Digging up Social Structures from Documents on the Web

Eleni Gessiou, Elias Athanasopoulos, and Sotiris Ioannidis
Institute of Computer Science
Foundation for Research and Technology - Hellas
N. Plastira 100, Vassilika Vouton, GR-700 13 Heraklion, Crete, Greece
{gessiou, elathan, sotiris}@ics.forth.gr

ABSTRACT

The world wide web is a medium used for a plethora of applications. Apart from traditional web browsing, the web is commonly used for sharing and exchanging documents. Unfortunately, a poorly configured web server may give document access to any user. In this paper, we performed a massive web crawl and collected over 15 million Microsoft Office documents and PDF files. We examined the metadata of each document and tried to extract any privacy-sensitive information associated with them. Our analysis revealed the existence, and exactly identified cliques of users that edit, revise and collaborate on industrial and military documents. Second, we grouped these cliques to produce a unified graph, and examined whether it contains the properties of a social graph. Third, we cross-correlated of all members identified in a clique with users of Twitter, and showed that it is possible to easily match them to their Twitter accounts. Our study raises major concerns about the risks involved in privacy leakage, due to metadata embedded in documents that are stored in public web servers.

Keywords

privacy leakage, metadata, social networks

1. INTRODUCTION

Millions of documents are created and shared over the Internet every day. Popular formats for these files are Microsoft Word, Excel, PowerPoint, as well PDF. Those documents contain much more data than what is intended to by their creator. This data is automatically generated by the applications, that is the word processors, the presentation managers, etc., and we refer to it as *metadata*. Most of the time, the author of a document is totally unaware of the existence of any metadata associated with it. Moreover, the popularity of the web has led many users to store their files in web servers, which are freely accessible from everywhere. Very often due to weak security configurations, these files become accessible to everyone.

Apart from the actual content of a document, which can be encoded or encrypted to defend against inspection, there is a metadata part. In Microsoft Office documents there are two types of properties included in their metadata, built-in and custom properties [7]. Custom document properties may be details about a file that helps identify it, such as the *date* completed and the *author name* or other important information in the file. Built-in document properties can not be deleted or even be changed. Built-in document properties are stored in the document and are not part of the author's content. They include information such as *title*, *keywords*, *subject* and *comments* that identify the document's subject matter and content. Similarly, PDF documents include metadata such as *viewer preferences*, *page mode*, etc.

The metadata may contain very sensitive information about the people who have authored or modified the document. There are several security issues that should be considered when thinking about metadata. First, revealing the *creator* of a document can be used for deriving possible usernames. Second, revealing the *application used* for the creation of the document may be helpful in determining potential attacks. For example, exploits or computer worms often target specific, known to be vulnerable, versions of an application [29, 23]. Thus, revealing the software and version used to create a document can narrow down an attack targeting a particular user.

This kind of information leakage may have very serious consequences. The most popular example is the case of Dodgy Dossier [1]: a document of the British government on Iraq, published in Microsoft Word format. An analysis on the *revision history* of the document revealed that much of the material of the dossier was actually plagiarized from a US researcher on Iraq. The incident raised many questions about the involvement of UK and the quality of British intelligence during the second Iraqi War. The importance of metadata associated with a document is also highlighted by a recent incident in Arizona [6]. The Supreme Court unanimously decided that metadata is part of public records and thus must be released when the records are also released. The Dodgy Dossier and the Arizona case are just a few real-world examples demonstrating that document metadata may contain very sensitive or even critical information.

In this paper we present a large-scale study of metadata associated with over 15 million, publicly accessible, on-line

documents, collected sporadically over a one year period. We use these documents to quantify the amount of metadata present in those on-line documents, and find sensitive information contained in the metadata of these documents. We employed existing libraries and tools to extract and visualize the degree of metadata diversity in several file formats.

Contributions. Our main contributions are the followings:

- We present an extensive analysis of all metadata embedded in over 15 million documents hosted on public web servers.
- We extract social cliques, composed of users that collaborate in the production of a particular document. From our analysis, it is evident that users of military and governmental institutions are not particularly concerned about protecting their privacy from possible risk related to metadata present in documents they create, edit and share. With our analysis we are able to associate many users with their colleagues, using information solely present in metadata.
- We search for the cliques identified in the document metadata analysis in Twitter. Our search successfully cross-correlated members of a clique with Twitter users. This unveiled that the members of a clique, form groups of followee and followers in Twitter.

Organization: This paper is organized as follows. In Section 2 we describe the methodology used to collect the data. In Section 3 we present the results of our analysis related to the various flavors Microsoft Office and PDF documents. We present our findings in extracting social cliques from information stored in metadata in Section 4. We review related work in Section 5 and conclude in Section 6.

2. METHODOLOGY

In this section we outline the basic methodology we use for the data collection. We first present the tools and techniques we employ for gathering all the samples and later we provide a short overview of some generic properties of all data collected.

2.1 Overview

It is necessary to generate a large collection of on-line documents in order to carry-out the metadata study. One rich source of on-line documents is a popular search engine, like Google. We created a custom crawler in the Python [30] scripting language, which is able to parse web pages with search results produced by Google. According to Google's policy, Google search engine does not serve more than 1,000 results per query [10]. We therefore used a dictionary to produce a series of queries, which can generate a large set of search results.

The query process works as follows. We take all words composed by more than three letters from the English dictionary, and use them to form a query for Google. Each query is composed of one English word, taken from the dictionary, and the *filetype* directive used by the Google search engine. This directive assists in producing a result set composed solely of specific filetypes.

