World Wide Web
DOI 10.1007/s11280-015-0327-3

Anonymizing multimedia documents

Bechara Al Bouna · Eliana J. Raad · Richard Chbeir · Charbel Elia · Ramzi Haraty

Received: 11 March 2014 / Revised: 18 October 2014 / Accepted: 6 January 2015
© Springer Science+Business Media New York 2015

Abstract Multimedia documents sharing and outsourcing have become part of the routine activity of many individuals and companies. Such data sharing puts at risk the privacy of individuals, whose identities need to be kept secret, when adversaries get the ability to associate the multimedia document's content to possible trail of information left behind by the individual. In this paper, we propose de-linkability, a privacy-preserving constraint to bound the amount of information outsourced that can be used to re-identify individuals. We provide a sanitizing MD^*-algorithm to enforce de-linkability along with a utility function to evaluate the utility of multimedia documents that is preserved after the sanitizing process. A set of experiments are elaborated to demonstrate the efficiency of our technique.

Keywords Data privacy · Anonymity · De-linkability · Multimedia document

B. Al Bouna (✉) · C. Elia
Ticket Labs, Antonine University, Baabda, Lebanon
e-mail: bechara.albouna@upa.edu.lb

C. Elia
e-mail: charbel.elia@upa.edu.lb

E. J. Raad
LE2I-CNRS, Bourgogne University, Dijon, France
e-mail: eliana.raad@u-bourgogne.fr

R. Chbeir
LIUPPA Laboratory, University of Pau and Adour Countries, Pau, France
e-mail: richard.chbeir@univ-pau.fr

R. Haraty
School of Arts, Sciences, Lebanese American University, Beirut, Lebanon
e-mail: rharaty@lau.edu.lb

Published online: 06 February 2015

1 Introduction

Large scale web applications are gaining increasing interest in recent times across a range of sectors, in both large and small firms. Companies are now constantly looking at what kind of data they have and what data they need in order to maximize their market position. In the era of big data, there are concerns about data privacy, and even the potential future value of data, as expressed in the White House Counsel John Podesta's 2014 report to the President on the challenges of Big Data [25]. The main added privacy risk is that this data – from voice calls, emails and texts to uploaded pictures, video, and music – is being reused and combined with other data in ways never before thought possible.

On the other hand, the ever-increasing amount of information flowing through social media and blogging sites has reflected the need for heightened privacy controls. More than 500 million photos are uploaded and shared every day, along with more than 200 hours of video every minute. In many situations, motivated by several campaigns such as politics, fraud fighting, cultural critics, and others, authors of some of these social media need to remain anonymous. Consequently, when a data provider outsources or publishes multimedia documents, it becomes extremely hard sometimes to maintain individuals' anonymity mainly due, but not limited, to: 1) the number of active social networks to which they actually participate, and 2) the trails of seemingly information they leave behind [20]. These trails of information make individuals victims of what is known by the Internet community as *cyberstalking* where an adversary clandestinely tracks the movements of an individual. The "Twitter Hunt"[1] scenario in which an adversary was able to re-identify the previous french prime minister *François Fillon* expresses clearly the risk of re-identifying anonymous individuals.

In this scenario, the adversary recognized the prime minister, who was using a fake account name "@fdbeauce" to remain anonymous, using the profile information available and previously published "tweets" which contained *enough* clues to disclose his identity. The first clue that made this attack successful is the username or the Twitter alias "@fdbeauce". This alias is based on his real information: *'F'* as the initial of his name and *'debeauce* is taken from the village name of Beaucé. The second clue is the profile image published on his account, and more precisely the GPS coordinates embedded in its metadata indicating that it was taken in *Beaucé*.Actually, *François Fillon* lives in a manor in *Beaucé*, in the department of Sarthe in western France. The Wikipedia page[2] about *François Fillon* contains this information, including a picture of the manor where he lives. The username information and profile image of *François Fillon*, combined with existing public knowledge from Wikipedia, allowed to re-identify him even when using a pseudonym.

It goes without saying that every anonymous multimedia document published can be put at risk and linked back to the individual without appropriate anonymization techniques. Indeed, exploiting inferable information can disclose anonymized identities where unrestricted access to online personal information remains a major threat. Most of the works done in the literature to preserve anonymity focus on structured relational data [10, 27, 30] while the only few techniques [12, 18] proposed to handle identity anonymization in multimedia documents assume textual data with no reference whatsoever to multimedia objects such as images and videos.

[1] http://www.euronews.com/2011/12/12/french-pms-shy-twitter-debut/

[2] http://en.wikipedia.org/wiki/Francois_Fillon

In a previous work [3], we proposed *de*-linkability, a novel technique for preserving individual privacy when outsourcing multimedia documents. *de*-linkability ensures that individuals' identifiable information composed of both textual and multimedia content cannot be used to infer his/her identity. In this paper, we extend the approach by including attributes to describe this information in a multimedia document and to quantify common information between multimedia documents. To do so, three operators were defined to:

1. compare text content between documents such as "@fdbeauce" and *Beaucé* in the Wikipedia page about *François Fillon*,
2. compare multimedia content according to multimedia objects such as the (geographical coordinates) metadata of the profile image of *François Fillon* and the metadata of the manor picture from Wikipedia page,
3. match text and multimedia data such as "@fdbeauce" and the (geographical coordinates) metadata of the profile image of *François Fillon*.

Our contributions can be summarized as follows:

– We formally define the identity anonymization problem in multimedia documents composed of textual and multimedia content.
– We quantify the re-identification threat which is highly dependent on how much information can be acquired from 1) adversaries' background knowledge and 2) external sources containing relevant information related to the anonymized individual.
– We present our sanitizing $\mathcal{MD}^*$-algorithm that allows to sanitize multimedia documents' content and preserve at the same time their utility in order to achieve the *de*-linkability.
– We provide a utility measure to determine to which extent a multimedia document remains consistent after the sanitizing process.

The remainder of this paper is organized as follows. In Section 2, we present the adversary model adopted in our study. In Section 3, we discuss some of the works on anonymous document outsourcing and privacy preserving. Our data model definitions and operators are presented in Section 4. In Section 5, we give a formal definition of the re-identification problem. Section 6 is dedicated to present the *de*-linkability privacy constraint and to show how it is possible to preserve individual anonymity using a multimedia document sanitizing algorithm (the MD^*-algorithm) and a utility measure. In Section 7, we evaluate our sanitizing algorithm to finally conclude and discuss our future research directions in Section 8.

2 Adversary model

In our adversary model, we assume that the adversary, that we call *cyberstalker*, knows that a given individual, that we call *cyberstalkee*[3], is hiding his/her identity (e.g., *François Fillon* in our scenario). We also assume that the *cyberstalker* has access to public information enabling him/her to link some personally identifying information, in a outsourced multimedia document, to the *cyberstalkee*. Thus, all *relevant* information

[3]Both terms *cyberstalkee* and *individual* will be used interchangeably in the remainder of this paper.

(identifying or quasi-identifying) extracted from the document is considered individually identifiable.

More subtle, we assume that the *cyberstalker* has no prior knowledge of specific values for the stalked individuals. For example, the *cyberstalker* described in our motivating example does not know a-priori that "*Château de Beaucé*" is the residence of the *cyberstalkee*.

