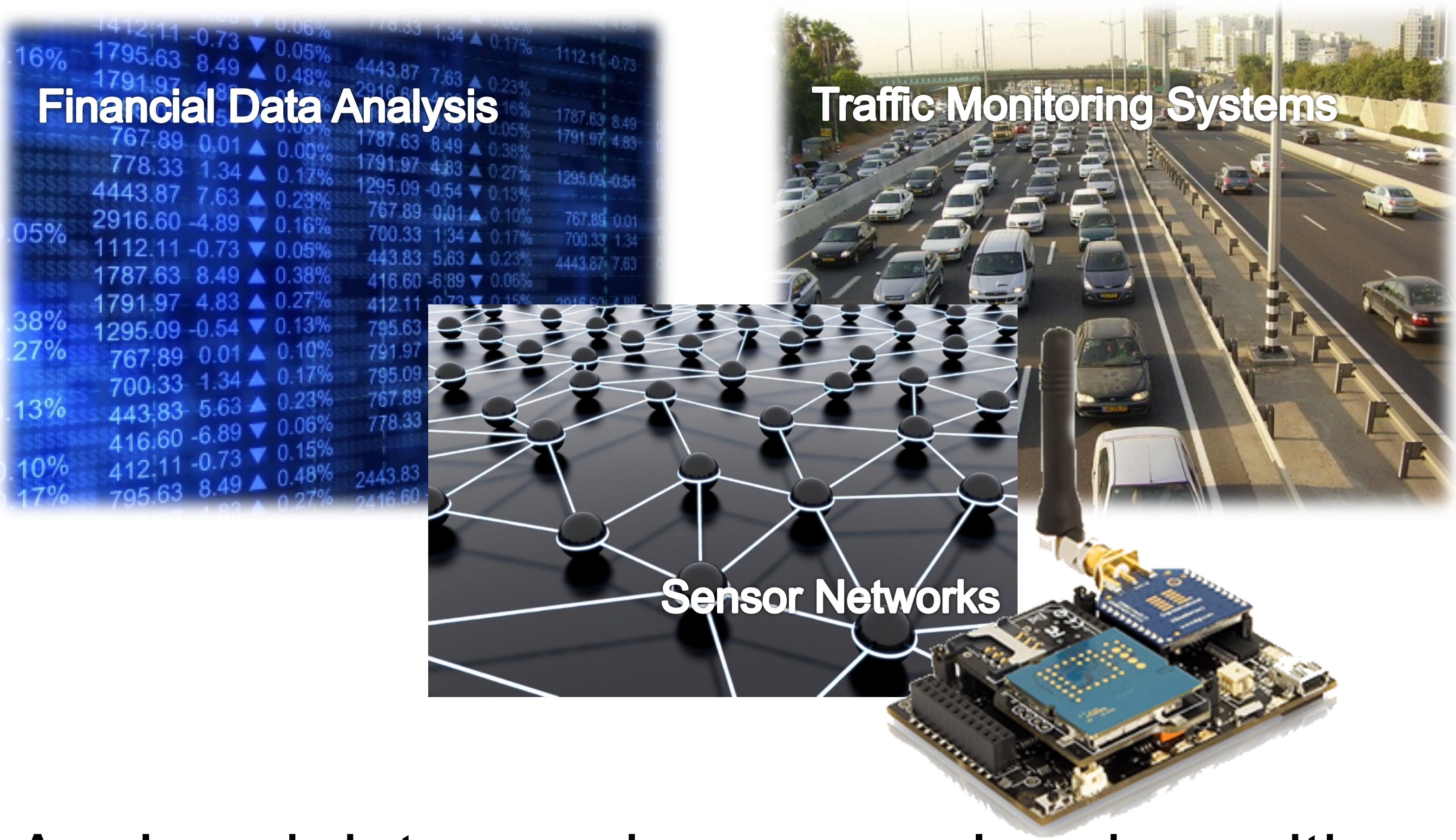
Privacy-Preserving Distributed Stream Monitoring

Arik Friedman[1], Izchak Sharfman[2], Daniel Keren[3], Assaf Schutser[2]

[1] NICTA, Australia [2] Technion, Israel [3] Haifa University, Israel

NICTA Funding and Supporting Members and Partners

Distributed Stream Networks



Analyzed data may be personal and sensitive

Related work…

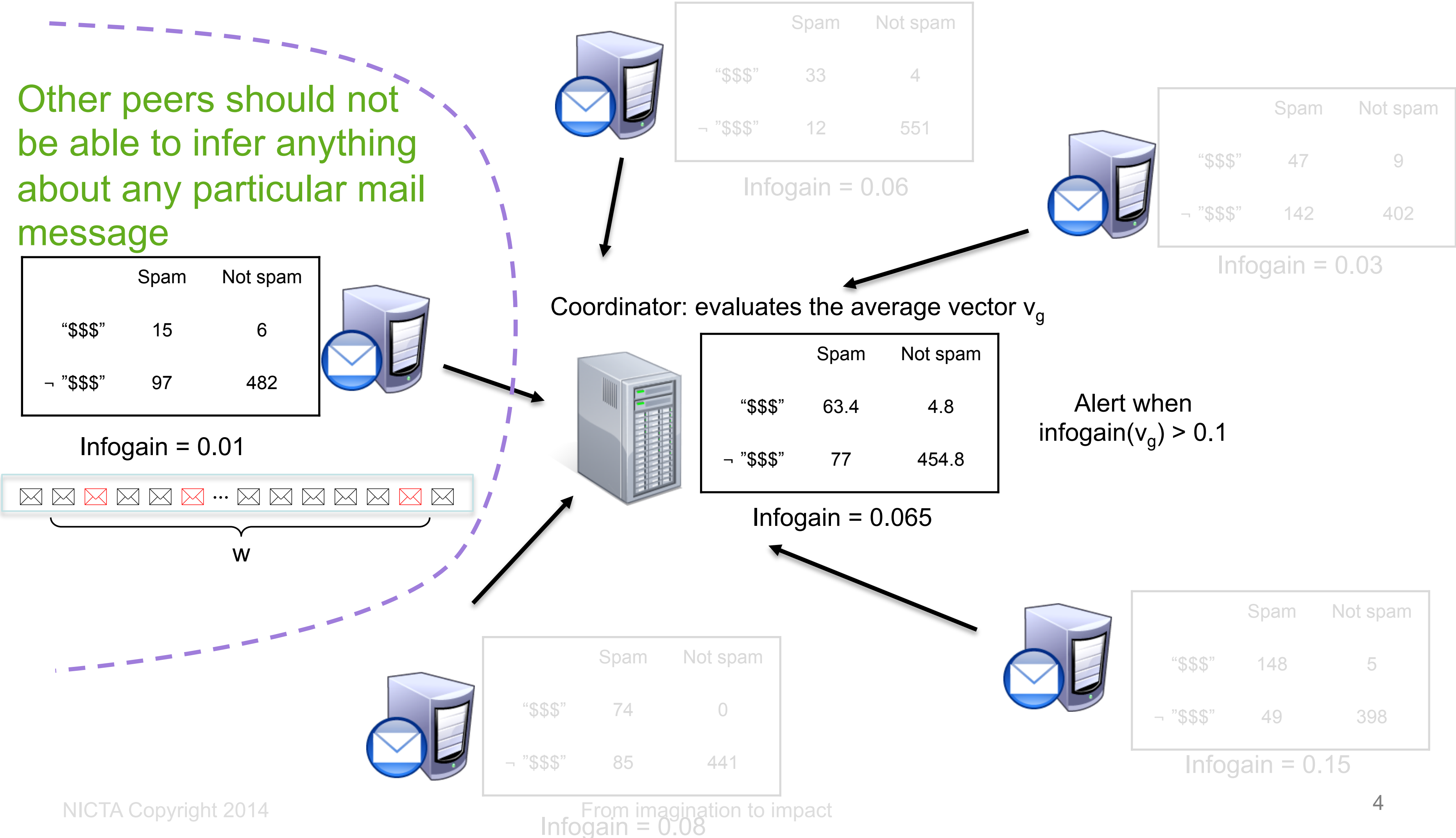
- Continuous monitoring in centralized settings
  - Differential privacy under continual observation [DPNR10]
  - Statistics on sketches [MMNW11]
  - Adaptive sampling [FX12]
- Computation in Distributed settings
  - Distributed noise generation [DKMMN06, CRFG12]
  - Distributed heavy hitters [HKR12]
- Distributed time series data
  - Historical time-series data [RN10]
  - Cryptographic protocols [SCRCS11]
  - Heavy hitters over a sliding window [CLSX12]

