kb-Anonymity

Data Protection & Privacy a.a. 2019/2020

Authors:

- Alessio Ravera S 4193948
- Andrea Canepa S 4185248

OVERVIEW

Why kb-Anonymity ?

kb-Anonymity VS k-Anonymity

Privacy Preservation Requirements

Behaviour Preservation Requirements

P-F & P-T

Dataset: INAIL Open Data

Performances of our solution

Proposal for Future Works

Why ***kb-Anonymity*** ?

01 Importance of ***test-cases*** in software development

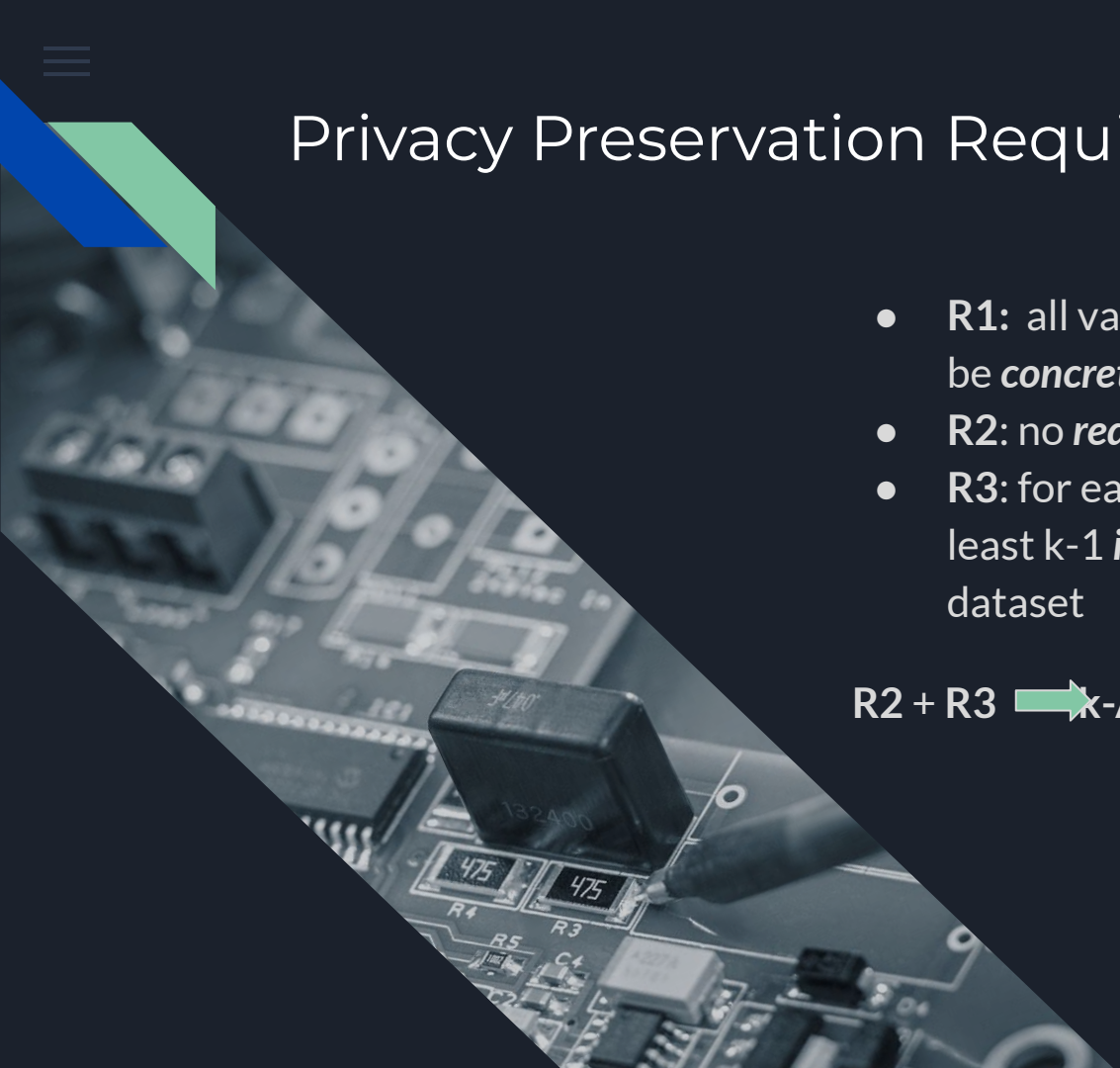
02 Useful test-cases are ***proprietary data*** dependent

