Sanitization's Slippery Slope: The Design and Study of a Text Revision Assistant

Richard Chow
PARC
rchow@parc.com

Ian Oberst[*]
Oregon State University
ian.oberst@gmail.com

Jessica Staddon
PARC
staddon@parc.com

ABSTRACT

For privacy reasons, sensitive content may be revised before it is released. The revision often consists of redaction, that is, the "blacking out" of sensitive words and phrases. Redaction has the side effect of reducing the utility of the content, often so much that the content is no longer useful. Consequently, government agencies and others are increasingly exploring the *revision* of sensitive content as an alternative to redaction that preserves more content utility. We call this practice *sanitization.* In a sanitized document, names might be replaced with pseudonyms and sensitive attributes might be replaced with hypernyms. Sanitization adds to redaction the challenge of determining what words and phrases reduce the sensitivity of content. We have designed and developed a tool to assist users in sanitizing sensitive content. Our tool leverages the Web to automatically identify sensitive words and phrases and quickly evaluates revisions for sensitivity. The tool, however, does not identify all sensitive terms and mistakenly marks some innocuous terms as sensitive. This is unavoidable because of the difficulty of the underlying inference problem and is the main reason we have designed a sanitization *assistant* as opposed to a fully-automated tool. We have conducted a small study of our tool in which users sanitize biographies of celebrities to hide the celebrity's identity both both with and without our tool. The user study suggests that while the tool is very valuable in encouraging users to preserve content utility and can preserve privacy, this usefulness and apparent authoritativeness may lead to a "slippery slope" in which users neglect their own judgment in favor of the tool's.

Categories and Subject Descriptors

H.2.0 [**General**]: Security, integrity and protection.

General Terms

Security, privacy.

Keywords

Sanitization, data loss prevention, inference detection, redaction.

1. INTRODUCTION

Organizations often need to distribute documents containing sensitive information. For example, government bodies must release documents to comply with the Freedom of Information Act (FOIA) and other legislation, and commercial organizations are frequently driven by economic pressure to outsource much of their document processing. Today, a prevalent approach to reconciling the competing goals of information sharing and privacy is *redaction*, that is the blacking-out or otherwise obscuring, of sensitive words in documents prior to their release. While an appealingly straightforward approach to protecting sensitive content, it is not clear that there exists a redaction strategy successfully balancing privacy and information sharing. Consider, for example, the practice of redacting by obscuring the exact set of sensitive words without reflowing the document. With this strategy, the length of the redacted words in conjunction with the font characteristics can often be used to determine the redacted words [7, 11, 10], causing a privacy failure. Alternatively, if the redaction is done more thoroughly (i.e. by redacting longer phrases and by reflowing to hide phrase length) then the utility of the document is severely diminished (see, for example, [18]) thus prohibiting the data analysis needs that often motivate document sharing (see, for example, [16]).

We study *sanitization* of text content as an alternative to redaction that preserves both privacy and information utility. Sanitization is a generalization of redaction increasingly advocated by government agencies [9][1], in which sensitive words need not be replaced with black bars, but instead may be replaced by words that are less sensitive but still convey useful information. Consider, for example, the sentence, "The study subject was born in Panama and currently resides in the 94304 zip code." Releasing the subject's zip code and ethnicity might be a privacy concern, since according to the 2000 U. S Census, there is only a single person with those attributes, however, redacting those attributes results in the following text that has little information utility: "The study

[*] Most of this work was done while this author was an intern at PARC.

[1] Note that this reference advocates sanitization on sensitive content stored internally.

Symposium on Usable Privacy and Security (SOUPS) 2009, July 15–17, 2009, Mountain View, CA USA
.

subject was born in [REDACTED] and currently resides in the [REDACTED] zip code.". Alternatively, the sanitized version: "The study subject was born in Central America and currently resides in Santa Clara County, California.", increases privacy as the attributes "Santa Clara County" and "Central America" match with more than 12,000 people according to the 2000 U. S. Census, and it preserves far more of the information in the original sentence.

In addition to the government applications already mentioned, sanitization may be useful in litigation, financial due diligence and for the end-user who wants to maintain an anonymous blog [19].

While sanitization is an attractive alternative to redaction in terms of information utility, it is also more challenging because revisions must be considered in the context of all the document information *and* information preservation constraints. Consider, for example, a document that states a person resides in Portland, Oregon. If for privacy reasons, residence should not be specified so precisely, we might consider replacing "Portland, Oregon" with simply "Portland" (a popular U. S. city name) or "Oregon". However, if a later sentence, that needs to be retained, refers to the region's "temperate climate", then Oregon is the more privacy-preserving choice, whereas a description of the residence city as "the largest in the state" indicates Portland is a better choice, since both Portland, Maine and Portland, Oregon are the largest in their respective states.

We present a tool that assists the user in sanitizing sensitive text. Given the AI-hard nature of automated sanitization, balancing both data utility and privacy, we do not attempt to completely automate the process (as is the current trend in redaction, [5, 14, 2]). Instead, we take a hybrid approach that leverages automation to help the user assess their revisions and guide them toward potential improvements in privacy, while leaving the actual revisions up to the user. Specifically, we leverage data mining and linguistic parsing to automate the privacy-risk analysis and make suggestions as to how to revise the document.

Our tool is designed to incentivize the user to preserve privacy while retaining as much information in the document as possible. In particular, the software employs the association mining algorithms of [3] to identify terms in the document that are likely to allow a sensitive topic to be inferred, and are thus a privacy risk, and highlights those terms using a color gradient. To discourage the user from broadly redacting terms to reduce privacy risk (and thus diminishing the utility of the document) we provide a scoring mechanism. The user's score increases with revisions that reduce the privacy risk of the document, but decreases with word deletions. Finally, the interface features a guide that points the user to the terms currently in the document with the most privacy risk, and suggestions for potentially less sensitive terms are available automatically through hypernyms recovered from WordNet [21] and Google Directory [8].

