Detecting Unintentional Information Leakage in Social Media News Comments

Inbal Yahav
Graduate School of Business
Bar-Ilan University
Inbal.Yahav@biu.ac.il

David G. Schwartz
Graduate School of Business
Bar-Ilan University
David.Schwartz@biu.ac.il

Gahl Silverman
Dept. of Interdisciplinary Studies
Bar-Ilan University
galsilverman@gmail.com

Abstract

This paper is concerned with unintentional information leakage (UIL) through social networks, and in particular, Facebook. Organizations often use forms of self-censorship in order to maintain security. Non-identification of individuals, products, or places is seen as a sufficient means of information protection. A prime example is the replacement of a name with a supposedly non-identifying initial. This has traditionally been effective in obfuscating the identity of military personnel, protected witnesses, minors, victims or suspects who need to be granted a level of protection through anonymity. We challenge the effectiveness of this form of censorship in light of current uses and ongoing developments in Social Networks showing that name-obfuscation mandated by court or military order can be systematically compromised through the unintentional actions of public social network commenters. We propose a qualitative method for recognition and characterization of UIL followed by a quantitative study that automatically detects UIL comments.

Keywords: Unintentional information leakage; online news; comments; censorship; privacy; text mining; social networks; social media.

1. Introduction

Information leakage is defined as the accidental or unintentional distribution of private or sensitive data to an unauthorized entity. Sensitive data in organizations include varied kinds of information. Information leakage and data misuse are considered an emerging security threats to organizations, as the number of leakage incidents and the cost they inflict continues to increase, whether caused by malicious intent or by an inadvertent mistake. Such leakage can occur in many forms and in any place [1].

Information leaks are an important concern of many organizations in today's era of social networks (SNS). Organizations have limited control over their employees' activity in public networks, and even less than that, on the activities of their friends and relatives and of the public at large. The latter two groups, on the other hand, often have limited understanding on what public information sharing is appropriate, and what is not. Many organizations, including businesses, the military and the courts, use forms of self-censorship in order to maintain security [2]–[4]. Non-identification of individuals, products, or places is commonly seen as a sufficient means of information protection. We contend that in the age of social networking and social media such non-identification is ineffective as a security measure. In fact, as we will show, the mere release of seemingly general information, and the discourses it arouses, can quickly lead to the exposure of facts that the releasing party intended to remain confidential. We seek to study and characterize such leaks in order to develop methods for their identification and possible prevention.

Traditional information leakage prevention technologies are mainly based on a physical domain, which is partitioned into different security domains according to data protection requirements and limits the data flow between security domains through firewall, encryption and terminal control [5]. Such technologies are irrelevant when the domain is a social network that supports private use, such as Facebook or Twitter.

In this paper, we focus on press releases – an established method of communications between an organization and the public. Press releases, which today are de facto 'public releases' not limited to members of the press, can be characterized by (a) the information that they include and equally if not more important (b) the information that they withhold. The choice of an organization to withhold information from a press release stems from the need to protect that information from a wide range of stakeholders including competitors, suppliers, customers, government authorities, adversaries or hostile groups.

Studies of internet users in the United States have shown that over 60% of Americans are consumers of online news articles, and a full 25% of internet users have posted a comment to an online news article [6]. These comments present a massive dynamic corpus of text to be studied for unintentional information leakage.

This paper has two main objectives: (1) to identify and qualify the nature of the Unintentional Information Leakage (UIL) problem, and (2) to generate an automated UIL detection and prevention system.

In the remainder of the paper we discuss the case study we analyze, the data collected, the qualitative learning phase, and our initial attempts to automatically detect UIL comments. We conclude and discuss future work.

1.1 Background

Previous works shows that active SNS users share large amounts of personal information - a phenomenon which has led to the creation of a treasure trove of data for many entities, from marketers and spammers to employers and intelligence agencies, and become a serious privacy concern. Previous works also addressed many aspects of privacy in SNS such as characterizing potential privacy leakage, possible ways for inferring sensitive private information, and appropriateness of default privacy settings; and in contrast, that by sending out friend requests to unknown other users, SNS users are willing to let a stranger, possibly an adversary, into their social network, thus granting their access to the users' personal information and to some extent to those of their friends [7].

