CS590: Data Security and Privacy **Spring 2022**

Mock Mid-term Exam

Due date & time: 11:59pm on March 7, 2022.
Please hand in your homework on Brightspace. No late day allowed. This aims to guide the review process. It has more problems than the actual mid-term exam. This counts for 5% of the course grade.

Your name (Last, First):

Filling in the Blanks (11 pts, 1pt each space) Fill in the underlined empty space with a number, a word, or a short phrase.

(a) When a process with effective user id 500 executes a file that is owned by the root user and has the SETUID bit set, the new euid is _____, and the new ruid is _____. and the new saved user id is _____.

(b) **(Continuing the previous problem, but fill each blank below with one of euid, suid, ruid.)** To temporarily drop the elevated privilege, a process can set the value of _____ to that of _____.

(c) Given 3 security levels and 3 categories, the lattice of security labels contains _____ labels. Given m security levels and n categories, the lattice of security labels contains _____ labels.

(d) In the RBAC96 family of models, RBAC0 is the basic RBAC. RBAC1 adds _____ to RBAC0, and RBAC2 adds _____ to RBAC0.

(d) In the kernelized deisgn for high-assurance systems, a reference monitor should be small enough to be analyzable, in addition it should be _____ and _____.

True/False Questions (38 pts, 2 pts each) Circle yes or no. When you think the question is a bit ambiguous, you can optionally add a sentence to explain your reasoning.

yes / no To delete a file f from a directory d in UNIX, one does not need to have read or write permission on the file f.

yes / no When using capability-based systems, rather than access control list, one has the ambient authority problem, that is, all the rights that a subject has are automated used without being selected.

yes / no A trusted subject in the Bell-LaPadula model can read and write any object in the system.

yes / no Using the Bell-LaPadula model, one does not need to trust human users to ensure that information at a high sensitivity level does not flow to persons who are not cleared at that level.

yes / no In the BLP model, the Star Property requires that for a subject S that is not trusted, it can write to an object O if and only if S's current security level is no higher than O's security level.

yes / no The tranquility principle in Bell-LaPadula (BLP) model is introduced to deal with BLP model's covert channel weakness.

yes / no The Chinese Wall policy aims at achieving the least privilege security principle.

yes / no Consider the Trust Computing Base (TCB) of a system for a desired security property, the larger the TCB is, the more secure the system is.

yes / no Whether a system component is trusted or not depends on how other components interact with the the component, as opposed to the intrinsic properties of the component.

yes / no In RBAC, static mutual exclusion constraints can be used to prevent one user from having too many permissions.

yes / no In k anonymity, the larger k, the less privacy is provided.

yes / no If one assumes that all attributes are quasi-identifiers, one can still apply k-anonymity, but one cannot apply ℓ-diversity or t-closeness.

yes / no The notion of ℓ-diversity is primarily motivated by the fact that k-anonymity does not prevent attribute disclosure.

yes / no The notion of t-closeness can be viewed as simulating an ideal world of privacy where all sensitive attributes are removed and only the quasi-identifiers are published.

yes / no The notions k-anonymity and ℓ-diversity are syntactic in nature in that they are defined **only** on the output.

yes / no Achieving the concept of "Privacy as Secrecy" in data privacy is impossible even if no data utility is provided.

The following questions ask you to judge whether the concept of Neighboring Dataset offers appropriate privacy in each of the settings below.

yes / no The dataset is similar to the Netflix movie rating dataset, where the dataset is a matrix, with each cell corresponding to a user's rating of a particular movie. Two dataset are neighboring if they differ in at most one cell.

yes / no The dataset contains hospital visit records, where each record corresponds to one visit of some patient. Two dataset are neighboring if they differ in at most one record.

yes / no The dataset contains communication pattern graph data, where each node represents an individual, and each edge between two nodes represents that the two users have communicated. Two dataset are neighboring if one can be obtained from the other by removing one node and all edges associated with the node.

Non-Deducability (4 pts) Given three boolean input variables x, y, and z, each taking values in {TRUE, FALSE}. For each of following, answer whether there is information flow in the non-deducibility sense. (Circle yes or no.)

