NORMAL FORMS

CS-A1153 - Databases (Summer 2020)

LUKAS AHRENBERG

Aalto University
School of Science

NORMAL FORMS ?

- Structuring a database
  - Removing redundancy
  - Avoiding anomalies
- Boyce-Codd Normal Form (BCNF)
- Fourth Normal Form

ANOMALIES

Bad design can lead to unintended behaviour when using the database. Such problems are called *anomalies*.

Redundancy

Repeated information over several tuples in a table (not over several tables)

Update Anomaly

Sensitivity to mistake in updating repeated information

Deletion Anomaly

If one part of a tuple needs to be deleted, information might be lost

ANOMALIES EXAMPLE

`M(title,year,length,genre,studioName,starName)`

(Information about movie and star in the same relation.)

title	year	length	genre	studioName	starName
Star Wars	1977	124	SciFi	Fox	Harrison Ford
Star Wars	1977	124	SciFi	Fox	Carrie Fisher
Star Wars	1977	124	SciFi	Fox	Mark Hamill
Waynes World	1992	95	comedy	Paramount	Mike Myers
Notting Hill	1999	124	comedy	MCA-Universal	Julia Roberts

GOALS ON THE ROAD TO NORMAL FORMS

- Functional Dependencies
- Keys
- Closure of attributes
- Decomposition of relations to
  - BCNF
- Multivalued Dependencies
- Fourth Normal Form

[U & W: 3:1 - 3:4, 3:6]

FUNCTIONAL DEPENDENCY

- A *functional dependency* (FD) indicates a dependency in a relation's attributes.
- Saying that if two tuples have the same values for some specific attributes, then they will also have the same values for some other ones
- Denoted **determinants** $\longrightarrow$ **dependants**
- In essence the **determinants** 'locks-in' the values of the **dependants** of an FD

(If you could write a function in your favourite programming language which took the left hand side as parameters and returned a unique right-hand side (by looking it up in the table), then you have a functional relation.)

FD - EXAMPLE 1

`M(title,year,length,genre,studioName,starName)`

- $title\ year \rightarrow length\ genre\ studioName$ is a functional dependency of `M`
  - Because there's only one movie with the same title every year, and $\{length, genre, studioName\}$ are in a sense properties of the movie
- On the other hand, $title\ year \rightarrow starName$ **does not hold**
  - Because there's more than one star in a movie!

title	year	length	genre	studioName	starName
Star Wars	1977	124	SciFi	Fox	Harrison Ford
Star Wars	1977	124	SciFi	Fox	Carrie Fisher
Star Wars	1977	124	SciFi	Fox	Mark Hamill
Waynes World	1992	95	comedy	Paramount	Mike Myers
Notting Hill	1999	124	comedy	MCA-Universal	Julia Roberts

AN FD IS NOT ABOUT THE DATA IN A TABLE, BUT ABOUT WHAT DATA *COULD* BE PRESENT

- Keep this in mind when tables are shown as examples - these are *samples*
- So, often a bit more abstract notation is used:
- There is 'some' relation `R`,
    - with attributes $A, B, C \ldots$, or
    - with attributes $A_1, A_2, \ldots, A_n$, or
    - with attributes $A_1, A_2, \ldots, A_n, B_1, \ldots, B_m, C_1, \ldots$
        - The latter version uses both different letters and indices to highlight different 'groupings' among the attributes

FD - EXAMPLE

- Given $\mathtt{R}$ with attribute set $\{A, B, C, D, E\}$
- Suppose there is a FD $A\,B \rightarrow D$, then
  - $A\,B$ are **determinants**
  - D is **dependant**
  - We know that if two rows agree on the values for A and B they **will** also have the same value for D
    - That is, for the two tuples $(a, b, c_1, d_1, e_1)\,, (a, b, c_2, d_2, e_2)$
      - $d_1 = d_2$
    - **Note:** The tuple *may still differ in the other columns* representing attributes $C\,E$, because these are not part of the FD

KEYS

- A functional dependency is a generalization of keys
- A set of one or more attributes of a relation is called a **key** of the relation
    1. They functionally determines *all* other attributes of the same relation
    2. Any attributes are removed from the key, 1. no longer holds. (I.e it is *minimal*.)
- A relation can have more than one key
    - In this case one of them is selected as **primary key**
- A **superkey** is a set of attributes containing a key, but *may also contain other attributes*
    - 'super' comes from *super set* (it doesn't mean that a they superkey is 'better', only larger)
    - A key is also a super key (it doesn't have to be a proper super set)