SNS are considered as organizations' weak point. One of the main characteristics of attacks through SNS is that they need not be technologically sophisticated to maximize effect. For example, a simple search on the Linked-in social network could reveal an organization's IT systems manager including all his personal details (name, duty, e-mail, phone number, social circles and friends, picture, etc., and mark him as a target [8].

This is an example of "Personally identifiable information" (PII), which is defined as information used to distinguish or trace an individual's identity either alone or when combined with other information that is linkable to a specific individual. The ability to link PII and combine it with other information falls into the scope of "leakage" as we have defined it. There are four types of PII leakage: (1) Transmission of the SNS identifier to third-party servers from the SNS; (2) Transmission of the SNS identifier to third-party servers via popular external applications; (3) Transmission of specific pieces of PII to third-party servers; (4) Linking of PII leakage within, across, and beyond SNS [9].

Social networks are increasingly used to generate conversations among people about news stories, with many news media finding that the volume of reader comments on a story posted on Facebook can exceed comments posted on the news organization's website. Moving news article commenting to Facebook is also being driven by the desire to reduce commenter anonymity and increase comment quality [10], [11].

2. Studying Press Releases

Our initial focus is on censored Israeli military press releases that are published on public news pages in Facebook (FB). Censorship in our study is the replacement of a name with a supposedly non-identifying initial (e.g. 'Corporal S.'). Information leakage is detected in the comments published by private users. In each article studied, the organization has attempted to protect the identity of an individual by referring to them by first initial, and this anonymity can be compromised through the social network of commenting parties.

The example in Figure 1 illustrates the type of data we study. The headline of the news article, as it appears on the FB page of a network news service, is: "*The Navy is satisfied with the appointment Colonel G. as the Commander of Flotilla 13: 'The natural choice, he has a unique character'*". Flotilla 13 is a naval Special Forces unit and personnel names are commonly withheld from the public. The second commenter, as a response writes: "*Congrats to G. I know him personally. Good choice, good luck*". Given the readily available identity of that commenter it only takes a few clicks, to find Colonel G., who is a Facebook friend of this commenter. We therefore treat this comment and other similar comments as *Unintentional Information Leakage* (UIL).

Figure 1: Information leakage through a comment on FB

(Hebrew original, translated by authors)

2.1 Data

The data were crawled from 37 FB news organization pages, covering the period of January 2012 and December 2013. A total of 325,527 press items were collected during a 4 day period in January 2014 to detect a dataset of censored articles meeting our criteria. All comments

from each matching article were collected. The data contain 50 censored press articles, found across 15 of the crawled sites, with a total of 3582 comments.

It should be noted that the data requires a massive qualitative (expert-based) analysis, which limits our ability to study large quantities of data at first. In the future, once we develop an algorithm that automatically classifies comments as UIL or non-UIL, we will extend our scope and enrich our data.

3. Qualitative Study

The qualitative data analysis followed a 2-stage methodology:

A: Comment Classification – applying discourse analysis and semiotic analysis. Discourse analysis attempts to uncover "how the socially produced ideas and objects that populate the world were created in the first place and how they are maintained and held in place over time" [12]. It aims to reveal the means in which social realities are produced. Discourse analysis can deal with linguistic or written discourse, drawn from of several kinds of media [13].

Semiotic analysis is "the study of signs, sign systems and their meanings" [13]. Semiotics "places particular importance on exploring the deeper meaning" of the data. A semiotics approach attempts to reveal the processes of making meaning and "how signs are designed to have an effect upon actual and prospective consumers of those signs" [14]. Semiotics is suitable for analyzing signs in our everyday life. Hence, it can be applied in not only documentary data analysis, but also to other data types primarily because of its "commitment to treating phenomena as texts" [15].

