Digging up Social Structures from Documents on the Web

Eleni Gessiou*, Stamatis Volanis†, Elias Athanasopoulos‡, Evangelos P. Markatos†, and Sotiris Ioannidis†
*Polytechnic Institute of New York University, US
gessiou@cis.poly.edu
†FORTH-ICS, Greece
{sebolani, markatos, sotiris}@ics.forth.gr
‡Columbia University, US
elathan@cs.columbia.edu

***Abstract*—We collected more than ten million Microsoft Office documents from public websites, analyzed the metadata stored in each document and extracted information related to social activities. Our analysis revealed the existence of exactly identified cliques of users that edit, revise and collaborate on industrial and military content. We also examined cliques in documents downloaded from Fortune-500 company websites. We constructed their graphs and measured their properties. The graphs contained many connected components and presented social properties. The a priori knowledge of a company's social graph may significantly assist an adversary to launch targeted attacks, such as targeted advertisements and phishing emails. Our study demonstrates the privacy risks associated with metadata by cross-correlating all members identified in a clique with users of Twitter. We show that it is possible to match authors collaborating in the creation of a document with Twitter accounts. To the best of our knowledge, this study is the first to identify individuals and create social cliques solely based on information derived from document metadata. Our study raises major concerns about the risks involved in privacy leakage due to document metadata.**

I. Introduction

Millions of documents are created and shared over the Internet every day. Popular formats for these files are Microsoft Word, Excel, PowerPoint and PDF. These documents contain much more data than what was intended by their creator. This data is automatically generated by the applications and we refer to it as *metadata*. In most cases, the author of a document is totally unaware of the existence of any metadata associated with it [20], while many users store their files on web servers. Often, due to poor security configurations, these files become accessible to everyone.

Microsoft Office documents include built-in and custom properties in their metadata [9]. Custom document properties, such as the *date completed* and the *author name*, identify the file itself. Built-in document properties, such as *title*, *keywords*, *subject*, and *comments*, identify the document's content. Unfortunately, metadata may contain sensitive information about the person that authored or modified the document. In this paper, we investigate several privacy issues that should be considered when thinking about metadata. First, revealing the *creator* of a document may be used for deriving possible usernames used in web applications, such as social networks and web e-mail. Second revealing the *application used* for the creation of the document may be helpful in determining potential attacks. Note that exploits or computer worms often target specific, known to be vulnerable, versions of an application [32], [24]. Thus, revealing the software and version used to create a document can narrow down an attack targeting a particular user. Third, in the context of forensics, *creator* and *last author* fields may reveal someone's real name, in case they use a nickname to hide their identity [18].

To highlight the importance of metadata we briefly discuss two real-world examples. The most notable example to date is the case of Dodgy Dossier [2], which refers to a document of the British government on Iraq published using Microsoft Word. 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 [7]. 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 cases 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 10 million publicly accessible online documents collected over a period of one year. We quantify the amount of metadata stored in online documents and find sensitive information associated with it. We employ existing libraries and tools to extract, visualize the degree of the metadata diversity and study the social graphs that emerge from this information. Finally, we successfully cross correlate the social graphs associated with metadata with actual graphs from social networks, such as Twitter. An adversary can take advantage of this information for launching targeted attacks, such as brute-force attacks against SSH or other services [16]. Moreover, by profiling individuals in social graphs, targeted spam campaigns could be deployed, focusing on individual or company characteristics.

Contributions. Our main contributions are the following:

- We collected a large dataset consisting of over 10 million

online documents and exposed all stored metadata. Using information solely present in metadata, we developed techniques for identifying social cliques, comprised of users that collaborate in the production of a particular document.
- We focused our study on social graphs derived from authors working in Fortune 500 companies.
- We searched Twitter for all exported social cliques identified in the documents' metadata. Our search successfully cross-correlated members of cliques with Twitter users. This unveiled that members of a clique form groups of followees and followers in Twitter.

