E6895 Advanced Big Data Analytics Lecture 8:

*Encrypted Domain Data Mining*

Ching-Yung Lin, Ph.D.
Adjunct Professor, Dept. of Electrical Engineering and Computer Science
Mgr., Dept. of Network Science and Big Data Analytics, IBM Watson Research Center
Q. How can users contribute their private data without compromising their privacy?
Information leak!!
TOO MUCH INFORMATION
How Do Recommendations Work in End-to-End Encrypted Scheme?

- End to End Encryption
- Similarity Measure in Encrypted Domain
- Encrypted Measurement Result Ranking
- Encrypted Domain Operations
- Secure Communication Channel
How Do Recommendations Work in End-to-End Encrypted Scheme?

- End to End Encryption
- Similarity Measure in Encrypted Domain
- Encrypted Measurement Result Ranking
- Encrypted Domain Operations
- Secure Communication Channel
1. Patients’ diagnosis results encrypted before doctor submits to database.

2. Sends to server via secure channels

III. Privacy preserving ranking

II. Server measures encryption-domain similarity at encrypted database

1. Doctor submits encrypted query to system (Private Information Retrieval)

Encrypted results returned to doctor
1. Patients’ diagnosis results encrypted before doctor submits to database.

2. Sends to server via secure channels.

I. Doctor submits encrypted query to system (Private Information Retrieval).

II. Server measures encryption-domain similarity at encrypted database.

III. Privacy preserving ranking.
How Do Recommendations Work in End-to-End Encrypted Scheme?

- End to End Encryption
- Similarity Measure in Encrypted Domain
- Encrypted Measurement Result Ranking
- Encrypted Domain Operations
- Secure Communication Channel
1. Patients’ diagnosis results encrypted before doctor submits to database.

2. Sends to server via secure channels.

I. Doctor submits encrypted query to provide treatment (Private Information Retrieval)

II. Server measures encryption-domain similarity at encrypted database

III. Privacy preserving ranking

Encrypted results returned to doctor
Using homomorphic encryption algorithm to calculate similarity in encrypted domain according to doctor’s request.
How Do Recommendations in an End-to-End Encrypted Domain?

- End to End Encryption
- Similarity Measure in Encrypted Domain
- Encrypted Measurement Result Ranking
- Encrypted Domain Operations
- Secure Communication Channel

Details To Be Continued Next Week.....
1. Patients’ diagnosis results encrypted before doctor submits to database.

2. Sends to server via secure channels.

I. Doctor submits encrypted query to system (Private Information Retrieval)

II. Server measures encryption-domain similarity at encrypted database

III. Privacy preserving ranking

Encrypted results returned to doctor.
Doctor chooses two private keys to conduct order-preserving encryption on encrypted domain similarity results so proxy can rank the encryption results. The proxy knows only the rank and not the true similarity measurements of each patient.

\[ \varepsilon (R_{1\text{doctor}}) \varepsilon (W_{ak}) + \varepsilon (R_{2\text{doctor}}) = \varepsilon_{op}(W_{ak}) \]
A subgroup of people with similar symptoms are selected at **proxy** and recommended to doctor, who applies **order-preserving encryption keys** to obtain the exact results.
1. Patients’ diagnosis results encrypted before doctor submits to database.

2. Sends to server via secure channels.

I. Doctor submits encrypted query to system (Private Information Retrieval)

II. Server measures encryption-domain similarity at encrypted database

III. Privacy preserving ranking

Encrypted results returned to doctor
How Do Recommendations Work in End-to-End Encrypted Scheme?

- End to End Encryption
- Similarity Measure in Encrypted Domain
- Encrypted Measurement Result Ranking
- Secure Communication Channel
1. Patients’ diagnosis results encrypted before doctor submits to database.

2. Sends to server via secure channels

I. Doctor submits encrypted query to system (Private Information Retrieval)

II. Server measures encryption-domain similarity at encrypted database

III. Privacy preserving ranking

Encrypted results returned to doctor
Encrypted-Domain Operations

- Ring homomorphic encryption
- Privacy-preserving ranking
- Secure channel
Algebraic operation supports both addition and multiplication.

