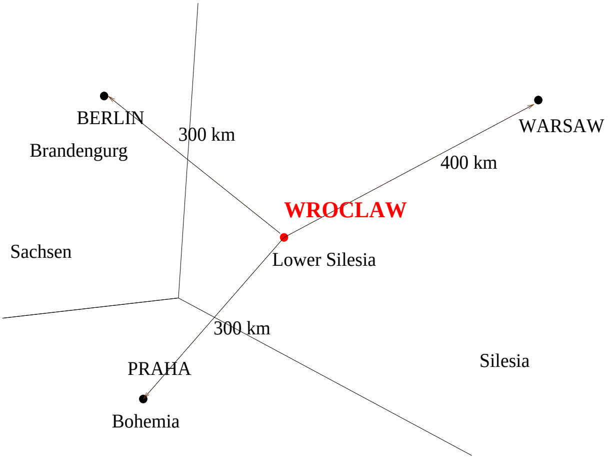
Anonymity and *k*-Choice Identities

Jacek Cichoń   Mirek Kutyłowski

Wrocław University of Technology
DELIS project

INSCRYPT'2007, Xining, China



BERLIN
Brandengurg
300 km
WARSAW
400 km
WROCLAW
Lower Silesia
Sachsen
300 km
PRAHA
Bohemia
Silesia

ID's and Privacy in Large Pervasive Systems

Pervasive Wireless Systems

Pervasive Systems

1. the number of mobile electronic artefacts increases, each artefact has its ID:
   - electronic tags for retail goods
   - electronic tags for library books
   - electronic tickets
   - electronic keys and personnel identification
   - ...
2. artefacts are carried by people
3. often anybody may read the artefact's ID in a wireless mode
   e.g. RFID technology

Pervasive Systems and Tracing

Tracing via Electronic Artefacts

1. tracing people by tracing their artefacts
2. cheap, easy and efficient
3. hard to prevent
4. hard to catch offenders tracing illegally

Privacy Problems

Possible Malicious Activities

1. violate personal privacy
2. derive consumer preferences in an unfair way
3. derive personal data on health condition (insuarance!)
4. unfair competition, business espionage
5. criminal and terrorist purposes

Legal Problems

Personal Data Protection Regulations

1. some countries impose strict rules on personal data protection
2. any data concerning **a person that can be identified** is **personal data**
   EU Directive
3. personal data protection obligatory,
   non-respecting is a crime, high penalties for system providers that do not fulfill data protection requirements

Legal Problems

Personal Data Protection Regulations

1. some countries impose strict rules on personal data protection
2. any data concerning **a person that can be identified** is **personal data**
   EU Directive
3. personal data protection obligatory,
   non-respecting is a crime, high penalties for system providers that do not fulfill data protection requirements

Practical consequences

Deploying useful systems can be blocked due to insufficient personal data protection offered by a current technologies.

Privacy Paradox

Two conflicting demands

1. an artefact **must show its ID** in most situations (e.g. a book returned to a library must show its ID upon arrival)
2. an artefact **must not show its ID** due to personal data protection

Some Consequences

privacy protection is the main usability problem of RFID technology in EU

example:

METRO company has withdrawn RFID tags from retail stores

Countermeasures

Solutions

killing **destroy RFID after use**
but then RFID's are not much useful

blocking **block RFID after use**
unblocking by legitimate readers only, but what a problem to capture a reader?

... **...**

Local Environments

Local Environments

Global Unique Identifiers

- ID collisions do not occur
- but no privacy

Local Environments

Global Unique Identifiers

- ID collisions do not occur
- but no privacy

Local Environments

- almost always a system has a limited scope and is relatively small
  (as in social networks)
- uniqueness required only within a local environment

Local Environments

Global Unique Identifiers

- ID collisions do not occur
- but no privacy

Local Environments

- almost always a system has a limited scope and is relatively small
  (as in social networks)
- uniqueness required only within a local environment

Idea

- uniqueness in local environments
- massive repetitions globally

Requirements

Basic Requirement

ID's can be set only at manufacturing time

Motivation

- resetting ID requires equipment
- possibility of resetting ID may be used for attacks, too
- sometimes there is a printed "hardcopy" of the ID -it is hard (or inconvenient) to change
  e.g. product tags in retail shops

k-choice IDs

Naive Solution - Random ID's

Algorithm

1. each artefact becomes an n-bit identifier chosen at random

Tradeoff

- long IDs make collisions in a small environment unlikely
- long IDs enable global tracing

Tradeoff for Random ID solution

Birthday Paradox

If a local environment has size 2^n and ID length is about $2n$, then a collision occurs with a fairly large probability.

Example

Tradeoff for Random ID solution

Birthday Paradox

If a local environment has size 2^n and ID length is about $2n$, then a collision occurs with a fairly large probability.

Example

- local environment of size $\approx 2^{10}$

Tradeoff for Random ID solution

Birthday Paradox

If a local environment has size 2^n and ID length is about $2n$, then a collision occurs with a fairly large probability.

Example

- local environment of size $\approx 2^{10}$
- required ID length $>$ 20 bits

Tradeoff for Random ID solution

Birthday Paradox

If a local environment has size 2^n and ID length is about $2n$, then a collision occurs with a fairly large probability.

Example

- local environment of size $\approx 2^{10}$
- required ID length > 20 bits
- many repetitions of a single ID occur provided that the global number of artefacts $\gg 2^{20}$

$\Rightarrow$ the method is useless for untracebility

Problem

Birthday Paradox

if a local environment as size 2^n and ID length is about $2n$, then a collision occurs with a fairly large probability.

Can we escape Birthday Paradox?

1. local environment should have appropriate size N (depending on application)
2. ID length should be not much higher than $\log N$, so many repetitions of the same ID occur globally
3. collisions should be unlikely

Algorithm description

Predistribution

A manufacturer preinstalls k (pseudo)random ID's in each artefact,
e.g.

$$H(K, T, 1), \ldots, H(K, T, k)$$

where

- K is the master key of the manufacturer
- T is a serial number
- H is a secure hash function truncated to the required length of the ID's.

Algorithm description

Registration in a local environment

When a new artefact arrives in a local system, then the system:

1. inspects which of the k ID's assigned to the artefact <u>has not been used yet</u> in this system,
2. and chooses one of them as the identifier of the artefact <u>for this</u> environment.

Collision-avoiding for 1-choice protocol

Theorem

Let B_N denotes the the first moment that a collision occurs in the random assignment process with N ID's. If $N \geq 20$ and $t \leq \sqrt{\frac{\pi N}{2}}$, then

$$1 - e^{\frac{t(t+1)}{2N}} \leq \Pr[B_N \leq t+1] \leq 1 - e^{\frac{t(t+1)}{2N}} + \frac{1}{\sqrt{N}} \; . \tag{1}$$

Collision-avoiding for 2-choice protocol

Theorem

Let C_N denote the first moment that a collision occurs for 2-choice protocol with N ID's during two-choices random assignment process. If $N \geq 5$ and $t < \sqrt[3]{3} \cdot \Gamma(\frac{4}{3}) \cdot N^{\frac{2}{3}}$, then

$$1 - e^{\frac{t(t+1)(2t+1)}{6N^2}} \leq \Pr[C_N \leq t+1] \leq 1 - e^{\frac{t(t+1)(2t+1)}{6N^2}} + \frac{1}{N^{\frac{2}{3}}} \ . \quad (2)$$

Collision-avoiding for 3-choice protocol

Theorem

... we get a similar expression:

$$1 - e^{\frac{t(t+1)(2t+1)}{6N^2}} \leq \Pr[C_N \leq t+1] \leq 1 - e^{\frac{t(t+1)(2t+1)}{6N^2}} + \frac{1}{N^{\frac{2}{3}}} \quad . \quad (3)$$

Corollaries

Let n be the length of ID's. The maximal size of local environment that still avoids collisions whp is

for 1-choice (random assignment): $\approx 2^{n/2}$

for 2-choice: $\approx 2^{2n/3}$

for t-choice: $\approx 2^{n \cdot t/(t+1)}$

Remark:
of course the local environment cannot have size $> 2^n$.

Applications

Municipal Ticket System

Period Tickets

1. each period ticket must have electronically readable ID
   goal: record the usage of the lines
2. it should not endanger passengers privacy

Example parameters for 2-choice

Table: N = minimal number of ID's , p = collision probability, t = number of passengers in a car

	$t = 50$	$t = 100$	$t = 200$	$t = 300$
$p = 10^{-2}$	2066	5802	16350	29999
$p = 10^{-3}$	6550	18389	51820	95081
$p = 10^{-4}$	20718	58166	163907	300742
$p = 10^{-5}$	65517	183942	518332	951052

Municipal Ticket System

Example parameters for 2-choice

Table: Estimated size of anonymity set for 1 million period tickets

	$t = 50$	$t = 100$	$t = 200$	$t = 300$
$p = 10^{-2}$	1936	689	244	133
$p = 10^{-3}$	610	217	77	42
$p = 10^{-4}$	193	68	24	13
$p = 10^{-5}$	61	21	7	4

Balancing Load in P2P Systems

Classical Approach

data M stored by a server at location indicated by $H(M)$, (H is a good hash function)

Improved Approach

allocate (very popular) data M at location indicated either by $H(M, 1)$ or by $H(M, 2)$

goal to achieve:
each server gets at most one heavy topic to serve

Conclusions

1. a simple, generic solution
2. but yet improves privacy a lot, if an adversary can trace only at some places and not all the time

Thanks for your attention!