Data Security and Privacy

Review for Final Exam

About 1/3 Will be From the Following Topics

- Topic 4: DAC Weakness
- Topic 5: Bell LaPadula Model
- Topic 6: Integrity models
- Topic 8: Role-Based Access Control
- Topic 10: Database Access Control
- Topic 14: Authentication and Key Establishment Protocols
- Topic 17: Non-interference and non-deducability

About 2/3 Will be From the Following Topics

- Topic 18: k-Anonymity, l-Diversity, t-Closeness
- Topic 19: Local Differential Privacy (using slides from Apr 12)
- Topic 20: Differential Privacy
- Topic 21: Publishing Histograms
- Topic 22: Meanings and caveats of DP
- Topic 23: Publishing Marginals
- Topic 24: Fully Homomorphic Encryption

K-Anonymity

- k-Anonymity
  - Attributes are separated into Quasi-identifiers (QIDs) and Sensitive Attributes (SAs)
  - Each record is indistinguishable from $\geq$ k-1 other records when only "quasi-identifiers" are considered
  - k-Anonymity ensures that linking cannot be performed with confidence > 1/k.
- Achieved by Generalization (Replace with less-specific but semantically-consistent values) and Suppression (Removing certain records)
- Weaknesses:
  - Doesn't prevent attribute inference
  - Syntactical, can be trivially satisfied.

L-Diversity

- The l-diversity principle
  - Each equivalent class contains at least l well-represented sensitive values
- Weaknesses:
  - difficult and unnecessary to achieve
  - Vulnerable to skewness and similarity attack
  - Ignore semantic meanings of attribute values

T-Closeness

- Principle: Distribution of sensitive attribute value in each equi-class should be close to that of the overall dataset (distance $\leq$ t)
- Uses Earth Mover Distance to capture semantic meaning of values
- Consider suppresion of all QI attributes as ideal world
- Limitation: requires specification of QI and sensitive attributes; vulnerable to inferences that use knowledge about algorithm

Towards Differential Privacy

- Syntactic versus Algorithmic Privacy Notions
- Definition: A mechanism *A* satisfies ε-Differential Privacy if and only if
  - for any neighboring datasets D and D'
  - and any possible transcript $t \in$ Range(A),
    $$\Pr[A(D) = t] \leq e^{\epsilon} \Pr[A(D') = t]$$
  - For relational datasets, typically, datasets are said to be neighboring if they differ by a single record.
- Differential between bounded and unbounded DP
- Impossibility of "Privacy as Secrecy"

Mechanisms for Satisfying DP

- Laplace Mechanism:
  - Understand global sensitivity of different kinds of queries
  - Understand Laplace distribution
- Exponential Mechanism not Required
- Understand Sequential Composability, Parallel Composability, Post-processing Invariance
- Intuitive understanding of what queries are easy and what queries are hard
- Four settings of using DP

Publishing Histograms

- Uniform grid, Adaptive grid
- Tradeoff between noise error and non-uniformity error
- PrivPFC
  - Criteria for choosing a grid

Meaning and Caveats of DP

- Personal Data Principle
- Caveats of applying DP
  - How neighboring datasets is defined?
  - What constitutes an individual’s data
  - One individual’s data or personal data under one individual’s control
  - Group privacy
  - Moral challenge
  - Choosing epsilon value
  - Learning models and applying to individuals

Publishing Marginals

- PriView
  - Goals: Marginal Queries
  - Direct method,
  - Flat method
  - Understand the middle ground's advantage
  - 4 steps in PriView
  - Other methods not required
- Membership privacy not required

Local Differential Privacy

- Difference from centralized setting in trust
- Generalized randomized response protocol (Direct Encoding)
- Unary encoding protocol (Basic Rappor)
- Optimized unary encoding, difference from above
- Frequent itemize mining
  - Padding and sampling
  - Two phases

Fully Homomorphic Encryption

- Homomorphic property of RSA and El Gamal
- Concept of somewhat homomorphic encryption
- The bit encryption nature of FHE schemes

Coming Up

- Final Exam