Computing a quantitative score for privacy

Imran Khan
imran.khan@insight-centre.org

Insight
Centre for Data Analytics

1. Introduction

- Organizations are exceedingly becoming prudent about customers privacy. However, these organizations need to share application data for a variety of reasons, including, for informed decision making, for the purpose of revenue generation, due to legal obligation or in general, for predictive analytics.
- To obtain richer insights from organizations application data, organizations frequently delegate this task to third parties who specializes in data analytics.
- Little attention has been paid to the measurement of risk to privacy in Database Management Systems, despite their prevalence as a modality of data access.
- A real-time measure of the (privacy) risk associated with queries by third parties against a Relational Database is desirable.
- The privacy score is a measure of the degree to which attribute values, retrieved by a principal engaging in an interactive query session, represent a reduction of privacy with respect to the attribute values previously retrieved by the principal.

2. Query Abstraction

i	Query (Q_i)	SQL Query Abstraction A(Q_i)
1	SELECT lastName, gender, city FROM hospitaldb WHERE firstName = 'Bob';	{SELECT, lastName, gender, city, hospitaldb, firstName_w}
2	SELECT lastName, gender FROM hospitaldb WHERE firstName = 'Bob';	{SELECT, lastName, gender, hospitaldb, firstName_w}
3	SELECT lastName, gender, department FROM hospitaldb;	{SELECT, lastName, gender, department, hospitaldb}
4	SELECT lastName, department FROM hospitaldb WHERE firstName = 'Alice' ;	{SELECT , lastName, department, hospitaldb , firstName_w}

Figure 1: Examples of SQL query is abstraction.

3. The Model

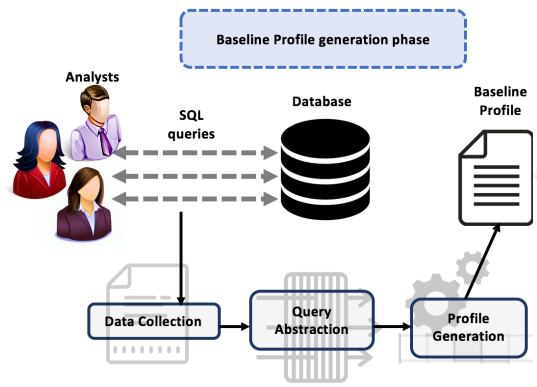

Figure 2: The architecture of privacy score systems: baseline profile generation phase. The first step is of data collection followed by query abstraction stage and subsequently the step a baseline profile generation.

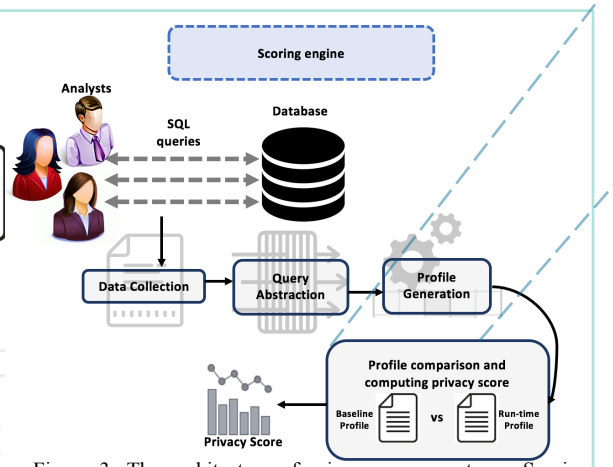

Figure 3: The architecture of privacy score systems: Scoring engine. The first three steps in scoring engine are same as are in first phase. The output of profile generation is a run-time profile.

4. Computing the Score

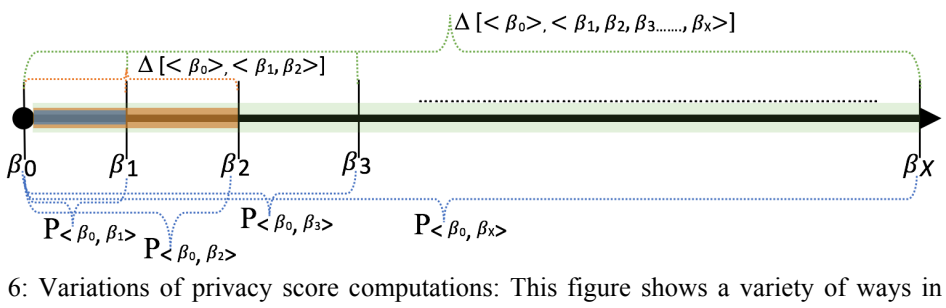

Figure 6: Variations of privacy score computations: This figure shows a variety of ways in which individual privacy scores and cumulative privacy scores can be computed. For example, Δ[< β_0 >, < β_1, β_2 >] represents the cumulative privacy for day 1 and day 2. P< β_0, β_3 > represents the individual privacy score for day 3.

