DATI, AI E ROBOTICA @POLITO

RICERCA, TRASFERIMENTO TECNOLOGICO E SUPPORTO ALLE AZIENDE SUI TEMI FONDAMENTALI DEI BIG DATA, INTELLIGENZA ARTIFICIALE, ROBOTICA E RIVOLUZIONE DIGITALE

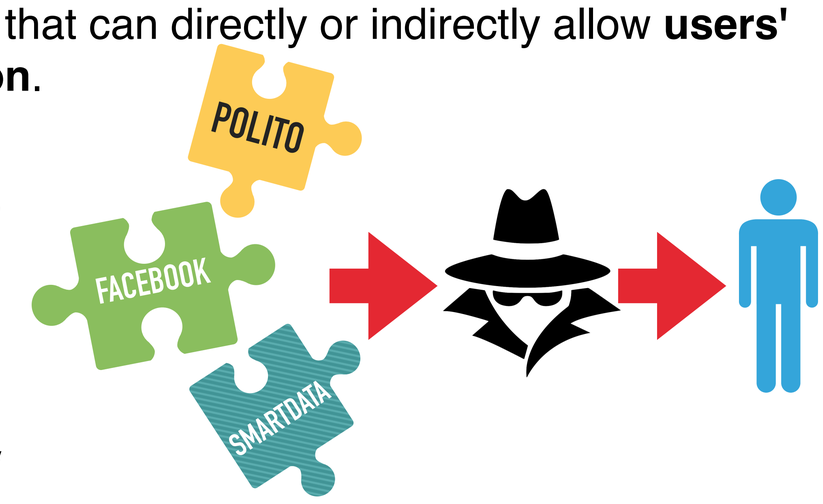
z-Anonymity & α-MON: how to protect users' privacy

Authors: Thomas Favale, Nikhil Jha

SCAN ME

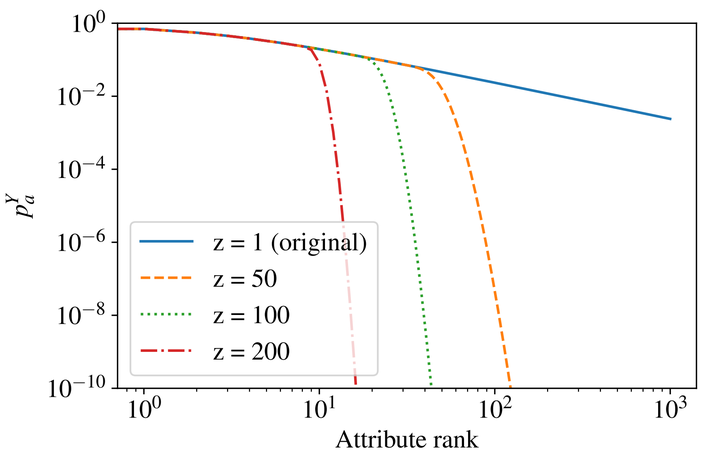
Motivation and background

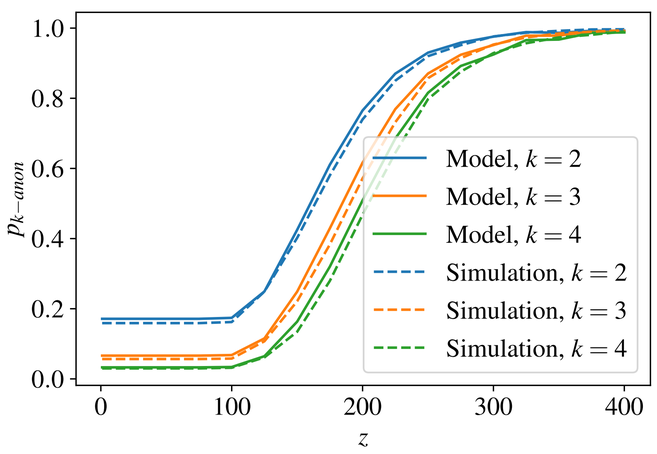
- With the advent of **big data** and the birth of the d**ata markets** that sell personal information, **individuals' privacy** is of utmost importance.
- The classical response is **anonymization**, i.e., sanitizing the information that can directly or indirectly allow **users' re-identification**.
- Network measurements must be performed with care to **avoid threatening users' privacy**