This work:

Monitoring complex functions
over statistics derived from streams

Problem Setting

Other peers should not be able to infer anything about any particular mail message

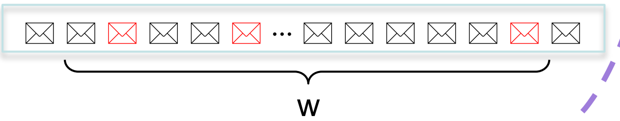


Problem Setting

Other peers should not be able to infer anything about any particular mail message

	Spam	Not spam
“$$$”	15	6
¬ ”$$$”	97	482

Infogain = 0.01



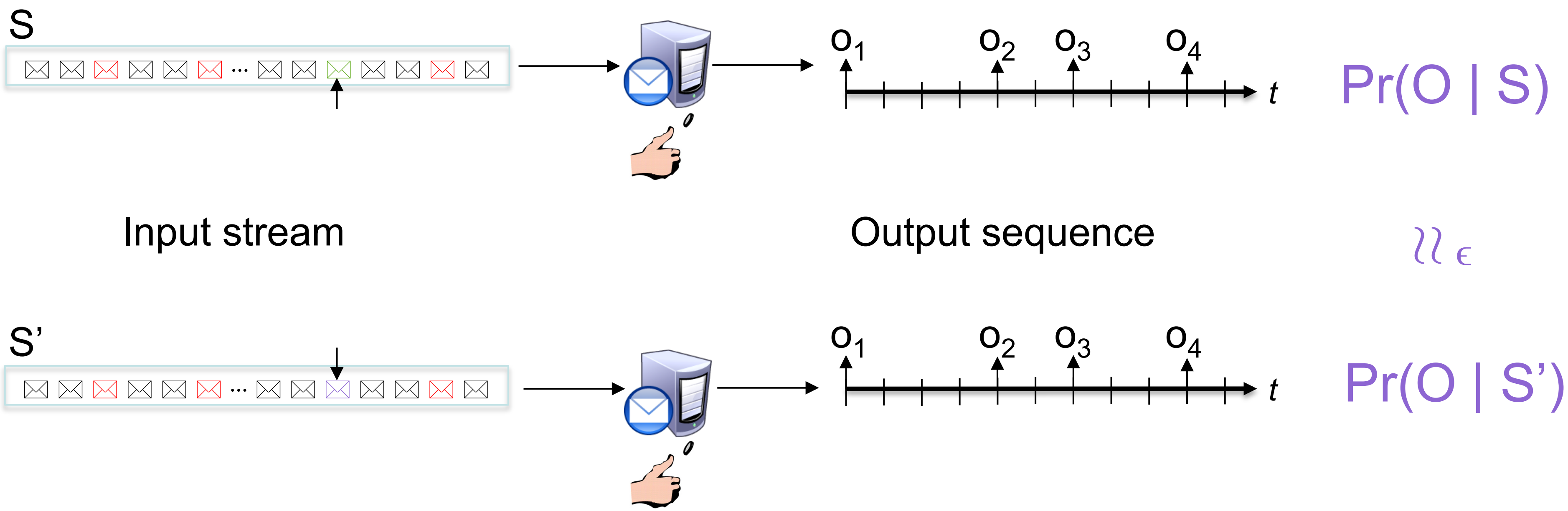
Cryptographic solutions:

☑ Confidentiality
☒ Inferences from the output still possibly

⇒Differential privacy addresses such leaks

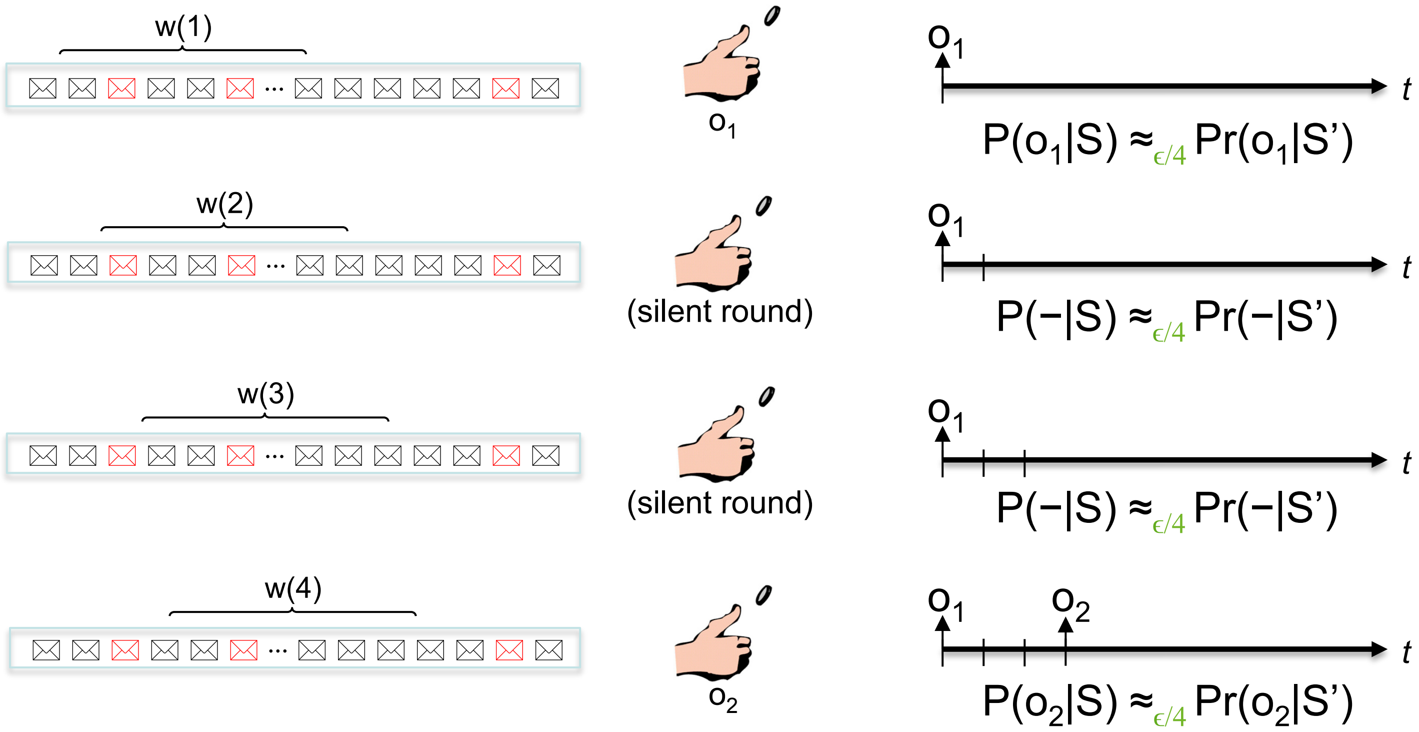
Differential privacy [DPNR10]