3 Related work

Several techniques have been defined in the literature [17, 19, 27, 30] to prevent information disclosure and eliminate possible linking attacks that are used for individual re-identification. These techniques assume that identifiable information and adversaries' background knowledge are stored in structured relational datasets. Specifically, they address linking attacks that can be established between (quasi)-identifying [31] and sensitive attributes of individuals stored in schema-based relational tables without referring to multimedia content.

Alternatively, techniques described in [12] and [18] preserve the individual's privacy in free text documents where data structure is missing. In [12], the authors measure sensitivity of identifiable information through a top-down propagation technique using prefixed sensitivity levels mapped to a reference ontology. According to these computed sensitivity levels, words are disseminated. In [18], the authors use a probabilistic-based algorithm to mine all searchable information concerning the individual. They use domain-specific ontologies to capture inferrable information and eventually provide more accurate results. Unfortunately, the ability of these techniques [12, 18] to deal with strong adversaries enforced with plausible background knowledge is limited when using domain-specific ontologies to compute sensitivity levels. These so-called levels of sensitivity should depend mainly on the knowledge that the adversary already has acquired which could be out of scope of a specific ontology. In [22], the authors propose a novel technique based on relevant occurrences to find user semantics. They assume that word co-occurrence is important to extract personal information from the Web. Similarly, in [13], the authors consider that the queries returning few results should be denoted as important. However, the amount of information is not always a relevant measure of dependency for privacy. For instance, two "tweets" with minimum co-occurrence might be issued by the same individual. Techniques described in [23], [28] and [6] are similar to a certain extent to our work. In [28], the authors propose a web-based solution to control undesired inferences. It first extracts relevant keywords from the document to be published and queries the web in order to capture additional knowledge contributing to a privacy breach. In [6], the authors present the notion of k-safety in which the identifying terms should be associated with at least k individuals. The authors in [23] sanitize sensitive parts of the document to measure information loss and risk disclosure. They assume that a relevant sanitizing process could be applied to maintain the utility of information in the document. As demonstrated in their experiments, these techniques are practical and promising, yet their ability to handle multimedia documents is limited. Unlike textual attributes, multimedia content cannot be approached without special processing to reduce uncertain decisions that overcome when similarity operators come to play. Here, we propose a technique to tackle individual re-identification threat caused by textual and multimedia content that can be linked to information obtained from external sources.

Table 1 Notations

u	an individual with anonymized identity
pf_u	an individual profile
mo	a multimedia object
$\mathcal{MD}_u$	a multimedia document related to u which should be sanitized
$\mathcal{MD}_\beta$	a multimedia document publicly accessible to adversaries extracted from an external source $\mathcal{E}$
$S_{\mathcal{MD}}$	a multimedia document signature
$\mathcal{E}$	an external source such as the social website, domain specific database, etc.
$\mathcal{A}$	a set of attributes relevant to the multimedia document content
α	an association threshold
β	an identification threshold
ϖ	an aggregation function such as average, minimum, max, etc.
$\mathcal{R}_w$	words relevance
$\mathcal{R}_m$	multimedia relevance
$\mathcal{U}$	multimedia document utility

4 Data model

In this section, we define the data model and the basic notations (Table 1) used in the remainder of this paper.

4.1 Data definition

Definition 1 (Attribute Set) $\mathcal{A}$ is a set of attributes where $\forall a_i \in \mathcal{A}$ for $(1 \leq i \leq |\mathcal{A}|)$, a_i can be any attribute of the dublin core metadata element set[4] such that $\{source, description, date, contributor, format\}$ or the MPEG-7 semantic set[5] $\{semantic\_place, concept, state, event, object\}$ or any domain specific attribute (e.g., spatial or temporal domain). We use $ma_i \in \mathcal{A}$ to denote a multimedia attribute whose values are of complex structure such as a BFILE/BLOB, an URL/URI, an URL/URL augmented with a primitive to represent a salient object (e.g., Minimum Bounding Rectangle, Circle) or a multimedia object (to be defined below).

Definition 2 (Multimedia Object) Let **mo** be any type of multimedia data such as an image, a video, or a salient object describing an object of interest (e.g., face of a person.). **mo** is formally represented as:

$$mo :\prec \mathcal{A}_m, V, O, MO, \varsigma \succ \tag{1}$$

where:

- $\mathcal{A}_m \subseteq \mathcal{A}$ is a subset of attributes of $\mathcal{A}$ whose values are used to identify a multimedia object mo.
- V is a set of values describing the multimedia object. $\forall v_i \in V$ for $(1 \leq i \leq |V|)$ $v_i \in \mathcal{D}(a_j)$ where a_j is an attribute of $\mathcal{A}_m$.
- O is the raw data of the multimedia object. $O \in \mathcal{D}(a_i)$ where a_i is a multimedia attribute of $\mathcal{A}_m$. $O(mo)$ denotes the raw data of multimedia object mo.

[4] http://dublincore.org/

[5] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

- MO is a set of multimedia objects *directly* contained in mo (i.e., we only consider multimedia objects at the first level of the hierarchy). Using such recursive definition helps better describe the hierarchy between the different components that could be hierarchically linked in a single multimedia object. This is the case of a video segment that is constituted of scenes, frames and salient objects as shown in Figure 2. $MO(mo)$ denotes the set of multimedia objects contained in mo.
- $\varsigma \subseteq \mathcal{A}_m \times X = \{(a_j, x_i) | a_j \in \mathcal{A}_m, x_i \in V \text{ or } x \text{ is } O\}$ is an association function that assigns each attribute a_j to its corresponding value which is either a textual $v_i \in V$ or a multimedia raw data O.

For example, Figure 1 shows multimedia objects mo_{beauce} and mo_{manoir} representing two images of "*Château de Beaucé*" where *keywords* is an attribute of mo, O contains the raw data and MO is the empty set of multimedia objects contained in mo. Figure 2 is another example of a video segment represented using our multimedia object representation. This is to express the sub-content of the video including scenes, frames and salient objects that are hierarchically linked.

Definition 3 (Individual Profile) Let u be a *cyberstalkee*, we denote by pf_u the profile of u formally defined as:

$$pf_u :\prec \mathcal{A}_p, PI, MO, \gamma \succ \tag{2}$$

where:

- $\mathcal{A}_p \subseteq \mathcal{A}$ is a subset of attributes of $\mathcal{A}$ whose values are used to identify an individual profile pf_u.
- PI is a set of values describing the individual's personal information. $\forall v_i \in PI$ for $(1 \leq i \leq |PI|)$, $v_i \in \mathcal{D}(a_j)$ where a_j is an attribute of $\mathcal{A}_p$.
- MO is a set of multimedia objects attributed to u such that $\forall mo_i \in MO$ for $(1 \leq i \leq |MO|)$ $mo_i \in \mathcal{D}(ma_j)$ where ma_j is a multimedia attribute $\in \mathcal{A}_p$.
- $\gamma \subseteq \mathcal{A}_p \times X = \{(a_j, x_i) | a_j \in \mathcal{A}_p, x_i \in PI \text{ or } x_i \in MO\}$ is an association function that assigns each attribute a_j to its corresponding value which is either a textual $v_i \in PI$ or multimedia $mo_i \in MO$.