KEYS - EXAMPLE

`M(title,year,length,genre,studioName,starName)`

- Claim: $\{title, year, starName\}$ is a key for `M`
  1. Holds
     - No two productions have the same title the same year
     - $length$, $genre$, and $studioName$ are all determined by the film
     - $starName$ may vary, but this is part of the key
  2. Holds
     - Can not remove $title$ - Many movies the same year with the same star
     - Can not remove $year$ - Remakes with the same stars in different years
     - Can not remove $starName$ - Most movies have more than one actor

SPLITTING AND COMBINING FD'S

We may **split the right hand side** of any FD with more than two dependants

- For example $A \rightarrow B\,C$ is split to
  - $A \rightarrow B$
  - $A \rightarrow C$

In general: $A_1\ A_2\ \ldots A_n \rightarrow B_1\ B_2\ \ldots B_m$ splits as

$$
\begin{gathered}
A_1\ A_2 \ldots A_n \rightarrow B_1 \\
A_1\ A_2 \ldots A_n \rightarrow B_2 \\
\vdots \\
A_1\ A_2 \ldots A_n \rightarrow B_m
\end{gathered}
$$

Reversible: *Singleton right hand side with the same left hand side may also be* ***combined*** *to a single expression.*

Aalto University
School of Science

TRIVIAL DEPENDENCIES

- If *all* right hand attributes (dependant set) are contained on among those on the left hand side (determinant set), the dependency is said to be **trivial**
- If *none* of the attributes on the right occurs on the left, the dependency is said to be **completely nontrivial**
- Otherwise it is just **nontrivial**
  - The right hand side can be simplified (made completely nontrivial) by *removing from the right attributes also occurring on the left*

ARMSTRONG'S AXIOMS

1. **Reflexivity**
   If $\{B_1, B_2, \ldots, B_m\} \subseteq \{A_1, A_2, \ldots, A_n\}$ then
   $A_1\ A_2 \ldots A_n \rightarrow B_1\ B_2 \ldots B_m$
2. **Augmentation**
   If $A_1\ A_2 \ldots A_n \rightarrow B_1\ B_2 \ldots B_m$ then
   $A_1\ A_2 \ldots A_n\ C_1\ C_2 \ldots C_k \rightarrow B_1\ B_2 \ldots B_m\ C_1\ C_2 \ldots C_k$ for some set of attributes $\{C_1, C_2, \ldots, C_k\}$ in the relation
3. **Transitivity**
   If $A_1\ A_2 \ldots A_n \rightarrow B_1\ B_2 \ldots B_m$ and $B_1\ B_2 \ldots B_m \rightarrow C_1\ C_2 \ldots C_k$
   then $A_1\ A_2 \ldots A_n \rightarrow C_1\ C_2 \ldots C_k$

THE CLOSURE OF ATTRIBUTES

- Taking one or more attributes in a relation, together with a set of FD's: which is the biggest possible set of attributes which can be affected?
  - This is called the **closure** of the original attribute(s)
- For a set of attributes $\mathcal{A}$ this is denoted $\mathcal{A}^+$
- Important concept
  - E.g: For `R(A,B,C,D)`, assume that $\{A, B\}^+$ (the closure of $\{A, B\}$) is $\{A, B, C, D\}$ under some FD's, then $\{A, B\}$ is a superkey of `R`

Aalto University
School of Science

THE CLOSURE ALGORITHM

- Starting with some set of attributes $\mathcal{A}$, and some functional dependencies $\mathcal{S}$, we can construct its **closure** $\mathcal{A}^+$ by 'growing' $\mathcal{A}$ as far as possible using $\mathcal{S}$:

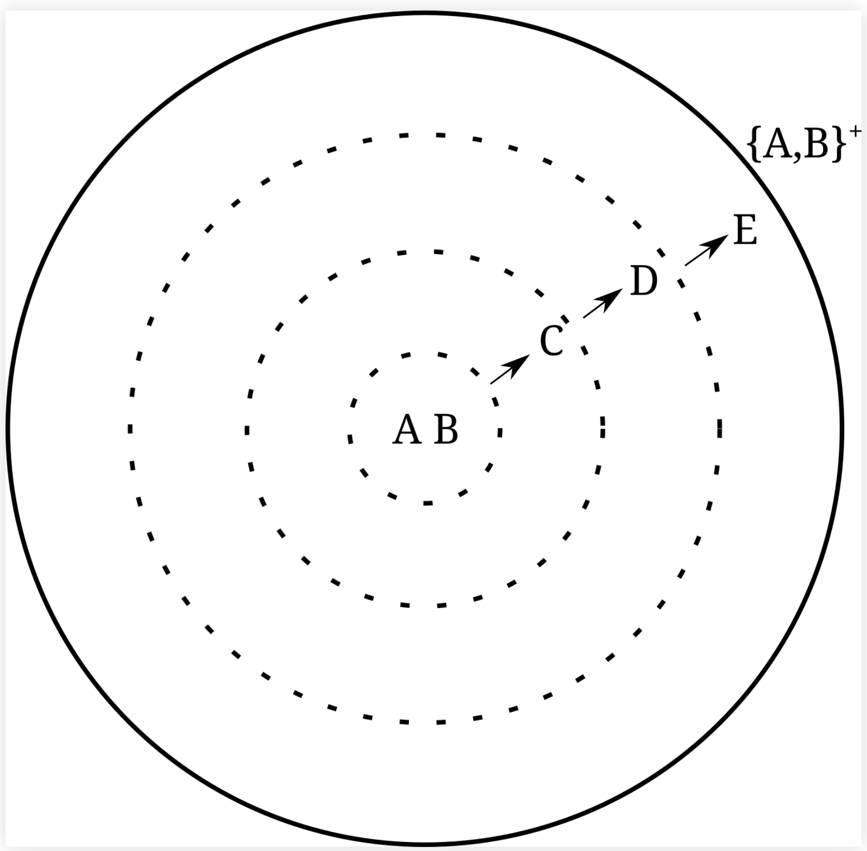
> - ***INPUT:*** *A set of attributes* $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ *and a set of FD's* $\mathcal{S}$
> - ***OUTPUT:*** *The closure* $\mathcal{A}^+ \supseteq \mathcal{A}$
>
> 1. *If necessary, split the FD's of* $\mathcal{S}$*, so that each FD in* $\mathcal{S}$ *has a single attribute to the right (splitting rule)*
> 2. *Let* $\mathcal{X}$ *be a set of attributes that eventually will become the closure. Initialize* $\mathcal{X} := \mathcal{A}$
> 3. *Look in* $\mathcal{S}$ *for any FD on the form* $B_1 B_2 \ldots B_m \to C$ *such that the left hand side* B_k *are* ***all*** *in* X*, but* C ***is not****.*
> 4. *If such an FD is found: Include* C *in the attribute set:* $\mathcal{X} := \mathcal{X} \cup \{C\}$*;* ***goto 3***
> 5. *else (an FD* ***is not found****):* $\mathcal{A}^+ := \mathcal{X}$*; stop.*

Aalto University
School of Science

EXAMPLE FROM U&W (P. 73)

- Relation `R(A,B,C,D,E,F)`
- FD's: $\mathcal{S} = \{BC \rightarrow AD, D \rightarrow E, AB \rightarrow C, CF \rightarrow B\}$

What is $\{A, B\}^+$, i.e. the closure of $\{A, B\}$?



FD's (split up):

$$BC \rightarrow A$$

$$BC \rightarrow D$$

$$D \rightarrow E$$

$$AB \rightarrow C$$

$$CF \rightarrow B$$

DECOMPOSITION

- A relation can be *decomposed* into two new relations by splitting its attributes
- This is done in an attempt to find a new schema which eliminates anomalies
- The goal is to replace a 'big' relation with several smaller ones that do not exhibit any anomalies

DECOMPOSITION EXAMPLE

$$\mathtt{M(title,year,length,genre,studioName,starName)}$$

$$\mathtt{M1} := \pi_{title,year,length,genre,studioName}(\mathtt{M})$$

title	year	length	genre	studioName
Star Wars	1977	124	SciFi	Fox
Waynes World	1992	95	comedy	Paramount
Notting Hill	1999	124	comedy	MCA-Universal

$$\mathtt{M2} := \pi_{title,year,starName}(\mathtt{M})$$

title	year	starName
Star Wars	1977	Harrison Ford
Star Wars	1977	Carrie Fisher
Star Wars	1977	Mark Hamill
Waynes World	1992	Mike Myers
Notting Hill	1999	Julia Roberts

Note that $M = M1 \bowtie M2$

title	year	length	genre	studioName	starName
Star Wars	1977	124	SciFi	Fox	Harrison Ford
Star Wars	1977	124	SciFi	Fox	Carrie Fisher
Star Wars	1977	124	SciFi	Fox	Mark Hamill
Waynes World	1992	95	comedy	Paramount	Mike Myers
Notting Hill	1999	124	comedy	MCA-Universal	Julia Roberts

PROJECTION OF FD'S

What happens to an FD when the relation it is defined for is decomposed?