We parse the search results using the custom crawler and we extract all included on-line documents. The crawler is able to recognize files produced by Microsoft Office, as well as PDF documents, based on their extension (`.doc`, `.xls`, `.ppt` and `.pdf`). Once a file is spotted in a set of Google results, the crawler downloads the file and verifies that the extension of the file matches the MIME type [5], which is advertised in the HTTP response issued by the host of the file. We discard all documents for which the file extension does not match the advertised MIME type for the following two reasons. First, it has been documented that many web servers are not configured properly [22] to serve all files with the correct MIME type. Second, it is a well known practice for web sites that host malware to advertise wrong MIME types in order to lure the user to open the malware, which is camouflaged under a fake extension. Thus, we remove all files downloaded with a discrepancy between the extension and the MIME type, since we do not want to have a biased sample due to issues not directly related with privacy leakage.

For each downloaded file we proceed and extract all possible metadata. We use the hachoir-metadata [3] and libextractor [2] libraries for extracting all metadata associated with Microsoft Office documents. As far as PDF files are concerned, we use the Poppler [8] rendering engine. All metadata extracted from Microsoft Office and PDF documents are stored in MySQL database for further processing.

2.2 Sample Properties

Using the technique outlined above we collected more than 5 million of MS Word documents, about 2.5 million of MS Spreadsheet and 2.5 million of PowerPoint and more than 5 million of PDF documents. Overall, our sample is over 15 million of distinct documents. All documents are hashed using the MD5 cryptographic hash function, to remove possible duplicates.

There is a fairly distinct distribution of the various filetypes. Notice that PDF and MS Word files dominate the set, compared to MS Excel and MS PowerPoint files. Our intuition is that PDF and MS Word files are more likely the user's choice for exchanging documents over the web. This may be also a result of the generic nature of MS Word and PDF format, which is ideal for embedding unstructured information. On the other hand, MS Excel and MS PowerPoint documents are more suitable for usage in a corporate environment, providing information structure (financial sheets or presentation slides), and thus less likely to find on public web servers. Nevertheless, our set includes substantial contribution from all of the four non-HTML filetypes considered the most popular [11] and thus we consider the metadata study carried out in this paper, highly representative.

3. CASE STUDY RESULTS

In this section we present our analysis in over 15 million online public documents. We divide the presentation of our results in three parts, each one associated with one of the three formats, namely Microsoft Word, Excel and PowerPoint included in our data set. We have conducted a similar

analysis for 5 million PDF documents, but we do not include it here due to space restrictions.

3.1 Microsoft Word Documents

We begin our analysis by looking at documents created with Microsoft Word. The Microsoft Word word processor creates files with the `.doc` extension. The extension however is not the only criterion for classifying a file as a Word document. We also verify that the file has been served by a web server using the "application/msword" MIME type. Our dataset contains over 5 million `.doc` files.

In our study, we choose not to consider the fields: *number of pages*, *number of paragraphs*, *number of lines*, *number of words* and *number of characters*. This is because the majority of files include these types of metadata and they have little significance as far as privacy is concerned. Nevertheless for completeness sake we present some general statistics taken from 99.7% of all sampled files: each Word file is approximately 2.8 pages long, includes 45 paragraphs, has 131 lines, 2,458 words and 13,828 characters.

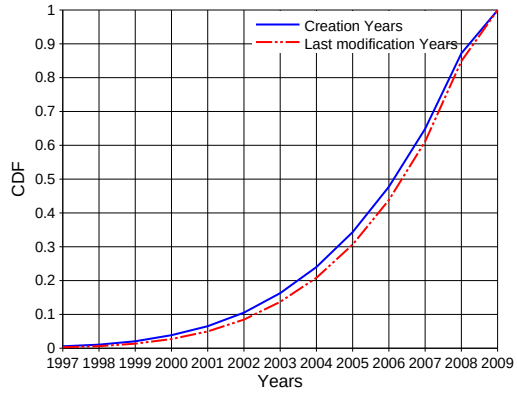

Figure 1: The CDF for creation and last modification years for Microsoft Word files. The blue, solid line is for *creation years* and the red, dashed line is for *last modification years*.

The CDF of *creation year* and *last modification year* of all Word documents in the sample are shown in Figure 1. Both, CDF of *creation year* and *last modification year*, present a huge raise in recent years. The intuition behind this is the following. First, people use Word documents more frequently in recent years compared to the past. Second, some users have become more familiar with the Internet and upload more documents. Third, the Google search engine returns the more recent documents than old ones. A slight shift is observed between creation year and last modification year. Apparently, a document that was created in year *X*, is expected to be modified in the years *X+1*, *X+2*, etc.

In Figure 2 we present the ten most popular metadata fields appearing in the entire set of Word files. Notice, that almost all Word documents contain the *application used* field. This is because all versions of Word fill in this information automatically in every document. Observe also, the *template used* field. Almost 93% of the files use the default template of Microsoft, Normal.dot. However, apart from the default

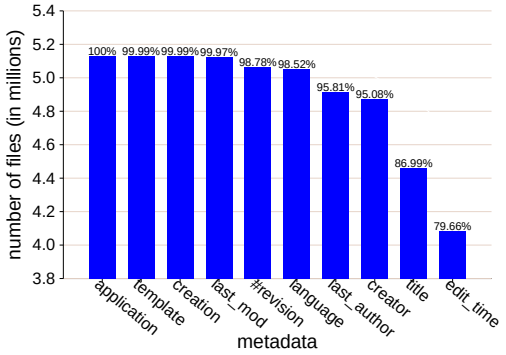

Figure 2: Top 10 metadata in Word documents. "last_mod" stands for *last modification date*.

Normal.dot, it seems that many organizations, especially the ones from the governmental sector, use their own custom templates. For example, nearly 1,500 `.doc` files, all downloaded from a City Council's site of a Canadian town, use the same custom template. These files have been modified by a set of different users, which can be identified through *name of creator* (8th place in Figure 2), *name of the person who last saved the document* (7th place in Figure 2) and *revision history* fields. More interestingly, all these names cannot be located in the City Council's site, using the site's search service. Thus, even though these names cannot be extracted from the actual web site, they can be extracted from metadata in files that the web site hosts. In another incident of an Australian governmental organization, about 99% of all documents, based on the same *templates used*, were last modified by a user who is identified, through the above mentioned metadata, as the organization's CEO.