(a)

mo_{beauce}	Keywords	O	MO
	Chateau de Beaucé	http://chateaubeauce.com/beauce.jpg	-

Multimedia object representing "*Château de Beaucé*"

(b)

mo_{manoir}	Keywords	O	MO
	Manoir de Beaucé	http://manoibeauce.com/beauce.jpg	-

Multimedia object representing "*Manoir de Beaucé*"

Figure 1 Examples of typical image descriptions using our multimedia object representation. Figure 1-a and Figure 1-b (a cropped image of the former) show similar content, contained in two different multimedia documents

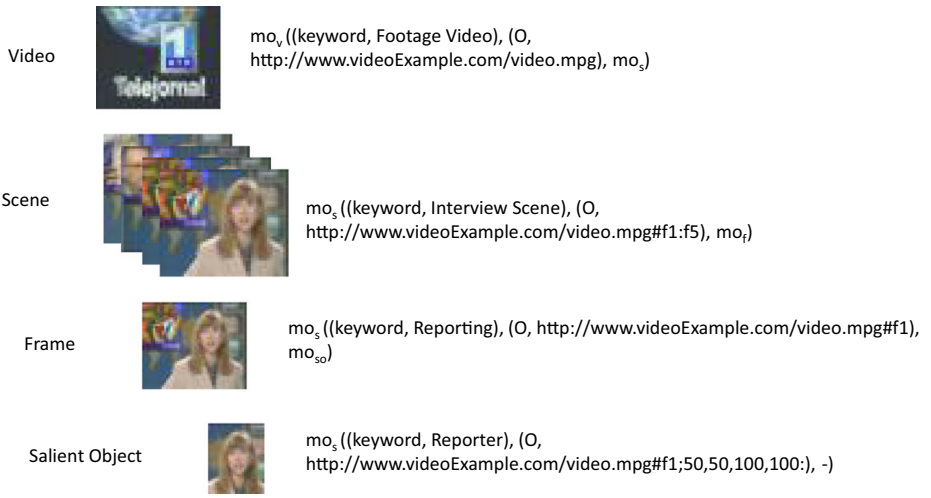


Figure 2 Example of a typical video representation using our multimedia object representation

Referring back to our scenario, a typical profile of the previous french Prime Minister *François Fillon* would be:[6]

$$pf_{Fillon} : ((\text{name}, \textit{François Fillon}), (\text{job}, \textit{Prime Minister}), (country, France), (email, fcafillon@wanadoo.fr), (home, mo_{beauce}))$$

Definition 4 (Multimedia Document) Let $\mathcal{MD}$ be a multimedia document. $\mathcal{MD}$ is two dimensional and composed of a set of words and multimedia objects. It is formally defined as follows:

$$\mathcal{MD} :\prec W, MO \succ \tag{3}$$

where:

- W is a base text that represents the document's content where $\forall w_i \in W$ for $(1 \leq i \leq |W|)$, w_i is a word contained in $\mathcal{MD}$ or a metadata (attribute information) of the document.
- MO is a set of multimedia objects where $\forall mo_i \in MO$ for $(1 \leq i \leq |MO|)$, mo_i is a multimedia object contained in $\mathcal{MD}$.

An example of a multimedia document could be, but not limited to, personal blogs, set of tweets, newspaper articles, etc. Typically, these documents are composed of words and multimedia objects.

Now that we have defined our multimedia document, we present in the following what we call a multimedia document signature ($S_{\mathcal{MD}}$).

Definition 5 (Multimedia Document Signature) Let $\mathcal{MD}$ be a multimedia document, a multimedia document signature denoted by $S_{\mathcal{MD}}$ is a subset of $\mathcal{MD}$ composed of textual and multimedia content. $S_{\mathcal{MD}}$ is created using $S_{\mathcal{MD}} = \mathcal{IA}(\mathcal{MD}, \mathcal{A}_s)$ where $\mathcal{IA}$ is a function

[6]It represents an attribute whose values are multimedia objects (e.g., pictureOf, imageOf, etc.).

used to retrieve from $\mathcal{MD}$ relevant words and multimedia objects related to the subset of attributes $\mathcal{A}_s \subseteq \mathcal{A}$.

We assume that not all attributes found in a multimedia object provide meaningful clues that could lead to re-identify the *cyberstalkee*. In essence, the idea is to generate signatures that are mainly related to the individuals. This could be done by determining the most significant and relevant attributes retrieved from the document's content and/or using some of the attributes from the individuals' profiles. These attributes help reducing the error rate of individual name disambiguation [14], particularly when the individual's profile is considered as a relevant source of attributes. For instance, it is unlikely for an individual working in a Health Care Department to be related to Computer Science. In other terms, some of the words and multimedia objects should more likely be related to the medical field instead of computing.

The followings are three sample multimedia documents' signatures generated based on the attributes *Country, Event*, *Location* and *Image* .

$$S_{\mathcal{MD}_{Fillon}} : ((visiting, \text{“}Japan\text{”}), (visit, \text{“}Meeting\text{”}), (annotation, \text{“@}beauce\text{”}), (home, mo_{beauce}))$$

$S_{\mathcal{MD}_{Fillon}}$ is the anonymous multimedia document signature of Prime Minister François Fillon.

$$S_{\mathcal{MD}_{\beta_1}} : ((name, \text{“}François\ Fillon\text{”}), (country, \text{“}France\text{”}), (visiting, \text{“}Japan\text{”}), (visit, \text{“}Meeting\text{”}))$$

$$S_{\mathcal{MD}_{\beta_2}} : ((name, \text{“}François\ Fillon\text{”}), (annotation, \text{“}Home\text{”}), (home, mo_{manoir}))$$

Both $S_{\mathcal{MD}_{\beta_1}}$ and $S_{\mathcal{MD}_{\beta_2}}$ are publicly available multimedia documents signatures related to Prime Minister *François Fillon*.

4.2 Data comparison

We provide in this section, the appropriate operators to address both multimedia and textual content of multimedia documents.

Definition 6 (Estimated Equality) Let W_1, W_2 be two sets of words over which an association function f can be used. Their estimated equality is computed as follows:

$$equ(W_1, W_2) = \varpi(f(w_1^1, w_1^2), \ldots, f(w_m^1, w_r^2)) \rightarrow [0, 1] \quad (4)$$

where:

- w_i^1, w_j^2 are two words of W_1 and W_2 respectively where $m = |W_1|$ and $r = |W_2|$.
- f is an association function defined as:

$$f(w_i^1, w_j^2) = \begin{cases} 1 & \text{if} w_i^1 \in W_1 \text{is the same as} w_j^2 \in W_2 \\ 0 & \text{otherwise} \end{cases}$$

- ϖ is an aggregation function (e.g., max, min, avg, etc.) used to aggregate association functions' scores.

In our example, the estimated equality takes the set of words in the Wikipedia article about *François Fillon* and compares them with the set of words on his Twitter page. If we

use the *substring* function SUBSTR between words from the two sets, we see that *beauce* (from the Wikipedia page) is the substring value we wish to find from *fdebeauce* (from the Twitter page).

The estimated equality is used to identify the amount of common textual values found in multimedia documents (or any subset of them). Alternatively, multimedia documents contain complex types such as images and videos which cannot be approached using traditional equality operators. We define in the following, an estimated similarity operator to process multimedia objects.

Definition 7 (Estimated Similarity) Let MO_1, MO_2 be two sets of multimedia objects over which n similarity functions $s_1, \ldots, s_n$ can be used. Their similarity score is computed as follows:

$$sim(MO_1, MO_2) = \varpi(s_1(mo_1^1, mo_1^2), \ldots, s_n(mo_m^1, mo_r^2)) \rightarrow [0, 1] \quad (5)$$

where:

- mo_i^1, mo_j^2 are two multimedia objects of MO_1 and MO_2 respectively where $m = |MO_1|$ and $r = |MO_2|$.
- s_k is a *unit* similarity function comparing multimedia objects $mo_i^1 \in MO_1$ and $mo_j^2 \in MO_2$. We note that $s_k(mo_i^1, mo_j^2)$ compares[7] mo_i^1 and mo_j^2 based on their attributes and raw data. s_k returns a score between [0, 1], where 0 expresses a total divergence and 1 a complete similarity.
- ϖ is an aggregation function used to aggregate the computed similarity scores.