II. Methodology

One rich source for finding online documents is Google. We created a custom web scraper using the Python scripting language, which is able to parse search results produced by Google. According to Google's policy, the search engine does not serve more than 1,000 results per query [12]. We therefore used an English dictionary to produce a series of queries which can generate a large set of search results. Each query was composed of one dictionary word and the *filetype* directive used by the Google search engine. This directive assists in producing a result-set composed solely of specific filetypes.

We extracted the URLs pointing to documents based on their extension (.doc, .xls and .ppt). Once a file was spotted in a set of Google results, we downloaded the file and verified that the extension of the file matches the MIME type [6] which is advertised in the HTTP response issued by the host of the file. It has been documented that many web servers are not configured properly [23] to serve all files with the correct MIME type. Also, 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. We discarded all documents for which the file extension did not match the advertised MIME type to avoid bias in the sample due to issues not directly related with privacy leakage. For each downloaded file we proceed and extract all possible metadata. We used the hachoir-metadata [4] and libextractor [3] libraries for extracting all metadata.

A. Sample Properties

Using the technique outlined above we collected more than 5 million MS Word documents, about 2.5 million MS Excel and more than 2.5 million MS PowerPoint documents. Overall, our sample contained over 10 million distinct documents. All documents were hashed using the MD5 cryptographic function, to remove potential duplicates. There was a fairly distinct distribution of the various filetypes. Notice that MS Word files dominate the set, compared to MS Excel and MS PowerPoint files. Our intuition is that 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 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 three non-HTML filetypes considered the most popular to date [13] and thus we consider our study highly representative.

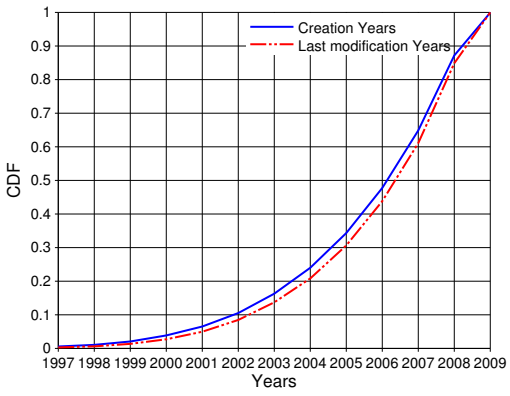

Fig. 1. The CDF for creation and last modification years for Microsoft Word files. The solid blue is the *creation year* and the red, dashed line is the *last modification year*.

B. A Peek into Document Metadata

We now present some of our findings relating to the collected metadata. Note, that due to space restrictions we do not describe further observations of the collected sample.

Figure 1 shows the CDF of *creation year* and *last modification year* of all Word documents in our sample. *Creation year* and *last modification year* seem to present an increase in recent years which may be due to several reasons. 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. Last, Google returns the more recent documents than old ones.

Another observation based on the sample, is that the application used for handling Word documents is mainly the Microsoft Word software, and especially, Microsoft Office 2000 (Office 9.0) is the most commonly used version in our dataset. We also notice that almost 93% of Word files use the default template of Microsoft, Normal.dot. However, apart from the default 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 *creator*, *last author* 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

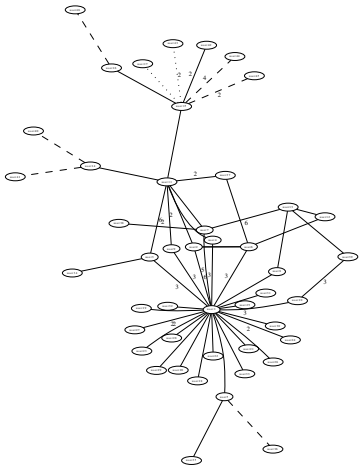

Fig. 2. Clique of company A. The dotted and the dashed edges are the connections of company A with other companies.

