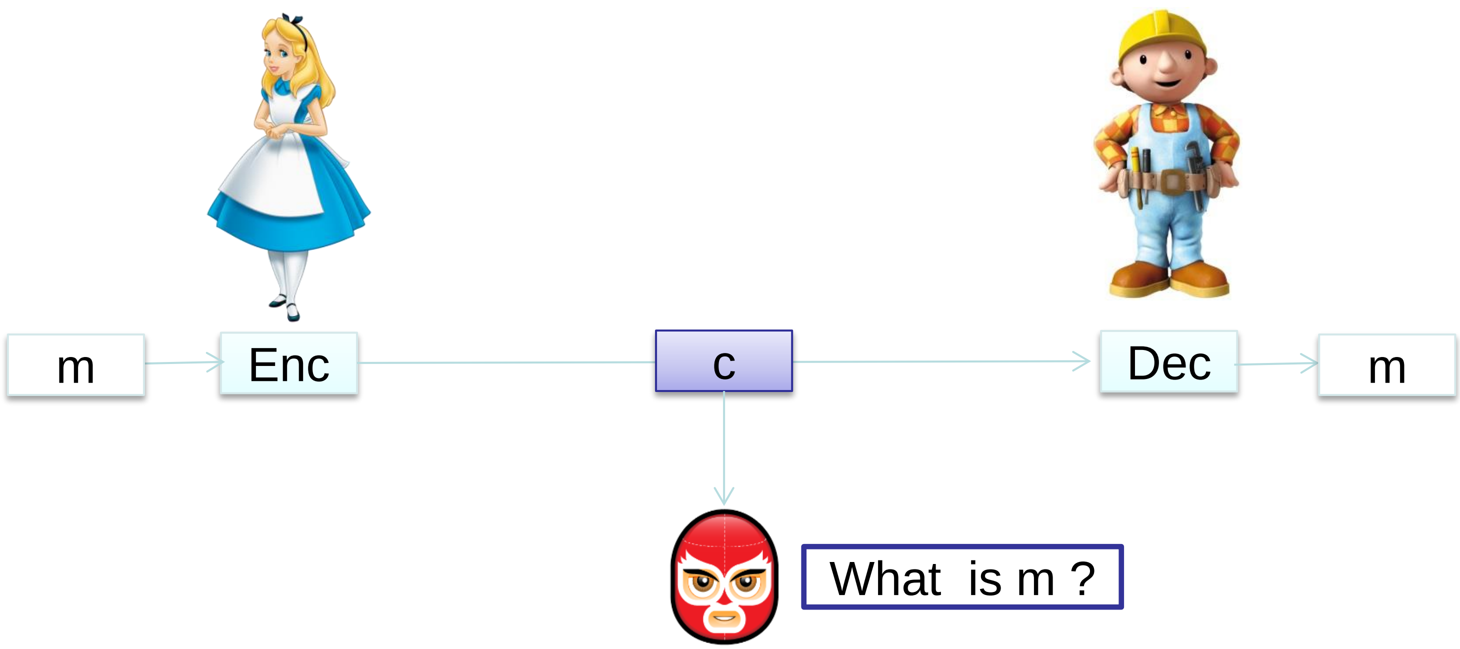
Privacy Preserving Data Collection and Analysis

