A Privacy Preserving Framework for Accuracy and Completeness Quality Assessment

Daniele Barone, Andrea Maurino, Fabio Stella, Carlo Batini

Sequoias group – SErvices and Quality Oriented InformAtion Systems
DISCo - Dipartimento di Informatica Sistemistica e Comunicazione
Università di Milano-Bicocca

Outline

- Motivation
- Main definitions
- Problem statement
- The P^{2}QA protocol
- Experimental Evaluation
- Conclusion

SEQUOIAS

DISCO
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Motivation

- The assessment process requires the data to be disclosed to the third-party.
- Often data disclosure to a third party is not allowed
  - ✓ business and legal reasons.
- **Cryptographic techniques do not be used**
  - ✓ **They do not preserve syntactic distances.**
- **There is the need of a new privacy preserving assessment technique**

SEQUOIAS

DISE
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Main Definitions 1

- Accuracy is defined as the closeness between a given element and the data value which is considered to be the correct representation of the entity to which x refers.
- The syntactic accuracy is defined to be its closeness to the elements belonging to a given domain lookup table Δ which is assumed to be error-free.

$$m_{acc}(X) = \frac{\sum_{i=1}^{N} acc(x_i, \Delta)}{N - N^*}$$

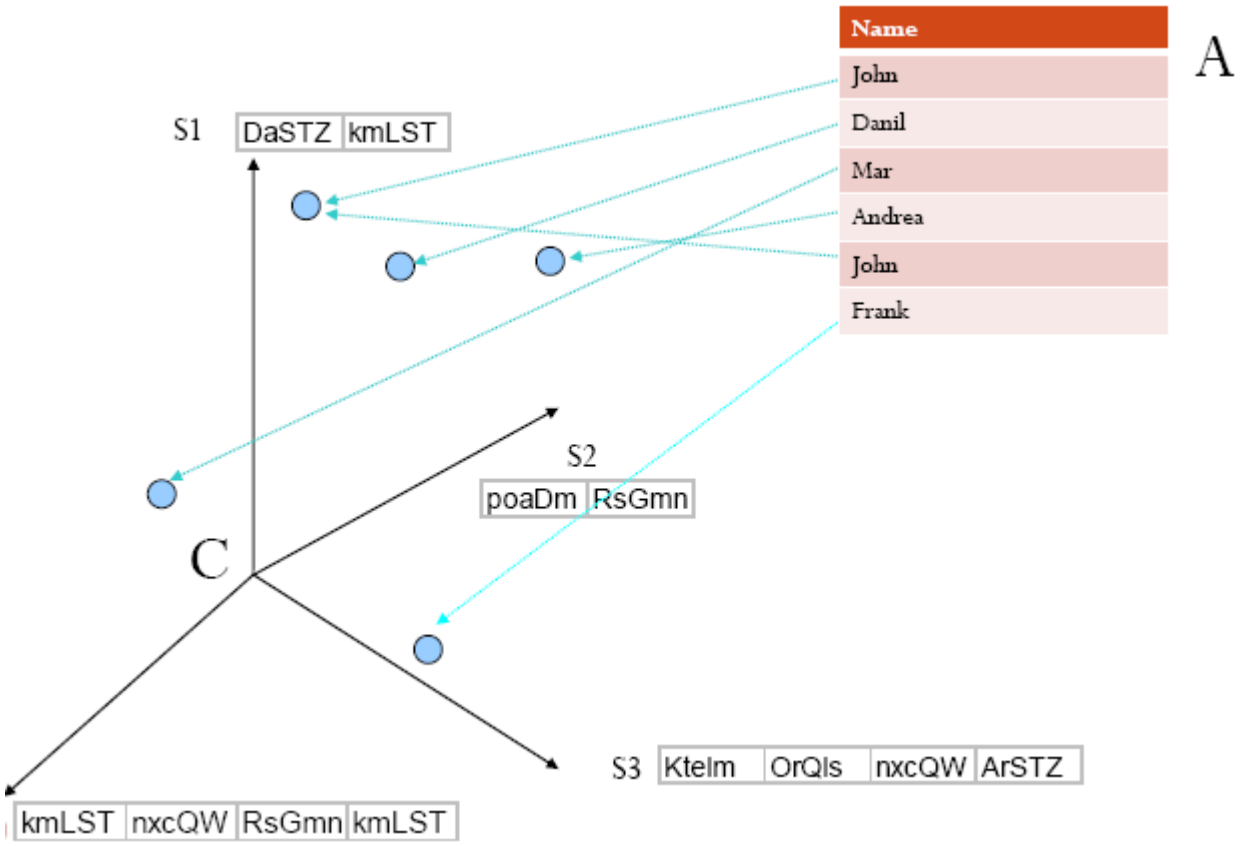
Main Definition 2

- Borugain Embedding is a mapping of a finite metric space into R^k
- Let Y be the native space consisting of n strings to be embedded, d be the edit distance
- Let $d'(y, Y') = \min_{z \text{ in } Y} \{d(y, z)\}$ be the edit distance from y to its nearest neighbor belonging to the subset Y'
- $t(y) = [d'(y, Y_1), \ldots, d'(y, Y_k)]$

