Composition for Statistical Databases

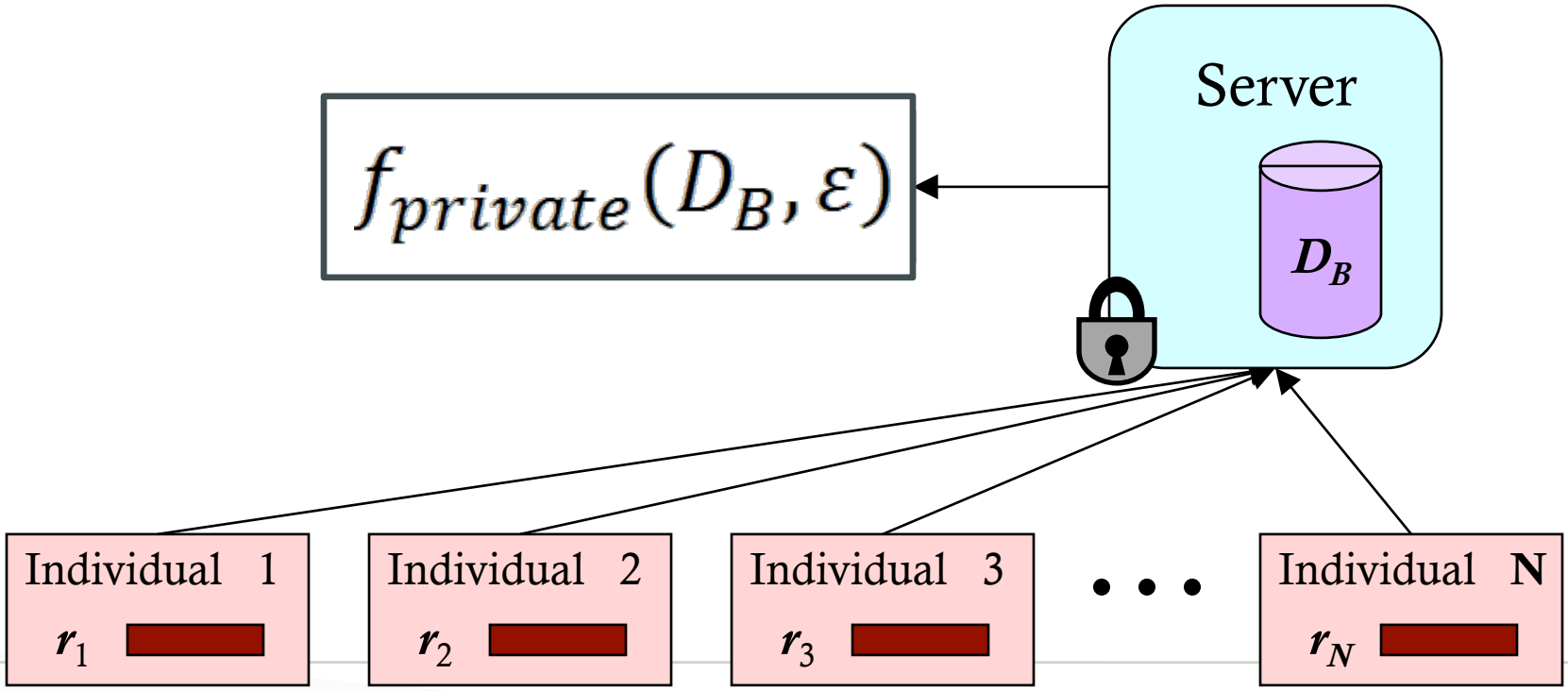
Ashwin Machanavajjhala
ashwin@cs.duke.edu