Iraklis Leontiadis

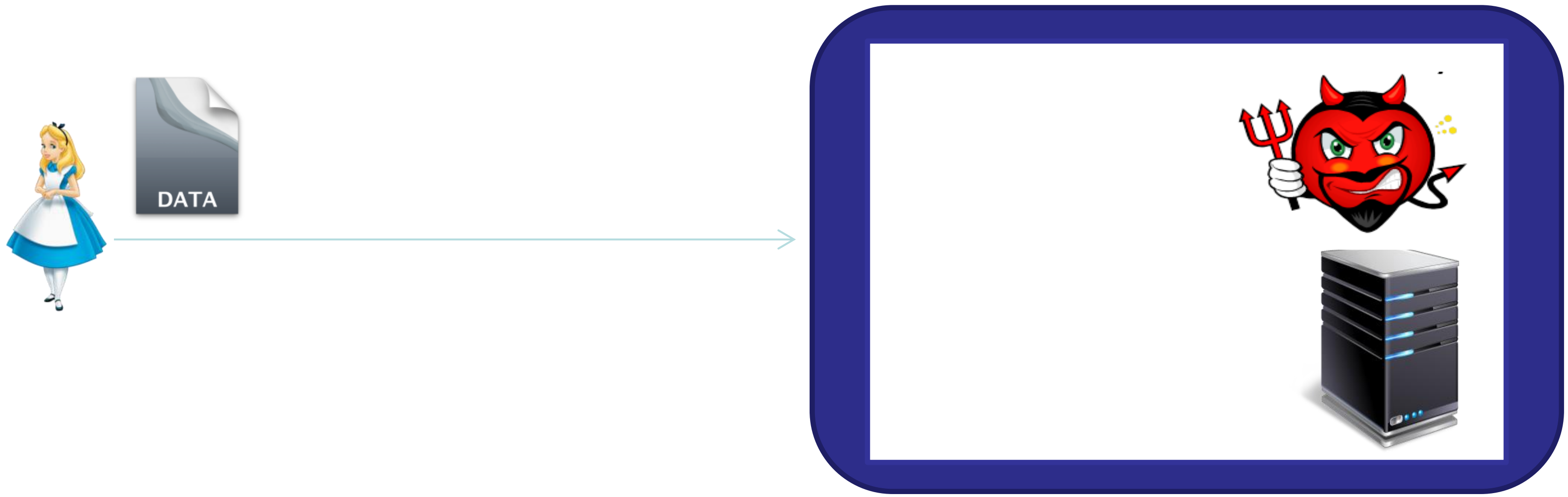
leontiad@eurecom.fr

The start



Nowadays

Server



Conventional Encryption for Data Storage

Homomorphic Encryption for Computations

Outline

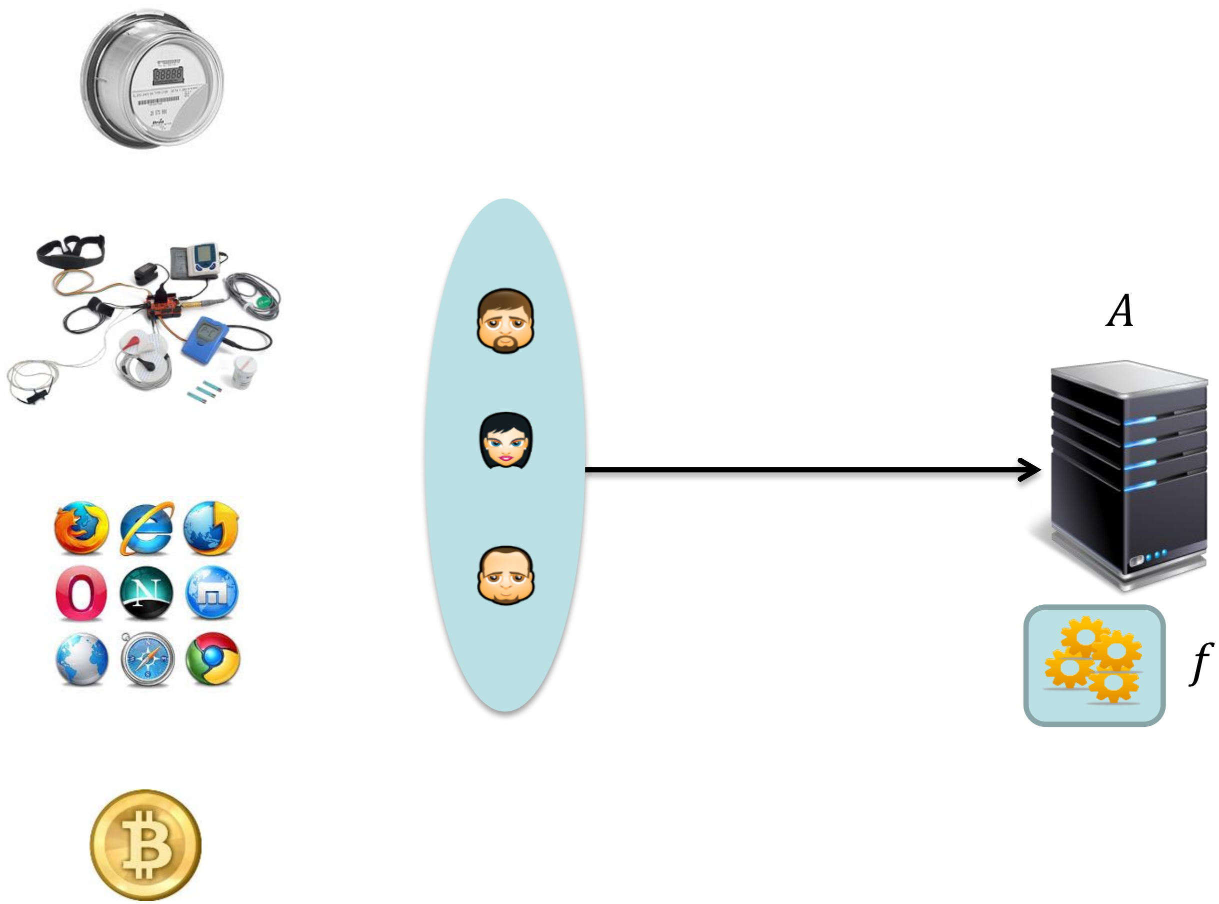
Generic problem
Related work
Shortcomings