03 Inability to release ***sensitive*** data to developers

kb-Anonymity VS ***k-Anonymity***

- Both replace data to ensure ***privacy***
- k-Anonymity have to preserve ***statistics*** of the original data, while kb-Anonymity may introduce ***fake values***
- Replaced data in the kb-Anonymity model must preserve the ***program behaviour*** of the original data

Privacy Preservation Requirements



- **R1:** all values in the released dataset must be ***concrete***
- **R2**: no ***redundant*** test-cases
- **R3**: for each released tuple there must be at least k-1 ***indistinguishable*** tuples in the raw dataset

R2 + R3 ⟹ k-Anonymity Modulo Uniqueness

Behaviour Preservation Requirements

Same Path

For each released tuple *b* and each raw tuple *t* mapped to *b*, *b* and *t* must exhibit the same behaviour in the SUT

Path preserving == Same execution path

Combined Privacy and Behaviour Preservation

P-F

Same Path, no Field Repeat

Given a raw dataset R, a released version X of R:

- X satisfies R1-R3
- $\forall i \in \{1,\ldots,n\}.\ \forall t \in R.\ \forall t' \in X.\ t[i] \neq t'[i]$

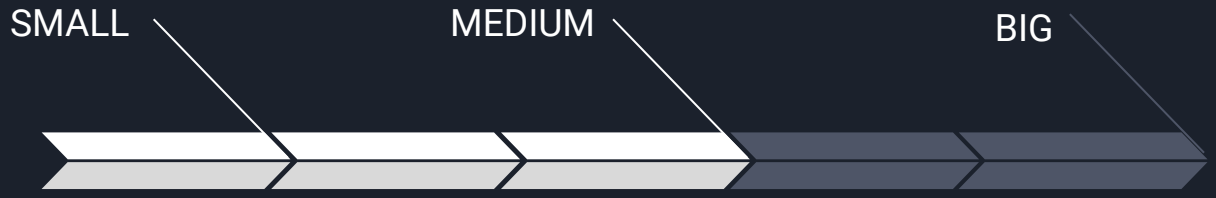
P-T

Same Path, no Tuple repeat

Given a raw dataset R, a released version X of R:

- X satisfies R1-R3
- $\forall t \in R.\ \forall t' \in X.\ t[i] \neq t'[i]$

Dataset: INAIL Open Data



1861 Tuples

Only accidents happened during November 2019

5229 Tuples

Accidents of last trimester of 2019

10458 Tuples

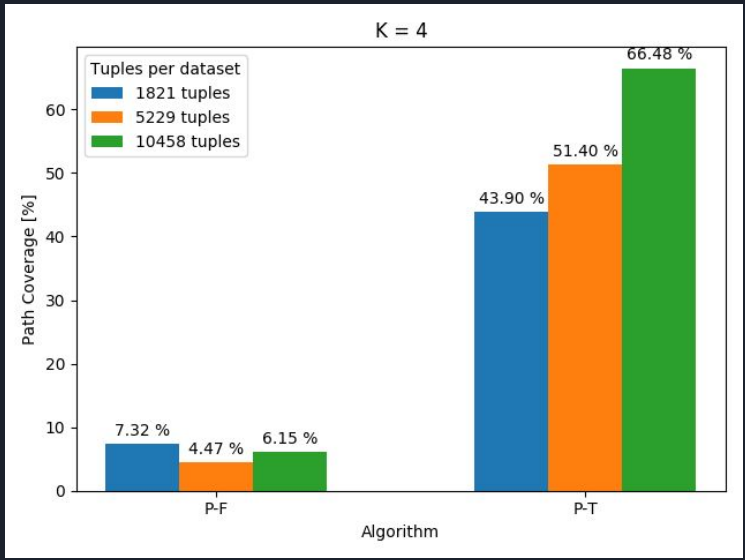
2x Medium dataset

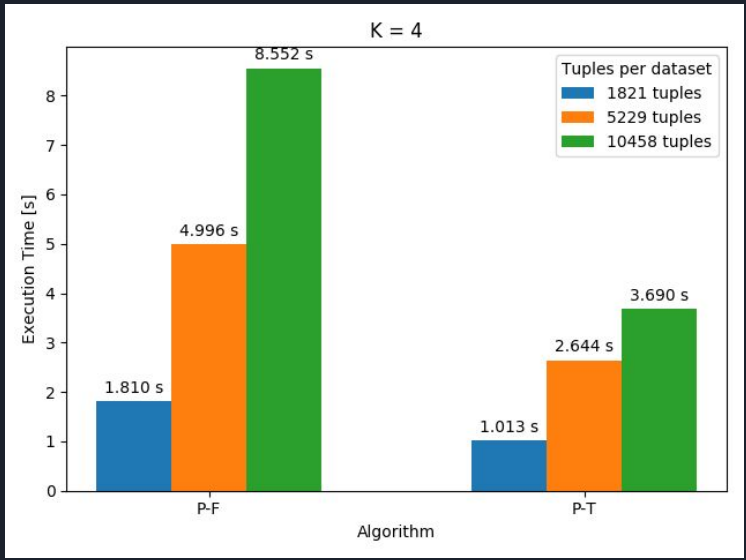
Accidents on workplace during the last quarter of 2019 in the Ligurian region

