Jeonghyeon Woo

CPTS 415

HW 6

1.a) Data consistency is validity and integrity of data. It is about how a random data piece in a database is agreeing to the rule. An example of inconsistent data is when there is a relation data model and the models have different records on the same attribute.

b) Data accuracy is about how close a value is to the true value of the entity that it represents. When there are two data with the same primary key, a data that makes less sense than the other is a relatively inaccurate data.

c) Data currency is about when the data is recorded. When a data is too old so it is not reliable, then it is an outdated data. An example of outdated data is that indicating the current president is Obama.

d) Entity resolution problem is about matching the same types of data from different databases. An example is checking if John living in Seattle is the same person as John who subscribed the Seattle newspaper.

e) Data managing for big data is too costly on time and effort because of dirty data and accuracy testing.

2. 1) t3 violated tp2 as the city name is different.

t1 and t4 violated tp1 since t1 and t4 have the same [country, area, phno] but different [street, city].

t2 violates tp1.

2) t1 violates CIND. Its type is book but it has different title then in the book table.

3) a) They are consistent. Since IND is always consistent, IND is consistent, and FD is also consistent.

b) In the given table, two tuples in R1 have the same a values but different b values. This is violation of FD, so it may not be terminated.

3. a) The requirements are confidentiality, integrity, availability, information quality, and completeness. When a user can not view the profile unless the user is logged-in, it is a confidentiality. When the user can not change his profile without sudo command or entering password, it is an integrity. When the system gives full authority to super user, it is an availability. Information quality and completeness is about the system is demanding specific information when a user creates an account.

b) K-anonymity is showing data except attributes that would distinguish a user from the others. It is critical at linking attack. When there are only two columns, name and sex, showing only sex is k-anonymity. l-diversity is having at least l values for sensitive attributes. It is critical at similarity attack and skewness attack. When an attacker reads tuples (woo, male, 12), (woo, male 15), (woo, male, 16), the attacker can not specify how old woo is. T-closeness is making distance between overall distribution of sensitive attribute values and distribution of sensitive attribute values in an equivalent class bounded by t. When an attacker reads (woo, male, 20, hello), (woo, male, 20, hi), (woo, male, 20, yo!), the attacker can not specify woo’s message since differences between the ages are very small.