ePPI: Locator Service in Information Networks with Personalized Privacy Preservation

Yuzhe Tang, Ling Liu, Arun Iyengar,
Kisung Lee and Qi Zhang

Outline

- **Background**
- ePPI: Personalized privacy preservation
- Practical ePPI construction
- Evaluation

Systems: Information networks

- Information networks arise in Health domain.
  - Health Information exchanges (HIE)

    commonwell HEALTH ALLIANCE

    NHIN Nationwide Health Information Network

    GEORGIA HEALTH INFORMATION NETWORK
  - Software

    the Direct Project

    CONNECT
- Information networks appear in other domains:
  - Social networks
  - Cloud computing
  - Enterprise networks

Georgia Tech

Application: Data exchange in HIE

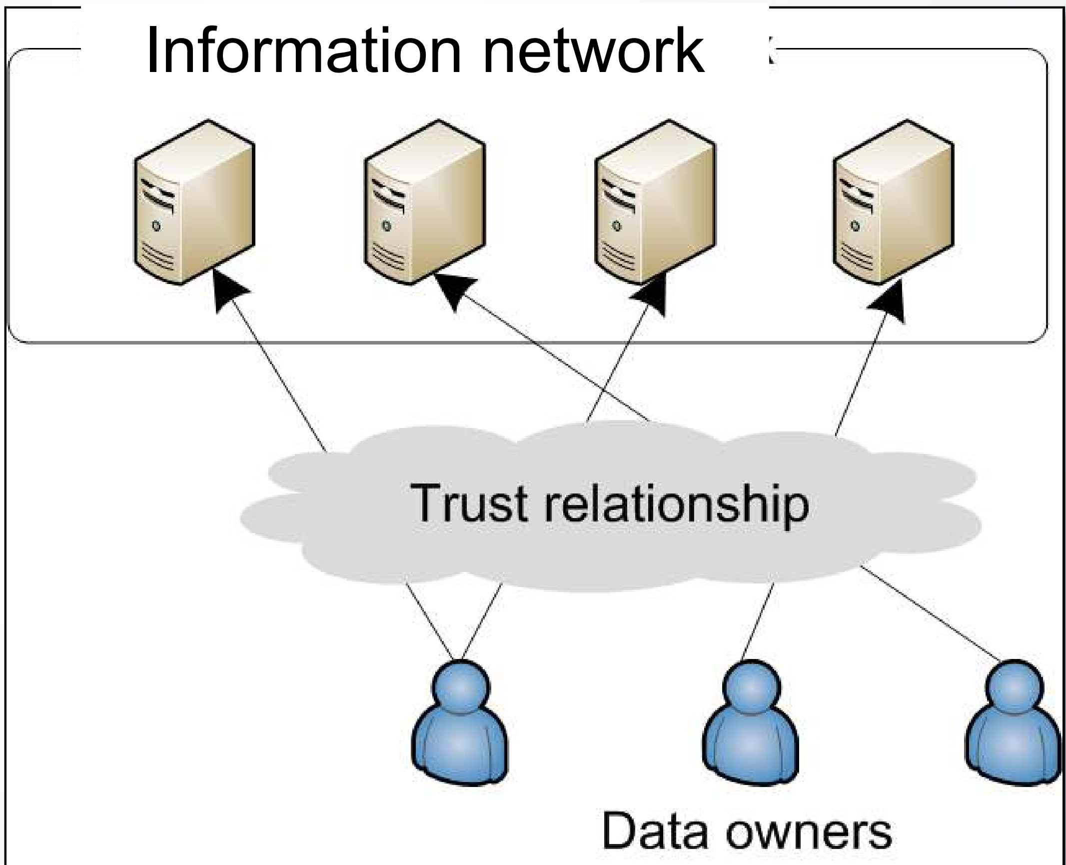
- Why exchange data? Boost the data value
- Example in HIE:
  - Patient in *Emory* hospital: “I just did my blood test in *Grady* hospital two days ago. Can I use that data?”
    - The case of unconscious patient
- Sharing information in HIEs creates privacy issues

Proposal: Privacy aspect of RLS

- Location of health care data should be private in certain cases.
  - Location of health care records could suggest type of medical condition a patient might be suffering from
- Privacy preservation is regulated.
  - HiPAA for privacy of healthcare records