PP clustering
PP ordering

Multi-User time-series data

Conclusion

Problem



Energy forecasting

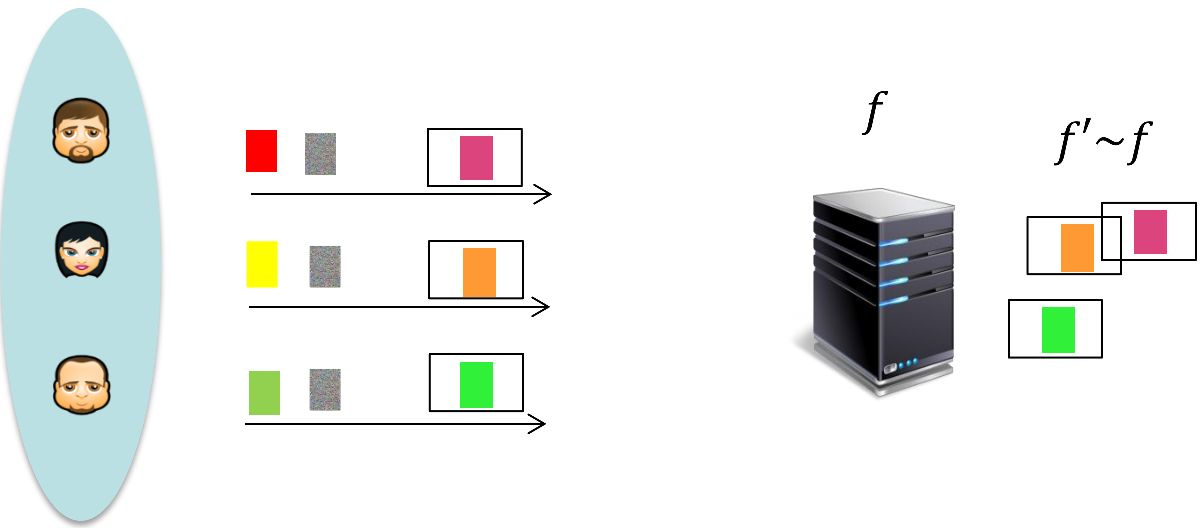
Medical data analysis

User Profiling

Overall Transactions

EURECOM
Sophia Antipolis

Ad-Hoc techniques



AA01 BFP05 AS00 KLR06 OZ10 BLJ08 KDWS03

Different Privacy definitions

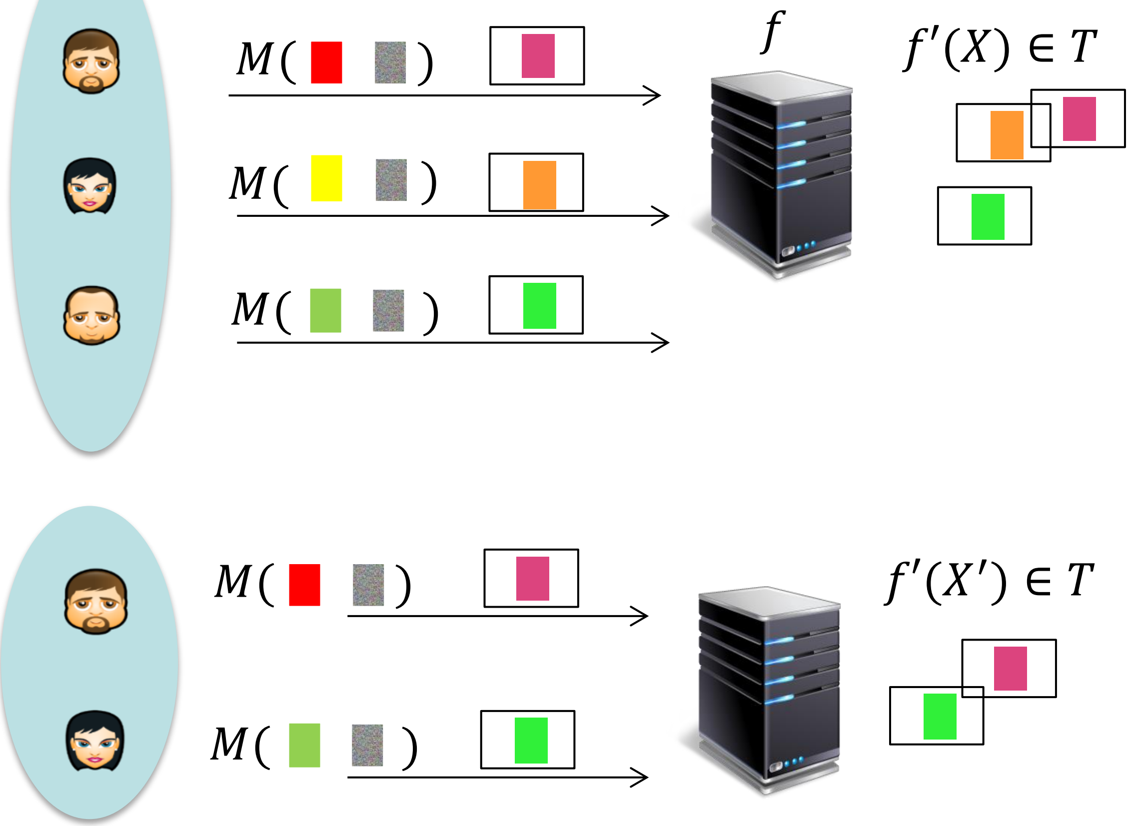
Entropy

Distributions' distance

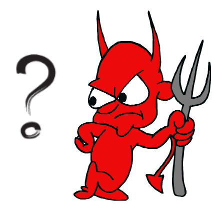
Confidence interval

Differential privacy

$$Pr[M(X) \in T] \leq e^{\varepsilon}\Pr[M(X') \in T]$$



DM06 HRK10 CSS11 BDGKB13 AC11 XWG11 GRC09



Noisy f

Cryptographic solutions

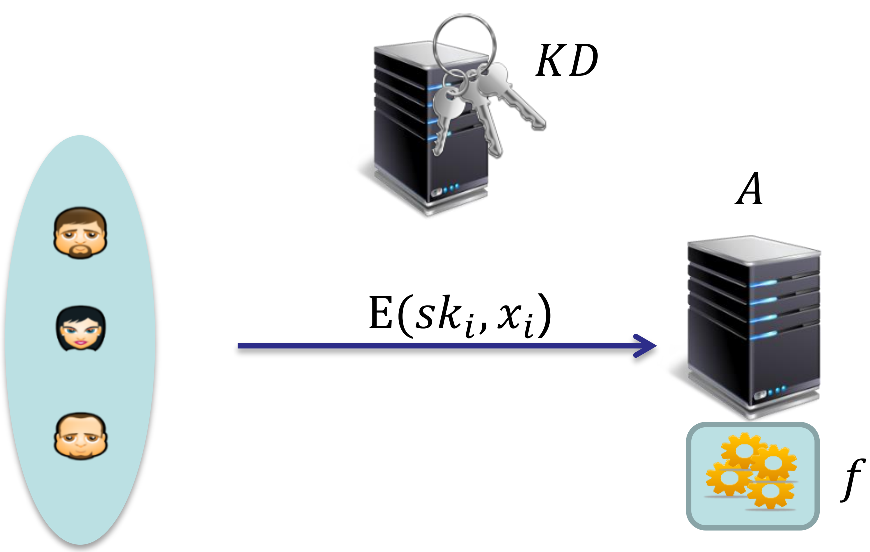
- **Trusted Key Dealer**
  - No support for dynamic population
  - Intolerant to failures

JL2013

SCRCS2011 GMP2014

- **No Key Dealer**
  - Communication cost

ET2012



Honest but Curious

EURECOM

Shortcomings with existing solutions

- **Noise-based**
  - No accuracy

PPC

PPSGS

- **Encryption-based**
  - Trusted key dealer
  - Honest but curious Aggregator

PDTDA

PUDA

Outline

Generic problem
Related work
Shortcomings

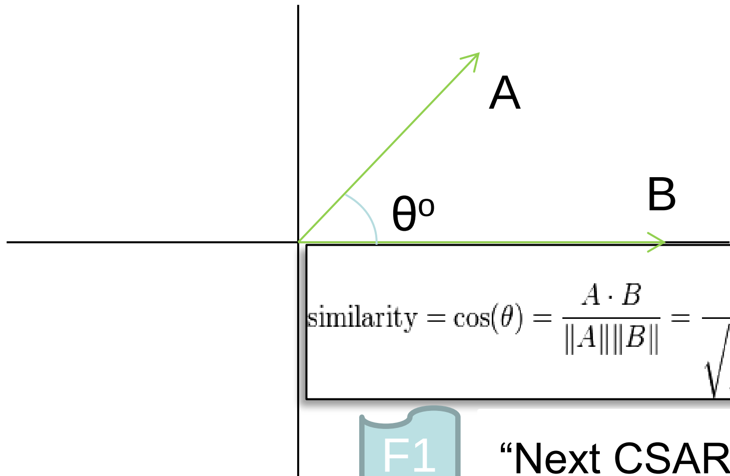
PP clustering
PP ordering

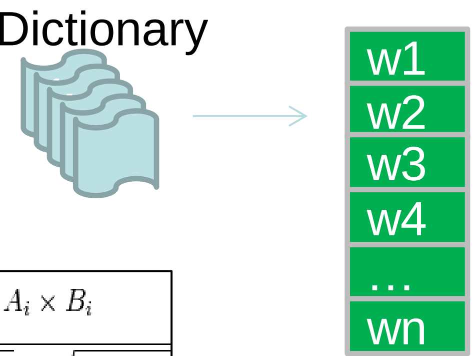
Multi-user time-series data

Conclusion

Privacy Preserving Clustering (PPC)

CSAR2013





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

F1 “Next CSAR workshop will be held in Karlsruhe”

F2 “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	…

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

EURECOM Sophia Antipolis

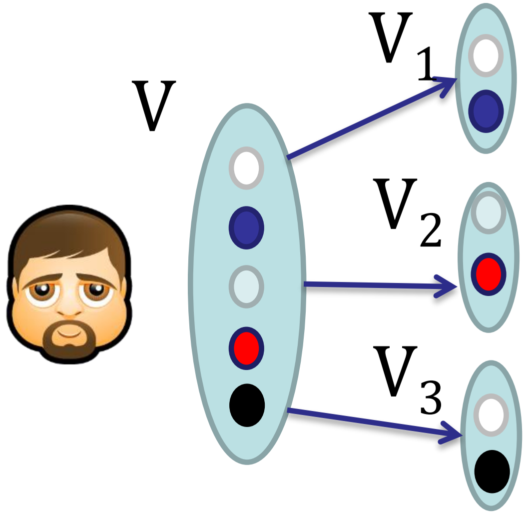
Privacy Preserving Clustering (PPC)

