Applying Differential Privacy to Text

Besat Kassaie
Supervised by: Frank Tompa
The Cheriton School of Computer Science, University of Waterloo

Research Problem

Differential Privacy (DP) has been shown to have desirable properties such as offering privacy quantification, being independent of an adversary's background knowledge, and providing an interpretable definition of privacy. Applying DP in medical domains entails many challenges resulting from the importance of data utility in medical domain, correlations between data items that should be preserved, finding and justifying the parameter values such as ϵ (Dankar & Emam, 2012), and dealing with unstructured data items. In this work we propose a solution to apply an appropriate variant of DP to medical text.

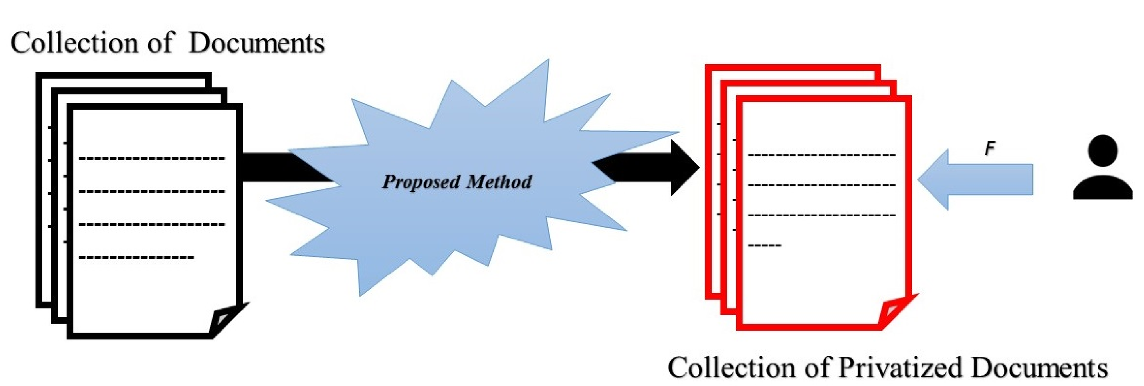

Figure 1:Researchers have a class of computations, F. We generate a privatized version of documents to compute F

Differential Privacy

A randomized algorithm M is ϵ-differentially private if for all $S \subset Range(M)$ and for all $x, y \in domain(M)$ such that $|x - y| \leq 1$: $Pr[M(x) \in S] \leq exp(\epsilon) Pr[M(y) \in S]$.
DP is a property of data access mechanisms that guarantees *indistinguishability*, i.e., expecting almost the same outputs on similar inputs.

Information Extraction

We should deal with the inevitable chaos in text to benefit from it. Utilizing *Information Extraction* techniques is a standard approach to make text machine-friendly. Information Extraction refers to the automatic extraction of structured information such as entities and relationships between entities from unstructured sources (Sarawagi, 2008).

Proposed solution

A step forward to solve the problem is to assume that researchers' information needs can be satisfied using structured records extracted from the documents. With this assumption, the problem can be illustrated as in Figure 2. We generate privatized documents in such a way that running the same I_E over them will result in the same private view V' which can be generated using extracted view V.

Domain-Preserving Functions

Let F be a set of domain-preserving functions, $F = \{f_i | f_i : W_i \rightarrow W_i, Domain(f_i(v_i)) = Domain(v_i)\}$. Each attribute A_i is associated with a function $f_i \in F$. Let the privatization function be domain-preserving, such that $r =< v_1, v_2, ..., v_{\mathcal{T}} >$ and $r'(j) =< v'_1, v'_2, ..., v'_{\mathcal{T}} >$ where:

$$v'_k = \begin{cases} f_k(v_k), & \text{if } k = j. \\ v_k, & \text{otherwise.} \end{cases} \quad (1)$$