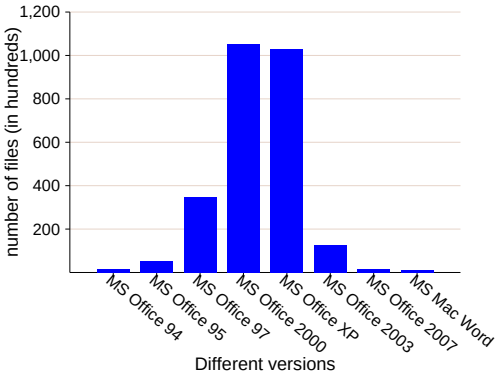

Figure 3: Versions of Microsoft Office used in Word documents.

The application used for creation and modification of Word documents is mainly the Microsoft Word software. Figure 3 shows the popularity of each version of Microsoft Word. The figure shows that Microsoft Office 2000 (Office 9.0) and Microsoft Office XP(Office 10.0 or Office 2002) are the versions most commonly used. Only a few thousand of Word files were created by MS Word for the Mac. An intriguing aspect to examine is if the most popular versions leak less

information or the least popular versions are also the more secure. Although Microsoft Word 2000 is the version used to create the majority of the documents included in our set, it is obviously the version which reveals the most information according to Figure 4. Especially in the case of *revision history*, it has a lot of difference when compared with the other versions and especially with the Word version for the Macintosh. Microsoft Word for the Macintosh presents low levels of metadata presence almost in all fields, and as far as the *revision history* is concerned no information is revealed. The following metadata types are not listed in Figure 4, because they are included by default in all versions of Microsoft Word: *creation date*, *last modification date*, *application used* and *template used*.

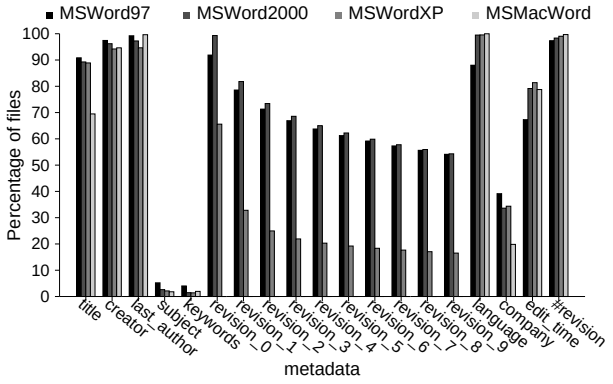

Figure 4: Existence of metadata among different versions of Microsoft Office used for the creation of Word documents.

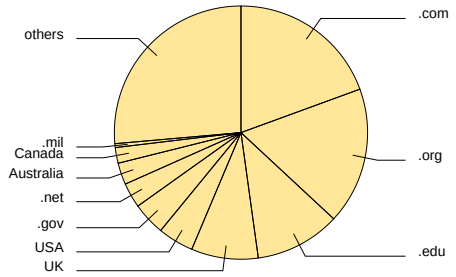

Figure 5: Distribution of Word documents based on Top Level Domains.

In Figure 5 we present a characterization of all collected Word documents based on their domain of origin. Half of them are from `.com`, `.org` and `.edu` sites. It is tempting to investigate whether documents that come from governmental and military sites are more secure than the rest of the collected documents, due to the fact that governmental and military documents are often considered to contain classified information. In Table 1 we present all types of metadata found in Word documents, along with the the percentages of documents that contain the metadata from `.mil` sites and from `.gov` sites. Obviously, these particular documents embed about the same amount of sensitive information and as a result they experience similar information leakage. The increased percentages in the cases of *subject* and *keywords* is apparently due to the need of taxonomy for finding documents relevant with a particular subject.

Metadata	% .mil	% .gov	% all
Title	83.97	85.09	86.39
Creator	89.93	88.88	92.32
Last saved by	90.58	91.90	93.08
Creation date	96.69	97.66	97.23
Last modification date	96.69	97.66	97.22
Application used	96.69	97.45	96.60
Subject	4.82	5.12	2.20
Keywords	0.76	3.29	1.54
Comments	0	0	0.0012
Template used	96.68	97.59	96.98
Format used	0	0.009	0.0011
Revision history 0	30.72	48.35	41.84
Revision history 1	25.9	39.55	30.30
Revision history 2	23.95	35.40	26.26
Revision history 3	22.43	33.22	24.24
Revision history 4	21.03	31.74	22.95
Revision history 5	21.28	30.58	21.97
Revision history 6	20.7	29.65	21.16
Revision history 7	20.20	28.87	20.42
Revision history 8	19.77	28.23	19.79
Revision history 9	19.34	27.61	19.22
Language	95.37	96.47	95.11
Company	45.02	35.26	31.90
Total editing time	75.76	72.82	77.04
Revision number	95.66	95.73	96.09

Table 1: The percentages of metadata fields in military and governmental Word documents in comparison with the total number of Word documents.

As far as military documents are concerned, every one in two includes *company* information, among them the more frequent are names for military departments. We found 1,500 distinct names of individuals who took part in the creation/modification of files and are associated with one of the fore-mentioned departments. All names are formatted in the same way: "name.surname", e.g, "john.doe". In case of common names an ascending number is added, e.g, "john.doe1", "john.doe2" etc. That is a perfect example of the format of possible usernames in the particular department.

3.2 Microsoft Excel Documents

Microsoft Excel is an application that produces Spreadsheet documents with the `.xls` extension. As in the case of Word files, we verify that a file has been served correctly by a web server using the "application/vnd.ms-excel" MIME type, before classifying it as an Excel document. Our dataset consists of over 2M `.xls` files.

