Toward Privacy in Public Databases

Shuchi Chawla (CMU)
Cynthia Dwork (Microsoft Research)
Frank McSherry (Microsoft Research)
Adam Smith (Weizmann Institute of Science)
Hoeteck Wee (UC Berkeley)

To appear at TCC 2005.
http://theory.csail.mit.edu/~asmith

Database Privacy

- Census problem
  - Individuals provide information
  - Census office publishes sanitized records
    - Allow extraction of global statistics
    - Protect individuals' privacy
- Inherent Privacy vs Utility Tradeoff
  - Extremes: Publish nothing, publish everything
- Goals:
  - Find a middle path: both privacy and utility
  - Hope: change the way privacy is approached
    - Framework for meaningful comparison of techniques
    - Encourage debate of what "privacy" means

Database Privacy

- Utility:
  - Users can extract "global statistics"
    - Means, variances, approximate clusters, …
  - Proof by algorithm + analysis
- Privacy:
  - What is required?
  - How to prove it?

Outline

- What do we mean by "Privacy"?
  - Geometric abstraction
  - Privacy breach $\approx$ "isolation"
- Example Sanitizations
- Conclusions and Future Work

Current solutions

- Extensively studied in statistics, data mining
  - Non-interactive: Suppress/aggregate cells, perturb data, synthesize new data, …
  - Interactive: monitor queries, perturb outputs
- Focus on utility
- Privacy claims unsatisfying *
  - Ad-hoc or unclear definitions
  - Unexpected leaks, e.g.
    - Erasure / refusal to answer can reveal info
    - Noise can cancel in interactive queries
  - Debate / criticism is difficult

* Recent exceptions: DN03,DN04,EGS03

Cryptographer's Approach

First:

- Define "privacy" in this context
  - "Privacy" is an overloaded term
  - How can we get a handle on it?

Second:

- Understand what kinds of information do – and don't – breach privacy

What do WE mean by privacy?

- Privacy is an overloaded term
  - What does it mean for databases?
- Intuition:
  privacy = blending into the crowd
- [Ruth Gavison] "Protection from being brought to the attention of others"
  - Inherently valuable
  - Attention invites further privacy loss
  - Also "chilling effect" on rights and speech
- Appealing definition; can be converted into a precise mathematical statement

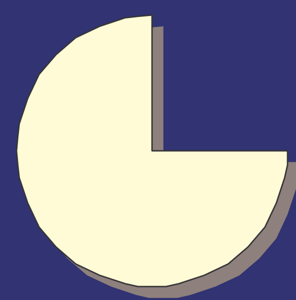
A Geometric Abstraction

- Database = vectors in metric space (e.g $\mathbb{R}^d$)
  - Points are unlabeled
    You are your collection of attributes
- Distance is everything
  - Points are similar if and only if they are close
- Highly abstracted version of problem
  - If we can't understand this,
    we can't understand real life
  - Assumption implicit in current literature
- For this talk: $\mathbb{R}^d$, with L_2 distance and large d