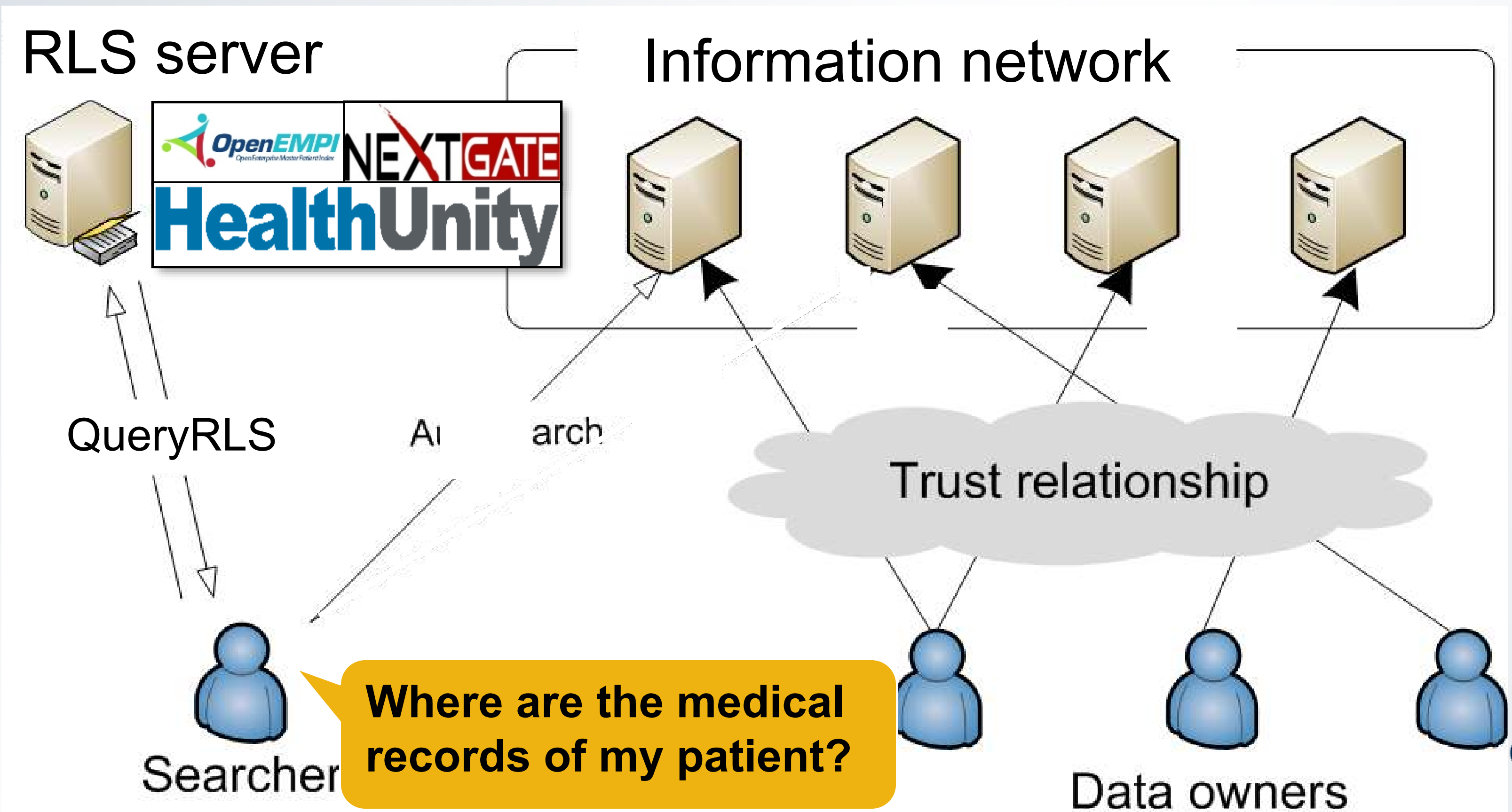
Abstract: System/trust model

- Owners to providers: Selected trust relationship
  - HIE: “A patient only trusts the hospitals s/he visited”
- Providers to providers: No mutual trust
  - Each provider in a separate domain
  - Different providers compete for the same customer base

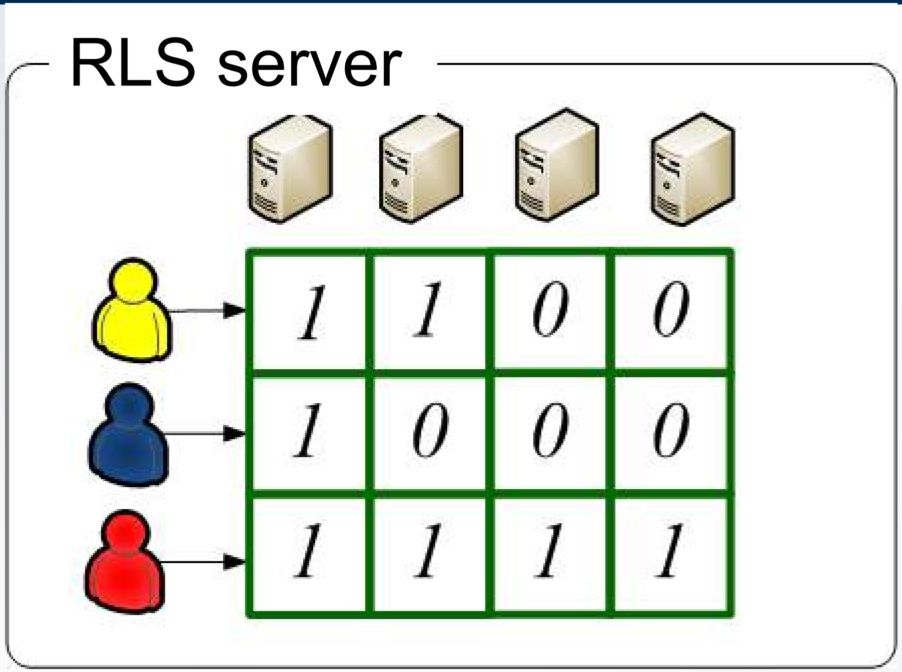


Record Locator Service (RLS)

- RLS: a standard procedure in HIE
  - "Given a patient ID, where are the medical records located?"



