Privacy preserving similarity detection for data analysis

CSAR 2013

Iraklis Leontiadis[1]
Melek Önen[1]
Refik Molva[1]
M.J. Chorley[2]
G.B. Colombo[2]

[1]Eurecom - France
[2]Cardiff - UK

Privacy vs Utility

Personality test

Clustering

Similarity

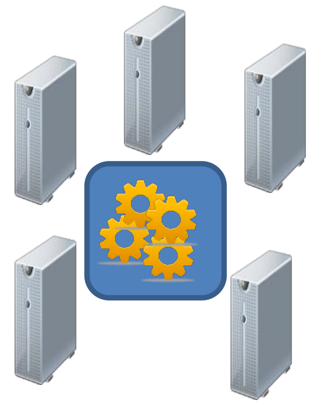
Data A_1,A_2,A_3,…A_n

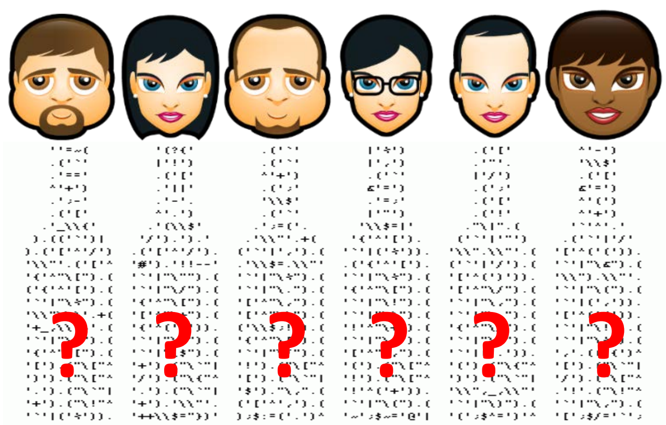
.

.

.

Data B_1,B_2,B_3,… B_n





Naïve solutions

- Encrypt data with standard crypto
  - Renders operations infeasible.
- Data separation
  - Vertical separation is not always applicable.
- Anonymizing techniques
  - Don’t protect individuals data.

Our Approach

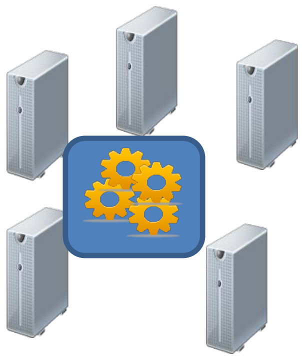
- Combine crypto with data processing

Data $A'_1, A'_2, A'_3, \ldots A'_n$

.
.
.

$$F(A_1, \ldots An) = F(A_1', \ldots An')$$

Data $B'_1, B'_2, B'_3, \ldots B'_n$

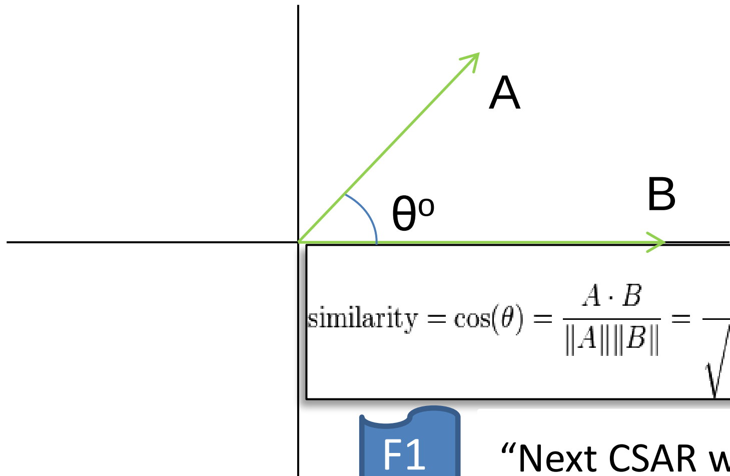


User	Data	Data analysis
Alice	$A_1', \ldots An'$	$F(A_1', \ldots An')$
Bob	$B_1', \ldots Bn'$	$F(B_1', \ldots Bn')$

