High performance. Delivered.

Assessing Disclosure Risk in Anonymized Datasets

Michele Bezzi (ATL) & Alexei Kounine (EPFL)

Outline

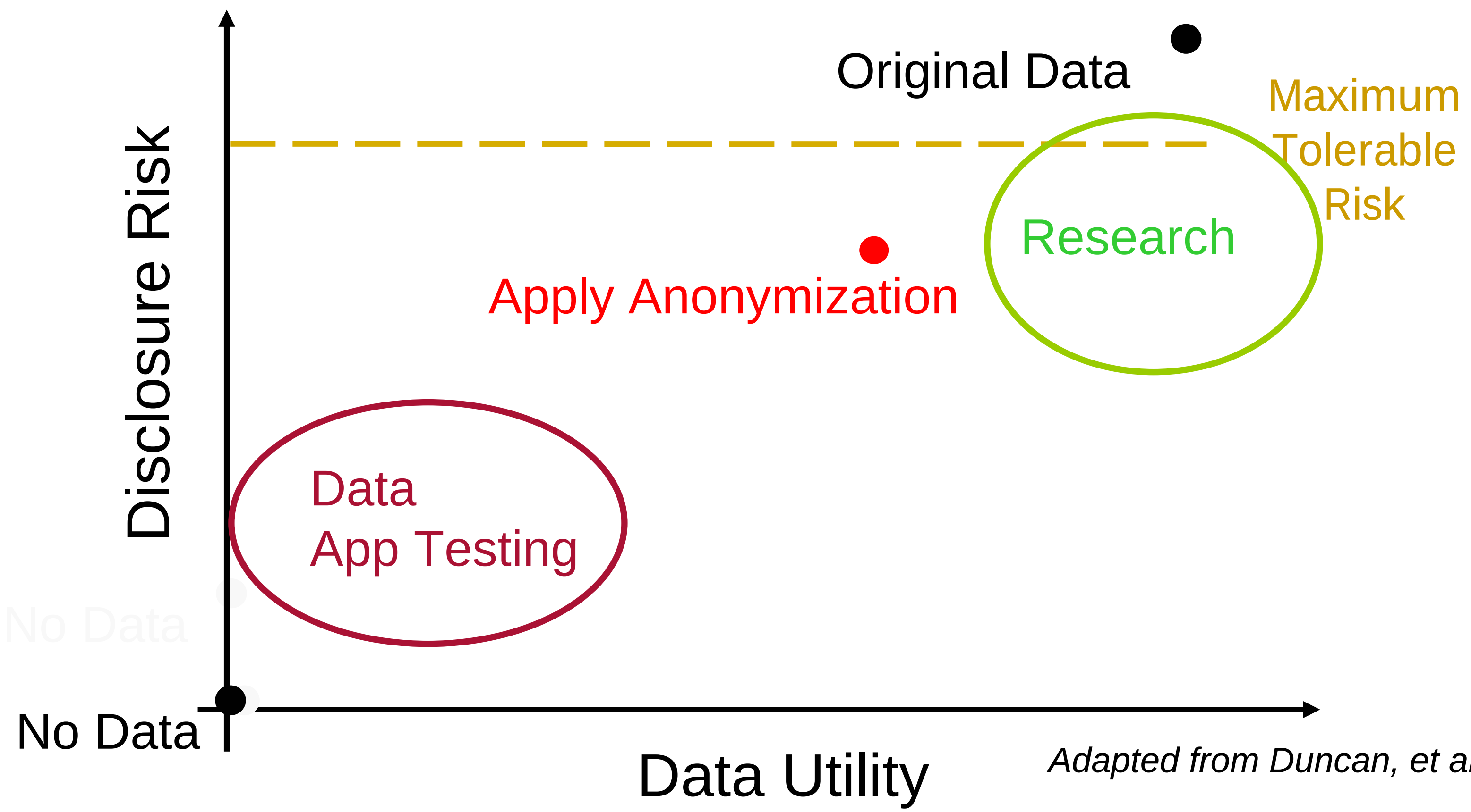
- Background & Motivation
- Anonymisation
- Disclosure Risk Estimation
  - Entropy measure
  - Properties
- Case Study: Flows
- Final remarks

