Building A PPML Application with HE

Dec 3rd, 2019

Wei Dai

Disclaimer: This is not a serious presentation. Read it for fun.

Find Ideas

Domain	Genomics	Health	National Security	Education	Social Security	Business Analytics	Cloud
Sample Topics	match maker	billing and reporting	smart grid	school dropouts	credit history	prediction	storage, sharing
Data Owner	medical institutions	clinics and hospitals	nodes and network	schools, welfare	government	business owners	clients
Why HE?	HIPAA	cyber insurance	privacy	FERPA	cyber crimes	data are valuable	untrusted server
Who pays?	health insurance	hospital	energy company	DoE	government	business owners	clients

homomorphicencryption.org

From News

- Strava, a fitness tracking app, released a heat map of users' activity.
- Human activities in Afghanistan reveals US military bases.
- Problem:
  - National security
  - No merger, acquisition, or takeover by a foreign entity
  - The company knows your daily commute/exercise route
  - People with access to database can lookup acquaintance's data
  - You receive targeted ads
  - More than route, heart rates are also collected
  - Etc.

Solutions

- Turn off data sharing → “A Turkish soldier forgot to turn off GPS.”
- Turn off data sharing by default → “How was my runs this month?”
- Take down heatmap → Data are private but not removed.
- Disable selected areas → “There is no U.S. base in this area.”
- Reduce accuracy → “There is a U.S. base within 1-mile.”
- Encrypt your data! → How does a service provider perform analysis?

A Service Provider

- Intel SGX (or other TEEs) → Attacks, the master key, memory size, etc.
- MPC → Battery life of a fitness device or a mobile device.
- HE → “It works. You should pay extra to turn off data sharing.”

Users

- Encryption on a fitness band → Incapable
- Encryption on a smart phone → Good
  - AES encryption → Costly to use HE
  - SEAL encryption → Limited data plan
    - Use batching
    - Use symmetric encryption
    - Use compression
- Can track running route
- Can see intensity

Computation

- Geolocations (x_i, y_i, z_i, t_i)
- Distance = $\sum \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2 + (z_i - z_{i+1})^2}$
- Use a linear model to predict intensity
  - Features can involve elevation, speed, duration, heart rate, etc.

- Let's have (x_i, y_i, z_i, t_i) encrypted on users' device before computation on server.

HE Design

- (x_i, y_i, z_i, t_i) → Huge size expansion
- $(x_0, y_0, z_0, t_0, \dots, x_i, y_i, z_i, t_i, \dots)$ → A lot of rotations
- $(x_0, \dots, x_i, \dots)$ → Better

- Square root → Approximation based on distance interval
  - Interval duration is decided by sampling frequency.
  - How far can you run within 1 second?
  - A 5 or 7 degree polynomial is good enough.

- Linear model → Vector and vector multiplication

Other Considerations

- Key generation:
  - Generate secret key on users' phone
  - Generate public key / relinearization key / Galois keys on users' phone
  - Send Galois keys to server can be troublesome
    - Slow
    - Limitation on data upload size (depend on cloud service provider)
- Sending ciphertexts:
  - How often
  - Compression
- Receiving ciphertexts:
  - Batch all results into a single ciphertext

POC

- A smart phone app:
  - Wraps Microsoft SEAL keygenerator/encoder/encryptor/decryptor
  - Reshape collected data and encrypt
  - Able to decrypt
  - Communication layer
  - Visualization: map, charts, etc.
- A cloud server:
  - Wraps Microsoft SEAL encoder/evaluator
  - Communication layer
  - Data storage
  - Computation

Challenges / Roadblockers

- Wrapper of Microsoft SEAL on smart phones
- Where to store secret key on phone
- Remove location info from app communication
- Encryption/sending on battery using cellular data?

Target Users / Use Scenarios

- Runners: track intensity
- Truckers: reminder to rest
- Drivers or passengers or insurance: rate drivers

- Make a story!

Impact

- Insurance company can provide better pricing
- Uber can reward defensive drivers
- Drivers get better policy/salary: xx% drivers get less/more payment

- xx% of accidents are caused by tired driving

- I pretend to workout.

Competitors

- No encryption: trucks carrying valuable assets
- Self-hosted service: not all companies can build cloud service
- MPC: truck drivers need unlimited data plan
- TEE: let’s not go there for this example
- Other HE: compare communication and computation cost

- Disrupters: cartels tracking routes and police checkpoints

Four Pillars

- Novelty: nothing similar in market, hopefully new to judges
- Soundness: server can guess my exercise frequency but not my geolocation
- Feasibility: we have a demo, or algorithm works as explained
- Impact: beneficial to insurance company and drivers