Differential Privacy Preserving Genomic Data Releasing via Factor Graph

Zaobo He[1], Yingshu Li[1], and Jinbao Wang[2]

[1]Georgia State University, USA

[2]Harbin Institute of Technology, China

Background

- Cost of DNA sequencing drops dramatically
  - Genomic study requires analyzing large amount of genetic information from many individuals.
  - Individuals are using their genomes to learn about disease predispositions, medicines, etc.
  - Voluntary and mandated sharing of genome data among hospitals, biomedical research organizations, and other data holders.

Privacy

Background (Cont.)

- Kin-Genomic Privacy
  - DNA sequences of relatives are even highly similar
  - No consent from one’s relatives is needed to release ones genome data
  - Individuals revealing genome data may threaten the relatives’ privacy besides their own.

Studied Problem

- Differential privacy guaranteed kin-genomic data releasing

Previous Privacy Preserving Methods

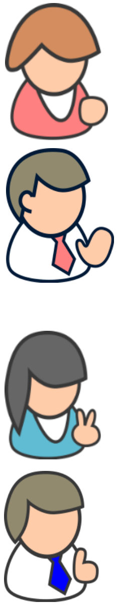
- Signal-to-noise ratio
  - Large scale of noise is required for high-dimensional genomic data
  - Degrade the utility of released data

Our Method

- Key idea: degrade genomic data sensitivity
  - Less noise is required to be injected
- Factorize the high-dimensional distribution of the original genomic data with a set of low-dimensional distributions
  - For low-dimensional distributions, signal-to-noise problem avoided
  - A sufficiently accurate approximation
- Data correlations
  - SNP-trait association
  - Mendelian Inheritance Probabilities

Genome-Wide Association Studies

- Objective of GWAS:
  - Analyze genomic data to find statistical correlations between SNPs and trait (e.g., disease)
  - Compare the genomes of patients with trait and the genomes of patients without trait (e.g., disease)



Case group:
with disease

AACTGTCCG

ACCTGTACG

Control group:
without disease

AATTGTACA

AATTGTCCA

Mendelian Inheritance Probabilities

- A child inherits one allele from mother and one from father.
- Each allele of a parent is inherited by a child with equal probability of 0.5.
- SNP position: BB, Bb, or bb.