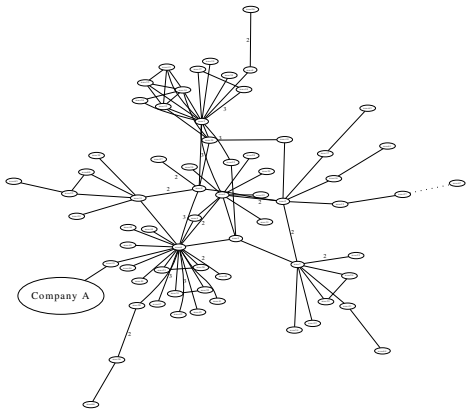

Fig. 3. Clique of company B. The node labeled "Company A" represents the graph of Company A depicted in Figure 2.

above mentioned metadata, as the organization's CEO.

In Table I we present all interesting types of metadata found in Word documents, along with 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.As far as military Word documents are concerned, every one in two includes *company* information and among them the more frequent are names for military departments. We also found 1,500 distinct names of individuals who took part in the creation/modification of documents, all downloaded from a specific `.mil` domain. All names are formatted in a similar fashion: "name.surname", e.g, "john.doe". In case of common names an ascending number is added, e.g, "john.doe1", "john.doe2", etc. Notice that the metadata of these documents reveal the scheme used in formatting usernames by this department, easing brute-force attacks against SSH [16].

It has been previously reported that companies create sample PowerPoint files which serve as *templates* for future use [30]. By inspecting our dataset we see that an initial *template* is used multiple times within a *company*. We calculated the average life time of PowerPoint files by finding the average difference between *last modification date* and *creation date*. PowerPoint files have a five times longer life-span than Word files. An interesting finding that justifies the longer life time of PowerPoint files is the following. We discovered several individuals who are the *authors* in more than one PowerPoint files. Those files have the same *creation date* but different *last modification dates*. So, we speculate 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. Moreover, we observe that specific individuals/professors create one initial presentation for their classes and each year they enhance their slides with new content. Considering the above, companies and academic lecturers seem to be among the main users of PowerPoint files.

TABLE I
THE PERCENTAGES OF METADATA FIELDS IN MILITARY AND GOVERNMENTAL WORD DOCUMENTS IN COMPARISON WITH THE TOTAL NUMBER OF WORD DOCUMENTS.

Metadata	**% .mil**	**% .gov**	**% all**
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
Template used	96.68	97.59	96.98
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
Company	45.02	35.26	31.90

III. DIGGING UP SOCIAL STRUCTURES

A detailed look in our collected dataset showed that a particular individual is the author in fourteen different PowerPoint documents, three different Word documents and three 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. This observation led us to investigate the possibility of extracting social structures by just inspecting documents' metadata. To conduct an initial study, we used all the Excel files of our dataset. For each document we located the metadata fields *creator* and *last author*. We searched for all documents that also list these authors in the respective metadata fields. If two documents listed the same *creator* or *author* and have

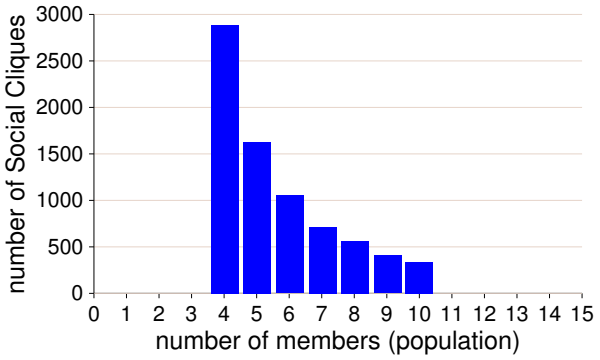

Fig. 4. Distribution of the populations of social cliques. The horizontal axis shows the number of members inside a social clique, and the vertical axis indicates the number of social cliques that correspond to each population.

been downloaded from the same web server (indicated by the domain of the URL) then we considered that these authors collaborate. In this way we created 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. A graph containing nodes from the same domain and their linkage is referred to as a *clique*.