We have conducted a small user study of our tool to evaluate its effectiveness as a sanitization assistant. In the study, users were asked to sanitize 2 short biographies to obscure the identity of the biography subject. One biography is sanitized with the tool and one without. We evaluated the privacy-preserving ability of the sanitized documents by asking another set of users (users from Amazon's Mechanical Turk service [12]) to guess the subject of each biography, and we evaluate the utility of the sanitized documents through the score and a manual review of the content changes. Our results indicate the tool can be used to produce documents that better preserve both privacy and utility, provided users regard the tool as a assistant and don't ignore their personal judgment. In particular, we present evidence indicating that when the document topic is less familiar, some users transition from using the tool as an assistant to relying on it as an automated "sanitizer", and tend not to exercise the same judgment they used when sanitizing by hand. These users tend to produce documents that suffer both in terms of privacy and utility. Hence, we believe user expectations around a sanitization tool need to be carefully managed.

Overview. We discuss related work in Section 2 and the design of the tool in Section 3. Section 4 describes the design of our user study and the results. We conclude in Section 5.

2. RELATED WORK

There are a number of commercial products for redaction (see, for example, [5, 14]) as well as a recent research prototype [2][2]. These tools aim to automate redaction and do not support the revision or sanitization of text. In contrast, we take a user-centric approach to sanitization with the software in the role of a guide that is capable of quickly assessing the value of the user's revisions.

In addition, the privacy goals of [2] are quite different. In [2], terms are redacted to ensure a form of k-anonymity [17]. Specifically, terms are only redacted enough to ensure that there are $k-1$ other records in the database with matching attributes. We strive for a more stringent privacy goal in which enough attributes are revised or removed to ensure there is not a strong association with the sensitive topic (i.e. it is not enough that there be strong associations between those attributes and other topics).

Our tool is part of a broader effort in the privacy research community to leverage natural language processing and data mining to assist users in meeting privacy goals. For example, [6] explores the use of natural language technology to help users author better privacy policy rules.

Finally, we note that our tool employs the Web-based association rule mining algorithms in [3] to detect privacy risk. These algorithms use the Web as a proxy for human knowledge, to show what attributes are likely to suggest a sensitive topic to the reader. One such (unsurprising) example, is the rule: US president ⇒ Barack Obama. The algorithms and their use in the tool are both explained in Section 3.1.

3. SANITIZATION TOOL

Our sanitization tool assists the user in understanding privacy risk and making revisions. We give a brief overview of the tool here and discuss the key features in more detail in subsequent sections. Screenshots of the tool are in Figure 1 and Figure 2.

The tool contains a toolbar to the right of the text being sanitized, for entering one or more sensitive topics. The tool operates in two distinct "views". The primary view is the sanitize view, which alerts the user to words or phrases that may allow sensitive topics to be inferred, and provides tools to help edit the document, look-up the context of sensitive

[2] Note that [2] uses the term "sanitization" as a synonym for redaction.

information, and find word replacements that might reduce privacy risk. The plaintext view provides basic free-text editing capabilities. The toolbar allows the user to toggle between the views. In Figure 1 we show the sanitize and plaintext views in the tool using an excerpt from the FBI document [10].

3.1 Conveying Privacy Risk

As difficult as it is to detect privacy breaches, the problem of conveying privacy risk to the user is perhaps even more challenging. It is particularly acute in the case of text content where the combination of seemingly innocuous words may lead to a privacy violation, and revisions of sensitive text may solve one problem while creating others.

Our tool builds on recently developed automated techniques for discovering privacy risk [3] and represents that risk using gradient-based highlighting within the sanitize view. We discover privacy risk by mining *association rules* (see, for example, [1]). In our setting, an association rule is an implication of the form $\{A_i\} \Rightarrow B$, where $\{A_i\}$ are terms in the document and B is a sensitive topic. For example, the introduction describes an association rule with, $A_1 =$"Panama", $A_2 =$"94304" and sensitive topic B equal to an individual's identity. Our software looks for high confidence association rules between the terms in the document and a sensitive topic input by the user using the algorithm in [3]. The confidence, c, of an association $A \Rightarrow B$ is measured as the ratio of the number of Web documents containing both A and B to the number of Web documents containing A.

Specifically, to measure the confidence of an association, $A \Rightarrow B$, we take the following steps:

1. Issue a search engine query: "A", and retrieve the number of returned hits, n_A.

2. Issue a search engine query: "B", and retrieve the number of returned hits, n_B.

3. Issue a search engine query: "A" "B", and retrieve the number of returned hits, $n_{A \wedge B}$.

Based on this, we estimate the confidence of the association $A \Rightarrow B$ using the same mechanism as in [3], namely:

$$\text{Confidence}(A \Rightarrow B) \approx n_{A \wedge B}/n_A$$

Prior to looking for association rules, our tool parses text using the Sentence Detector and Part of Speech (PoS) Tagger available from OpenNLP [13]. It links text based on the sentence structure (as derived from PoS information). Terms A and B are linked for topic T if A and B are linguistically related, e.g. one of the terms describes the other, and

$$\Pr(T|A, B) > \max(\Pr(T|A), \Pr(T|B)).$$

The goal of the linking is to produce association information that is more intelligible and to improve the efficiency of the association mining algorithm [3] by reducing the number of Web queries. As an example of the former, while a user might be surprised to see the suggestion that the word "white" be revised when anonymizing a document about George Bush, they are unlikely to be confused when "white" is linked with "house". In this case, the confidence of {white $\Rightarrow$ George Bush} and {house $\Rightarrow$ George Bush} are both less than the confidence of {white AND house $\Rightarrow$ George Bush}.

The tool then searches for pairs of linked terms that allow the sensitive topic to be inferred using the algorithm described above.

A potential danger of linking is that some sensitive associations that don't involve the entire linked set may be missed. While our experiments indicate this is a rare situation, the tool allows linking to be turned off if the user so desires. In addition, linking allows us to better model human reasoning in some situations. For example, when trying to determine the subject of a sanitized document, a human is unlikely to evaluate whether each individual attribute is closely associated with a potential topic, rather they consider whether all attributes together suggest a potential topic. Linking allows us to accomplish this more efficiently.

Figure 2 shows an extract from an FBI document about Osama Bin Laden that was redacted poorly and promptly unredacted [10].[3] The sensitive terms are highlighted with a color gradient that reflects the degree of sensitivity. The guide button takes the user to the terms the software deems should be revised first. In this example, the guide button has taken the user to the highlighted phrase "holy mosques" (two terms linked together by the tool) because this phrase has a strong association with Bin Laden, especially when paired with another term in the document, "magnate" (i.e. magnate + holy mosques $\Rightarrow$ Bin Laden, with high confidence). The rationale and algorithms behind the guide are discussed more in a subsequent section.

