Accessing Data while Preserving Privacy

Kobbi Nissim

Georgetown University and CRCS@Harvard

Based on joint work with Georgios Kellaris (Harvard and Boston University), George Kollios (Boston University) and Adam O'Neill (Georgetown University)

DIMACS Workshop on Outsourcing Computation Securely

July 6 – 7, 2017

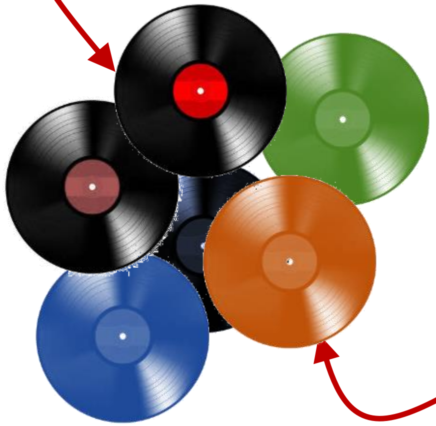
Outsourced database systems

I need all records of clients named "Gina" — Point query

… clients whose age is between 32 and 52 — Range query

… clients with Sex = M — 1-way attribute query

… clients with Sex = M *and* Married = F — 2-way attribute query



Name	ZIP	Sex	Age	Balance
George	52525	M	32	20,012
Gina	02138	F	30	80,003
⋮	⋮	⋮	⋮	⋮
Greg	02246	F	28	20,500

Search keys

Outsourced database systems

Dealing with this database myself is so tiring!

Delegate your data to me!

Outsourced database systems

Delegate your data to me!

But, I can't trust you with my customers' personal information!

We will use crypto!

* In this talk we only consider privacy (not correctness)

We have the power

Great! Can we use SFE [Yao ’82, GMW ‘84], ORAM [Gol ’87, GO ‘96], FHE [Gen 09], computational PIR [KO 97], searchable encryption [Song, Wagner, Perrig ‘01], …

This is the real world

Great! We can use SFE [Yao ’82, GMW ‘84], ORAM [Gol ’87, GO ‘96], FHE [Gen 09], computational PIR [KO 97], searchable encryption [Song, Wagner, Perrig ‘01], …

I’m convinced

Hell, no!

We should use a system that is secure and practical!

I will use order preserving and deterministic encryption* schemes

* Kobbi’s plea: Let’s call these *encodings* instead of encryptions

This is the real world

- Implemented systems use relaxed notions of encryption
  - Allows use of existing database indexing mechanisms → efficient querying
- Examples: CryptDB [PRZB’11], Cipherbase [ABEKKRV’13], …
- Security/privacy not well understood

- Attacks exist:
  - Utilizing leaked access pattern and auxiliary info about data: [Hore, Mehrotra, Canim, Kantarcioglu ’12] [Islam, Kuzu, and Kantarcioglu ’12], [Islam, Kuzu, Kantarcioglu ‘14], [Naveed, Kamara, Wright ’15]
  - Utilizing leaked access pattern: [Dautrich, Ravishankar ’13], [KKNO ‘16]

Is this just fantasy?

Great! We canuse SFE [Yao ’82, GMW ‘84], ORAM [Gol ’87, GO ‘96], FHE [Gen 09], computational PIR [KO 97], searchable encryption [Song, Wagner, Perrig ‘01], …

Great idea!

We will protect not only the access pattern, but all aspects of the computation!

Leaked communication volume

00101

001010010110

01101

110101

Oh! This shouldn't be a problem!