We give an example to illustrate the estimated similarity between two sets of multimedia objects. Therefore, we propose a unit similarity function that takes coordinates of two images and returns 1 if the input coordinates belong to the same geographical location. We see in our example that one of the images published on the Twitter account of François Fillon was captured in the same geographical coordinates (*Latitude* : 48.357483, *Longitude* : -1.116662) as the image of "*Château de Beaucé*" found on the Wikipedia page of Fillon.

Definition 8 (Cross-Matching Score) Let $S_{\mathcal{MD}_1}$, $S_{\mathcal{MD}_2}$ be two distinct multimedia documents signatures. The cross-matching score between their components (W and MO) is computed as follows:

$$match(S_{\mathcal{MD}_1}, S_{\mathcal{MD}_2}) = \lambda_m \times f(W_1, MO_2) + (1 - \lambda_m) \times f(W_2, MO_1) \rightarrow [0, 1] \quad (6)$$

where:

- $f(W_1, MO_2)$ and $f(W_2, MO_1)$ are association functions to determine the association of a set of words contained in $S_{\mathcal{MD}_1}$ ($S_{\mathcal{MD}_2}$ respectively) with the set of multimedia objects contained in $S_{\mathcal{MD}_2}$ ($S_{\mathcal{MD}_1}$ respectively). $f(W_1, MO_2)$ is defined as:

$$f(W_1, MO_2) = \varpi(f_1(w_1^1, mo_1^2), \ldots, f_n(w_m^1, mo_r^2)) \rightarrow [0, 1] \quad (7)$$

where:

  - f_k is a *unit* association function used to determine whether there is an association between a word $w_1 \in W_1$ with a multimedia object $mo_1 \in MO_2$.

[7] We invite the reader to consult our work on multimedia objects similarity computation in [2].

We note that $f_k(w_i^1, mo_j^2)$ determines the association between w_i^1 and mo_j^2 based on the attributes and values of the latter. For example, it may associate a GPS coordinates, specified as one of the multimedia objects' metadata, with a word representing the corresponding location. f_k returns a score between [0, 1], where 1 represents a perfect association of the attributes and 0 represents the absence of association.

- ϖ is an aggregation function used to aggregate the computed association scores.

- $\lambda_m \in [0, 1]$ allows to assign priorities to $f(W_1, MO_2)$ and $f(W_2, MO_1)$, based on the relevance of multimedia objects in the multimedia documents.

To illustrate the influence of the cross-matching, we take as input the set of words from the Wikipedia page about *François Fillon* and the set of multimedia objects from his Twitter page. We define a function that maps words from the first set to geographical coordinates from the second set. By doing so in our example, we found that the profile image published on the Twitter account and "*Château de Beaucé*" found on the Wikipedia page refer to the same location.

We show in the following how multimedia documents intersection can be determined using selective intersection.

Definition 9 (Selective Intersection) Let $S_{\mathcal{MD}_1}, S_{\mathcal{MD}_2}$ be two distinct multimedia documents signatures, their selective intersection is defined as:

$$SelInt(S_{\mathcal{MD}_1}, S_{\mathcal{MD}_2}) = ||\sum_{a_i} wa_i \times equ_{a_i}(W_1, W_2) + \sum_{a_j} wa_j \times sim_{a_j}(MO_1, MO_2) + \sum_{a_k} wa_k \times match_{a_k}(S_{\mathcal{MD}_1}, S_{\mathcal{MD}_2})|| \quad (8)$$

where:

- a represents an attribute for which an equality, similarity and/or cross-matching score should be computed. Such attributes, defined in the attribute set, can be used to selectively choose relevant content in multimedia documents. For instance, it is possible to capture the amount of common information related to the attribute *Person*. This refers to computing equality, similarity and cross-matching of words and multimedia objects that are related to this attribute for both multimedia document signatures $S_{\mathcal{MD}_1}, S_{\mathcal{MD}_2}$.
- wa is the weight assigned to attribute a where its magnitude depends on the normalizing assumptions.

Selective intersection returns a normalized score $\in [0, 1]$ computed based on equality, similarity and cross-matching of multimedia documents content. For instance, let us compute the selective intersection between $S_{\mathcal{MD}_{Fillon}}$ and both $S_{\mathcal{MD}_{\beta_1}}$ and $S_{\mathcal{MD}_{\beta_2}}$. We adopt the *max* aggregation function to compute the equality, similarity and/or matching scores for each attribute and finally determine their average score. The selective intersection based on

the attributes *Country, Event* and *Location* is detailed below:

$$Sel\,Int(S_{\mathcal{MD}_{Fillon}}, S_{\mathcal{MD}_{\beta_1}}) = \frac{\frac{1+1+0}{3} + 0 + 0.3}{3} = 0.32$$

$$Sel\,Int(S_{\mathcal{MD}_{Fillon}}, S_{\mathcal{MD}_{\beta_2}}) = \frac{\frac{0}{3} + 0.8 + 0.5}{3} = 0.44$$

We assume, in this case, that the estimated similarity between multimedia objects mo_{beauce} and mo_{manoir} in $S_{\mathcal{MD}_{Fillon}}$ and $S_{\mathcal{MD}_{\beta_2}}$ returns a 0.8 score. The cross-matching score between multimedia object mo_{beauce} and $France$ in $S_{\mathcal{MD}_{Fillon}}$ and $S_{\mathcal{MD}_{\beta_1}}$ respectively returns a 0.3 score for semantic similarity. This matching returns a 0.5 score between $S_{\mathcal{MD}_{Fillon}}$ and $S_{\mathcal{MD}_{\beta_2}}$ based on the matching between keywords attribute of the multimedia object mo_{manoir} and $@beauce$.

Unlike mutual information metric [26] which is based on joint probability measures, our selective intersection is a non-correlation based metric where the count of each value in the signatures has minimum influence on the overall computation score. Specifically and for privacy reasons, this assumption is useful to determine "minimum" intersection between multimedia documents where weighted attributes reflect relevant association measure if processed efficiently. We will show in the following definition, the premise of multimedia documents association.

Definition 10 (α-*association*) Let $\mathcal{MD}_1$, $\mathcal{MD}_2$ be two distinct multimedia documents. We say that an α-association exists between $\mathcal{MD}_1$ and $\mathcal{MD}_2$ if their selective intersection $Sel\,Int(S_{\mathcal{MD}_1}, S_{\mathcal{MD}_2})$ is greater than α where:

- $S_{\mathcal{MD}_1}$ and $S_{\mathcal{MD}_2}$ represent corresponding multimedia documents signatures.
- $\alpha \in [0, 1]$ is the association threshold.

α-*association* expresses the presence of a possible association between two multimedia documents represented by their signatures. It measures the strength of an association between two multimedia document signatures based on their common information composed of both textual and multimedia content.