Goal

- Problem:
  - The goal is to transform original data records so that no sensitive personal data are disclosed, whereas preserving the maximum amount of relevant information (*anonymity vs. utility* trade off), data integrity and consistency.
- Application
  - Creating datasets for application testing, whenever production DB contains sensitive data. (Our original goal)
  - Allowing researchers to share data and run analytical models on micro-data (e.g., log files), preserving privacy.

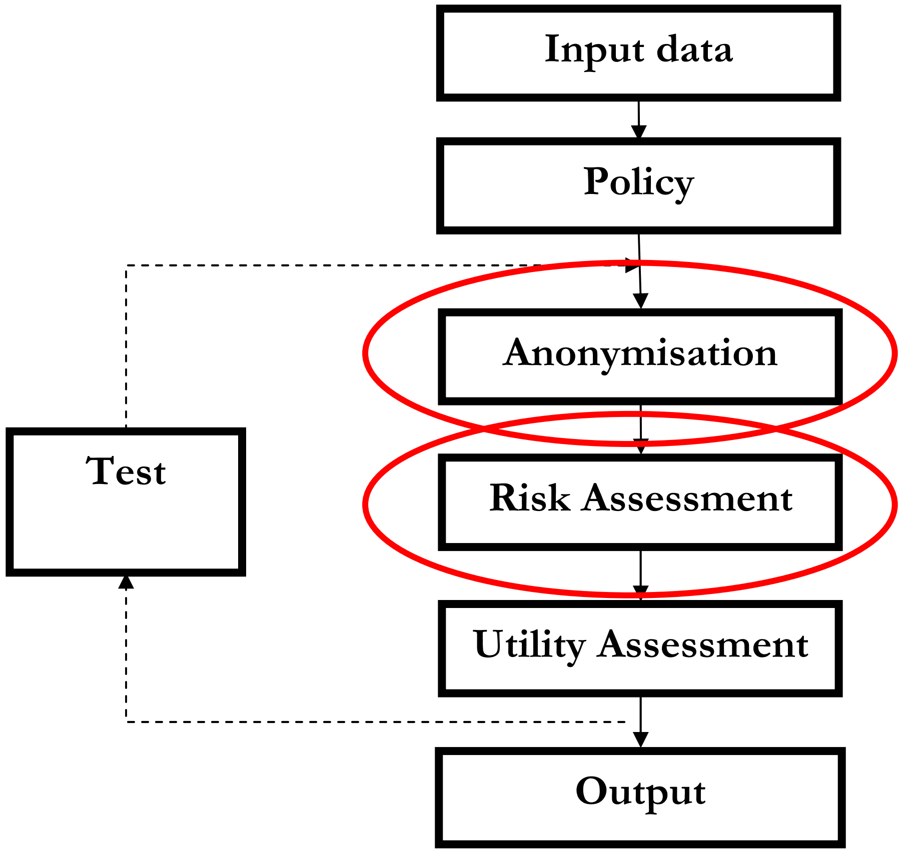
Goal

Risk-Utility Confidentiality Map



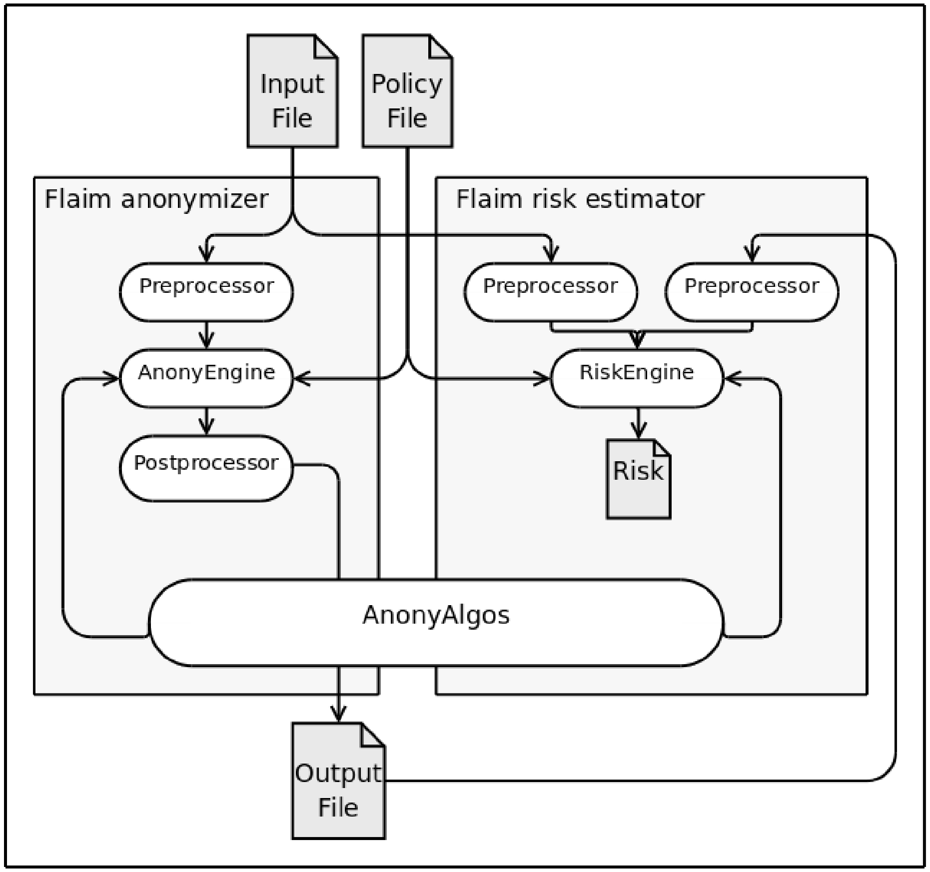
Adapted from Duncan, et al. 2001

Anonymisation engine & risk estimation



Implementation

- Using FLAIM (Framework for Log Anonymization and Information Management), developed by NCSA
- FLAIM anonymization engine (adapted) + risk module



Anonymisation primitives

IPs

- Black Marker (16 bits):
- Random Permutation (one-to-one mapping)
- Prefix-preserving (random permutation, but preserving structure)

IP Address	**Black Marker (16-bit)**	**Random Permutation**	**Prefix-preserving**
168.125.96.167	*168.125.0.0*	*124.12.132.37*	*12.131.102.67*
168.125.96.18	*168.125.0.0*	*231.45.36.167*	*12.131.102.17*
168.125.132.37	*168.125.0.0*	*12.72.8.5*	*12.131.201.29*

Port number

- Bilateral Classification: Replace with 0 or 65535 (the port smaller or larger than 1024): E.g., *27 -> 0 , 2048->65535*

Number of packets/bytes

- Add random noise (zero-average)
- Classification

Attack scenario

• The attacker aims at re-identifying released data by linking them with some background knowledge, which has some overlapping attributes with the released dataset.

• Estimating *P(r|s):* knowing data masking transformations, distance based similarity

• More uncertain mapping is - lower risk

• Because the data holder does not know in advance which records and attributes might be available to the attacker, it must run the risk analysis on the whole released dataset and assume a set of key attributes the attacker might know and use for re-identification.