B: Lexical Identifier Mapping – applying content analysis. Content analysis is a research method for categorization and systematic encoding of text, that allows exploring a large amount of textual information in order to find trends and patterns of use of words, the frequency of words, their relationship, and the structure and discourse of the media using words [13], [16].

By applying discourse analysis and semiotic analysis followed by lexical identifier mapping, the comments were first classified into one of three main categories as follows: 1) UIL comments - comments that potentially lead to the identification of the personnel who is the subject of the article. 2) Other relevant comments- comments with other significance. 3) Non relevant comments- comments with no significance.

The UIL comments were then classified into 9 a-posteriori sub categories according to severity level of information exposure, presented here with a typical UIL comment example:

1. **Explicit Identification** - Revealing the name of the subject. (e.g.: News Item-"*Meet Captain D., a distinguished officer in electronic warfare, who had been saved from Iranian military service*"; UIL comment- "*Captain David Bachshian…*").
2. **Direct Acquaintanceship** - Exposing a personal acquaintance with the subject by clear mention of a relationship. (e.g.: News Item - "*The Navy is satisfied with the appointment Colonel G. as the Commander of Flotilla 13: "the natural choice, he has a unique character"*"; UIL comment- "*Congrats to G. I know him personally. Good choice, good luck*").
3. **Transitive Acquaintanceship** - Exposing a personal acquaintance with another person from same circles or activities of the subject. (e.g.: News Item - "*Two years ago… A. was badly wounded…today against all odds, he was flying again…*" UIL comment- "*way to go, regards to the parents*").
4. **Ascriptive Association** - Exposing a personal acquaintance with the subject through use of a term of ascription. (e.g.: News Item - "*At this very moments, first lieutenant B. is getting his pilot's wings…as a helicopter combat pilot*"; UIL comment-"*…Our pride!*").
5. **Reminiscence** - Exposing a personal acquaintance with the subject through reminiscing. (e.g.: News Item - "*Introducing cadet D. …finished the course and has been chosen as battalion's star*"; UIL comment-"*Well done bro', I have been always believed in you and I won't forget that day…which I surprised you after my activity*").
6. **Common Affiliation** - Exposing an acquaintance with the subject, by mentioning place of residence, work, leisure activity, etc. held in common with the subject. (e.g.: News Item - "*... first lieutenant T. got his pilot's wings today….*"; UIL comment- :"*My village, he is a superstar*").
7. **Expressions of Warmth/Intimacy** - Exposing a personal acquaintance with the subject by expressing warm thoughts or intimate emotions. (e.g.: News Item - "*At this very moments, first lieutenant B. getting is pilot's wings…as a helicopter combat pilot*"; UIL comment- "*Nothing like this charming!! Exciting to tears! ...*").
8. **Expressions of Humor** - Exposing a personal acquaintance with the subject by expressing humor (giggle, smiley). (e.g.: News Item -"*…this is the story of Sargent N. ...*"; UIL comment- "*Haha 'as if' Sargent N. Bro.*").
9. **Semiotic Indicators** - Exposing a personal acquaintance with the subject by repeating the obfuscated name with semiotic markings indicating recognition. (Dots/ exclamation marks/slang). (e.g.: News Item - as above; UIL comment- "*Sargent N.…way to go bro. Good luck")*.

Other relevant comments were classified into one of two a-posteriori sub categories of comments:

Pro-Censor - Expressing a positive opinion about the necessity of censoring the relevant article. (e.g.: "*A little secrecy won't be harmful in our lives*").

Anti-Censor - Expressing a negative opinion about the necessity of censoring the relevant article. (e.g.: "*Why don't they publish? Even the name of secret service's chief is published*").

Using the data set that resulted from the qualitative analysis we proceeded to test a series of text mining techniques.

4. Quantitative Study

In this section we use text mining techniques to automatically detect UIL comments on censored press articles. We present here our preliminary analysis, which we plan to extend in future work.