Encrypted $a$ × Encrypted $b$ + Encrypted $c$ = Encrypted $a \times b + c$

Plaintext $a \times b + c$
Ring Homomorphic Encryption

- Polynomial rings setting
- Key generation
- Encryption scores
- Encrypted domain similarity measure
- Decryption results
An Example of Ring Homomorphic Encryption
Encryption Domain Scalar Product

\[ m_1^a m_1^k + m_2^a m_2^k + m_3^a m_3^k = \]

\[ 3 \times 7 + 5 \times 6 + 10 \times 1 = 61 \]
Mapping to Polynomial

\[ m_1^a = 3, \quad m_1^a(x) = x + 1, \]
\[ m_2^a = 5, \quad m_2^a(x) = x^2 + 1, \]
\[ m_3^a = 10, \quad m_3^a(x) = x^3 + x, \]
\[ m_1^k = 7, \quad m_1^k(x) = x^2 + x + 1 \]
\[ m_2^k = 6, \quad m_2^k(x) = x^2 + x \]
\[ m_3^k = 1, \quad m_3^k(x) = 1 \]

\[ m_1^a m_1^k + m_2^a m_2^k + m_3^a m_3^k = 3 \times 7 + 5 \times 6 + 10 \times 1 = 61 \]
Polynomial rings setting

Taking $N=7$, $p = 29$, and $q = 491531$,

\[ f(x) = x^5 - x^4 - x^2 + x + 1, \]
\[ g(x) = -x^6 + x^3 + x^2 - 1, \]
Key generation

The inverses exist in both $R_p[x] = Z_p[x]/(x^N - 1)$ and $R_q[x] = Z_q[x]/(x^N - 1)$

Compute the two inverses of $f(x), F_p(x), F_q(x)$, respectively in $R_p[x]$ and $R_q[x]$

$$f(x)F_p(x) \equiv 1 \mod R_p[x]$$

$$f(x)F_q(x) \equiv 1 \mod R_q[x]$$
Key generation

The inverse $F_p(x)$ of $f(x)$ modulo $p$ is $F_p(x) = 7x^6 + 6x^5 + 22x^4 + 10x^3 + 24x^2 + 4x + 15$

The inverse $F_q(x)$ of $f(x)$ modulo $q$ is $F_q(x) = 359995x^6 + 214612x^5 + 477685x^4 + 48461x^3 + 76153x^2 + 470762x + 318457$. 
Key generation

*Alice* randomly selects a trinary polynomial \( g(x) \in \mathcal{L}_g \) in which \( g(x) \in T(d, d) \) and computes

\[
h(x) = F_q(x)g(x) \mod R_q[x]
\]

*Alice* uses the polynomial \( h(x) \) as her public key and \( f(x) \) and \( F_p(x) \) becomes her private key for decryption.
Public key generation

The public key \( h(x) \) is

\[
h(x) = F_q(x)g(x) \mod R_q[x] = 339225 + 41538x^5 + 346149x^4 + 263073x^3 + 62307x^2 + 27692x + 394609.
\]
Encryption

\[ e_i^a(x) = (p r_i^a(x) h(x) + m_i^a(x)) \mod R_q[x] \]

\[ e_i^k(x) = (p r_i^k(x) h(x) + m_i^k(x)) \mod R_q[x] \]
Encryption

\[ e_i^a(x) = (p_{r_i}^a(x)h(x) + m_i^a(x)) \mod R_q[x] \]

\[ e_i^k(x) = (p_{r_i}^k(x)h(x) + m_i^k(x)) \mod R_q[x] \]
Encryption

\[ m_1^a = 3, \; m_1^a(x) = x + 1, \]
\[ m_2^a = 5, \; m_2^a(x) = x^2 + 1, \]
\[ m_3^a = 10, \; m_3^a(x) = x^3 + x, \]
\[ m_1^k = 7, \; m_1^k(x) = x^2 + x + 1 \]
\[ m_2^k = 6, \; m_2^k(x) = x^2 + x \]
\[ m_3^k = 1, \; m_3^k(x) = 1 \]
Encryption

\[ r_1^a(x) = x^3 + x^2 - x - 1, \quad r_1^k(x) = -x^4 - x^2 + x + 1 \]

\[ r_2^a(x) = -x^6 + x^3 - x + 1, \quad r_2^k(x) = x^6 + x^5 - x^3 - 1 \]