In Figures 2 and 3 we show two example cliques. Note that these two example-graphs have been constructed manually. 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. We proceeded and automatically constructed 10,000 social cliques with at least four members and at most four hops depth. This means that the maximum route-length connecting two individual authors, if such route exists, is of length four. By setting the threshold of four in both cases (#members and #hops), we aim at both meaningful and easy veriable results. The distribution of the population of the social cliques extracted is shown in Figure 4. There are 1,481 social cliques having more than 15 members each. We choose to exclude these groups from the graph. Only 6 social cliques, not shown in the graph, consist of more than 500 members. The most populated social cliques are one with 3,886 members and another with 3,923, not shown in the graph as well.

IV. Fortune-500 Companies

We applied the techniques outlined in Section II-B in documents associated with high profile companies. We did this for two reasons. First, we seek to identify if major companies do expose sensitive information via documents' metadata, which may have serious security implications. If a social graph of a high profile company is exposed, then an attacker can send a malicious document to the most highly-connected nodes of the graph. Thus, the adversary increases the probability for spreading the malicious document fast in the company network. Also, a social graph can be of valuable help for spammers sending targeted phishing emails. Another reason for our decision about high profile companies is based on our intuition that large companies may collaborate with each other. We aim at exposing these collaborations by studying just information found in metadata.

We used the Fortune-500 company sites of 2010 as listed in CNN.com. [1] We selected and extracted from our original dataset all Word documents associated with these companies. For each of the Fortune-500 company sites, we first gathered all Word files that were downloaded from the company's web server, indicated by the domain of the URL. For each document in the set, we located the metadata fields *creator*, *last author* and *revision history*. The *revision history* fields have the following format: Author 'name' worked on 'computer's location' (e.g., Author 'User' worked on 'C:\My Documents\confidential.doc'). If two documents list the same name in one or more aforementioned fields, we assume that these authors collaborate. Note, that although the queries we used for collecting our dataset were not targeted towards any particular company, we managed to extract a total of 79 cliques out of the Fortune-500 companies. A determined adversary could potentially target the site of a particular company to achieve optimal results, by downloading a very precise set of documents.

Each created clique consists of more than two nodes. The average number of nodes is $\sim$29 nodes per clique and the average degree is $\sim$1.08 edges per node. The low average degree per node suggests that cliques are not connected. The most populated clique contains 860 nodes, 899 edges and 246 connected components, and belongs to a leading producer of computer software. The largest connected component of this clique is depicted in Figure 6. Nodes correspond to company's employees and edges to social or person-to-person relationships among employees of the particular company. Note, that all graphs are anonymized for privacy reasons. Overall, 50 out of the 79 cliques contain more than one connected components.

It is interesting to identify whether the metadata graphs depict social networks or random graphs. We considered all the connected components that consist of more than 4 nodes, to examine their properties. The average clique degree is $\sim$3 and the average diameter is $\sim$3 with $\sim$13 nodes and $\sim$20.5 edges per clique, on average. Also, they have a very high average clustering coefficient equal to $\sim$0.54.

Apart from the social structure, we were also interested to see the document distribution among the authors of each company and find the most frequent document publishers. By creating the document distribution, we can extract a broader set of employees and the number of documents they have worked on. An adversary could take advantage of such a distribution, since it could ease him choose his targets. The most active author would have many documents and thus more information