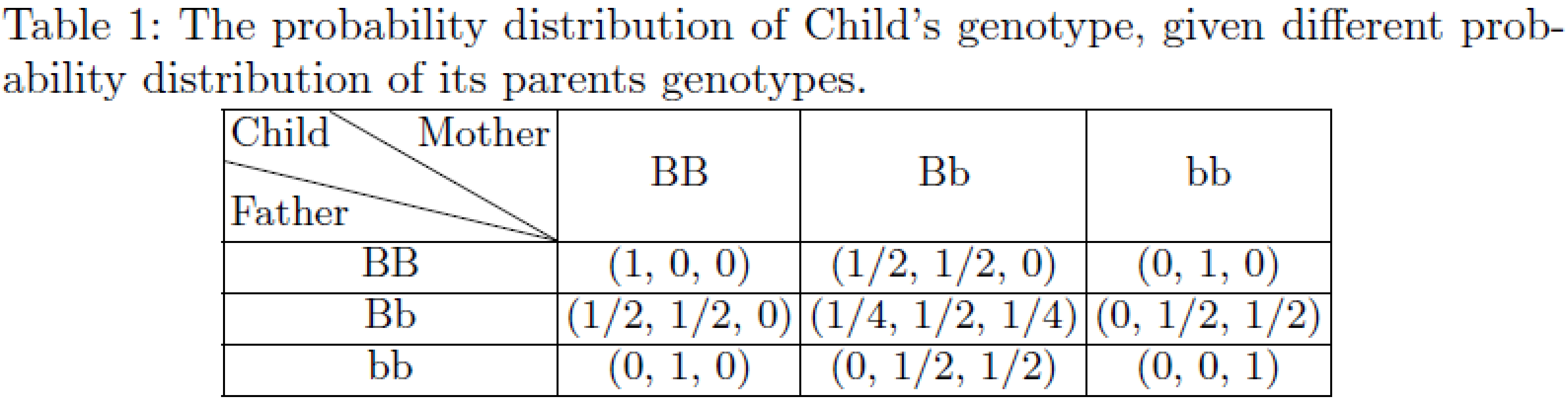
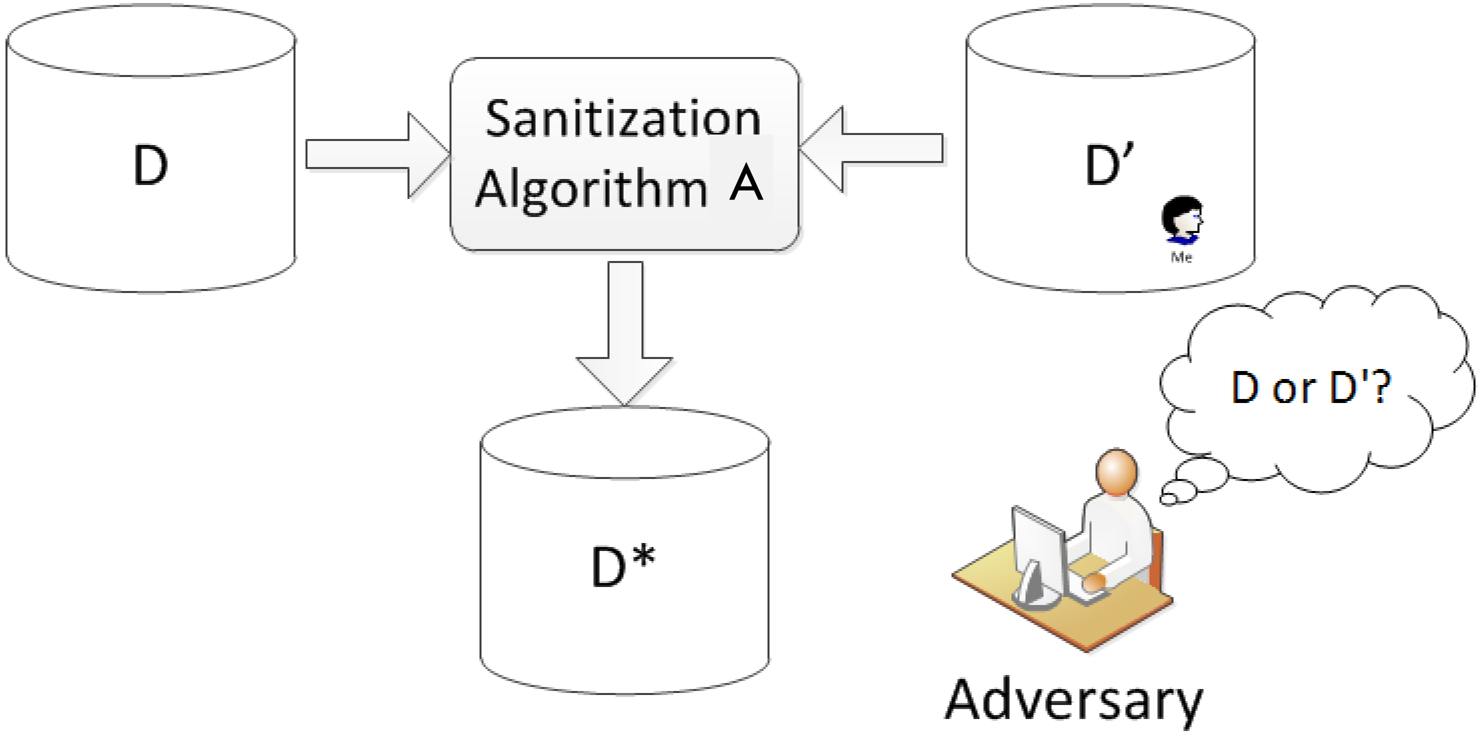


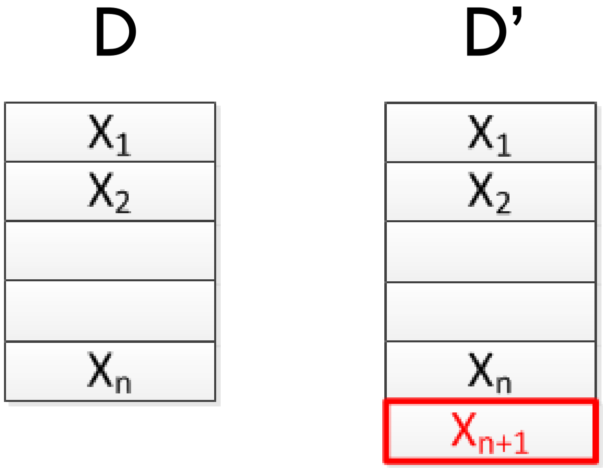
Table 1: The probability distribution of Child's genotype, given different probability distribution of its parents genotypes.