We choose to not consider the fields: *number of pages*, *number of paragraphs*, *number of lines*, *number of words* and *number of characters*, since they do not apply in the `.xls` format. This is confirmed by our dataset, as these fields are filled for only four files, and furthermore they are incorrect. *Revision history* is also omitted because no Excel file in the sample has any information about it.

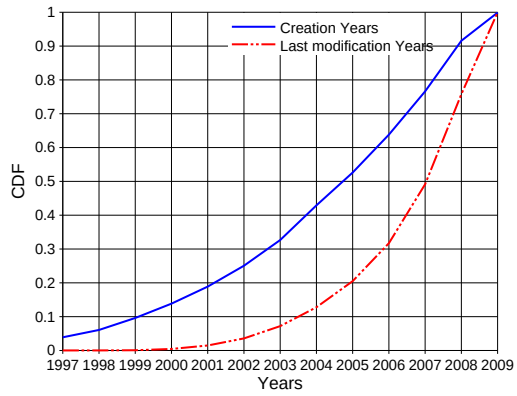

Figure 6: The CDF for creation and last modification years for Spreadsheet files. The blue, solid line is for *creation years* and the red, dashed line is for *last modification years*.

In Figure 6 we plot the CDF of *creation dates* in Spreadsheet documents. We observe a rapid increase in 1996 with more than 50,000 files (not shown in the CDF figure). This amount decreases by half in 1997 and thereafter we have a stable accession until 2008. This increase may have been caused by the release of Windows 95 in 1995 or the release of Microsoft Office 4.0 in 1994 when computers started becoming more accessible to end users.

With respect to the *last modification dates* in Excel documents, in Figure 6, we observe that they are mostly accumulated in the period 2000-2008, although older documents do exist. The intuition is that Excel documents may be more meaningful and thus have a longer life-span than Word files. To verify this assumption we find the average difference between *last modification date* and *creation date*. In other words, for every file containing both *last modification* and *creation date*, we subtract these two values and calculate the average of the results. This shows that the Spreadsheet files have an average life-span of 2.1864 years in contrast to Word files that have an average life-span of 0.2558 years. Thus, Excel documents have an average life 2 years while Word documents only 3 months. In addition, the functionalities provided by this type of documents is ideal for long-term storage of data (for example, Spreadsheets that contain the salaries and/or the employees of a company).

As we can see from Figure 7, the top 10 metadata found in Excel documents are quite different from the ones in Word documents. *Application used*, exists in the majority of Excel documents, that are created by Microsoft Excel, except for around one thousand files that are created by Microsoft Access software. *Creation date* and *last modification date* of the documents are in the 2nd and 4th place respectively, while the *name of the person who last saved the document* and the *name of creator* are in 3rd and 5th places. We found that one fourth of the documents are created and last saved by the same individual. In 6th place, *company* metadata field is present in about 500K Excel files. Next, *title* is followed by *subject*, *keywords* and last *revision number*.

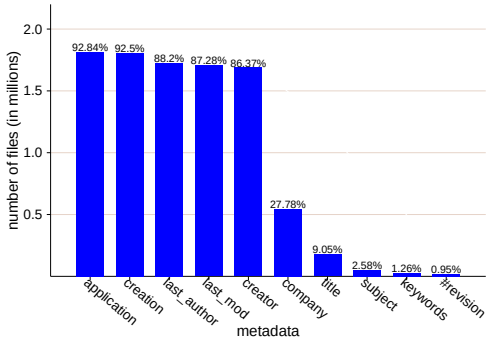

Figure 7: Top 10 metadata in Spreadsheet documents. "last_mod" stands for *last modification date*.

Based on our dataset, Spreadsheets seem to be used by many statistics offices, as the companies most popular in the *company* field of metadata are statistics organizations, in the Czech Republic, in the United Kingdom and in the United States.

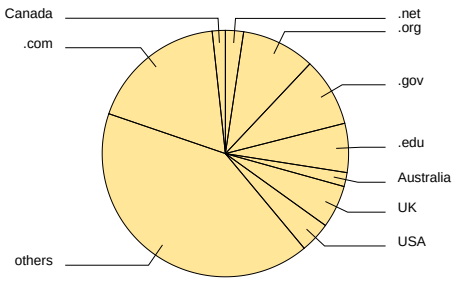

Figure 8: Distribution of Spreadsheet documents based on Top Level Domains.

Figure 8 is a distribution of the Excel documents according to the domains of their origin. The Excel files are more evenly distributed amongst TLDs than Word files. We can see that the majority are from `.com` sites and a very large chunk comes from governmental sites (`.gov`). This leads us to explore the metadata in documents found on `.gov` sites in more detail.

Table 2 presents the percentages of fields of metadata in governmental Spreadsheet files in comparison to all the Spreadsheet documents in our dataset. Governmental documents seem to follow the same distribution of metadata as in non-governmental documents. As in the case of Word files, *subject* and *keywords* occur more frequently in `.gov` files than in documents from other domains, as well as *title* that exhibits double the probability to occur in a document from governmental site than from any other site.

3.3 Microsoft PowerPoint Documents

In the final analysis section we focus on Microsoft PowerPoint files. All files have the `.ppt` extension and as before we verify that the file has been served by a web server us-

Metadata	% .gov	% all
Title	11.43	5.97
Creator	75.19	81.44
Last saved by	88.44	92.58
Creation date	94.27	94.49
Last modification date	89.93	90.85
Application used	80.26	83.32
Subject	3.77	1.87
Keywords	2.41	0.99
Comments	0	0.0002
Template used	0.01	0.019
Format	0.003	0.001
Language	0.010	0.0018
Company	25.87	24.63
Total editing time	1.21	0.79
Revision number	2.05	2.13

Table 2: The percentages of metadata fields in governmental Spreadsheet documents in comparison with the total number of Spreadsheet documents.

ing the "application/vnd.ms-powerpoint" MIME type. The total number of PowerPoint files in our dataset exceeds two million.

