CS 380S - Theory and Practice of Secure Systems
Fall 2009

Homework #4

Due: 2pm CST (in class), December 3, 2009

NO LATE SUBMISSIONS WILL BE ACCEPTED

YOUR NAME: ______________________________

Collaboration policy

No collaboration is permitted on this assignment. Any cheating (*e.g.*, submitting another person's work as your own, or permitting your work to be copied) will automatically result in a failing grade. The Computer Sciences department code of conduct can be found at `http://www.cs.utexas.edu/academics/conduct/`

Homework #4 (30 points)

Problem 1 (3 points)

Let X be some database, and assume that each element $x \in X$ is drawn from some probability distribution D over integers between 0 and 100. The database is perturbed and then published. Let R be the randomization operator applied to each element of the database to preserve privacy prior to publication. R operates as follows: for each $x \in X$, $R(x) = x + \xi$, where ξ is an integer which is distributed uniformly at random between -20 and 20. Define $R(X) = \bigcup_{x \in X} \{R(x)\}$.

Give an example of a distribution D such that $R(X)$ completely reveals X for *any* X.

Problem 2

In this problem, we consider *online* query monitoring and auditing, *i.e.*, instead of publishing a perturbed database, the database owner interactively receives queries and, for each query, decides whether it is safe to answer it using some auditing or monitoring algorithm.

Let $X = \{x_1, \ldots, x_n\}$ be the database. Each element x_i is associated with some integer value v_i. The questioner specifies any subset $X' \subseteq X$. If the query is safe, the response is the *second highest* value among those associated with the elements of the requested subset. Unsafe queries are denied.

Problem 2a (4 points)

Suppose the database owner uses a naive auditing algorithm. Given a query, the naive auditor denies it if the responses to all previous queries, taken together with the response to the current query, would reveal the value associated with some element of the database X.

Give an example of a database X and a sequence of queries that, if processed by the naive auditor, completely reveals the value associated with some element of X.

Problem 2b (4 points)

Suppose the database owner has both a safe query **monitor** and a simulatable auditor at his disposal. Give an example of a database X and a sequence of queries such that the query monitor will deny the last query, but a simulatable auditor will allow it.

Problem 3 (5 points)

A significant percentage of people living in the US are uniquely identified by the (ZIP code, birthdate, gender) quasi-identifier. If an anonymized dataset contains ZIP code, birthdate, and gender attributes, it is possible to determine whether a particular person is present in the dataset simply by checking if this person's quasi-identifier occurs in one of the records.

Recall that k-anonymity relies on generalizing and suppressing quasi-identifiers until every quasi-identifier occurs in at least k records in the anonymized dataset. Therefore, given any person's record, it is guaranteed that the same quasi-identifier occurs in the records of at least $k-1$ other people.

A common interpretation of this property is that k-anonymity hides whether a particular person is present in the anonymized dataset.

Assuming that the attacker knows only people's quasi-identifiers, is this interpretation accurate? Explain.

Problem 4 (4 points)

Differential privacy is closely related to the concept of function *sensitivity*. Give examples of low-sensitivity functions (other than those from the "Differential Privacy" paper) that one may want to compute on a database, and give a mathematical explanation of why their sensitivity is low.

Problem 5 (5 points)

Consider a deterministic algorithm for generalizing and suppressing quasi-identifiers, such as those used to achieve k-anonymity, l-diversity, t-closeness, and similar properties.

Assuming that the attacker knows only people's quasi-identifiers, can generalization and suppression be used to achieve differential privacy? Explain.

Problem 6 (5 points)

Many privacy policies talk about the "purpose" for which a sensitive piece of data may or may not be used. For example, a medical privacy policy may state that a person's DNA may be released for the purpose of making a treatment decision, but not for the purpose of approving or denying health-insurance coverage.

Consider a software program operating on sensitive data (you may assume that they are stored in program variables). Explain how techniques for information-flow control can be used to express the concept of "purpose" and to ensure that data are not released for purposes forbidden by the privacy policy.