I'm making uniformly random range queries

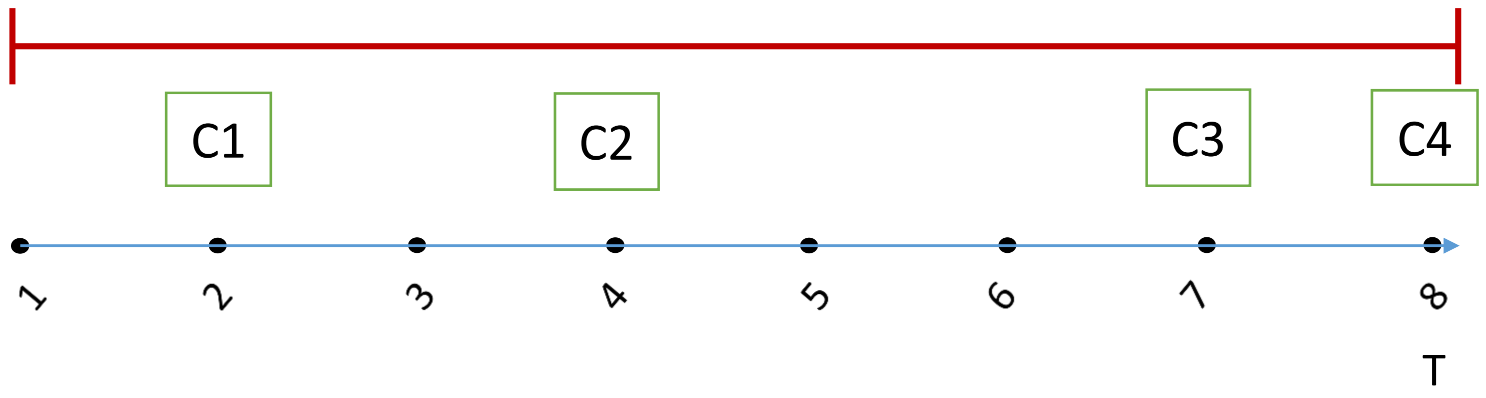
2 records

1 record

An exact reconstruction attack based on communication volume

Recovering positions:

- Find # queries (out of $\binom{T}{2} + T$) that return i records
  - Can be well estimated given O(T^4) queries

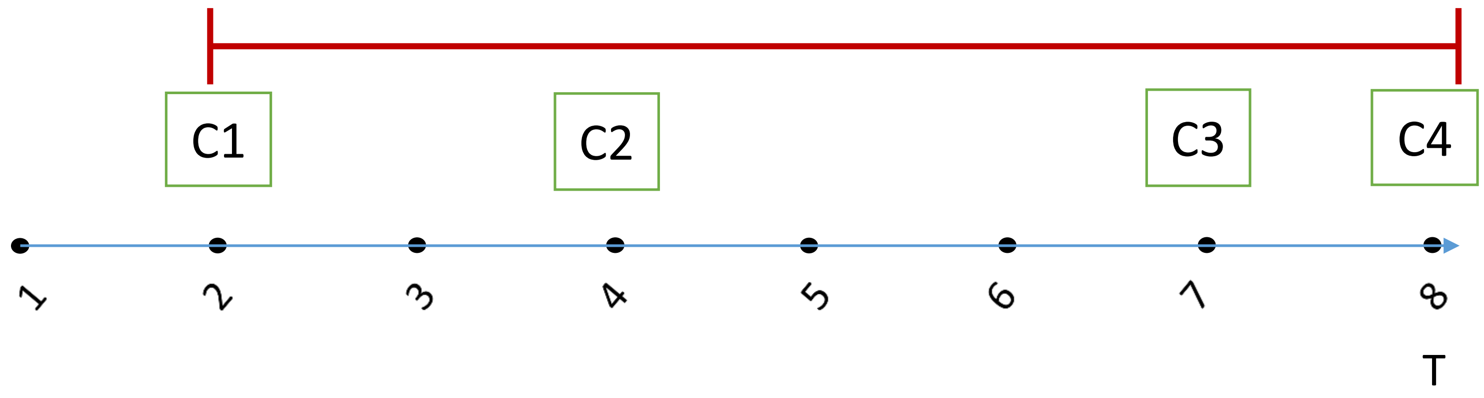


# records	# queries
4	
3	
2	
1	
0	

An exact reconstruction attack based on communication volume

Recovering positions:

- Find # queries (out of $\binom{T}{2} + T$) that return i records
  - Can be well estimated given O(T^4) queries

