Information Processing Letters 128 (2017) 38–45

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Privacy risks ensuing from cross-matching among databases: A case study for soft biometrics

Debanjan Sadhya [a,*], Sanjay Kumar Singh [b]

[a] Indian Institute of Technology Roorkee, Roorkee, India
[b] Indian Institute of Technology (Banaras Hindu University), Varanasi, India

ARTICLE INFO

Article history:
Received 11 September 2016
Received in revised form 10 July 2017
Accepted 6 August 2017
Available online 10 August 2017
Communicated by L. Viganò

Keywords:
Soft biometrics
Databases
Information theory
Privacy

ABSTRACT

In this digital era, it is a very common practice for individual users to submit their data in multiple databases. However, the existence of correlated information in between these databases is a major source of privacy risk for the database respondents. In our study, we investigate such situations regarding soft biometric databases. A majority of modern biometric recognition systems utilize soft biometric traits in concurrence with primary biometric features due to the multiple gains incurred in the overall performance of the systems. In our work, a theoretical model has been developed which captures the notion of the user's privacy in the case of a soft biometric database leakage. In a broader sense, our work proposes a framework which quantifies the privacy levels of individuals supposing some form of correlation based attack has been successfully executed by an adversary. The modeling process itself is based upon elements of information theory such as conditional entropy (equivocation) and mutual information.

1. Introduction

The utilization of biometric systems for human authentication has been gradually getting proliferated. A rudimentary biometric system essentially depends on some biometric traits, which in turn must possess certain characteristics. These properties primarily include universality, distinctiveness, permanence, and collectability. However, it has been observed and reported that individually, a single biometric trait cannot satisfy all the required characteristics mentioned previously [1]. To overcome these and more problems, the concept of multimodal biometrics was introduced. A multimodal biometrics system utilizes more than one biometric trait for recognition and verification purposes. Although the application of this novel idea mitigated some difficulties associated with the unimodal systems, some new problems came to the forefront. The *cost* and *verification time* greatly increased since the multimodal framework utilized more than one modality. The novel idea of using *soft* biometric traits was subsequently initiated for exploiting the advantages of both the unimodal and multimodal systems. These new ancillary data are those properties which furnish a limited amount of information about an individual but lack the distinctiveness and permanence to adequately differentiate any two. Typical examples of these traits include height, weight, skin color, eye color, age, ethnicity, gender etc. On their own, these properties are not sufficient enough to accurately distinguish between a genuine user and an impostor. However when they are used in combination with a primary biometric trait (fingerprint, face etc.) in a multimodal environment, the overall performance greatly improves.

A common feature in most biometric systems is the requirement of a biometric database. This database is used for storing the biometric data of every successfully enrolled user. In a soft biometric based fusion framework, the

[*] Corresponding author.
E-mail address: debanjan.sadhya@gmail.com (D. Sadhya).

http://dx.doi.org/10.1016/j.ipl.2017.08.001
0020-0190/© 2017 Elsevier B.V. All rights reserved.

soft features of the users get stored alongside the primary ones in the same database. In our study, we consider the privacy risks which could potentially arise on the leakage of such information. Intuitively it can be understood that the privacy issues of the users emerge due to the high degree of correlation that exists between their soft biometric information and other external databases. This ultimately increases the net amount of publicly available information associated with the users.[1]

In our work, we have tried to theoretically quantify the privacy levels (or alternatively the loss in privacy) on the leakage of a soft biometric database. The principal factor that increases the privacy risks associated with such databases is the various linkages (attribute wise) which exist between soft biometric and micro databases. These links assist an adversary in performing various cross-matching or correlation based attacks in between the databases. The privacy risks further increase when the adversary possesses some auxiliary background information about either any particular targeted user or about the entire database as a whole.

To summarize, our work attempts to quantify a user's privacy guarantees by providing a theoretical framework under the various attack scenarios possible in case of a soft biometric database leakage. As per knowledge of the authors, there exists no work in the literature which attempts such an evaluation. In a sense, our study can be generalized as the assessment of privacy risks due to the presence of correlation among two databases. Our study is motivated by the pioneering work of [2], in which the authors gave an information theoretical analysis of the tradeoff between utility and privacy in micro databases.

2. Background concepts

This section briefly introduces some core concepts which are related to our work.

2.1. Soft biometrics

As briefly mentioned in the previous section, soft biometrics traits are those attributes which provide some information about an individual, but are not able to separately authenticate the person. This problem arises mainly due to the lack of distinctiveness and permanence in the traits themselves. However using these ancillary properties have some advantages of their own, the primary ones being reduced cost and effort in computation, sensing at a distance and pruning the search space prior to primary biometric based matching stages (thus reducing the operation time of the recognition system). Soft biometric traits are classified on the basis of *nature of value*, *permanence*, *distinctiveness* and *subject perception*. The first property (i.e. nature of value) refers to the fact that whether any particular trait is continuous or discreet. Furthermore, permanence indicates the ability of the trait to remain invariant with time, distinctiveness signifies the degree of variation of the trait between subjects and subject perception points to the ability of humans to unambiguously recognize the specific trait.

A majority of studies related to soft biometrics focus on innovative applications involving computer vision and conventional biometric systems. For instance, [3] dealt with automatic extraction of soft biometric features from videos, whereas the role of soft biometrics in the problem of facial recognition from distance was studied in [4]. An exhaustive list about such works is efficiently compiled in [5]. Regarding analytic works, a theoretical analysis on the reliability of soft biometric systems was performed in [6]. The authors centered their study in those situations were identification error occurs due to different subjects sharing similar soft biometrics. In addition, they also provided asymptotic bounds for interpreting their statistical model. This work basically establishes a useful mathematical framework for predicting the benefits associated with using a greater number of soft biometric traits, albeit subjected to some constraints.

2.2. Privacy

Privacy has been used as a metric for measuring the level of uncertainty of information corresponding to an individual within a database. Privacy preservation was first described in [7] as the guarantee that an adversary learns nothing extra about any target if the adversary gains access to published data. Regarding databases, public and private attributes are generally modeled as random variables having a specific joint probability distribution. The privacy of an individual remains intact (i.e. there is no privacy loss) if the disclosure of the associated public attributes provides no additional information about the corresponding private attributes. Conventionally, privacy has been accounted for in an information theoretic way. The uncertainty about a piece of undisclosed information is related to its information content. The information content of a source S is measured by its entropy H which is defined as –

$$H(S) = \sum_i p_i \log \frac{1}{p_i}$$

where p_i is the probability with which a character s_i is emitted from the source S.

Let X_{prv} denote the set of random variables which represent sensitive attributes in a database. Similarly, let X_{PUB} characterize the set of random variables corresponding to the public information which is accessible by the adversary. As demonstrated in subsequent sections, the source of this information can be external public attributes as well as correlated auxiliary side information. Furthermore, let's assume that X_{prv} and X_{PUB} are correlated by a joint probability distribution function $p_{(X_{prv}, X_{PUB})}(y, x)$ where $\forall (y, x) | y \in X_{prv}$ and $x \in X_{PUB}$. Under such a naive scenario, the privacy of the individual ($\mathcal{P}$) can be quantified as –

$$\mathcal{P} = H(X_{prv} | X_{PUB})$$

where $H(X_{prv}|X_{PUB})$ represents the conditional entropy (equivocation) of X_{prv} given X_{PUB}. The parameter $\mathcal{P}$ accurately captures the essence of privacy as it represents

[1] Zero leaked information translates to the highest level of privacy whereas complete disclosure of information relates to zero privacy.