Anonymity Protocols as Noisy Channels

Kostas Chatzikokolakis, Catuscia Palamidessi and Prakash Panangaden

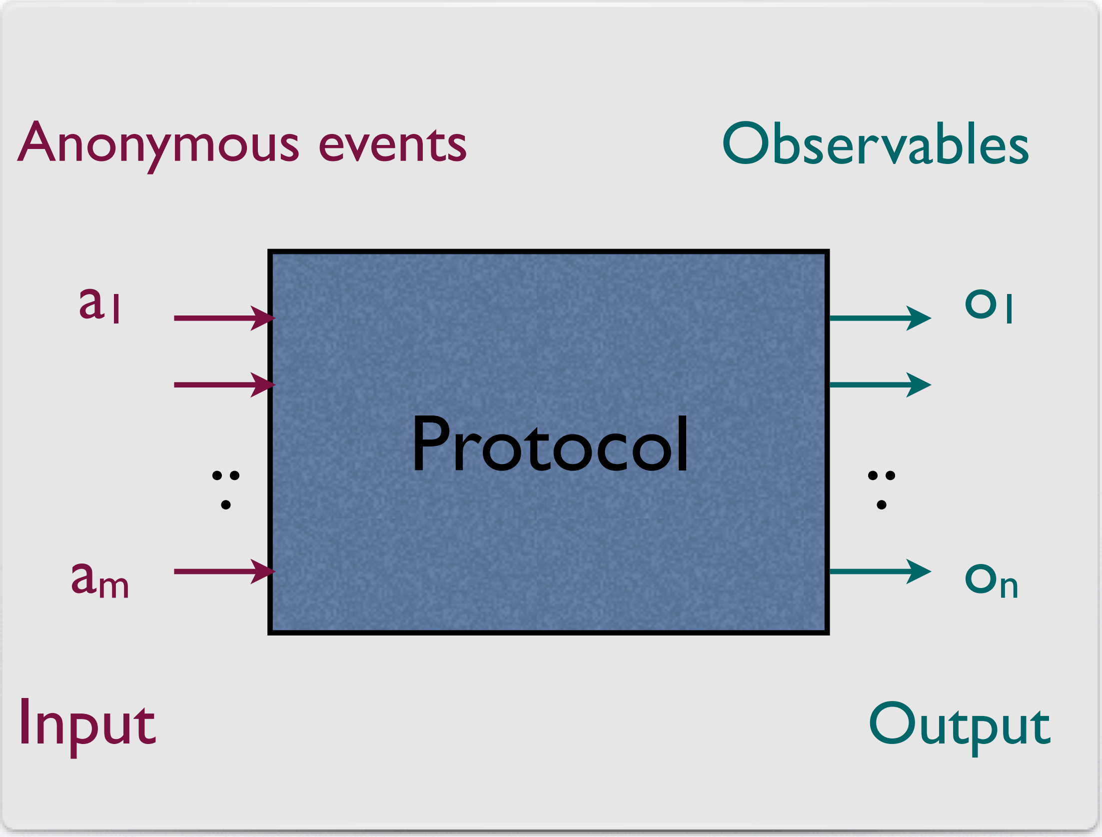
Plan of the talk

- Motivation
- Protocols as channels
- Preliminary notions of Information Theory
- Anonymity as converse of channel capacity
- Statistical inference and Bayesian error
- Relation with other notions in literature

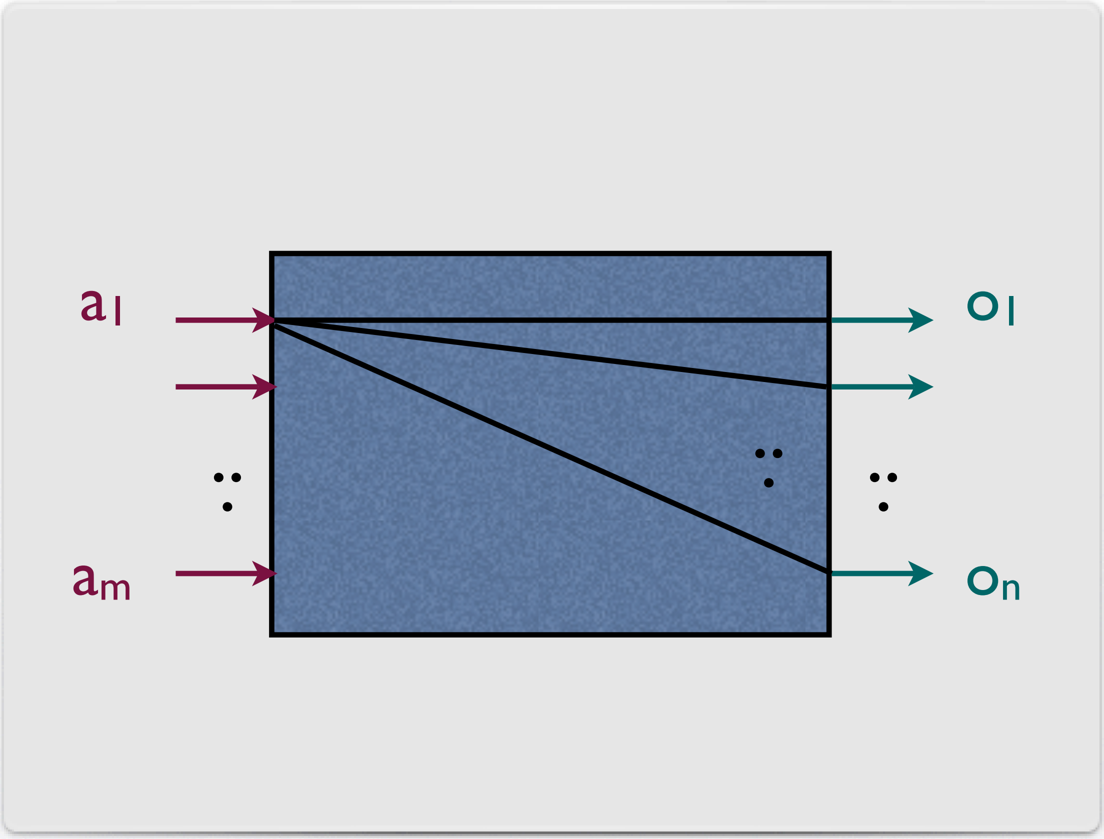
Motivation

Protection of information:

- Identity protection (Anonymity)
  - Hide the link between the data and its sender/receiver
  - The action of sending itself can reveal one's identity
  - Many applications
    - Anonymous message-sending
    - Elections
    - Donations
- Data protection
  - Information flow
  - ...

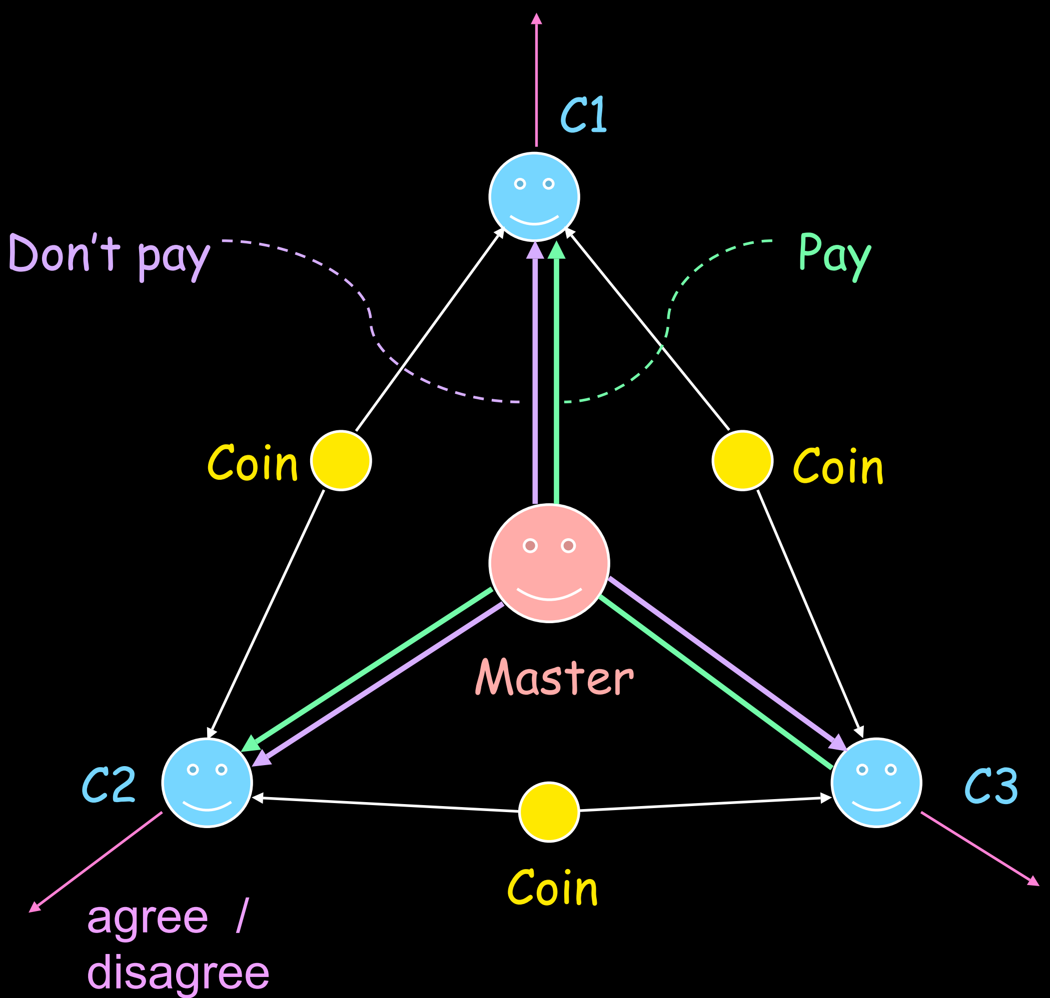


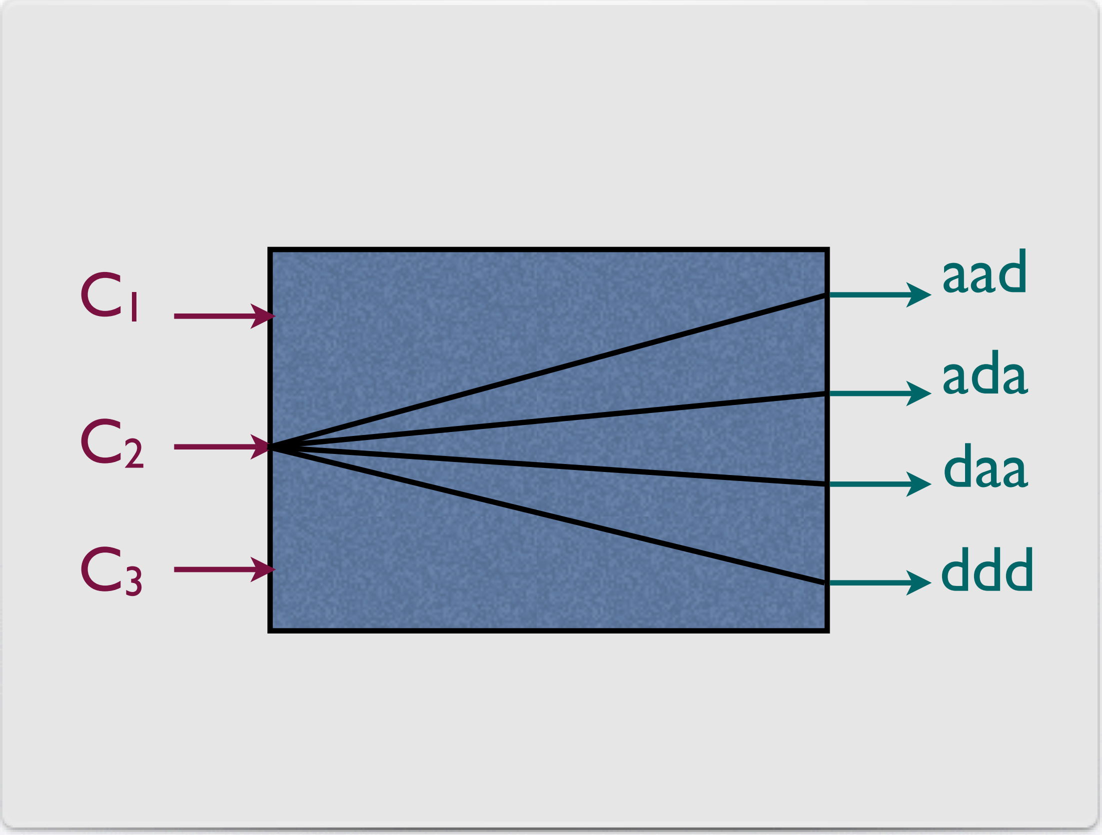
Protocols as channels



Protocols as noisy channels

Example: the dining cryptographers

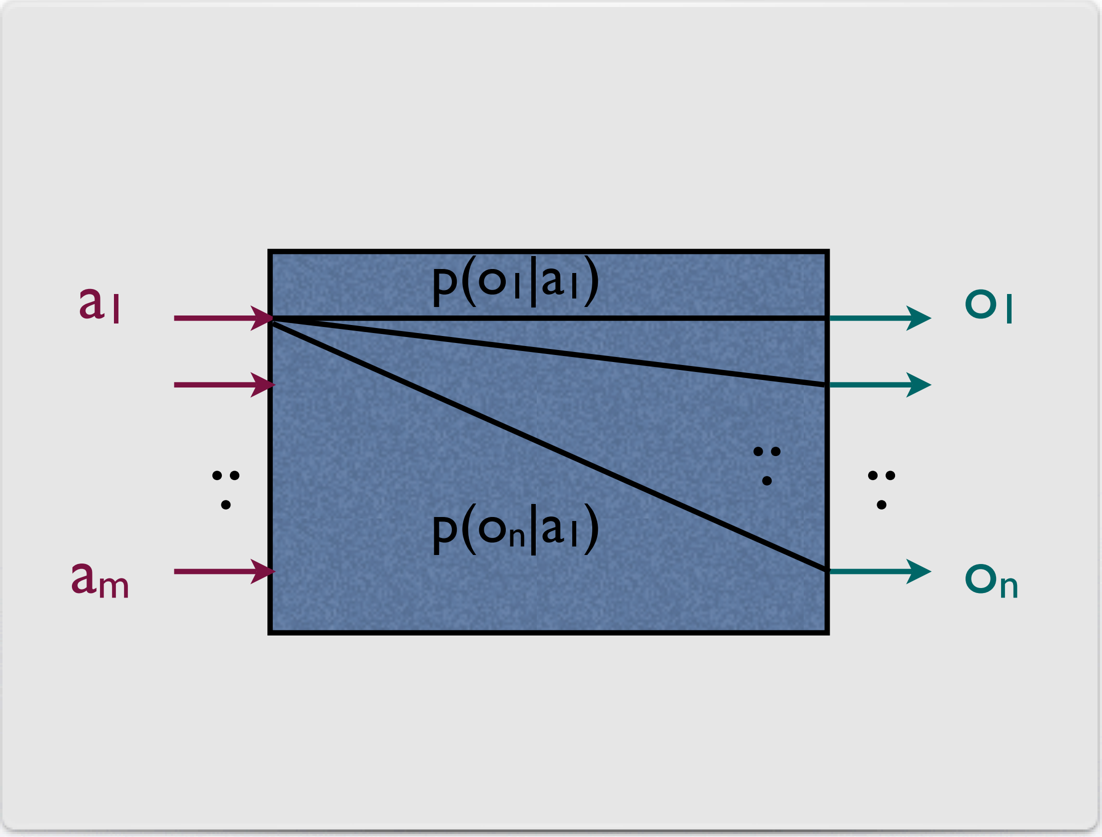




The protocol of the dining cryptographers

Protocols as noisy channels

- We consider a probabilistic approach
  - Inputs: elements of a random variable A
  - Outputs: elements of a random variable O
  - For each input a_i, the probability that we obtain an observable o_j is given by $p(o_j \mid a_i)$
- We assume that the protocol receives exactly one input at each session
- We want to define the degree of anonymity independently from the input's distribution, i.e. the users



The conditional probabilities

	o_1	...	o_n
a_1	$p(o_1 \mid a_1)$	...	$p(o_n \mid a_1)$
$\vdots$	$\vdots$		
a_m	$p(o_1 \mid a_m)$		$p(o_n \mid a_m)$

The channel is completely characterized by the matrix of conditional probabilities

Preliminaries of Information Theory

- The entropy $H(A)$ measures the uncertainty about the anonymous events:

$$H(A) = -\sum_{a \in \mathcal{A}} p(a) \log p(a)$$

- The conditional entropy $H(A|O)$ measures the uncertainty about A after we know the value of O (after the execution of the protocol).
- The mutual information $I(A;O)$ measures how much uncertainty about A we lose by observing O:

$$I(A;O) = H(A) - H(A|O)$$

Degree of Anonymity

- We define the degree of anonymity provided by the protocol as the converse of the capacity of the channel:

$$C = \max_{p(a)} I(A;O)$$

- Note that this definition is independent from the distribution on the inputs, as desired

Relative anonymity

- Some information about A may be revealed intentionally
- Example: elections

- We model the revealed information with a third random variable R

R = number of users who voted for c

Relative anonymity

- We use the notion of conditional mutual information

$$I(A;O|R) = H(A|R) - H(A|R,O)$$

- And define the conditional capacity similarly

$$C_R = \max_{p(a)} I(A;O|R)$$

Partitions: a special case of relative anonymity

- We say that R partitions $\mathcal{X}$ iff $p(r|x)$ is either 0 or 1 for every r, x
- Examples: elections, group anonymity

Theorem

If R partitions $\mathcal{A}$ and $\mathcal{O}$ then the transition matrix of the protocol is of the form

$$\begin{array}{c|cccc} & \mathcal{O}_1 & \mathcal{O}_2 & \dots & \mathcal{O}_l \\ \hline \mathcal{A}_1 & M_1 & 0 & \dots & 0 \\ \mathcal{A}_2 & 0 & M_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}_l & 0 & 0 & \dots & M_l \end{array}$$

and

$$C_R \leq d \quad \Leftrightarrow \quad C_i \leq d, \forall i \in 1..l$$

where C_i is the capacity of matrix M_i.

Statistical inference

- An adversary tries to infer the hidden information (input) from the observables (output)

- We assume that the adversary can force the re-execution of the protocol (with the same input). Intuitively this increases his inference power

Statistical inference

- $o = o_1, o_2, \ldots, o_k$: a sequence of observations
- f : the function used by the adversary to infer the input from a sequence of observations
- Error region of f for input a: $E_f(a) = \{\boldsymbol{o} \in \mathcal{O}^n \mid f(\boldsymbol{o}) \neq a\}$
- Probability of error for input a: $\eta(a) = \sum_{\boldsymbol{o} \in E_f(a)} p(\boldsymbol{o}|a)$
- Bayesian probability of error for f :

$$P_{f_n} = \sum_{a \in A} p(a)\eta(a)$$

Bayesian decision functions

- f is a Bayesian decision function if $f(o) = a$ implies
  $p(o \mid a)\, p(a) \geq p(o \mid a')\, p(a')$ for all a, a' and o
- **Proposition:** Bayesian decision functions minimize the Bayesian probability of error
- Note that the property of being Bayesian depends on the input's distribution

Independence from the users

- However, for large sequences of observations the input distribution becomes negligible:
- **Proposition:** A Bayesian decision function f can be approximated by a function g such that $g(o) = a$ implies

  $$p(o \mid a) >= p(o \mid a') \qquad \text{for all } a, a' \text{ and } o$$

- "approximated" means that the more observations we make, the smaller is the difference in the error probability of f and g

Relation with existing notions

Strong probabilistic anonymity

$p(a) = p(a|o) \quad \forall a, o$ — [Chaum, 88], aka "conditional anonymity" [Halpern and O'Neill, 03].

$p(o|a_i) = p(o|a_j) \quad \forall o, i, j$ — [Bhargava and Palamidessi, 05]

Proposition

An anonymity protocol satisfies strong probabilistic anonymity iff $C = 0$.

Example: Dining cryptographers

	100	010	001	111
a_1	1/4	1/4	1/4	1/4
a_2	1/4	1/4	1/4	1/4
a_3	1/4	1/4	1/4	1/4

Strong anonymity and Bayesian inference

- When the rows of the matrix associated to the protocol are all the same, the adversary has no criteria for defining the decision function.

- The Bayesian probability of error is maximal:

$$P_E = \frac{|A|-1}{|A|}$$

Probable Innocence

- A weaker notion of anonymity
- Verbally defined [Reiter and Rubin, 98] as:

"from the attacker's point of view, the sender appears no more likely to be the originator of the message than to not be the originator"

- Can be formally defined [Chatzikokolakis and Palamidessi, 05] as:

$$(n-1) \geq \frac{p(o|a)}{p(o|a')} \quad \forall o \in \mathcal{O}, \forall a, a' \in \mathcal{A}$$

Probable Innocence

- Can be generalized into a more general concept of partial anonymity:

$$\gamma \geq \frac{p(o|a)}{p(o|a')} \quad \forall o \in \mathcal{O}, \forall a, a' \in \mathcal{A}$$

Theorem

If a protocol satisfies partial anonymity with $\gamma > 1$ then

$$C \leq \frac{\log \gamma}{\gamma - 1} - \log \frac{\log \gamma}{\gamma - 1} - \log \ln 2 - \frac{1}{\ln 2}$$

Future work

- Challenging problem (not much investigated in statistical inference): infer the input distribution without the power of forcing the input to remain the same through the observations
- Investigate characterizations for other (weaker) notions of information hiding, which are easy to model check (i.e. they do not require to analyze the capacity as a function of the input distribution)
- Develop a logic for efficient model checking