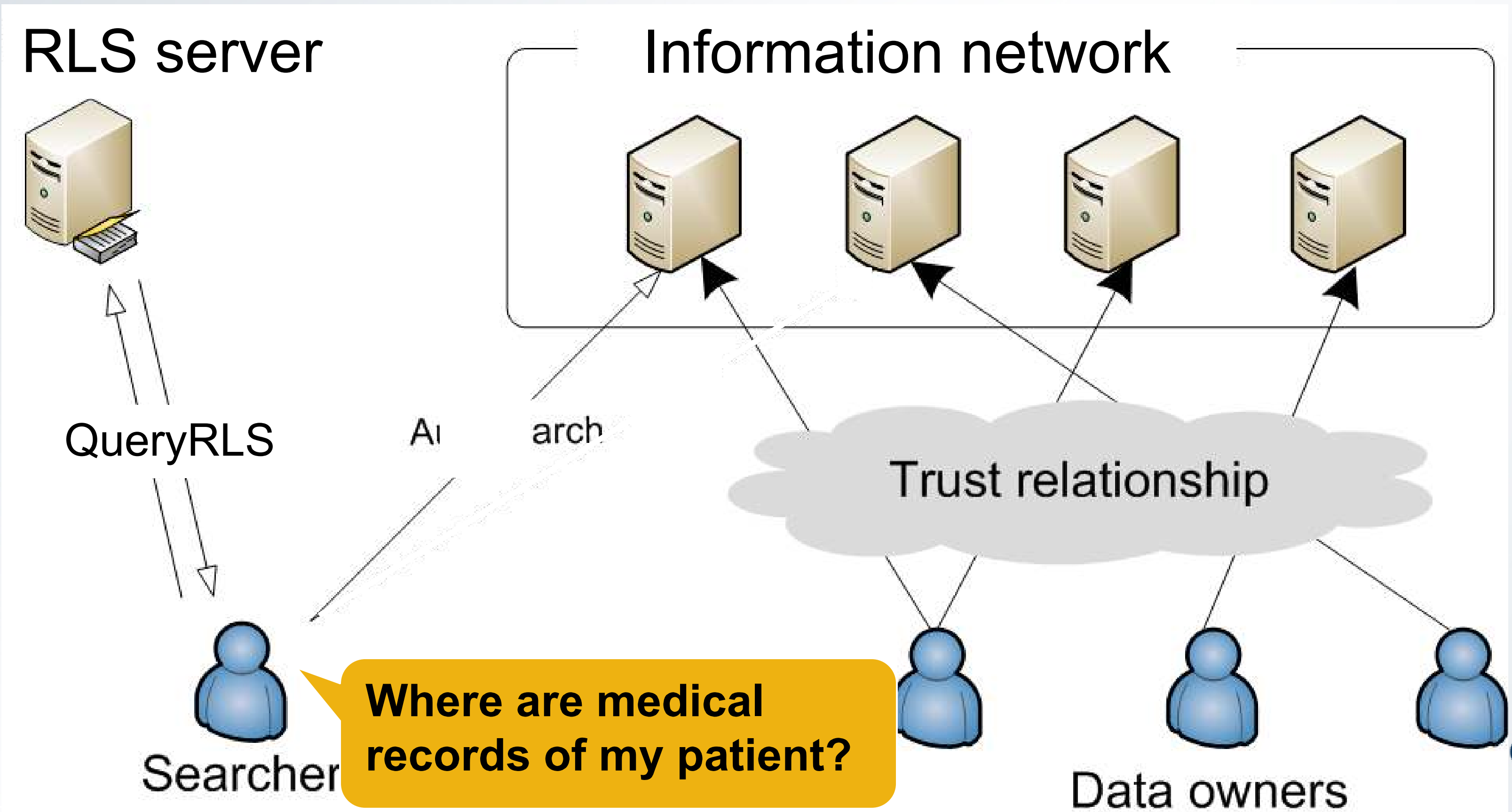
RLS: Data model and privacy



- Essentially an inverted index.
  - Mapping between a patient/owner and a provider.
- Assumption:
  - Owner/patient has the same ID globally
  - Related work: Record linkage/MPI (UTD, Vanderbilt)

Proposal: Privacy-preserving index in information networks

- PPI is a Privacy-Preserving Index for RLS.



Georgia Tech

Previous Approach: k-Anonymity Using Groups

- Organize providers into disjoint groups
- Satisfy query with a group containing a valid provider
- Providers in same group are indistinguishable by searchers
  - Valid searcher may need to contact each provider in a group to find a record
- Drawbacks
  - Assumes providers are willing to share private local indices
  - Cannot provide privacy levels personalized to individual patients
  - Cannot specify quantitative privacy guarantees

Contribution

- We are the first to consider an untrusted RLS with privacy preservation.
  - Traditional RLS server requires trusts from participating hospitals and providers.

- We are the first to study the following two problems:
  - Personalized privacy preservation
  - Practical ePPI construction.

Outline

- Background
- **ePPI: Personalized privacy preservation**
- Practical ePPI construction
- Evaluation

Problem 1: Personalized privacy preservation

- Different people have different levels of privacy concerns.

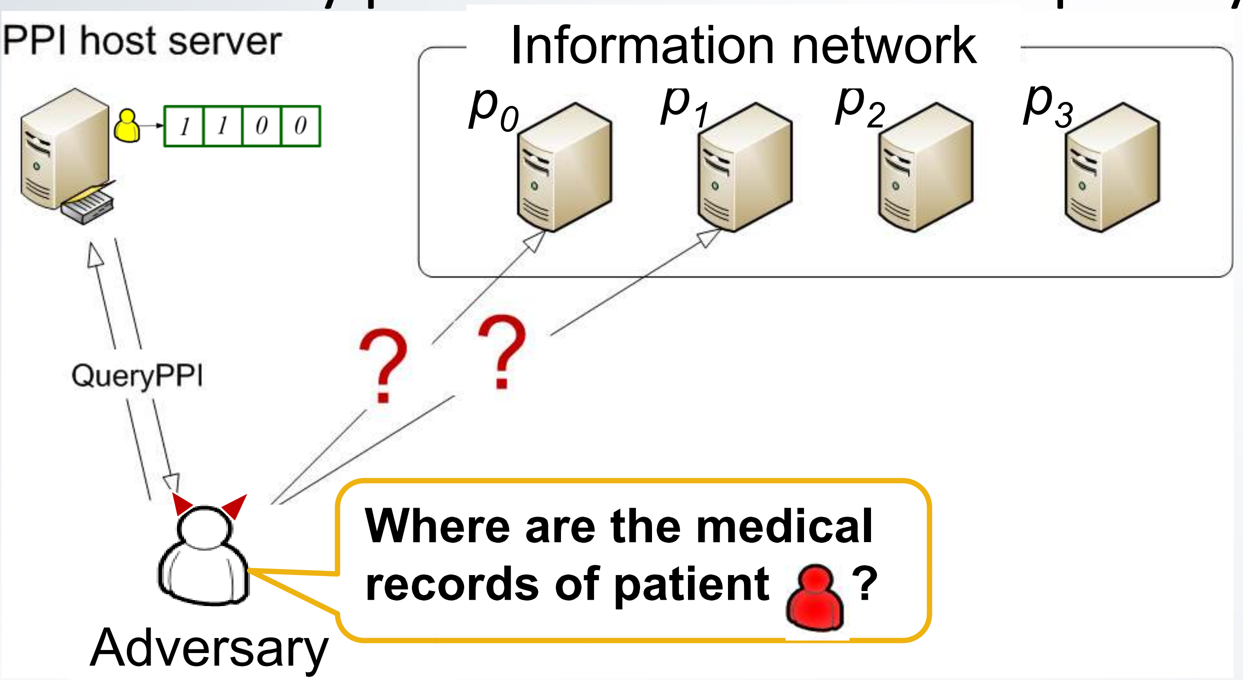
Famous athlete/ politician visited a hospital

>

An average person visited a hospital

ePPI: Personalized privacy protection

- *e*-privacy: e is privacy degree=> proportion of false positives.
  - Moderately-private: *e* =0.5 for balanced perf./privacy prsvn.
  - Non-private: *e* =0 for best search performance.
  - Extremely private: *e* =0.75 for best privacy preservation.



e=# /#(+)

=1/2=0.5

=3/4=0.75

- *k*-anonymity does not apply here.
  - Grouping *k* providers is agnostic to patients.

How to specify e?