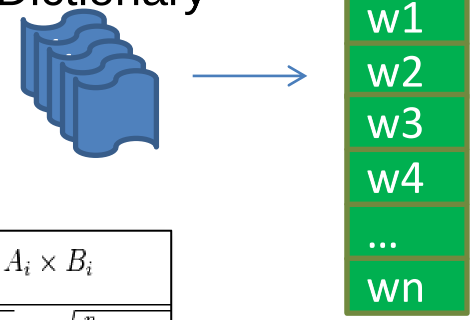
Outline

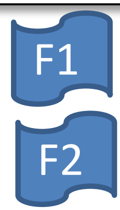
- Our solution
  - Cosine similarity
  - Privacy with Geometrical Transformations
- Security Analysis
- Performance Evaluation
  - Hierarchical clustering
  - Results
- Looking Ahead

Cosine similarity



$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$





"Next CSAR workshop will be held in Karlsruhe"

"Next CSAR workshop will be held in London"

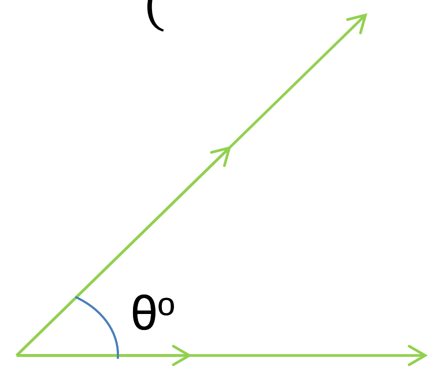
		Karlsruhe	be	London	held	in	workshop	Next	CSAR	will	
A=	…	1	1	0	1	1	1	1	1	1	…
B=	…	0	1	1	1	1	1	1	1	1	…

$$\sigma = \frac{\sum \mathrm{Ai} * \mathrm{Bi}}{\sqrt{\sum Ai^2} * \sqrt{\sum Bi^2}} = 0.875$$

Random Scaling

- Data encoded as unique vectors in $\mathbb{R}^n$
- φ_{r}: $\mathbb{R}^n \rightarrow \mathbb{R}^n$ s.t:

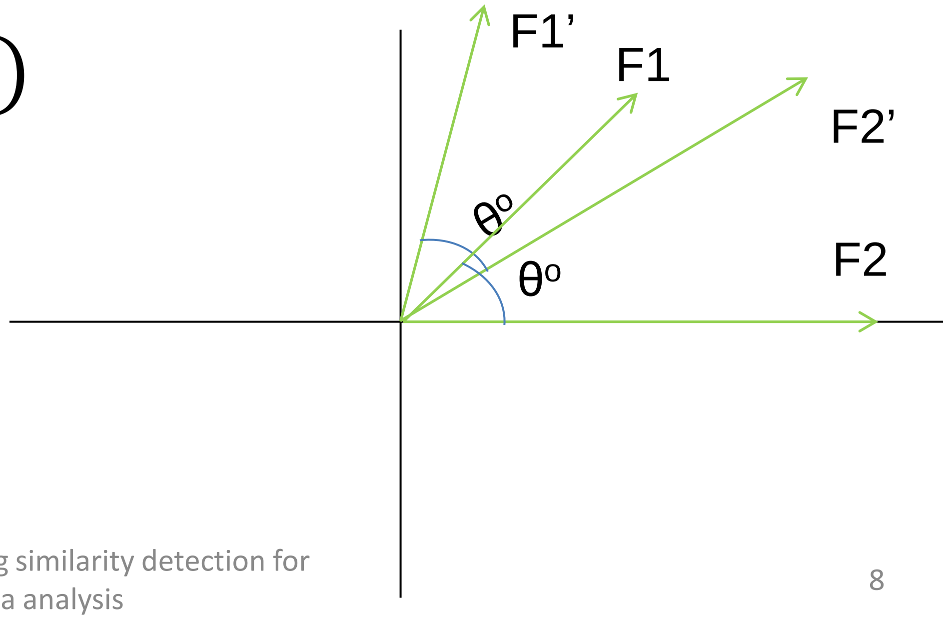
$$\cos(\mathrm{a}, \mathrm{b}) = \cos(\varphi_{\mathrm{r1}}(\mathrm{a}), \varphi_{\mathrm{r2}_{(}}\mathrm{b}))$$



