Database Design Theory and Normalization Part II

CS 377

Recap: Last Class

- Normal form: set of properties that relations must satisfy
  - Relations exhibit less anomalies
  - Successively higher degrees of stringency
- 1NF: most basic normal form with atomic attributes
- Functional dependencies: X —> Y
  - Armstrong's axioms to derive additional FDs to find good relational decompositions

Finding Keys of Relation R

- Bad news: find all keys of a relation is NP-complete
  - Running time of algorithm to solve the problem exactly is exponentially increasing with the problem size
  - Large NP-complete problems are difficult to solve!
  - No efficient solution to find all the keys
  - Brute force algorithm: Check every subset of attributes for super key strategy — tests every possible solution
- Solution: use heuristics to find all the keys of a relation
  - Turn towards closures to help us find keys in a relation

Attribute Closure Set

- If X is an attribute set, the closure X^+ is the set of all attributes B such that X —> B
  - X is subset of X^+ since X —> X
  - X^+ includes all attributes that are functionally determined from X
- Importance: If $X^+ = R$, then X is a superkey
  - Closure can tell us if set of attributes X is a superkey

Example: Closure

- **Product**(name, category, color, department, price)
  - name —> color
  - category —> department
  - color, category —> price
- Attribute Closure:
  - $\{name\}^+$ = {name, color}
  - $\{name, category\}^+$ = {name, color, category, department, price}

Finding a Key after Closure

- If X^+ not equal to the relation, we must augment more attributes to X to obtain a key
- If $X^+ = R$, then X is superkey — check for minimality
  - Remove one or more attributes A
  - Compute the closure of X - A to see if $(X - A)^+ = R$
  - X is a key if $(X - A)^+$ not equal R for any attribute A

Closure Algorithm

- Input: A set F of FDs on a relation schema R, and a set of attributes X, which is a subset of R

- Algorithm:
  Initialize $X^+ := X$
  repeat
      old $X^+ := X^+$
      for each functional dependency $Y \longrightarrow Z$ in F
          if X^+ superset Y, then $X^+ := X^+$ union Z
  until (X^+ = old X^+)

Example: Closure Algorithm

EmpProj(SSN, FName, LName, PNo, PName, PLocation, Hours)

- SSN —> FName, LName
- PNo —> PName, PLocation
- SSN, PNo —> Hours

Example: Closure Algorithm (2)

- Initialize SSN^+ := SSN
- Repeat loop (for each FD)
  - SSN —> FName, LName
    => SSN^+ := SSN, FName, LName
  - PNo —> PName, PLocation
    => no change
  - SSN, PNo —> Hours
    => no change

Since there were changes, repeat another loop through FDs, which results in no changes => done

- Result: SSN^+ := SSN, FName, LName

Example: Closure Algorithm (3)

- Initialize PNo^+ := PNo
- Repeat loop (for each FD)
  - SSN —> FName, LName
    => no change
  - PNo —> PName, PLocation
    => PNo^+ := PNo, PName, PLocation
  - SSN, PNo —> Hours
    => no change
- Result: PNo^+ := PNo, PName, PLocation

Since there were changes, repeat another loop through FDs, which results in no changes => done

Example: Closure Algorithm (4)

- Initialize $(SSN, PNo)^+$:= SSN, PNo
- Repeat loop (for each FD)
  - SSN —> FName, LName
    => $(SSN, PNo)^+$:= SSN, PNo, FName, LName
  - PNo —> PName, PLocation
    => $(SSN, PNo)^+$:= SSN, PNo, FName, LName, PName, PLocation
  - SSN, PNo —> Hours
    => $(SSN, PNo)^+$:= SSN, PNo, FName, LName, PName, PLocation, Hours
- Result: $(SSN, PNo)^+$:= SSN, PNo, FName, LName, PName, PLocation, Hours

Example: Closure Algorithm (4)

- Summary of results:
  - SSN^+ := SSN, FName, LName
  - PNo^+ := PNo, PName, PLocation
  - $(SSN, PNo)^+$:= SSN, PNo, FName, LName, PName, PLocation, Hours
- (SSN, PNo) is a superkey!
- (SSN, PNo) is minimal superkey
  - $\{(SSN, PNo) - (SSN)\}^+ = (PNo)^+$
  - $\{(SSN, PNo) - (PNo)\}^+ = (SSN)^+$

Finding Keys: Heuristic 1

- Increase/decrease until you find keys
- Step 1: Compute closure of all functional dependencies in F
- Step 2:
  - If deficient, then add missing attributes to the LHS until the closure is equal to the relation
  - If sufficient, then remove extraneous attributes from the LHS until set is minimal