Original data *S*

SrcIP	SrcPort	DestIP	DestPort	Packets
168.125.253.2	80	147.81.124.1	3157	40
39.109.219.43	7310	142.68.22.108	59959	126
35.187.130.82	161	213.48.19.68	22	83

Anonymised data *R*

SrcIP	SrcPort	DestIP	DestPort	Packets
168.125.253.0	1023	10.1.1.1	65535	42
39.109.219.0	65535	10.1.1.1	65535	132
35.187.130.0	1023	10.1.1.1	0	81

P(r|s)

P(r|s')

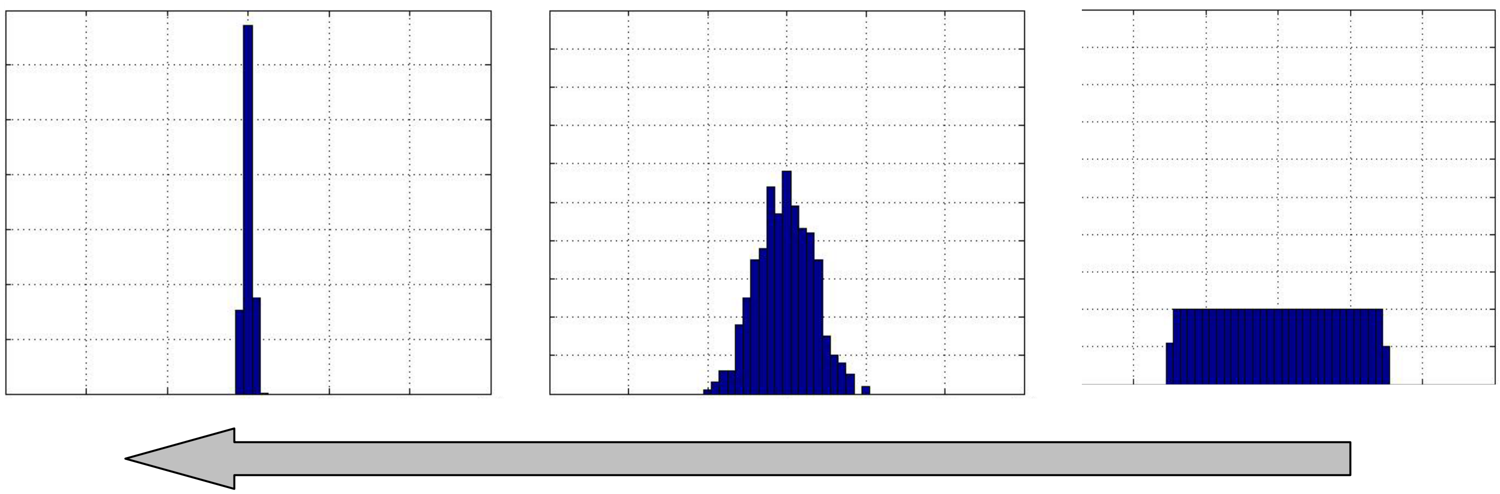
Original data *S*

SrcIP	SrcPort	DestIP	DestPort	Packets
168.125.253.2	80	147.81.124.1	3157	40
39.109.219.43	7310	142.68.22.108	59959	126
35.187.130.82	161	213.48.19.68	22	83

Background knowledge *S'*

SrcIP	SrcPort	DestIP	DestPort	Packets
39.109.219.43	7310	142.68.0.0		

Estimating risk



High Risk
of re-identification

Low Risk
of re-identification

H=1.2 *H=3.7* *H=4.9*

Shannon entropy: Average # of binary questions to identify *s*

Small: risky

Large: safe

Entropy as a risk measure

Shannon entropy: Average # of binary questions to identify a *single s*

$$H(\mathcal{R}|s) = -\sum_{r \in \mathcal{R}} P(r|s) \log_2 P(r|s)$$

Global risk:

Expected number of correct matches

$$E_{CM} = \sum_{s \in S} \frac{1}{2^{H(\mathcal{R}|s)}}$$

k-anonymity condition

Some properties

- Directly linked to information loss (utility):

$$I(\mathcal{S},\mathcal{R}) = H(\mathcal{R}) - \sum_{s\in\mathcal{S}} P(s)H(\mathcal{R}|s)$$