- Random scaling
  - $\mathrm{r} \longleftarrow \mathbb{R}^n$
  - $\mathrm{S}(\mathrm{r}, \mathrm{A}) = \mathrm{r} \cdot \mathrm{A} = \begin{bmatrix} \mathrm{r} & \cdots & \\ \vdots & r & \vdots \\ & \cdots & r \end{bmatrix} \cdot \mathrm{A}$

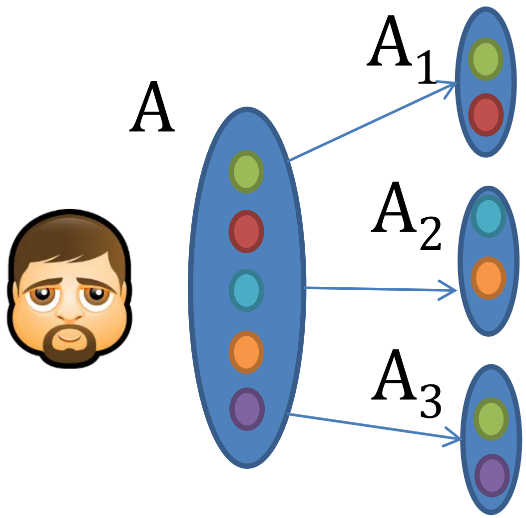
Vector Rotation

- Rotation by a common angle $\lambda°$
  - $\mathrm{R}_{\lambda°}(\mathrm{a}) = \mathrm{a} \cdot \begin{bmatrix} \cos(\lambda°) & \cdots & \sin(\lambda°) \\ \vdots & & \vdots \\ -\sin(\lambda°) & \cdots & \cos(\lambda°) \end{bmatrix}$
- $\varphi_{\mathrm{r}} = \mathrm{a} \cdot \mathrm{R}_{\lambda°}(\mathrm{a}) \cdot Sr(\mathrm{a})$



Our solution

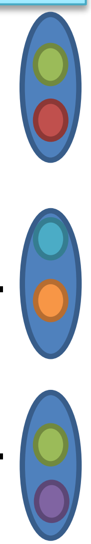
Dimension reduction	Random Scaling	Rotation



$S(r_1, A_1) = r_1 \cdot$

$S(r_2, A_2) = r_2 \cdot$

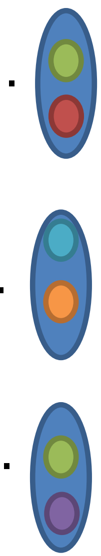
$S(r_3, A_3) = r_3 \cdot$



$R_{\lambda^\circ}(r_1 \cdot A_1) = R_{\lambda^\circ} \cdot r_1 \cdot$

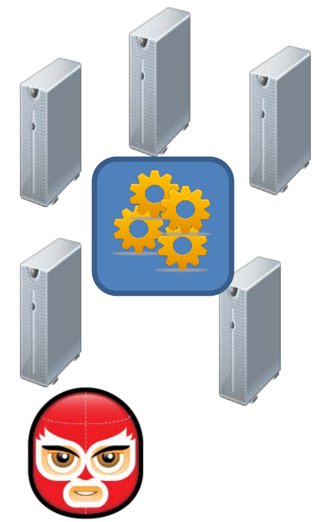
$R_{\lambda^\circ}(r_2 \cdot A_2) = R_{\lambda^\circ} \cdot r_2 \cdot$