The availability of textual data in today's social networks era has kindled an academic interest in text mining. A common text mining application in business related disciplines is the study of opinion and sentiment analysis in blogs and micro-blogs [17],[18]. Other literature addresses the need to mine emotions and expressions such as sarcasm or fear [19], [20]. Our goal is to examine whether text mining can help us detect UIL comments. The main challenge we face is the availability of the data (very low), and the need to detect UIL comments in new press articles, that may or may not use the same vocabulary we see in our training data. To that extend, our goal is to define and detect emotions related to acquaintance and affection.

Our analysis explores the relationship between UIL comments and features of the comments. Two families of features are examined: (1) features extracted from the qualitative analysis stage, referred to as Guided Qualitative (GQ)-based features; and (2) standard text mining (TM) features.

The two families represent two mining approaches. The first, guided qualitative approach is semi-automated. Here knowledge is gained from our experts in the qualitative stage. In essence, according to the GQ model, probability of a comment to be UIL is assigned based on set of pre-defined rules. For examples, our experts noticed that UIL comments are usually posted close to the release of the press article, and therefore their rank in the comment list is lower than non UIL comments. They also found that UIL comments are shorter, and commonly contain words related to 'love' or 'family'. The advantage of this approach is that it enables us to model complex rules such as *"short comment that contains the word 'bro' followed by 3 dots, is usually UIL"*. We believe that this approach is more likely to capture out-of-sample UIL suspects. The second approach (TM) is completely data driven, and not based on any prior knowledge. The GQ and TM list of features are described in Table 1.

Table 1: List of Comment Features

Family	Feature	Description
GQ	Rank	Comment rank (1st, 2nd etc.)
	Number of words	Number of words in the comment
	GQ-based UIL indicator	Whether the comment contains common UIL words/sentence structure, as learnt by our experts (e.g., words related to family, residency, etc.)
	Semiotics	Number of semiotic symbols ("!", "…", etc.)
TM	W-based UIL score	Word-based UIL score, see below*
	W-based non-UIL score	Word-based non-UIL score, see below*
	G1-based UIL score	First level grammar-based, UIL score, see below**
	G1-based non-UIL score	First level grammar-based, non-UIL score, see below**
	G2-based UIL score	Second level grammar-based, UIL score, see below**
	G2-based non-UIL score	Second level grammar-based, non-UIL score, see below**

* Word-based scores are computed as follows:

First, word importance ($\widetilde{W}_i$) for each class (*UIL*, *non-UIL*) is computed by its inverse-to-frequency. Equation 1 computes the importance of word *i* in class *UIL*. Inverse-to-frequency is commonly used to reduce the impact of common speech words such as *"in"* and *"at"*. Second, words' scores are determined by their relative importance in each class (Equation 2). Finally, a score of a comment equals to the sum of the scores of it words, computed separately for classes *UIL* and *non-UIL*.

$$(1)\quad \widetilde{W}_i^{UIL} = \frac{\left|W_i^{UIL}\right|}{\left|W_i^{UIL}\right| + \left|W_i^{non\text{-}UIL}\right|},$$

$$(2)\quad S\widetilde{W}_i^{UIL} = \frac{\widetilde{W}_i^{UIL}}{\sum_j \widetilde{W}_j^{UIL}}$$

** In first level grammar analysis, we replace each word by its lexical part-of-speech (e.g., Noun, Verb, etc). Scores are then computed similarly to Word-based scores. Second level analysis contains information about gender,

tense, number (singular, plural), and person (1st, 2nd, 3rd). Grammatical analysis of comments is based on [21].

We model class membership (UIL, non UIL) with two separate logistic regressions, one for each feature family: GQ and TM. We chose to run each family separately in order to see the marginal contribution of the approaches. Tables 2 and 3 present the outcome of the logistic regression models. In the GQ model, we find that Semiotics is insignificant. Since our experts did mention the importance of semiotics, we believe that we need to find a better model to capture this feature, or alternatively, to extend our GQ-based rules. The other rules defined by our experts, namely rank, number of words, and the existence of typical UIL sentence structure of words are shown to be highly significant in the model. According to the TM model, first level grammar is a non-significant indicator. This result is reasonable, as Hebrew sentences, similar to English, have a well-defined structure, regardless of their meaning. Words and second level grammar, however, are highly significant in classifying comments.