Example: Key Heuristic 1

- **R**(A, B, C, D, E, F)
  - A —> B, C
  - B, D —> E, F
  - F —> A
- Step 1: Closure of all functional dependencies
  - A^+ = A, B, C
  - $(B, D)^+$ = A, B, C, D, E, F
  - F^+ = F, A, B, C

Example: Key Heuristic 1 (2)

- Step 2: Insert / remove attributes
  - A^+ = A, B, C — insufficient so add
    - Add D: $(A, D)^+$ = A, B, C, D, E, F —> key!
    - Add E: $(A, E)^+$ = A, B, C, E
    - Add F: $(A, F)^+$ = A, B, C, F
    - Add E, F: $(A, E, F)^+$ = A, B, C, E, F
    - No more so done

Example: Key Heuristic 1 (3)

- Step 2: Insert / remove attributes
  - $(B, D)^+ = A, B, C, D, E, F$ — sufficient so try deleting
    - Delete B: $(D)^+ = D$
    - Delete D: $(B)^+ = B$
    - No more so done

B, D is minimal and thus a key!

Example: Key Heuristic 1 (4)

- Step 2: Insert / remove attributes
  - F^+ = F, A, B, C — insufficient so add
    - Add D: $(D, F)^+$ = A, B, C, D, E, F —> key!
    - Add E: $(E, F)^+$ = A, B, C, E, F
    - No more so done

Keys are: (A, D), (B, D), and (D, F)!

Finding Keys: Heuristic 2

- Find necessary attributes first
- Find the irreplaceable attributes
    - Attribute is replaceable if it appears in the RHS of some functional dependency
- A key must include every irreplaceable attribute
- Base set is set of all irreplaceable attributes
- Add other attributes to base set until you have a key

Example: Key Heuristic 2

- **R**(A, B, C, D, E, F)
  - A —> B, C
  - B, D —> E, F
  - F —> A
- Step 1: Find irreplaceable attributes and construct base set
  Base set = {D}

Example: Key Heuristic 2 (2)

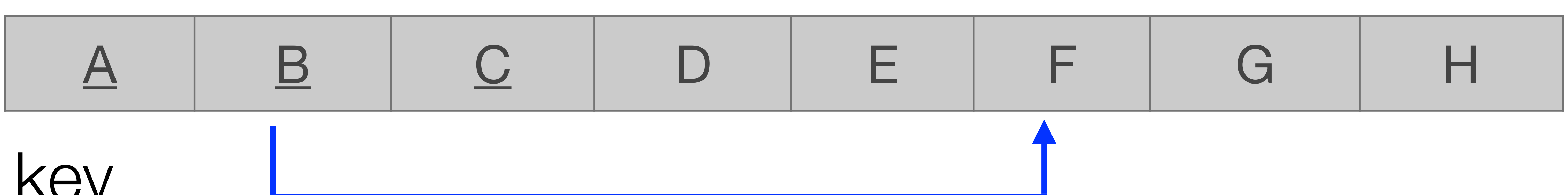
- Step 2: Add other attributes until you have key
  - Add A: $(A, D)^+$ = A, B, C, D, E, F —> key!
  - Add B: $(B, D)^+$ = A, B, C, D, E, F —> key!
  - Add C: $(C, D)^+$ = C, D
  - Add E: $(D, E)^+$ = D, E
  - Add F: $(D, F)^+$ = A, B, C, D, E, F —> key!

Example: Key Heuristic 2 (3)

- Step 2: Add other attributes until you have key (do not expand known keys)
    - Add C: $(C, D, E)^+ = C, D, E$
    - No more to add, so done!

Second Normal Form (2NF)

- (Definition) A relation schema R is in 2NF if every non-prime attribute (i.e., not a member of any candidate key) A in R is not partially dependent on any key of R
    - Relation is 1NF (attributes are atomic)
    - No non-key attribute that is functionally determined by only a (proper) subset of a key

