A Systems Approach to Cyber Security
A. Roychoudhury and Y. Liu (Eds.)

doi:10.3233/978-1-61499-744-3-107

Scholarly Digital Libraries as a Platform for Malware Distribution

Nir Nissim[1], Aviad Cohen[1], Jian Wu[2], Andrea Lanzi[3], Lior Rokach[1], Yuval Elovici[1] and Lee Giles[2]
[1]*The Malware Lab at the Cyber Security Research Center (CSRC), Ben-Gurion University*
[2]*The Pennsylvania State University*
[3]*University of Milano*

Abstract. Researchers from academic institutions and the corporate sector rely heavily on scholarly digital libraries for accessing journal articles and conference proceedings. Primarily downloaded in the form of PDF files, there is a risk that these documents may be compromised by attackers. PDF files have many capabilities that have been widely used for malicious operations. Attackers increasingly take advantage of innocent users who open PDF files with little or no concern, mistakenly considering these files safe and relatively non-threatening. Researchers also consider scholarly digital libraries reliable and home to a trusted corpus of papers and untainted by malicious files. For these reasons, scholarly digital libraries are an attractive target for cyber-attacks launched via PDF files. In this study, we present several vulnerabilities and practical distribution attack approaches tailored for scholarly digital libraries. To support our claim regarding the attractiveness of scholarly digital libraries as an attack platform, we evaluated more than two million scholarly papers in the CiteSeerX library that were collected over 8 years and found it to be contaminated with a surprisingly large number (0.3%-2%) of malicious scholarly PDF documents, the origin of which is 46 different countries spread worldwide. More than 55% of the malicious papers in CiteSeerX were crawled from IP's belonging to USA universities, followed by those belonging to Europe (33.6%). We show how existing scholarly digital libraries can be easily leveraged as a distribution platform both for a targeted attack and in a worldwide manner. On average, a certain malicious paper caused high impact damage as it was downloaded 167 times in 5 years by researchers from different countries worldwide. In general, the USA and Asia downloaded the most malicious scholarly papers, 40.15% and 27.9%, respectively. The top malicious scholarly document downloaded is a malicious version of a popular paper in the computer forensics domain, with 2213 downloads in a worldwide coverage of 108 different countries. Finally, we suggest several concrete solutions for mitigating such attacks, including simple deterministic solutions and also advanced machine learning-based frameworks.

Keywords Scholarly, Digital, Library, Paper, PDF, Malware, Malicious, Attack, Distribution.

1. Introduction

The number of scholarly documents (English-language) accessible on the Web is enormous, estimated at 114 million PDF documents in 2014 [9], of which over 27 million (~24%) can be freely accessed without payment or subscription [9]. These documents are freely accessible in part because researchers publish draft versions of their papers on

their professional homepages (often within the domains of universities), before the final versions are published by the publishers. Researchers also publish their research on their homepages to increase exposure, reach researchers around the world, and gain citations and recognition for their work [10], [11]. In order to assist researchers, many scholarly digital libraries and search engines collect and index the author's version. Thus, the papers can be freely downloaded worldwide. This free collection of scholarly documents is a valuable resource for most researchers and academics who may not have a comprehensive subscription to all publishers' content.

Figure 1 presents a snapshot of search results for a searched paper using Google Scholar. At the bottom of the page, one can access all 15 versions of the paper, already indexed by Google Scholar, simply by clicking on the blue "All 15 versions" link; thereby, free and convenient versions, are literally at the user's fingertips, as seen in Figure 2.

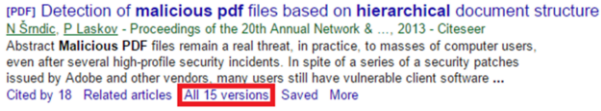

Figure 1. Google Scholar's search results for a given academic paper, including 14 additional versions of the paper.

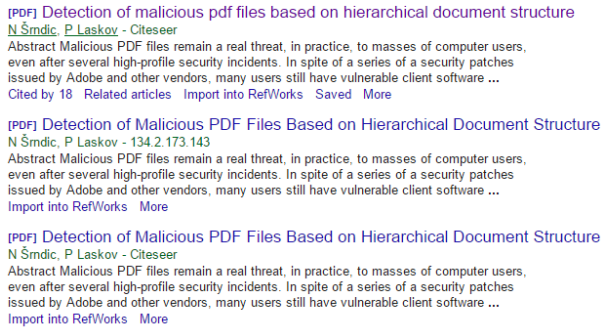

Figure 2. Some of the additional versions of the searched paper, including those available for free.

Researchers heavily use scholarly digital libraries to access and download scholarly documents. For example, according to a survey by EBLIDA[1], the total number of academic libraries in Europe is 5,974; however, this number is far from complete given that it is based on information provided by only 25 countries participating in the survey. Nevertheless, the number of registered users of these libraries is 39,328,294. As Europe represents only part of the world's research activity, the global use of scholarly digital libraries is much higher.

Universities are considered to be highly reputable institutions that primarily focus on research and the goal of which is to contribute new and valuable knowledge to the world. Therefore, they are considered a trusted content source without malicious intent. Correspondingly, the Websites of their academics and researchers (which reside on the institution's network domain) are also considered to contain only trusted content, free of

[1] http://www.eblida.org/activities/kic/academic-libraries-statistics.html

malicious PDF files. In a circular fashion, academic digital libraries tend to harvest these allegedly trusted sites without hesitation or fear and therefore do not even scan them to detect malicious content[2]. In addition, this reputation as sources of trusted scholarly documents makes digital libraries an attractive platform from which to take advantage of and distribute malicious PDF files; researchers' Webpages have become a target that can be used to launch attacks[3]. In addition, researchers, professors, and research students are naturally attractive candidates for attack, because, due to the nature of their work, they have access to confidential and sensitive information, such as nuclear knowledge, medical records [32][33], aviation, and educational records and materials (student data, exams). Moreover, researchers collaborate with governmental agencies and industry, which allows them access to national and confidential information from governments (such as computational criminology), national institutions, and companies (such as strategic information).