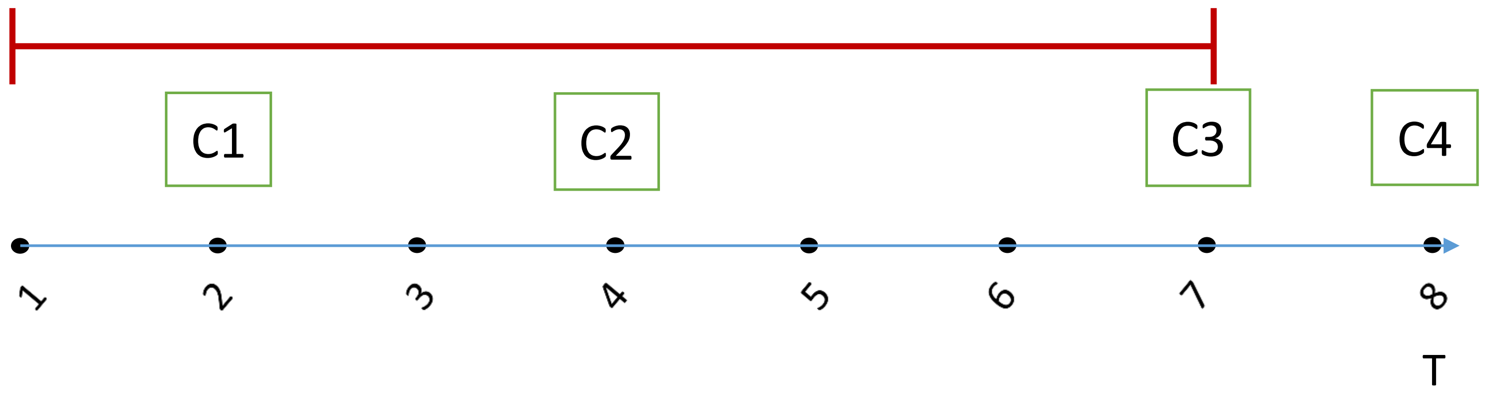


# records	# queries
4	2
3	
2	
1	
0	

An exact reconstruction attack based on communication volume

Recovering positions:

- Find # queries (out of $\binom{T}{2} + T$) that return i records
  - Can be well estimated given O(T^4) queries

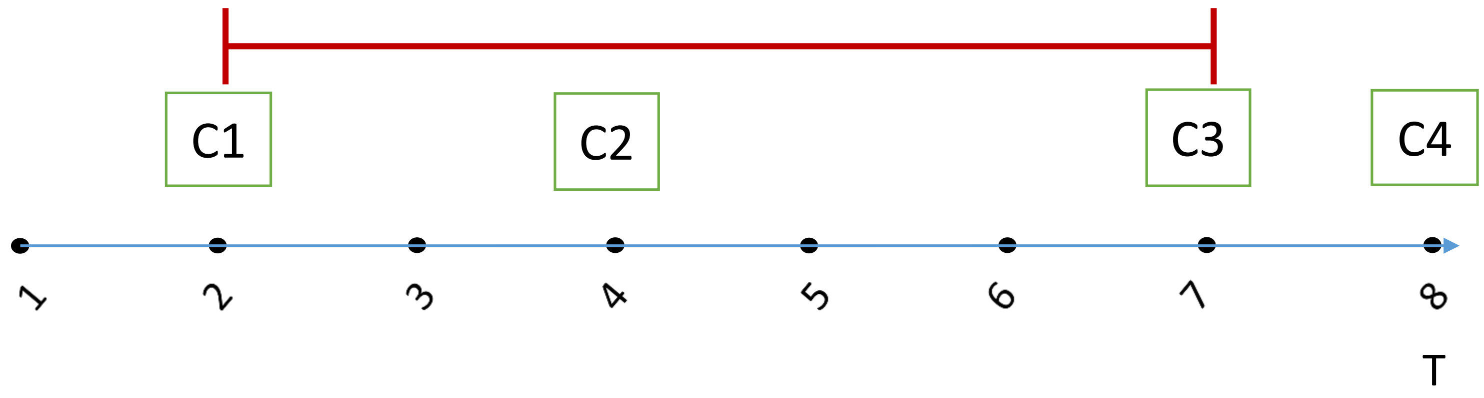


# records	# queries
4	2
3	
2	
1	
0	

An exact reconstruction attack based on communication volume

Recovering positions:

- Find # queries (out of $\binom{T}{2} + T$) that return i records
  - Can be well estimated given O(T^4) queries