- $w = x$ OR y OR z

  yes / no there is information flow between x and w

  yes / no there is information flow between $\{x, y\}$ and w

- $w = x$ XOR y AND z

  yes / no there is information flow between x and w

  yes / no there is information flow between $\{x, y\}$ and w

Non-interference (3 pts) Explain why non-interference between certain inputs and outputs of a program cannot be satisfied by monitoring the execution of the program and denying illegal actions or terminate the program if illegal action is detected.

Integrity Models (6 pts) For each question below, choose one or more from the four policies in the Biba integrity model: (1) Strict Integrity Policy; (2) Subject Low-Water Mark Policy; (3) Object Low-Water Mark Policy; (4) Ring Policy.

- Which one policy is the main one being used in the Clark-Wilson integrity model?
- In which policy or policies are the high-level subjects trusted to be benign?
- In which policy or policies are the high-level subjects trusted to be correct in addition to being benign?
- In which policy or policies can the integrity levels of objects represent only the quality of information contained in the objects, but not the levels of importance of these objects?

Integrity (4 pts) (a) Explain why integrity differs from confidentiality in that one has to trust some subjects for integrity, but not so for confidentiality. (b) Explain why this means that one does not need to be concerned with covert channels for integrity protection.

Discretionary Access Control (3 pts) Which aspect/feature of the implementation of UNIX DAC indicate that it implicitly assumes that all programs are benign? Which aspect/feature of the implementation of UNIX DAC indicate that it implicitly assumes that all programs are correct?

Anonymization [3 pts] Describe a method in which one can generate a dataset that satisfies k-anonymity, has the same size as the input dataset, offers reasonable utility, yet offers very weak privacy protection.

Reconstruction attack (5 pts) Use your words to prove the following result (from the Dinur–Nissim Paper): Given a dataset that is a vector of N bits. For any mechanism M that can answer subset sum queries with error bounded by E. Then there exists an adversary that can reconstruct the database to within $4E$ positions.

Differential Privacy: Laplace Mechanism (9 pts) Consider a dataset of salary figures, and assume that all numbers are in the range $(0, 900K]$. Under Unbounded Differential Privacy, what is the global sensitivity of publishing each of the following?

- The number of people with salary above \$100,000: _____.
- The histogram of the number of people with salary in each of $(0, 100K]$, $(100K, 200K]$, $\cdots$, $(800K, 900K]$: _____
- The histogram of the number of people with salary in each of $(0, 10K]$, $(10K, 20K]$, $(20K, 30K]$, $\cdots$, $(880K, 890K]$, $(890K, 900K]$: _____
- Sum of total salary: _____
- The median salary: _____
- The mode, i.e., the number which appears most often in the set: : _____

Local Differential Privacy: Frequency Oracle [8 pts] We want each user to report a value that has a domain of $d = 100$ values, in a way that satisfy ϵ-local differential privacy for $\epsilon = \ln 4$.

- (2 pts) When using generalized randomized response, what probability should one report the value without change? Express your answer using a common fraction.

- (2 pts) When using generalized randomized response, suppose that each value is preserved with probability p. If a server collects 100000 responses, among which 3000 has a particular value, what is the best estimate of the number of respondents who actually have that value? Express your answer using a formula involving p.

- (2 pts) When using unary encoding, each value is encoded using a 100-bit string with one bit being 1 and the other bits being 0. Every bit is randomly perturbed independently before being transmitted. When using the basic Rappor protocol, what is the probability that a 1-bit is not changed? What is the probability that a 0 bit is not changed?

- (2 pts) When using the Optimized Unary Encoding protocol, what is the probability that a 1-bit is not changed? What is the probability that a 0 bit is not changed?

Local Differential Privacy: Heavy Hitter Discovery [6 pts] Assuming that you are designing a LDP protocol to identify around 150 most frequent values using the Prefix Extending protocol. The input domain is 15 bytes, and you want to limit the total number of frequency oracle queries to no more than 2^{28}. How to use LDP frequency oracles to achieve high accuracy.