  – Minimal info loss:

$$\sum_{s\in S} P(s)\log_2 H(R|s) \qquad \text{with constraint } H(R|s) \geq h_{min}$$

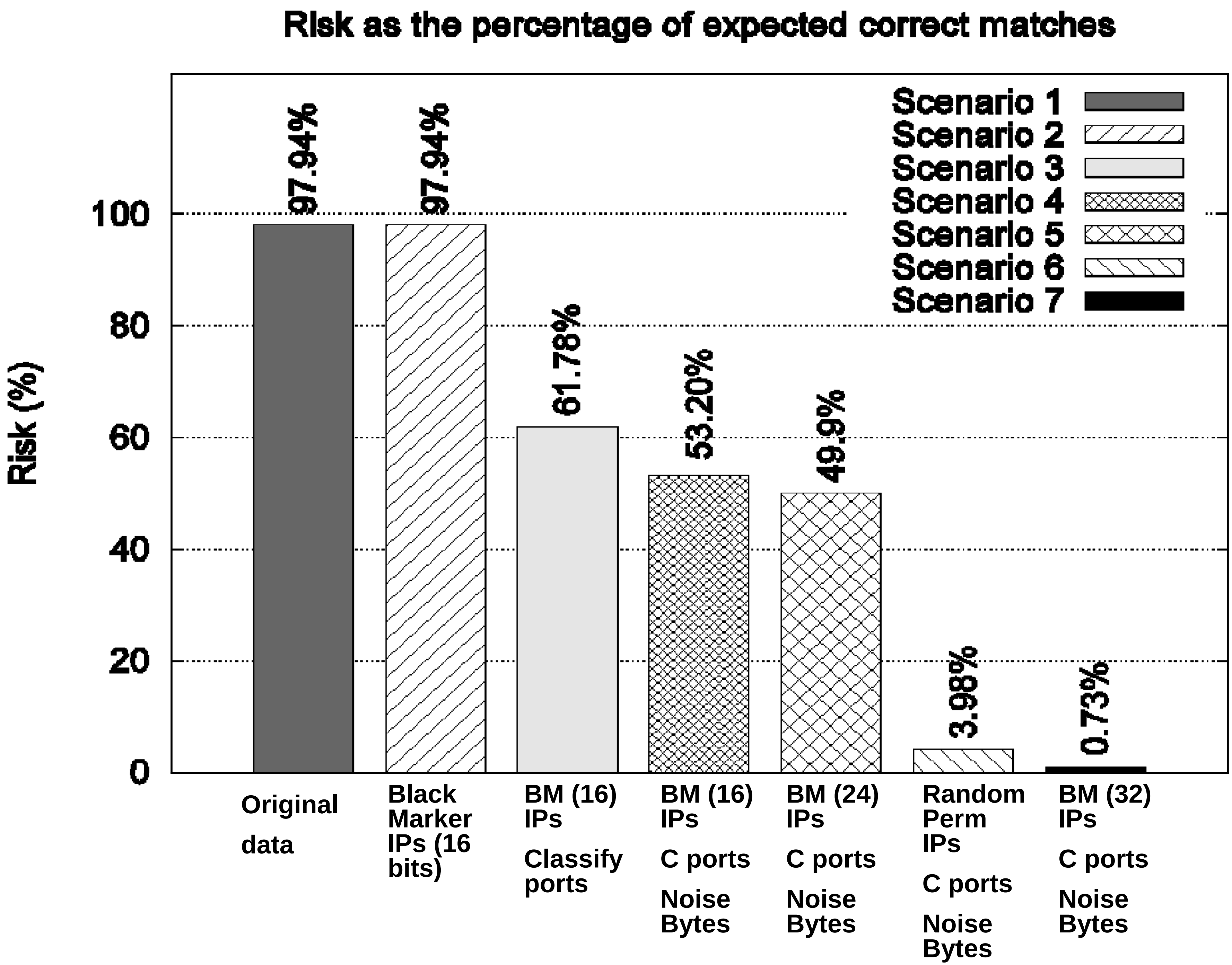
- Additivity

$$H(\mathcal{R}_1,\mathcal{R}_2|s) = H(\mathcal{R}_1|s) + H(\mathcal{R}_2|\mathcal{R}_1, s)$$