Terms with low privacy risk are highlighted in green. The user can thus get a rough assessment of the privacy risk from the color alone and can hover over the terms for a deeper analysis. For example, by hovering over the "holy mosques", the user learns the precise level of sensitivity of the phrase when it is paired with "magnate". When the user clicks the phrase "holy mosques", an editing window is displayed below, from which they can view a Web page about Bin Laden that uses those terms (see Figure 2).

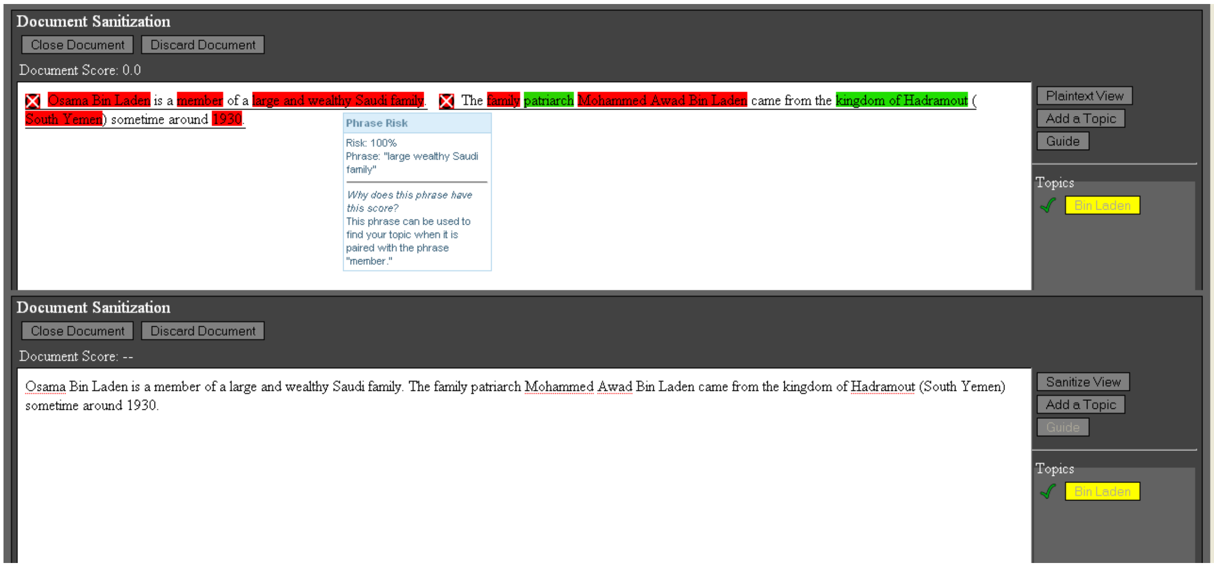
Information is also available about entire sentences. The icon at the beginning of the sentence indicates it can be found in its entirety on the Web (e.g. at the following URL, [10]). Similar to individual words or phrases, clicking on the sentence icon displays an editing window. Our intention with this feature is to draw the user's attention to the fact that the information in such a sentence is exactly present on the Web, rather than just likely present based on high-confidence associations with individual terms in the sentence. This indicates that the sentence may need to be reformulated even if every term appears innocuous.

A potential issue with this approach of using search engines to help discover privacy risk is that information may be leaked to the search engine provider through the nature of the queries. One approach is to create noise in the queries, adopting the approach of TrackMeNot [20]. Another is to periodically download a representative snapshot of the Web and do local queries.

3.2 Encouraging Information Retention Through Scoring

We display a score to the user as he edits a document. The score is intended to provide a game-like aspect to sanitizing. It aims to steer the user to do the right thing in an enjoy-

[3] We replaced "Osama Bin Laden" in the figure with "his" to show how the document might appear during the sanitization process.



Document Score: 0.0

☒ Osama Bin Laden is a member of a large and wealthy Saudi family. ☒ The family patriarch Mohammed Awad Bin Laden came from the kingdom of Hadramout (South Yemen) sometime around 1930.

Phrase Risk

Risk: 100%
Phrase: "large wealthy Saudi family"

Why does this phrase have this score?
This phrase can be used to find your topic when it is paired with the phrase "member."

Document Score: --

Osama Bin Laden is a member of a large and wealthy Saudi family. The family patriarch Mohammed Awad Bin Laden came from the kingdom of Hadramout (South Yemen) sometime around 1930.

Figure 1: The top screenshot shows the application in sanitize view. Words and phrases that allow the sensitive topic (Bin Laden) to be inferred, are highlighted in red. The user can hover over the highlighted text to get more information about the privacy risk associated with the word or phrase. The bottom screenshot shows the same text in plaintext view. In plaintext view the user may revise the sensitive terms.

able way. Consistent with the game-like aspect, the score starts at 0 and moves higher as privacy risk decreases. To encourage information retention, the score also moves lower if information is removed that is unrelated to the topic. The score encourages the model of substituting sensitive terms with other, less sensitive terms.

The score we implemented has three components:

$$\begin{aligned}\text{Score} = {} & (\text{Change in document risk}) \\ & + (\text{Risk removal bonus}) \\ & + (\text{Overall deletion penalty})\end{aligned}$$

The "Change in document risk" is the change in the sum of all term confidences, c_i, in the original and current documents:

$$\text{Change in document risk} = \sum_{\text{orig doc}} c_i - \sum_{\text{current doc}} c_i$$

The "Risk removal bonus" is the sum of $(c_i - 1/10)$ over all terms deleted from the original document :

$$\text{Risk removal bonus} = \sum_{\text{deleted terms}} (c_i - 1/10)$$

Note that the value of 1/10 essentially defines a threshold for the confidence. If the user replaces a term that has confidence less than 1/10, he is penalized. On the other hand, the user is rewarded for replacing terms with confidence greater than 1/10.

The "Overall deletion penalty" encourages retention of terms by subtracting the overall number of terms deleted:

$$\text{Overall deletion penalty} = \min(\text{change in \# terms}, 0)$$

Note that this score only imperfectly enforces the goal of information retention: one can simply delete all terms and replace them with the same number of meaningless 0-confidence terms. It is possible to make a more complex score if this is an issue (for instance, by adding a bonus for retaining low-confidence terms). However, we believe that in most uses of the tool, the user will also have other incentives to retain information.