5 Identity anonymization problem

In the presence of adversaries with sophisticated tracking abilities, privacy and ownership preserving of outsourced data tends to be a complex task. Such adversaries, armed with plausible background knowledge and a wide range of accessible web-based social information, compromise anonymization techniques and put at risk individuals' privacy. Here, we express the identity anonymization problem that could arise when outsourcing multimedia documents as the amount of information accessible by the adversary and that can be, at the same time, associated with the owner of the outsourced multimedia documents. It is formally defined as follows:

Definition 11 (Identity Anonymization Problem) Let $\mathcal{MD}_u$ be the multimedia document of an individual u. We say that an adversary is able to re-identify u from $\mathcal{MD}_u$ if $\exists \mathcal{MD}_\beta$, a *publicly available* multimedia document, such that:

- $\mathcal{MD}_u$ and $\mathcal{MD}_\beta$ are α-associated and,

– The knowledge related to u that can be obtained from $\mathcal{MD}_\beta$ is greater than β. It is expressed as a β-association between $\mathcal{MD}_\beta$ and the individual profile pf_u where α is the association threshold, β is an identification threshold and both α and β are user-defined.

It is difficult to know how much the adversaries know and to what extent their ability to disclose individuals' identities can be compromising. Here, we only avoid leaking information to the *cyberstalker* except for what he/she already has. Such assumption is no different than the one adopted by differential privacy [10] where our main objective is essentially providing constraints on the release of the data.

6 Privacy preservation prior to publication

Preserving privacy requires that the *cyberstalker* remains unable to detect the anonymized identity of the *cyberstalkee*, owner of the multimedia document to be published. As we have stated in the previous section, a re-identification threat occurs mainly due to:

– the link between his/her related multimedia document $\mathcal{MD}_u$ and a multimedia document $\mathcal{MD}_\beta$ accessible by the *cyberstalker* and,
– the amount of information extracted from $\mathcal{MD}_\beta$ and associated with u.

Controlling the latter can be a burden or eventually unrealizable due to accessibility issues while, on the other hand, breaking the link between multimedia documents is achievable and can be done using de-linkability.

***de*-linkability**. *Given a cyberstalkee u and a multimedia document $\mathcal{MD}_u$, the de-linkability privacy-preserving constraint is satisfied if $\forall \mathcal{MD}_\beta \in \sigma_{E_u}(\mathcal{E})$ that is β-associated with pf_u, $\mathcal{MD}_u$ cannot be linked to $\mathcal{MD}_\beta$ through an α-association, where $\sigma_{E_u}(\mathcal{E})$ is a selection on an external source $\mathcal{E}$ based on a conjunctive set of words and/or multimedia objects (E_u) related to u.*

de-linkability breaks the link between a multimedia document (to be published or outsourced) and any other document *accessible* to a *cyberstalker* and that can be linked to u. It is important to note that the content of E_u that is used to retrieve multimedia documents $\mathcal{MD}_\beta$ from the external source should be considered carefully in order to reduce the scope of potential error. A straightforward assumption is to consider this content as a subset of the individual's profile including both identifying and quasi-identifying values.

6.1 Achieving *de*-linkability

de-linkability can be achieved in textual documents in a straightforward way using extension of traditional anonymization techniques such as suppression, substitution or generalization relationships between domains and values [27, 29, 30] for textual values in $\mathcal{MD}_u$ as long as there is no $\mathcal{MD}_\beta$ that can be α-associated with $\mathcal{MD}_u$. Unsurprisingly, multimedia objects need a special interest. Eventually, the objective is to break linkable objects that could contribute in re-identifying the anonymized individual. More subtle is to hide and/or disseminate multimedia objects content while at the same time preserving a minimum semantic or visual coherence. In this paper, we do not provide an in-depth details on how multimedia objects content could be protected. This matter is left for future work. We

only use traditional techniques to protect salient objects as in [4] where the authors protect textual and image data through flexible low-level adapted security rules, while in [11] object substitution is adopted. In [5], blurring proved efficiency, and objects removal from images and videos were addressed in [8, 9, 15, 24, 32, 33].

Here, we refer to this process as document sanitization which we formally define as follows:

Definition 12 (Multimedia Document Sanitization) Let $\mathcal{MD}_u$ be the multimedia document related to a *cyberstalkee* u. Given $\widetilde{\mathcal{G}}_W$ and $\widetilde{\mathcal{G}}_{MO}$ two corresponding sanitizing functions, we say that $\mathcal{MD}_u$ is sanitized, denoted by $\mathcal{MD}^*_u = \widetilde{\mathcal{G}}_{(W,MO)}(\mathcal{MD}_u)$ if both words and multimedia objects are sanitized $\widetilde{\mathcal{G}}_W(W_{MD_u})$ and $\widetilde{\mathcal{G}}_{MO}(MO_{MD_u})$.

Multimedia document sanitization ensures that the specified content (W, MO) is either removed, suppressed, generalized and/or protected in the multimedia document $\mathcal{MD}_u$ based on the sanitization function $\widetilde{\mathcal{G}}$.

6.2 Multimedia document sanitization: $\mathcal{MD}^* - algorithm$

$\mathcal{MD}^*$-algorithm is used to sanitize a multimedia document and protect the *cyberstalkee*'s identity. As mentioned in the pseudo-code, the algorithm takes a multimedia document $\mathcal{MD}_u$, a set of attributes $\mathcal{A}_s$ (used to extract multimedia document signature), the *cyberstalkee* profile pf_u along with E_u and both association and identification thresholds α, β. It returns a sanitized multimedia document ($\mathcal{MD}^*_u$).

The $\mathcal{MD}^*$-algorithm extracts in Step 1 the multimedia document signature $S_{\mathcal{MD}_u}$ using the extraction function $\mathcal{IA}$. It sanitizes $\mathcal{MD}_u$ from Step 2 to 10.

In Step 3, it extracts the signature of a multimedia document $\mathcal{MD}_\beta$ retrieved from an external source $\mathcal{E}$ based on the set of entities E_u related to u. In order to determine the amount of information related to u and that can be obtained from $\mathcal{MD}_\beta$, we compute the selective intersection on $\mathcal{MD}_\beta$ and the *cyberstalkee* profile pf_u. If their selective intersection $Sel\,Int(S_{\mathcal{MD}_\beta}, pf_u)$ is greater than β, the link between $\mathcal{MD}_u$ and $\mathcal{MD}_\beta$ should be anonymized

Algorithm 1 $\mathcal{MD}^*$-algorithm

Require: a multimedia document $\mathcal{MD}_u$, set of attributes $\mathcal{A}_s$ over $\mathcal{MD}_u$, an individual profile pf_u, conjunctive set of words and/or multimedia objects E_u, association threshold α and identification threshold β

Ensure: Multimedia Document Sanitization $\mathcal{MD}^*_u$

1: $S_{\mathcal{MD}_u} = \mathcal{IA}(\mathcal{MD}_u, \mathcal{A}_s)$ ▷ Generate Multimedia Signature on $\mathcal{MD}_u$
2: **for each** $\mathcal{MD}_\beta$ **in** $\sigma_{E_u}(\mathcal{E})$ **do**
3: $S_{\mathcal{MD}_\beta} = \mathcal{IA}(\mathcal{MD}_\beta, \mathcal{A}_s)$ ▷ Generate Multimedia Signature on $\mathcal{MD}_\beta$
4: **if** $Sel\,Int(S_{\mathcal{MD}_\beta}, pf_u) > \beta$ **then**
5: **while** $Sel\,Int(S_{\mathcal{MD}_u}, S_{\mathcal{MD}_\beta}) > \alpha$ **do**
6: Retrieve least significantly threatening W_β and MO_β
7: $\mathcal{MD}^*_u \leftarrow \widetilde{\mathcal{G}}_{(W_\beta, MO_\beta)}(\mathcal{MD}_u)$ ▷ Sanitize $\mathcal{MD}_u$ based on W_β and MO_β
8: **end while**
9: **end if**
10: **end for**

as done from Step 5 to 8. That is, as long as they are α-associated the **least**[8] significant W_β and MO_β are sanitized in $\mathcal{MD}_u$.