The performance on our sample is summarized in the Receiver-operating characteristic (ROC) Curves in Figure 2. ROC Curves are visual aids to depict the trade-off between False Positive Rate (1-Specificity) and True Positive Rate (Sensitivity) for different probability cutoffs. In other words, the output of each logistic regression is a probability for a comment to be UIL. If we set the cutoff to zero (meaning all comments with probability of being UIL higher than zero are classified as UIL), than the True Positive Rate is one, but also the False Positive Rate. If we set the cutoff to one, both rates become zero. Cutoffs in midrange generate tradeoffs, which is what the plot depicts. It is not surprising that the TM approach outperforms the GQ approach on our sample, as it is optimized to the data at hand. Splitting the data to training and validation sets, as customarily done in data mining, is infeasible in our case, as our experts received the entire dataset to learn from and extract the GQ rules. Testing our approaches on out-of-sample sets will be done in future work. The performance of the TM approach on validation set, when splitting the data, is statistically similar to the performance reported below.

Table 2: Guided Qualitative-Based Regression Coefficients

Variable	Estimate	Sig.
(Intercept)	-3.71192	< 2e-16
ln(rank)	-0.191	0.001956
ln (number of words)	0.44041	0.000231
GQ-based UIL indicator	1.25211	7.50E-08
ln(semiotics)	0.06422	0.623626

Table 3: Text Mining-Based Regression Coefficients

Variable	Estimate	Sig.
(Intercept)	-2.1957	6.30E-09
ln(W-based UIL score)	2.8893	< 2e-16
ln(W-based non-UIL score)	-1.741	< 2e-16
ln(G1-based UIL score)	-0.7071	0.8342
ln(G1-based non-UIL score)	0.2102	0.9479
ln(G2-based UIL score)	7.5987	0.0239
ln(G2-based non-UIL score)	-7.5308	0.0312

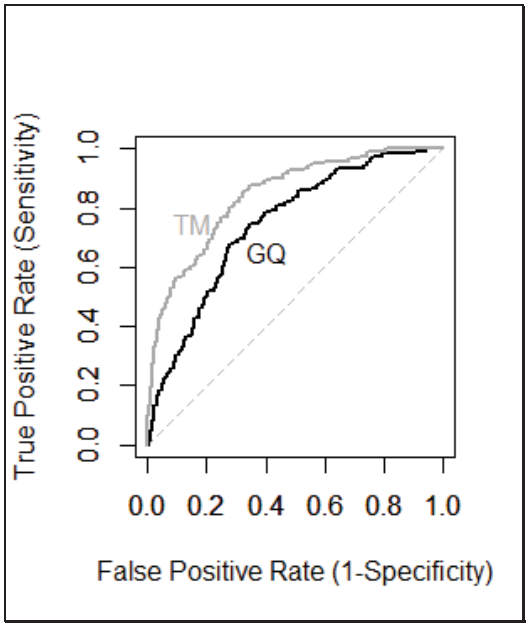

Figure 2: ROC Curves of the Logistic Models

5. Conclusions and Future Work

We have identified a form of unintentional information leakage in social networks that will continue to grow over time. Our first domain of study is news releases from organizations that are trying to control certain identifying information. Beyond the organizational realm there are vast concerns regarding the anonymity of stakeholders in the court system [3], [4] that remains to be studied. Once sufficient progress in UIL identification is made, the next steps to be considered in parallel are (a) prevention mechanisms to be put in place prior to a comment being released, and (b) implications for organizational information release policy.

Our qualitative study identified 9 distinct classes of UIL comment in the data studied. These classes represent the 'way' commenters hint, or implicitly state that they personally know the subject of the press release. Our

quantitative stage attempts to quantify these utterances and detect them automatically. Two quantitative approaches were presented: Guided Qualitative approach and Text Mining approach. Both were found to have significant positive performance, with the latter outperforming the former for the current data set. Further studies on larger data sets from different subject domains will help validate our model as effective for UIL identification, and enable tuning it for other domains.