Another issue with the score is that the tool may simply be wrong and the human right, discouraging the human from doing the right thing.

The score is displayed just above the content window on the left side (see Figure 1).

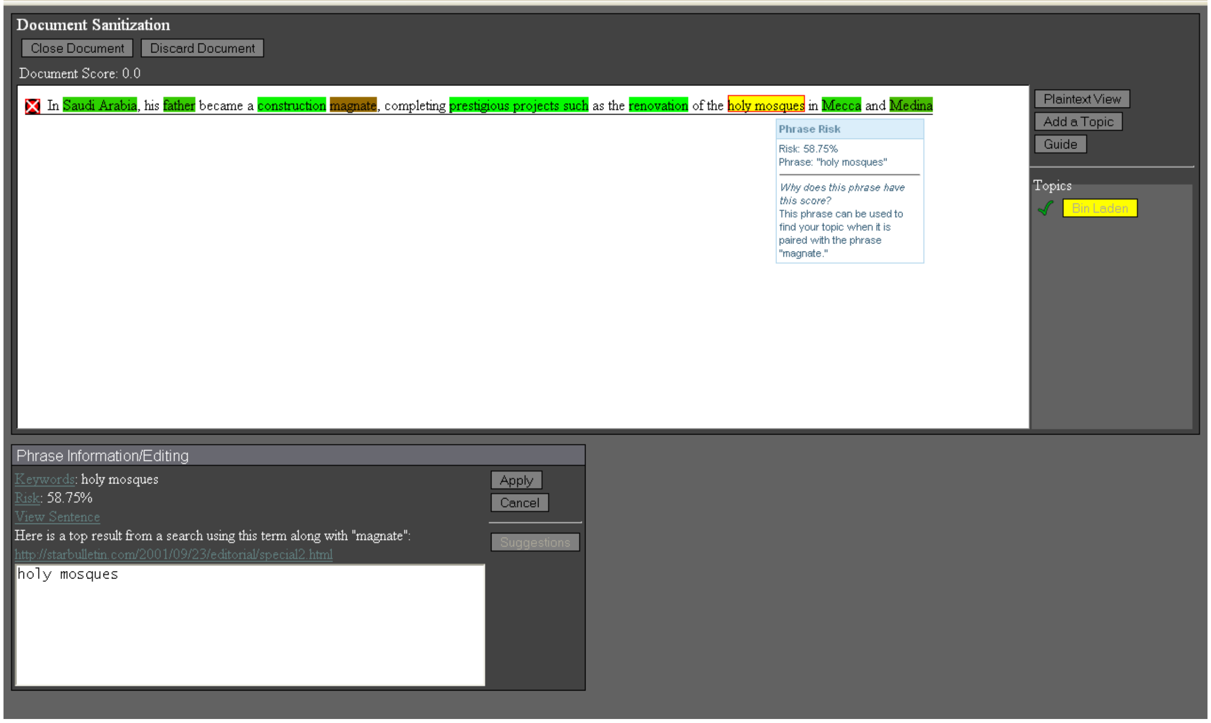
3.3 Making Suggestions to Guide the User

The act of sanitizing a document can be daunting if there are a large number of sensitive terms. An interface filled with red may lead to information overload because the best action is not clear. To mitigate this, we offer the user a "Guide" button in our interface that suggests the current term to work on.

The heuristic behind our guide performs two tasks. First, it attempts to locate sensitive terms in the document that are likely to cause other terms to become more sensitive when they are used together. Second, it tries to select terms that help users achieve their goal of producing a better sanitized document, and thus achieve a higher score.

The guide algorithm calculates a "guide value" for each term. First, the algorithm calculates the difference between the original confidence (before any pairing) of a term, and the confidence it achieves as a result of being paired. This difference is calculated for all other terms in its search space. All the differences are then summed to produce the guide value for a particular term.

Using this methodology, we can rank all information in the document from highest to lowest, with the highest being the most beneficial to remove, both in terms of lowering the risk of the document as well as raising the score of the user. The increase in the user's score is readily apparent. Since the term causes many other terms to have higher risk, the



Document Score: 0.0

In Saudi Arabia, his father became a construction magnate, completing prestigious projects such as the renovation of the holy mosques in Mecca and Medina

Risk: 58.75%
Phrase: "holy mosques"

Why does this phrase have this score?
This phrase can be used to find your topic when it is paired with the phrase "magnate."

Keywords: holy mosques
Risk: 58.75%
View Sentence
Here is a top result from a search using this term along with "magnate":
http://starbulletin.com/2001/09/23/editorial/special2.html

```
holy mosques
```

Figure 2: The guide takes the user to the phrase "holy mosques" first because when paired with "magnate" these terms have the strongest association with the sensitive topic. By clicking on "holy mosques" the user can view more information on its sensitivity and make edits.

term with the highest score will cause the highest reduction in risk possible at that time for sanitizing a single term.

In addition, the tool suggests replacement options for sensitive terms. The user can access these suggestions by clicking on the sensitive nouns and noun phrases. Suggestions are generated through a combination of Google Directory [8] and WordNet [21]. In particular, if a noun is in WordNet, the tool provides nouns listed by WordNet as similar, as suggestions. Proper nouns are looked up in Google Directory and a portion of the listing information is returned to the user. For example, "Tom Cruise" might appear in the category "Actor" which is in turn a part of the category, "Actor and Actresses", etc.. The tool typically excludes the first and last categories (with the reasoning being the former is too specific and the latter too general) and offers the rest as revision suggestions.

4. USER STUDY

There were 12 users in our study, all are associated with PARC (aka Xerox PARC). 9 are computer science researchers, 2 are social science researchers and 1 is an administrator.

In the study, users were asked to sanitize two biographies about actors. One biography they sanitized "by hand" meaning they could not use our tool, but could use the Web if they so desired. For the other biography they used our tool and they had the option of using the Web as well.