- Evaluation:

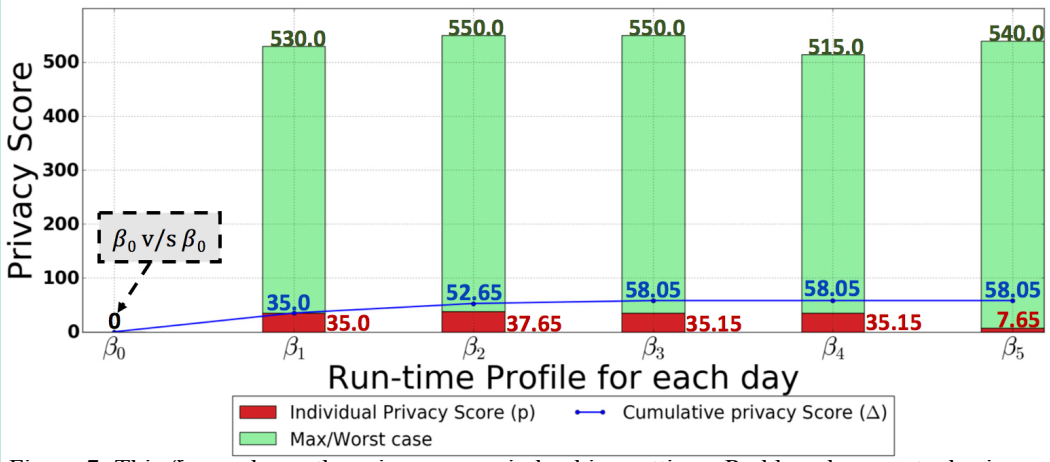

Figure 7: This figure shows the privacy score in banking settings. Red bar shows actual privacy score. Green bar in the figure indicates the maximum possible privacy score $\mathcal{M}$. Blue line shows the cumulative privacy score.

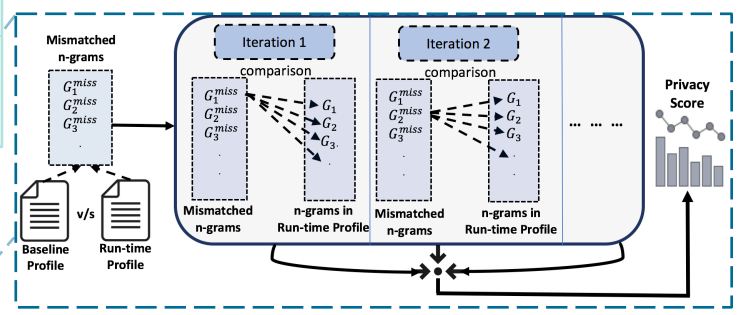

Figure 4: The run-time profile is compared with the baseline profile resulting in a privacy score. Each n-gram from the set of mismatched n-grams is compared with each n-gram in the baseline profile.

5. Privacy Equivalence Relation ($\overset{p}{\equiv}$)

- Cases where one SQL query is a subset of another SQL query in-terms of privacy.
- To unravel this we proposed an *Privacy-equivalence relation* based on *Privacy-Aware attribute relationship diagram* using on Discrimination Rate Privacy metric.
- A privacy equivalence relation based on Discrimination Rate (DR) Privacy metric enables to compare two queries in-terms of privacy.
- DR for an attribute is given by:

$$DR_X(Y) = 1 - \frac{H(X|Y)}{H(X)} \quad (1)$$

$$H(X) = -\sum_{x \in \mathcal{S}} p(x) log(p(x)) \quad (2) \qquad H(X|Y) = -\sum_{x \in \mathcal{S}_1} \sum_{y \in \mathcal{S}_2} p(x,y) log(p(x|y)) \quad (3)$$

- DR for a combination of attributes is given by:

$$CDR_X(Y_1, Y_2, \ldots Y_n) = 1 - \frac{H(X|Y_1, Y_2, \ldots Y_n)}{H(X)} \quad (4)$$

- Identification capability of SQL query is computed as $DRSQL(A(Q_i)) = CDR_X(A(Q_i))$.

firstName	lastName	gender	department	city	departmentHead
John	Smith	Male	Oncology	Vancouver	Dr.George
Bob	Lopez	Male	Oncology	Vancouver	Dr.George
Alice	Miller	Female	Oncology	Vancouver	Dr.George
Bob	Smith	Male	Cardiology	Vancouver	Dr.Albert
John	Wilson	Male	Cardiology	Vancouver	Dr.Albert

Figure 5: An example table having records of five individuals.

5. Related work

- Majority of literature in context of privacy research focuses on anonymizing the data in the database - several privacy definitions, including, *k*-anonymity, l-diversity, t-closeness and differential privacy.
- Syntactic definition of privacy deals with Privacy-preserving data publishing where one can use them to anonymize the data and subsequently publish it thus preserving individuals privacy. Differential privacy is deigned for statistical databases - allows aggregate queries.
- The literature lacks on approaches that measure privacy in interactive settings within the RDBMS framework that provides a quantitative score enabling the concerned to monitor and detect any exploitation of their application data.

This material is based upon works supported by the Science Foundation Ireland under Grant No. 12/ RC/ 2289 which is co-funded under the European Regional Development Fund.

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN Science Foundation Ireland sfi For what's next