- The original FD's are not necessarily valid
- New FD's may result due to the projection of the original set

Let $\mathtt{R}$ *be a relation, decomposed into the relation* $\mathtt{R1}$ *(and some other relation).*

Let $\mathcal{S}$ *be the set of FD's for* $\mathtt{R}$*.*

Then valid FD's for $\mathtt{R1}$ *can be determined as*

For each possible subset of attributes $\mathcal{A}$ *of* $\mathtt{R1}$*, and some attribute* B *of* $\mathtt{R1}$*,* $\mathcal{A} \rightarrow B$ *is an FD, if the following conditions hold*

1. B *is included in* $\mathcal{A}^+$ *(with respect to* $\mathcal{S}$*)*
2. B *is not included in* $\mathcal{A}$

Aalto University
School of Science

EXAMPLE 13, U&W (P. 78)

Given `R(A,B,C,D)` and FD's $A \to B, B \to C, C \to D$. Decompose into `R1(A,C,D)` and some other relations, which FD's hold in `R1`?

- Look at subsets of the attribute set $\{A, C, D\}$
- $\{A\}^+ = \{A, B, C, D\}$
  - $A \to C, A \to D$
- $\{C\}^+ = \{C, D\}$
  - $C \to D$
- $\{D\}^+ = \{D\}$
  - No new FD's
- $\{A, C\}^+ = \{A, B, C, D\}$
  - No new FD's
- $\{C, D\}^+ = \{C, D\}$
  - No new FD's
- And so on…

BOYCE-CODD NORMAL FORM - BCNF

> *A relation,* $\mathtt{R}$ *is said to be in BCNF if and only if: for any non-trivial FD* $A_1 A_2 ... A_n \rightarrow B_1 B_2 ... B_m$ *for* $\mathtt{R}$*, the attribute set* $\{A_1, A_2, ..., A_n\}$ *is a superkey for* $\mathtt{R}$*.*

- In other words, the closure of the left hand side of any non-trivial FD contains all the attributes
- A relation on BCNF does not exhibit the previously mentioned anomalies

BCNF EXAMPLE

- `M(title,year,length,genre,studioName,starName)`
  - FD $title\ year \rightarrow length\ genre\ studioName$ holds in `M`
  - But the left side of the FD **is not a superkey** in `M`
  - Thus `M` **is not** BCNF

-`M1`, `M2`, on the other hand are **both** BCNF

$$M1 := \pi_{title,year,length,genre,studioName}(M)$$

title	year	length	genre	studioName
Star Wars	1977	124	SciFi	Fox
Waynes World	1992	95	comedy	Paramount
Notting Hill	1999	124	comedy	MCA-Universal

$$M2 := \pi_{title,year,starName}(M)$$

title	year	starName
Star Wars	1977	Harrison Ford
Star Wars	1977	Carrie Fisher
Star Wars	1977	Mark Hamill
Waynes World	1992	Mike Myers
Notting Hill	1999	Julia Roberts

HOW TO DECOMPOSE A RELATION TO BCNF ?

1. Pick any non-trivial FD violating BCNF for **R**
   - (If none is found **R** is on BCNF)
2. Use it to create two (partially overlapping) sets of attributes
   - Set 1 : The closure of the determinants (the left side) of the violating FD
   - Set 2 : The union of the set of determinants with any attributes in **R** not already in Set 1.
3. These two sets are the attributes of two new relations: **R1**, **R2**
4. Apply the same procedure to **R1** and **R2**.

Note: The choice of which FD to use in step 1 can lead to different partitions (all valid). In the exercises it is sufficient to present one of them.

EXAMPLE - U&W (P. 87)

`M(title,year,studioName,president,presAddr)`

Schema: $\{title, year, studioName, president, presAddr\}$

Functional Dependencies:

- $title\ year \rightarrow studioName$ [BCNF]
- $studioName \rightarrow president$ [Violation]
- $president \rightarrow presAddr$ [Violation]

EXAMPLE - U&W (P. 87)

Schemas: $\{title, year, studioName\}$ and $\{studioName, president, presAddr\}$

Functional Dependencies:

$\{studioName, president, presAddr\}$

- $studioName \rightarrow president$ [BCNF]
- $president \rightarrow presAddr$ [Violation]

$\{title, year, studioName\}$

- $title\ year \rightarrow studioName$ [BCNF]