6. Acknowledgements

This research was funded by Israel Ministry of Science and Technology research grant 3-9770 "Data Leakage in Social Networks: Detection and Prevention". The crawler and related programs were written by Yossi Tokash.

7. References

[1] A. Shabtai, Y. Elovici, and L. Rokach, *A survey of data leakage detection and prevention solutions*. Springer, 2012.

[2] R. T. Davis, *The US Army and the Media in the 20th Century*, vol. 31. Government Printing Office, 2009.

[3] M. Johnson, "Of public interest: How courts handle rape victims' privacy suits," *Commun. Law Policy*, vol. 4, no. 2, pp. 201–242, 1999.

[4] J. M. Schumm, "No Names, Please: The Virtual Victimization of Children, Crime Victims, the Mentally Ill, and Others in Appellate Court Opinions," *Ga. Law Rev.*, vol. 42, p. 471, 2008 2007.

[5] J. Wu, J. Zhou, J. Ma, S. Mei, and J. Ren, "An Active Data Leakage Prevention Model for Insider Threat," 2011, pp. 39–42.

[6] K. Purcell, L. Rainie, A. Mitchell, T. Rosenstiel, and K. Olmstead, "Understanding the participatory news consumer," *Pew Internet Am. Life Proj.*, vol. 1, pp. 19–21, 2010.

[7] S. Ghorbani and Y. Ganjali, "Will you be my friend? privacy implications of accepting friendships in online social networks," in *2012 International Conference on Information Society (i-Society)*, 2012, pp. 340–345.

[8] S. Herskovitz, "Ha'Meida Dolef Befek (in Hebrew. 'The information leaks freely')," *Status Magazine*, no. January, 2012.

[9] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *SIGCOMM Comput. Commun. Rev.*, 2010, vol. 40, pp. 112–117.

[10] M. Glaser, "Facebook Pushes Comments Upgrade, But Will Publishers Bite? | Mediashift | PBS," *PBS.org MediaShift*, 02-Mar-2011. [Online]. Available: http://www.pbs.org/mediashift/2011/03/facebook-pushes-comments-upgrade-but-will-publishers-bite061. [Accessed: 29-Jun-2014].

[11] M. Shanahan, "More news organizations try civilizing online comments with the help of social media," *Poynter.org*, 16-Jul-2013. [Online]. Available: http://www.poynter.org/latest-news/top-stories/218284/more-news-organizations-try-civilizing-online-comments-with-the-help-of-social-media/. [Accessed: 29-Jun-2014].

[12] N. Phillips and C. Hardy, *Discourse Analysis: Investigating Processes of Social Construction*. Sage, 2002.

[13] C. Grbich, *Qualitative data analysis: An introduction*. London: Sage Publications, 2007.

[14] A. Bryman, *Social Research Methods.*, 3rd ed. Oxford: Oxford University, 2008.

[15] P. Liamputtong, "Qualitative data analysis: conceptual and practical considerations," *Health Promot. J. Austr.*, vol. 20, no. 2, pp. 133–139, 2009.

[16] N. . Kondracki and N. . Wellman, "Content analysis: Review of methods and their applications in nutrition education," *J. Nutr. Educ. Behav.*, no. 34, pp. 224–230, 2002.

[17] A. Ghose, P. G. Ipeirotis, and B. Li, "Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content," *Mark. Sci.*, vol. 31, no. 3, pp. 493–520, 2012.

[18] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *Knowl. Data Eng. IEEE Trans. On*, vol. 23, no. 10, pp. 1498–1512, 2011.

[19] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business: Market-structure surveillance through text mining," *Mark. Sci.*, vol. 31, no. 3, pp. 521–543, 2012.

[20] O. Tsur, D. Davidov, and A. Rappoport, "ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews.," in *ICWSM*, 2010.

[21] Y. Goldberg, "Automatic syntactic processing of Modern Hebrew," Ben-Gurion University of the Negev, 2011.