[1] http://money.cnn.com/magazines/fortune/fortune500/

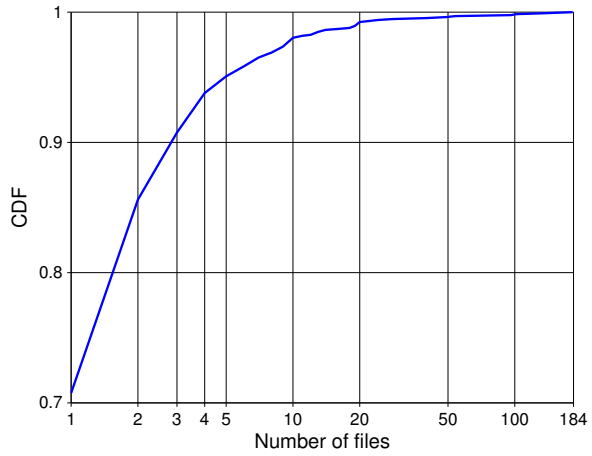

Fig. 5. CDF of document distribution in the most populated clique, which consists of 860 nodes and 899 edges. The majority of the nodes has participated only in one document. There are some nodes that have worked on some tens, even hundreds of documents.

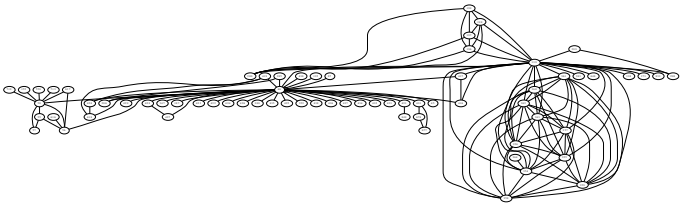

Fig. 6. Example of a connected component which is part of the most populated clique. The nodes are denoted in the graph with increasing numbers. It consists of 70 nodes, 139 edges and has an average degree equal to 3.9714. The node with the highest betweenness(=0.84) and closeness(=0.66) is #46, while #139 node has the highest degree centrality(=25).

about the company in his possession. The attacker could use the victim's name for guessing the username or the password using a brute-force attack [27]. A representative distribution is depicted in Figure 5, indicating that only a few members of the clique have actually collaborated in many documents (>10).

Examples of Fortune-500 company entire graphs are depicted in Figure 7 and Figure 8. More specifically, Figure 7 presents the graph of one leading producer of personal computer and related equipment. In the figure, node #2 has the highest degree, betweenness and closeness centrality [28]. An interesting fact is that we were not able to find any information in this company's web site about the individual represented by node #2. A more in-depth examination shows that this node participates in the graph because it is present in the *revision history* fields. This leads us to assume that it is a company's contractor or collaborator, rather than an employee. Thus, this case suggests that *revision history* fields could disclose collaborations between two companies: the initiator-company creates a document (*creator*), this document is then modified from both sides (*revision history*), and finally returns to the owner for inspection (*last author*). Finally, the owner has the right to upload the document to their server, where it can be downloaded by anyone.

V. Identifying Users in Social Networks

We seek to identify if we can efficiently fingerprint users [34] that collaborate in the production of documents by locating them in Twitter[2]. 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, out of which 51,709 are unique. For the rest of our experiments we exclude the 27 most frequently appearing names, like "Preferred Customer" and 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 include names that contain generic words such as "user", "administrator" and "department", 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. We searched all these individuals at Twitter and found their followers and those they follow. We sought to extract correlation that would verify that people collaborating in the editing of a particular document can be identified in Twitter.

Overall, we examined 575 cliques, containing at total 14,969 people. We found that 1,911 people among them own a Twitter account. Overall, we managed to find 115 social cliques hosting people who have common friends in Twitter. For example, in one case 2 out of 3 individuals belonging to a social clique are Twitter friends and also have 39 friends in common. In another case, 2 people out of 19 in a social clique are Twitter friends and have 4 friends in common. There are also 3 social cliques, containing 131, 500 and 297 individuals

[2]Twitter is also used for professional purposes: http://www.nytimes.com/2009/08/26/technology/internet/26twitter.html