Case study: flow

- nfdump testing dataset provided by FLAIM group
- 10000 records
- Src/Dst IPs, Src/Dst ports, Bytes used

Date flow start	Duration	Proto	Src IP Addr:Port		Dst IP Addr:Port	Packets	Bytes	Flows
2007-09-15 21:09:11.401	0.392	TCP	93.82.215.84:36073	->	215.177.13.213:80	8	672	1
2007-09-15 21:09:12.491	0.000	UDP	57.28.244.23:48549	->	204.23.1.67:33467	1	40	1
2007-09-15 21:09:12.431	0.000	UDP	57.28.244.23:48549	->	204.23.1.67:33465	1	40	1
2007-09-15 21:09:12.356	0.354	TCP	89.240.246.94:60717	->	190.0.95.202:3128	7	1253	1
2007-09-15 21:09:12.127	0.000	UDP	154.159.232.119:56395	->	204.23.1.67:33524	1	40	1
2007-09-15 21:09:11.617	0.000	UDP	72.252.1.23:53	->	191.69.116.86:4489	1	165	1
2007-09-15 21:19:20.043	4294216.796	UDP	151.117.100.51:111	->	106.243.186.60:967	5	280	1
2007-09-15 21:19:21.348	1430.067	UDP	111.96.210.161:61718	->	70.114.202.209:161	2	154	1
2007-09-15 21:19:22.694	0.000	UDP	169.53.207.33:53	->	247.215.39.74:3337	1	329	1
2007-09-15 21:19:20.074	0.000	TCP	141.245.94.187:39414	->	217.242.169.109:479	1	60	1
2007-09-15 21:19:21.323	4293905.249	UDP	111.96.210.161:51937	->	80.187.116.29:161	2	154	1
2007-09-15 21:19:21.314	1388.111	UDP	111.96.210.161:53427	->	80.187.116.29:161	3	231	1
2007-09-15 21:19:19.139	0.000	UDP	169.53.207.33:53	->	99.74.24.233:51878	1	284	1
2007-09-15 21:19:19.321	0.000	UDP	169.53.207.33:53	->	99.74.24.233:51879	1	284	1
2007-09-15 21:19:21.321	0.000	UDP	111.96.210.161:53877	->	80.187.116.29:161	1	77	1
2007-09-15 21:19:26.305	4294392.436	UDP	169.53.207.33:53	->	98.14.24.3:50999	2	348	1
2007-09-15 21:19:15.297	69.143	TCP	121.191.230.139:25	->	135.219.55.50:1674	4	291	1
2007-09-15 21:19:21.375	5.023	TCP	103.6.42.145:20144	->	88.118.84.209:51024	552	28712	1


Risk as the percentage of expected correct matches
Risk (%)
0
20
40
60
80
100
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5
Scenario 6
Scenario 7
97.94%
97.94%
61.78%
53.20%
49.9%
3.98%
0.73%
Original data
Black Marker IPs (16 bits)
BM (16) IPs Classify ports
BM (16) IPs C ports Noise Bytes
BM (24) IPs C ports Noise Bytes
Random Perm IPs C ports Noise Bytes
BM (32) IPs C ports Noise Bytes

Final remarks

Quantifying disclosure risk is essential for finding the optimal trade-off between privacy and utility.

Measure disclosure risk using entropy:

- General: applicable to any anonymization algorithm (unlike k-anonymity)
- Stable: depends on shape of the distribution
- Linked to Information Theory

Future works (a lot…):

- More realistic testing (larger dataset, correlation across fields/records)
- Utility, Optimisation, **…**

Thanks for the attention

Michele.bezzi@accenture.com