For any two *adjacent* streams and for any output sequence O



Large ϵ allows bigger difference between the probabilities
$\Rightarrow$ reflects the input more accurately, less private

Privacy as a Budget - Naïve Solution



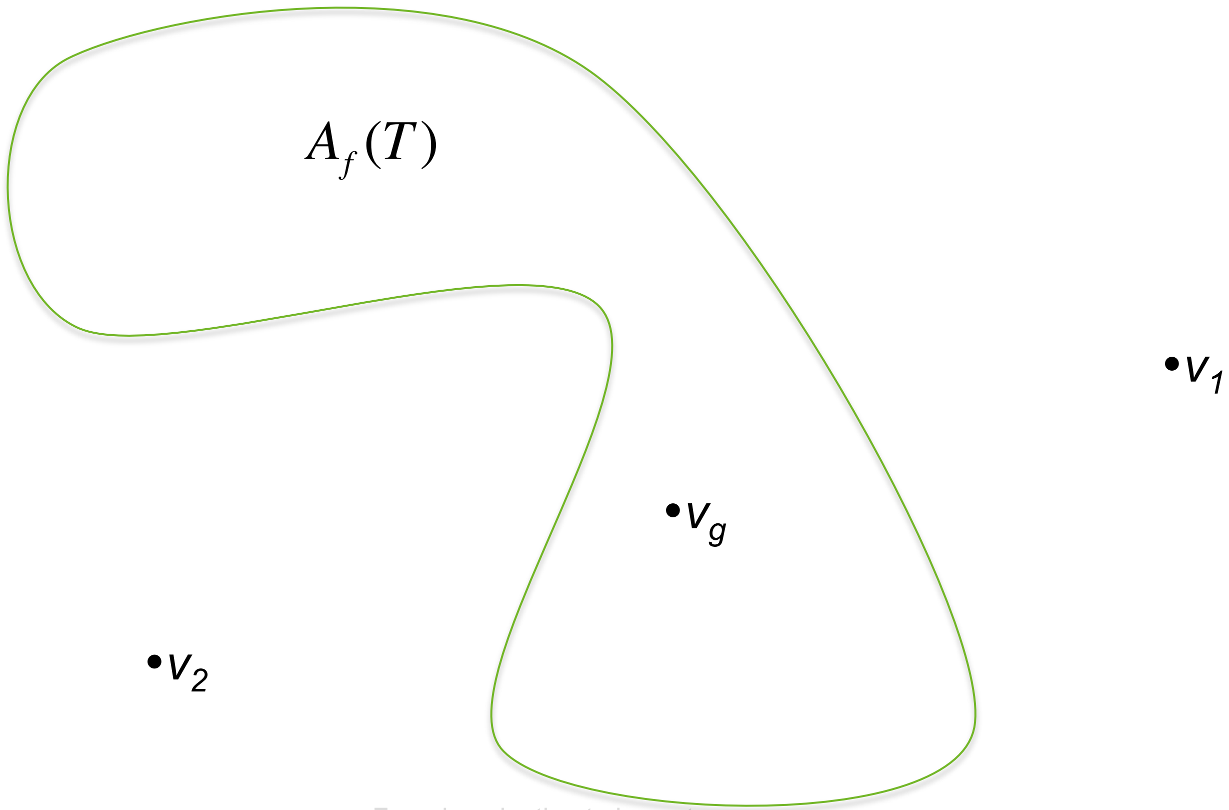
$$\Rightarrow P(o_1{-}{-}o_2|S) \approx_{\epsilon} Pr(o_1{-}{-}o_2|S')$$

Privacy loss in each time period ⇒ wasteful, outputs are not independent
Instead, privacy cost can be *amortized*

Efficient stream monitoring [SSK'06, KSSL'12]

Recall the problem: detect $f(v_g) > T$ for $v_g = \frac{1}{k}\sum_k v_i$

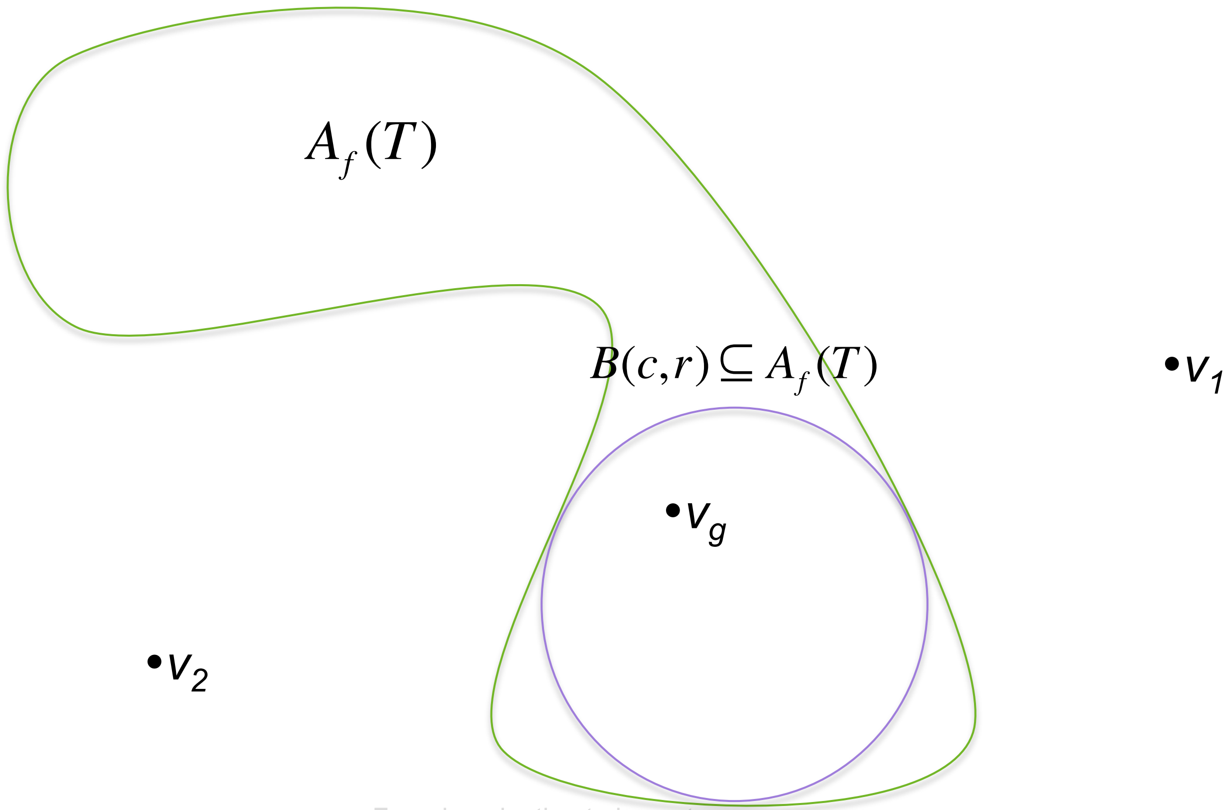
The admissible region: $A_f(T) = \{v \mid f(v) \le T\}$



Efficient stream monitoring [SSK'06, KSSL'12]

Recall the problem: detect $f(v_g) > T$ for $v_g = \frac{1}{k}\sum_k v_i$

The admissible region: $A_f(T) = \{v \mid f(v) \leq T\}$



Efficient stream monitoring [SSK'06, KSSL'12]

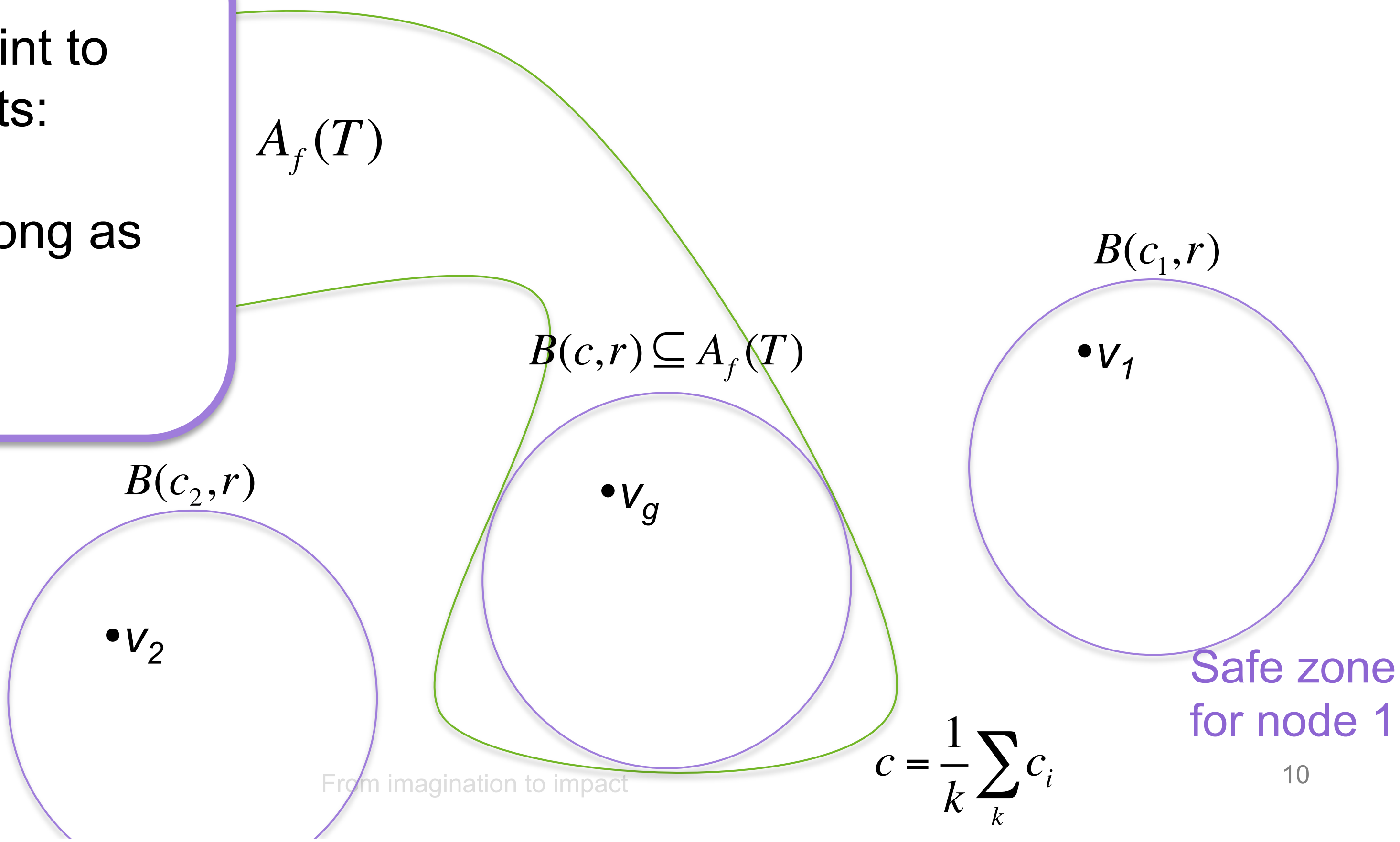
Recall the problem: detect $f(v_g) > T$ for $v_g = \frac{1}{k}\sum_k v_i$

le region: $A_f(T) = \{v | f(v) \le T\}$

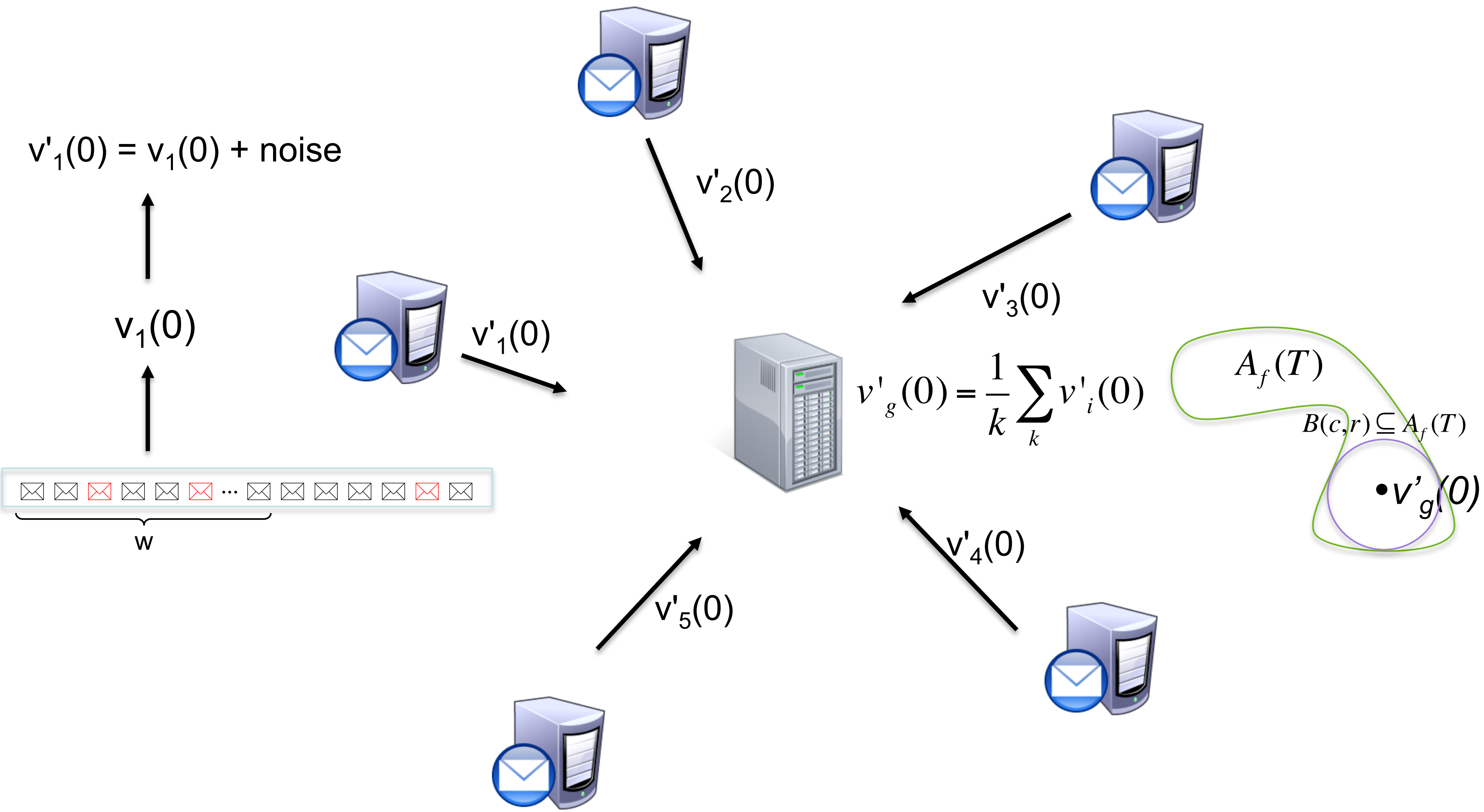
Global constraint to local constraints:

$v_g \in B(c,r)$ as long as

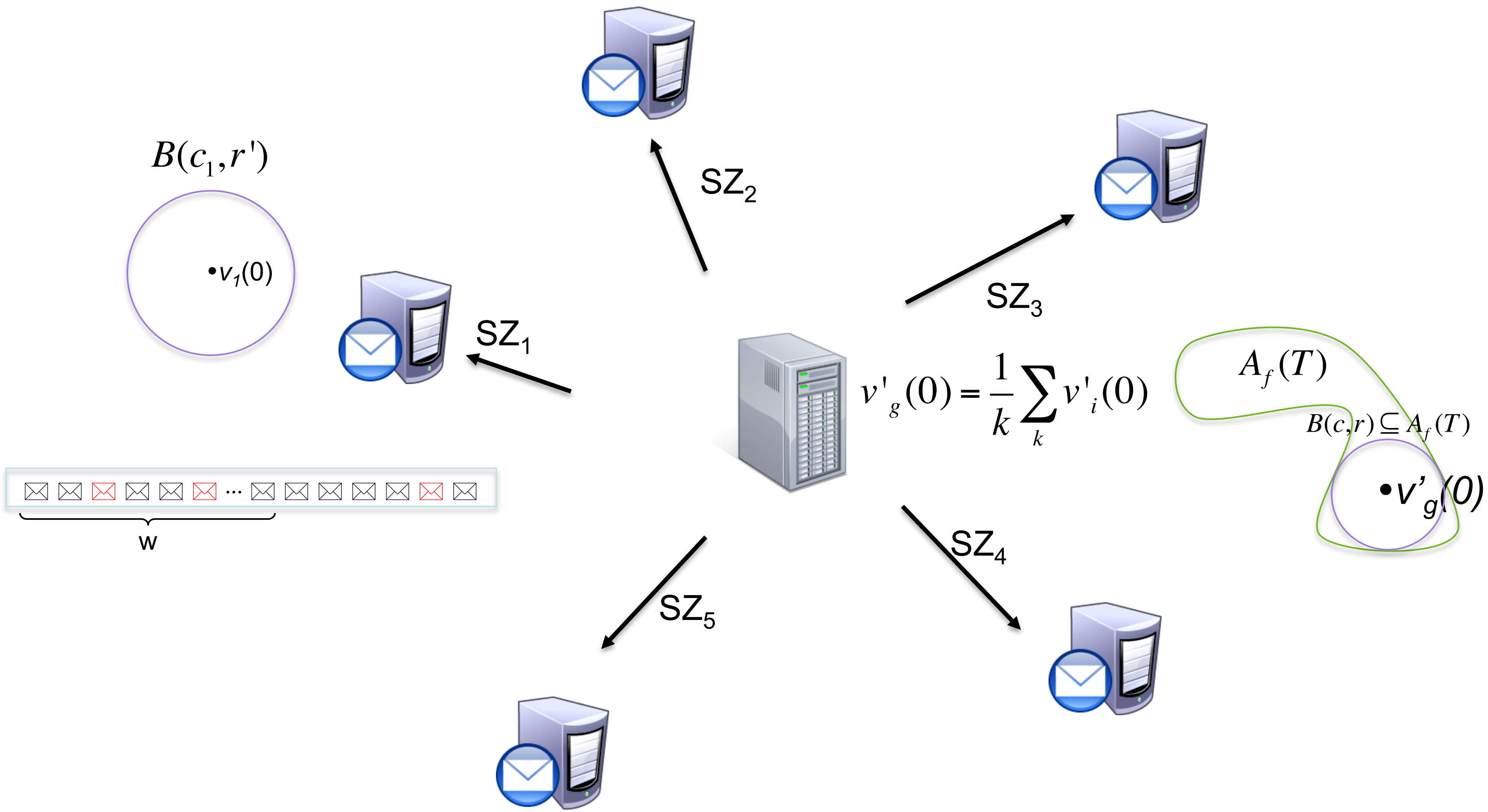
$\forall i : v_i \in B(c_i,r)$



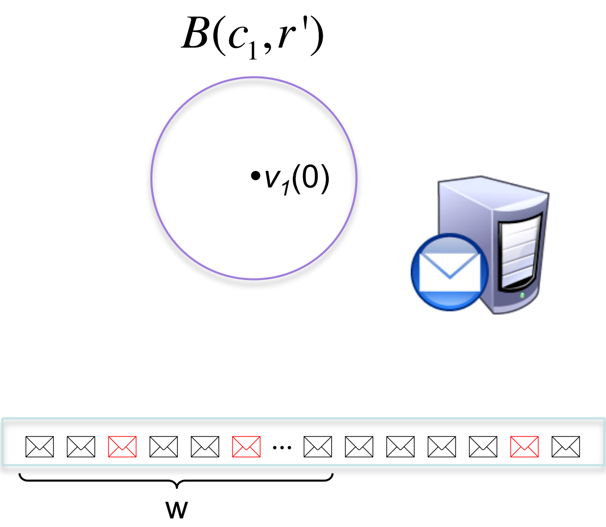
Our Algorithm

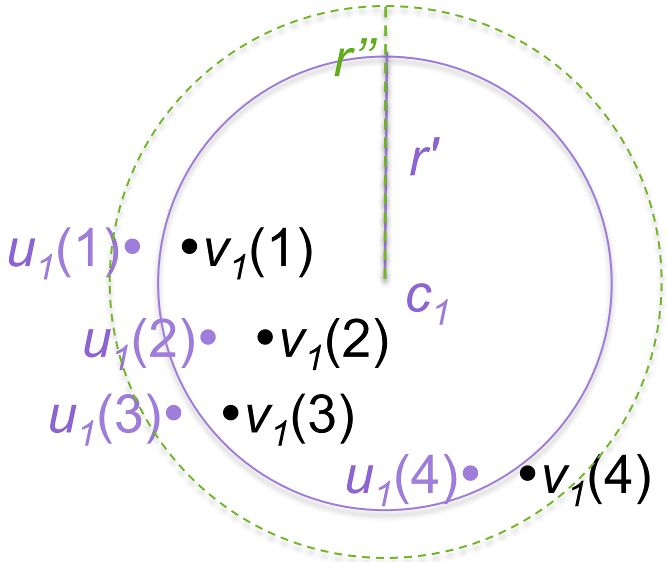


Our Algorithm



Privacy at the Node Level





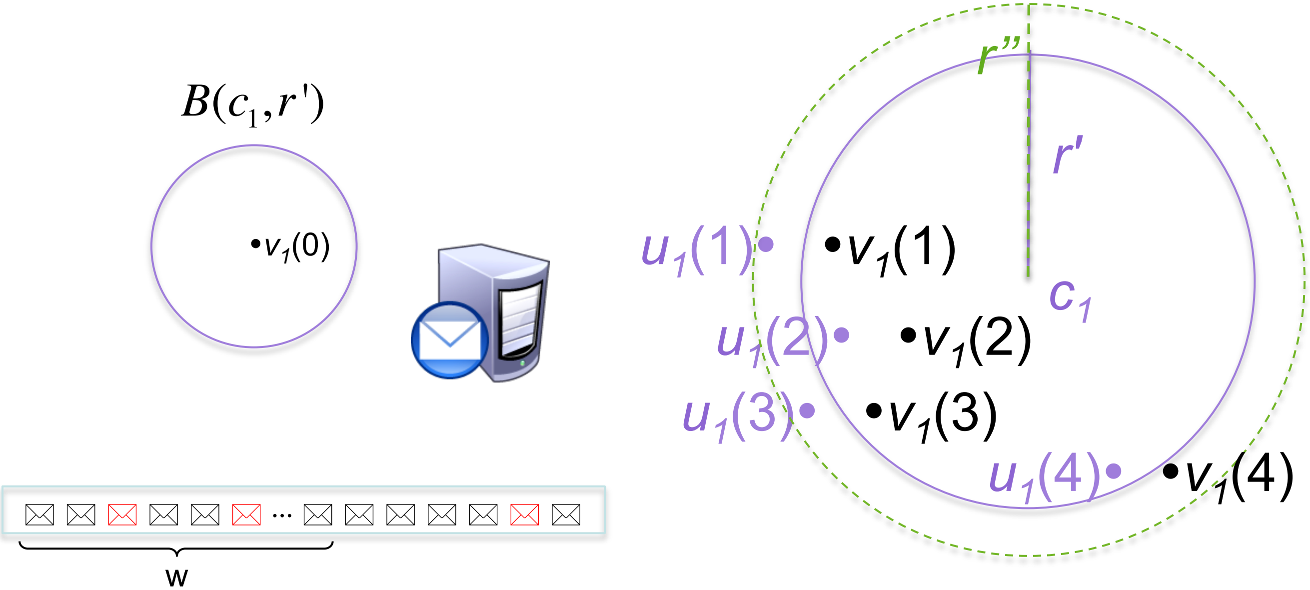
Noise added to the safe zone will protect the privacy in all silent rounds, until a new safe zone is assigned!

Evaluating $v_1(t)$ against
<u>the safe zone in Stream S:</u>
t=1: silent round
t=2: silent round
t=3: silent round
t=4: safe zone breach

Evaluating $u_1(t)$ against
<u>the safe zone in Stream S':</u>
t=1: ~~silent round~~ breach!

⇒ Addressed by adding randomness to the safe zone radius (Laplace mechanism)
Pr(silent | S) ≈ Pr(silent | S') because Pr(r') ≈ Pr(r'')

Privacy at the Node Level

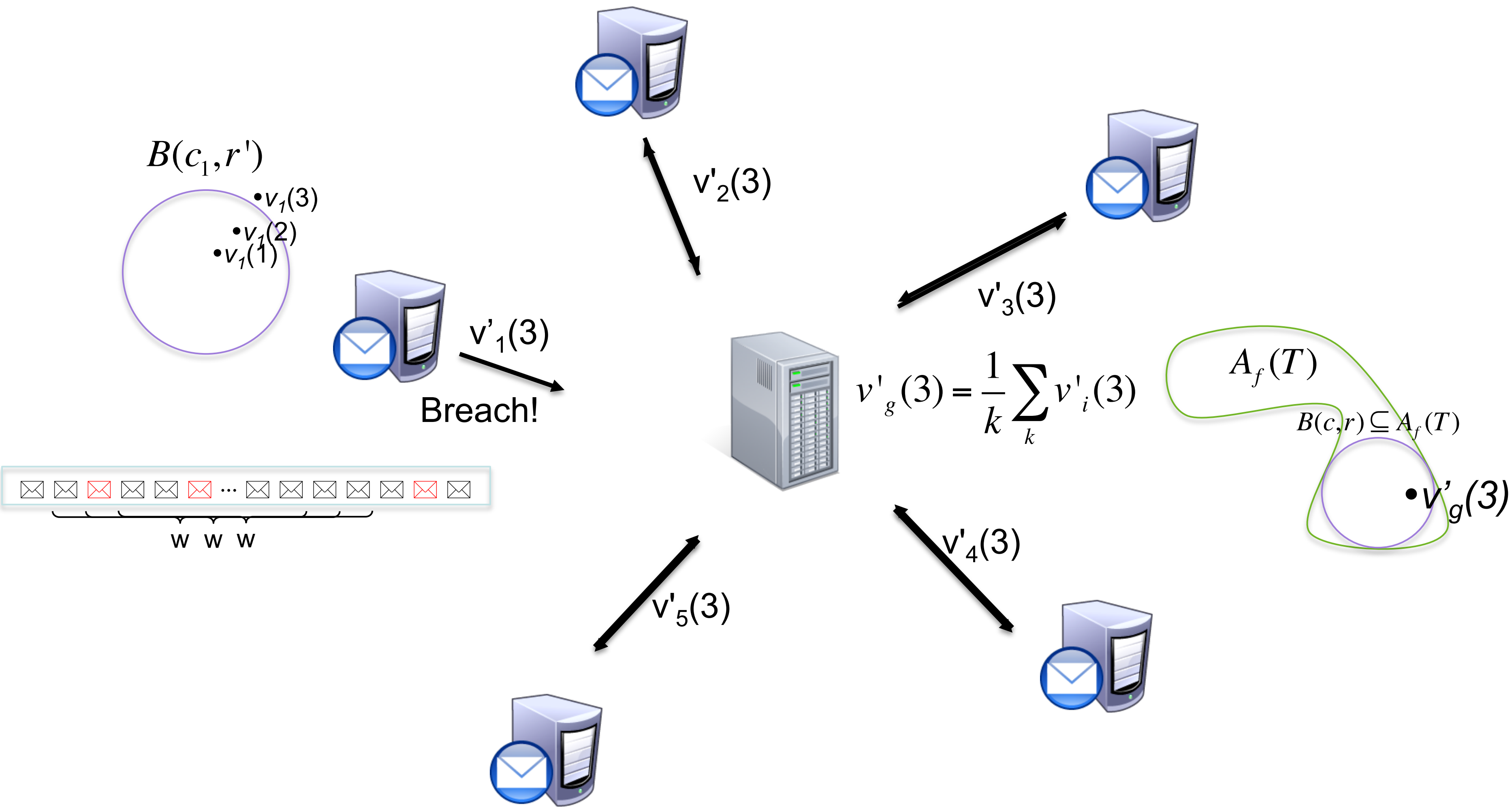


Evaluating $v_1(t)$ against
<u>the safe zone in Stream S:</u>
t=1: silent round
t=2: silent round
t=3: silent round
t=4: safe zone breach

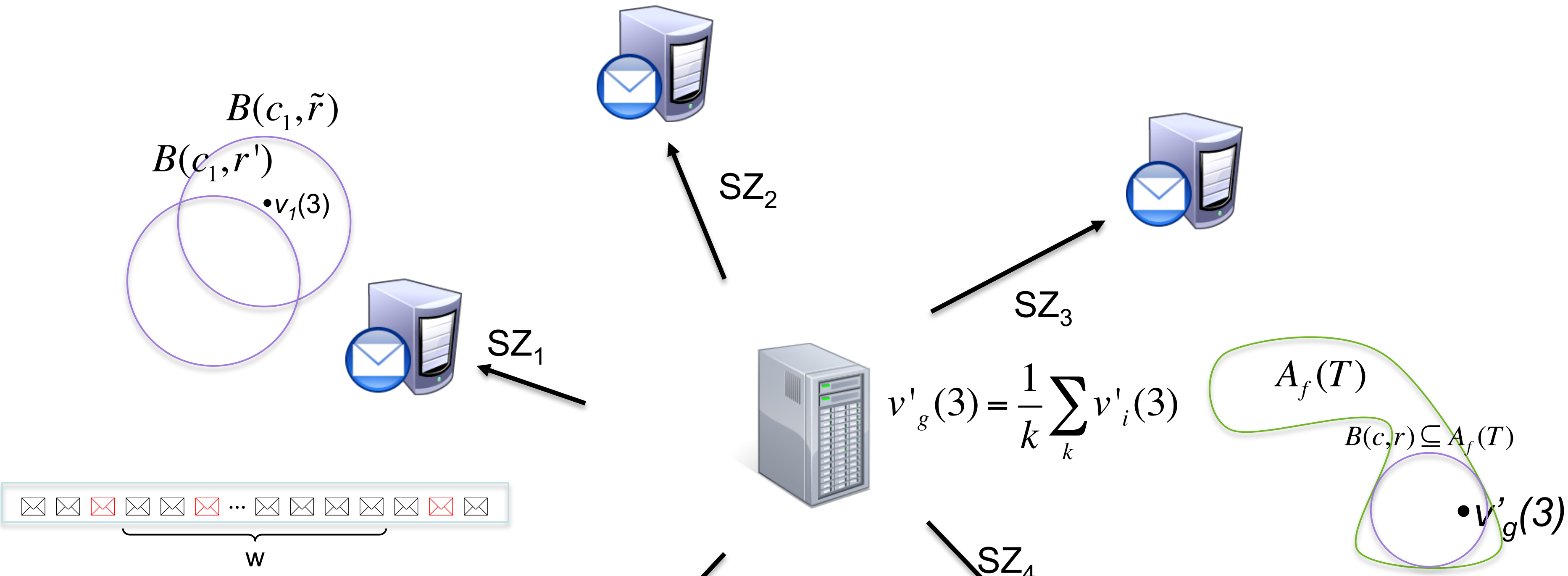
Evaluating $u_1(t)$ against
<u>the safe zone in Stream S':</u>
t=1: silent round
t=2: silent round
t=3: silent round
t=4: ~~safe zone breach~~ silent round

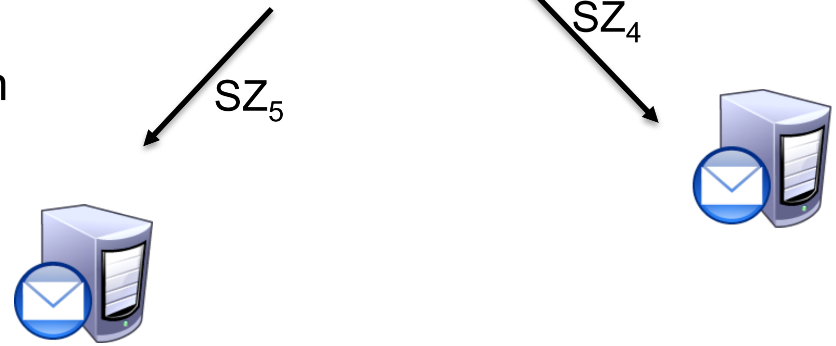
$\Rightarrow$ Addressed by adding randomness (exponential mechanism) when evaluating $v(t) \in_{\varepsilon} B(c, r')$

Our Algorithm

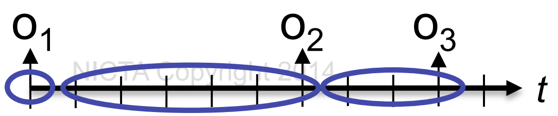


Our Algorithm





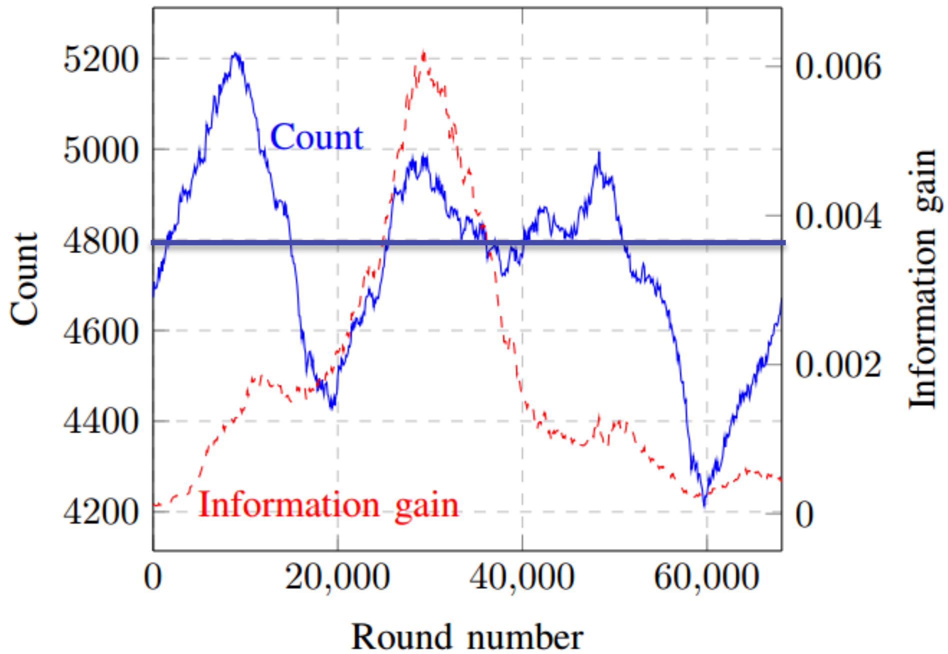
Continue as long as within privacy budget…
(system liftetime:
b local breaches allowed)



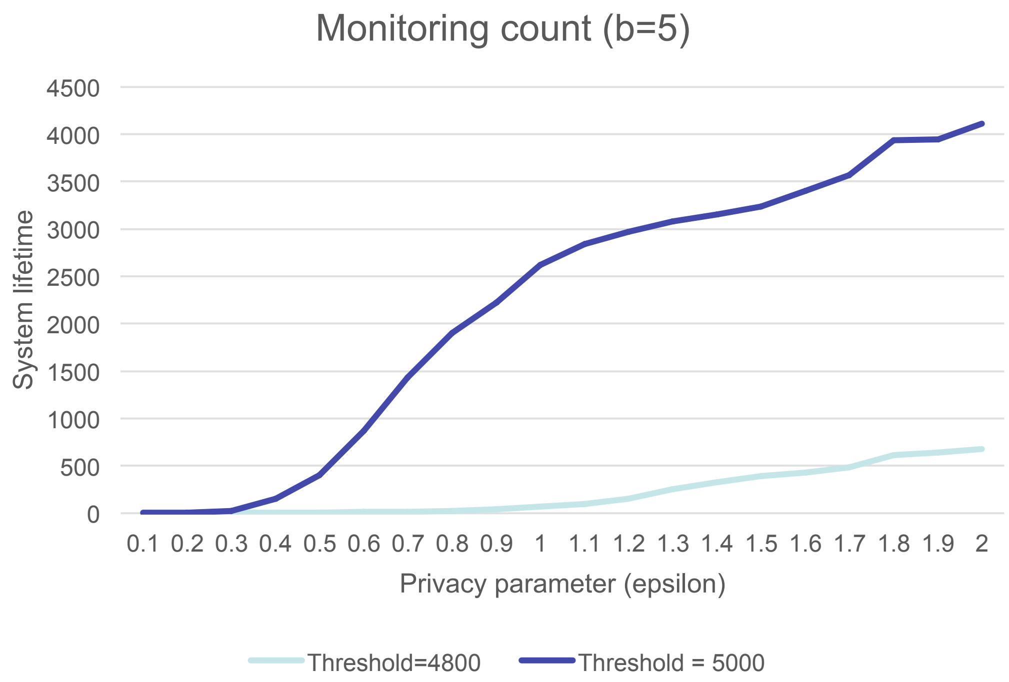
Experimental evaluation

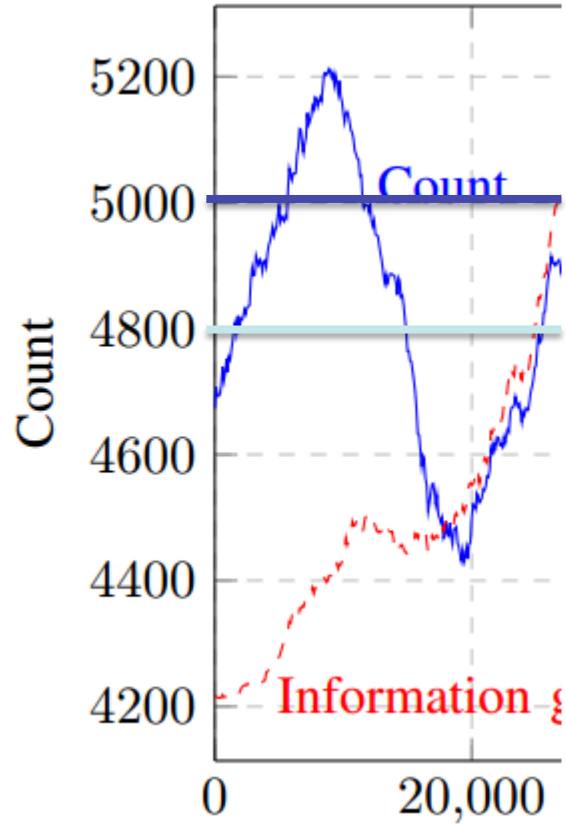
Reuters corpus:

- 781,265 labelled news stories
- Distributed by round robin between 10 nodes
- Each node monitors a window of 10,000 stories
- “CCAT” category denotes spam, “febru” feature a monitored term



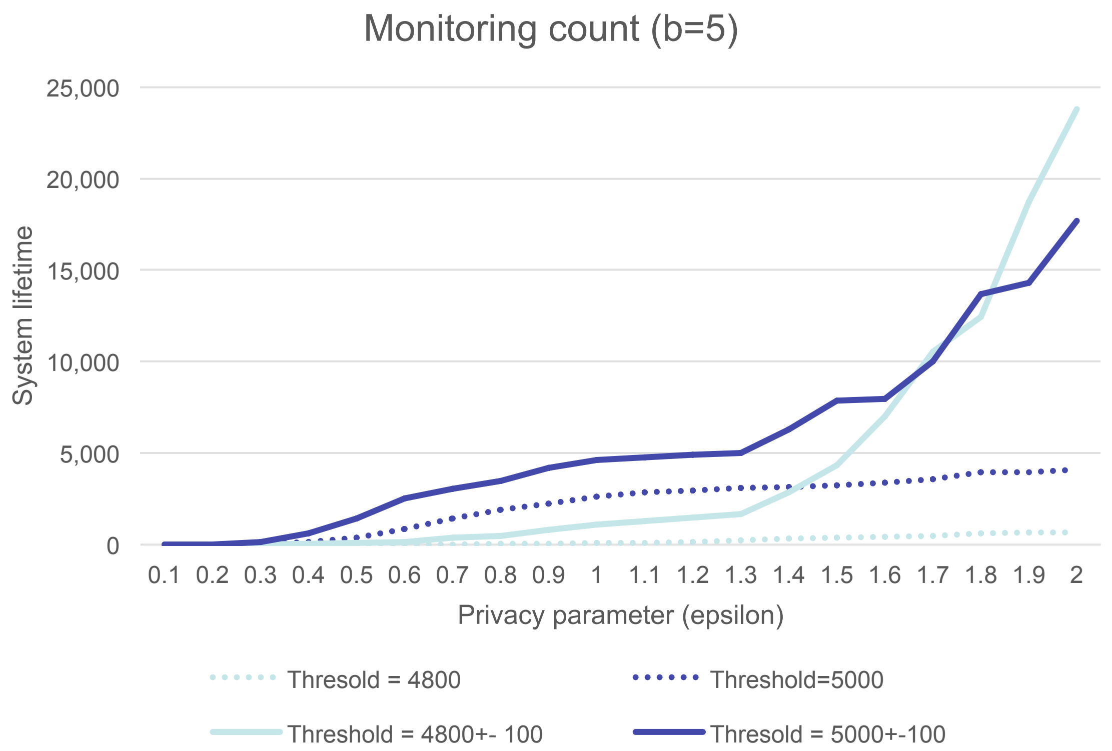
Monitoring count

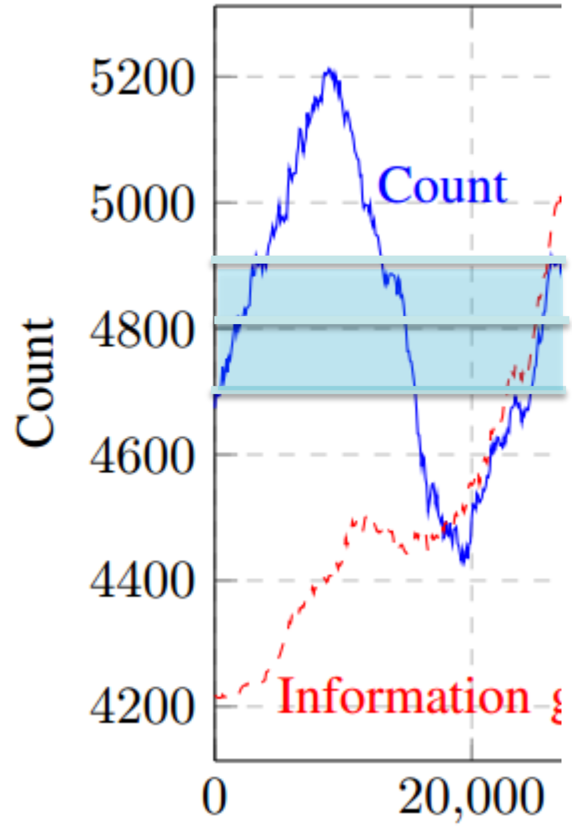




Likelihood of local breach higher when closer to the threshold

Adding error margins





Error margins trade accuracy for longer system lifetime

Additional results in the paper…

- Infogain evaluation
  - Tradeoff between System lifetime, threshold and privacy: we pay for privacy mainly when close to the threshold.
- Error margins trade-offs
- Violation rounds (local breaches *b*) trade-off
- Costs of distributed vs. centralized

Summary and future directions

Communication efficiency translates to better privacy

- Possible enhancements:
  - Local communication between nodes could allow further mitigation of privacy loss
  - Prediction models that tailor safe zones to nodes can reduce the probability of local breaches
  - As the processing window advances, the privacy budget can be replenished

NICTA

Thank you

NICTA @ Sydney
(we hire!)