DJoin: Differentially Private Join Queries Over Distributed Databases

Arjun Narayan Andreas Haeberlen

What is the problem?
1. Useful information exists in various databases today, but getting access is hard because the data is private.

Differential Privacy
Key idea: To protect individuals, it is okay to answer only questions that are about aggregates.

How does it work?
Adds carefully calculated noise to query result.

Result:
Provable privacy guarantees. Get to use the data!

Challenge: Distributed Data
Existing differential privacy runtimes like PINQ and Fuzz are limited to a single database.

Insight: Use Set Intersections
Some Joins can be rewritten as set intersections.

Rewriting JOIN queries
SELECT COUNT(X) FROM HOSPITAL JOIN AIRLINE WHERE Destination= "Elbonia" AND Diagnosis = "Malaria"

How does DJoin Perform?
Simple queries on three databases with 15,000 rows each take between 1 and 7.5 hours. Previous best was Secure Multiparty Computation - i.e. years.

First distributed differentially private system capable of executing JOIN queries. Scalable: All operations parallelize very well. Extensible to multi-way JOINs ≥3 parties.