Comments, *format* and *revision history* are absent from all PowerPoint files in our dataset, so we do not include them in our results, as well as the *number of pages*, the *number of paragraphs*, the *number of lines*, the *number of words* and the *number of characters* fields.

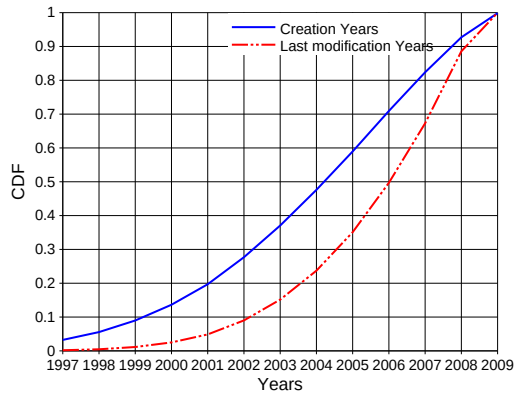

Figure 9: The CDF for creation and last modification years for PowerPoint files. The blue, solid line is for *creation years* and the red, dashed line is for *last modification years*.

In Figure 9 we present the CDF of PowerPoint files among the years, based on their *creation year* and *last modification year*. Generally the number of documents increases rapidly throughout the years. In contrast with the other filetypes, *creation year* of PowerPoint documents experience a smooth distribution in recent years. We see that there are only a few documents up until 1998 but after that there is a steady increase.

Many companies create sample PowerPoint files, which serve as *templates* for future files [27]. We can verify that an initial *template* is used multiple times within a *company* by looking at the collected documents. Thus, we can calculate the average life time of PowerPoint files, which in our sample is estimated to 1.6096 years. An interesting finding that may justify the longer life time of PowerPoint files, is that we discovered several individuals who are the *authors* in more than one PowerPoint files. The files, in these cases, have the same *creation date* but different *last modification dates*. So, our assumption is that the *authors* use the first version of the files as a seed to create new presentation files. In other words, the first PowerPoint file serves as a *template* for future presentations, and as a result these initial PowerPoint files increase the average life time of the files. Another reason that explains the long life time of PowerPoint files is that many of them, are used for lectures in university classes. We observe that specific individuals/professors create one initial presentation for their classes and each year that they teach the same course, they enhance their slides. This way, a PowerPoint file can live for many years. Considering the above, companies and academic lecturers are among the main users of PowerPoint files.

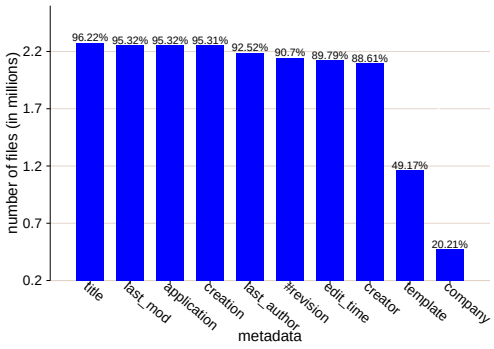

Figure 10: Top 10 metadata in PowerPoint documents. "last_mod" stands for *last modification date*.

In Figure 10 we present the metadata fields which occur more frequently in PowerPoint files. We see that almost all PowerPoint documents contain information about their *title*, *last modification date*, *application used* and *creation date*. We found that tens of thousands of PowerPoint users create documents with older versions of PowerPoint many years after the initial launch, e.g. PowerPoint 4.0. Fewer files contain *name of last author*, *revision number*, *total editing time* and *name of creator*. In contrast with spreadsheets, half of the PowerPoint files contain information about the *template used*, as templates are heavily used in PowerPoint files[1]. In Table 3 we list the most popular design *templates* used in our sample.

The *company* field is present in fewer than one forth of all documents. The distribution of PowerPoint files based on their TLD in figure 11, verifies that most of the files are used for academic purposes as the dominant TLD is `.edu`. Apart from the educational domain, `.org` and `.com` are also used.

[1]All design templates are listed at: `http://office.microsoft.com/en-ca/templates/`

Design Template	#PowerPoint files
Blends	36,423
Blank Presentation	25,231
Stream	23,073
Pixel	19,086
Edge	18,807
Ocean	18,402
Textured	15,326
Capsules	15,031
Beam	12,301
Globe	12,275
Profile	12,145
Network	12,029
Ripple	11,404
Layers	11,338
Soaring	11,097
Shimmer	10,823

Table 3: Popular design templates in PowerPoint files.

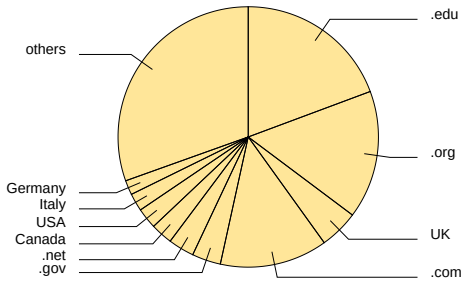

Figure 11: Distribution of PowerPoint documents based on Top Level Domains.

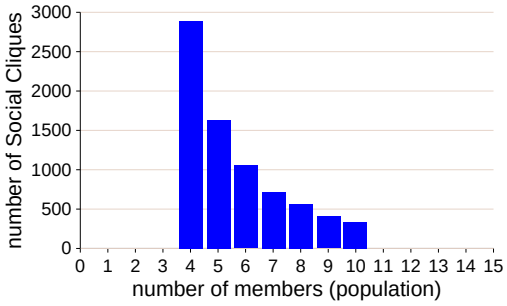

Figure 14: Distribution of the populations of social cliques. The x-axis shows the number of members inside a social clique, and the y-axis indicates the number of social cliques that correspond to each population.