$R_{\lambda^\circ}(r_3 \cdot A_3) = R_{\lambda^\circ} \cdot r_3 \cdot$



Security analysis

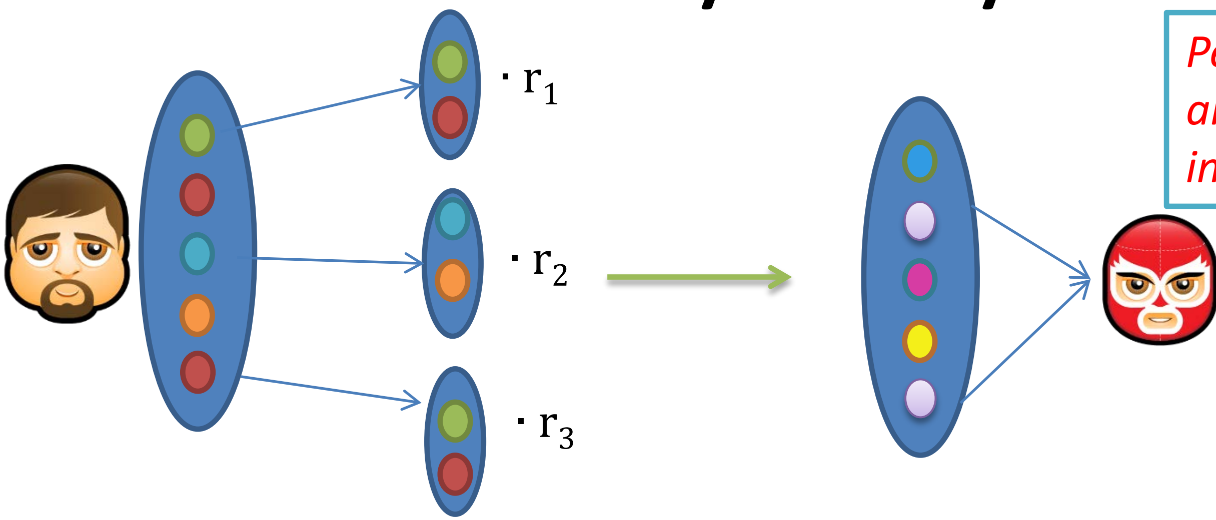
$V'_1 = \mathrm{R}_{\lambda^\circ}(\mathrm{S}(\mathrm{r}_1, d_1, d_2), \mathrm{S}(\mathrm{r}_2, d_3, d_4), \mathrm{S}(\mathrm{r}_3, d_1 d_5))$



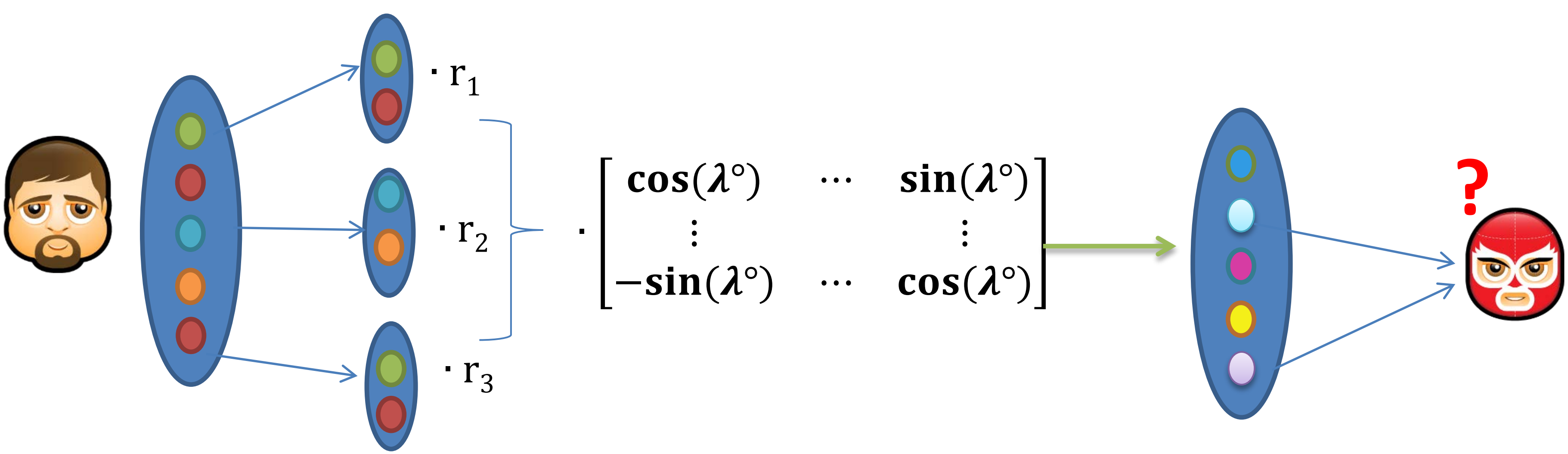
- External:
  - Rotation angle remains unknown.