- **Model:**
  - Trustworthy users
  - HbC Aggregator

- **Technique:**
  - Cosine preserving transformations

Our solution

Dimension reduction	Random Scaling	Rotation



$S(r_1, V_1) = r_1 \cdot$

$S(r_2, V_2) = r_2 \cdot$

$S(r_3, V_3) = r_3 \cdot$

$R_{\lambda^\circ}(r_1 \cdot V_1) = R_{\lambda^\circ} \cdot r_1 \cdot$

$R_{\lambda^\circ}(r_2 \cdot V_2) = R_{\lambda^\circ} \cdot r_2 \cdot$

$R_{\lambda^\circ}(r_3 \cdot V_3) = R_{\lambda^\circ} \cdot r_3 \cdot$

Clustering approach

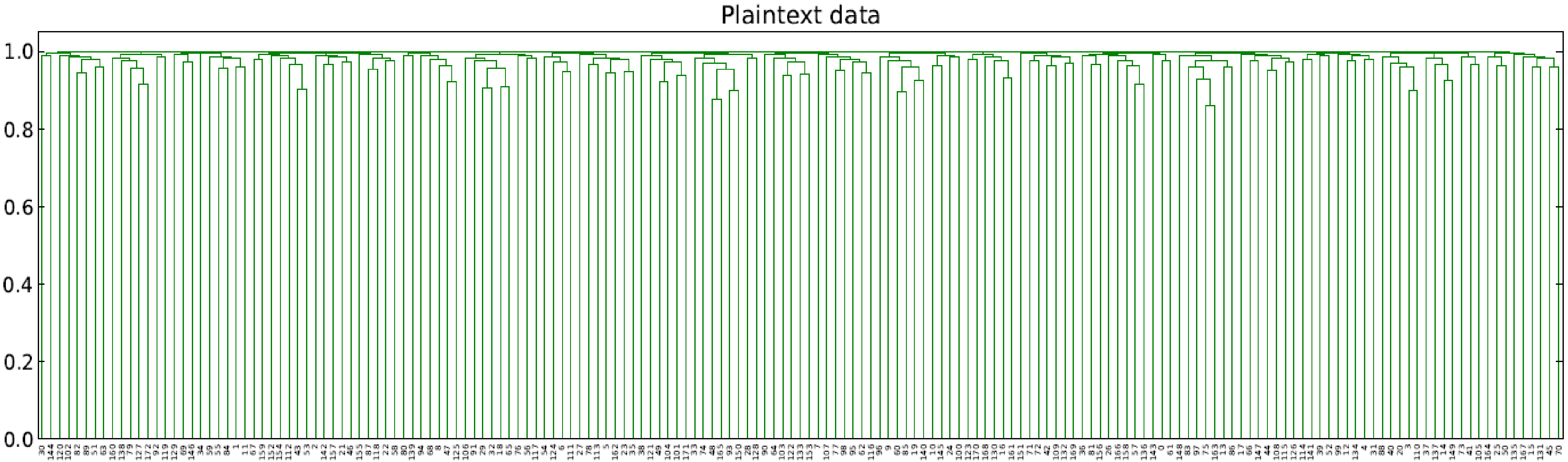
- **Hierarchical Agglomerative clustering (HAC)**
  - Input: n points and N*N similarity matrix
  - Output: Single cluster containing all n points

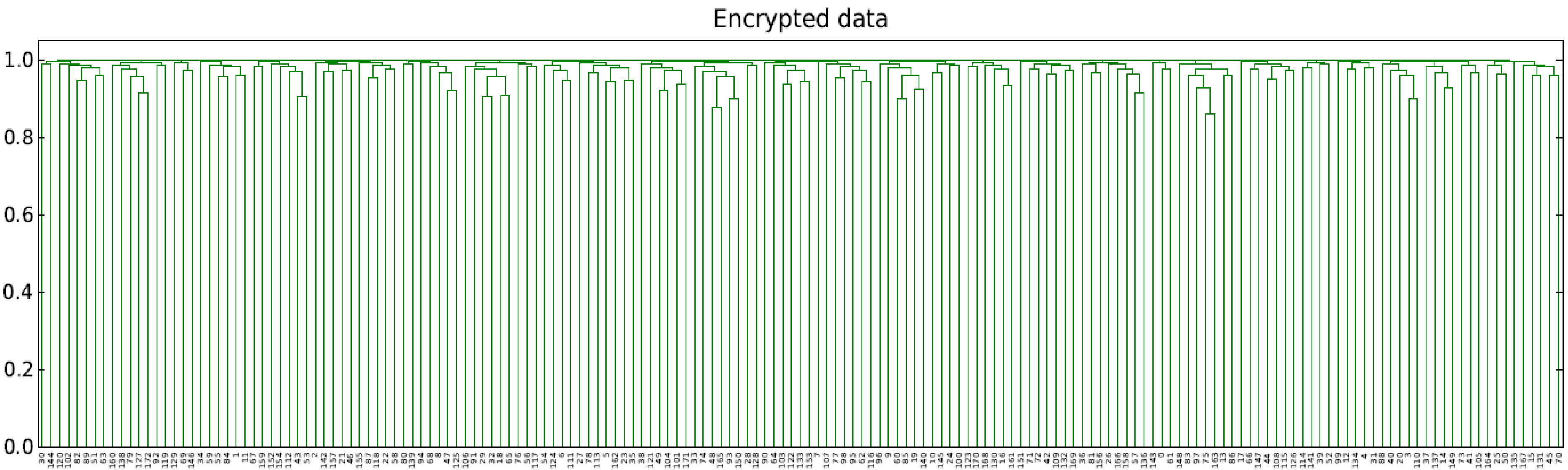
```
C=MakeSingletonClusters();
for i=0 to i=n:
   Find “closest” clusters c1,c2;
      Merge(c1,c2);
      RecomputeDistances(C);
      if #C=1 exit();
```

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

Cosine Similarity

Analysis


Plaintext data
1.0
0.8
0.6
0.4
0.2
0.0

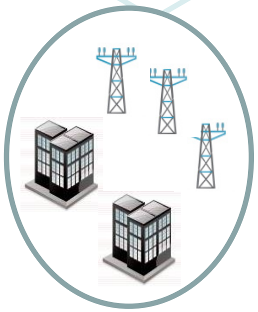

Encrypted data
1.0
0.8
0.6
0.4
0.2
0.0

Privacy Preserving Smart Grid Statistics (PPSGS)