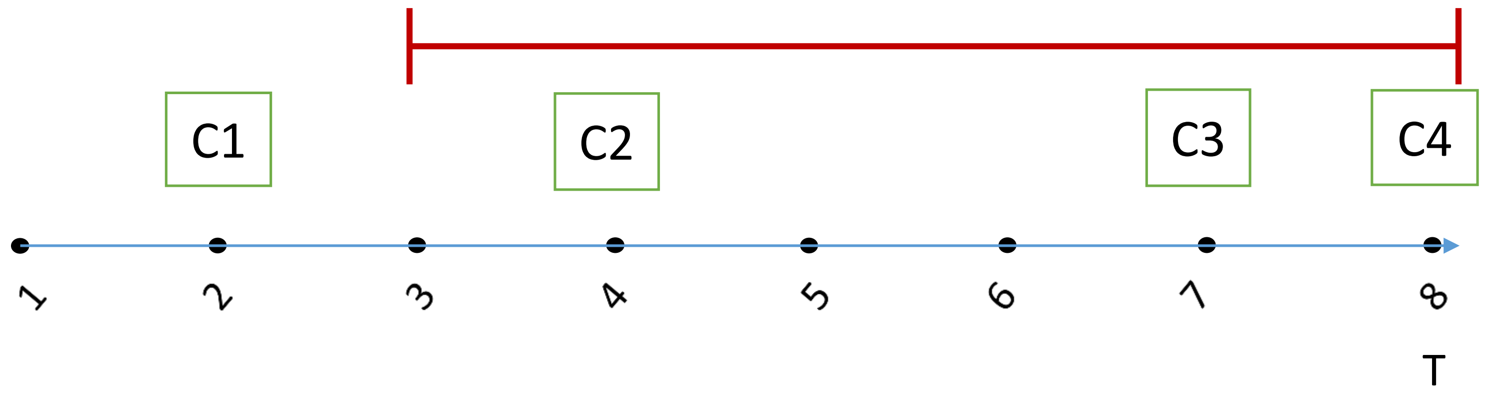


# records	# queries
4	2
3	
2	
1	
0	

An exact reconstruction attack based on communication volume

Recovering positions:

- Find # queries (out of $\binom{T}{2} + T$) that return i records
    - Can be well estimated given O(T^4) queries

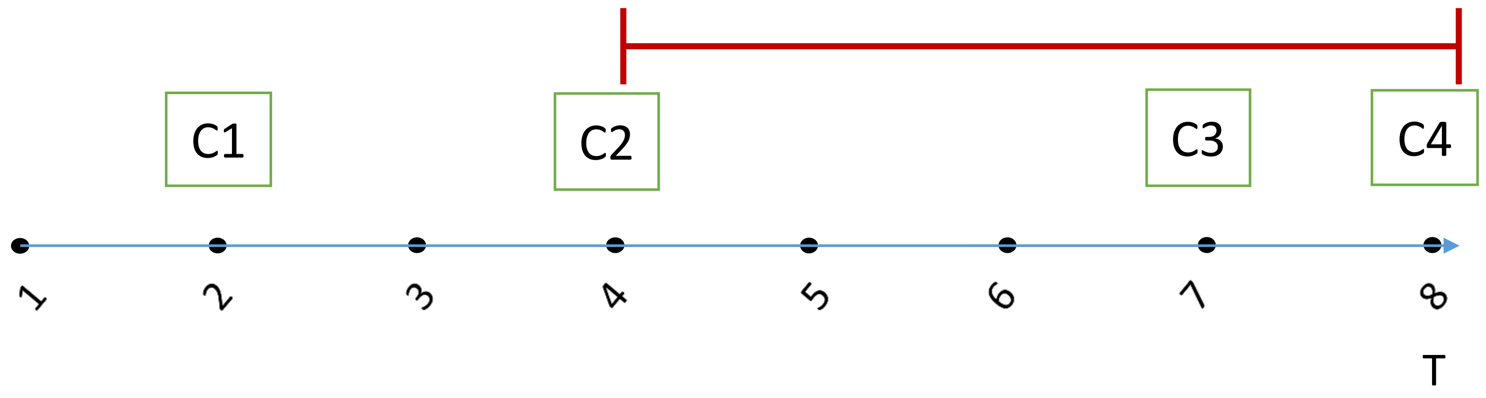


# records	# queries
4	2
3	
2	
1	
0	

An exact reconstruction attack based on communication volume

Recovering positions:

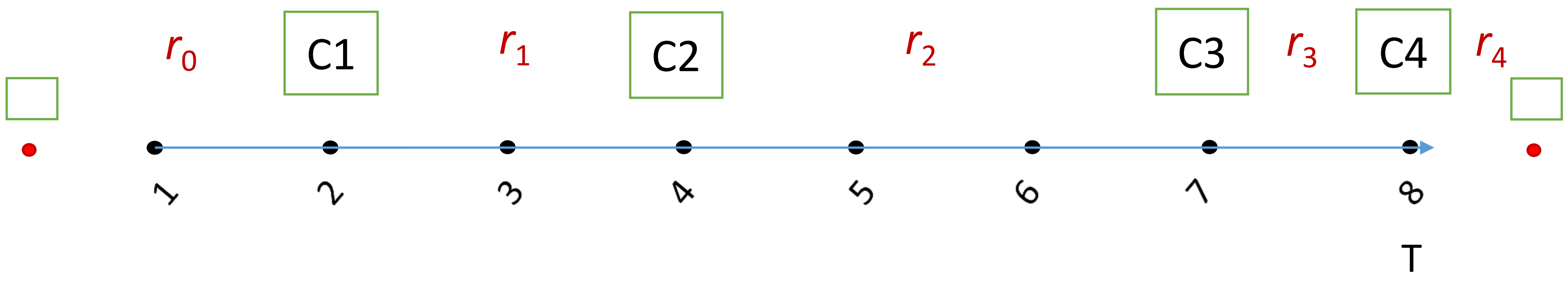
- Find # queries (out of $\binom{T}{2} + T$) that return i records
  - Can be well estimated given $O(T^4)$ queries



# records	# queries
4	2
3	4
2	11
1	14
0	5

An exact reconstruction attack based on communication volume

Recovering positions:



# records	# queries
4	2
3	4
2	11
1	14
0	5

An exact reconstruction attack based on communication volume

Recovering positions:

- We get:

$$r_0 \cdot r_4 = f_4$$
$$r_0 \cdot r_3 + r_1 \cdot r_4 = f_3$$
$$r_0 \cdot r_2 + r_1 \cdot r_3 + r_2 \cdot r_4 = f_2$$
$$r_0 \cdot r_1 + r_1 \cdot r_2 + r_2 \cdot r_3 + r_3 \cdot r_4 = f_1$$

- Let

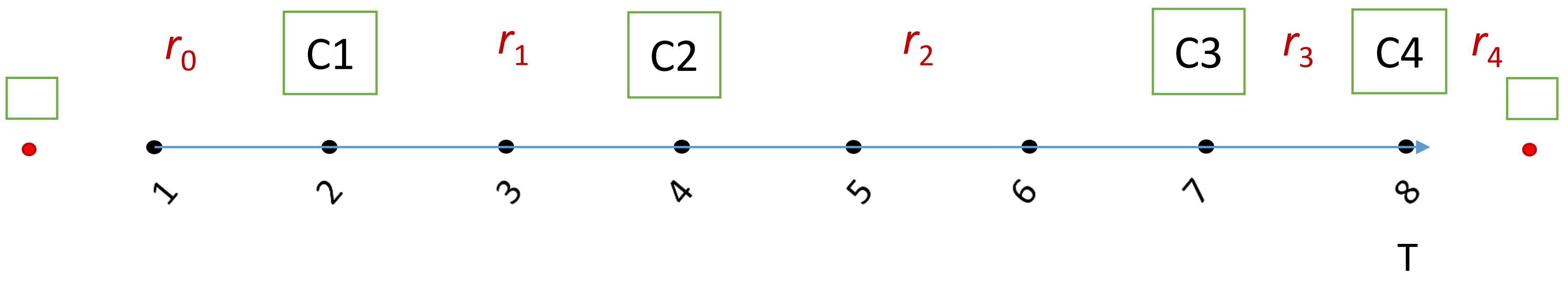
$$r_0^2 + r_1^2 + r_2^2 + r_3^2 + r_4^2 = 2c_0 + T + 1 = f_0$$

- Note:

$$r(x)\, r^R(x) = f_4 + f_3 x + f_2 x^2 + f_1 x^3 + f_0 x^4 + f_1 x^5 + f_2 x^6 + f_3 x^7 + f_4 x^8 = F(X)$$