A Geometric Abstraction

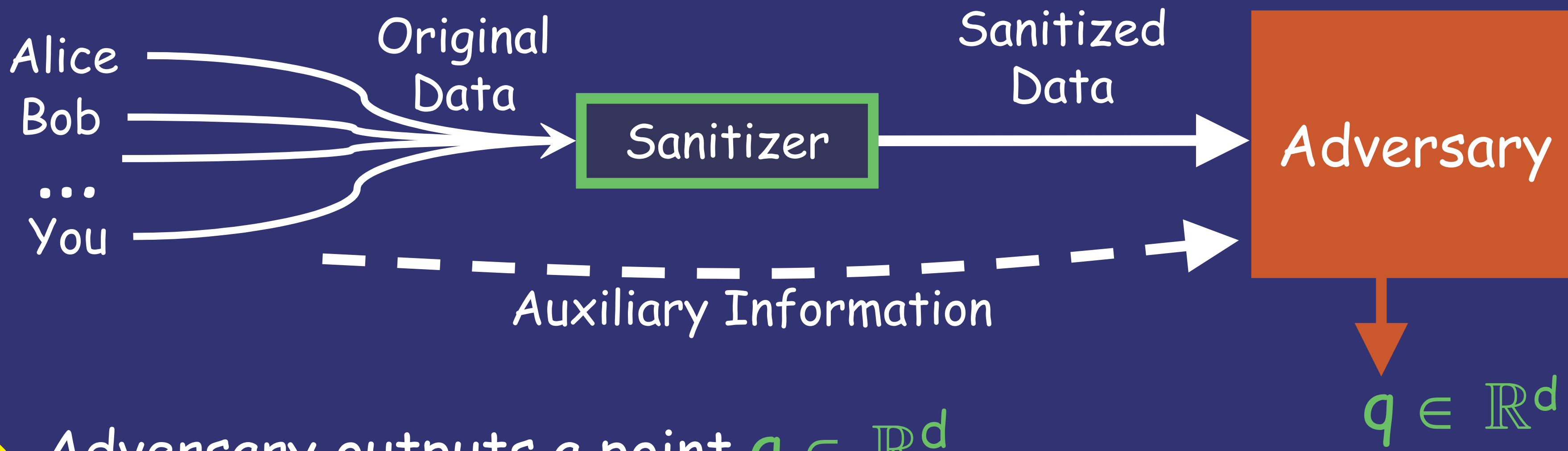
n points in $\mathbb{R}^d$

Charts, tables, etc released by census

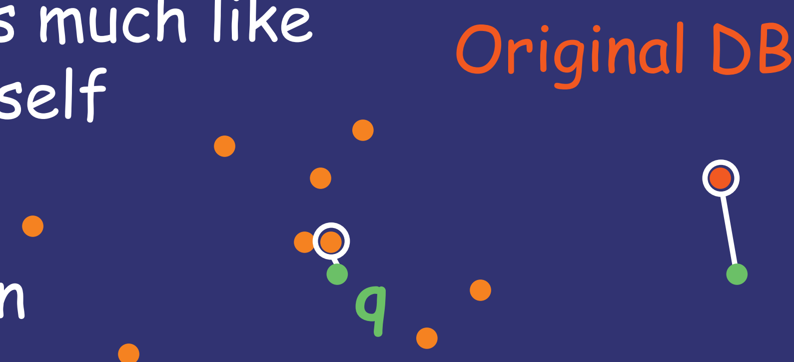


(non-interactive)

The Adversary as Isolator - Intuition



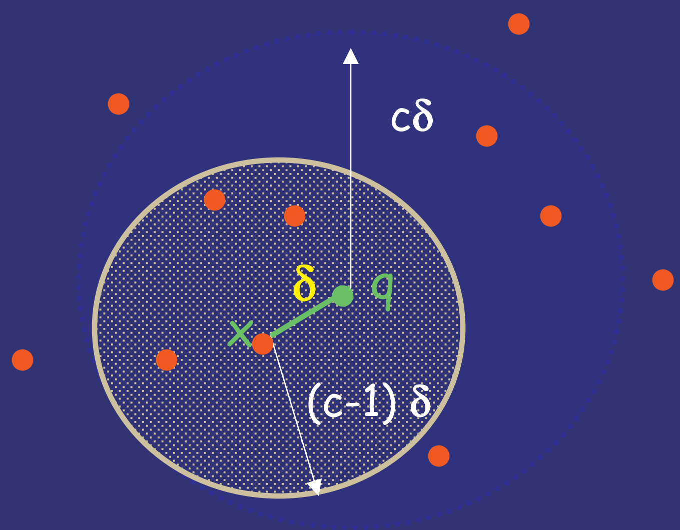
- Adversary outputs a point $q \in \mathbb{R}^d$
- q "isolates" an original DB point x, if it is much closer to x than to x's near neighbors
- q fails to isolate x if q looks as much like x's neighbors as looks like x itself
- Tightly clustered points have smaller radius of isolation



Isolation – the definition

- I(Sanitized DB,aux) = q
- x is isolated if $B(q,c\delta)$ contains fewer than T other points from Original DB
- T-radius of x – distance to its T^{th}-nearest neighbor
- x is "safe" if δ_x > (T-radius of x) / (c-1)

  $B(q,c\delta_x)$ contains x's entire T-neighborhood



c – privacy parameter; eg, 4

Requirements for the sanitizer

- Intuition: side info may allow isolating points apriori
  - Emulate definition of semantic security of encryption
- Sanitization breaches privacy if giving the adversary access to the SDB considerably increases its probability of success
- Forgiving def: "Considerably" $\approx 1/n^{1/2}$, or 1/1000
- Roughly: For a particular distrib. $\mathcal{D}$ on DB and aux: $\forall$ I, $\exists$ "simulator" I' , w. high pr. over $\mathcal{D}$,

  Pr[I(San.DB,aux) isolates pt.]
  — Pr[I'(aux) isolates pt.] < ε
- Framework for measuring sanitization methods

What About Utility?