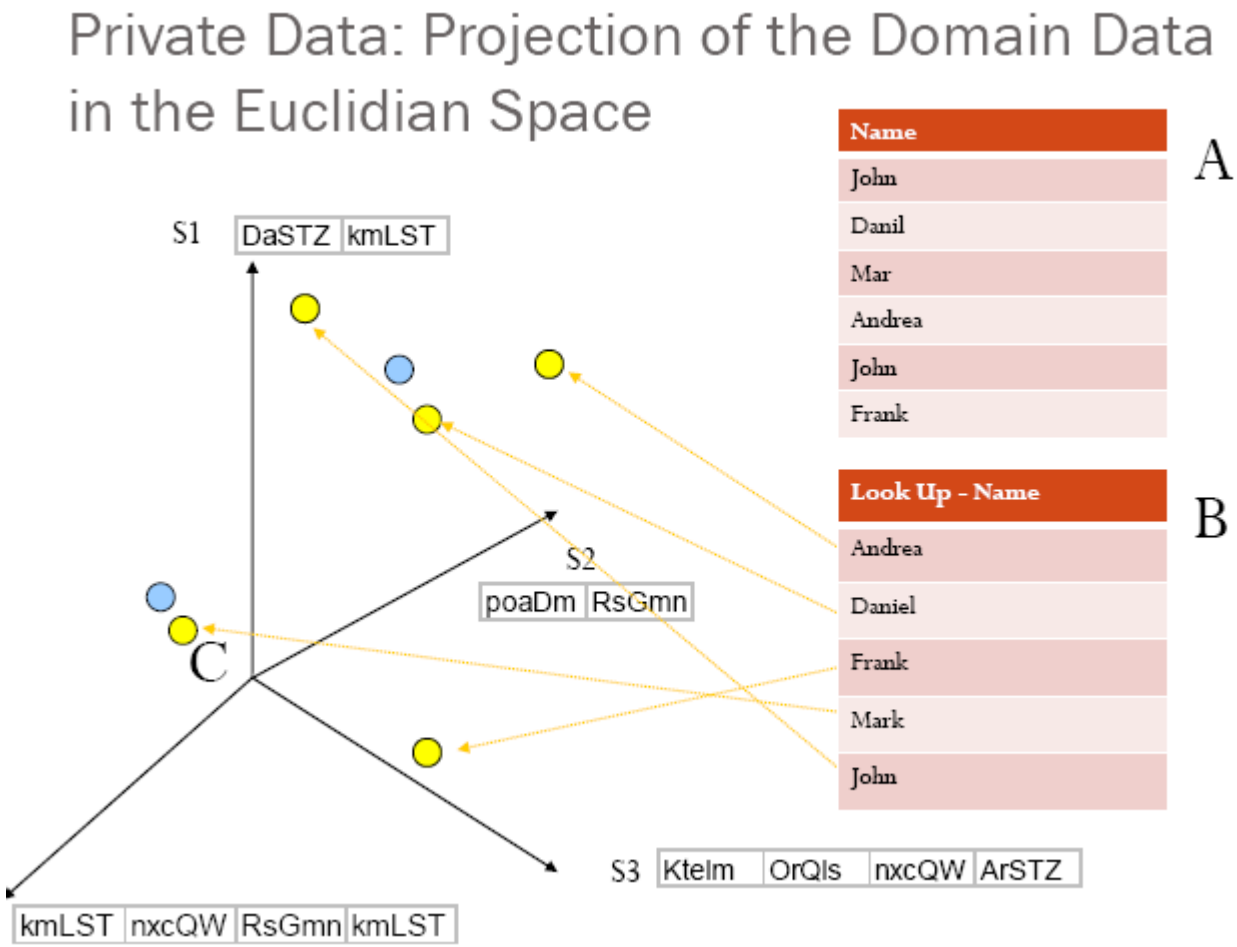
SEQUOIAS

DISCO
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Main Definition 3



Main Defintion 4



Main definition 4

Private Data: The Problem of Distortion

- A **distortion** [15] exists between distance in the Euclidian space with respect to distance in the original space:

$$\text{EuclidianDistance}(o_1, o_2) \leq \text{EditDistance}(o_1, o_2)$$

- a **distortion** exists in the privacy assessment metric:

$$\text{PrivacyAccuracy}^{\text{App}}(X^E, D^E) \leq \text{Accuracy}^{\text{App}}(X, D)$$

SEQUOIAS

DISE
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Problem Statement

- let **DP** be the Data Provider, **DLTP** be the Domain Look-up Table Provider and **C** be the Certifier.
- DP owns the set of data X to be certified
- C offers a data quality certification service while accessing the domain look-up tables owned by DLTP.
- The privacy-preservation data quality assessment problem is the problem of assessing the quality of X without requiring any data disclosure to DP and to DLTP.

SEQUOIAS

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

The P²QA protocol

- The protocol is composed by three phases
- Set-up
- Request
- Usage

SEQUOIAS

DISE
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Phase 1

- Given the domain look-up table Δ, DLTP acts as follows:
- generates the set R of reference sets,