The theory behind

- The most famous **anonymization** paradigm is **k-anonymity**, which aims at assuring that every user in the dataset has other **k-1 users with identical** quasi-identifying attributes, by generalizing/suppressing the data
- **Classic** anonymization paradigms **suffer the stream** scenario, hence the need of a ***zero-delay* anonymization technique**
- z**-anonymity** tends to **protect rarest attributes** reducing their **probability** of being published when occurring (p_a^Y), that could easily bring to user re-identification



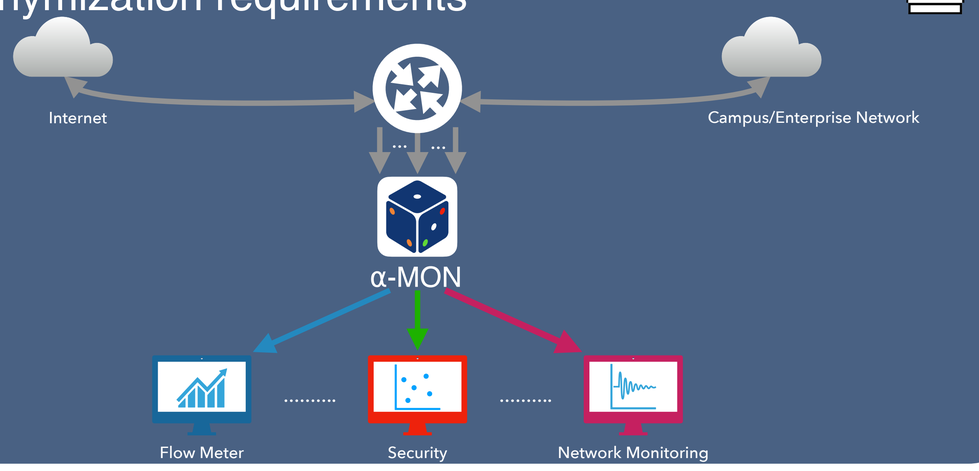


- Applying z**-anonymity** allows to **k-anonymize** a user with a **given probability** p_{k-anon}, which can be evaluated analytically

The practice

- Starting from what we learned from z**-anonymity**, we have created a **high-performance tool** that implements it: α**-MON**

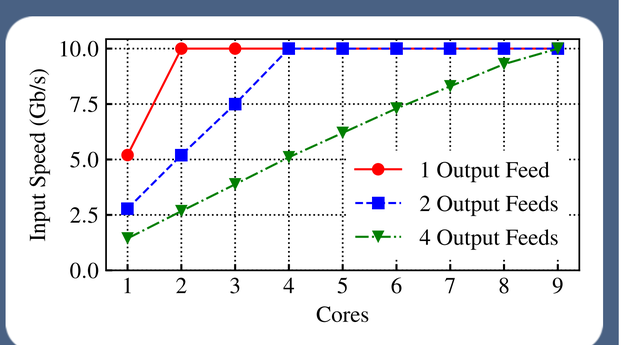
- **Supports** z-anonymity to **hide** private **quasi-identifiers** with custom z and ΔT
- **Flexible** set of anonymization policies
- **Scalable** and deployable in high-speed links
- Support multiple **legacy applications** with different anonymization requirements



Is it feasible?

- α-MON with a **single core** can sustain **10 Gbit/s** with a single consumer
- The performance is reduced when α-MON has to feed **multiple consumers**

- The **throughput scales linearly** with the number of cores in all cases



Yes, it is feasible!

Conclusions and future work

- A **zero-delay algorithm** to reduce the risk of re-identification on a stream of data
- The only *zero-delay* algorithm which does **not require the injection** of false data to protect users' privacy
- **Many possible improvements**:
  - auto-adjustment of z
  - generalizing private data instead of suppressing