PWSCUP-2021

Abyssec (静岡大学 大木研究室/西垣研究室)

匿名加工フェーズ

本選の方針

方針①

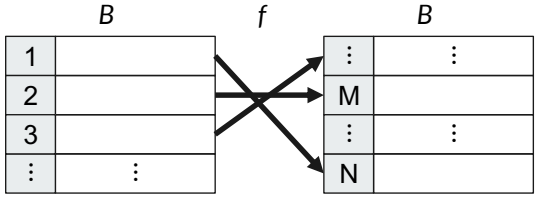
- 識別フェーズにてすべての行に対して適当な行番号（-1以外）を出力すると、安全性指標のうちPrecisionは1.0+ε、Recallは0.5+εを叩き出すことができてしまう
- すなわち、残りのTop-kが勝敗を左右するので極限までTop-kを０に近づけるべきである

方針②

- BからCを作成し、そこからさらにDを作成するのが大変
  - 第１匿名化したものから第２匿名化を行う必要があるので、有用性を満たすDを作成することが難しい
- 先に有用性を満たすDを作成し、それをもとにCを作成すれば簡単に有用性をクリアできる？

方針①の実装

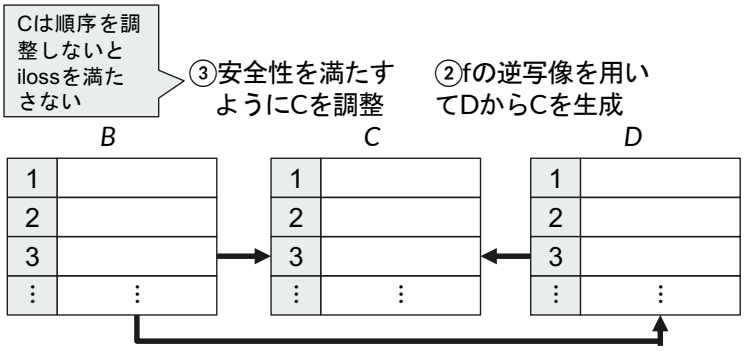
- 攻撃者はTop-kの推定にレコードリンケージ（RL）攻撃を用いると想定
- BとDの対応するレコードについて、なるべく距離が遠くなるように加工する必要がある
- ノイズ付与等をしなくても、写像fを用いてCからDを作成すると、RL攻撃のTop-k推定精度を低く抑えることができる



距離の遠いレコード同士の対応付け

方針②の実装

- BからCを作ってCからDを作るのは大変
- それならば、Bから先にDを作成し、BとDから安全性&有用性を満たすCを作成



識別フェーズ

予備選の方針

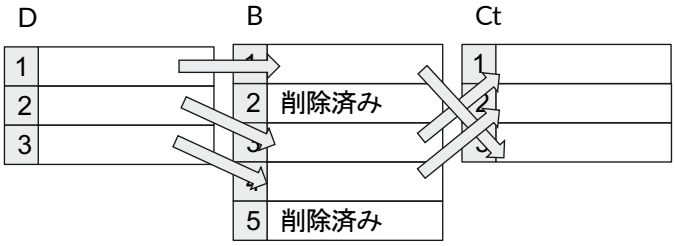
Dの順番がシャッフルされてない

⬇

B↔Dで照合が行える

⬇

D→B→Ctの順番でDとCtの行が対応付けられる



具体的なB↔Dの照合方法

1. Bとのiloss制限を満たすDの候補を列挙
2. BとDの距離上位1000件のインデックスを列挙
3. 1と2のインデックスのANDをとる
4. Dのn行目について、Bのn-1行目以下の候補を削除
5. 上下のインデックス候補との関係を考慮して選定

本戦の方針

Bが使えず、DとCtのみからPrecを高めるのは困難

⬇

ilossを満たすCtの全ての行についてtop3を予測

今回の制限下ではDからBを十分に離すことが可能

⬇

ほとんどのチームが距離を十分に離してくると予想

⬇

Ctの各行とilossを満たす中で最も距離が遠いDの20行の中から乱択でtop3を選択