- Define:

$$r(x) = r_0 + r_1 x + r_2 x^2 + r_3 x^3 + r_4 x^4$$
$$r^R(x) = r_4 + r_3 x + r_2 x^2 + r_1 x^3 + r_0 x^4$$



# records	# queries	
4	2	f_4
3	4	f_3
2	11	f_2
1	14	f_1
0	5	c_0

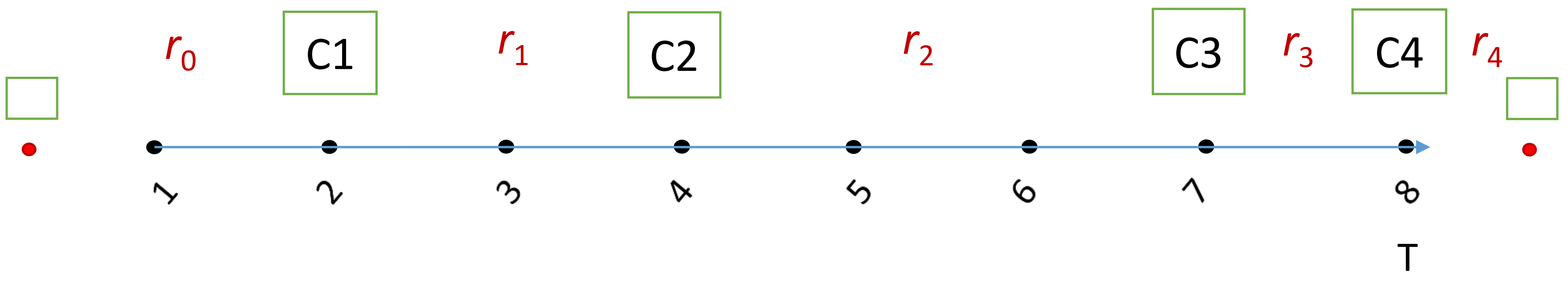
An exact reconstruction attack based on communication volume

Recovering positions:

- We defined: $r(x) = r_0 + r_1x + r_2x^2 + r_3x^3 + r_4x^4$
  $r^R(x) = r_4 + r_3x + r_2x^2 + r_1x^3 + r_0x^4$

  and $r(x)\, r^R(x) = f_4 + f_3x + f_2x^2 + f_1x^3 + f_0x^4 + f_1x^5 + f_2x^6 + f_3x^7 + f_4x^8 = F(X)$

- Factoring F(x) (over integers) can be done in polynomial time [Berlekamp 67]
  - If the factors are two irreducible polynomials, we found $r(x)$, $r^R(x)$



# records	# queries
4	2
3	4
2	11
1	14
0	5

A more efficient heuristic

- Factorization may be slow for a large number of records
- Equations:

$$r_0 \cdot r_4 = f_4$$
$$r_0 \cdot r_3 + r_1 \cdot r_4 = f_3$$
$$r_0 \cdot r_2 + r_1 \cdot r_3 + r_2 \cdot r_4 = f_2$$
$$r_0 \cdot r_1 + r_1 \cdot r_2 + r_2 \cdot r_3 + r_3 \cdot r_4 = f_1$$

- Heuristic algorithm: DFS search for a solution
  - For $m < n/2$:
    - For all integers r_m and r_{n-m} that satisfy the equation, find all feasible r_{m+1} and r_{n-m-1}
  - Otherwise:
    - Prune the combinations that do not satisfy the equation

Is the reconstruction unique? Factors of $F(x)$

- Not necessarily!
  - $r(x)=(x+2)(x+3) = x^2+5x+6$; $r^R(x)=(2x+1)(3x+1) = 6x^2+5x+1$
  - $F(x)=(x+2)(x+3)(2x+1)(3x+1) = 6x^4+35x^3+62x^2+35x+6$
  - *$F(x)$ can also be factored as*
    $r(x)=(x+2)(3x+1) = 3x^2+7x+2$; $r^R(x)=(2x+1)(x+3) = 2x^2+7x+3$

Experiments

- 2 HCUP Nationwide Inpatient Sample datasets
- ~1,500 Hospitals, each having ~6,000 patient records
- Indexed attributes: length of stay (T=365) and age (T=27)
- Simulation
  - Reconstruction always successful (up to mirroring)
  - Speed after retrieving T^4 queries: 40ms on average (max: 3.5 sec)
- Real system
  - CryptDB
    - mySQL server
    - Client
    - Packet sniffer
  - Total attack time for age attribute: 15 hours
- Demonstrates an overlooked weakness that needs to be investigated

What went wrong?

