**CSDS 234 Structured and Unstructured Data  
Department of Computer & Data Sciences, Fall 2023**

Assignment 6



1. **[Dirty data] (30)** Recall the five common types of data quality issues we discussed: **inconsistency** (violation of data constraints), **duplication** (redundant information), **missing data** (incomplete data), **data currency** (outdated data), and **data accuracy** (inaccurate data).

You are reviewing medical record data (structured data stored in tables) for a hospital, which records patient information, and their pharmacy and insurance providers. Annotate the following descriptions with \*all\* the type of data quality issues that can describe them. Briefly clarify the reason.

* 1. There are multiple medical records and none of them has required identifiers called “medical record numbers” (MRN).
* Issues:
  + Missing Data with the MRNs
    - They are required identifiers, and without them, the data is incomplete
    - It’s a violation of the data constraint that requires every record to have a unique MRN
  1. A patient named “John Carroll” is recorded as “J.Carroll” in the database. His case is confused with another patient named “Joe Caroll” who has the same birthdate as John Carroll. The latter patient has a recorded address in “Nimrod St, Solon” but he said he has already moved to “Wintergreen Dr, Solon”.
* Issues:
  + Inconsistency of how patients are recorded in the database
    - This violates the constraint of unique patient identification
  + Duplicate Records
    - The two people might have redundant or overlapping records
  + Data Accuracy
    - The address for Joe Caroll is incorrect and there may be errors in the names
  + Data Currency
    - The address for Joe Caroll is incorrect as he has moved
  1. A patient claimed he has an insurance provider A, but A does not exist in any existing provider records; there is a data constraint that states “any insurance provider of a patient must be from a set of registered providers”.
* Issues:
  + Inconsistent:
    - There is a violation of the data constraint where the insurance provider must exist in the set
  + Data Accuracy:
    - The data inaccurately lists a provider that is non-existent
  1. There are two patient records with different MRNs but the same name, birthdate, SSN, and home address. Later it is confirmed that both refer to the same patient.
* Issues:
  + Duplication
    - There are two records that refer to the same patient
  + Inconsistency
    - The patient should not have two MRN’s
  1. A transferred patient has an age “65” in her profile from another healthcare provider and an age “70” in the hospital records. Her birth year is missing in both records.
* Issues:
  + Inconsistency:
    - Different ages for the same patient
  + Missing Data:
    - Birth year is missing
  + Data Accuracy:
    - One or both age values may be incorrect
  1. A pharmacy provider has relocated to a new address; the patient records use the old address.
* Issues:
  + Data Currency:
    - The provider’s address is outdated
  1. The registrar types in the name “Jessica Smith” and several medical records pop up on the screen, all pertaining to the same patient with the same MRN.
* Issues:
  + Duplication:
    - Several records redundantly represent the same patient
  + Inconsistency:
    - Duplicate records violate the rule that MRNs must uniquely identify patients
  1. A patient visited his local pharmacy to get his medicine and was told that he would need to pay the full bill because the pharmacy’s record and the insurance provider’s record of that patient had “age” mismatched.
* Issues:
  + Inconsistency:
    - The age mismatch indicates a violation of data alignment between systems
  + Data Accuracy:
    - Either or both records have the wrong age

1. **[Data errors] (40)** Consider two databases from a bank with the following schema:

**Customer (FN, LN, St, City, CC, Country, tel, gd);   
Tran (FN, LN, St, City, CC, Country, phn, when, where)**

**Customer** record specifies a credit card holder identified by first name (FN), last name (LN), street (St), city, country code (CC), country, phone number (tel) and gender (gd). A **tran** tuple is a record of a purchase paid by a credit card at time when (local time) and place where, by a customer identified by first name (FN), last name (LN), street (St), City, country code (CC), Country and phone number (phn). *An inclusion dependency* states that for any tuple in **Tran** with value **Tran(FN,LN,phn),** there is a tuple in **Customer** with the same value **Customer(FN,LN,Tel).**

The records below are correct, consistent, and up to date. They log two customers and their matching transaction records.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **FN** | **LN** | **St** | **City** | **CC** | **Country** | **Tel** | **gd** |
| David | Jordan | 12 Holywell St | Oxford | 44 | UK | 66700543 | Male |
| Paul | Simon | 5 Ratcliffe Terrace | Oxford | 44 | UK | 44944631 | Male |
| **FN** | **LN** | **St** | **City** | **CC** | **Country** | **phn** | **when** | **where** |
| David | Jordan | 12 Holywell St | Oxford | 44 | UK | 66700543 | 1 pm, 10/18/2019 | Netherlands |
| Paul | Simon | 5 Ratcliffe Terrace | Oxford | 44 | UK | 44944631 | 6 am, 11/25/2019 | US |

