Practical linking of databases using secure multiparty computation

Riivo Talviste
riivo@cyber.ee
Cybernetica, Team Sharemind

February 21, 2014
Problem statement

• State is interested in data-driven decisions, thus needs to analyze combined data

• Databases contain sensitive information
• Combined “super databases” are risky
• Replace combined databases with MPC
Practical application

• 2013: Income analysis of public sector
  • Data sources: local governments and ministries
  • All sources have same data structure:
    (job_title, count, salary)

• Web-based MPC application running on public cloud
Practical application

• 2013: Income analysis of public sector
  • Data sources: local governments and ministries
  • All sources have same data structure:
    (job_title, count, salary)

• Web-based MPC application running on public cloud
  • Public demo: https://sharemind.cyber.ee/clouddemo/
Practical application: PRIST

- More complicated scenario: different data structure
- Universities vs. companies: *Does employment during studies have a negative effect?*
- PRIST: “Privacy-preserving statistical studies on linked databases”
  - Answer that question for information technology
  - How? Link income data from Tax Office with education data from Education Information System
**Practical application: PRIST**

- Classical approach
  - Requires approval from Data Protection Agency
  - Tax Office gives out data with $k$-anonymity
  - Data loss with grouping education info, sex, age is 76 – 98%

---

**No approval required as no sensitive information is gathered.**

**MPC platform: Sharemind Application Server**
Practical application: PRIST

- Classical approach
  - Requires approval from Data Protection Agency
  - Tax Office gives out data with $k$-anonymity
  - Data loss with grouping education info, sex, age is 76 – 98%

- Using secure multiparty computation
  - Input data sets are secret shared
  - Data sets are linked using privacy-preserving database linking using personal ID codes
  - Data Protection Agency: “No approval required as no sensitive information is gathered.”
  - MPC platform: Sharemind Application Server
• Modular design with **protection domains**
  • e.g. additively shared 2-party with active security, \( n \)-party with Shamir sharing

• Additive 3-party protection domain:
  • Data types: \([u]int\{8,16,32,64\}\), boolean, float\{32,64\}, strings (known- and bounded-length)
  • Oblivious sorting (sorting networks, radix sort [BLT13]; quicksort \([H^{+}12]\))
  • Oblivious shuffle [LWZ11]
  • Privacy-preserving database linking (SQL equi-join) [LTW13]

---


• Data persistence layer
• Programmable
Hybrid model

// Import a PDK module called 'additive3pp'
import additive3pp;

// Create a domain 'private' from the PDK
domain private additive3pp;

void main () {
    // Perform secure computations using the PDK
    private int a = 2, b = 3;
    private int c = a * b;
    // We need a special function to publish 'c'
    print (declassify (c));
}
// Protection domain kind polymorphism

template <domain D>
D uint sum (D uint [[1]] vector) {
    D uint result = 0;
    for (uint i = 0; i < size (vector); i++) {
        result [i] = result [i] + vector [i];
    }
    return result;
}

// Specialization to a PDK

template <domain D: additive3pp>
D uint sum (D uint [[1]] vec) {
    D uint result = 0;
    __syscall ("additive3pp::sum_uint64_vec",
               __domainid (D),
               vec, result);
    return result;
}
SecreC standard library

- EU-funded project “Usable and Efficient Secure Multiparty Computation” (UaESMC)
- Statistics suite
  - table filtering, linking and sorting,
  - descriptive statistics (mean, variance, standard deviation),
  - percentiles (minimum, maximum, mean, custom percentiles),
  - five-number summary and box-plots, histograms,
  - t-tests, paired t-tests, $\chi^2$-tests, Wilcoxon tests.
- All algorithms support oblivious filters
- Build an R-like statistics application
- Why statistics?
  - We performed 25 interviews internationally
  - We explained what MPC can do and asked where it could be applied?
  - Most popular answers: statistics and optimization
Thank you!

https://sharemind.cyber.ee

The work of Riivo Talviste is supported by European Social Fund Doctoral Studies and Internationalisation Programme DoRa.

“Privacy-preserving statistical studies on linked databases” (PRIST) project is funded by the European Regional Development Fund through the Implementing Agency Archimedes Foundation. http://cyber.ee/en/research/research-projects/prist/

“Usable and Efficient Secure Multiparty Computation” (UaESMC) project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-284731. http://www.usable-security.eu/en