6.3 Utility estimation

To ensure safety, there is trade-off to be made at the stake of utility in order to meet strong privacy requirements. While this could be limited in general, it is considered an absolute necessity in order to establish trust between data owners and data providers. This issue has been the essence of several works [7, 16, 34] that provide data anonymization. Here, we determine to what extent a multimedia document remains consistent after the sanitizing process. In particular, we provide an estimation of utility based on the relevance of both words and multimedia objects sanitized.

Definition 13 (Words Relevance $\mathcal{R}_w$) Let W^* be the set of words sanitized from $\mathcal{MD}_u$, we define words relevance, denoted by $\mathcal{R}_w(W^*)$, as the raw frequency of words addressed by the sanitizing process. It is computed as follows:

$$\mathcal{R}_w(W^*) = \sum_{w_i \in W^*} \frac{c(w_i)}{N_w} \tag{9}$$

where:

- W^* is the set of sanitized words from $\mathcal{MD}_u^*$.
- $c(w_i)$ is the number of times w_i appeared in the multimedia document.
- N_w is the total number of words in the multimedia document.

Note that $\mathcal{R}_w$ assigns importance to individual words sanitized from the multimedia document. It determines the relevance despite the adopted anonymization technique (generalization, suppression or encryption).

Unlike words, determining the relevance of multimedia objects depends on the raw data of the multimedia object.

Definition 14 (Multimedia Relevance $\mathcal{R}_m$) Let MO^* be the set of multimedia objects sanitized from $\mathcal{MD}_u$, we define multimedia relevance, denoted by $\mathcal{R}_m(MO^*)$, as the importance of multimedia objects sanitized from the multimedia document $\mathcal{MD}_u$. $\mathcal{R}_m(MO^*)$ is computed based on multimedia objects raw data. It is determined as follows:

$$\mathcal{R}_m(MO^*) = \frac{\sum_{mo_i \in MO^*} r_m(O(mo_i))}{\sum_{mo_i \in MO^*} \rho_i} \tag{10}$$

where:

- MO^* is the set of sanitized multimedia objects from $\mathcal{MD}_u^*$.
- $r_m(O(mo_i)) = \rho_i \frac{sizeOf(O(mo_i))}{sizeOf(O(mo_j))}$ is the relevance of mo_i's raw data.
- ρ_i is the importance threshold of the multimedia object mo_i. It can be computed based on the association of mo_i's raw data with words and/or multimedia objects from the individual's profile.

[8]The importance of retrieved W_β and MO_β is determined based on the priority thresholds prefixed in the selective intersection function.

- $sizeOf(O(mo_i))$ is the size of the raw data of multimedia object mo_i in terms of width and height.
- $sizeOf(O(mo_j))$ is the size of the *container* mo_j where $mo_i \in MO(mo_j)$.

Now that we have defined word and multimedia relevance measures, we provide in the following a formal definition of the utility of a multimedia document.

Definition 15 (Multimedia Document Utility $\mathcal{U}$) Let $\mathcal{MD}_u^*$ be a sanitized multimedia document of an individual u, we denote by $\mathcal{U}(\mathcal{MD}_u^*)$ the utility measure of $\mathcal{MD}_u^*$ which is the estimated coherence given the relevance of sanitized words and multimedia objects from $\mathcal{MD}_u$. It is formally defined as follows:

$$\mathcal{U}(\mathcal{MD}_u^*) = \frac{1 - \mathcal{R}_w(W^*) \times \mathcal{R}_m(MO^*)}{1 + \mathcal{R}_w(W^*) \times \mathcal{R}_m(MO^*)} \tag{11}$$

where $\mathcal{R}_w(W^*)$ and $\mathcal{R}_m(MO^*)$ are word and multimedia relevance metrics defined in equations 13 and 14.

$\mathcal{U}$ is used to express the trade-off between privacy and utility. It shows at which point a sanitized multimedia document can be considered useless according to the amount of relevant information it contains.

7 Experiments

In this section, we present a set of experiments to evaluate the efficiency of our approach. We implemented the $\mathcal{MD}^*$-algorithm code[9] in Java and conducted experiments using a 3.4 GHz Intel Core i7 with 16 GB RAM.

7.1 Dataset configuration

We used 200 individuals of the dataset published[10] by the authors of [1]. For each individual, we grouped 100 of his/her tweets to form his/her $\mathcal{MD}_u$. These $\mathcal{MD}_u$ have been filtered to remove identifying names. OpenCalais api[11] is used to extract attributes from multimedia documents $\mathcal{MD}_u$ and $\mathcal{MD}_\beta$. We actually used the most relevant attributes extracted based on a predefined threshold that we have set to 0.5 (this threshold can be used to fine-tune the evaluation results and include relevant attributes).

Alternatively, we limited our use of multimedia objects to images. We specifically used the Zemanta api[12] to retrieve and associate images with their related words contained in $\mathcal{MD}_u$. As a matter of fact, the images that were mainly retrieved from the web, compensate the lack of metadata that could be used to link words to their corresponding images. That being said, the use of the Zemanta api enriched the content of $\mathcal{MD}_u$ with multimedia objects that could be used to re-identify individuals.

[9] The source code of the prototype can be downloaded from http://sourceforge.net/p/pmi1/code/HEAD/tree/trunk/MDanon/

[10] http://wis.ewi.tudelft.nl/umap2011/

[11] http://www.opencalais.com/

[12] http://developer.zemanta.com/

Individual profiles pf_u were downloaded using the Twitter api[13]. For our assessment, we only focused on four profile attributes namely *name, screen name, location and profile_image_url*.

As per *cyberstalkee*, we retrieved up to 10 relevant multimedia documents $\mathcal{MD}_\beta$ using the Google api[14] applying to the individual *name* combined to relevant content from his/her related $\mathcal{MD}_u$. This way, we can assert that the retrieved multimedia documents $\mathcal{MD}_\beta$ are related to the *cyberstalkee* at hand at least through their names.

To compare images, we used the phash function[15] and assigned a weight of 0.5 to the estimated similarity for the selective intersection $S_{el}\,I_{nt}$.

7.2 Evaluation Results

We elaborated a set of measurements to evaluate the efficiency of the $\mathcal{MD}^*$-algorithm. These measurements can be summarized as follows:

- Evaluating the identity anonymization problem represented by the percentage of individuals re-identified.
- Determining uncertainty raised after the sanitizing of multimedia documents.
- Evaluating the utility of multimedia documents after the sanitizing process.
- Determining the computational cost of our $\mathcal{MD}^*$-algorithm.

7.2.1 Evaluating Privacy

In this test, we evaluated the identity anonymization problem represented by the percentage of individuals identified according to what they have published in their $\mathcal{MD}_u$ and their related multimedia documents $\mathcal{MD}_\beta$. We fixed the identification threshold $\frac{1}{\beta} = 10$ in order to capture significant number of multimedia documents related to individual u and used various association thresholds $\frac{1}{\alpha} = 2, 4, 6, 8$ and 10. The results shown in Figure 3 show the percentage of re-identified individuals (in Figure 3(a)) and the number of threatening MD_β (in Figure 3(b)) .