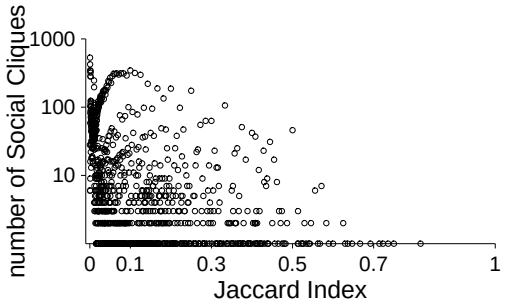

Figure 15: Distribution of jaccard indices. The x-axis shows the ditribution of jaccard indices and the y-axis indicates the number of social cliques corresponding to each jaccard index.

4. UNDIGGING SOCIAL STRUCTURES

In this section, we demonstrate how someone can extract social structures by inspecting the authors collaborating in editing documents. We apply our techniques in a subset of all documents collected from public web servers and we identify the nature of the social structures produced. First, we present our techniques for constructing cliques of users that collaborate in the production of a document and second we present how to locate these produced cliques in popular social networks, such as Twitter [21].

4.1 Identifying Cliques

A detailed look of our collected data showed that a particular individual was the author in fourteen different PowerPoint documents, three different Word documents and three more Excel documents. In PowerPoint documents, he collaborated with seven different individuals, in Word documents with three different individuals and in Excel documents with another two figures. This observation drove us to investigate the possibility of extracting social structures by inspecting the metadata embedded in documents publicly available in web servers.

For this we use all Excel files we have collected by crawling the web. For each document we locate the metadata fields *name of creator* and *name of last author*. These fields, as it has been already stated, identify the creator and author of the document. We then search for all documents that also list these authors in the respective metadata fields. If two documents list the same creator or author and have been downloaded by the same web server - indicated by the domain of the url - then we consider that these authors collaborate. In this way we create graphs which have all identified authors as nodes. Each node is linked with another node if and only if these two authors are collaborating on a particular document. We extracted 10,000 social cliques with at most four hops depth. This means that the maximum route-length connecting two individual authors, if such route exists, is of length four.

In Figures 12 and 13 we show two example cliques, derived from the metadata information found in our sample. In each graph, nodes represent authors and solid edges represent that two authors are collaborating in editing a particular document. Dashed and bold edges represent a connection where members of one clique collaborate with members of another clique. The weights on the edges indicate the number of the documents that the two authors collaborate on. If no weight is indicated on an edge, assume as being one.

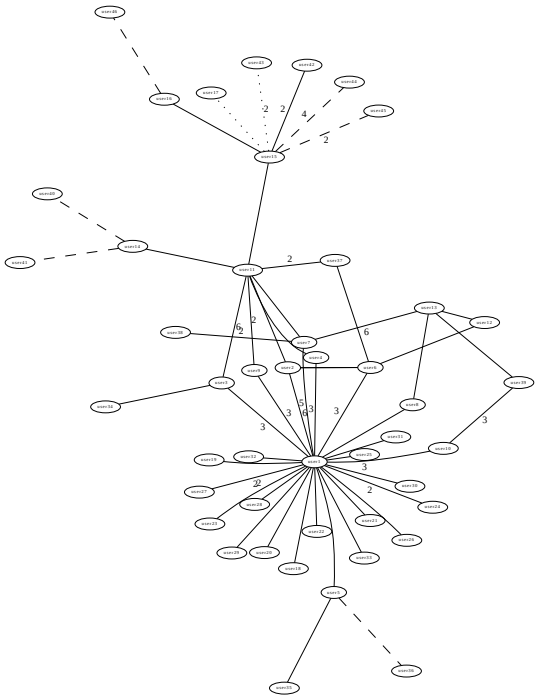

Figure 12: Clique of company A. The dotted, the dashed and the bold dashed edges are the connections of company A with smaller cliques belonging to other companies.

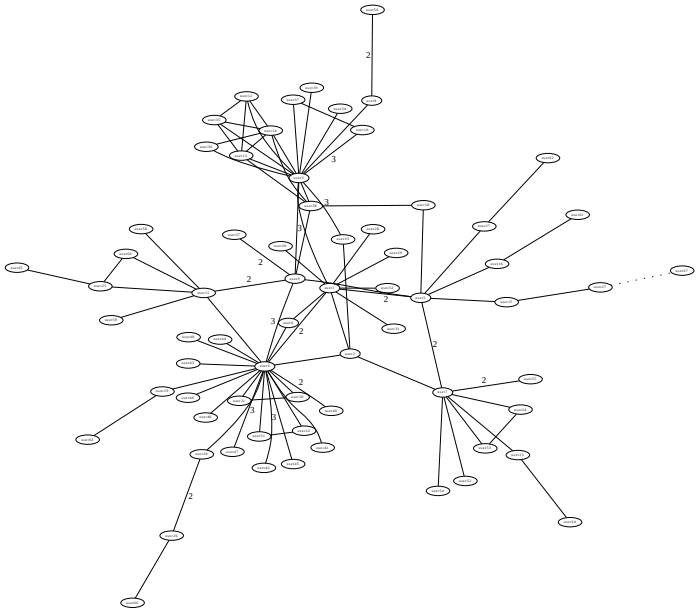

Figure 13: Clique of company B. The bold edge stands for the company's connection with company A and the dotted edge is for its connection with another smaller company.