- Observation: *"It is clear that if the computed function leaks information on the parties' private inputs, any protocol realizing it, no matter how secure, will also leak this information."* [BMNW '07]
  - In our case: Exact #records leaks significant information

- Sounds familiar?
  - Observation partly motivated research into (differential) privacy

- Can differential privacy help?

DP Storage

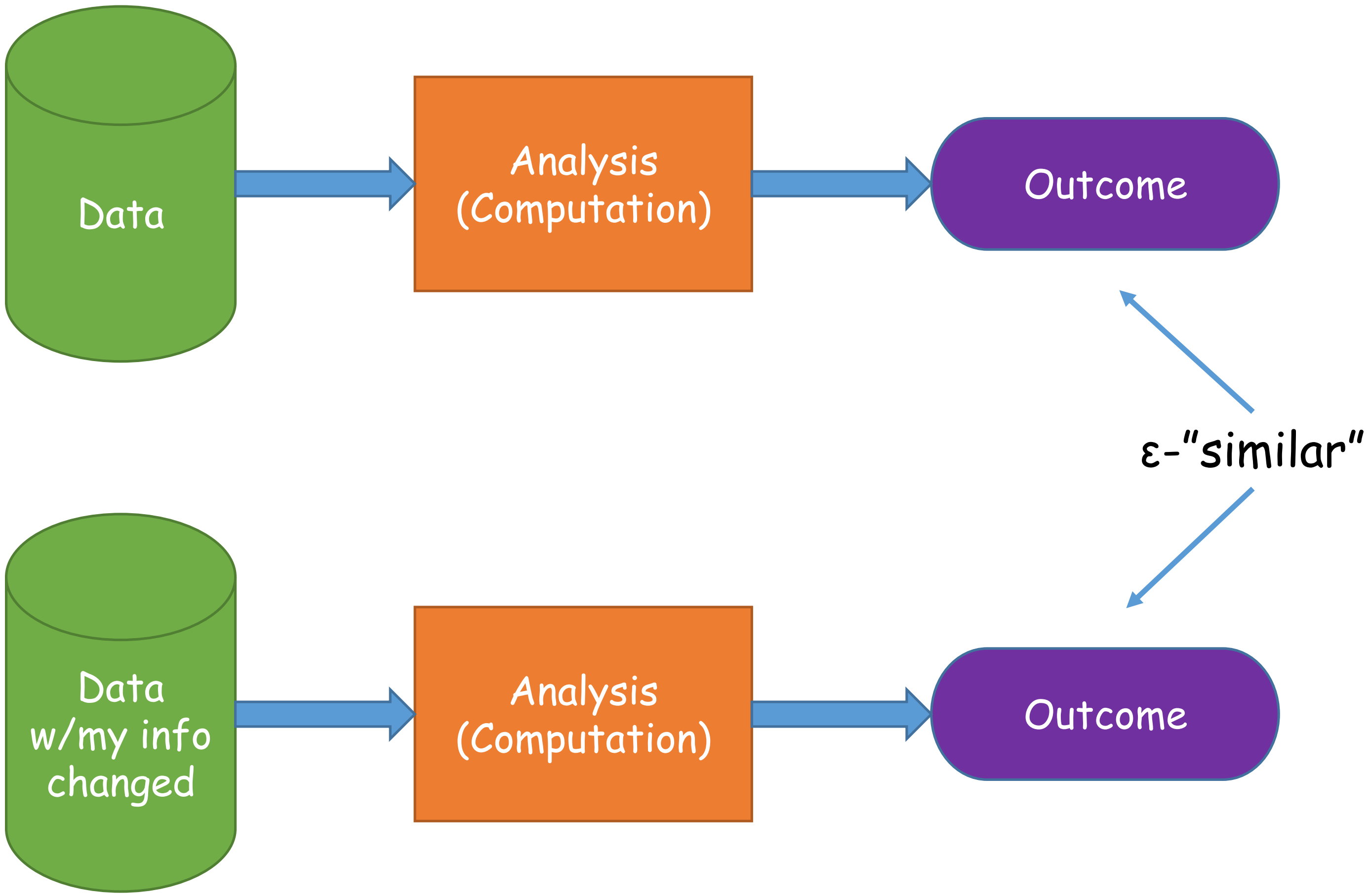
- General construction:
  - Use ORAM, inflate communication to preserve privacy
  - DP storage given a DP-sanitized version of the data
  - Can do updates
- Atomic model:
  - Multiple copies of same encrypted record
    - Only require semantic security
  - DP storage for point queries, range queries
- In both no/limited protection for queries

