CPTS 415 Big Data

Assignment 6

Instructor: Yinghui Wu

---

1. [**Data Quality**] (40) We have introduced the central aspects of data quality issues.
   a. What is data consistency? Give an example of inconsistent data.
   b. What is data accuracy? Give an example of inaccurate data.
   c. What is data currency? Give an example of outdated data.
   d. What is the entity resolution problem? Give an example.
   e. What are new challenges introduced by big data to data quality management?

   ** You may give a running example (e.g, a single relational table) for all the above questions.*

2. [Dependencies for Data Quality] (30)
   (1) Consider a CFD: **(cust(country, area-code, phno → street, city, zip**), Tp) with pattern tableau Tp as:

id	country	area-code	phon	street	city	zip
tp1	44	131	_	_	Edi	_
tp2	01	908	_	_	MH	_
tp3	_	_	_	_	_	_

   Find all tuples in the table below that violate the CFD, and explain why.

id	country	area-code	phon	street	city	zip
t1	44	131	1234567	Mayfield	Edi	EH4 8LE
t2	44	131	3456789	Mayfield	NYC	19082
t3	01	908	3456789	Mountain Ave	NYC	19082
t4	44	131	1234567	Chrichton	EDI	EH8 9LE

   (2) Consider an CIND: (**order[asin, title, price; type] ⊆ book[asin, title, price; nil],** Tp) with Tp:

type
book

   Consider the following three tables:

order

id	asin	title	type	price	country	county
t1	a23	H. Porter	book	17.99	US	DL
t2	a12	J. Denver	CD	7.94	UK	Reyden

book

asin	isbn	title	price
a23	b32	Harry Porter	17.99
a56	b65	Snow white	7.94

CD

asin	title	price	genre
a12	J. Denver	17.99	country
a56	S. White	7.94	a-book

   Find all tuples that violates the CIND, and explain why.

   (3) Consider two tables R1(A, B), R2(B, C), with
   - FD: R1[A] → R1[B]
   - IND: R2[B] ⊆ R1[B]

   (a): Is the set of the above FD and IND ***consistent***? That is, can you construct a pair of instance of R1(A,B) and R2(B,C) that satisfy both constraint?

   (b): Consider table D(R1): {(1, 2), (1, 3)}, D(R2) = {(2, 1), (3, 4)}. Explain why a constraint-by-constraint repair (fixing individual constraints one by one) may not terminate.

3. [**Data privacy and security**] (30)

a. (10) What are the requirement of information security? Give an example to illustrate each requirement.
b. (20) [Data anonymization] What are the concepts of k-Anonymity, I-Diversity and t-Closeness? Give an example to explain each concept.