Telco Data Anonymization

Introduction

Anonymization is the process of protecting private or sensitive information by erasing or encrypting identifiers that connect an ‘entity’ to stored data. If the anonymization is fool-proof, the process of Deanonymization should not reveal personal identifiable information. Typically, anonymization could be just patt ern anonymization, or just the value anonymization, or both. In this work, we want to do just the value anonymization, so as to preserve the predictive /detective power. Just like any other data, even in Telco data, we will have to deal with both the categorical Variables and the numerical variables. There are various approaches under anonymization:

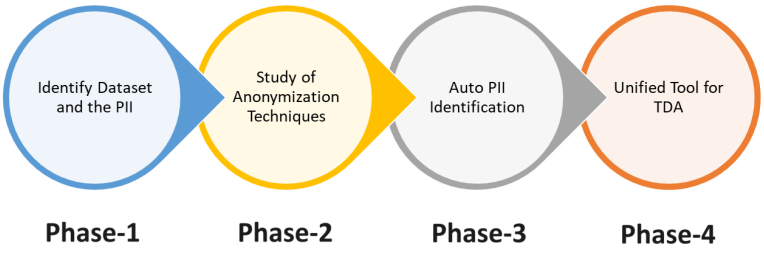
- Suppression
- Masking
- Pseudonymization
- Generalization
- Data Swapping
- Data Perturbation
- Synthetic Data

Typically, Synthetic data generation is considered as the most fool-proof among all the other approaches.

In this work, we want to answer the following questions:

1. Do we have the datasets, which consists of the all the sensitive information?
    a. Well used, freely available, significant size, etc.
2. What kind of sensitive information is well suited for each of the different techniques?
    a. Mapping of a type of sensitive information to a technique.
3. Is there a single technique that is applicable to all kinds of sensitive information?
4. Can we automatically identify the PII in the given dataset?
5. Can we build a tool that takes in the dataset and anonymizes it automatically (with no manual intervention) using the best technique ?

Phases



Phase-1

In this phase, we want to (a) agree on what constitute the ‘sensitive’ (PII) data. (b) agree on the exact dataset to use. (c) find the 'gaps' in datasets and try to fill it somehow.

What constitutes the PII information:

- Names (Systems, Domain, Individuals, Organizations, Places, etc.)
- Address (IP and MAC)
- Telco Fields - IMSI, IMEI, MSIN, MSISDN, MCC+MNC
- Location Data (Cell-ID, Count, etc.). GPS Data on its own is not a sensitive information. The context around that, such as 'names', are sensitive.

PII Type	Dataset (links)
Names (Systems, Domain, Individuals, Organizations, Places, etc.)	ServerLOG1, ServerLog2, LogHub, Campus
Address (IP and MAC)	Internet Traffic Dataset: EX1, EX2
Telco Fields - IMSI, IMEI, MSIN, MSISDN, MCC+MNC	Adult Dataset enhanced with Telco-Fields Adult Dataset: Generate random IMEI/IMSI* fields and add it to this dataset
Location Data (GPS, Cell-ID, Count, etc.)	OpenCellID, GPS, IEEE-dataport (crawdad)

