Secure Data Processing

Prof. Dr. E. Rahm und Mitarbeiter

Seminar, WS 2018/19

NEW CHALLENGES

- Cloud
- Social media
- CryptoCurrencies
OUTSOURCED PRIVATE DATA

Security Concerns?
Confidentiality of Data

Solution:
Encryption

BUILDING TOOLS: ENCRYPTION

Deterministic
AES + EBC
Electronic Codebook Mode

Non-deterministic
AES + CBC
Cipher Block Chaining Mode
**SECURE SQL?**

**GOAL:** Developing algorithms that can answer queries over securely outsourced data without fetching all data

```sql
SELECT SUM(price) AS total
FROM orders
WHERE 10 <= price AND city = 'Vienna'
GROUP BY order_id
HAVING total > 20
```

- comparison of entries for equality
- keyword search: search for pattern
- range query: comparison of numerical value
- aggregation

---

**ENCRYPTION**

- **Homomorphic encryption:** allows some computations on ciphertext without decrypting
  - partially homomorphic (e.g. additive / multiplicative)
  - fully homomorphic (quite inefficient)

- **Order-preserving encryption** (Agrawal et al, Sigmod2004)
  - standard database indexes can be used
  - vulnerable to statistical attacks
Microsoft “Always Encrypted“ database (based on Cipherbase)

Privacy
- right of individuals to determine by themselves when, how and to what extent information about them is communicated to others (Agrawal 2002)

Privacy threats
- extensive collection of personal/private information / surveillance
- Information dissemination: disclosure of sensitive/confidential information
- Invasions of privacy: intrusion attacks to obtain access to private information
- Information aggregation: combining data, e.g., to enhance personal profiles or identify persons (de-anonymization)

Challenge:
- preserve privacy despite need to use person-related data for improved data analysis
 PRIVACY FOR BIG DATA

- need for comprehensive privacy support ("privacy by design")
- privacy-preserving publishing of datasets
  - anonymization of datasets
- privacy-preserving data mining
  - analysis of anonymized data without re-identification
- privacy-preserving record linkage
  - object matching with encoded data to preserve privacy
  - prerequisite for privacy-preserving data mining

RE-IDENTIFICATION OF „ANONYMOUS DATA“ (SWEENEY 2001)

- US voter registration data
  - 69% unique on postal code (ZIP) and birth date
  - 87% US-wide with sex, postal code and birth data

- solution approach: K-Anonymity
  - any combination of values appears at least k times
  - generalize values, e.g., on ZIP or birth date
**DIFFERENTIAL PRIVACY**

- Statistical approach to derive accurate analysis results despite systematic changes to data/query answers, e.g. randomized response
  
- Example: randomize Yes/No answer (e.g. „Have you ever used illegal drugs?“)
  - throw coin - if head: answer correctly, if tail: throw coin again and answer correctly if head
  
- Most approaches assume trusted central party
  - Has access to raw data and performs data perturbation for query answers

- **Local differential privacy**: eliminate trusted party
  
  - Google Rappor prototype for browser accesses
  - Apple smartphones (since iOS 10)

---

**PRIVACY-PRESERVING DATA MINING**

- Physically integrated data (e.g. data warehouse) about persons entails greatest privacy risks

- Data mining over distributed data can better protect personal data by limiting data exchange, e.g., using SMC (secure multiparty computation) methods
compute a function across several parties, such as no party learns the information from the other parties, but all receive the final results

example 1: millionaire problem
- two millionaires, Alice and Bob, are interested in knowing which of them is richer but without revealing their actual wealth.

example 2: secure summation

Step 0: Z = 999
Party 1: x₁ = 55
Step 1: Z + x₁ = 1054

Step 4: s = 1169 - Z = 170
Party 2: x₂ = 73
Step 2: (Z + x₁) + x₂ = 1127

Step 3: ((Z + x₁) + x₂) + x₃ = 1169
Party 3: x₃ = 42

SECURE MULTI-PARTY COMPUTATION (SMC)

Bitcoin & Blockchain
26.10.2018

- decentralized, peer-to-peer electronic cash system
- digital identities/signatures
  - public/private key pair
- ledger
  - the balance of each identity – saved in a blockchain (instead of a central bank database)
- transactions
  - move bitcoins from one to another
  - concurrency control to serialize transactions
  - typically backed by transaction log (blockchain)
- log is persistent (replicated across all network nodes), immutable and tamper-free

BITCOIN

- a bitcoin is a chain of digital signatures
- coin owners digitally sign their coins to transfer them to other recipients
- Alice wants to move a bitcoin to Bob