\[ r_3^a(x) = x^6 + x^5 - x^4 - x, \quad r_3^k(x) = x^5 - x^4 - x^3 + 1 \]
Encryption

\[ e_1^a(x) = 235408x^6 + 159221x^5 + 179994x^4 + 353054x^3 + 484568x^2 + 269992x + 283889 \]

\[ e_1^k(x) = 179965x^6 + 353083x^5 + 484626x^4 + 270020x^3 + 283860x^2 + 235380x + 159193 \]

\[ e_2^a(x) = 394600x^6 + 339244x^5 + 41546x^4 + 346120x^3 + 263058x^2 + 62319x + 27708 \]

\[ e_2^k(x) = 186928x^6 + 83034x^5 + 200738x^4 + 34641x^3 + 124639x^2 + 55386x + 297698 \]

\[ e_3^a(x) = 387666x^6 + 117646x^5 + 325376x^4 + 90027x^3 + 422307x^2 + 242314x + 380790 \]

\[ e_3^k(x) = 221511x^6 + 207672x^5 + 256152x^4 + 332339x^3 + 311566x^2 + 138448x + 6906 \]
Encrypted domain similarity measure

\[
(e_1^a(x)e_1^k(x) + e_2^a(x)e_2^k(x) + e_3^a(x)e_3^k(x)) \mod R_q[x]
= 464241x^6 + 345580x^5 + 79051x^4 + 9314x^3
+ 130865x^2 + 131419x + 314135
\]
Apply first private key for decryption

\[ \alpha \equiv f(x)^2 e_i^a(x) e_i^k(x) \mod R_q[x] \]
\[ = (p^2 r_i^a(x) r_i^k(x) g(x)^2 \]
\[ + pf(x) g(x) \left( r_i^a(x) m_i^k(x) + r_i^k(x) m_i^a(x) \right) \]
\[ + f(x)^2 m_i^a(x) m_i^k(x) \mod R_q[x] \]
Apply second private key for decryption

\[
(F_p(x))^2 (p^2 r^a_i(x) r^k_i(x) g(x)^2 \\
+ p f(x) g(x) \left( r^a_i(x) m^k_i(x) + r^k_i(x) m^a_i(x) \right) \\
+ (F_p(x)^2 f(x)^2 m^a_i(x) m^k_i(x)) \mod R_q[x] (\mod p) \\
1
\]

\[m^a_i(x) m^k_i(x)\]
Apply second private key for decryption

\[ \alpha = -22886x^6 - 16282x^5 + 1531x^4 + 18383x^3 + 23015x^2 + 8623x - 12372 \]

\[ F_p(x)^2((-22886x^6 - 16282x^5 + 1531x^4 + 18383x^3 + 23015x^2 + 8623x - 12372) \mod R_q[x]) \mod p \]
\[ = x^4 + 3x^3 + 3x^2 + 4x + 1 \]
Recover the exact plaintext similarity measurement results

\[ x^4 + 3x^3 + 3x^2 + 4x + 1 = 1 \times 2^4 + 3 \times 2^3 + 3 \times 2^2 + 4 \times 2^1 + 1 = 61 \]

\[ m_1^a m_1^k + m_2^a m_2^k + m_3^a m_3^k = 3 \times 7 + 5 \times 6 + 10 \times 1 = 61 \]
Encrypted-Domain Operations

- Ring homomorphic encryption
- Privacy-preserving ranking
- Secure channel
1. Patients’ diagnosis results encrypted before doctor submits to database.

II. Server measures encryption-domain similarity at encrypted database

II. Doctor submits encrypted query to system *(Private Information Retrieval)*

III. Privacy preserving ranking

2. Sends to server via secure channels

Encrypted results returned to doctor

Alice

Bob

Carol

Surgeons

Physicians

Doctors
Using the affine transformation for privacy-preserving ranking

Encrypted Domain Operation

- Ring Homomorphic Encryption
- Privacy Preserving Ranking
- Secure Channel
1. Patients’ diagnosis results encrypted before doctor submits to database.