- "Pointwise" approach: we prove that specific functionalities can be learned
  - averages, medians, clusters, singular value decomposition,…

For now…

- Goal: large class of interesting tests for which there are good approximation procedures using sanitized data
  - Work in progress
  - Ideal: everything learnable with "noise" is learnable privately

Outline

- ✓ What do we mean by "Privacy"?

- Example Sanitizations (non-interactive)
  - Recursive Histogram - privacy
  - Density-based Perturbation - utility
  - Hybrid: Cross-training

Use local density

- Conclusions and Future Work

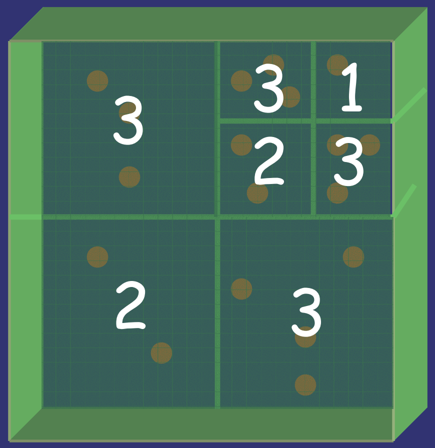
Histogram Sanitization

- $U = [-1\ ,\ 1]^d$
  - d-dim cube, side = 2
- Cut into 2^d subcubes
  - split along each axis
  - subcubes have side = 1
- For each subcube

  if number of RDB points > 2T

  then recurse
- Output: list of cells and counts

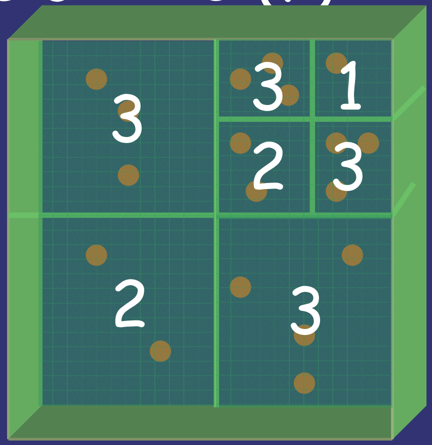
  E.g. "The subdivisions were … .
  Cell 1 had 3 points, Cell 2 had 2 points, …"



Histogram Sanitization

Theorem: Recursive histograms are safe
if database points uniform in $[-1,1]^d$

- Pr[I(SDB) c-isolates] $\leq 2^{-\Omega(d)}$, where c $\approx$ 10 (?)

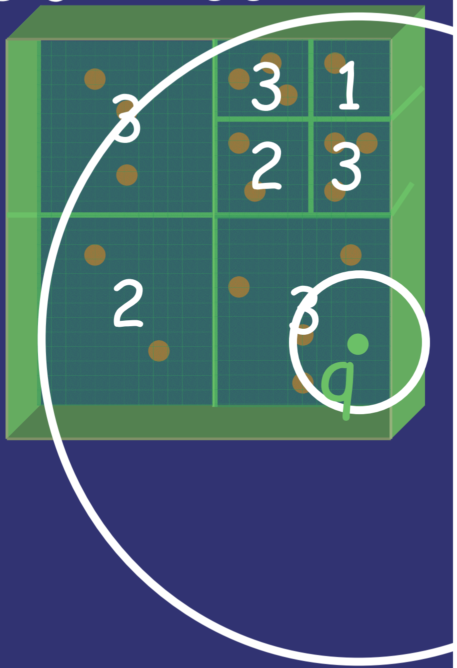


- Strong assumptions!
  - Specific distribution
  - No auxiliary information
- Assumptions can be relaxed…

Histogram Sanitization

Theorem: Recursive histograms are safe
if database points uniform in $[-1,1]^d$

- $\Pr[\, I(SDB) \text{ c-isolates}] \leq 2^{-\Omega(d)}$, where $c \approx 100^*$