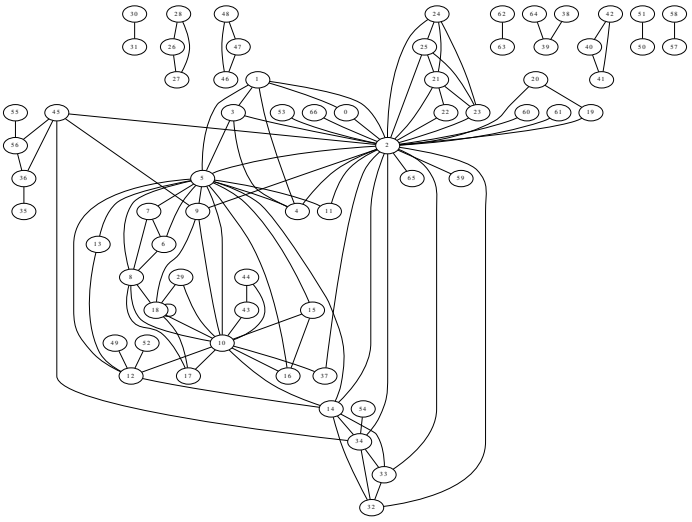

Fig. 7. Example of a populated graph which consists of 67 nodes, 108 edges and 9 connected components. The largest component, as it being depicted, has 47 nodes, 93 edges and an average degree of 3.9574. Node #2 has the highest degree(=26), betweenness(=0.6) and closeness centrality(=0.62).

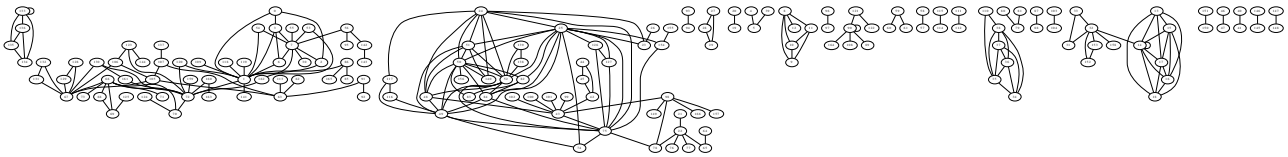

Fig. 8. Example of a populated graph which consists of 167 nodes, 228 edges and 24 connected components.

each, that their members have common Twitter friends and moreover there is a couple of individuals in each clique that connects to each other with direct Twitter friendship. We also found 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 an identity derived from metadata with one registered in Twitter.

A. Fortune 500 Companies

The same procedure is conducted for the cliques created by the Fortune 500 companies, which is described in Section IV. We used the most populated cliques, 12 in number, that correspond to major companies. These 12 cliques consist of 1,561 unique names. Again, we excluded names that contain generic words such as "company", "employee", "computer", etc. Finally, 1,508 unique full names were checked in Twitter for examining if they own an account. Out of them, 798 individuals seem to have a Twitter account. We got all their friends and followers, and checked for direct and indirect connections between them. In order to avoid false positives, we excluded popular Twitter users [22]. The results show that in the most populated clique, which consist of 860 nodes, there are 1,843 indirect and 11 direct connections. In other words, there are 11 pairs of people that both belong to the same clique and are friends at Twitter, and moreover there are 1,843 more pairs that have 1 common friend. The rest of the cliques contain 6 to 318 indirect connections. An indirect connection defines an implicit friendship between two Twitter users that share common friends, but they are not directly connected. Users that share a significant amount of common friends have high probability of being also friends.

VI. Leakage Reduction

Throughout this paper, we have highlighted various privacy risks stemming from the exposure of information stored in metadata associated with documents. We will briefly now discuss techniques for reducing the risk and the privacy leakage. First, the sanitization techniques offered by various tools for extracting and scrubbing metadata can significantly reduce metadata leakage [10], [1], [8]. These tools support a wide variety of file formats and can automatically eliminate all metadata information stored in documents. However, metadata has many legitimate uses for sorting, categorizing and indexing user files. Eliminating all metadata is not the optimal solution in all cases. On the other hand, encryption can be applied to ensure that only certain people within a company have access to each corresponding document. Metadata analysis done, during our study, on PDF documents has revealed that they contain dramatically less metadata information than all other formats. For example, PDFs do not contain *revision history* in the format that MS Office documents do. Thus, one can convert Microsoft Office documents to PDFs. However, using PDF is sometimes hard for collaborative editing. Also, in cases that it is suitable, the usage of RTF files, instead of Word documents, can significantly reduce the leakage, although RTF files support a limited set of text decoration and customization. In our initial dataset, there were some documents that had *.doc* extension but were actually RTF files. We noticed that none of them contained any metadata. Finally, a good practice is to carefully review all configuration files associated with web servers and either prevent directory listing in folders hosting sensitive documents, or offer to serve only files that are already sanitized.