Duke University

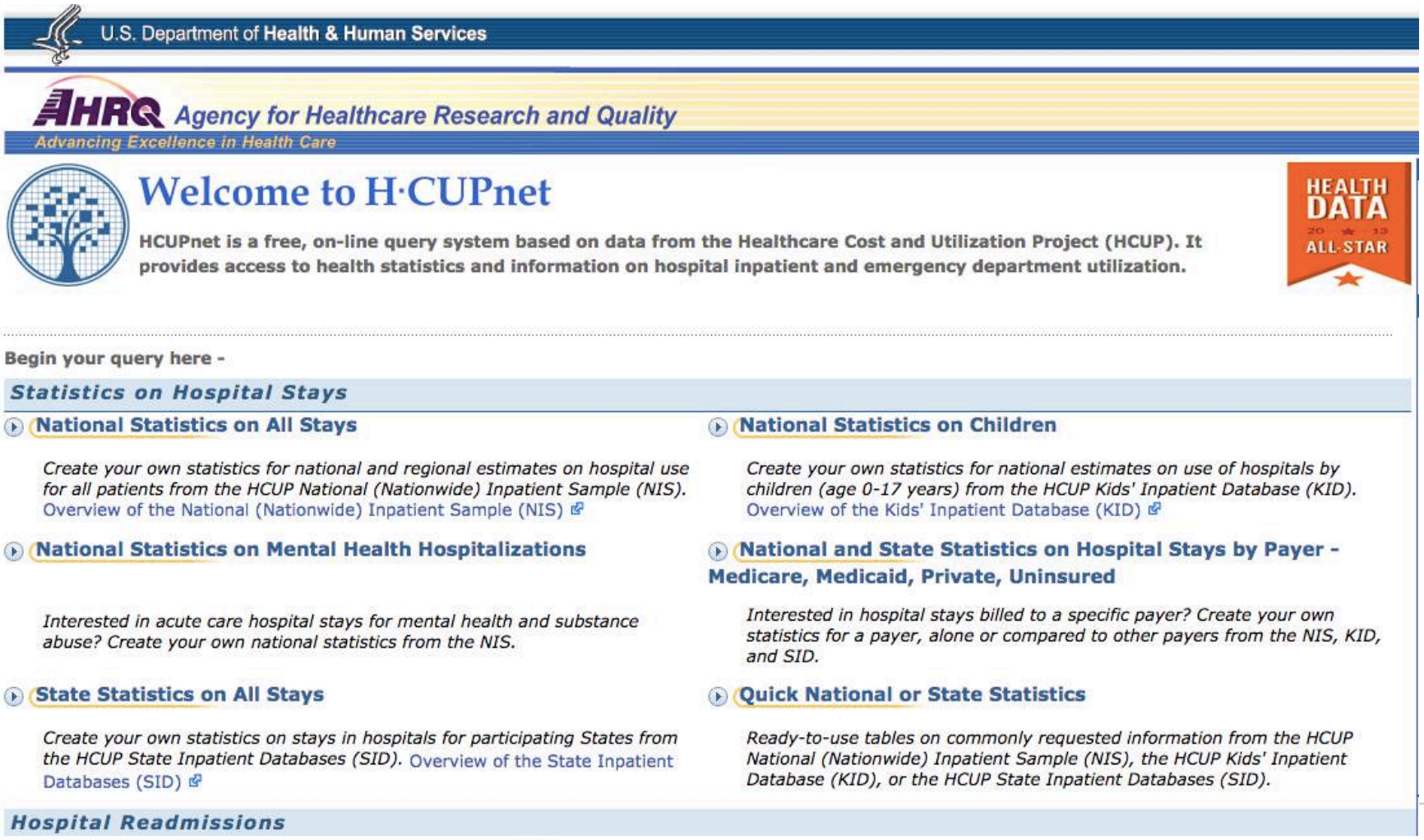
Statistical Databases

Utility: $f_{private}$ *approximates* f

Privacy: No breach about any individual



Example: HCUPnet



#Hospital discharges in NJ of ovarian cancer patients, 2009

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	*	19	22
1-17	*	*	*	*	*	*	*	*
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

