A Cryptography-Flavored Approach to Privacy in Public Databases

Drineas, Dwork, Goldberg, Isard, Redz, Smith, Stockmeyer

Think “Census”

- Method for sanitizing a database
  - Meaningful statistical analysis
  - Preservation of individuals’ privacy
- What do we mean?

“Privacy” in English

- Protection from being brought to the attention of others [Gavison]
  - inherently valuable
  - attention invites further privacy loss, eg info
- One’s privacy is maintained to the extent that one blends in with the crowd.
- Crowd size exceeds threshold T

Focus on Geometric Data

- Real database (RDB) consists of n points in d-dimensional space (say, unit ball)
  - points are unlabeled
- Publish sanitized database (SDB)
  - candidate sanitization procedure (later)

Adversary: The Isolator

- Inputs to a c-isolator:
  - SDB
  - auxiliary information z
- Output $I(SDB, z) = q \in \mathcal{B}$
- Success occurs if

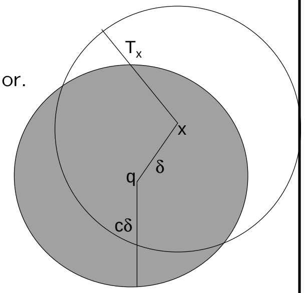
$$|B(q, c\delta) \cap RDB| \le T$$

where δ is distance from q to closest RDB point

Relative Notion of Isolation

Lemma:
Let $T_x =$ distance from x to T^{th} neighbor.
If $B(q, c\delta)$ contains $\le T$ RDB points
then $\delta = D(q, x) \le T_x/(c-1)$.
(So q is "close" relative to T_x.)

Proof:
$c\delta < T_q$ and $T_q \le \delta + T_x$



Isolation Does Not Imply Failure of Sanitization

- Cynthia publishes her point p on web
  - I(SDB,Cynthia's web site) = p
  - δ = 0 and ball of radius cδ contains only one RDB point
- Not the fault of the sanitization procedure!
  - I'(Cynthia's web sit) = p

Cryptographic Flavoring

- SDB shouldn't help the isolator "too much"
- Definition of "not too much" should be fairly forgiving, eg, advantage obtained from seeing the SDB may be, say, $n^{1+\varepsilon}$

Candidate: Effective Sanitization

$\forall^* z \forall \mathcal{D} \forall I \exists I'$ whp over RDB $\in_R \mathcal{D}$:

$$\Pr[I(SDB, z)] - \Pr[I'(z)] \leq n^{-(1+\varepsilon)}$$

Alternatively, worst case over RDBs:
$\forall z \forall I \exists I' \forall \text{RDB} \ldots$

*Need to constrain z somewhat.

Distribution on Databases?

- Don't want to deal with crypto-like definitions, in which, say, sum of every 7th elements is congruent to 23 mod 51
- Take statistician's approach: each point in the RDB is an independent sample from a single fixed distribution

Candidate Sanitization Procedure

- For each x ∈ RDB
  - □ Find T_x = distance to T^{th} nearest neighbor
  - □ Choose x' $\in_R$ $B(x,T_x)$
- Complements definition of c-isolation
  - □ if q c-isolates x then $D(q,x) \leq T_x/(c-1)$
  - □ consequence: high dimensionality is our friend
- Intuition:
  - □ perturb minimally to prevent isolation
  - □ outliers randomized to oblivion
    - kills isolated anomalies, maintains group anomalies

Meaningful Statistical Analysis

- Dream: find a large class of algorithms that "perform well" on sanitized data
- Start with clustering
  - □ clusterings have measures of quality (diameter, conductance, etc.)
  - □ See how measures are preserved
    - under sanitization
    - under de-sanitization