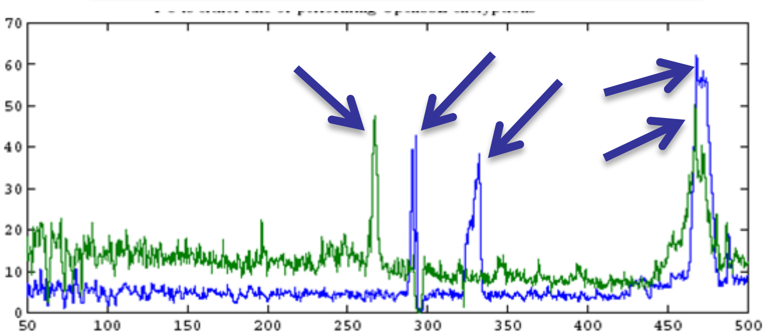
DASEC2014

- **Provide accurate individual statistics**

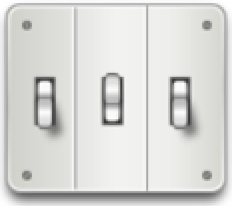
When did a home consume the maximum energy?



Obfuscate real values but reveal the order



Augment functionality by filtering spurious spikes



Promote awareness

EURECOM Sophia Antipolis

PPSGS

$\mathrm{E_{OPE}}(\mathrm{sk_i}, x_{i,t}) = c_{i,t}$

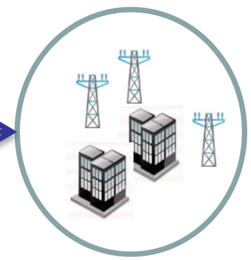
t

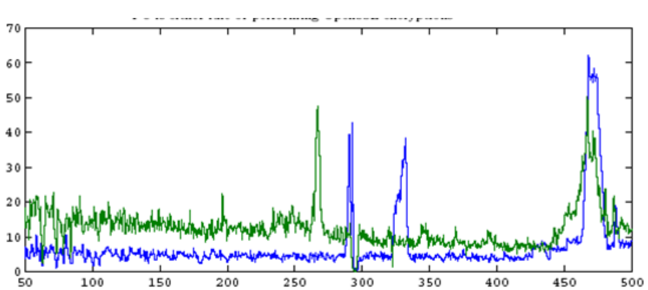
$t_0 \quad t_1 \quad t_2 \quad t_3 \quad t_4$

- $\mathrm{MAX}(\{c_{i,t}\})$ at $t_{t'}$
- min interval $= \mathrm{m}$

$c_{i,t}$

$d_{1,} d_{2,} d_3 \ldots, d_m$





$$\sum_{t=1}^{m} d_t \overset{?}{=} 0$$

EURECOM
Sophia Antipolis

Analysis

- **Feasibility (1 day)**
  - Device [Texas Instruments MSP430 Microcontrollers]
  - 16-bit RISC MSP430X MCU
  - 256KB Flash
  - 20 MHz clock rate
  - AES Accelerator

Period (seconds)	#Meterings	Flash(KB)	Time (Mcb)
1	86400	172.8	13.33
2	43200	86.4	6.32
3	28800	56.6	4.08
4	21600	43.2	2.99
5	17280	34.5	2.35
6	14400	28.8	1.93
7	12343	24.6	1.63
8	10800	21.6	1.41
9	9600	19.3	1.24
10	8640	17.2	1.10

- **Security**
  - Reductionist proof from POPF-CPA

Outline

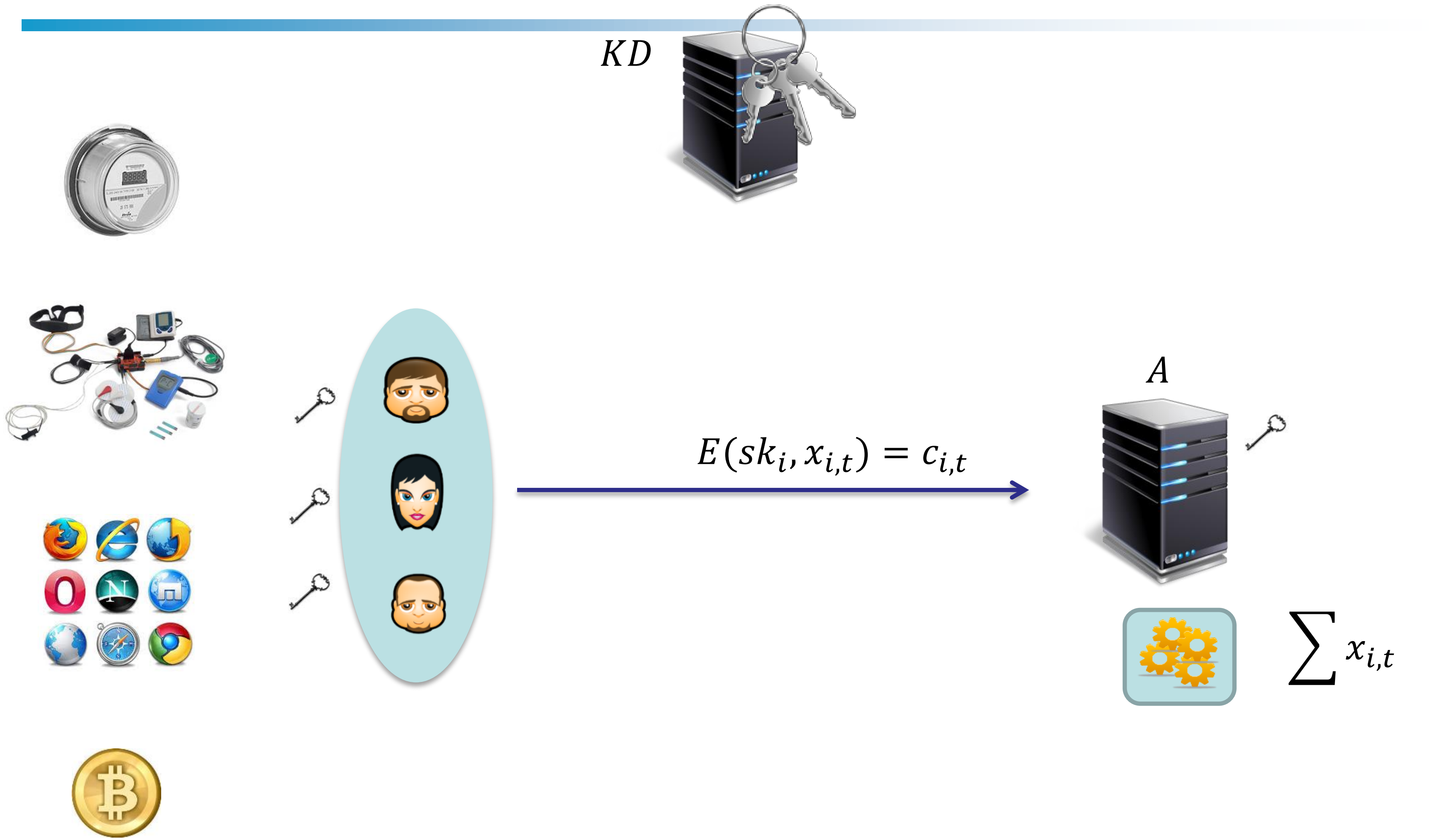
Generic problem
Related work
Shortcomings