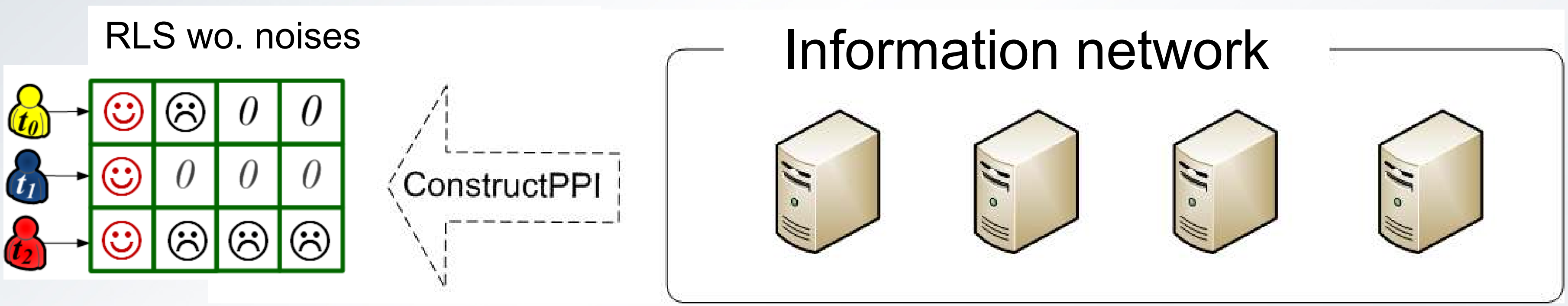
- Heuristics:
  - Value *e* depends on how famous the person is?
  - "Average person"   big *e*
  - "Average person"   small *e*
- Use social network analysis to recommend *e* automatically.
  - Social users with big degree   big *e*
  - Social users with small degree   small *e*

Outline

- Background
- ePPI: Personalized privacy preservation
- **Practical ePPI construction**
- Evaluation

Secure ePPI construction

- ePPI construction:
  - Input: sensitive mapping data on untrusted providers
  - It needs to be secure



  - Add noises ( ) quantitatively

Problem 2: Efficient ePPI construction

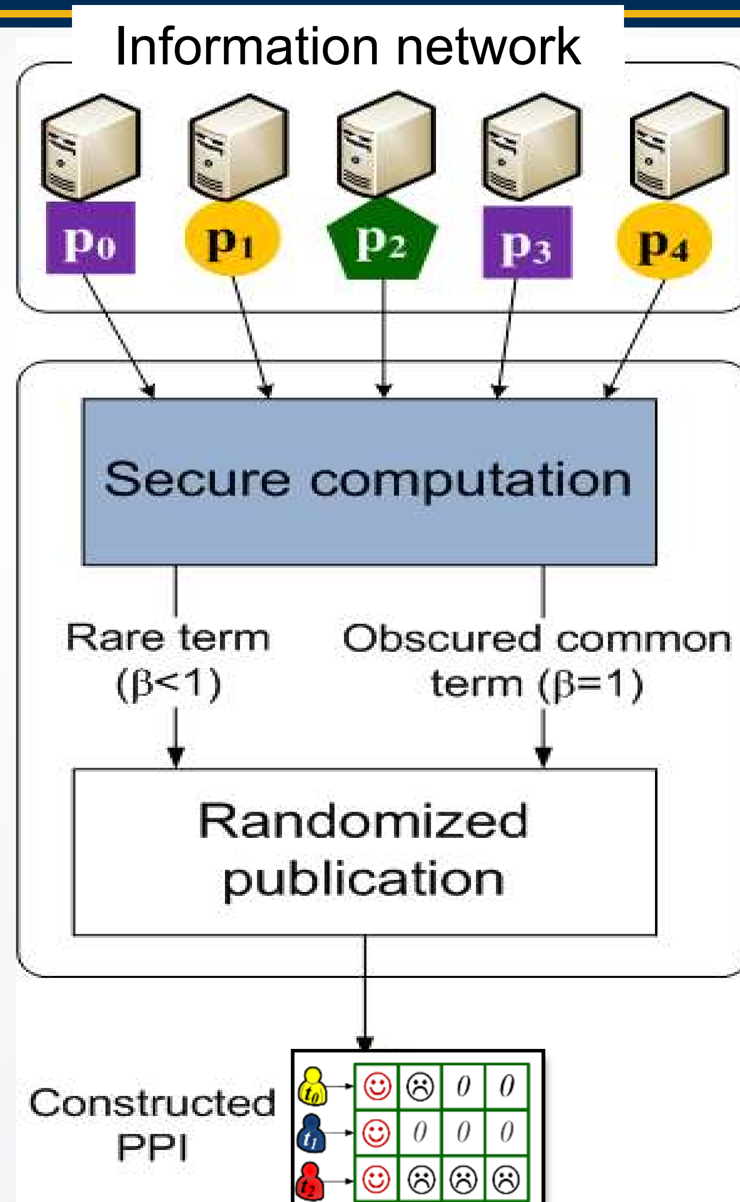
A challenge for the large-scale index construction:

- Traditional technique: MPC (multi-party computations).
  - Sample Problem: Answer "Who is the richest person in this room?" while keeping financial data private
- MPC is very expensive for big data and computations (DJoin [OSDI 2012: Narayan & Haeberlen])

FairplayMP [4], need about 10 seconds to evaluate (very simple) functions that can be expressed with 1,024 logic gates.