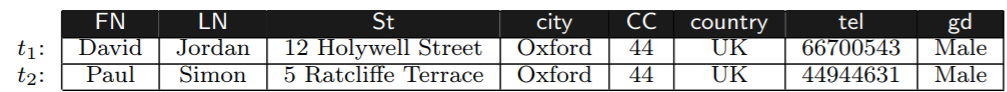
If You are an “attacker” and have gained access to the table. For each of the five data quality issues,

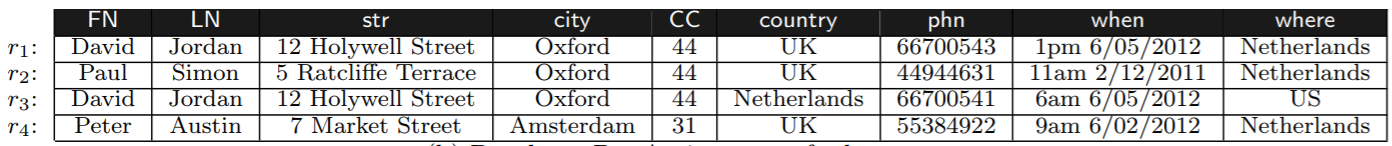
* 1. give at least one example modification (Insert, delete, update) that will “pollute” the above clean data by causing that data quality issue (no need to give SQL; just describe how you will modify the tuples);
  2. give a query Q such that the results of Q from the original tables and the polluted tables will be different. You may state the query in SQL, relational algebra, or natural language.

For example, removing “Oxford” from City causes missing data in Tran; a query SELECT City FROM Customer WHERE CC=’44’ and LN=’Simon’ will be affected.

* Inconsistent Data:
  + Update the phn in the Tran table for David Jordan to 111-111-1111. This ends up not matching the corresponding Tel in the Customer table, which violates the inclusion dependency
  + Query:
    - SELECT Tran.FN, Tran.LN FROM Tran WHERE Tran.phn NOT IN (SELECT Tel FROM Customer)
* Duplicate Data:
  + Insert a new record into the Customer table which creates a duplicate record for one of the persons (ex, David Jordan)
  + Query:
    - SELECT FN, LN, COUNT(\*) FROM Customer GROUP BY FN, LN HAVING COUNT(\*) > 1
* Missing Data:
  + Delete the phn value for Paul Simon’s transaction record in the Tran table and leave it blank
  + Query:
    - SELECT FN, LN, phn FROM Tran WHERE phn IS NULL;
* Outdated Data:
  + Update the St field in the Customer table for Paul Simon to 26 Bull Street but leave the address in the Tran table unchanged
  + Query:
    - SELECT Customer.FN, Customer.LN, Customer.St, Tran.St FROM Customer JOIN Tran ON Customer.FN = Tran.FN AND Customer.LN = Tran.LN WHERE Customer.St <> Tran.St
* Incomplete Data:
  + Update the gd field in the Customer table for David Jordan, which makes it incorrect. This works for this table and these records
  + Query:
    - SELECT FN, LN, gd FROM Customer WHERE gd != ‘Male’

1. **[Data cleaning] (30)** Consider the following instances for the same schema in Problem 1.





The security department of the bank now has provided several data quality rules to you as follows. For each of the rule, give an error that can be captured by the rule (for example, tuple r3.country), and suggest how to fix it by modifying the attribute values (for the last rule, suggest if there is a possible fraud that should be further inspected by security experts).

* (on table Tran): if a customer’s CC is 31, but his/her country is neither Netherlands nor Holland, then the country is wrong and should be Netherlands;
  + Error: Tuple r4 where CC = 31 and country is UK
  + How to fix: Modify tuple r4.country to “Netherlands”
* (on tables Customer and Tran): if the same person from different tables has different phones, the phone number from table Customer is more reliable;
  + Error: Tuple r3, where FN = David, LN = Jordan, phn = 66700541 differs from t1.tel = 66700543
  + How to fix: Update r3.phn to 66700543
* (on table Tran) a country code (CC) uniquely determines a country;
  + Error: Tuples r3 and r1 both have CC = 44, but the countries are different, (Netherlands versus UK respectively)
  + How to fix: Update r3.country to “UK” to have consistency
* (on table Tran) if two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour (assuming 6 hours’ time difference between these two countries), these two purchases are either a fraud or were erroneously recorded.
  + Error: Tuple r3 (country = Netherlands, when 6am 6/05/2012) and r4 (country = US, when 9am 6/02/2012)
  + How to fix: Look into it, because it might be a recording error because of timestamps or because of the country, or it may be fraudulent

1. Bonus Question [3 pts]: Where do you see knowledge graphs being implemented in the future?

* I see knowledge graphs being implemented in the future in traffic optimization, especially with traffic lights. Eventually, there will be a use for analyzing data from vehicles and traffic lights to improve road safety and stop congestion and traffic.