PP clustering
PP ordering

Multi-user time-series data

Conclusion

Multi-user time series aggregation



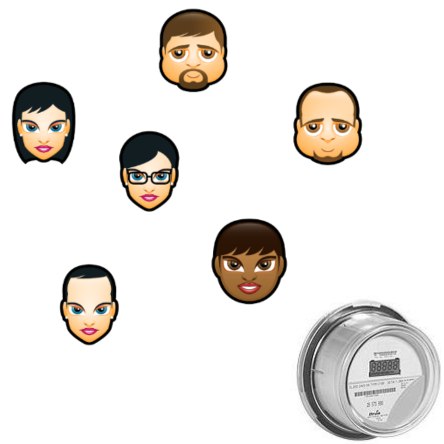
Shortcomings

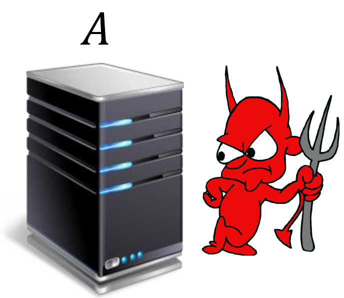
- **Fully trusted key dealer**
- **No support for dynamic population**
- **Intolerant to failures**

PDTDA

- **Lack of a stronger security model**

PUDA





Private and Dynamic Time-Series Data Aggregation

CANS2014

- **Goals**
  - No trusted key dealer
  - Dynamicity
  - Resiliency to failures
- **Ideas**
  - User-generated keys
  - Responsibility splitting mechanism

EURECOM
Sophia Antipolis

JL2013

- **Setup(k):**
  - $\mathbf{N} = \mathbf{pq}$ for primes $\mathbf{p}, \boldsymbol{q}$ (l the size of $\mathbf{N}$)
  - Trusted Dealer distributes:
    - secret keys $\mathrm{sk_i} \in \{0,1\}^{2l}$ to the users.
    - $sk_0 = -\sum_{i=1}^{n} sk_i$ to the Aggregator.
    - $H(\quad): \mathbb{Z}_N \to \mathbb{Z}_{N^2}^{*}$
- **Encrypt($\boldsymbol{sk_i}, \boldsymbol{x_{i,t}}$):**
  - $c_{i,t} = (1 + x_{i,t}N)H(t)^{sk_i} mod\ N^2$
- **Aggregate:**
  - $V_t = H(t)^{sk_0} \prod_{i=1}^{n} c_{i,t} = (1 + \sum_{i=1}^{n} x_{i,t}\, N) mod\ N^2$
  - $\sum_{i=1}^{n} x_{i,t} = \frac{V_t - 1}{N} \in \mathbb{Z}$

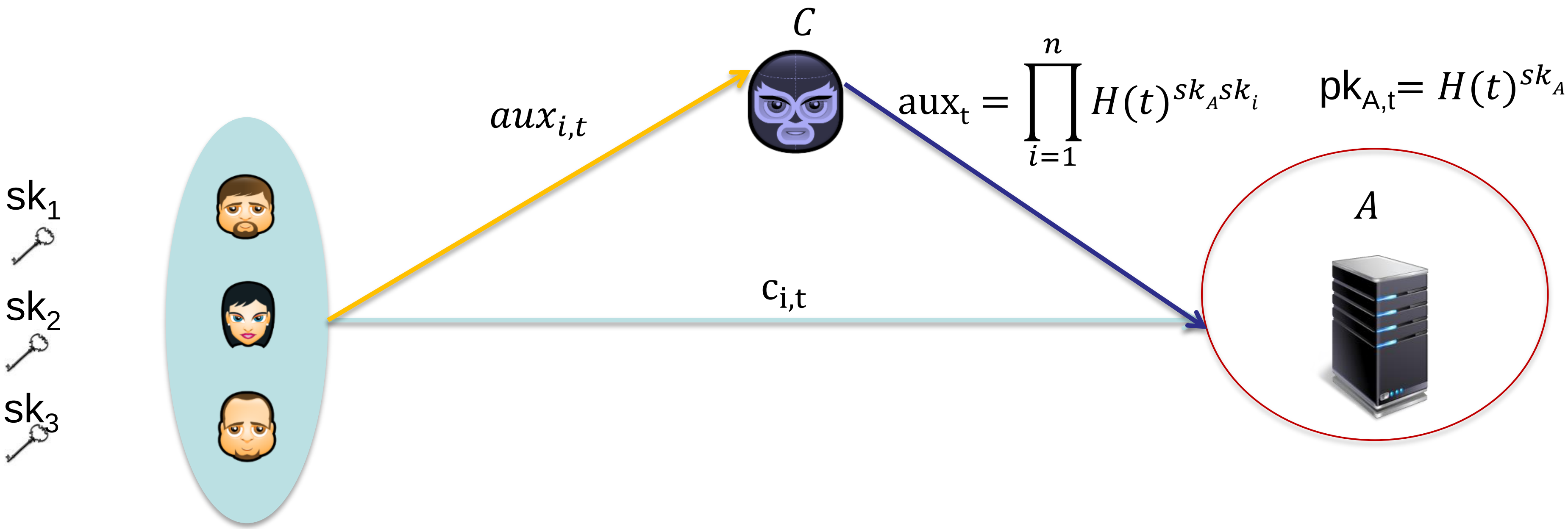
EURECOM Sophia Antipolis

PDTDA

$$aux_{i,t} = H(t)^{sk_A sk_i}$$

$$c_{i,t} = (1 + x_{i,t}N)H(t)^{sk_i} mod\ N^2$$

$sk_A \in \mathbb{Z}_N^*$



1. $P_t = \prod_{i=1}^{n}(c_{i,t})^{sk_A} = (1 + N\sum_{i=1}^{n} x_{i,t})^{sk_A} H(t)^{sk_A \sum_{i=1}^{n} x_{i,t}} \bmod N^2$
2. $I_t = \frac{\frac{P_t}{aux_t} - 1}{N}$
3. $\sum_{i=1}^{n} x_{i,t} = I_t sk_A^{-1} \bmod \mathbb{Z}_N$