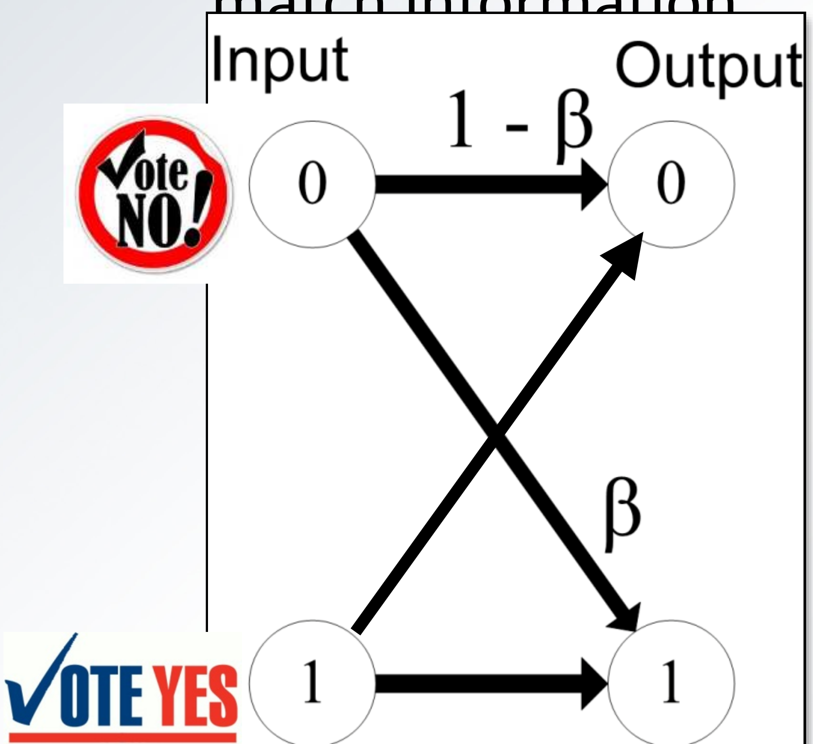
ePPI construction overview

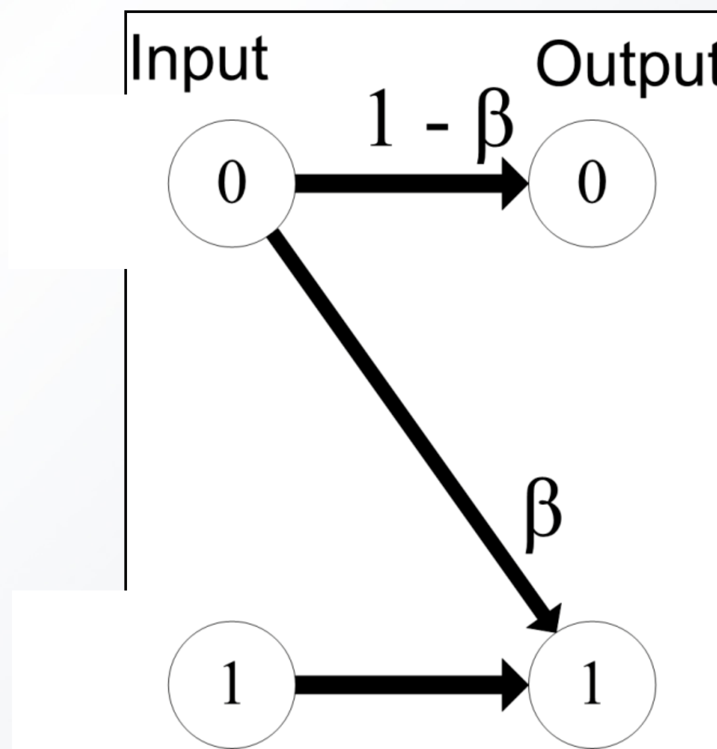
- Design: Separate secure and non-secure computations
  - Minimize secure computations
- Index construction framework:
  1. Secure computation producing a probability β
  2. Randomized publication based on β [link]
  3. Generate a false positive for a provider which does not store a record with probability β.



Randomized publication

- Inspired by the privacy preserving voting technique
  - Voting: “Vote for/against President Obama wo. disclosing my decision”
  - ePPI: “Releasing match/non-match data wo. disclosing match information”





Randomized publication

- Randomized publication: given a probability β, each provider flips their "coins" to decide tell a truth or lie.
  - Essentially, a process of *Bernoulli trials.*
  - Provide quantitative privacy guarantees with *Chernoff bounds*.

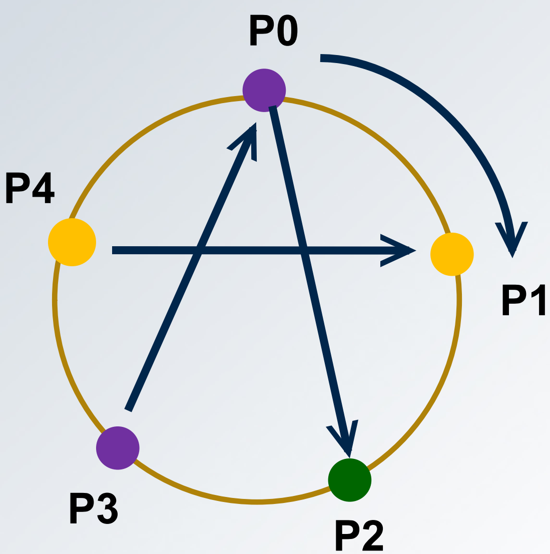
Theorem 4.1: Given desired success rate $\gamma > 50\%$, let $G_j = \frac{\ln \frac{1}{1-\gamma}}{(1-\sigma_j)m}$ (where m is the number of providers) and