Aalto University
School of Science

IS BCNF ALWAYS ENOUGH ? (NO)

- A relation on BCNF has no anomalies due to functional dependencies
- But, there may still be redundancy issues present in such relations
- Multivalued dependencies
    - Often occurring when a relation has to contain combinations of possible attribute values

EXAMPLE FROM U&W 3:6.1 (P. 102)

```
S(name, street, city, title,year)
```

A relation containing movie star addresses and films they've stared in.

name	street	city	title	year
C. Fisher	123 Maple St.	Hollywood	Star Wars	1977
C. Fisher	5 Locus Ln.	Malibu	Star Wars	1977
C. Fisher	123 Maple St.	Hollywood	Empire Strikes Back	1980
C. Fisher	5 Locus Ln.	Malibu	Empire Strikes Back	1980
C. Fisher	123 Maple St.	Hollywood	Return of the Jedi	1983
C. Fisher	5 Locus Ln.	Malibu	Return of the Jedi	1983

- One star, two addresses, and three movies: six tuples
- This relation is on BCNF (the key consists of all attributes)
- Yet, much redundancy

MULTIVALUED DEPENDENCIES (MVD'S)

- Generalization of a functional dependency
- A statement not only about the determinants and dependants, but also about the determinants and all attributes *not* in the dependant set
- Expressed using a two headed arrow: $\twoheadrightarrow$

MVD STATES THAT

- Relation `R`
  - Attribute set $\{A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_m, C_1, C_2, \ldots, C_k\}$
- MVD $A_1\ A_2\ \ldots\ A_n \twoheadrightarrow B_1\ B_2\ \ldots\ B_m$

For any two tuples t, u in a relation agreeing on the $A_1, \ldots, A_n$, there is another tuple v such that

1. v agrees with both t and u on $A_1, \ldots, A_n$
2. v agrees with one of t, u on $B_1, \ldots, B_m$
3. v agrees *with the other one of* t, u on for *every other* attribute in `R`: $C_1, \ldots, C_k$

MVD PATTERN

Capture patterns like

	A	B	C
t	x	y	*
u	x	*	z
v	x	y	z

for MVD $A \twoheadrightarrow B$

MVD'S - EXAMPLE

$name \twoheadrightarrow street\ city$ is a MVD in

`S(name, street, city, title,year)`

Take tuples

C. Fisher	5 Locus Ln.	Malibu	Star Wars	1977
C. Fisher	123 Maple St.	Hollywood	Empire Strikes Back	1980

The MVD declares that also tuples

C. Fisher	123 Maple St.	Hollywood	Star Wars	1977
C. Fisher	5 Locus Ln.	Malibu	Empire Strikes Back	1980

are valid in the relation.

Aalto University
School of Science

MVD'S - EXAMPLE CONT.

name	street	city	title	year
C. Fisher	123 Maple St.	Hollywood	Star Wars	1977
C. Fisher	5 Locus Ln.	Malibu	Star Wars	1977
C. Fisher	123 Maple St.	Hollywood	Empire Strikes Back	1980
C. Fisher	5 Locus Ln.	Malibu	Empire Strikes Back	1980
C. Fisher	123 Maple St.	Hollywood	Return of the Jedi	1983
C. Fisher	5 Locus Ln.	Malibu	Return of the Jedi	1983

NONTRIVIAL MVD'S

$$A_1\ A_2\ \ldots\ A_n \twoheadrightarrow B_1\ B_2\ \ldots\ B_m$$

trivial

if $\{B_1, B_2, \ldots, B_m\} \subseteq \{A_1, A_2, \ldots, A_n\}$

nontrivial

if $\{B_1, B_2, \ldots, B_m\} \not\subseteq \{A_1, A_2, \ldots, A_n\}$, and there are some other attributes of the relation than those of the MVD

Aalto University
School of Science

FOURTH NORMAL FORM (4NF)

A relation, $\mathtt{R}$*, is said to be in the fourth normal form if whenever some MVD* $A_1\ A_2\ \ldots\ A_n \twoheadrightarrow B_1\ B_2\ \ldots\ B_m$ *is nontrivial then* $\{A_1, A_2, \ldots, A_n\}$ *is a superkey of* $\mathtt{R}$*.*

- In other words:
  - It has no nontrivial functional dependencies
  - nor nontrivial multivalued dependencies,
  - which are not superkeys.
- A relation in 4NF is always in BCNF
  - The opposite is not necessarily the case
  - 4NF is stricter than BCNF