#Hospital discharges in NJ of ovarian cancer patients, 2009

Age	#discharge	White	Black	Hispanic	Asian/Pcf	Native American	Other	Missing
#discharges								2
1-17								
18-44								
45-64								
65-84								
85+	34	29	*	*	*	*	*	*

Privacy condition: K-Anonymity

Blending in a crowd of size at least k

Algorithm: Suppression

Suppress answers with count at most k

#Hospital discharges in NJ of ovarian cancer patients, 2009

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	1	19	22
1-17	3	1	*	*	*	*	*	*
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

Can reconstruct tight bounds on rest of data

[Vaidya et al AMIA 2013]

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	1	19	22
1-17	3	1	[0-2]	[0-2]	[0-1]	[0]	[0-1]	[0-1]
18-44	70	40	13	[9-10]	[0-6]	[0]	[0-6]	[1-8]
45-64	330	236	31	32	[10]	[0]	11	[10]
65-84	298	229	35	13	[2-8]	[1]	[2-8]	[4-10]
85+	34	29	[1-3]	[1-4]	[0-1]	[0]	[0-1]	[0-1]

Can reconstruct tight bounds on rest of data

[Vaidya et al AMIA 2013]

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges								2
1-17								0-1]
18-44								1-8]
45-64								10]
65-84								4-10]
85+	34	29	[1-3]	[1-4]	[0-1]	[0]	[0-1]	[0-1]

K-Anonymity does not compose well with itself.

Can reconstruct tight bounds on rest of data

[Vaidya et al AMIA 2013]

Infact, when linked with queries giving other statistics, we can figure out that exactly 1 Native American woman diagnosed with ovarian cancer went to a privately owned, not for profit, teaching hospital in new Jersey with more than 435 beds in 2009. Furthermore, the woman did not pay by private insurance, had a routine discharge, with a stay in the hospital of 33.5 days, with her home residence being in a county with 1 million plus residents (large fringe metro, suburbs), and her age was exactly 75 years.

Why Composition?

- Attackers can use non-trivial algorithms
  - Reasoning about privacy of a complex algorithm is hard.

- Helps software design
  - If building blocks are proven to be private, it would be easy to reason about privacy of a complex algorithm built entirely using these building blocks.

Composition in Databases

- What are limits on composition?
- Are there privacy notions that compose well with themselves?
- What about composition across privacy notions?

A Lower Bound

- In order to ensure utility, a statistical database must leak some information about each individual
- We can only hope to bound the amount of disclosure
- Hence, there is a limit on number of queries that can be released

Dinur Nissim Result

[Dinur-Nissim PODS 2003]

- A vast majority of records in a database of size n can be reconstructed when $n \log(n)^2$ queries are answered by a statistical database …

  … even if each answer has been arbitrarily altered to have up to $o(\sqrt{n})$ error

.

Privacy as Constrained Optimization

- Three axes
  - Privacy
  - Error
  - Queries that can be answered

- E.g.: Given a fixed set of queries and budget on the amount of privacy, what is the minimum error that can be achieved?

Composition in Databases

- What are limits on composition?

- **Are there privacy notions that compose well with themselves?**
  - Differential Privacy
- What about composition across privacy notions?

Self-Composition

- Let P be some privacy criterion that takes ε as a privacy parameter.

- If M_1, M_2, …, M_k are algorithms that access a private database D such that each M_i satisfies the privacy criterion P with param ε_i

  then the combination of their outputs satisfies privacy parameter P with param $\varepsilon = f(\varepsilon_1, \ldots, \varepsilon_k)$

Differential Privacy

[Dwork ICALP 2006]

An algorithm A satisfies ε-differential privacy if:

For **every pair** of *neighboring tables* D_1, D_2
(that differ in one individual)
For **every output** O

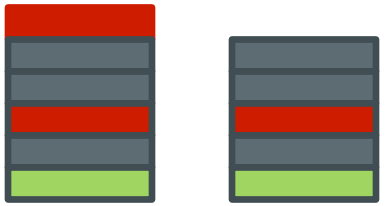
$$\Pr[A(D_1) = O] \leq e^{\varepsilon} \Pr[A(D_2) = O]$$