The users were given a brief training session before doing any sanitization on their own. In the training session they were asked to imagine they were employees tasked with revising company documents prior to their release to remove sensitive information. If the user was to sanitize by hand first, they were then shown the biography to sanitize as a text file and shown a Web search engine available for use. Alternatively, if they were to use the tool first, they were then walked through the various features of the tool (Section 3) using a biography of the actor Tom Cruise as an example. After the user completed their first sanitization task they were trained for their second task.

After the training, the users were each given 2 short biographies to sanitize (see Table 1), one of actor Harrison Ford and one of actor Steve Buscemi. The assignment order in which the subjects sanitized (tool vs. by hand) and the choice of biographies, were randomized. The assignments for each user are shown in Table 2.

The user's desktop was video recorded during both sanitization tasks.

When recruiting participants we requested 1 hour of their time for the study. However, we did not restrict the time the users spent on any part of the study. Almost all the subjects sanitized each biography in less than 30 minutes. The exact timings are in Table 2.

After users completed both sanitizations they were briefly questioned. We list the questions and summarize the answers in Table 3.

As discussed in the introduction, we measure the success of sanitization against 2 metrics: privacy and utility. Privacy is measured as the ability of an "adversary" to identify the topic of a sanitized biography (i.e. either Harrison Ford or Steve Buscemi). To measure this, we used Amazon's Mechanical Turk service [12]. Each of 119 "Turkers" was given 2 randomly selected biographies sanitized by the study participants, one that was sanitized by hand and one that was

Harrison Ford	**Steve Buscemi**
Harrison Ford was born in July of 1942 and grew up in the Midwest. He struggled as a student, but developed a passion for acting early in his college career. He may have been inspired in this direction by his parents, as both were former actors. Ford attended Ripon college in Wisconsin but failed to graduate; some reports indicate he was expelled due to poor grades. After college he moved to Los Angeles supporting himself from time to time as a carpenter. For many years he played bit parts in movies and television shows. His breakthrough role came in 1977 with Star Wars, and he has since gone on to star in the top-grossing Indiana Jones film series as well as "Patriot Games", "The Fugitive" and "Air Force One" earning him a strong fan-base as an action movie star. His acting accolades include an Oscar for "Witness" and a star on the Hollywood Walk of Fame.	Steve Buscemi was born in New York in December of 1957. Buscemi's interest in acting stems back to his high school days during which he was part of the school drama troupe. After a brief period at a local community college, he attended the prestigious Lee Strasberg Institute in Manhattan. The Strasberg institute is the former school of a number of popular Hollywood actors included Angelina Jolie and Robert De Niro. Before his acting career took off, Buscemi worked as a firefighter in New York. Buscemi's movie career began in 1988 with "Call Me" and has including numerous supporting roles that have garnered rave reviews. He is well-known for a string of 6 movies made with Joel and Ethan Coen, including "Millers Crossing", "Barton Fink" and "Fargo". Other notable performances include Tarantino's debut feature "Reservoir Dogs" and Buscemi's recurring role on the acclaimed HBO series, "The Sporanos". Buscemi also has directing credits including episodes of "The Sopranos" and HBO's "Oz".

Table 1: Users were given each of the above biographies and asked to sanitize one with the tool and one without the tool but with optional use of the Web. We term the latter approach, sanitization "by hand". The order in which the user was asked to sanitize (by tool or by hand) and the biography on which they used each approach, were randomly selected. 6 users sanitized the Harrison Ford biography by hand, and the Steve Buscemi with the tool, and 6 used the opposite approach on those biographies. Note that the typo, "Sporanos", did not substantially effect the risk assessment because Yahoo! by default corrected for the misspelling.

User	First Bio	Method	Time (Min:Sec)	Score	Turker Success Rate	Second Bio	Method (Min:Sec)	Time	Score	Turker Success Rate
1	Harrison Ford	By Hand	18:00	NA	.87 (13/15)	Steve Buscemi	By Tool	24:57	40.8	.18 (2/11)
2	Harrison Ford	By Tool	13:02	26.95	.8 (8/10)	Steve Buscemi	By Hand	8:06	NA	.08 (1/13)
3	Steve Buscemi	By Tool	31:28	36.91	.57 (4/7)	Harrison Ford	By Hand	15:23	NA	1 (13/13)
4	Steve Buscemi	By Hand	23:40	NA	0 (0/11)	Harrison Ford	By Tool	30:21	35.01	.56 (5/9)
5	Steve Buscemi	By Hand	21:19	NA	.11 (1/9)	Harrison Ford	By Tool	27:56	11.74	.77 (10/13)
6	Steve Buscemi	By Tool	30:21	19.48	1 (4/4)	Harrison Ford	By Hand	9:47	NA	.57 (4/7)
7	Harrison Ford	By Tool	15:35	21.79	.57 (4/7)	Steve Buscemi	By Hand	Missing	NA	.1 (1/10)
8	Steve Buscemi	By Tool	25:20	38.72	.67 (4/6)	Harrison Ford	By Hand	13:52	NA	.23 (3/13)
9	Steve Buscemi	By Hand	25:38	NA	.31 (4/13)	Harrison Ford	By Tool	30:25	14.09	.8 (4/5)
10	Steve Buscemi	By Tool	9:09	22.02	.54 (7/13)	Harrison Ford	By Hand	9:16	NA	.55 (6/11)
11	Harrison Ford	By Hand	9:05	NA	.91 (10/11)	Steve Buscemi	By Tool	9:54	28.12	0 (0/12)
12	Harrison Ford	By Tool	14:43	27.13	.33 (3/9)	Steve Buscemi	By Hand	31:35	NA	0 (0/9)

Table 2: The sanitization tasks are listed for all users and their quantitative attributes (completion time, score, and Turker success rate) are given. In parentheses following each Turker success rate is the number of Turker who guessed correctly out of the total Turkers who were given that biography. The video of user 7 sanitizing by hand was damaged, hence we do not have the completion time for that user.