2. Sends to server via secure channels

I. Doctor submits encrypted query to system (*Private Information Retrieval*)

II. Server measures encryption-domain similarity at encrypted database

III. Privacy preserving ranking

Encrypted results returned to doctor
Diffie-Hellman key exchange scheme for secure channel

A

common paint

$g$

+ secret color A

$r_a$

B

common paint

$g$

+ secret color B

$r_b$

assume that mixture separation is expensive

A

+ secret color A

+$g^{r_a}$

B

+ secret color B

+$g^{r_b}$

public transport

$g^{r_a}g^{r_b}$

common secret

$g^{r_a}g^{r_b}$
Challenge with response for security

Alice

$N_A, r_a, \quad t_a = g^{r_a}$

Bob

$E_B(N_A), t_a$

$k = H(N_A, N_B)$

MAC verification

$S_{AB} = t_b^{r_a}$

$E_A(N_B), t_b$  

$MAC_k(t_a, t_b)$  

$MAC$ verification  

$k = H(N_A, N_B)$

$S_{AB} = t_a^{r_b}$  

$g^{ra} \rightarrow g^{rb}$
Outline

- Introduction
- Privacy-Preserving Recommendation System
- Encrypted Domain Operations
- Experiments
- Conclusion
Experiment Sets

- Prototype system
  - Microsoft Windows 7 (32-bit) operation system
  - Intel Core 2 Quad CPU, and 2 GB RAM.
- Jester dataset from UC Berkeley
- Book-Crossing dataset
Experiments

- Performance
- Accuracy
- Security
Experiments

- Performance
- Accuracy
- Security
Experiments

- Performance
- Accuracy
- Security

Dot Products in Random Perturbation

\[ A' \cdot B' = \sum_{i=1}^{n} (a_i + r_i)(b_i + v_i) \]

\[ = \sum_{i=1}^{n} a_i b_i + \sum_{i=1}^{n} a_i v_i + \sum_{i=1}^{n} r_i b_i + \sum_{i=1}^{n} r_i v_i \]

≈ \sum_{i=1}^{n} a_i b_i

Zero Mean
Drawbacks of User Profile Distribution

- Distribution
- Clustering based approaches

Experiments

- Performance
- Accuracy
- Security
<table>
<thead>
<tr>
<th></th>
<th>Our approach</th>
<th>Randomized perturbation</th>
<th>User profile distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vulnerable to attack</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Trusted third party</td>
<td>No</td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td>Accuracy</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Security</td>
<td>High</td>
<td>Fair</td>
<td>Fair</td>
</tr>
<tr>
<td>End-to-end encryption</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Efficacy</td>
<td>Low</td>
<td>High</td>
<td>Fair</td>
</tr>
</tbody>
</table>
Contributions

- Ring homomorphism operation
- Secret sharing key management
- End-to-end encryption
Conclusion

- High accuracy recommendation
  - Lossless number theory approaches
  - Statistical approaches
- Practical and implementable
- Well-suited for cloud computing applications
- Highly Secure
  - Post-quantum cryptography
Future Work

- More complex mathematical operations in the encryption domain
- Robust privacy-preserving data mining
Multimodality Intelligent Sensor for Homecare

Sensors in the home to detect or prevent accident

Fall down
Stroke
Tension
Sleep monitoring

S: sensors
Multimodality Intelligent Sensor for Homecare

Long-Term Monitoring and Logging of Personal Activities for Chronic Disease and Health Maintenance

S: sensors

Watching TV

Talking on the phone

Sleep monitoring
Intelligent Multimodality Sensor

Commodity:
- Fixed Low-Cost Multimodality Recognition Sensor installed under the ceiling or mounted on the wall.
- A Replacement or extension of the smoke detector, intruder detector, baby monitor, etc.

Technical Innovation:
- Distributed Intelligence:
  - Integrated recognition-driven feature extraction module in the sensor
  - Only transmitting required features through wireless channel (e.g., MVs, Color histogram, Sound MFCC coefficients, etc.)
  - Separate inference engines for behavior and event recognition.

Benefits:
- Low-Energy Consumption
- Low Data Transmission Bandwidth
- High Privacy Protection
Intelligent Multimodality Behavior-Recognition Software Engine

**Commodity:**
- Extensible software applications residing in PC or other computing devices for behavior or event inference.
- Standard-compliant wireless signal receiver for interfacing with the wireless sensor.