Access pattern leakage is not always a problem!

Differential privacy [Dwork McSherry N Smith 06]

Real world:

My ideal world:



Differential privacy [Dwork McSherry N Smith 06]

A (randomized) algorithm $M: X^n \to T$ satisfies (ϵ, δ)-differential privacy if

$\forall x, x' \in X^n$ that differ on one entry,

$\forall S$ subset of the outcome space T,

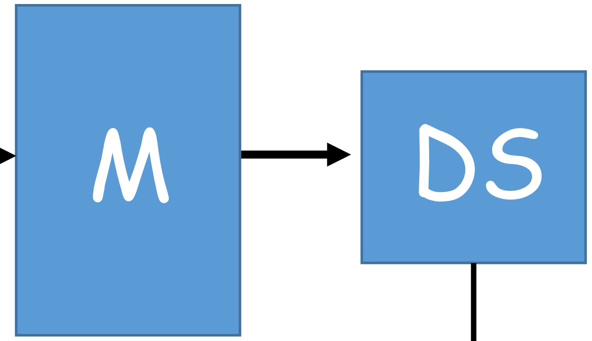
$$\Pr_{M}[M(x) \in S] \leq e^{\epsilon} \Pr_{M}[M(x') \in S] + \delta$$

Prevents reconstruction (and more)

Data sanitization [BLR'08]

- Q: A collection of statistical queries
- Sanitization:

Name	ZIP	Sex	Age	Balance
George	02139	M	32	20,000
Gina	02138	F	30	80,000
⋮	⋮	⋮	⋮	⋮
Greg	02134	F	28	20,000



For all $q \in Q$: $q(x) \approx_{\alpha} q(DS)$

$q(DS) - q(x) \in [0, \alpha]$

- [BLR 08]: $\alpha \approx (VC(Q) \log|X|)^{1/3} n^{2/3}$

Data sanitization of specific query classes

- Point queries:
  - Index: element of [1, T]
  - Query: $a \in [1, T]$; answer: # records with index = a
- Range queries:
  - Index: element of [1, T]
  - Query: $[a, b] \subseteq [1, T]$; answer: # records with index $\in [a, b]$
- 1-way attribute queries:
  - Index: element of $\{0, 1\}^k$
  - Query: $i \in [1, k]$; answer: # records with i^{th} bit of index = 1

	Pure DP	Approx. DP
Point queries	$O(\log T)$	$O(1)$ [BNS'13]
Range queries	$O(\log T)$ [BLR'08, DNPR'10, CSS'10, DNRR'15]	$O(2^{\log^* T})$ [BNS'13, BNSV'15]
1-way attribute queries	$O(k)$	$O(k^{1/2})$

DP Storage : a generic construction

- Idea: combination of a DP sanitizer for the query class and ORAM
- Setup:
  - Sanitizer is applied to the data to create a data structure DS, to be stored on the server
  - ORAM used to store all records (+indexing information as needed)
- Answering a query q:
  - q(DS) computed to get a number t of records to retrieve
    - t surpasses the real record number for q by at most α
  - ORAM used to retrieve t records
    - Including the real number of records + fake records
- Efficiency:
  - Optimally efficient for storage
  - Communication overhead = α

Summary

- Need a rigorous analysis of inherent security/privacy – efficiency tradeoffs for outsourced database systems
  - Optimal efficiency → reconstruction attacks (access pattern and/or communication volume) even with very limited adversaries
  - Can be mitigated by combining ORAM with differential privacy

- Question:
  - What is/are the right notion(s) of privacy we should pursue in this context?
  - Things to consider: privacy of data, privacy for inquirer