- generates the learning sets that will be used to train the predictive model
- computes, for each element of the learning data set the edit distance from x to its nearest element belonging to Δ
- uses R to embed Δ and the learning set
- develop the predictive model O

SEQUOIAS

DISE
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Phase 2

- Given a DP that wants to certify X (e.g. a set of English family names), it:
- (a) looks for the DLTP storing the domain look-up table,
- (b) receives from DLTP the reference set R and preprocessing tasks to prepare DP data according to DLTP requirements,
- (c) calculates the lengths for the elements belonging to X in the array L(X),
- (d) uses R to embed its data X producing t(X)
- (e) sends the set and the identification of the DLTP used for validation to C

SEQUOIAS

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Phase 3

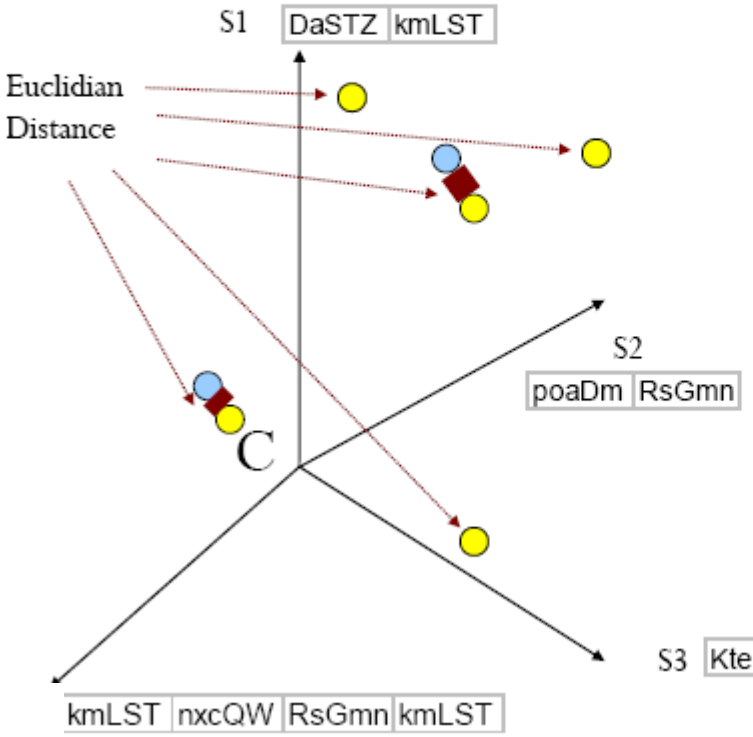
- C receives t(X), L(X) from DP and t(Δ) together with O from the selected DLTP and:
- (a) uses the privacy preserving quality assessment framework for assessing syntactic accuracy of t(X) against t(Δ)
- (b) forecasts, for inaccurate data belonging to t(X) the error entity by using the predictive model O,
- (c) sends to DP the certification of the data quality level together with a report which includes the forecast of the corresponding error for inaccurate data.