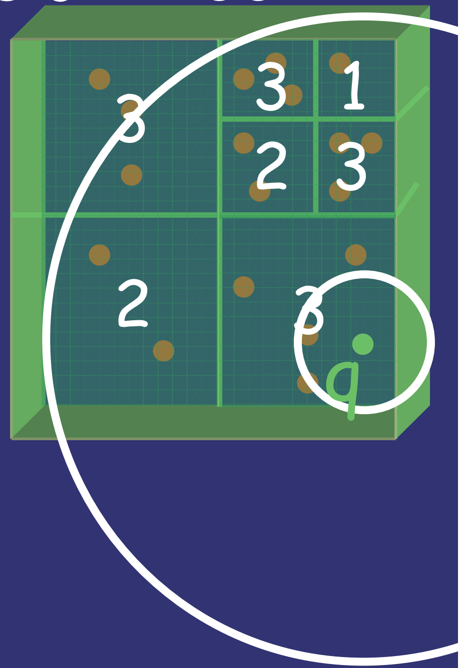
- If $n = 2^{o(d)}$, proof is simple:
  - Let adversary pick q,
  - s = side-length of sub-cube C of q
  - Dist. to nearest point = $O(s\, d^{\frac{1}{2}})$ w.h.p.
  - Increasing the distance by c captures C and most of its parent cell
  - Parent of C contains 2T points $\Rightarrow$ q doesn't c-isolate anyone

Histogram Sanitization

Theorem: Recursive histograms are safe

if database points uniform in $[-1,1]^d$

- Pr[I(SDB) c-isolates] $\leq 2^{-\Omega(d)}$, where $c \approx 100^*$



- If $n = 2^{o(d)}$, proof is simple
- If $n = 2^{\Omega(d)}$, proof is harder…

For Very Large Values of n

- Wlog can switch to ball adversaries: (q,r)

  I wins if B(q,r) contains at least one RDB point and B(q,cr) contains fewer than T RDB points

- Define a probability density f(x) that captures adversary's view of the RDB

  Ball Lemma: To win with probability ε, I needs:

  $$Pr_f[B(q,r)] \geq \varepsilon/n$$

  $$Pr_f[B(q,cr)] \leq 2T/n$$

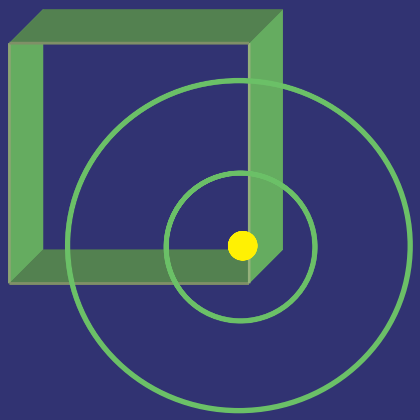
  $$Pr_f[B(q,r)] \,/\, Pr_f[B(q,cr)] \geq \varepsilon/2T$$

- Bound ε by bounding ratio by $2^{-\Omega(d)}$

Bounding $Pr_f[B(q,r)]/Pr_f[B(q,cr)]$

Inflation Lemma: If B = ball with small* radius, C = cube $[-1,1]^d$,

$$\frac{\mathrm{Vol}(B \cap C)}{\mathrm{Vol}(2B \cap C)} \leq 2^{-\Omega(d)}$$



Proof (outline):

- Approximate uniform over ball by Gaussian
- Crunch numbers

(Nicer proof?)

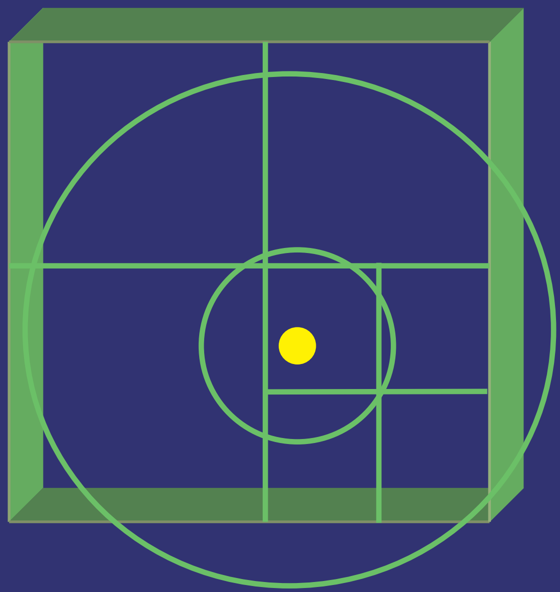
* Small = $\beta d^{\frac{1}{2}}$side-length(C), where $\beta \approx 1/60$

Bounding $Pr_f[B(q,r)]/Pr_f[B(q,cr)]$

- $f(x) = (n_C/n)\ (1\ /\ Vol(C))$
  - fraction of RDB points landing in cell C, spread uniformly within C
- If r is small, Inflation Lemma says that big ball captures exp(d) more mass in each subcube it touches than small ball

Thus,
  - Total mass increases exponentially

$\Rightarrow$ Ratio is small



yields $\varepsilon < 2^{-\Omega(d)}$

Bounding $Pr_f[B(q,r)]/Pr_f[B(q,cr)]$

- $f(x) = (n_C/n)\ (1 / Vol(C))$
  - fraction of RDB points landing in cell C, spread uniformly within C
- If r is small, Inflation Lemma says that bigger ball captures exp(d) more mass in each subcube than smaller ball
- If r is large, the small ball captures nothing or the bigger ball captures parent cube
- Either way isolation cannot occur ($c \approx 100$? 10?)