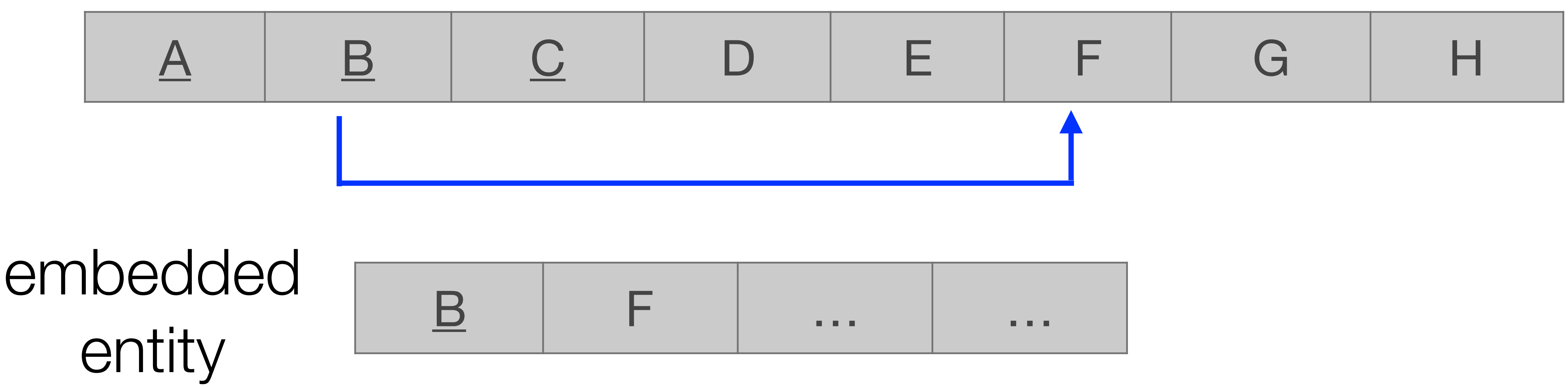


key
(A, B, C)

B —> F means F is functionally dependent on subset of key => violation of 2NF

2NF Meaning

A relation that violates 2NF contains another embedded autonomous entity



Example: Violation of 2NF

- **EmpProj**(<u>SSN</u>, FName, LName, <u>PNo</u>, PName, Hours)
  - SSN —> FName, LName
  - PNo —> PName
  - SSN, PNo —> Hours
- FName is not part of any key
- SSN is (proper) subset of a key
- Violation since Employee entity is embedded (SSN, FName, LName)

Decomposition for Normal Form Violations

- Break a relation into two or more relations
- One possibility for **EmpProj**(<u>SSN</u>, FName, LName, <u>PNo</u>, PName, Hours):
  - R1(PNo, PName, Hours)
  - R2(SSN, FName, Lname)
- Another possibility for EmpProj
  - R3(SSN, FName, Lname)
  - R4(SSN, PNo, PName, Hours)

Are these good or bad decompositions?

Decomposition Effect

- Populate the new relations using data of the original relation
  - Achieve this by using projection operation on the original relation
  - Example:

$$R1 = \pi_{\text{SSN},\text{FName},\text{LName}}(\text{EmpProj})$$
$$R2 = \pi_{\text{PNo},\text{PName},\text{Hours}}(\text{EmpProj})$$

Decomposition Effect (2)

- Can we obtain the same information stored in the original relation?

- Reconstruction algorithm:
  If ($R1 \cap R2 \neq \emptyset$) {
  reconstruction = R1 * R2 // Natural join
  } else {
  reconstruction = R1 x R2 // Cartesian product
  }

Example: Decomposition Effect

SSN	FName	LName	PNo	PName	Hours
111-11-1111	John	Smith	pj1	ProjectX	20
111-11-1111	John	Smith	pj2	ProjectY	10
333-33-3333	Jack	Rabbit	pj1	ProjectX	5

⇩

SSN	FName	LName
111-11-1111	John	Smith
333-33-3333	Jack	Rabbit

PNo	PName	Hours
pj1	ProjectX	20
pj2	ProjectY	10
pj1	ProjectX	5

Example: Reconstructing After Decomposition

SSN	FName	LName
111-11-1111	John	Smith
333-33-3333	Jack	Rabbit

x

PNo	PName	Hours
pj1	ProjectX	20
pj2	ProjectY	10
pj1	ProjectX	5

SSN	FName	LName	PNo	PName	Hours
111-11-1111	John	Smith	pj1	ProjectX	20
111-11-1111	John	Smith	pj2	ProjectY	10
111-11-1111	John	Smith	pj1	ProjectX	5
333-33-3333	Jack	Rabbit	pj1	ProjectX	20
333-33-3333	Jack	Rabbit	pj2	ProjectY	10
333-33-3333	Jack	Rabbit	pj1	ProjectX	5

Extraneous tuples that weren't present in original relation!