We can see that when the association threshold increases, there is a higher chance of linking individuals to the multimedia documents $\mathcal{MD}_\beta$ retrieved from the external source and eventually leading to their re-identification.

7.2.2 Evaluating uncertainty

We evaluate the MD^*-algorithm to determine the increasing uncertainty raised due to the sanitizing process[16]. To do so, we calculate the average entropy [21] of individuals' multimedia documents $\mathcal{MD}_u$ in a pre- and post-sanitizing process. As a matter of fact, for each individual's multimedia document, we compute its entropy based on the most relevant attributes used to generate its own multimedia document signature (see Definition 5) as:

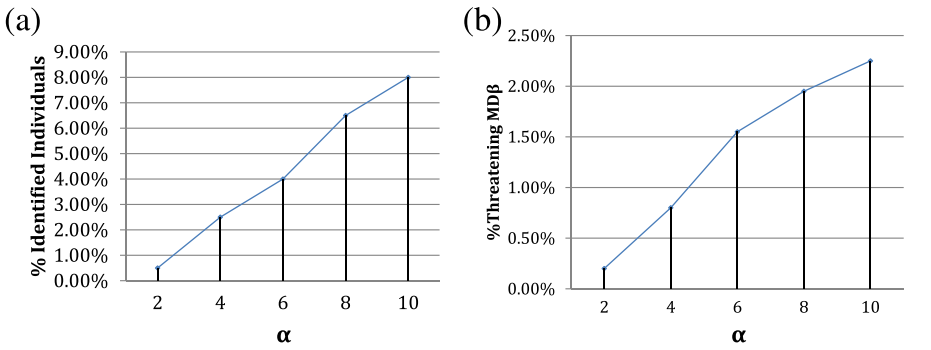
$$Entropy(\mathcal{MD}_u) = -\sum_{a \in \mathcal{A}} Pr(a) log(Pr(a))$$

[13] https://dev.twitter.com/

[14] https://developers.google.com/custom-search/v1/overview

[15] http://phash.org/docs/howto.html

[16] we have omitted the threatening values and objects from our evaluation process



Privacy violation evaluation in terms # of Individuals

Privacy violation evaluation in terms # of threatening MD_{β}

Figure 3 Privacy violation evaluation

where a is the related attribute. We estimate the uncertainty to be: $|Entropy(\mathcal{MD}_u) - Entropy(\mathcal{MD}_u^*)|$ where $\mathcal{MD}_u^*$ is the sanitized individual's multimedia document. The results are shown in Figure 4.

Figure 4 shows that the uncertainty caused by the sanitizing process is relatively small. This uncertainty could get even smaller if sanitizing multimedia objects was approached differently using blurring or pixelizing techniques that preserve the semantic and coherence of images' content. This process is left for a future work.

7.2.3 Evaluating utility

To evaluate the utility of multimedia documents that have been subject to a sanitization process, we sanitized a specific percentage (20, 40, 60 and 80 %) of words and multimedia objects chosen at random from the multimedia document signatures of 100 $\mathcal{MD}_u$. Here, the salient objects which are represented using our multimedia object representation have been

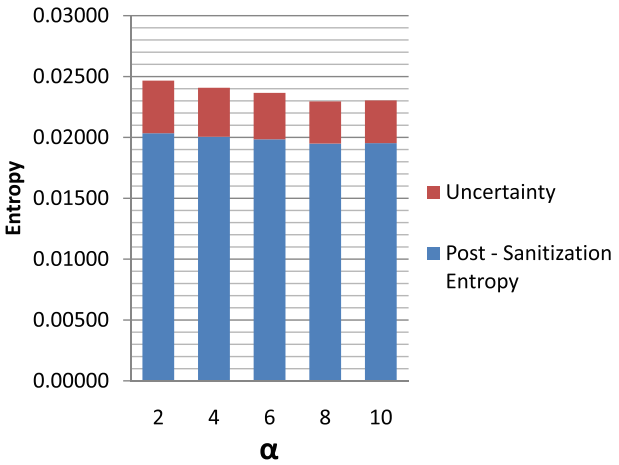


Figure 4 Uncertainty evaluation

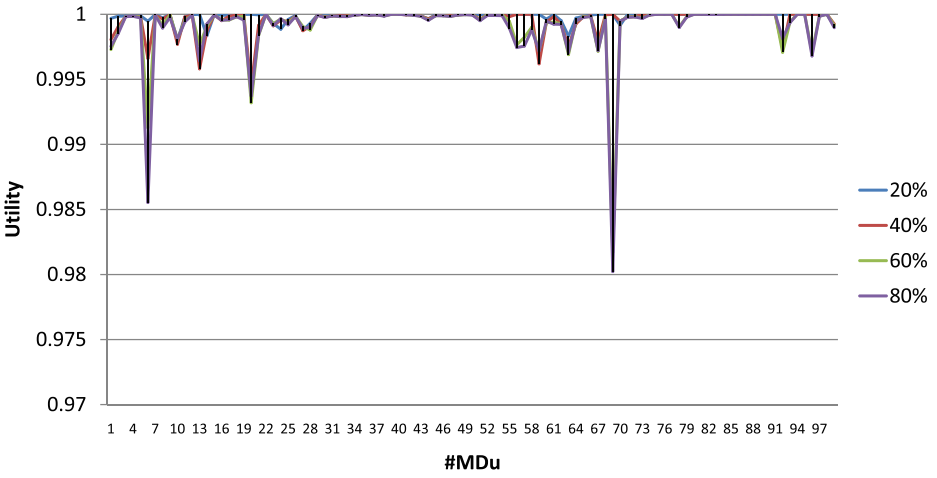


Figure 5 Utility evaluation

sanitized. The resulted utility computed in terms of words and multimedia relevance metrics of each of the multimedia documents is shown in Figure 5.

As one can notice, the trade-off between privacy and utility is explicitly shown in the results where the increased percentage of anonymized words and multimedia objects decreases the utility of the multimedia documents. Nonetheless, such decrease of utility remains bounded by the number of words and multimedia objects contained in the multimedia documents signatures where only their content is subject to sanitization.

7.2.4 Evaluating computational cost

The $\mathcal{MD}^*$-algorithm's time complexity is polynomial and of

$$O(|\sigma_{E_u}(\mathcal{E})| \times (|W^*| + |MO^*|) \times z) \approx O(|\sigma_{E_u}(\mathcal{E})| \times (|W^*| + |MO^*|))$$

where $|\sigma_{E_u}(\mathcal{E})|$ is the number of relevant multimedia documents retrieved from the external source, $|W^*| + |MO^*|$ is the number of sanitized words and multimedia objects from $\mathcal{MD}_u$ and z is the number of attributes used by the selective intersections. This can also be seen experimentally in Figure 4.

The resulting computational time depends on: 1) the conjunctive set of words and/or multimedia objects in E_u that are used to query the external source, 2) the external source from which multimedia documents ($\mathcal{MD}_\beta$) are retrieved (e.g., the Web in our case). This is what we call fetching time which in some cases can be unpredictable as noticed between $\frac{1}{\alpha} = 4$ to 6 where the time to retrieve the individuals' data from the external source has increased (Figure 6).