Differential Privacy

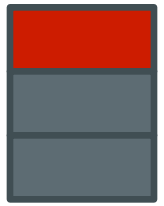
[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

Adversary should not be able to distinguish between any D_1 and D_2 based on any O

$$\log\left(\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]}\right) < \varepsilon \quad (\varepsilon > 0)$$

Differential Privacy composes well with itself

- If $M_1, M_2, \ldots, M_k$ are algorithms that access a private database D such that each M_i satisfies ε_i-differential privacy,

  then the combination of their outputs satisfies ε-differential privacy with $\varepsilon = \varepsilon_1 + \ldots + \varepsilon_k$

Sequential Composition

Differential Privacy composes well with itself

- If $M_1, M_2, \ldots, M_k$ are algorithms that access disjoint databases $D_1, D_2, \ldots, D_k$ such that each M_i satisfies ε_i-differential privacy,

  then the combination of their outputs satisfies ε-differential privacy with $\varepsilon = \max\{\varepsilon_1, \ldots, \varepsilon_k\}$

Parallel Composition

Other privacy notions?

- Extending differential privacy with any relation defining neighboring databases satisfies *linear* self composition. **[Kifer-M TODS 2014]**

  - E.g., Blowfish Privacy **[He-M-Ding SIGMOD 2014]**
    used in a US Census data product **[Haney et al -EuroStat 2015]**

Composition in Databases

- What are limits on composition?
- Are there privacy notions that compose well with themselves?
- **What about composition across privacy notions?**

Why composition across privacy notions?

- Organization publishes exact counts about the data (say using k-anonymity)
- Then, organization publishes differentially private counts at finer resolutions.
- What can we say about the privacy of data now?

Composition across privacy notions

- Need to understand semantics of privacy
  - **Pufferfish Privacy Framework** **[Kifer-M TODS 2014]**

- An algorithm is differentially private if and only if an adversary, who believes the database rows are independent, does not learn too much additional information about any one row using the output of the algorithm.

Composition across privacy notions

- Need to understand semantics of privacy
  - **Pufferfish Privacy Framework** **[Kifer-M TODS 2014]**

- **Differential Privacy** if and only if
  rows are safe from IND adversary

Composition with prior releases

- Can we say …

  Prior release + Differential privacy if and only if **rows are safe from IND' adversary**

  where **IND'** adversary assumes rows are independent conditioned on prior release

Composition with prior releases

[He-M-Ding SIGMOD 2014]

- NO. But we can say …

  Prior release + Blowfish privacy only if
  rows are safe from IND' adversary

  where **Blowfish** extends the neighbor definition of Differential privacy to take into account prior releases.
  - Stronger than DP.

Summary & Open Questions

- Composition is an important tool for private algorithm design by non-experts.
- Differential privacy satisfies linear self composition
  - **What other privacy notions satisfy self composition?**
- Differential privacy does not compose well with other privacy notions or prior data releases.
  - **Understand composition across privacy notions is an important open question.**

References

[Dinur-Nissim PODS 2003] Dinur, Nissim, "Revealing information while preserving privacy", PODS 2003

[Dwork ICALP 2006] Dwork, "Differential Privacy", ICALP 2006

[Haney et al EuroStat 2015] Haney, Machanavajjhala, Abowd, Graham, Kutzbach, Vilhuber, "Formal privacy protection for data products combining individual and employer frames", UNECE-Eurostat Joint Workshop 2015

[He-M-Ding SIGMOD 2014] He, Machanavajjhala, Ding, "Blowfish Privacy", SIGMOD 2014

[Kifer-M TODS 2014] Kifer, Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions", ACM TODS 2014

[Vaidya et al AMIA 2013] Vaidya, Shafiq, Jiang, Ohno-Machado, "Identifying inference attacks against healthcare data repositories", AMIA 2013