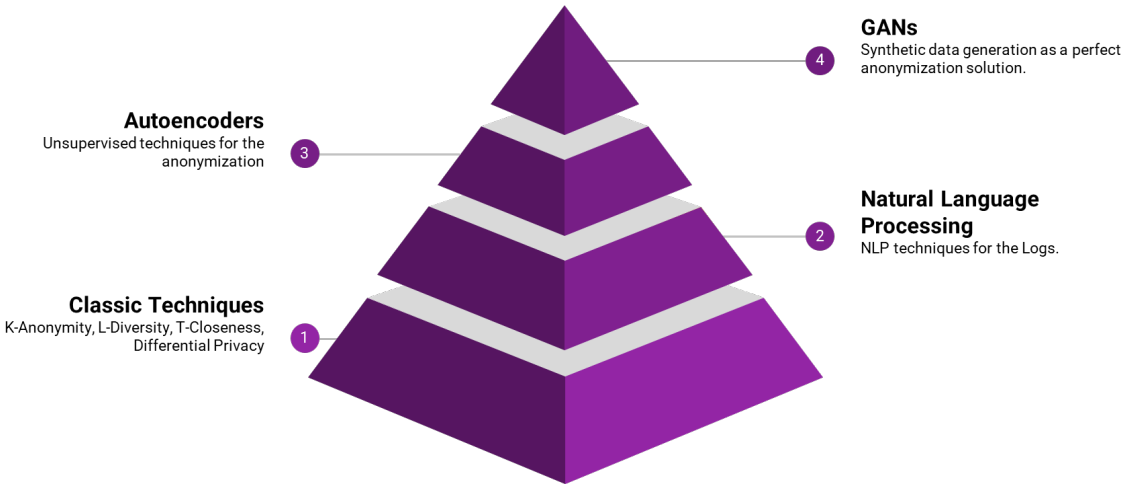
Phase-2

In this phase we would want to:

1. Try available tools (Libraries) and techniques (implementations) on the chosen datasets.
2. Find the gaps in tools and techniques.
3. Fill those gaps considering the

we will be evaluating the following techniques

- Classic (and its variations): K-Anonymity, L-Diversity, T-Closeness, Differential Privacy
- Data Anonymization with Autoencoders
- NLP approaches for data anonymization
- Generative AI (GANs)



Anonymizing Names and Telco-Fields

We have found that the classic-techniques do well when it comes to anonymizing both Names and telco-fields (Nouns and Numbers) - when it is in a structured (columns) format.

In this repo, you can find the techniques that we have tried for these fields: https://github.com/sknrao/anonymization

Anonymizing Packet Fields

Anonymizing the packet fields is a very well researched area. Works are available from early 2000. The most recent ones are using condensation-based differential privacy.

References:

- RFC6235: IP Flow Anonymization Support. https://www.rfc-editor.org/rfc/rfc6235.txt
- PCAPLIB : Y.-D. Lin, P.-C. Lin, S.-H. Wang, I.-W. Chen, and Y.-C. Lai, "Pcaplib: A System of Extracting, Classifying, and Anonymizing Real Packet Traces," IEEE Systems Journal, vol. 10, no. 2, pp. 520-531, 2014.
- CRYPTOPAN : J. Fan, J. Xu, M. H. Ammar, and S. B. Moon, "Prefix-Preserving Ip Address Anonymization: Measurement-Based Security Evaluation and a New Cryptography-Based Scheme," Computer Networks, vol. 46, no. 2, pp. 253-272, 2004.
  - Newer Version: https://ant.isi.edu/software/cryptopANT/index.html
  - Using with Python: https://github.com/certtools/cryptopanlib
- TCPANON : F. Gringoli. (2009, 11/10/2020). Tcpanon. Available: http://netweb.ing.unibs.it/~ntw/tools/tcpanon/
- SCRUB-TCPDUMP: D. Koukis, S. Antonatos, D. Antoniades, E. P. Markatos, and P. Trimintzios, "A Generic Anonymization Framework for Network Traffic," in 2006 IEEE International Conference on Communications, 2006, pp. 2302-2309

- TRACEWRANGLER: J. Bongertz. (2013). Sec-4 Trace File Sanitization, the Sharkfest Challenge. Available: https://sharkfestus.wireshark.org/sharkfest.13/presentations/SEC-04_Trace-File-Sanitization-NG_Jasper-Bongertz.pdf
- PKTANON : https://github.com/KIT-Telematics/pktanon

Currently the team is working on

(a) implementing the condensation-based differential privacy.

(b) Developing containers to test and evaluate the above techniques.

Anonymizing location information (cell-ID, count, etc.).

We are currently working on this and exploring different techniques.

Anonymizing Log-Data.

The team is currently exploring use of NLP for this. Once there is a progress, we will update this section.

Phase-3

The team is currently working on building a tool that auto-detects of the PII data to picks the best technique to use on the data.

Phase-4

The team is currently building a container-based architecture for a unified tool.