Child Mother Father	BB	Bb	bb
BB	(1, 0, 0)	(1/2, 1/2, 0)	(0, 1, 0)
Bb	(1/2, 1/2, 0)	(1/4, 1/2, 1/4)	(0, 1/2, 1/2)
bb	(0, 1, 0)	(0, 1/2, 1/2)	(0, 0, 1)

Differential Privacy [DMNS, TCC 06]



Differential Privacy

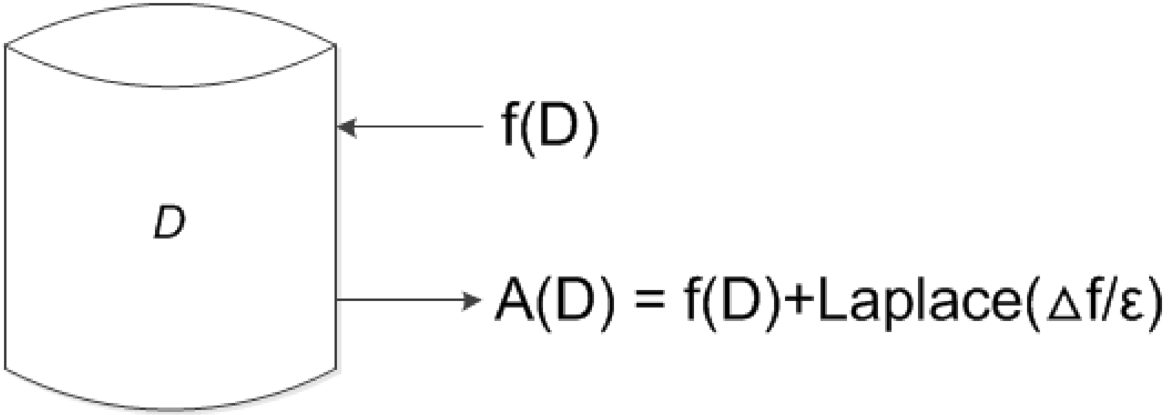


D and *D'* are neighbors if they differ on at most one record

A randomized algorithm *A* satisfies *ε*-differential privacy, if for any two neighbours *D* and *D'*, and for any possible output O of *A*, we have:

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

Laplace Mechanism

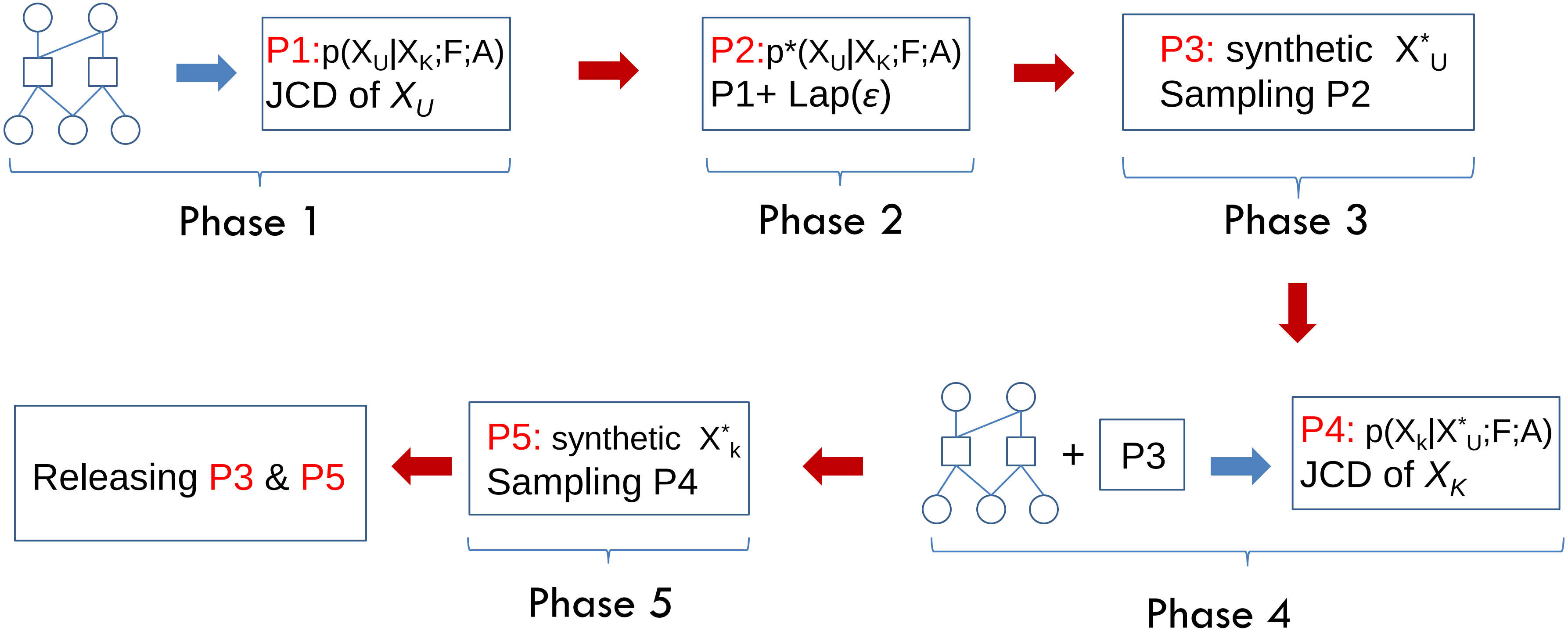


$\Delta f = \max_{D,\, D'} ||f(D) - f(D')||_1$

For a counting query f: $\Delta f = 1$

- E.g., for counting query Q over dataset *D*, returning Q(*D*) + Laplace(1/*ε*) maintains ε-differential privacy.

Our Method: High-Level Overview



P1: $p(X_U|X_K;F;A)$ JCD of X_U

P2: $p^*(X_U|X_K;F;A)$ P1+ Lap(ε)

P3: synthetic X^*_U Sampling P2

+ P3

P4: $p(X_k|X^*_U;F;A)$ JCD of X_K

P5: synthetic X^*_k Sampling P4

Releasing P3 & P5

JCD: Joint Conditional Distribution

X_U: sensitive variables

X_K: non-sensitive variables

F: Mendelian inheritance probabilities

A: SNP-trait associations

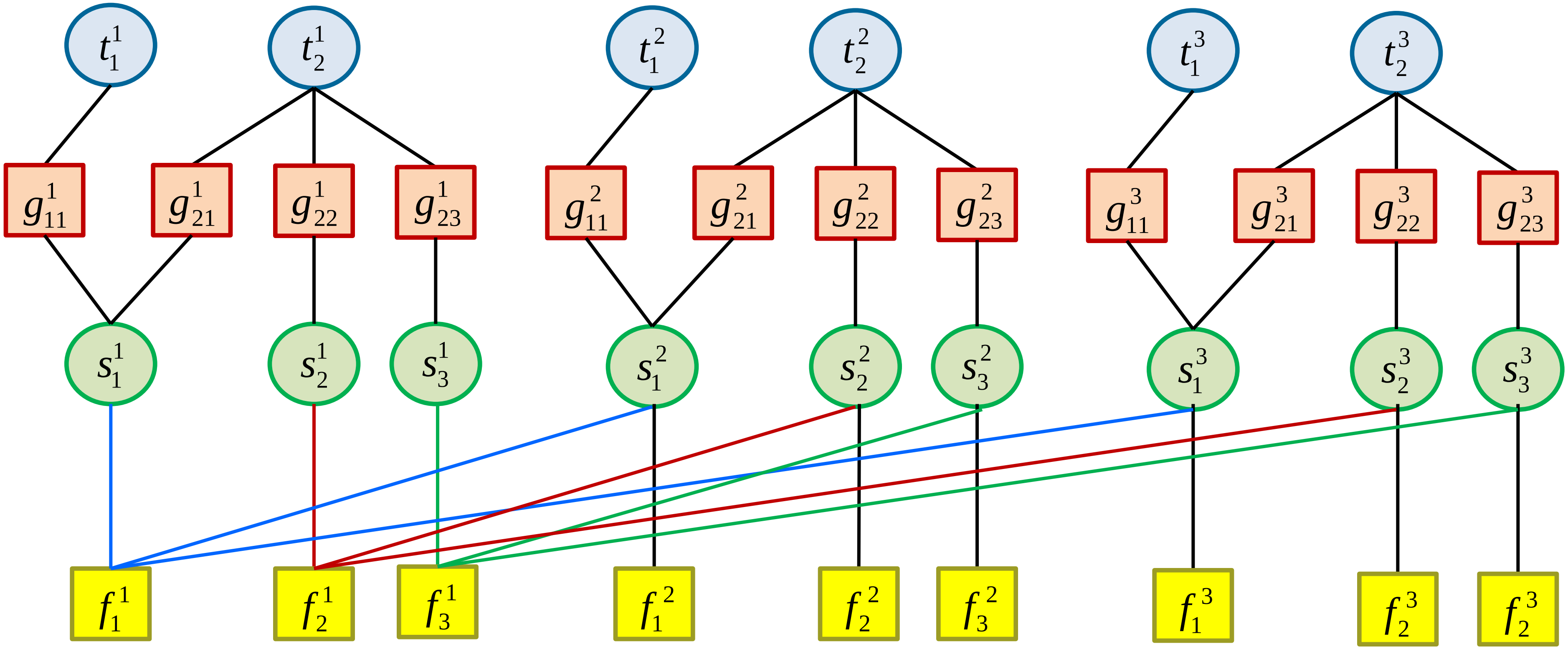
Two Challenges

- Computation of JCD is non-trivial, considering the scale of human genomes
  - Tens of millions of SNPs
  - Large scale of potential traits
- Inject differential privacy noise into genomic data (including SNPs and traits) to derive a close approximation
  - Data privacy: poor scalability; expensive
  - Data utility: high sensitivity