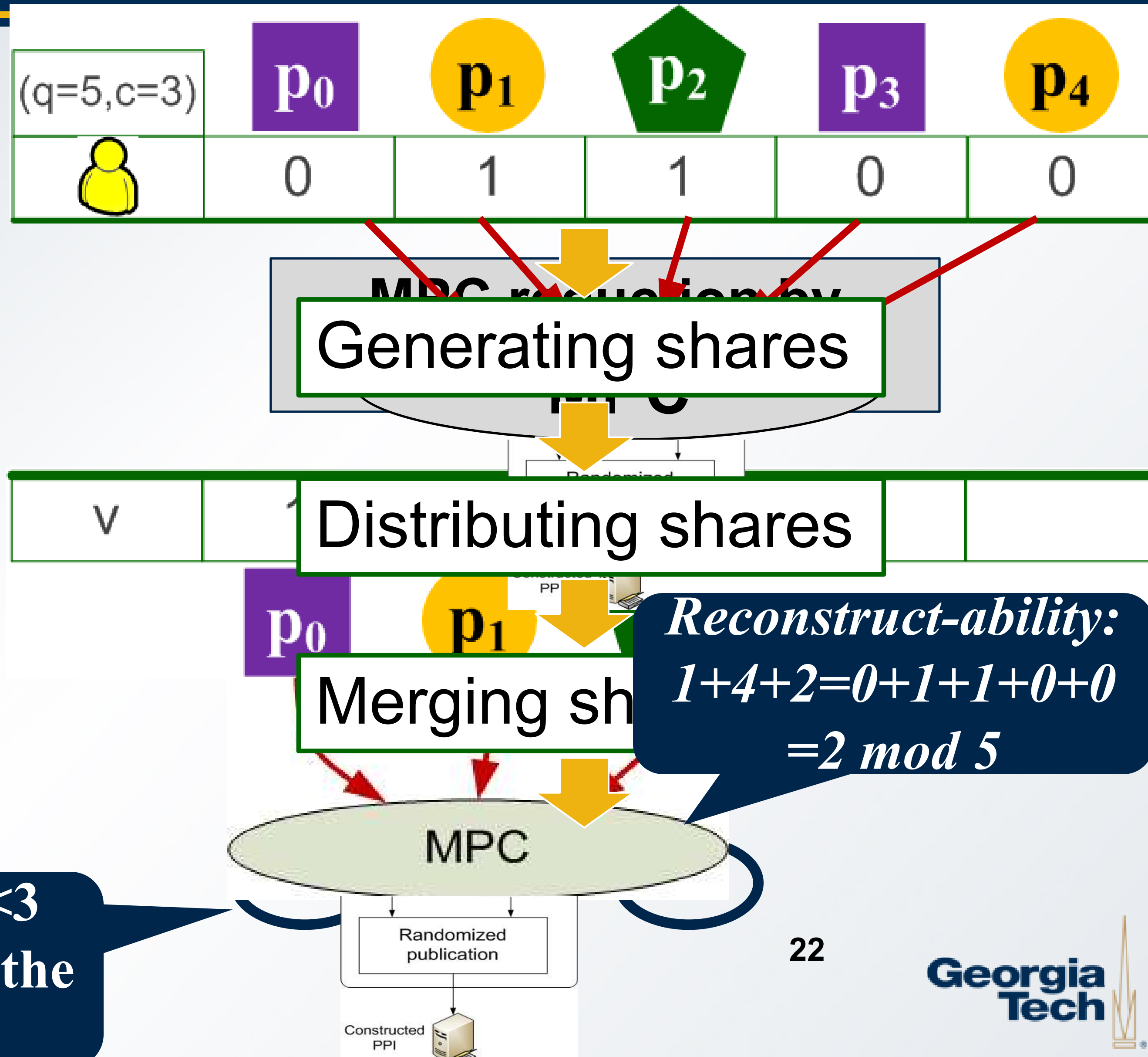
$$\beta_c(t_j) \geq \beta_b(t_j) + G_j + \sqrt{G_j^2 + 2\beta_b(t_j)G_j} \qquad (3)$$

Then, the randomized publishing with $\beta(t_j) = \beta_c(t_j)$ statistically guarantees that the actual false positive rate in the published ϵ-PPI is larger than ϵ with success rate $p_p \geq \gamma$.

Proof in ePPI paper [link]