Performance of our solution

Test #1

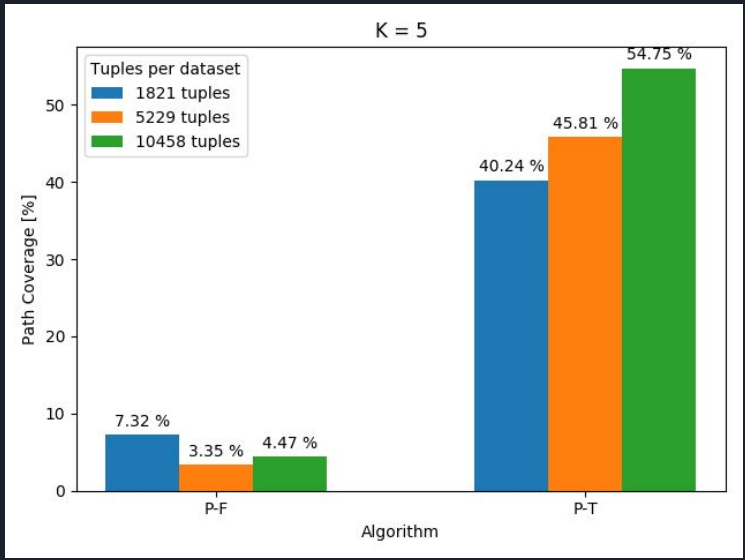
K = 4

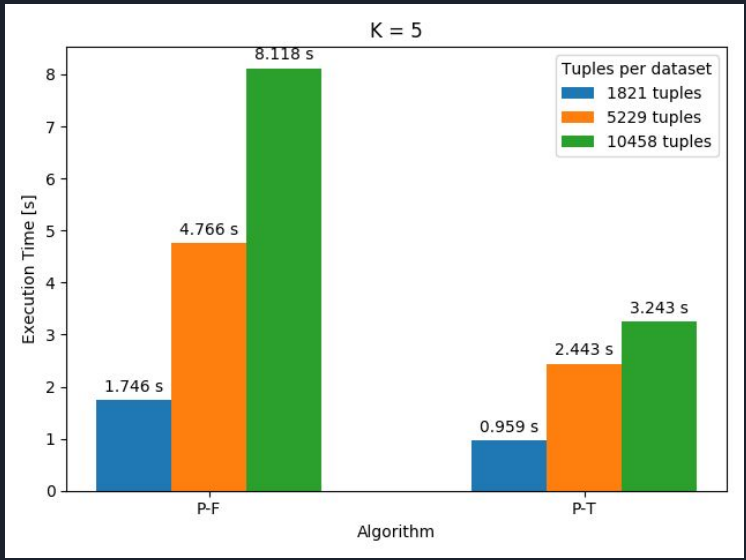

K = 4
Tuples per dataset
1821 tuples
5229 tuples
10458 tuples
Path Coverage [%]
0
10
20
30
40
50
60
7.32 %
4.47 %
6.15 %
43.90 %
51.40 %
66.48 %
P-F
P-T
Algorithm


K = 4
Tuples per dataset
1821 tuples
5229 tuples
10458 tuples
Execution Time [s]
0
1
2
3
4
5
6
7
8
1.810 s
4.996 s
8.552 s
1.013 s
2.644 s
3.690 s
P-F
P-T
Algorithm

Test #2

K = 5

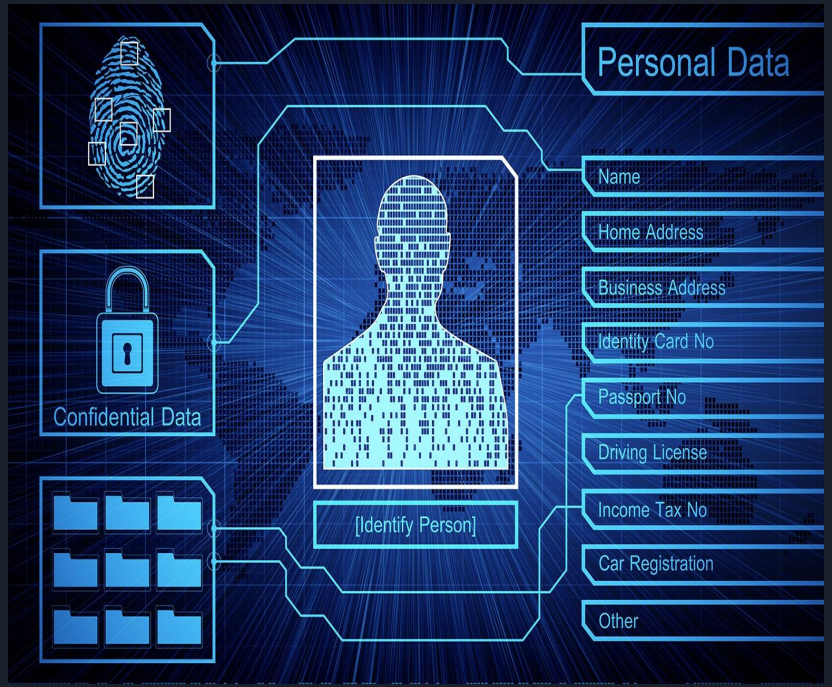

K = 5
Tuples per dataset
1821 tuples
5229 tuples
10458 tuples
Path Coverage [%]
0
10
20
30
40
50
7.32 %
3.35 %
4.47 %
40.24 %
45.81 %
54.75 %
P-F
P-T
Algorithm


K = 5
Tuples per dataset
1821 tuples
5229 tuples
10458 tuples
Execution Time [s]
0
1
2
3
4
5
6
7
8
1.746 s
4.766 s
8.118 s
0.959 s
2.443 s
3.243 s
P-F
P-T
Algorithm

Proposal for Future Works

An alternative to P-F and P-T techniques, kb(R)-Anonymity:

- Pick a ***random*** value for each field in a tuple that satisfy the *path conditions* and *range constraints*
- At least equal ***path coverage***, but it could improve
- ***Behaviour*** is preserved
- What about ***privacy*** ?



Thank you for your attention!

References:

- Budi, Aditya & Lo, David & Jiang, Lingxiao & Lucia, Lucia. (2011). kb-Anonymity: A Model for Anonymized Behavior-Preserving Test and Debugging Data. ACM SIGPLAN Notices. 46. 10.1145/1993316.1993551.