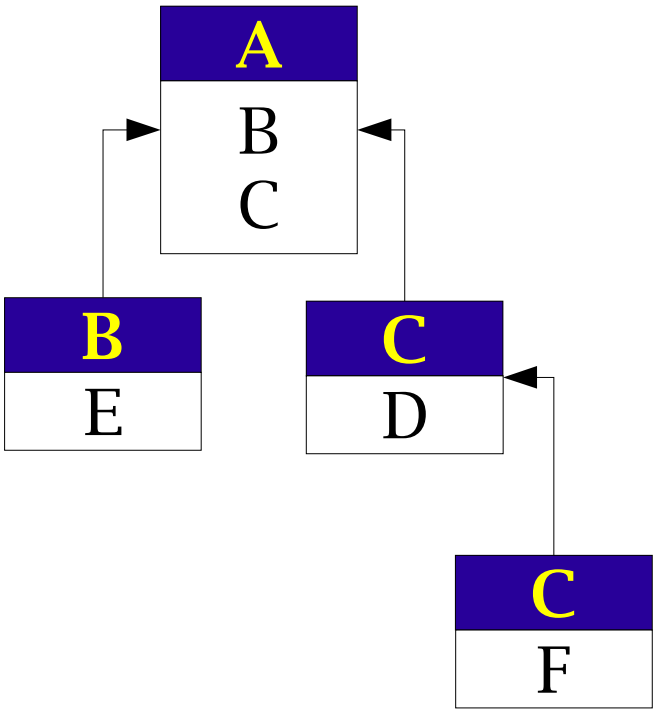
Každý řádek představuje relaci.
Atribut řádku je primárním klíčem relace
Relaci dále tvoří všechny atributy s +1.

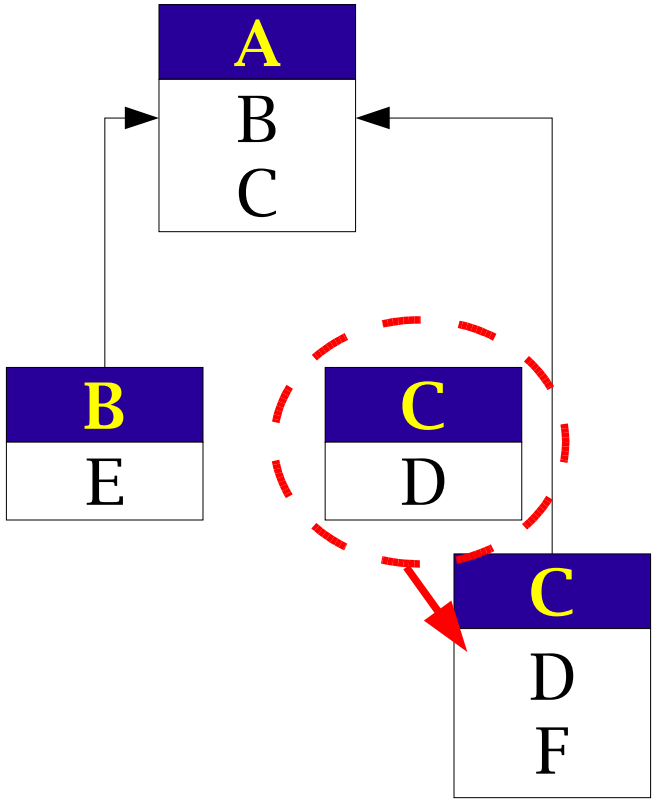


	A	B	C	E	F
A	0	1	1	0	0
B	-1	0	0	1	0
C	-1	0	0	0	1
E	-1	-1	0	0	0
F	-1	0	-1	0	0

Rekontrukce relací

Relaci vzniklou ze vzájemných relací je možné zakomponovat do relace o úroveň výše nebo níže.





Oba modely splňují kritéria 3. normální formy.

Motivace pro fuzzy

Pokud rekonstruujeme nekompletní data, klasická definice funkční závislosti je příliš restriktivní.