- Internal:
  - Rotation angle is known.

Security analysis cont'd



Per user equivalent coefficient are exposured as auxiliary information



Evaluation

- 173 users willing to run 4sqPersonality test
- 5 factor personality test
  - Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism.

Clustering approach

- Hierarchical Agglomerative clustering (HAC)
  - Input: n points and N*N similarity matrix
  - Output: Single cluster containing all n points

```
C=MakeSingletonClusters();
for i=0 to i=n:
   Find “closest” clusters c1,c2;
      Merge(c1,c2);
      RecomputeDistances(C);
      if #C=1 exit();
```

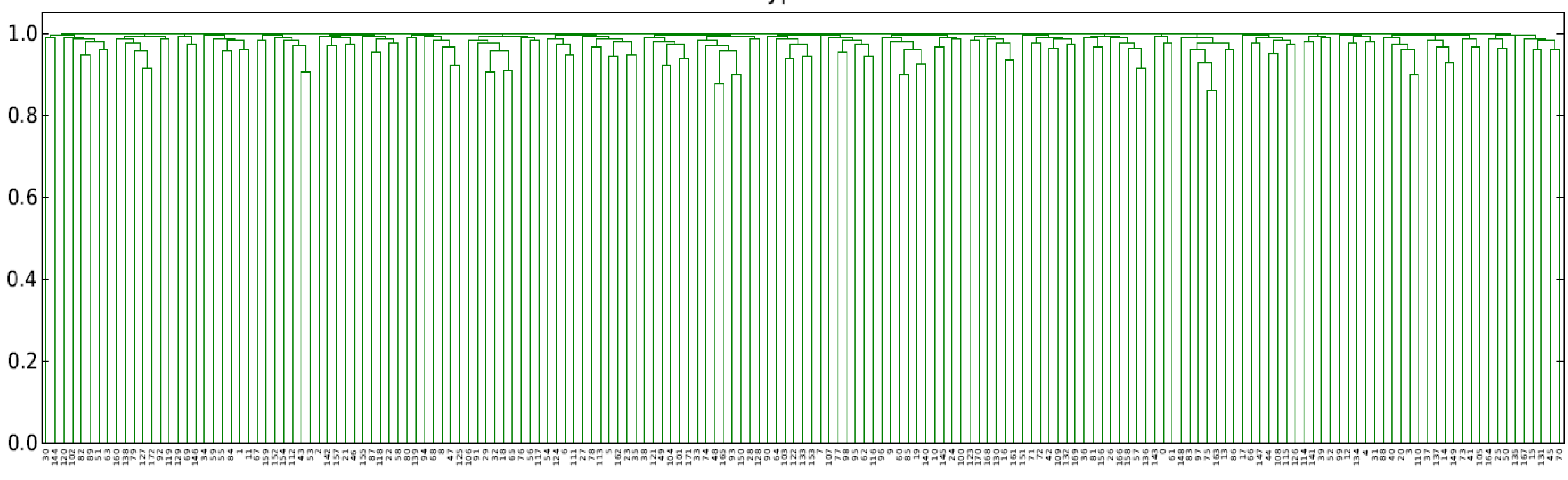
Agglomerative: $O(n^3)$
Divisible: $O(2^n)$

Cosine Similarity

Results

Plaintext data

Encrypted data



Recap

1. Pairwise cosine similarity for multidimensional vectors.
2. Geometrical transformations compatible with cosine similarity.

Looking Ahead

- Other privacy preserving similarity detection algorithms.
- Privacy preserving data analysis algorithms:
  - MAX,MIN

Thank you!

Iraklis Leontiadis
leontiad@eurecom.fr