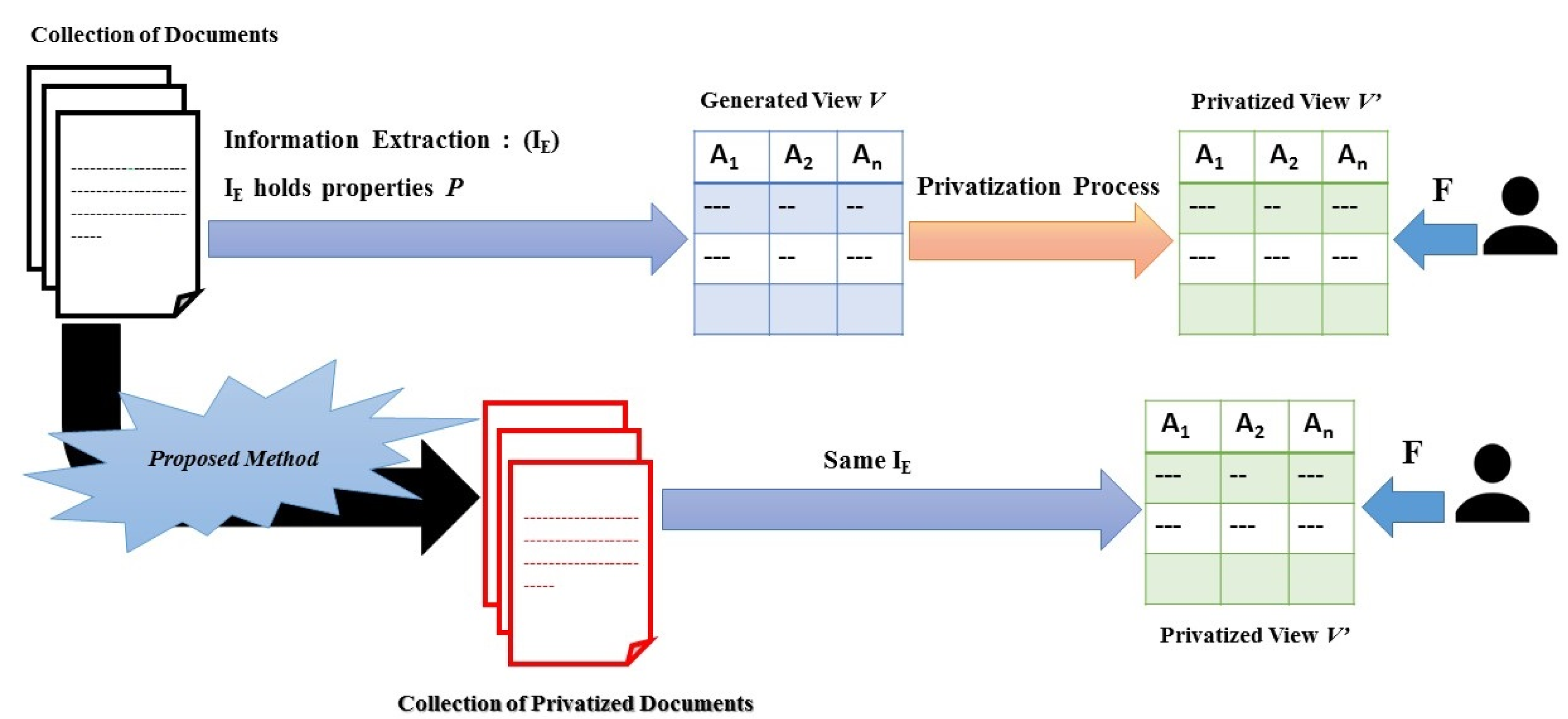

Figure 2:The proposed solution for generating privatized documents.

Strict Extractor

An IE algorithm is *Strict* if the set of extracted values in a record is a subset of words appearing in the corresponding document, $\{v_1, v_2, ..., v_{\mathcal{T}}\} \subseteq \{w_1, w_2, w_3, ..., w_{\mathcal{N}}\}$. Let $P_D(j) \subseteq \{p | w_p = v_j\}$, i.e., a subset of positions in $D =< w_1, w_2, ..., w_{\mathcal{N}} >$ where $w_p = v_j$ (the position(s) from which v_j is extracted).