EURECOM Sophia Antipolis

Privacy analysis

- **Aggregator Obliviousness based on:**
  - DCR in $\mathbb{Z}_{N^2}^*$
- **Collector Obliviousness based on:**
  - DCR in $\mathbb{Z}_{N^2}^*$
  - QR in $\mathbb{Z}_N^*$
  - DDH in the subgroup of QR in $\mathbb{Z}_N^*$

Benchmarks (sec)

- Intel Core i5 CPU M 2430 @ 2.40GHz x4, 6GB

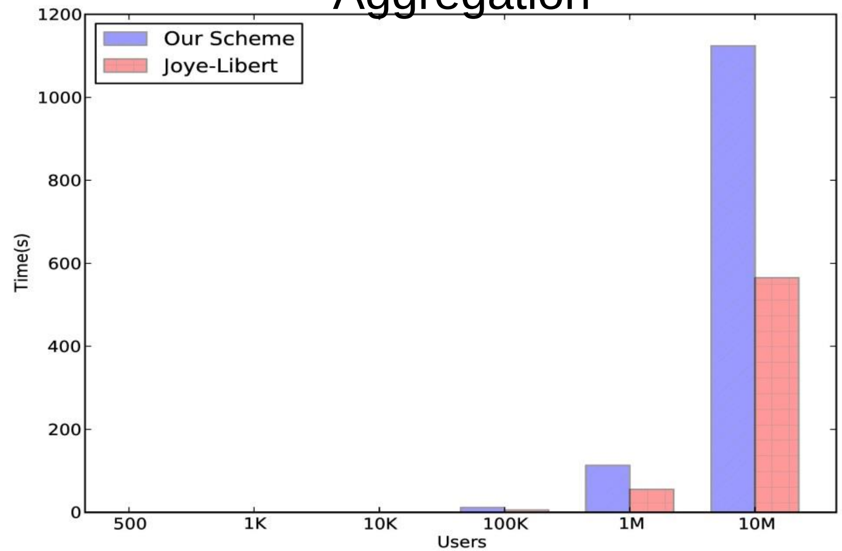
Algorithm \ |N|	2048	4096
Encrypt	0.116	0.4
Aux	0.123	0.44
Total	0.239	0.84

Entity \ #Users	500	1K	10K	100K	1M	10M
Collector	0.030	0.056	0.556	5.60	59.72	562.66
Aggregator	0.159	0.190	0.690	5.73	59.22	569.19