Post-Study Question	Answer Summary
Were you familiar with Harrison Ford prior to the study?	Yes (10), No (2)
Were you familiar with Steve Buscemi prior to the study?	Yes (2), No (10)
Were there times that you disagreed with the software? For example, were there words that you thought were sensitive but that weren't highlighted in red or vice versa?	I found highlighted text that wasn't sensitive (5) I found sensitive text that wasn't highlighted (4)
Would you want to use the software for actual document sanitization?	Yes (7), No (1) Unsure (4)
Do you have any suggestions for us?	Improve text suggestions (4) Explain score more and make previous scores viewable (3) Explain what types of information are ok to leave in (e.g. profession)(2)

Table 3: The questions from the post-study interview. The right column summarizes the responses; the number of people who gave a certain response follows the response in parentheses.

sanitized with our tool. They were asked to use any resources (including the Web) to identify the subject of the bio and to explain their answers; some explained their answers by listings the attributes they found most identifying.

Figure 3 shows a sample Turker task. Turkers were paid between \$.02 and \$.1 based on the amount of uptake we were getting on the tasks. The percentage of Turkers able to correctly identify the subject from a given sanitized biography is in Table 2 (see "Turker Success Rate").

Content utility is difficult to gauge because it depends on context. That is, the context in which a document is used often indicates what information is most necessary in the document, and this in turn should influence the sanitization strategy. Because of this, we don't attempt to represent utility with a single numerical value, rather we consider 2 metrics: the score given to a document sanitized with the tool and a summary of edits to each biography organized by category. We describe each of these metrics in turn.

Recall that the score encourages users to preserve content utility by penalizing them for deleting terms. During training, users were encouraged to get as high a score as they could. The score was not perfectly designed, indeed it can be artificially inflated by "padding" the sanitized biography with words that aren't closely associated with the subject (and may even be nonsensical). Only 2 users can be said to have taken advantage of this by inserting information that was both inaccurate and not sensitive (i.e. not closely associated with the biography subject). In both cases, these users inserted movie titles for nonexistent movies. The first such user received a score of 28.12 on the Harrison Ford biography, where the average score across all users was 23.66 and the second user received a score of 13.38 on the Steve Buscemi biography, where the average score was 31.

As another gauge of utility, we decomposed the biographies into 7 categories of information that together cover most of the content of the biographies: name, date of birth, hometown, alma mater, childhood experience, previous careers and movie/tv roles. We manually reviewed the 24 sanitized biographies to determine in which of the categories the user had modified information (if any). We list these categories together with their rough risk level (low = approximately 30% risk and lower, medium = approximately 50% risk, high = approximately 80% risk and higher) as assessed by the tool, in Table 4.[4] Recall that because semantically related terms are sometimes leaked in order to measure risk, some surprising terms can be highlighted as high risk. For example, in the Harrison Ford biography, "Witness" (a term with many uses) has risk 35% because when paired with "Patriot Games" it is associated with Harrison Ford.

4.1 Results

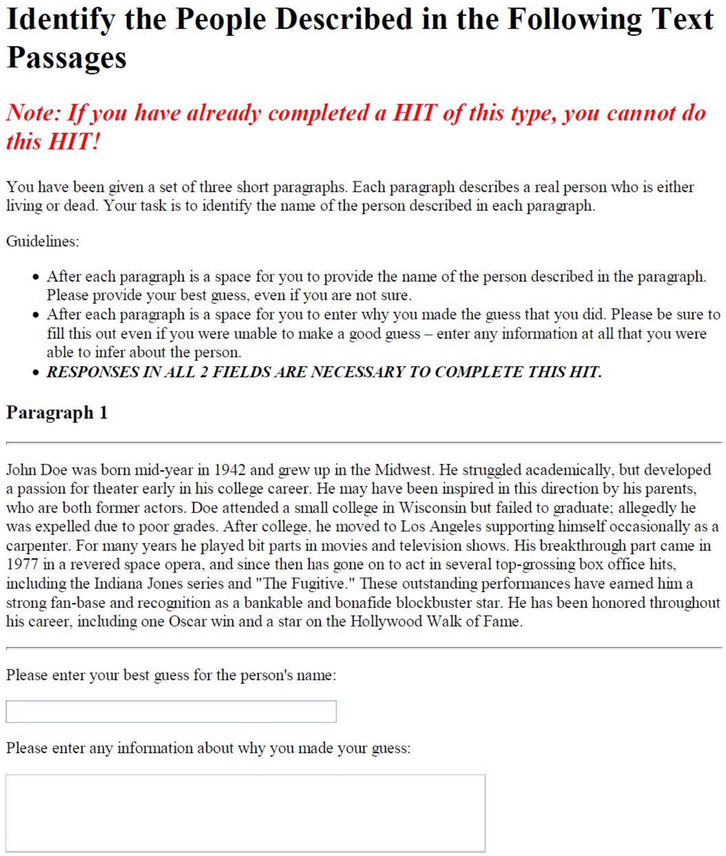
There are many interesting issues to analyze in a study of this kind, including for example, the correlation between user demographics or Web familiarity with sanitization success. Given the scale of our study we can only speak to a small number of the interesting questions and we can't provide overwhelming evidence to validate any conjecture. Rather, we view our study as providing indications of interesting effects that merit a more focused and extensive study.

We organize our observations into the areas in which the tool worked, the areas in which it didn't work, and data demonstrating the apparent difference in sanitization strategy when sanitizing with and without the tool found amongst some users. In particular, we found that users are often inconsistent in their judgment of what type of information is likely to be identifying when sanitizing an unfamiliar subject, seemingly leaning on the tool's judgment more when the subject is unfamiliar. This indicates that designing a sanitization tool that users consistently use as an *assistant*, rather than a solution provider, may be a considerable research challenge, as the trust users place in the tool may vary with the subject matter.

TOOL SUCCESSES. The highlighting feature made it difficult for users to miss explicitly identifying information like parts of the biography subject's name. Indeed, the only person who failed to remove all mentions of a biography subject's name was someone sanitizing by hand. User feedback on the highlighting feature was uniformly positive in terms of ensuring they wouldn't miss content to consider for revision. One user said, "[when sanitizing by hand] I noticed I would occasionally skip a line or something like skip a sentence and I needed to go back and make sure I got all those. With having them highlighted, that never happened to me.".

As intended, the tool lessened the need for manual search engine queries by the users. Of the 11 users for whom we have video of their manual sanitization, 8 users made search

[4] There was some variation in risk assessments across users for the same biography because of variations in search engine results. What we provide in the table is an average risk estimate.



Identify the People Described in the Following Text Passages

Note: If you have already completed a HIT of this type, you cannot do this HIT!