VII. Related Work

Byers et al. were one of the first to conduct research for metadata, counting the hidden words in a few thousand of documents, but did not take into account all available kinds of metadata, and their sample was much smaller than the one used in this paper [17]. LeakHunter is a tool which finds personally identifiable information that may be stored in documents. It addresses similar problems to the ones we have highlighted in this paper. In a similar fashion, researchers have explored metadata collected by the Operating System's filesystem [15]. In the context of privacy risks due to metadata, several incidents that demonstrate a series of security breaches and sensitive information disclosures that have recently become a serious threat to many organizations around the world are presented in [14]. Among other findings [11] indicates that business users in Asia are unaware of the risk of metadata. Similarly, the authors of [19] 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.

Symantec [5] shows that the majority of malicious Trojans exploiting file formats in 2009 was primarily in Word documents (67%), PowerPoint files (17%), SpreadSheet files (3%) and PDF documents (3%). This observation was one of the reasons that led us to select these particular formats for our study. Many real-life and potential incidents concerning hidden data in these formats are presented in the 13th chapter of [33]. Problems due to revision history in Word's metadata 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 quantifies the amount of this information.

The authors of [31] developed the PRIX (PPT Residual Information eXtractor) tool. Its aim is to identify residual information in PowerPoint documents. In a followup work [30], PRIIX (Powerpoint Residual Information & Identifiers eXtractor) adds slide and object identifiers extraction. Data concealment and detection in Microsoft Office 2007 files that use Office Open XML (OOXML) is studied in [29]. The paper is proving that someone can indeed hide data in such files and presents algorithms for finding them. In a similar fashion, metadata has been also used for steganography [18].

There is a considerable amount of previous work in the fields of extraction and analysis of social networks. P. Mika presents Flink [26], which constructs and visualizes social networks by using information from sources such as web pages, emails, and publication archives. Polyphonet [25] presents a series of methods for obtaining a social network using a web search engine and are used in order to enhance scalability. Recently, some steps towards characterizing social networks emerged from e-mail exchange have been done, such as [21] that presents behavioral profiles of it and how the augmentation of contact lists may be succeed, through adding contacts-of-contacts.

VIII. Conclusions

In this paper we have presented an in-depth analysis of the metadata hidden inside 10 million documents present in public web sites. We highlighted a series of privacy risks involved in sharing documents that contain sensitive information in their metadata. Additionally, we showed that it is possible, using information found in metadata fields, to extract social cliques of users that collaborate in the creation and editing of documents. We were able to escalate our attack by successfully identifying some of these cliques in Twitter. Our study raises major concerns about the risks involved in privacy leakage, due to metadata embedded in online documents.

Acknowledgements

The project was supported in part by the Operational Programme "Competitiveness & Entrepreneurship", Measure "COOPERATION", Marie Curie Reintegration Grant project PASS and by the ForToo Project funded by the Directorate General Home Affairs.