The distribution of the population of the social cliques extracted is shown in Figure 14. There are 1,481 social cliques having more than 15 members each (not shown in this graph). Only 6 social cliques consist of more than 500 members. The most populated social cliques are one with 3,886 members and another with 3,923. For each individual social clique, we seek to identify if it exceeds in similarity compared to the rest of the social cliques created. We use the Jaccard similarity metric [19], which is defined as:

$$J = \frac{|A \cap B|}{|A \cup B|}. \quad (1)$$

The Jaccard indices of the social cliques are shown in Figure 15. More than 350,000 Jaccard indices are equal to 0 (not depicted in the graph). The majority of the rest have value lower than 0.1, which means that the corresponding social cliques do not have more than 10 members in common. Note that many author names, as listed in the metadata fields of a document are common pseudonyms, such as "Preferred Customer", "Valued Gateway Client", "Unknown User", etc.

We also managed to merge social cliques into larger ones, automatically. For example, we merged a social clique containing 78 edges with 70 other social cliques of different companies. The resulting clique contains 1,564 edges among 2,172 nodes, with an average degree of 1.44. The most connected node has 55 edges, meaning that the particular individual has collaborated with other 55 individuals, on writing/modifying some documents. There are 3 heavily connected nodes, 14 nodes having more or equal than 10 edges, some tens nodes having 9 to 2 edges. The majority of the nodes have one edge, indicating that there are cliques of 2 individuals that work only one with the other. The average clustering of this social clique is 0.013841, there are 686 connected components, 1,491 maximal cliques and the size of the largest clique is 4.

4.2 Identifying Users in Social Networks

In this section we seek to identify if we can efficiently fingerprint users [32] that collaborate in the production of documents, by locating them in popular social networks, such as Twitter. We try to match the cliques we have already identified with users following and are followed by particular users in Twitter.

First, we adjust all identified cliques by filtering out the most frequently occurring names in the documents' metadata. All, the 10,000 identified social cliques include 124,779 names in total, from which 51,709 of them are unique. For the rest of our experiments we exclude the 27 most frequently appearing names, such "Preferred Customer", names that do not contain at least 2 words of at least 2 letters (we want a full name and not just a pseudonym). Also, we do not select names that contain generic words such as "bureau", "department", "service", "city", "user", "customer", "administrator", "school", "student", "staff", as they are popular pseudonyms selected by different organizations and thus they dilute the results. The experiments and results that are described below use full names of people that wrote/modified at least 9 files and at most 47 files.

Using the social cliques that the above frequent full names, 1,307 in number, were included, we identify these particular individuals at Twitter and find their followers and those they follow. We seek to extract correlation that would verify the people collaborating in the editing of a particular document can be identified in Twitter.

Overall, we examine 575 cliques, containing at total 14,969 people. We find that 1,911 people among them own a Twitter account. We manage to find 115 social cliques that a subset of their members correlate with each other through Twitter. People in these social cliques seem to have common friends in Twitter. We also find one case that 2 out of 3 individuals belonging to a social clique, are friends and also have 39 friends in common. In another case, 2 people out of 19 in initial social clique, are friends and moreover have 4 common friends. There are also 3 social cliques, containing 131, 500 and 297 individuals each, that their members have common friends in Twitter, and moreover there is a couple of individuals in each clique that connect to each other with direct friendship in Twitter. We also find that 2 people follow the group of the company that they work in, based on the url in our dataset. This fact verifies that we can correctly match the identity derived from the metadata and the one registered in Twitter.

5. RELATED WORK

The first paper that highlights the risks due to metadata in documents found on the Internet is [15]. Despite Byers et al. counting the hidden words in in a few thousand documents, they do not take into account all available kinds of metadata and their sample is much smaller than the one used in this paper. A new tool which finds personally identifiable information that may be stored in documents is introduced in [14]. LeakHunter tries to solve the problems that metadata may cause to companies and individuals, similarly to the ones we have highlighted in this paper. The ideas in [13] and our work are very similar. The metadata used in the above study however, is somewhat different from the ones we use. That is because these metadata are from the Operating System's filesystem. The filesystem's metadata do not normally contain personal information, and that is why they mostly focused on temporal changes and not in privacy issues.

In the context of privacy risks due to metadata, [12] presents several incidents that demonstrate that security breaches, and sensitive information disclosure has recently become a serious threats to many organizations around the world. Among other findings, [9] indicates that business users in Asia are unaware of the risk of metadata. Similarly, the authors of [18] support that the overall amount of metadata associated with documents is increasing. Their assessment and results, suggest that a more detailed analysis of metadata may reveal more associations between individuals, e.g. the existence of social networks; a fact that our study confirms.

In another context, the authors of [16] take advantage of the space that metadata and generic unwanted data take up in a document, and utilize this space for steganography.

Symantec [4] shows that the majority of malicious Trojans exploiting file formats in 2007 was primarily in Word documents (67%), PowerPoint files (17%), SpreadSheet files (3%) and PDF documents (3%). This observation and the fact that these file formats are considered to be widely used from our experience, led us to select them for our study. Many real-life and potential incidents concerning hidden data in Word, SpreadSheet, PowerPoint and PDF documents, are presented in the 13th chapter of [31]. The problems that the revision history in Word files can cause, are the first to be mentioned. Overall, although much work has been done to identify and to remove sensitive information from documents, our study is the first that tries to quantify the amount of this information. Using existing libraries and tools, we show the amount of the different kinds of metadata in several file formats. The novelty of our work is that we analyze metadata in an empirical study on a very large set of documents. The number of documents that we use demonstrates that our findings are quite representative.

The authors of [28] developed the PRIX (PPT Residual Information eXtractor) tool. Its aim is to identify the residual information in PowerPoint documents. Residual information is created when the option "allow fast saves" is selected. In a followup work [27], apart from text residual information, PRIIX (PowerPOint Residual Information & Identifiers eXtractor) extracts slide and object identifiers, too. Data concealment and detection in Microsoft Office 2007 files that use Office Open XML (OOXML) as their basis is studied in [26]. The paper starts by proving that someone can indeed hide data in such files and it also presents algorithms for finding hidden data in these. The retrieval of any object or text previously deleted or modified, from the creation time of the document to its most recent version, is attempted in [17] for PDF documents.