You have been given a set of three short paragraphs. Each paragraph describes a real person who is either living or dead. Your task is to identify the name of the person described in each paragraph.

Guidelines:

- After each paragraph is a space for you to provide the name of the person described in the paragraph. Please provide your best guess, even if you are not sure.
- After each paragraph is a space for you to enter why you made the guess that you did. Please be sure to fill this out even if you were unable to make a good guess – enter any information at all that you were able to infer about the person.
- ***RESPONSES IN ALL 2 FIELDS ARE NECESSARY TO COMPLETE THIS HIT.***

Paragraph 1

John Doe was born mid-year in 1942 and grew up in the Midwest. He struggled academically, but developed a passion for theater early in his college career. He may have been inspired in this direction by his parents, who are both former actors. Doe attended a small college in Wisconsin but failed to graduate; allegedly he was expelled due to poor grades. After college, he moved to Los Angeles supporting himself occasionally as a carpenter. For many years he played bit parts in movies and television shows. His breakthrough part came in 1977 in a revered space opera, and since then has gone on to act in several top-grossing box office hits, including the Indiana Jones series and "The Fugitive." These outstanding performances have earned him a strong fan-base and recognition as a bankable and bonafide blockbuster star. He has been honored throughout his career, including one Oscar win and a star on the Hollywood Walk of Fame.

Please enter your best guess for the person's name:

Please enter any information about why you made your guess:

Figure 3: The first part of an example HIT available to the Mechanical Turk users. By scrolling down, the user could view another paragraph with an identical set of questions.

engine queries, with an average number of queries of 6.5. Whereas, with the tool, only 3 users made manual search engine queries. In addition, the queries users did make may have been more effective, when using the tool. In particular, there is a correlation of $-.44$ between the use of manual Web queries with the tool and Turker success rate (the correlation is negative because queries tend to lead to a better sanitization thus reducing Turker success) and a correlation of only $-.34$ between the use of manual Web queries and Turker success rate when sanitizing by hand.

Utility of the sanitized content appears higher with the tool. In particular, with respect to the 7 content categories listed in Table 4, 7 users revised content in more categories when sanitizing by hand than when sanitizing with the tool; 3 users revised content in the same number of categories and 2 revised content in fewer categories when revising by hand. As an example of the change in behavior, 83% of users revised information about the biography subject's previous career when sanitizing by hand, but only 42% did so when using the tool. Similarly, *all* the users revised information about movie and TV roles when sanitizing by hand, as opposed to 83% when using the tool.

Tool Shortcomings. Perhaps the most significant shortcoming is the Turker success rate. When sanitizing the Harrison Ford biography, a celebrity familiar to almost all the users in the study and probably almost all Turkers, the tool provided little advantage. Turkers had an average success rate of 70% on Harrison Ford biographies that were sanitized by hand, and an average success rate of 64.15% on biographies sanitized with the tool. More surprisingly, the average success on Steve Buscemi biographies sanitized by hand was 10.8% in comparison with 39.6% when sanitized with the tool. One reason for this is that, as an early-stage research prototype, the tool would benefit from far more engineering. In particular, the tool's suggestion mechanism, while popular, only occasionally yielded useful results. Users requested a total of 47 word or phrase suggestions, but only accepted 11 of them (23%), and almost half of the users who requested suggestions did not use any. In addition, only 5 of the 13 users made use of the Guide feature, that sought to simplify the process by stepping the user through the document in order of risk of words and phrases. Rather most users preferred to work on the document linearly, pausing to edit portions that made a threshold risk level (typically text highlighted in a shade of red).

While these engineering shortcomings are significant, we don't think they are the entire explanation. Understandably, users appear to rely more on the tool when sanitizing a less familiar topic, Steve Buscemi, to the point that their judgment of attribute sensitivity was inconsistent between the 2 sanitization tasks (by hand and with the tool). We

NAME	DATE OF BIRTH	HOMETOWN	ALMA MATER	CHILDHOOD	PREVIOUS CAREER	MOVIES,TV
Harrison Ford	**July, 1942**	Midwest	Ripon College	*Poor Student*	**Carpenter**	**Indiana Jones**, *Star Wars*, **Patriot Games**, *The Fugitive*, *Air Force One*, Witness
6/6 BH 6/6 BT	5/6 BH 6/6 BT	0/6 BH 1/6 BT	6/6 BH 4/6 BT	1/6 BH 1/6 BT	4/6 BH 4/6 BT	6/6 BH 6/6 BT
Steve Buscemi	**December** 1957	New York	**Strasberg Institute**	High School Drama Troupe	Firefighter	*Millers Crossing*, *Barton Fink*, Fargo, Call Me, *Reservoir Dogs*, *Sopranos*, Oz
6/6 BH 6/6 BT	6/6 BH 5/6 BT	3/6 BH 0/6 BT	6/6 BH 6/6 BT	0/6 BH 0/6 BT	6/6 BH 1/6 BT	6/6 BH 4/6 BT

Table 4: A summary of the attributes for each biography organized by category. A bolded attribute is one the tool considered high risk, an italicized attribute is one the tool considered medium risk, and other attributes had low or no risk. The second and fourth rows of the table show the fraction of users who revised content in that category when sanitizing By Hand (BH) and By Tool (BT).

	Ave. No. Turker Success Rate, Harrison Ford	**Ave. No. Turker Success Rate, Steve Buscemi**	**Ave. No. Categories Revised**	**Ave. No. Score, Harrison Ford**	**Ave. No. Score, Steve Buscemi**
Consistent Users with Tool	55.6	9.1	4.3	35	34.5
Consistent Users By Hand	88.8	0	4.3	NA	NA
Inconsistent Users with Tool	71.35	73.5	4.2	17.6	26.8
Inconsistent Users By Hand	44.9	17.3	5.2	NA	NA
All Users (with Tool or By Hand)	**66.3**	**29.6**	**4.6**	**23.66**	**31**

Table 5: A summary of how the users with the most consistent sanitization strategy perform when using the tool and without, in contrast to the users with the least consistent sanitization strategy. The most consistent users (users 1, 4 and 11) edited exactly the same categories when using the tool and with hand. The least consistent users (users 5-10) each behaved differently (edited or didn't) on 2 of the 5 categories. "NA" indicates that no score was calculated for biographies sanitized by hand.

discuss this in more detail below.