X	Y
1	1
2	1
2	1
2	2
3	2

Jediný chybný záznam rozhodne o neplatnosti funkční závislosti

Klasický přístup
X-/>Y

4/5 záznamů funkční závislosti vyhovují

Fuzzy přístup
X->Y (@80%)

Matice fuzzy závislostí

$M_{ij} = c_{ij} / C - c_{ji} / C$

c_{ij} počet záznamů splňujících $A_i \rightarrow A_j$

c_{ji} počet záznamů splňujících $A_i \rightarrow A_j$

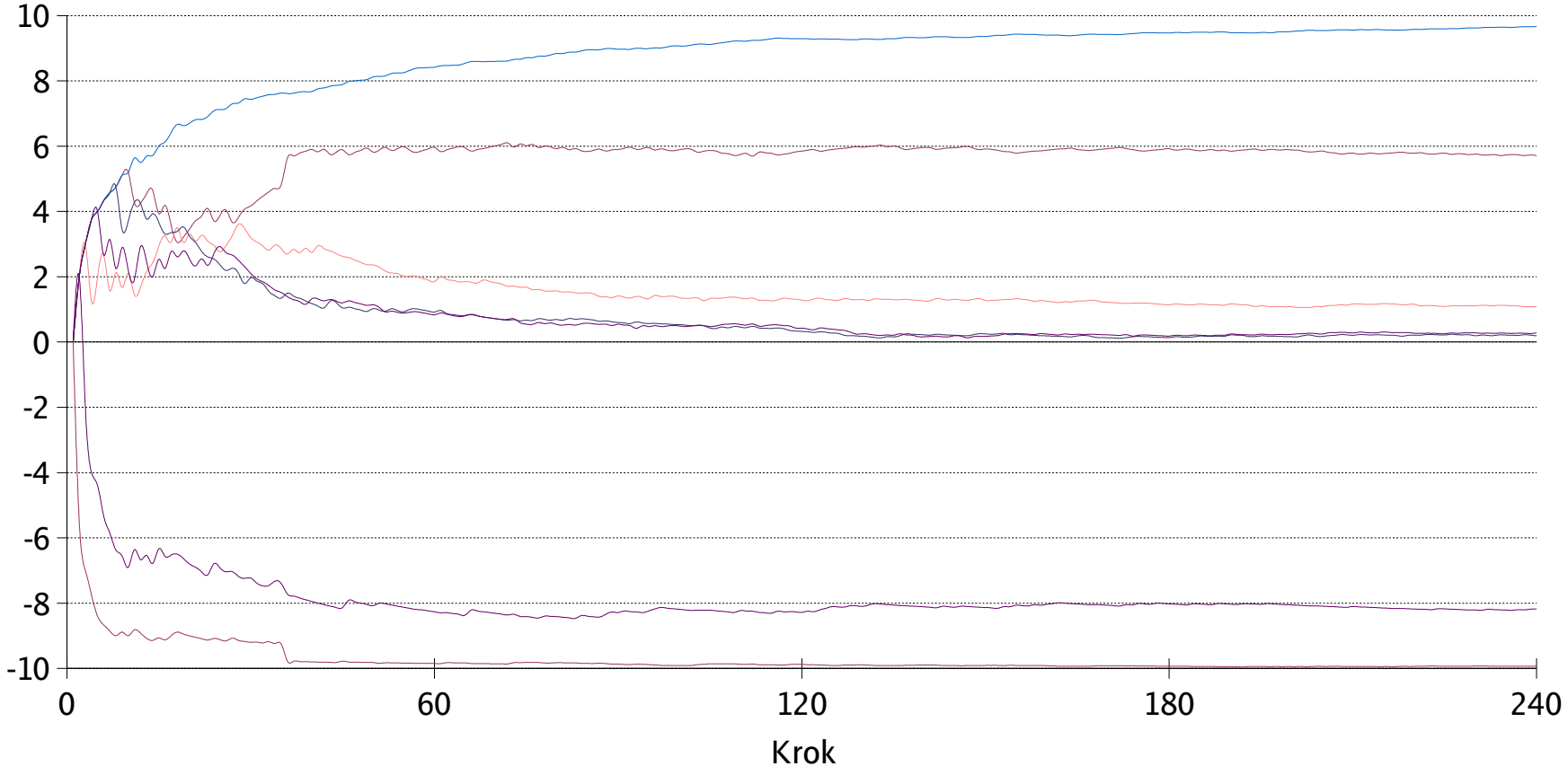
C celkový počet záznamů

Uspořádání atributů pro fuzzy

Kritérium pro uspořádání shodné s nefuzzy verzí:

$$\sum_j m_{ij}$$

Stabilita součtu fuzzy závislostí



Shrnutí

Nefuzzy závislosti

+ Implementovaná metoda s ohledem na možnost fuzzyfikace
+ Extenzivní přístup k datům
+ Ohled na výpočetní složitost, redukce stavového prostoru
- Nevhodné pro nekompletní data
- Výsledek: 1 model (náhodně vybraný z více)

Fuzzy varianta

+ Intenzivní přístup k datům
+ Vychází z nefuzzy metody
+ Možnost odhadu extenzivního řešení - stabilita
+ Výsledek: Ohodnocená množina modelů
(vhodné kvůli nejednoznačnosti)