8 Conclusion

In this paper, we proposed *de*-linkability, a privacy-preserving constraint that ensures the safe outsourcing of multimedia documents to semi-trusted third parties. *de*-linkability addresses the privacy threat in its broader aspect while considering both textual and multimedia content. We provided a sanitizing algorithm to protect against violating content

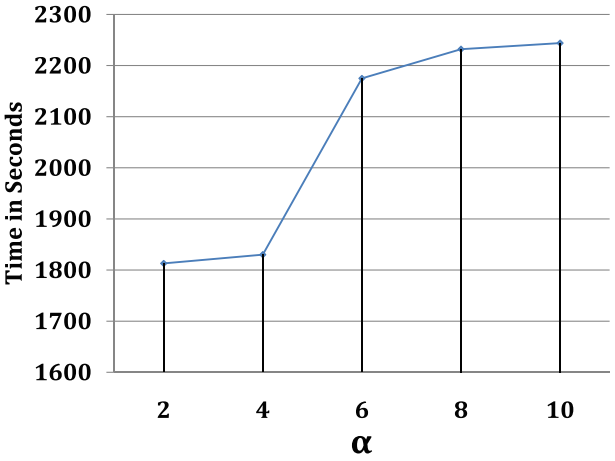


Figure 6 Computational cost evaluation

and preserve at the same time a minimum quality through an adapted sanitization process that takes into consideration the complex nature of multimedia objects. In the near future, we expect to provide more tests to demonstrate the efficiency of *de*-linkability. We also intend to extend our technique to include an in-depth quality assessment and evaluation for both multimedia and textual attributes. As for metadata, we plan to integrate metadata similarity processing to be able to compare multimedia documents semantically.

Acknowledgments This study is funded by the Lebanese CNRS Research Grant Program NCSR project 506 fund 1003. It is also partly funded by the CEDRE research collaboration program, project AO 2011, entitled: Easy Search and Partitioning of Visual Multimedia Data Repositories, jointly funded by the French CNRS (National Center for Scientific Research)

References

1. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: Konstan, J., Conejo, R., Marzo, J., Oliver, N. (eds.) User Modeling, Adaption and Personalization, volume 6787 of Lecture Notes in Computer Science, pages 1–12. Springer Berlin Heidelberg (2011)
2. Al Bouna, B., Chbeir, R., Marrara, S.: A multimedia access control language for virtual and ambient intelligence environments. In: Proceedings of the 2007 ACM workshop on Secure web services, SWS '07, pages 111–120, New York, NY, USA, ACM (2007)
3. Al Bouna, B., Raad, E.J., Elia, C., Chbeir, R., Haraty, R.: De-linkability: A privacy-preserving constraint for safely outsourcing multimedia documents. In: Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems, MEDES '13, pages 68–75, New York, NY, USA, ACM (2013)
4. Bouna, B.A., Chbeir, R., Gabillon, A.: The image protector - a flexible security rule specification toolkit. In: SECRYPT, 345–350 (2011)
5. Boyle, M., Edwards, C., Greenberg, S.: The effects of filtered video on awareness and privacy. In: CSCW, pages 1–10, Philadelphia, Pennsylvania, ACM (2000)
6. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.K.: Efficient techniques for document sanitization. In: Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08, pages 843–852, New York, NY, USA, ACM (2008)

7. Chow, R., Golle, P., Staddon, J.: Detecting privacy leaks using corpus-based association rules. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08, pages 893–901, New York, NY, USA, ACM (2008)
8. Chun, B.T., Bae, Y., Kim, T.-Y.: A method for original image recovery for caption areas in video. In: Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on, volume 2, pages 930–935 (1999)
9. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. Image Process., IEEE Trans. on **13**(9), 1200–1212 (2004)
10. Dwork, C.: Differential privacy. In: ICALP (2), pages 1–12 (2006)
11. Fan, J., Luo, H., Hacid, M.-S., Bertino, E.: A novel approach for privacy-preserving video sharing. In: CIKM, pages 609–616, Bremen, Germany, ACM (2005)
12. Geng, L., You, Y., Wang, Y., Liu, H.: Privacy measures for free text documents: Bridging the gap between theory and practice. In: TrustBus, pages 161–173 (2011)
13. Gessiou, E., Vu, Q.H., Irild, S.: Ioannidis: An information retrieval based method for information leak detection. In: Computer Network Defense (EC2ND), 2011 Seventh European Conference on 33–40 (2011)
14. Huang, J., Ertekin, S., Giles, C.L.: Efficient name disambiguation for large-scale databases. In: PKDD, pages 536–544. Springer-Verlag (2006)
15. Komodakis, N.: Image completion using global optimization. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 1, pages 442–452, june (2006)
16. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Proceedings of the 22Nd International Conference on Data Engineering, ICDE '06, pages 25–, Washington, DC, USA, IEEE Computer Society (2006)
17. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: ICDE, pages 106–115 (2007)
18. Ma, R., Meng, X., Wang, Z.: Preserving privacy on the searchable internet. In: iiWAS, pages 238–245 (2011)
19. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. In: Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on, pages 24–24 (2006)
20. Malin, B.: Trail Re-identification and Unlinkability in Distributed Databases. PhD thesis, Carnegie Mellon University (2006)
21. Martin, N.F., England, J.W.: Mathematical theory of entropy, vol. 12. Cambridge University Press, Cambridge (2011)
22. Mori, J., Matsuo, Y., Ishizuka, M.: Finding user semantics on the web using word co-occurrence information. In: Proceedings of the International Workshop on Personalization on the Semantic Web (PersWeb05) (2005)
23. Nettleton, D.F., Abril, D.: Document sanitization: Measuring search engine information loss and risk of disclosure for the wikileaks cables. In: Domingo-Ferrer, J., Tinnirello, I. (eds.) Privacy in Statistical Databases volume 7556 of Lecture Notes in Computer Science, pages 308–321. Springer Berlin Heidelberg (2012)
24. Patwardhan, K., Sapiro, G., Bertalmio, M.: Video inpainting under constrained camera motion. Image Process., IEEE Trans. on **16**(2), 545–553 (2007)
25. Podesta, J., Pritzker, P., Moniz, E., Holdren, J., Zients, J.: Big data: seizing opportunities, preserving values. Executive Office of the President, The White House Washington, Study (2014)
26. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes 3rd Edition: The Art of Scientific Computing. Cambridge University Press, New York, NY, USA, 3 edition (2007)
27. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
28. Staddon, J., Golle, P., Zimny, B.: Web-based inference detection. In: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium, SS'07, pages 6:1–6:16, Berkeley, CA, USA, USENIX Association (2007)
29. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. J. Uncertain., Fuzziness and Knowl.-Based Syst. **10**(5), 571–588 (2002)
30. Sweeney, L.: k-anonymity: a model for protecting privacy. J. Uncertain., Fuzziness and Knowl.-Based Syst. **10**(5), 557–570 (2002)
31. Vimercati, S., Foresti, S.: Quasi-Identifier. In: van Tilborg, H.C.A., Jajodia, S. (eds.) Encyclopedia of Cryptography and Security, pp. 1010–1011. Springer, US (2011)

32. Wang, L., Jin, H., Yang, R., Gong, M.: Stereoscopic inpainting: Joint color and depth completion from stereo images. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8 (2008)
33. Wexler, Y., Shechtman, E., Irani, M.: Space-time completion of video. Pattern Analysis and Machine Intelligence, IEEE Trans. on **29**(3), 463–476 (2007)
34. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: Proceedings of 32nd International Conference on Very Large Data Bases (VLDB 2006), Seoul, Korea, Sept. 12–15 (2006)