An Analysis Of Protected Health Information Leakage In Deep-Learning Based De-Identification Algorithms

Salman Seyedi[1,*], Li Xiong[2], Shamim Nemati[3], Gari Clifford[1,4]
[1]Department of Biomedical Informatics, Emory University, Atlanta, GA, [2]Department of Computer Science, Emory University, Atlanta, GA,
[3]Department of Biomedical Informatics, University of California San Diego Health, La Jolla, CA, [4]Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA
*sseyedi@dbmi.emory.edu
https://arxiv.org/abs/2101.12099

Goal

Investigating potential leakage of sensitive information from a de-identification algorithm

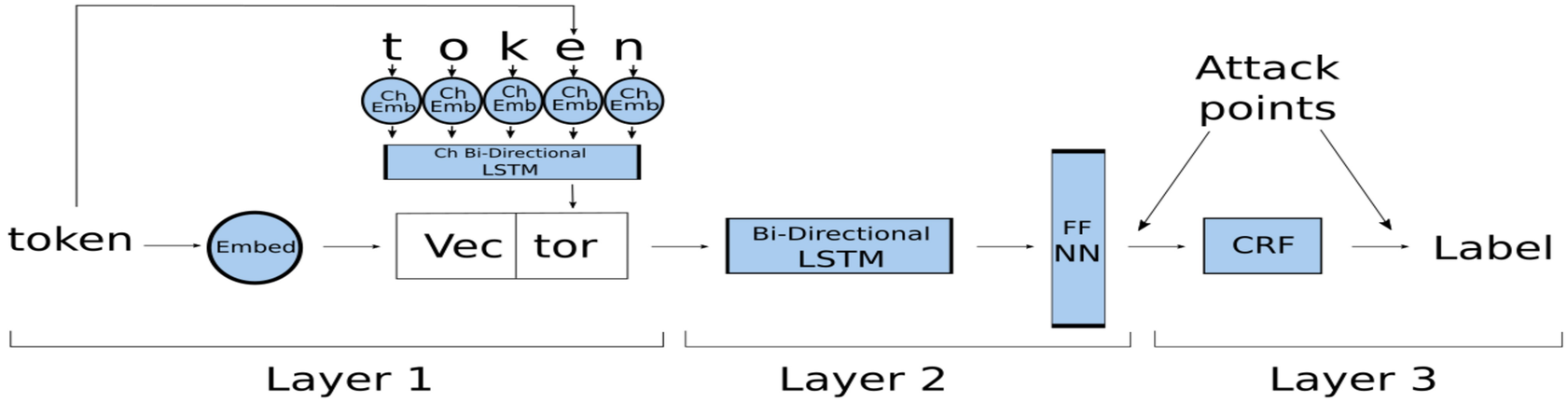
Investigated Model (NeuroNER)

NeuroNER
(State of the art De-identification Model)
Layer 1: Text tokenizer
Layer 2: Neural Networks
Layer 3: Conditional Random Field

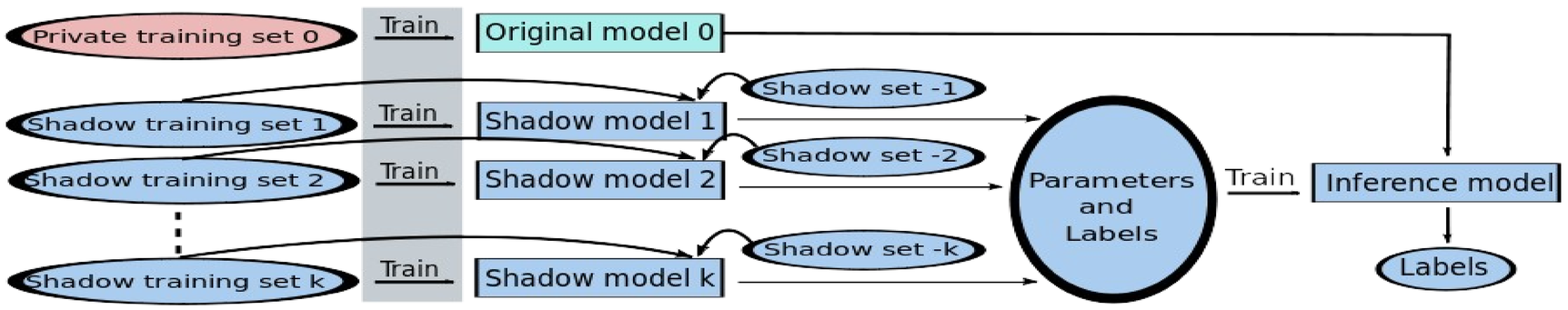
Re-identification Attacks

White Box Attempts:
- Naive cut-off
- Brute-force cut-off
- Membership inference attack

De-Identification Deep Neural Network Model (NeuroNER) And Investigated Attack Points



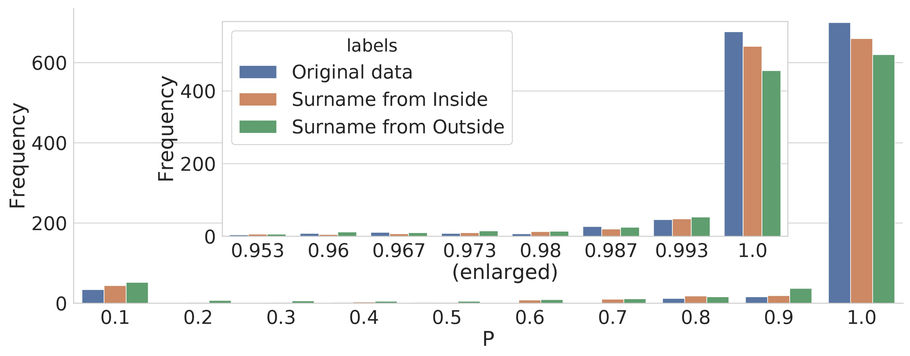
Membership Inference Attack



Results and Conclusions

- Despite different distributions, zero successful re-identification
- Model not prone to several implemented attacks
- Statistically different distributions but with overwhelming overlap for successful cut-off attacks

Histogram of Probabilities for Surnames



Acknowledgments

National Science Foundation, grant # 1822378
`Leveraging Heterogeneous Data Across International Borders in a Privacy Preserving Manner for Clinical Deep Learning',
The National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378.

NSF

NIH National Center for Advancing Translational Sciences