Evidence of the Slippery Slope. Users were remarkably consistent when sanitizing a familiar topic, Harrison Ford. [5] For example, all users revised information about his college when sanitizing by hand or by tool even though the tool marked Ripon College as having no risk.

However, when sanitizing a less familiar subject, Steve Buscemi, there was a noticeable shift in behavior. For example, the attribute of "Fargo" was considered by all subjects sanitizing by hand to be identifying of Steve Buscemi and was removed, whereas 25% of the users (3/12) left this information in when using the tool, perhaps because of the low risk assigned by the tool to this attribute. Similarly, the attribute of "firefighter" was also considered identifying by *all* subjects sanitizing by hand, but only a *single* user revised it when using the tool, most likely because the tool gave the attribute low risk. It appears that the users sanitizing by hand were right as the mention of Fargo and Firefighter were the top 2 pieces of evidence cited by the Turkers for the answer of Steve Buscemi.

We summarize the other attributes and the different treatment they experienced when the subject used the tool or sanitized by hand in Table 4.

The users who exercised a consistent sanitization strategy (as evidenced by the categories they revised) whether sanitizing with the tool or by hand, generally produced more privacy-preserving biographies. Specifically, the 3 most consistent users created sanitized documents with the tool that were harder on average for the Turkers to guess. In particular, users 1 and 11 had Turker success rates of 18.18% and 0%, respectively, on a biography topic (Steve Buscemi) with average Turker success rate 29.6% overall, and user 4 had a success rate of 55.6% on a biography topic (Harrison Ford) with an average Turker success rate of 66.3% overall. In contrast, out of the 6 least consistent users, 5 created sanitized biographies that were easier for the Turkers to guess on average. In particular, the Harrison Ford biographies of the inconsistent users were on average 5 percentage points easier to guess and the Steve Buscemi biographies were on average 44 percentage points easier to guess.

Similarly, the users with the most consistent sanitization strategy scored higher on average, perhaps indicating their biographies also preserved more utility. Specifically, they have an average score of 34.5 on the Steve Buscemi biography (which has an average score overall of 31) and an average score of 35 on the Harrison Ford biography (which has an average score of 23.66 overall. The least consistent users averaged 17.6 on Harrison Ford and 26.8 on Steve Buscemi, both less than average.

We summarize these data in Table 5. The groups of users who appear to do best at sanitization are the consistent users when using the tool, and the inconsistent users when sanitizing by hand, however, it looks quite likely that the latter group preserves less biography utility.

At least anecdotally, the users who expressed views of the tool consistent with it being an *assistant* as opposed to an automated sanitization agent, did exercise their own judgment both when using the tool and when not, as evidenced by revision consistency in the sanitization tasks. For example, user 11 (one of the 3 most consistent users) said, "I would use it just I like I use Google. Like, it's a kind of a reference of how well. Do I trust it 100%? No. Do I trust Google? No. But I use it somehow...it's useful."

In contrast, the users who indicated the tool knew more about sanitization than they, were much less consistent. For example, one such user said this about the tool, "The tool actually taught me to do sanitization."

5. CONCLUSION

We have presented an early prototype of a text sanitization assistant. The tool employs Web-based data mining to estimate privacy risk, online directories to suggest more general phrasing to lessen privacy risk and a scoring mechanism to encourage the user to retain as much content as possible while reducing privacy risk. A small user study indicates the tool has promise in terms of saving work at the user end and in improving the utility and privacy of the sanitized document. However, our study also indicates that the trust users put in the tool may vary based on the user's familiarity with the subject matter. This is problematic as users seem to create more privacy-preserving sanitizations when they leverage their own judgment as well as the tool's, and it is also difficult to design for as this change seemingly depends on user background knowledge. An interesting research challenge is to confirm this effect with a larger study and understand its impact on other security applications.

Acknowledgments

The authors thank the anonymous reviewers for their thoughtful comments on this paper. The authors gratefully acknowledge the support of Xerox for this research.

6. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1994.

[2] V. Chakaravarthy, H. Gupta, P. Roy and M. Mohani. Efficient techniques for document sanitization. *CIKM 2008.*

[3] R. Chow, P. Golle and J. Staddon. Detecting privacy leaks with corpus-based association rules. *KDD 2008.*

[4] K. Crawford. Have a blog, lose your job? CNN/Money. February 15, 2005.

[5] IntelliDact. CSI Computing Systems Innovations. http://www.csisoft.com

[6] C. Karat, J. Karat, C. Brodie and J. Feng. Evaluating interfaces for privacy policy rule authoring. *CHI 2006.*

[7] D. Lopresti and A. Spitz. Information leakage through document redaction: attacks and countermeasures. *Proceedings of Document Recognition and Retrieval XII.* January 2005.

[8] Google Directory. http://www.google.com/dirhp

[9] C. Johnson, III. Memorandum M-07-16, "Safeguarding against and responding to the breach of personally identifiable information". FAQ. May 22,2007.

[10] Judicial Watch. FBI protects Osama bin Laden's "Right to Privacy" in document release. April 20, 2005. http://www.judicialwatch.org/printer_5286.shtml

[5] Indeed, this familiarity is likely what made sanitization of the Harrison Ford biography so difficult with respect to thwarting the Turkers. With very little information, Harrison Ford was a natural guess by the Turkers whether or not such a guess was supported by the text.

[11] J. Markoff. Researchers develop computer techniques to bring blacked-out words to light. *The New York Times*. May 10, 2004.

[12] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome

[13] OpenNLP. http://opennlp.sourceforge.net/

[14] RapidRedact. http://www.rapidredact.com/

[15] S. Shane. Spies do a huge volume of work in invisible ink. *The New York Times*. October 28, 2007.

[16] B. Sullivan. California data leak raises questions. Experts wonder: Why do agencies share SSNs? *MSNBC*. October 27, 2004.

[17] L. Sweeney. K-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[18] V. Plame Wilson. *Fair Game: My life as a spy, my betrayal by the White House*. Simon and Schuster, 2007.

[19] A. Witt. Blog Interrupted. The Washington Post. August 15, 2004.

[20] TrackMeNot. http://mrl.nyu.edu/ dhowe/trackmenot/

[21] WordNet. http://wordnet.princeton.edu