Decomposition Relation Requirements

- Must be able to obtain all tuples in the original relation R using the reconstruction algorithm
  - Missing tuples means that we have lost information which is unacceptable
- Must not obtain extraneous tuples that were not present in the original relation R using the reconstruction algorithm
  - Invalid information in the relation which is also unacceptable

Lossless Decomposition

- A decomposition of relation R into 2 relations R1 and R2 is called lossless if and only if content(R1) * content(R2) = content(R) or content (R1) x content(R2) = content(R)

- 2 lemmas that provide needed guidelines to decompose R to guarantee lossless

  - Lemma 1: $\text{content}(R) \subseteq \text{content}(R_1) * \text{content}(R_2)$

  - Lemma 2: If either $R_1 \cap R_2 \rightarrow R_1$ or $R_1 \cap R_2 \rightarrow R_2$, then $\text{content}(R) = \text{content}(R_1) * \text{content}(R_2)$

Example: 2NF via Lemma 2

- **EmpProj**(SSN, FName, LName, PNo, PName, Hours)
  - SSN —> FName, LName
  - PNo —> PName
  - SSN, PNo —> Hours
- At least one violating FD
  - SSN —> FName
  - SSN —> LName

Remove all attributes functionally dependent on SSN => compute closure of SSN

Example: 2NF via Lemma 2 (2)

- $R1(SSN^+)$ = **R1**(SSN, FName, LName)
- R2(R - R1) = **R2**(PNo, PName, Hours)
  - To satisfy lemma 2, add SSN to R2 => **R2**(SSN, PNo, PName, Hours)
  - $R1 \cap R2$ = SSN, and SSN —> R1

Are R1 and R2 in the 2NF?

Example: 2NF via Lemma 2 (3)

- **R1**(SSN, FName, LName)
  - SSN —> FName, FName — key = good dependency
- **R2**(SSN, PNo, PName, Hours)
  - SSN, PNo —> Hours — key = good dependency
  - PNo —> PName — not key = bad!

Remove all attributes functionally dependent on PNo => compute closure of PNo

Example: 2NF via Lemma 2 (4)

- R21(PNo^+) = **R21**(<u>PNo</u>, PName)
- R22(R2 - R21) = **R22**(<u>SSN</u>, Hours)
  - To satisfy lemma 2, add PNo to R22 => **R22**(SSN, PNo, Hours)
- Resulting decomposition:
  R1(<u>SSN</u>, FName, LName)
  R21(<u>PNo</u>, PName)
  R22(<u>SSN</u>, <u>PNo</u>, Hours)

Are R1, R21, and R22 in the 2NF?

Example: 2NF Complaint

- **Employee2**(<u>SSN</u>, FName, LName, DNo, DName, MgrSSN)
  - SSN —> FName, LName, DNo
  - DNo —> DName, MgrSSN
- Employee2 is 2NF as DNo is not a subset of any key and neither of the functional dependencies violate 2NF criteria
- But…
  - Insert anomaly — adding new department results in NULL values
  - Delete anomaly — deleting an employee may delete information about department
  - Update anomaly — changing department name results in updates of multiple tuples

Transitive Functional Dependency

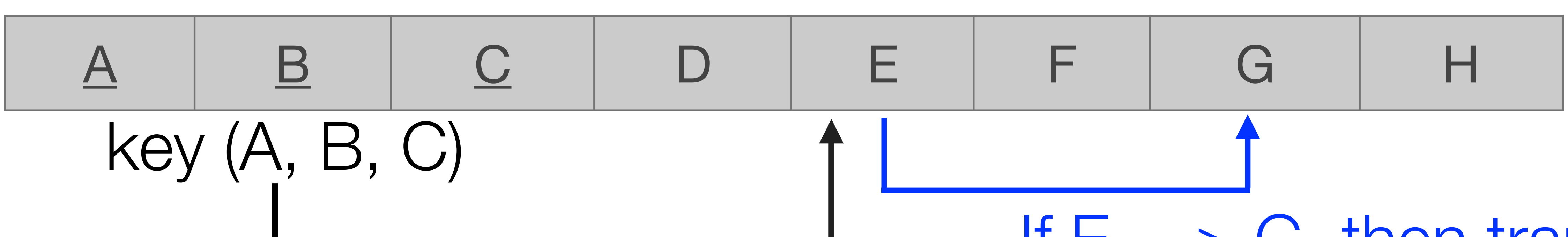
A functional dependency A —> B is a transitive functional dependency in relation R if there is a set of attributes X such that:

- A —> X
- X —> B
- X is not a super key

Third Normal Form (3NF)

(Definition) A relation schema R is in 3NF if, whenever a nontrivial functional dependency X —> A holds in R, either (a) X is a super key of R, or (b) A is a prime attribute of R

- R is in 2NF
- Every non-key attribute is non-transitively dependent on all the keys