There is a considerable amount of previous work in the field of extraction and analysis of social networks. P.Mika presents Flink [25], which constructs and visualizes social networks of semantic web researchers by using information from sources such as web pages, emails, publication archives and FOAF profiles. Polyphonet [24] presents of series of methods for obtaining a social network using a web search engine are described and used in order to enhance scalability. Polyphonet is the implementation of these algorithms, that are enabled at Japan Society of Artificial Intelligence conferences over three years and at the UbiComp conference. Recently, some steps to email social networks have been done, such as [20] that presents behavioral profiles of it and how the augmentation of contact lists may be succeed, through adding contacts-of-contacts.

6. CONCLUSIONS

In this paper we have presented an in-depth analysis of the metadata hidden inside almost 16 millions of documents, which we obtained through public web servers. We highlighted a series of privacy risks involved in sharing documents that carry sensitive information in their metadata section. Additionally, shown that it is possible to extract social cliques of users that collaborate in the production of documents, by simply correlating the author fields found in the metadata of documents. We were able to escalate our attack on privacy by successfully identifying some of these cliques on Twitter. This allows us to cross-correlate the

public activities of someone on Twitter with their private activities, like their contribution in the editing of a particular document. Our study raises major concerns about the risks involved in privacy leakage, due to metadata embedded in documents that are stored in public web servers.

7. REFERENCES

[1] Dodgy dossier: Microsoft word bytes tony blair in the butt. http://www.computerbytesman.com/privacy/blair.htm.

[2] Gnu libextractor. http://www.gnu.org/software/libextractor/.

[3] Hachoir projects. http://bitbucket.org/haypo/hachoir/wiki/Home.

[4] The hunt for file format vulnerabilities. http://www.symantec.com/connect/blogs/hunt-file-format-vulnerabilities.

[5] Iana application media types. http://www.iana.org/assignments/media-types/application/.

[6] Metadata in arizona public records can't be withheld. http://yro.slashdot.org/story/09/10/30/1539241/Metadata-In-Arizona-Public-Records-Cant-Be-Withheld?from=rss.

[7] Microsoft office metadata. http://www.document-metadata.com/microsoft-office-metadata.html.

[8] Poppler. http://poppler.freedesktop.org/.

[9] The risk of sharing in asia - may 2005. http://www.workshare.com/downloads/whitepapers/.

[10] Search protocol reference. http://code.google.com/apis/searchappliance/documentation/64/xml_reference.html.

[11] What are the most popular non-html format files on the web? http://www.google.com/help/faq_filetypes.html#popular.

[12] Workshare global security threat report january - april 2007. www.workshare.com/go/research/07aprilthreats.pdf.

[13] N. Agrawal, W. J. Bolosky, J. R. Douceur, and J. R. Lorch. A five-year study of file-system metadata. *Trans. Storage*, 3(3):9+, 2007.

[14] T. Aura, T. A. Kuhn, and M. Roe. Scanning electronic documents for personally identifiable information. In *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 41–50, New York, NY, USA, 2006. ACM.

[15] S. Byers. Information leakage caused by hidden data in published documents. *IEEE Security and Privacy*, 2(2):23–27, 2004.

[16] A. Castiglione, A. De Santis, and C. Soriente. Taking advantages of a disadvantage: Digital forensics and steganography using document metadata. *J. Syst. Softw.*, 80(5):750–764, 2007.

[17] A. Castiglione, A. D. Santis, and C. Soriente. Security and privacy issues in the portable document format. *Journal of Systems and Software*, 83(10):1813 – 1822, 2010.

[18] A. J. Clark. Document metadata, tracking and tracing. *Network Security*, 2007(7):4 – 7, 2007.

[19] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[20] T. Karagiannis and M. Vojnovic. Behavioral profiles for advanced email features. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 711–720, New York, NY, USA, 2009. ACM.

[21] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[22] H. Lin-Shung, W. Zack, E. Chris, and J. Collin. Protecting Browsers from Cross-Origin CSS Attacks. In *CCS 10: Proceedings of the 17th ACM Conference on Computer and Communications Security*, New York, NY, USA, 2010. ACM.

[23] L. Lu, V. Yegneswaran, P. Porras, and W. Lee. Blade: an attack-agnostic approach for preventing drive-by malware infections. In *CCS '10: Proceedings of the 17th ACM conference on Computer and communications security*, pages 440–450, New York, NY, USA, 2010. ACM.

[24] Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. Polyphonet: An advanced social network extraction system from the web. *Web Semant.*, 5(4):262–278, 2007.

[25] P. Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):211–223, October 2005.

[26] B. Park, J. Park, and S. Lee. Data concealment and detection in microsoft office 2007 files. *Digital Investigation*, 5(3-4):104 – 114, 2009.

[27] J. Park and S. Lee. Forensic investigation of microsoft powerpoint files. *Digital Investigation*, 6(1-2):16 – 24, 2009.

[28] J. Park, B. Park, S. Lee, S. Hong, and J. H. Park. Extraction of residual information in the microsoft powerpoint file from the viewpoint of digital forensics considering percom environment. In *PERCOM '08: Proceedings of the 2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications*, pages 584–589, Washington, DC, USA, 2008. IEEE Computer Society.

[29] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose. All your iFRAMEs point to us. In *Proceedings of the 17th USENIX Security Symposium*, pages 1–16, 2008.

[30] G. Rossum. Python reference manual. Technical report, Amsterdam, The Netherlands, The Netherlands, 1995.

[31] S. Smith and J. Marchesini. *The Craft of System Security*. Addison-Wesley Professional, 2007.

[32] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *SP '10: Proceedings of the 2010 IEEE Symposium on Security and Privacy*, pages 223–238, Washington, DC, USA, 2010. IEEE Computer Society.