**Technical Innovation:**
- **Distributed Intelligence:**
  - Developing Machine Learning and Data Mining Algorithms for Human Behavior or Environmental Context Recognition
  - Recognition is based on the received feature signals

**Benefits:**
- **Scalability:** Consumers can buy any combination of software modules for different applications – surveillance, healthcare, etc.
- **Low Maintenance Cost**
Simple Multimodality Sensors for Sleep Situation Inference [Peng et al., ISCAS 2006] [Peng et al., D2H2 2006]

- Understand human night-time activity – *Sleep*

- What we have achieved:
  - Using visual-audio sensors to monitor a person’s sleep patterns
  - Measurement of sleep quality
Develop Hardware Multimodality Semantic Sensor

- Extract Features of Audio/Visual/PIR Signals using FPGA.

e.g. video smart sensor
Issues about Sensor Network

- The coverage range (required number) of sensors installed in the target environment

- The power consumption of mobile sensors; Different strengths (power fading) of transmitted signals from mobile sensors in different regions

- Hand-off of signals/communication in mobile sensor networks

- Information Fusion of Sensors

- Data storage, management, and retrieval in sensor networks
Issues about Activity Understanding

- How to learn human activity automatically – Need to develop novel learning algorithms

- Hybrid choices of supervised learning and unsupervised learning

- Frameworks for activity object tracking, multimodality feature extraction.

- Multimodality Joint Activity and Behavior Understanding

- Retrieval and Logging of Activity Understanding
Distributed Sensor Information Processing, Mining and Management

- **Distributed Signal Processing:**
  - Decide the optimal solutions to distribute the stages of recognition modules into multiple sensors and servers based on resource constraints

- **Sensor Information Management:**
  - Record, Index, and Manage the recognition results and the original sensor signals for activity mining

- **Sensor Network Information Aggregation:**
  - Federated information reasoning based on aggregated recognition results from sensors and sensor networks at multiple sites
Sleep-Monitoring and Sensors

- Sleep Research Laboratory equipped with polysomnographic sleep recording system
  (http://www.son.washington.edu/departments/bnhs/bnhs-tg/sleep-lab.asp)
Sleep Monitoring

- Sleep occupies almost $\frac{1}{3}$ of human life!

- Sleep deprivation due to sleep-related disorders may introduce severe health impairments [WHO 2004]

- Long-term sleep monitoring can provide information for detection of early symptoms of sleep related disorders, and response to treatment
Previous Studies in Sleep Monitoring

- Previous works in sleep monitoring
  - Static charge sensitive bed, heuristic thresholding, detection of wake, quiet-sleep, and active-sleep [Salmi et al. 1986]
  - EEG, HMM, detection of arousal [Huang et al. 1996]
  - Air cushion, finite-state machine, detection of 6-stages of sleep [Watanabe et al. 2004]

- Specially developed sensors
Traditional Methods to Measure Sleep (1)

- **Subjective approach**

  - Self-rated questionnaires and sleep diaries
    - **Pittsburgh Sleep Quality Index (PSQI)**
      - How many hours of actual sleep did you get at night?
      - Cough or snore loudly?
  
  - Limited capabilities
  - Less reliable than objective approach
Traditional Methods to Measure Sleep (2)

- Objective approach
  - Polysomnography (PSG)
    - EEG, ECG, EMG, and EOG
    - Data scored by human examiners
    - ~$1000/study
    - Intrusive
  - Actigraphy
    - Limb movement via accelerometer
    - Less-intrusive than PSG
    - >$1000
Innovation and Advantage

- **Single vs. Multimodality**
  - Using multimodality data fusion, the accuracy can be increased or the FA can be reduced compared to previous works using single data modality
  - Example: sleep-wake detection using PIR & HR sensors

<table>
<thead>
<tr>
<th>Modality</th>
<th>Fusion</th>
<th>Motion only</th>
<th>Heart-rate only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miss Rate</td>
<td>0.1610</td>
<td>0.1867</td>
<td>0.0737</td>
</tr>
<tr>
<td>FA</td>
<td>0.0615</td>
<td>0.3604</td>
<td>0.2693</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.9282</td>
<td>0.6497</td>
<td>0.7454</td>
</tr>
</tbody>
</table>

- **Specific vs. Generic framework**
  - Compared to previous works which impose many conditions and thus are specific to certain activity recognition, our method would be a framework for general activities understanding
Publications