Relaxing Assumptions

- Extends to many interesting cases
  - non-uniform but bounded-ratio density fns
  - isolator knows constant fraction of attribute vals
  - isolator knows lots of RDB points
  - isolation in few attributes
    (very weak bounds)

- Can be adapted to "round" distributions

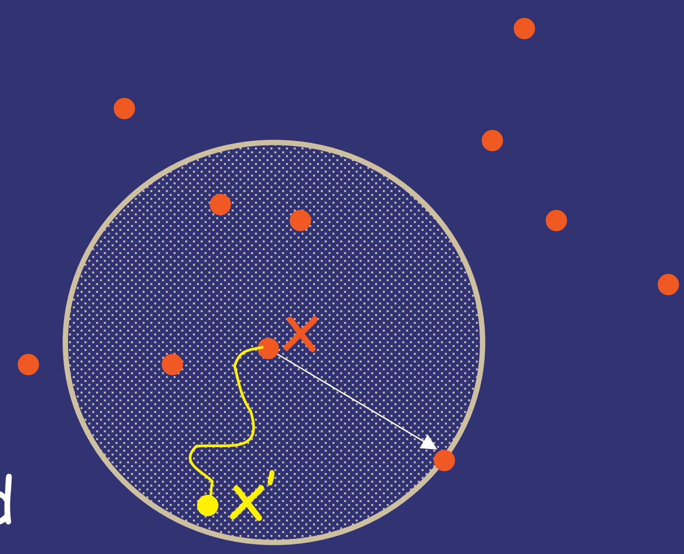
  balls, spheres, mixtures of Gaussians,
  with effort; [work in progress]

Outline

- What do we mean by "Privacy"?
- Example Sanitizations (non-interactive)
  - Recursive Histogram - privacy
  - Density-based Perturbation - utility
  - Hybrid: Cross-training
- Conclusions and Future Work

Round Sanitizations

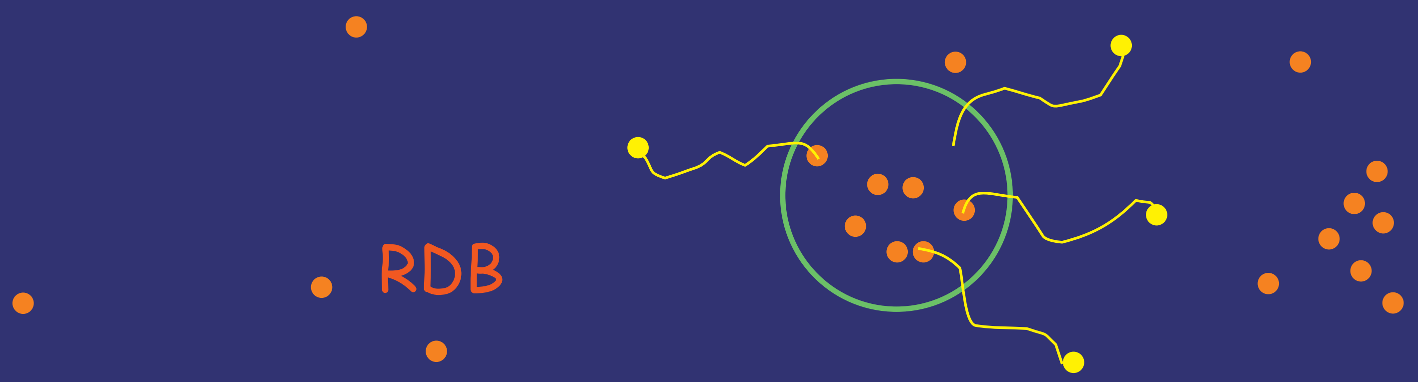
- The privacy of x is linked to its T-radius
  $\Rightarrow$ Randomly perturb it in proportion to its T-radius
- $x' = San(x) \in_R B(x, T\text{-}rad(x))$
  - alternatively, N(x, T-rad(x)), d-dim Gaussian
- Intuition:
  - Blend x in with its crowd
  - Adding random noise with mean zero to x,
    $\Rightarrow$ means, correlations should be preserved.