Solutions

- Factorize the high-dimensional JCD into the product of simpler local functions
  - Capturing data correlations (SNP-trait associations; Mendelian inheritance probabilities)
- Inject differential privacy noise into these local functions
  - Low-dimensional local functions incur low sensitivity

Computation of JCD



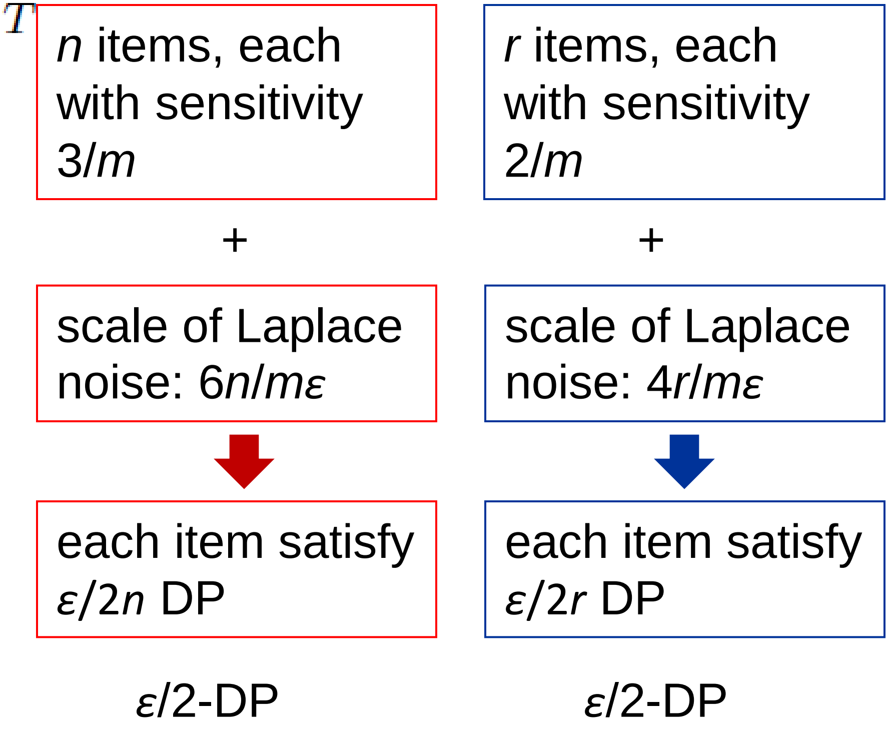
$$p(X_U|X_K,\mathcal{F},\mathcal{A}) = \frac{1}{Z}\prod_{i\in S}\prod_{j\in T} f_i(s_i^C, s_i^F, s_i^M, \mathcal{F})\, g_{ij}(s_i, t_j, \mathcal{A})$$

Belief propagation in a factor graph obtains exponential gains in efficiency.

Injection of Differential Privacy Noise

$$p(X_U|X_K,\mathcal{F},\mathcal{A}) = \frac{1}{Z}\prod_{i\in S}\prod_{j\in T}\boxed{f_i(s_i^C,s_i^F,s_i^M,\mathcal{F})}\boxed{g_{ij}(s_i,t_j,\mathcal{A})}$$

To construct approximate distribution p*($X_U|X_K$, *F*, *A*).
m: # individuals in the target family
n: # SNPs
r: # traits



n items, each with sensitivity 3/*m*

+

scale of Laplace noise: 6*n*/*mε*

each item satisfy *ε*/2*n* DP

ε/2-DP

r items, each with sensitivity 2/*m*

+

scale of Laplace noise: 4*r*/*mε*

each item satisfy *ε*/2*r* DP

ε/2-DP

Compensability property:
ε-DP is satisfied!

Conclusions

- Differential-privacy preserving kin-genomic data releasing

- Key ideas of the solution
  - Belief propagation in a factor graph for dimension reduction
  - Differential privacy noise directly injected into low-dimensional local distributions

THANK YOU!

Q&A