If E —> G, then transitive dependency

(A, B, C) —> E —> G

Example: 3NF Violation

- **Employee2**(SSN, FName, LName, DNo, DName, MgrSSN)
  - SSN —> FName, LName, DNo
  - DNo —> DName, MgrSSN
- Since DNo is not a super key, there is a transitive dependency SSN —> DNo —> DName, MgrSSN

Simpler Form of 3NF

- A relation R is 3NF if and only if for every functional dependency X —> B in relation R, one of the following must be true:
  - X is a superkey, or
  - B is a key attribute (part of some key)
- Violation detection: Check every functional dependency X —> B for:
  - B is a non-key attribute, and
  - X is not a superkey

Example: 3NF Violation Take 2

Employee2(SSN, FName, LName, DNo, DName, MgrSSN)

- SSN —> FName, LName, DNo
  - FName, LName, and DNO are non-key attributes => YES
  - SSN is not superkey => NO
  - FD is good
- DNo —> DName, MgrSSN
  - Name and MgrSSN are non-key attributes => YES
  - DNo is not superkey => YES
  - FD is bad and a 3NF violation

Example: 3NF Decomposition

- Solution: remove the violation by removing X^+ from the original relation
- R(A, B, C, D, E, F)
  - A —> B, C, D
  - D —> E, F
- Step 1: Find all keys
  - A^+ = (A, B, C, D, E, F)

Example: 3NF Decomposition (2)

- Step 2: Is R 2NF?
  - Key(s): A
  - Non-key attributes: B, C, D, E, F
  - Is any of the non-key attributes functionally dependent on subset of (A)? NO
  - Relation is 2NF

Example: 3NF Decomposition (3)

- Step 3: Is R 3NF?
  - Key(s): A
  - Non-key attributes: B, C, D, E, F
  - Is any of the non-key attributes functionally dependent on attributes that are not super key? YES!
    - D —> E, F where D is not a superkey

Example: 3NF Decomposition (4)

- Step 4: Extract offending functional dependence
  - $D^+ = (D, E, F)$
  - R1(<u>D</u>, E, F)
    R2(<u>A</u>, B, C, D)
- Step 5: Check the new relations if they are 3NF?
  - R1: D —> E, F doesn't violate 3NF criteria
  - R2: A —> B, C, D doesn't violate 3NF criteria

Summary of 1NF, 2NF, 3NF

Normal Form	Test	Normalization (Remedy)
1NF	Relation should have no multi-valued attributes or nested relations	Form new relation for each multivalued attribute or nested relation
2NF	For relations where primary key contains multiple attributes, no nonkey attribute should be functionally dependent on a part of the primary key	Decompose and set up a new relation for each partial key with its dependent attributes using lossless decomposition
3NF	Relation should not have a nonkey attribute functionally determined by another nonkey attribute	Decompose and set up a relation that includes the nonkey attribute(s) that functionally determine(s) other nonkey attributes

Boyce-Codd Normal Form (BCNF)

(Definition) A relation schema R is in BCNF if whenever a nontrivial functional dependency X —> A holds in R, then X is a superkey of R

- Difference from 3NF: 3NF allows A to be prime attribute
- Every relation in BCNF is also in 3NF
- Most relation schemas that are in 3NF are also BCNF but not all
  - Example: R(<u>A</u>, <u>B</u>, C)
    - A, B —> C
    - C —> A

Example: BCNF Violation

- TSS(Teacher, Subject, Student)
  - Student, Subject —> Teacher
  - Teacher —> Subject
- Keys in TSS
  - (Student, Subject)
  - (Student, Teacher)

Example: BCNF Violation (2)

- Is TSS in the 3NF?
  - Student, Subject —> Teacher — superkey = okay
  - Teacher —> Subject
    - Is teacher a superkey? NO
    - Is subject a key attribute (part of key)? YES — okay
- Even though TSS is 3NF…
  - Duplicate information is stored in relation (teacher, subject)

Example: BCNF Violation (3)

- Problem arises when 2 or more composite keys are in a relation
- Is relation BCNF?
  - Student, Subject —> Teacher — superkey = okay
  - Teacher —> Subject
    Teacher is not a superkey => BCNF violation!
- Solution: Decompose the violating FD
  - T1(Teacher, Subject)
    R2(Teacher, Student)

Is Normalization Always Good?

- Example: Suppose A and B are always used together but normalization says they should be in different tables
  - Decomposition might produce unacceptable performance loss (always joining tables)
  - For example, data warehouses are huge historical DBs that are rarely updated after creation — joins are expensive or impractical
- Everyday DBs: aim for BCNF, settle for 3NF!

Database Design: Recap

- Closure algorithm to find keys
- Lossless decomposition
- 2NF
- 3NF
- BNCF