Computable Extractor

An IE algorithm is *Computable* if for all $j, j' \in [1 \ldots \mathcal{T}]$

$$if \begin{cases} P_D(j) \text{ is explicit(given)} \\ P_D(j) \text{ and } P_D(j') \text{ are pairwise disjoint.} \end{cases} \quad (2)$$

Stable Extractor

Let $g(D, j) =< w'_1, w'_2, w'_3, ..., w'_{\mathcal{N}} >$ where:

$$w'_k = \begin{cases} f_j(w_k), & \text{if } k \in P_D(j) \ . \\ w_k, & \text{otherwise.} \end{cases} \quad (3)$$

An IE algorithm is *Stable* if $\forall j \in [1 \ldots \mathcal{T}]$ $P_D(j) = P_{g(D,j)}(j)$ and $I_E(g(D, j)) = r'(j)$.

Theorem

For any function $I_E : \mathcal{D} \rightarrow R$ having the aforementioned properties, there exists an algorithm $A(F, P_D(j))$ such that for an arbitrary set of functions $F = \{f_i | f_i : W_i \rightarrow W_i, \ i \in [1 \ldots \mathcal{T}]\}$ and any document $D \in \mathcal{D}$, $A(F, P_D(j))$ produces $D_F^{\mathcal{P}}$ in such way that, $F(I_E(D)) = I_E(D_F^{\mathcal{P}})$.

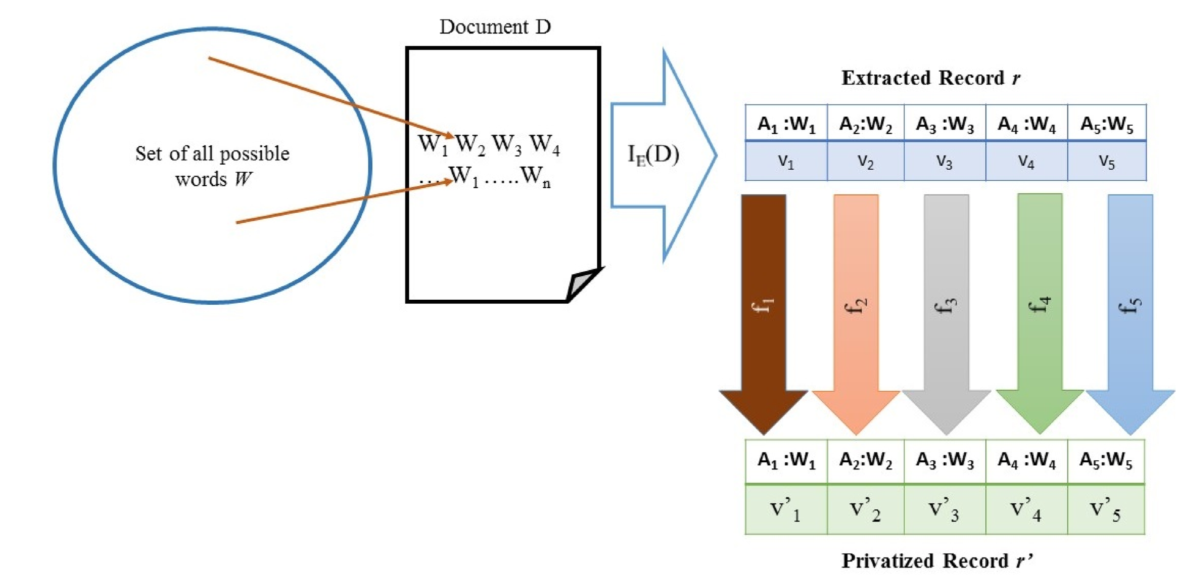

Figure 3:$I_E(D)$ extracts a record r. Then a privatized record r' is generated.

Claim

For any function I_E having the aforementioned properties, algorithm 1 produces $D_F^{\mathcal{P}}$ in such a way that $F(I_E(D)) = I_E(D_F^{\mathcal{P}})$.

Algorithm 1 PrivateGen

Input: $F, \{P_D(j) | j \in [1 \ldots \mathcal{T}]\}$
Output: $D_F^{\mathcal{P}}$

```
1: for j ∈ [1...T] do
2:   for every i in P_D(j) do
3:     substitute w_i ∈ D with f_j(w_i)
4:   end for
5: end for
```

References

Dankar, F. K., & Emam, K. E. (2012). The application of differential privacy to health data. In *Proceedings of the joint EDBT/ICDT workshops*.
Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*.
Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases*, *1*(3).

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science