Secure computation: secret sharing

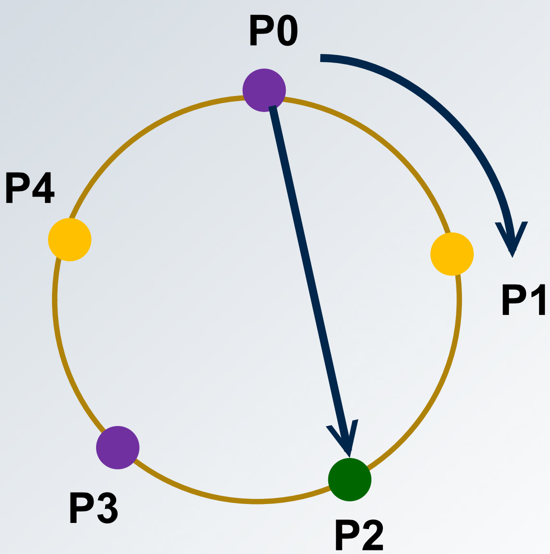


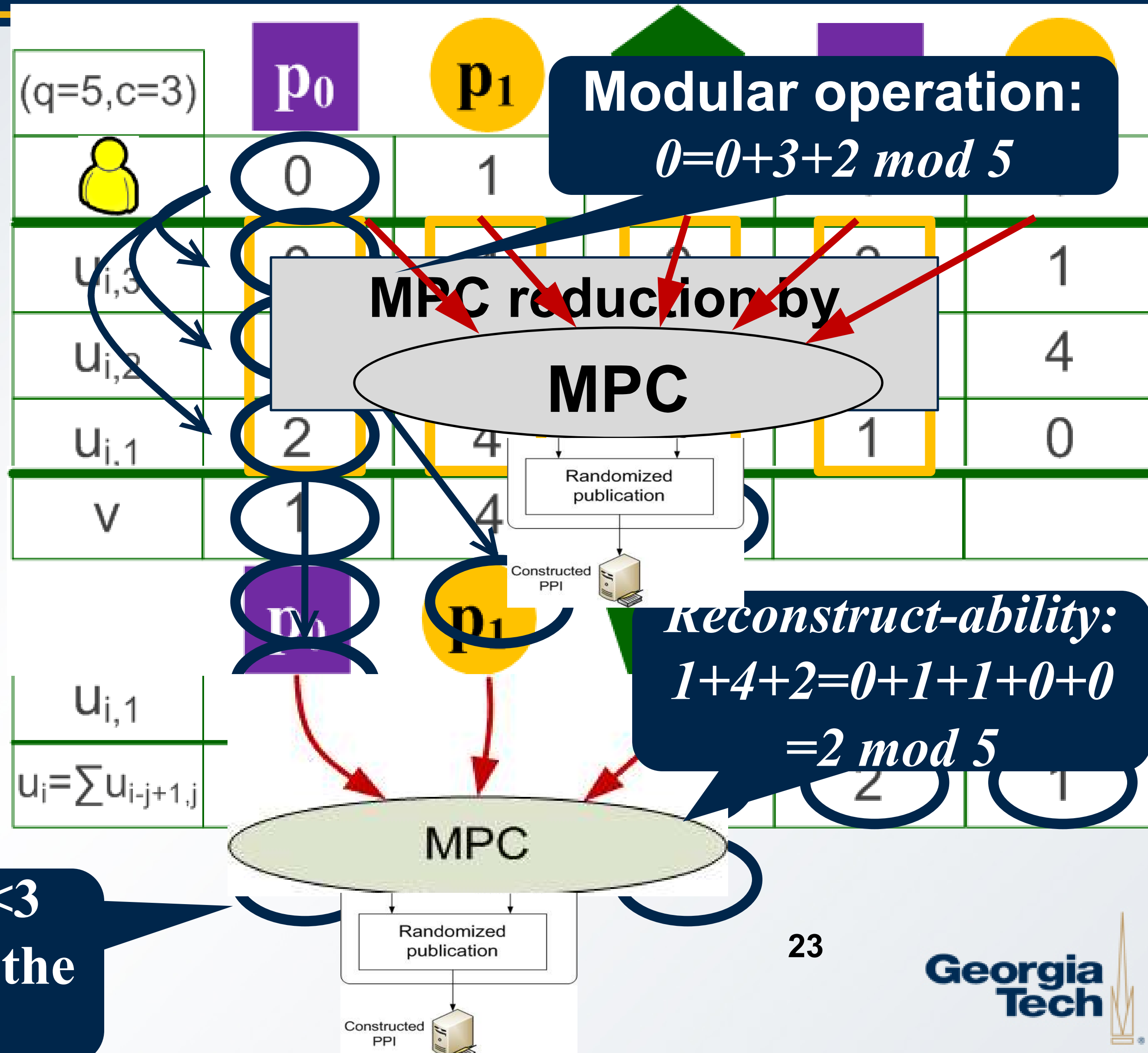


Reconstruct-ability:
1+4+2=0+1+1+0+0
=2 mod 5

***Secrecy*: knowing <3 shares can't deduce the secret sum, 2.**

Secure MPC reduced by secret sharing





Modular operation:
0=0+3+2 mod 5

Reconstruct-ability:
1+4+2=0+1+1+0+0
=2 mod 5

Secrecy: knowing <3 shares can't deduce the secret sum, 2.

Georgia Tech

Outline

- Background
- ePPI: Personalized privacy preservation
- Practical ePPI construction
- **Evaluation**

Evaluation

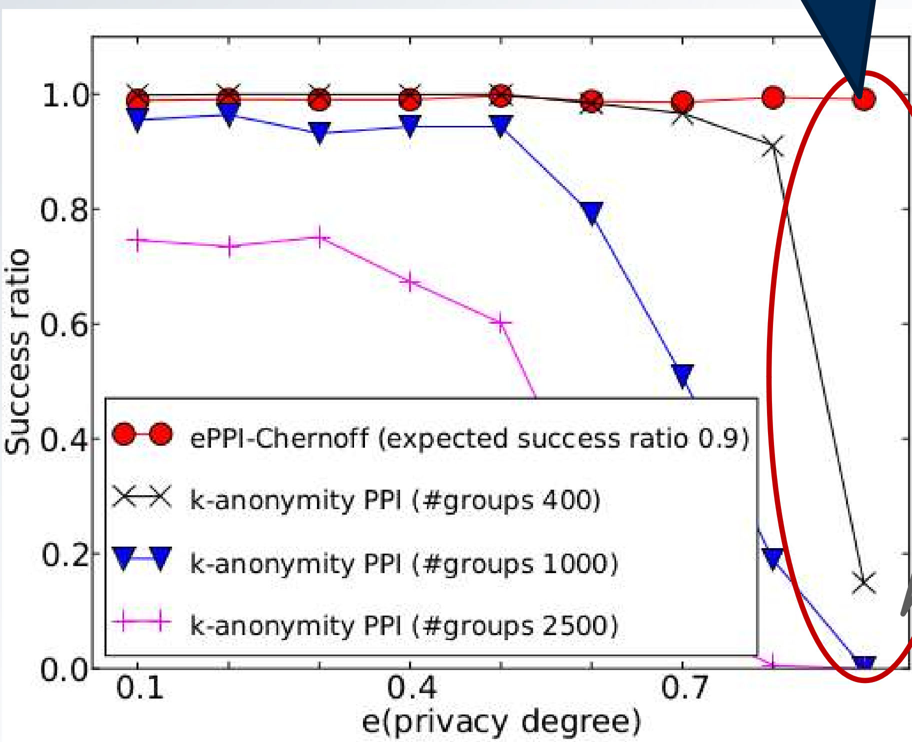
- Exp-1: Privacy (Problem 1)
  - By simulation
- Exp-2: Performance (Problem 2)
  - By real system implementation.

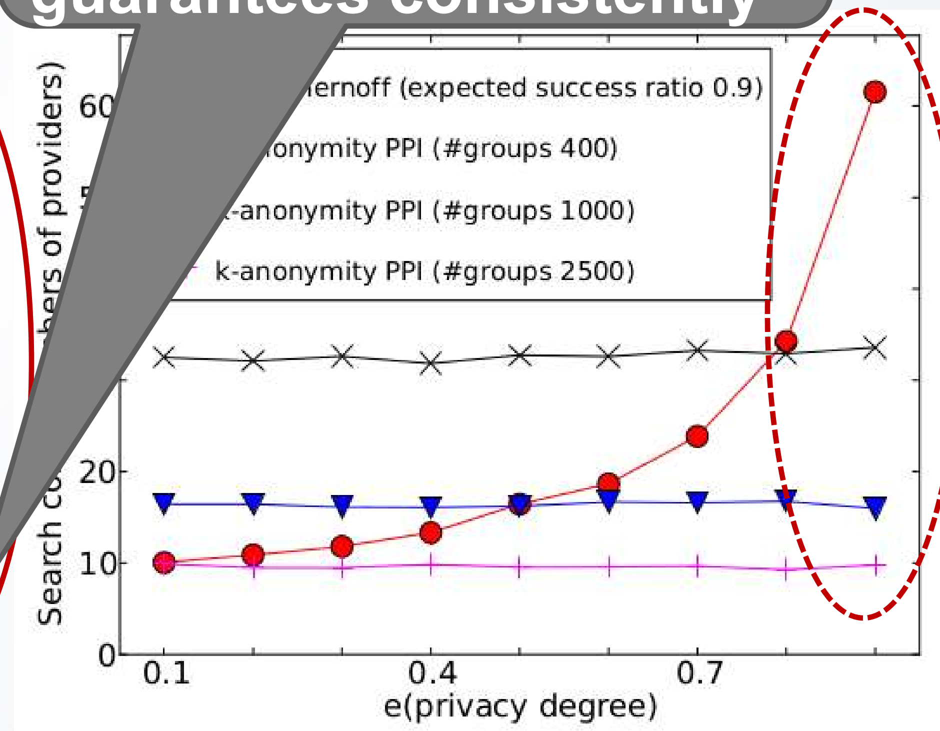
Comparing ePPI with *k*-anonymity based PPIs