References

[1] "Doc scrubber," http://www.javacoolsoftware.com/docscrubber.
[2] "Dodgy dossier: Microsoft word bytes tony blair in the butt," http://www.computerbytesman.com/privacy/blair.htm.
[3] "Gnu libextractor," http://www.gnu.org/software/libextractor/.
[4] "Hachoir projects," http://bitbucket.org/haypo/hachoir/wiki/Home.
[5] "The hunt for file format vulnerabilities," http://www.symantec.com/connect/blogs/hunt-file-format-vulnerabilities.
[6] "Iana application media types," http://www.iana.org/assignments/media-types/application/.
[7] "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.
[8] "Metareveal," http://www.beclegal.com/products.aspx?id=64.
[9] "Microsoft office metadata," http://www.document-metadata.com/microsoft-office-metadata.html.
[10] "Remove hidden data," http://support.microsoft.com/kb/834427.
[11] "The risk of sharing in asia - may 2005," http://www.workshare.com/downloads/whitepapers/.
[12] "Search protocol reference," http://code.google.com/apis/searchappliance/documentation/64/xml_reference.html.
[13] "What are the most popular non-html format files on the web?" http://www.google.com/help/faq_filetypes.html#popular.
[14] "Workshare global security threat report january - april 2007," www.workshare.com/go/research/07aprilthreats.pdf.
[15] N. Agrawal, W. J. Bolosky, J. R. Douceur, and J. R. Lorch, "A five-year study of file-system metadata," in *In Proceedings of the 5th USENIX Conference on File and Storage Technologies. USENIX Association*, 2007.
[16] H. Berghel and D. Hoelzer, "Pernicious ports," *Commun. ACM*, vol. 48, 2005.
[17] S. Byers, "Information leakage caused by hidden data in published documents," *IEEE Security and Privacy*, vol. 2, 2004.
[18] A. Castiglione, A. De Santis, and C. Soriente, "Taking advantages of a disadvantage: Digital forensics and steganography using document metadata," *J. Syst. Softw.*, vol. 80, 2007.
[19] A. J. Clark, "Document metadata, tracking and tracing," *Network Security*, vol. 7, 2007.
[20] T. S. Guidelines, "The sedona guidelines: Best practice guidelines and commentary for managing information & records in the electronic age," in *The Sedona Guidelines*, 2005.
[21] T. Karagiannis and M. Vojnovic, "Behavioral profiles for advanced email features," in *Proceedings of the 18th International Conference on World Wide Web*. New York, NY, USA: ACM, 2009, pp. 711–720.
[22] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010, pp. 591–600.
[23] H. Lin-Shung, W. Zack, E. Chris, and J. Collin, "Protecting Browsers from Cross-Origin CSS Attacks," in *Proceedings of the 17th ACM Conference on Computer and Communications Security*. New York, NY, USA: ACM, 2010.
[24] L. Lu, V. Yegneswaran, P. Porras, and W. Lee, "Blade: an attack-agnostic approach for preventing drive-by malware infections," in *Proceedings of the 17th ACM Conference on Computer and Communications Security*. New York, NY, USA: ACM, 2010, pp. 440–450.
[25] 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.*, vol. 5, 2007.
[26] 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*, vol. 3, no. 2-3, pp. 211–223, October 2005.
[27] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. New York, NY, USA: ACM, 2007, pp. 29–42.
[28] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, vol. 32, pp. 245 – 251, 2010.
[29] B. Park, J. Park, and S. Lee, "Data concealment and detection in microsoft office 2007 files," *Digital Investigation*, vol. 5, 2009.
[30] J. Park and S. Lee, "Forensic investigation of microsoft powerpoint files," *Digital Investigation*, vol. 6, pp. 16 – 24, 2009.
[31] 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 *Proceedings of the 2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 584–589.
[32] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose, "All your iFRAMEs point to us," in *Proceedings of the 17th USENIX Security Symposium*, 2008, pp. 1–16.
[33] S. Smith and J. Marchesini, *The Craft of System Security*. Addison-Wesley Professional, 2007.
[34] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, "A practical attack to de-anonymize social network users," in *Proceedings of the 2010 IEEE Symposium on Security and Privacy*. Washington, DC, USA: IEEE Computer Society, 2010, pp. 223–238.