T=4

Round Perturbations Provide Utility

Distributional/ Worst-case	Objective	Assumptions	Result
Worst-case	Find K clusters minimizing largest diameter	-	Diameter increases by a factor of 3
Distributional	Find k maximum likelihood clusters	Mixture of k Gaussians	Spectral clustering is correct w.h.p. when centers are well separated

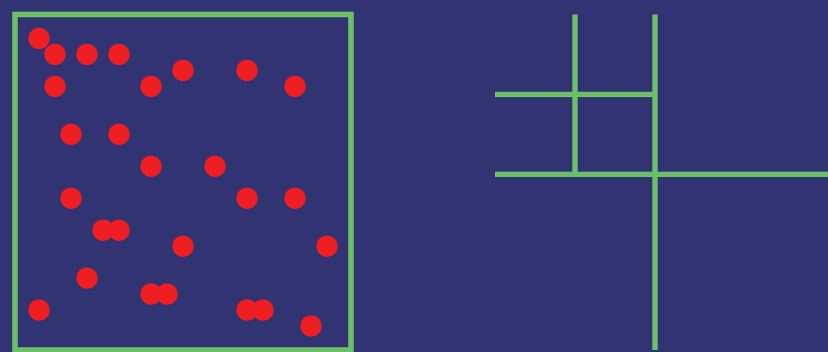


Privacy for n Sanitized Points?

- Given n-1 points in the clear, the probability of isolating the nth is $\exp(-\Omega(d))$
- Intuition for extension to n points is wrong!
  - Privacy of x_n given x_n' and n-1 points in the clear does not imply privacy of x_n given all n sanitized points!
  - Sanitization of other points reveals information about x_n
  - Worry is for safety of the reference point (the neighbor defining the T-radius), not the principal

Combining the Two Sanitizations

- Partition RDB into two sets *A* and B
- Cross-training
  - Compute histogram sanitization for B
  - $v \in A$: ρ_v = side length of *C* containing v
  - Output GSan(v, ρ_v)



Cross-Training Privacy

- Privacy for B: only histogram information about B is used
- Privacy for A: harder version of proof for histograms
  - so far, proof works only for $|A| = 2^{o(d)}$
- Immediate Next Goals:
  - Extend privacy proofs to more distributions
  - Not all utility results have carried over
    - Spectral techniques work; not all clustering does

Outline

- ✓ What do we mean by "Privacy"?
- ✓ Example Sanitizations (non-interactive)
  - Recursive Histogram - privacy
  - Density-based Perturbation - utility
  - Hybrid: Cross-training
- Conclusions and Future Work

Future Research (Abstract Model)

- This talk: Abstract Model
  - Many interesting questions remain
  - Strengthen existing results?
- Work in Progress
  - [DwNa +]
    - Impossibility of all-purpose sanitizers
    - Interesting utilities that have no privacy-preserving sanitization (cf. why secure protocols don't suffice)
  - [DuSm +]
    - Low-dimensional data: combining our perspective with techniques from statistics ("density estimation")
  - [NiSm +]
    - Extending approach to categorical data (no distance?)

What About the Real World?

- Lessons from the abstract model
  - We can prove meaningful statements
  - High dimensionality is our friend
    - Treat data as whole (not component-wise)
  - E.g.: we can bound re-identification risk

- Moving towards real data
  - Problem: Why Euclidean distance?
    - Rescale coordinates, use other metrics…
    - Addressed by follow-up work
    - Easy…
  - Problem: Auxiliary information
    - What happens when adversary knows other databases?
    - Hard (provably impossible in general)

What About the Real World?

- Hard to provide good sanitization in the presence of arbitrary auxiliary info
  - Provably impossible in general
  - Suggests we need to control aux
- How to quantify what adversary knows?
  - "Smoothness"?
- How should we redesign the world?
  - Leave data in hands of users
  - Dwork: "Our Data, Ourselves"
  - Aggarwal et al: "Privacy for the Paranoids"

Conclusions

- Goals:
  - Cryptographer's approach to database privacy
  - Proposed formalism for abstract problem
  - Concrete sanitizations, results
    - Statistical / algorithmic techniques
- Many challenges remain
  - Bring approach closer to real world
- Merits attention of wider community