- Dataset: A d ... 03].
- Success ratio meas ... th goals are met (regar ... g

ePPI preserves privacy with high success ratio on large *e*

***k*-anonymity based PPI can not deliver privacy guarantees consistently**





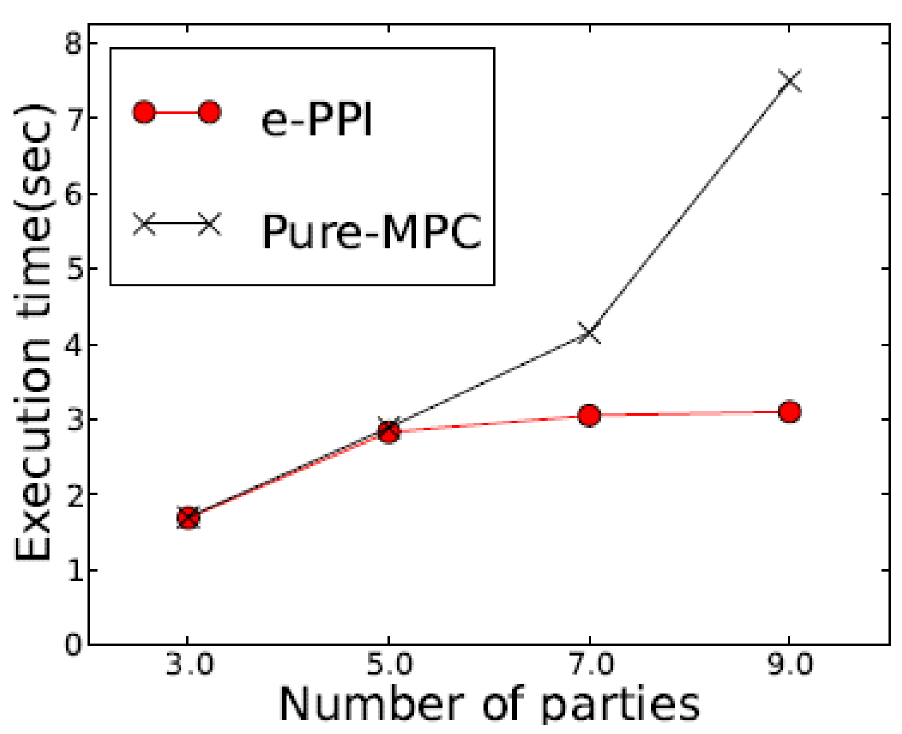
Experiment setup for performance evaluation

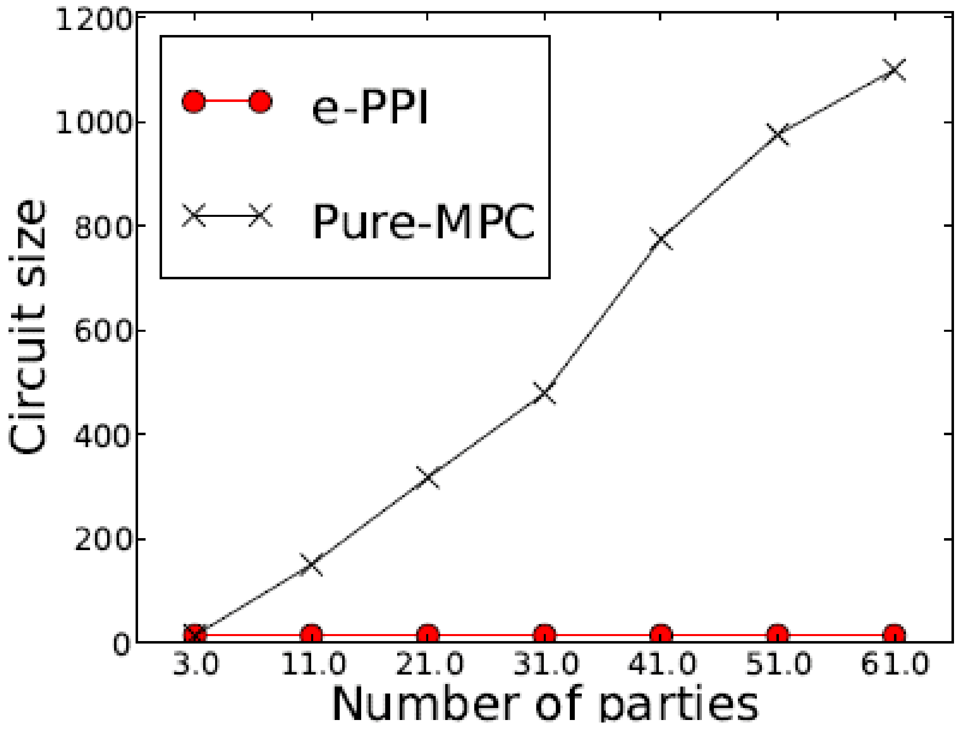
- Implementation:
  - Secret sharing reduction with limited MPC using:
    - Protocol Buffers for object serialization.
    - Netty for network communication.
  - MPC by FairplayMP[CCS08]

- Evaluation platform:
  - Emulab: with 10 machines
  - Machine with a 2.4GHz core and 12G RAM

Performance

- ePPI construction incurs time constant to the number of parties.
- Pure-MPC construction incurs exponentially growing time.





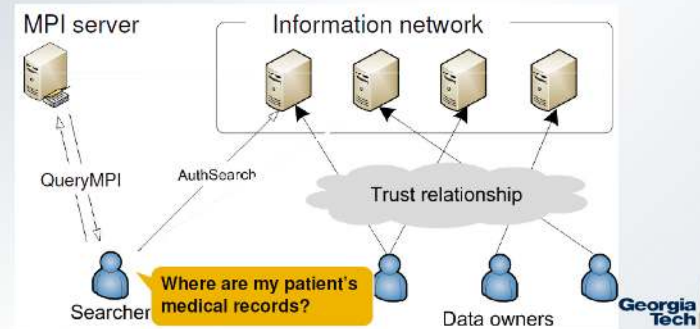
Talk summary for QA

Systems: Information networks

- Information networks arise in many application areas.
  - Health: Information exchanges (HIE)
  - Distributed social networks
  - IT business: Multi-clouds

Privacy-preserving index in information networks

- PPI is a Privacy-Preserving Master Patient Index.
- PPI is public, without access controls.



Problem 1: Personalized privacy preservation

- Different people have different levels of privacy concerns.

Tiger Woods (or VIP) visited a hospital > **An average person visited a hospital**

12

ePPI construction overview

- Design: Separate secure and non-secure computations
- Index construction framework:
  1. Secure computation producing a probability β
  2. Randomized publication based on β [link]