DIGITAL SIGNATURES AND BITCOIN

[Diagram showing the process of signing and verifying signatures]
**HASHING H(X)**

- Signatures and public keys are combined using hashing
- Takes any string x of any length as input
- Fixed output size (e.g., 256 bits)
- Efficiently computable
- Bitcoin uses SHA-256
  
  \[
  \text{SHA256}(\text{Signature}_{\text{Alice-Bob}} \ || \ P_{k_{\text{Diana}}}) = \text{256-bit (32-byte) unique string}
  \]

- Satisfies:
  - **Collision free**: No two x, y s.t. H(x) = H(y)
  - **Hiding**: Given H(x) infeasible to find x (one-way hash function)

---

**WHAT IS THE LEDGER?**

- Blockchain
- Transactions are grouped into blocks
  - Blocks are chained to each other through pointers (hence blockchain)
Where is the ledger stored?
- each network node maintains its copy of the ledger

How is the ledger tamper-free?
- blocks are connected through hash-pointers
  - each block contains the hash of the previous block
  - this hash gives each block its location in the blockchain
  - tampering with the content of any block can easily be detected
SEMINAR GOALS

- Learn to know about a new topic of scientific and practical importance
  - can be basis for bachelor/master thesis
- student tasks
  - study scientific literature to prepare presentation and written summary on 1 topic
  - presentation
  - discussion
  - summarizing article
- mentoring co-worker provide help and feedback

SEMINAR

- Master Informatik
  - part of module „Modern database technologies“
  - seminar module
- Bachelor Informatik
  - seminar module
SEMINAR DETAILS

- presentation with discussion (45 minutes)
  - slides should be in English
  - talk in German or English
  - discuss slides with mentor beforehand
- article/report (ca. 15 pages)
  - discuss/iterate with mentor
  - final deadline March 31, 2019
- active participation in all presentations
  - module workload: 30 h presence, 120 h self study
- successful seminar requires both: talk/discussion + report

SEMINAR (3)

- Topic assignment
  - meet mentor within two weeks, i.e., until Nov. 9th, 2018
  - otherwise seminar registration will become void
  - voluntary leave also until Nov. 9th, 2018
- Presentation dates
  - fridays, Ritterstr, starting Jan. 11th 2019
  - max. 4 presentations starting at 1:30 pm
<table>
<thead>
<tr>
<th>Themen</th>
<th>Betreuer</th>
<th>#Themen</th>
<th>Termin</th>
<th>Studenten</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td>Kricke</td>
<td>2</td>
<td>11.10.</td>
<td>Fuchs</td>
</tr>
<tr>
<td>• intro security/encryption</td>
<td>Christen</td>
<td></td>
<td></td>
<td>Wegner</td>
</tr>
<tr>
<td>• public key &amp; one-way encryption</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Secure knowledge bases</strong></td>
<td>Sehili/F S/Franke Franke Sehili</td>
<td>5</td>
<td>18.10.</td>
<td>Kühl/ste Möller</td>
</tr>
<tr>
<td>• searchable encryption /secure indexes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• privacy-preserving fuzzy search</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• priv.-pres. search of chemical compounds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• encryption on labeled graphs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• encrypted dbs / Cipherbase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Privacy-preserving data mining</strong></td>
<td>Christen Christen Rost Rostami Zschache Zschache</td>
<td>7</td>
<td>18.11.</td>
<td>Möller</td>
</tr>
<tr>
<td>• methods, metrics, applications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• RAPPORT (randomized aggregation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• calibrating noise in private data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• LinkMirage: pr.-pres. an. of social relations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Cryptonets: neural networks on encr. data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Differential privacy for SQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• privacy-preserving deep learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Intrusion detection</strong></td>
<td>Grimmer/K G/Kricke Gomez Nentwig</td>
<td>4</td>
<td>1.12.</td>
<td>Kügelke/Kleber Nimmergau/Klein Lege Sager</td>
</tr>
<tr>
<td>• Host Intrusion Detection Systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• HIDS: sequence / argument analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• secure data processing in the cloud &amp; SGX</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• SpnPectre attacks via speculative execution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Blockchain</strong></td>
<td>Peukert Saeedi Alkhouri Wilke</td>
<td>4</td>
<td>8.2.</td>
<td>Musial/Krigger Ho</td>
</tr>
<tr>
<td>• introduction crypocurrency blockchains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• scalability of blockchains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• using blockchains for smart contracts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• risks of blockchain technology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>