SEQUOIAS

DISCO
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Phase 3

Private Data: A Certifier Calculates Euclidian Distance



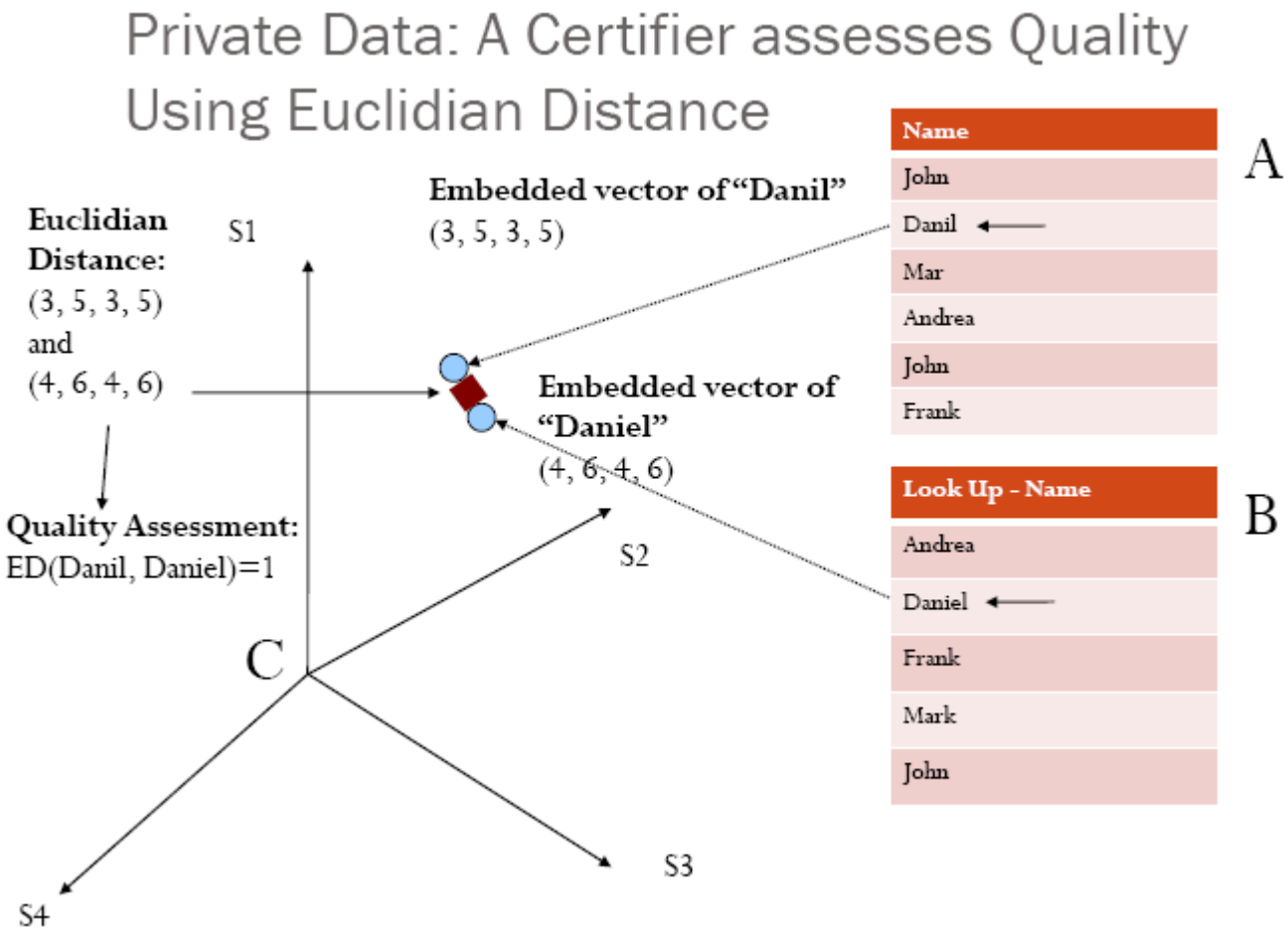
Name
John
Danil
Mar
Andrea
John
Frank

A

Look Up - Name
Andrea
Daniel
Frank
Mark
John

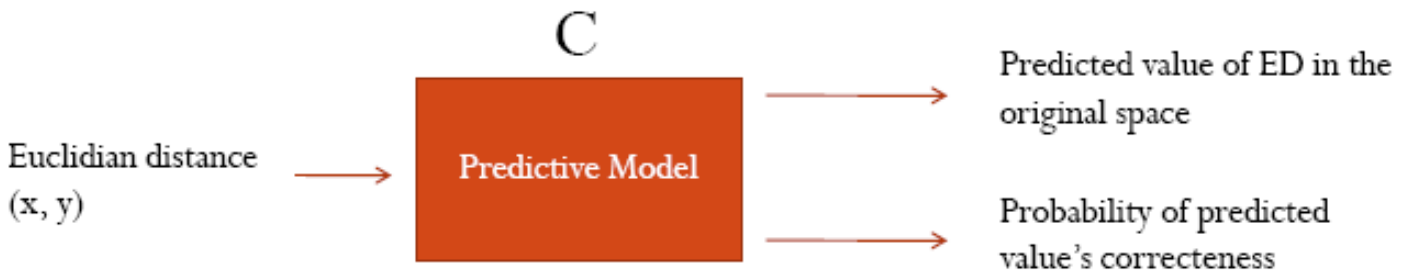
B

Phase 3


Private Data: A Certifier assesses Quality Using Euclidian Distance
Euclidian Distance:
(3, 5, 3, 5)
and
(4, 6, 4, 6)
S1
Embedded vector of "Danil"
(3, 5, 3, 5)
Embedded vector of "Daniel"
(4, 6, 4, 6)
Quality Assessment:
ED(Danil, Daniel)=1
C
S2
S3
S4
Name
John
Danil
Mar
Andrea
John
Frank
A
Look Up - Name
Andrea
Daniel
Frank
Mark
John
B

Experimental Evaluation

Private Data: The Solution is a (Tree Augmented Naive Bayes) Data Mining Predictive Model



Testing Dataset	ED	Precision	Recall	Accuracy
English First Names	0	100	100	100
	1	80.12	95.14	
	2	64.58	58.81	75.34
	3	66.15	14.40	
Italian First Names	0	100	100	100
	1	79.97	94.82	
	2	64.58	61.21	74.83
	3	66.22	15.23	

ACCURACY 84%(*)

Canada Citizens DB :
[15,792 records]:
- First name,
- Surname,
- City of birth.

SEQUOIAS

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Call for Paper: Special issue on Data Quality in the Internet Era on Internet Computing

- This special issue of IC seeks original research contributions, industrial experience, and case studies relating to all aspects of information quality in the Internet era.
- Appropriate topics include
  - information integration and fusion;
  - information quality in multimedia, semistructured, and unstructured data;
  - information quality in Web applications (including Web 2.0 tools);
  - information quality in sensor networks;
  - information quality assessment; and
  - information quality methodologies

DEADLINE 15 NOVEMBER 2009

www.computer.org/internet/cfp.htm

SEQUOIAS

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Conclusion

- Introduce the problem of assess data in a privacy preserving way
- Define and implement a technique to estimate the distorsion
- For the future...
  - ✓ Improve the performance
  - ✓ Consider other data quality dimension
  - ✓ Analyze other approach e.g. Cryptographic hashing

SEQUOIAS

DISCO
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Thank you! Questions?

(For the CFP send an email to maurino@disco.unimib.it)

SEQUOIAS

DISCO
DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE