PriSec Research Group

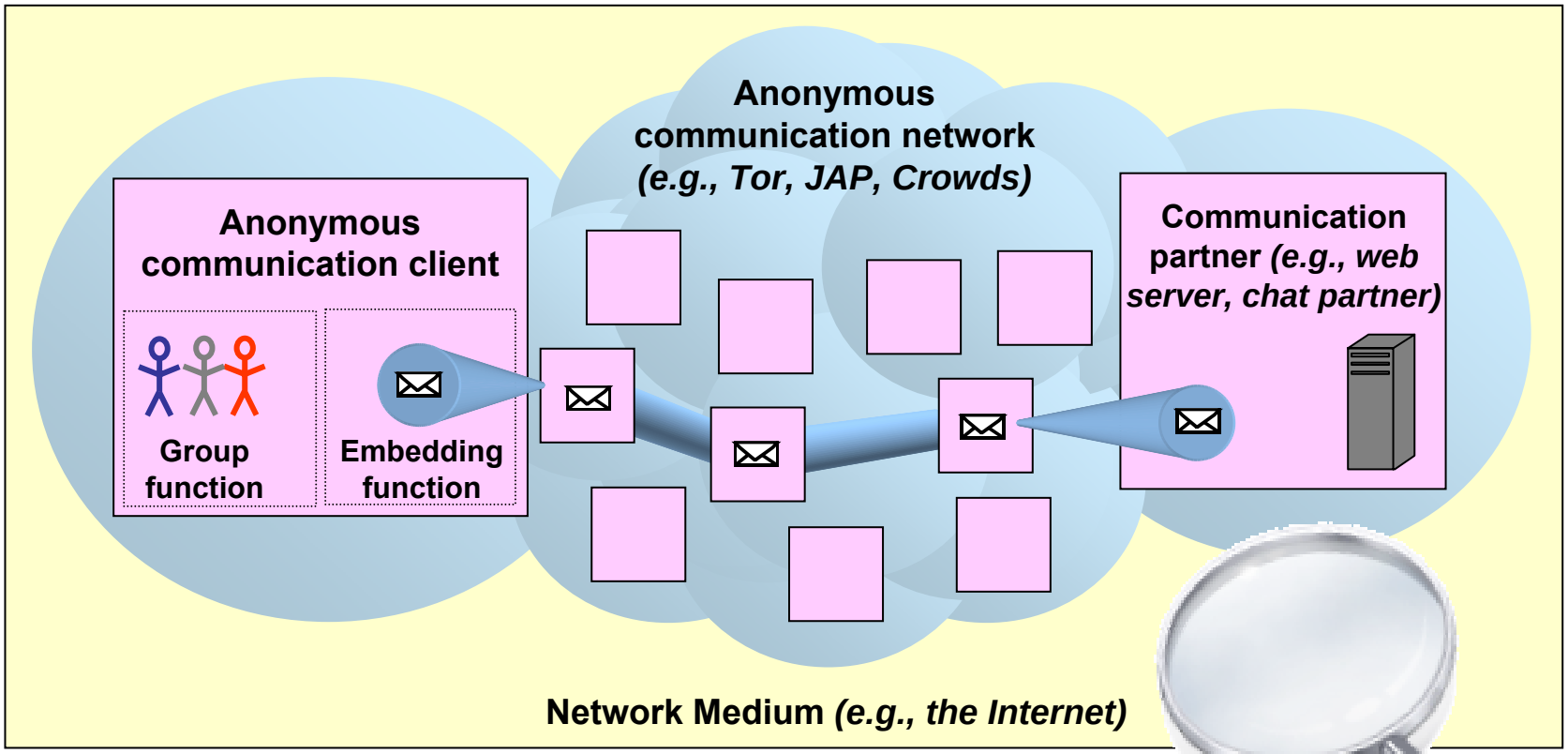
Datavetenskap, Karlstads universitet

Christer Andersson, Reine Lundin

On the Fundamentals of Anonymity Metrics

Christer Andersson
IFIP Summerscool 2007,
6 – 10 th Aug, 2007

Introducing Paper Context



Anonymity Metrics **quantify the degree of (network level) anonymity in a certain scenario**

Methodology in Paper

- 1 Evaluate a set of example scenarios using a selection of state-of-the-art anonymity metrics
- 2 Use the evaluation results of the scenarios together with some basic theory of measurement to formally define a set of criteria for anonymity metrics
- 3 Evaluate the same earlier studied anonymity metrics against these criteria
- 4 If necessary, propose an anonymity metric better suited for fulfilling these criteria

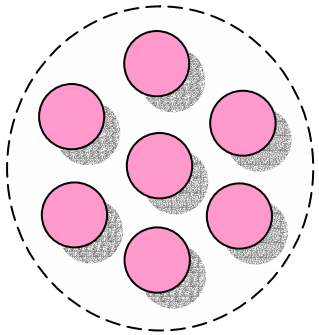
Methodology in Paper

- 1 Evaluate a set of example scenarios using a selection of state-of-the-art anonymity metrics
- 2 Use the evaluation results of the scenarios together with some basic theory of measurement to formally define a set of criteria for anonymity metrics
- 3 Evaluate the same earlier studied anonymity metrics against these criteria
- 4 If necessary, propose an anonymity metric better suited for fulfilling these criteria

Studied Anonymity Metrics

- Anonymity set size (Chaum, 1988)

  The anonymity is quantified as the number of users in the user base – *the anonymity set*

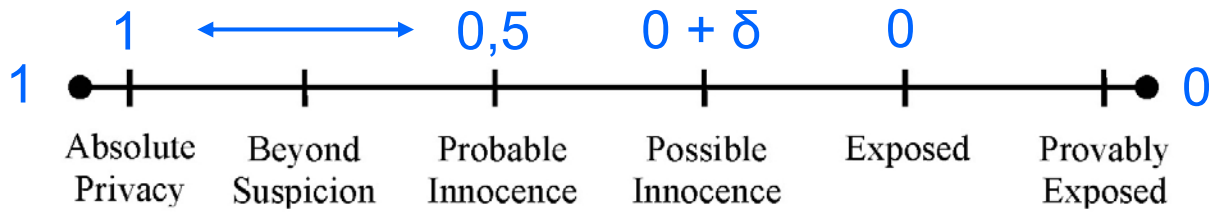


A = 7

- Crowds-based metric (Reiter & Rubin, 1997)

  The *degree of anonymity* is quantified on a continuous scale between “absolute privacy” and “provably exposed”

  This metric can be made more detailed by explicitly by presenting the result as $A = 1 - p_i$



Studied Anonymity Metrics

- The source hiding property (Tóth & Hornák, 2004)

  The anonymity is quantified as the *maximum probability* an attacker can assign the a sender (recipient) regarding the linkability to a certain message

$$\Theta = \max(\mathcal{P}) \qquad 1 \leq \Theta \leq \frac{1}{n} \qquad \text{best case} \qquad \Theta = \frac{1}{n}$$

Example of a probability distribution $\mathcal{P}$

$$\mathcal{P} = \{0.56, \frac{0.44}{8}, \frac{0.44}{8}, \frac{0.44}{8}, \frac{0.44}{8}, \frac{0.44}{8}, \frac{0.44}{8}, \frac{0.44}{8}, \frac{0.44}{8}, 0\}$$

Studied Anonymity Metrics

- Entropy based metric (Serjantov & Danezis, 2002)

  The *effective anonymity set size* is the remaining information the attacker needs to obtains to identify the sender (recipient)

$$\mathcal{S} = H(\mathcal{P}) = -\sum_{i=1}^{n} p_i log_2(p_i)$$
$$0 \leq H(\mathcal{P}) \leq log_2(n)$$

- Entropy based metric (Claudia Diaz *et. al*., 2002)

  The *degree of anonymity* is quantified as the normalized entropy regarding who is the sender (recipient) of a message

$$d = \frac{H(\mathcal{P})}{H(\mathcal{U})} = \frac{H(\mathcal{P})}{log_2(n)} \quad \text{where} \quad 0 \leq d \leq 1$$

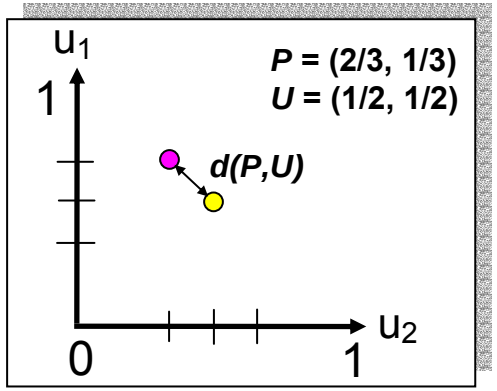
Studied Anonymity Metrics

- Euclidian distance in n-space (our proposal)

  An alternative way of measuring the uniformity of the probability distribution *P*. It outputs the ordinary distance between *P* and *U* when plotted in an *n*-dimensional space.

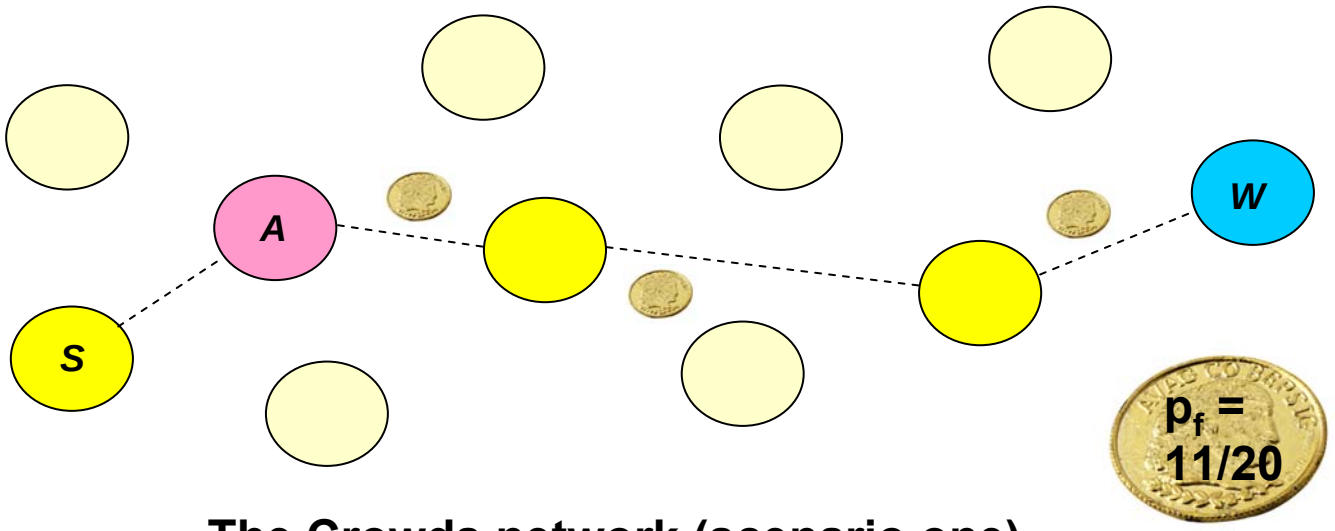
  As a comparison, *H(P)/H(U)* is also an alternative measure of the uniformity of *P*. Another option would be *H(U) – H(P)*

$$d(\mathcal{P}, U) = \sqrt{\sum_{i=1}^{n} (p_i - u)^2}$$



Evaluation of Scenarios (Summary #1)

- Calculate the degree of sender anonymity (recipient anonymity in the extended version of the paper) against malicious jondos and the web server



The Crowds network (scenario one)

Evaluation of Scenarios (Summary #2)

- Some observations:
  - All metrics except anonymity set size yielded a higher degree of anonymity against the web server (this was because *P*, from the perspective of the web server, was uniformly distributed)
  - Although stated so, we do not think that the entropy based metric by Serjantov & Danezis represents the “effective anonymity set size”
  - We observed that the measuring the *Euclidian distance in n-space* behaved fairly similar to the probability based anonymity metrics (future work)

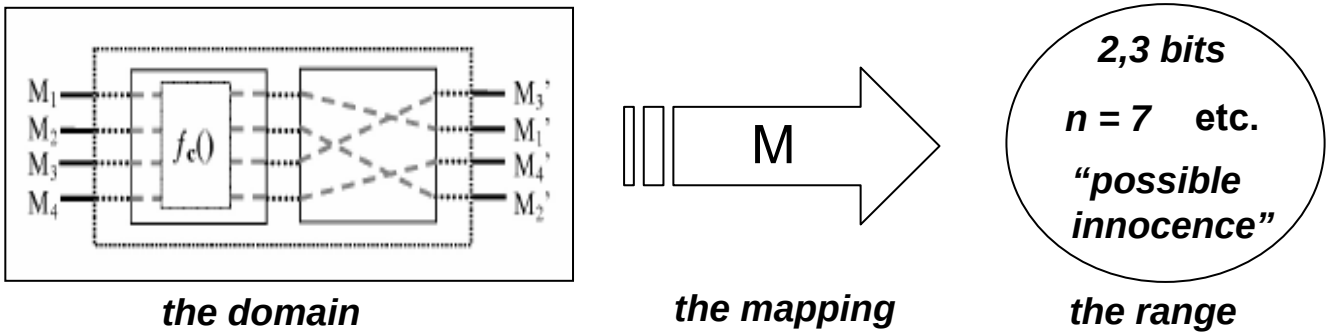
Methodology in Paper

- 1 Evaluate a set of example scenarios using a selection of state-of-the-art anonymity metrics
- 2 Use the evaluation results of the scenarios together with some basic theory of measurement to formally define a set of criteria for anonymity metrics
- 3 Evaluate the same earlier studied anonymity metrics against these criteria
- 4 If necessary, propose an anonymity metric better suited for fulfilling these criteria

Basic Theory of Measurements

- An anonymity metric is a mapping from the empirical world (*the domain*) to the mathematical world (*the range*) where numbers or symbols are assigned to entities in a system to describe the degree of anonymity
- *The representation condition:*

  "A measurement mapping must map entities into numbers and empirical relations into numerical relations in such a way that the empirical relations are preserved by the numerical relations"



Criteria for Anonymity Metrics

- C1 – An anonymity metric should base its analysis on probabilities
- C2 – An anonymity metric must have well defined and intuitive endpoints
- C3 – The more uniform the distribution *P*, the higher the degree of anonymity (*rep. cond.*)
- C4 – The more the users in the anon. set, the higher the degree of anonymity (*rep. cond.*)
- C5 – The elements in the metric's value domain should be well defined
- C6 – The value domain of the metric should be ordered and not too coarse

Methodology in Paper

- 1 Evaluate a set of example scenarios using a selection of state-of-the-art anonymity metrics
- 2 Use the evaluation results of the scenarios together with some basic theory of measurement to formally define a set of criteria for anonymity metrics
- 3 Evaluate the same earlier studied anonymity metrics against these criteria
- 4 If necessary, propose an anonymity metric better suited for fulfilling these criteria

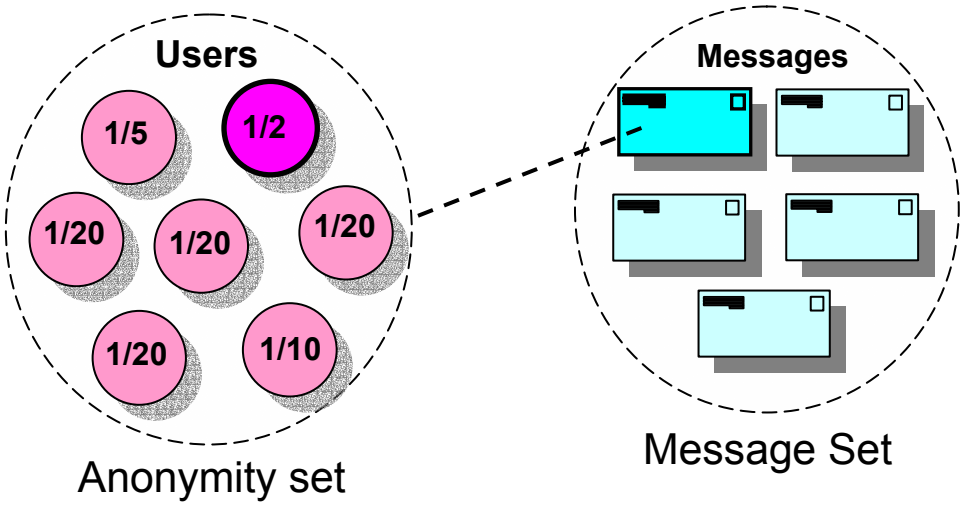
Summary of Survey Results

	C1	C2	C3	C4	C5	C6
Anonymity Set	-	-	-	+	+	+
Crowds-based metric	+	+	-	+	+	-
Entropy-based *(Diaz et al.)*	+	+	+	-	+	+
Entropy-based *(Serjantov & Danezis)*	+	-	+	+	+	+
Source-hiding property	+	-	-	+	+	+

Examples of Survey Results

- C1 – An anonymity metric should base its analysis on probabilities

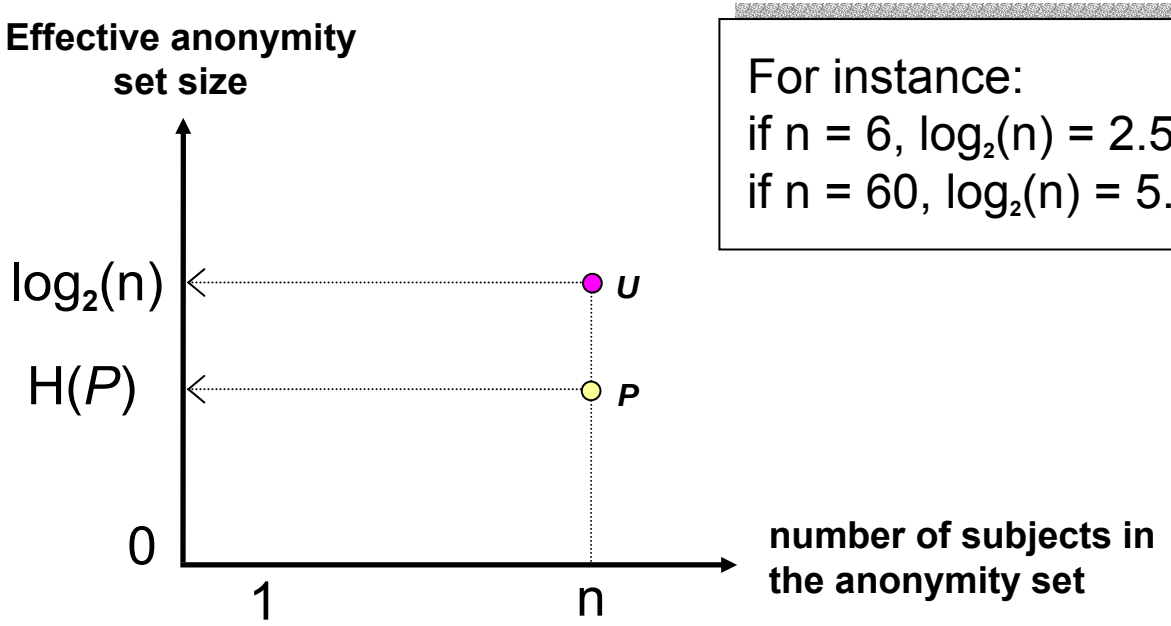
  The anonymity set size metric does not consider probabilities



Examples of Survey Results

- C2 – An anonymity metric must have well defined and intuitive endpoints

  We don't think the endpoints of the entropy-based metric by Serjantov & Danezis are not intuitive. In any case, the theoretical max ($\log_2(n)$) should always be made explicit

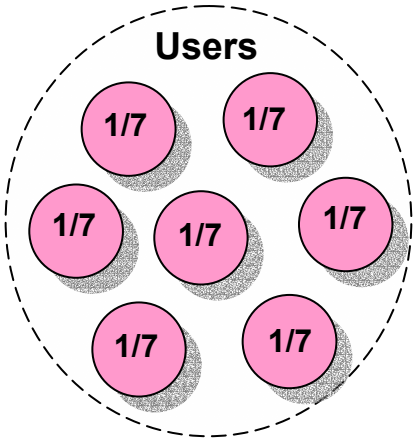


For instance:
if n = 6, $\log_2(n) = 2.58$
if n = 60, $\log_2(n) = 5.91$

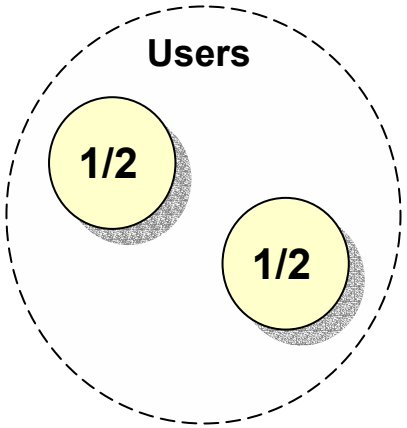
Examples of Survey Results

- C4 – The more the users in the anonymity set, the higher the anonymity

  This is not necessarily the case for the Entropy-based metric by Diaz *et al.*, as the degree of anonymity is normalized and the output is in the range of 0 and 1



Anonymity set #1



Anonymity set #2

Methodology in Paper

- 1 Evaluate a set of example scenarios using a selection of state-of-the-art anonymity metrics
- 2 Use the evaluation results of the scenarios together with some basic theory of measurement to formally define a set of criteria for anonymity metrics
- 3 Evaluate the same earlier studied anonymity metrics against these criteria
- 4 If necessary, propose an anonymity metric better suited for fulfilling these criteria

Scaled Anonymity Set Size

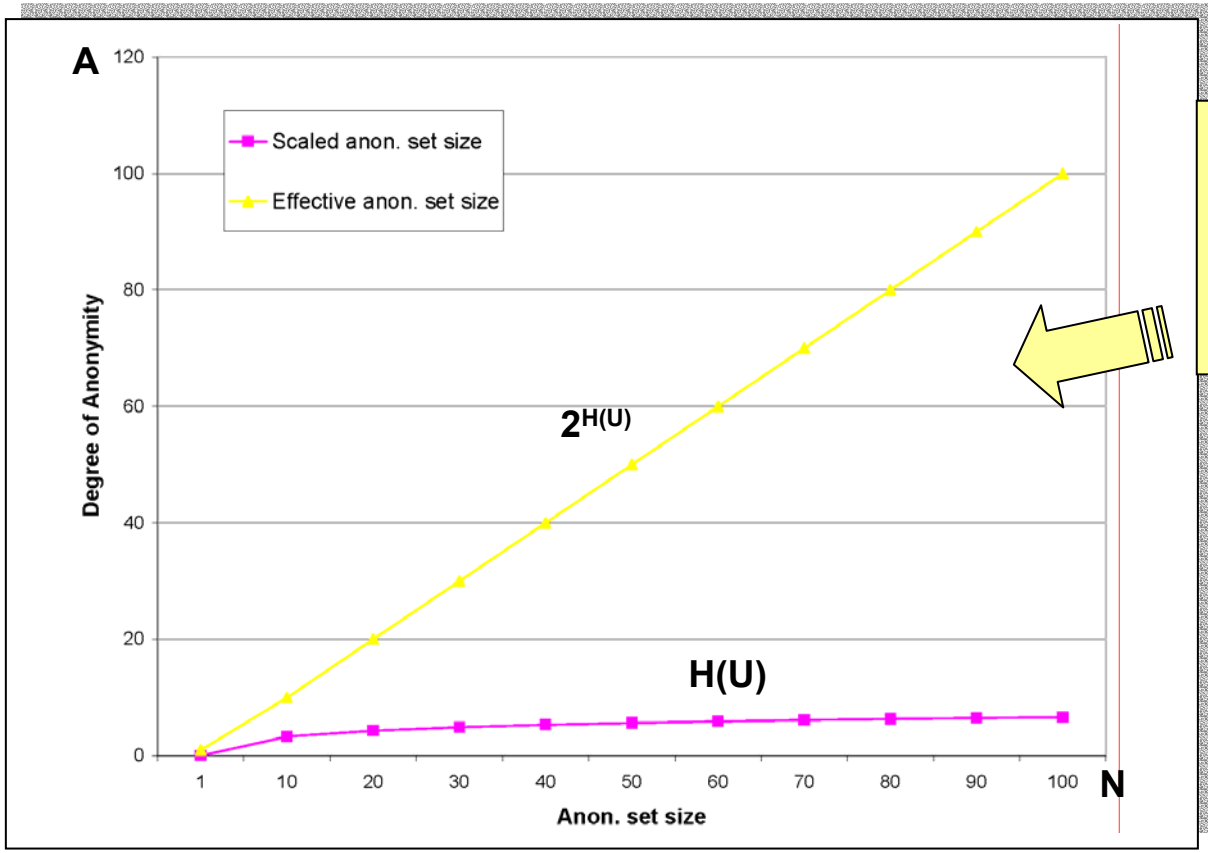
$$A = 2^{H(\mathcal{P})}$$

- H(P) is (a lower bound for) the expected amount of binary questions the attacker needs to answer to identify the sender

→ **$2^{H(P)}$ is the expected number of possible outcomes given H(P)**

- Based on probabilities (*C1*)
- The endpoints overlap with those of the anonymity set size, $1 \leq A \leq n$ (*C2*),
- Increases with an increasing uniformity of P and a growing number of users (*C3, C4*)
- Well defined semantics (*C5*)
- The degree of anonymity is ordered and continuous (*C6*)

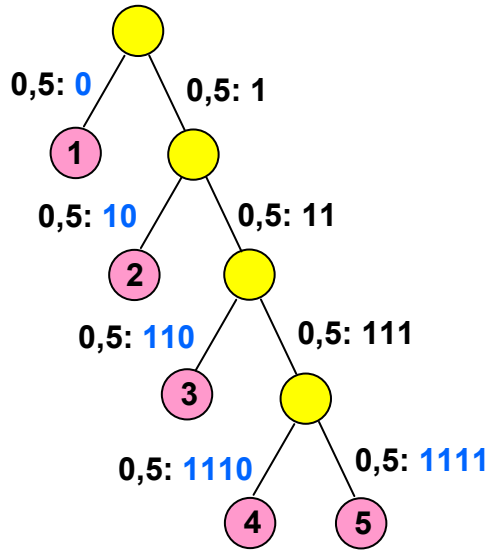
Scaled Anonymity Set Size



Comparison of the entropy-based metric by Serjantov & Danezis and the scaled anonymity set size metric, assuming that P = U (the uniform distribution),

Numerical Example #1

- What would be an optimal strategy for an attacker given P?



Huffman Tree

P = (1/2, 1/4, 1/8, 1/16, 1/16)

p(0) = 1/2, p(10) = 1/4, p(110) = 1/8, p(1110) = 1/16, p(1111) = 1/16

H(P) = 1,875

EQ = 15/8 = 1,875

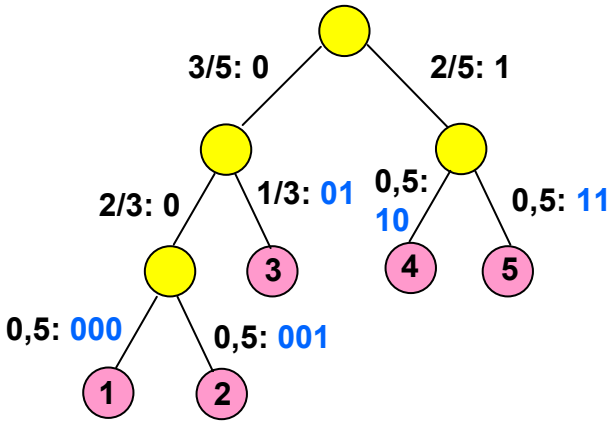
$A = 2^{H(P)} = 3{,}67$

EQ = Expected number of binary questions

$H(P) \leq EQ < H(P) + 1$ (*source coding theorem*)

Numerical Example #2

- What would be an optimal strategy for an attacker given P?



Huffman Tree

P = U = (1/5, 1/5, 1/5, 1/5, 1/5)

p(01) = 1/5, p(10) = 1/5, p(11) = 1/5

p(000) = 1/5, p(001) = 1/5

$H(P) = H(U) = \log_2 5 = \underline{2{,}32}$

$EQ = 12/5 = \underline{2.4}$

$A = 2^{H(P)} = 5$

EQ = Expected number of binary questions

$H(P) \leq EQ < H(P) + 1$ (*source coding theorem*)

Future Work

- Open questions
  - What does $2^{H(P)}$ *really* measure?

    (what does H(P) *really* measure?)
  - Compare H(P) and EQ. How do they differ?
  - What does 2^{EQ} measure?

H(P)	**≤**	**EQ**	**<**	**H(P) + 1 →**
$2^{H(P)}$	**≤**	**2^{EQ}**	**<**	**$2^{H(P)+1}$ =**
$2^{H(P)}$	**≤**	**2^{EQ}**	**<**	**$2*2^{H(P)}$ (w.c. = 2n)**

  - There are many metrics that measures the uniformity of *P* and/or the number of users in the anonymity set. Is this the same as measuring *anonymity?*
  - *Euclidian distance in n-space* yet another metric?