Encryption

|N| \ Scheme	Our scheme	Joye-Libert
2048	0.239	0.156
4096	0.84	0.4

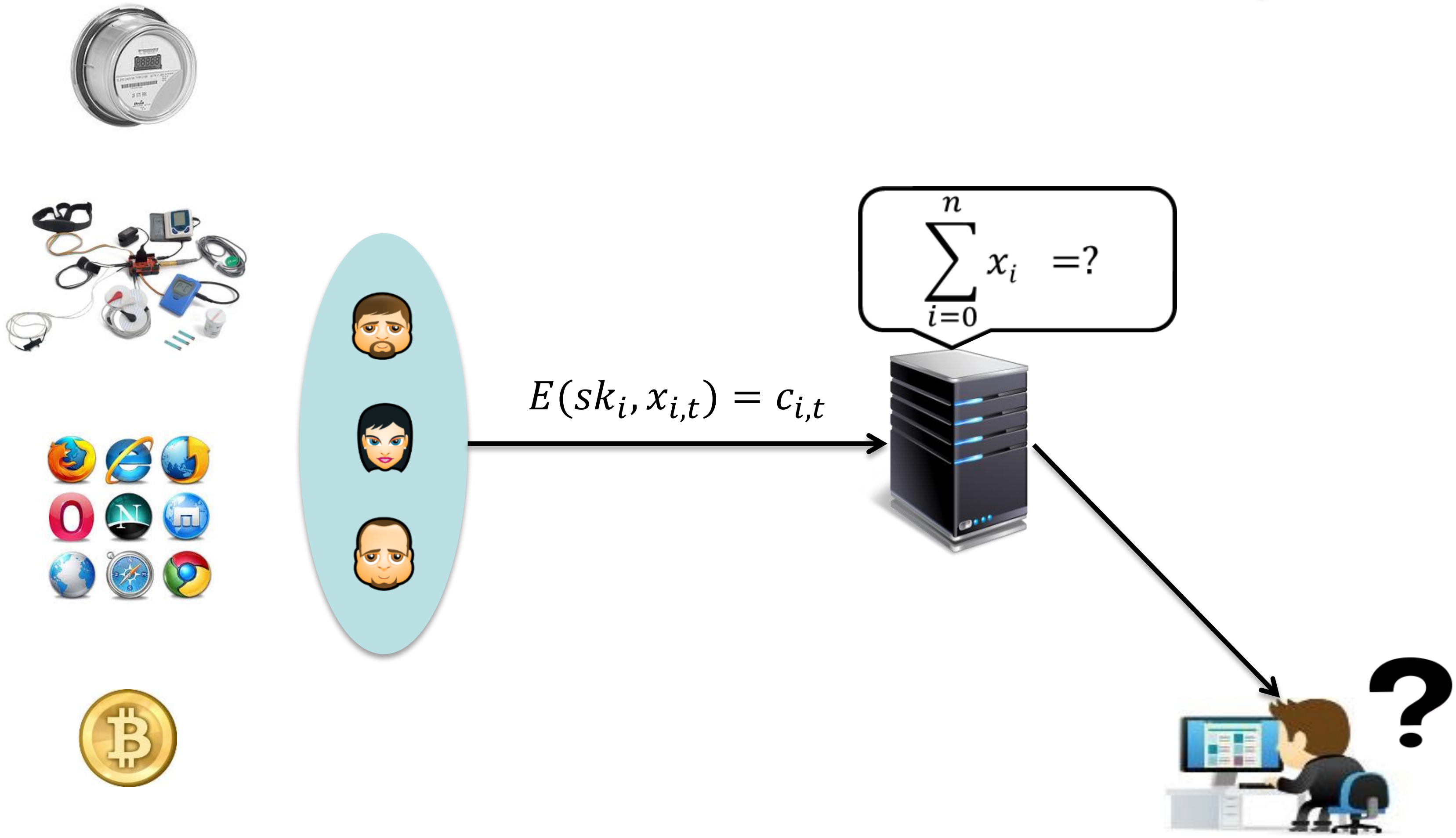
Aggregation



- Cubieboard ARM Cortex-A7 Dual-Core, 1GB

Private and Unforgeable Data Aggregation

CANS2015



Private and Unforgeable Data Aggregation

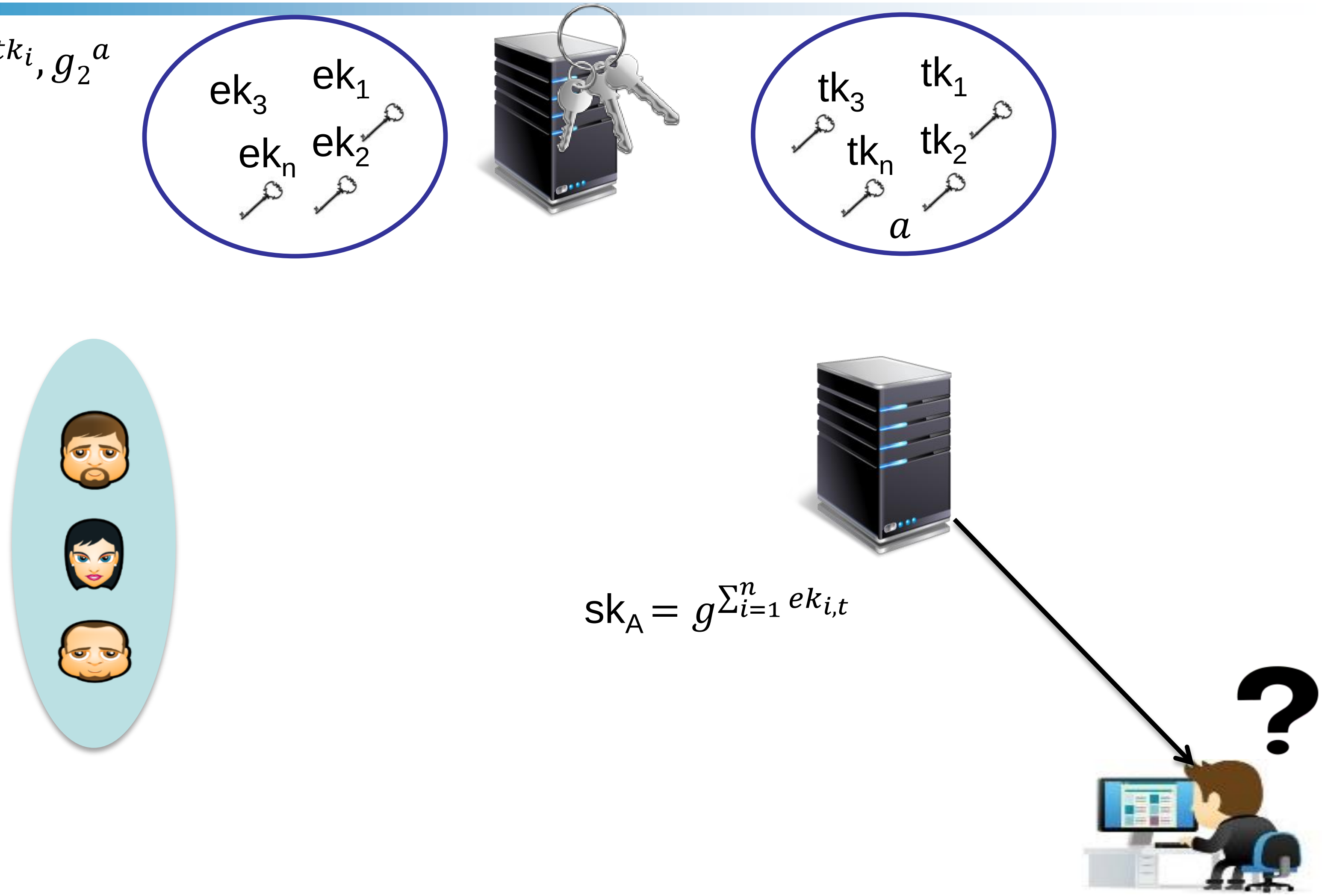
- **Goals**
  - Public Aggregate Verification
  - Obliviousness
  - Multi-user
- **Idea**
  - Homomorphic tags
  - Homomorphic encryption

SCRCS2011

- **Setup(k):**
  - $\mathbb{G}$ a cyclic group with a generator $\boldsymbol{g}$ and prime order p
  - Trusted Dealer distributes:
    - secret keys $\mathrm{sk_i} \in \mathbb{Z}_p$.
    - $sk_0 = -\sum_{i=1}^{n} sk_i$ to the Aggregator.
    - $H(\quad): \{0,1\}^* \rightarrow \mathbb{G}$
- **Encrypt($\boldsymbol{sk_i, x_{i,t}}$):**
  - $c_{i,t} = g^{x_{i,t}} H(t)^{sk_i} \in \mathbb{G}$
- **Aggregate:**
  - $V = H(t)^{sk_0} \prod_{i=1}^{n} c_{i,t} = g^{\sum_{i=1}^{n} x_{i,t}} g \in \mathbb{G}$
  - $\sum_{i=1}^{n} c_i = \log_g(V)$

Private and Unforgeable Data Aggregation

$\mathrm{vk} = g_2^{\sum_{i=1}^{n} tk_i}, g_2^{a}$



$\mathrm{sk_A} = g^{\sum_{i=1}^{n} ek_{i,t}}$

Private and Unforgeable Data Aggregation

$\text{vk} = g_2^{\sum_{i=1}^{n} tk_i}, g_2^{a}$

$$\sigma_t = \prod_{i=1}^{n} tag_{i,t} = H(t)^{\sum_{i=1}^{n} tk_i} \, g^{a sum_t}$$

$ek_1\ tk_1\ a$

$ek_2\ tk_2\ a$

$ek_3\ tk_3\ a$

$c_{i,t} = g^{x_i} H(t)^{ek_i}$

$tag_{i,t} = g^{a x_{i,t}} H(t)^{tk_i}$

$\text{sk}_A = g^{\sum_{i=1}^{n} ek_{i,t}}$

sum_t, σ_t

Constant verification time

$$e(\sigma_t, g_2) = e(H(t), g_2^{\sum_{i=1}^{n} tk_i}) e(g^{sum_t}, g_2^{a})$$

EURECOM Sophia Antipolis

Security Analysis

- **Aggregator Obliviousness based on:**
  - DDH
- **Aggregate Unforgeability based on:**
  - BCDH
  - New LEOM assumption

Secure under GGM

Outline

Generic problem
Related work
Shortcomings

PP clustering
PP ordering

Multi-user time-series data

Conclusion

Recap

- **New aggregation functions + accuracy**
- **No key dealer + dynamicity**
- **Verifiability**

PPC

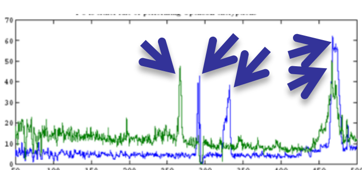
Dimension reduction

Random Scaling

Rotation

PPSGS

OPE + differences



PDTDA

No key distribution

JL

CO

PUDA

Homomorphic Tags

Shi et al.

Future work

- **Verifiability in presence of untrustworthy users**
- **Verifiability + No key dealer**
- **Standard Model**

Questions?

Thank you!!!