Recent studies have presented many methods of improving the detection of malicious PDF files [1], [2]. These studies focused on detection techniques based on analyzing the malicious PDF files when they have already been downloaded to the host machine. To the best of our knowledge, no study addressed the issue at the stage one step before downloading, a step at which it might be possible to prevent malicious PDF files from being mass-distributed through legitimate channels and exiting platforms, and thus, markedly improve the detection of malicious PDF files, including those found on popular, well-known, and extensively used sources of PDF files, such as scholarly digital libraries. These libraries can be intentionally used as a free and very successful platform for distributing PDF malware quickly and easily to a desired group of victims with access to valuable information. An academic paper arouses little suspicion, particularly if an attacker wants to distribute a new 0-day attack quickly in the shape of a benign PDF file. 0-day attack[4] utilizes new attack techniques or new vulnerabilities[5] that are difficult to detect, particularly by the antivirus tools commonly used by organizations such as universities and academic digital libraries for scanning PDF files. Thus, these libraries can easily be used as a new and convenient platform for distributing 0-day attacks. The contributions of our paper include:

1. A demonstration of the vulnerability of digital libraries and also an estimation of the extent of malicious use of scholarly digital libraries. Specifically, we perform a retro-perspective analysis of the papers that were collected by CiteseerX over a period of eight years. Using current antiviruses, we can assess which paper contained malware when it was indexed.
2. An evaluation of the impact damage of malicious documents published in a scholarly digital library.
3. Additional distribution attack approaches that can be used by attackers to leverage these digital libraries.

In addition to the above contributions we suggest methods for mitigating the problem we identified:

[2] According to CiteseerX team which are part of the authors in this paper.

[3] http://www.nytimes.com/2013/07/17/education/barrage-of-cyberattacks-challenges-campus-culture.html?pagewanted=all&_r=0

[4] http://www.bullguard.com/bullguard-security-center/pc-security/computer-threats/what-are-zero-day-attacks.aspx

[5] http://www.pctools.com/security-news/zero-day-vulnerability/

- Compatibility check of PDF files so they can be correctly opened by users before their publication. (96.5% of malicious PDF files are incompatible).
- Re-check of re-uploaded PDF Files
- Periodic review of a library's files in the library when a new PDF malware/vulnerability-exploitation is identified.
- Machine learning-based methods for enhancing the detection of malicious PDF files.

2. Background

As indicated previously, the Web contains more than 114 million scholarly documents [9], and this number represents a significant attack power for adversaries who want to utilize the fact that scholarly digital libraries are considered trusted and that their content (PDF files) is used and downloaded by many users worldwide. In order to grasp the potential harm that can be done by malicious PDF files existing in a scholarly digital library, we briefly present targeted attacks through scholarly digital libraries using malicious PDF files. Then, we present the possible attacks that can be launched by a malicious PDF file mistakenly considered a benign scholarly document, and the techniques used to achieve this. Thereby, we aim to raise the awareness of scholarly digital libraries, as well as of innocent researchers and readers, of the power of a malicious PDF file, so that they will improve their security level.

2.1. Targeted Attacks via Scholarly Digital Libraries using Malicious PDF Files

Sophisticated attackers interested in sensitive and novel knowledge about a specific domain, such as nuclear energy, can launch a targeted attack by inserting an attractive, yet malicious, paper that addresses nuclear energy into digital libraries, engaging and tempting researchers to download the paper. It is noteworthy that the attacker does not need to be a co-author of the paper. Our investigation showed that most scholarly digital libraries (such as Google Scholar) crawl academic Websites and index the papers they find, disregarding any mismatches between the author's affiliation and the Website that stores the paper. Thus, an attacker can take a popular paper written by someone else, inject malicious code into it, and upload it to a Website. When the victim opens the malicious PDF file, a malicious code will be executed in the computer. This malicious code will allow the attacker to extract data from the victim's machine and send it to a remote server controlled by the attacker.

This attack is within the realm of reality, for the previously mentioned reasons, as well as because users consider non-executable files safer than executables, and thus are less suspicious of PDF files, especially when downloaded from popular and trusted scholarly sources. Unfortunately, non-executable files such as PDF files are as dangerous as executable files, since their readers can contain vulnerabilities that, when exploited, can allow an attacker to execute malicious actions on the victim's computer. F-Secure's 2008-2009 report[6] indicates that the most popular file types for targeted attacks in 2008-2009 were PDF and Microsoft Office files. Note that since that time, the number of targeted attacks on Adobe Reader has almost doubled. In the following section, we elaborate on several of the most common techniques and attacks involving the use of malicious PDF files.

[6] http://www.f-secure.com/weblog/archives/00001676.html

To demonstrate the damage that can be caused by malicious PDF files, we refer to a famous incident involving the Israeli Ministry of Defense (IMOD) that took place on January 15, 2014, which provides an example of a new type of targeted cyber-attack. According to various media reports[7] published on January 26, 2014, the Seculert[8] Company reported that it had identified an attack in which attackers sent email messages, allegedly from the IMOD, with a malicious PDF file attachment posing as an IMOD document. When opened, the PDF file installed a Trojan horse that enabled the attacker to take control of the computer.

2.2. Possible Attack Techniques using PDF Files

Before explaining how scholarly digital libraries can be easily used as a platform to leverage and distribute attacks worldwide, we now present some of the many ways PDF files can be used maliciously when created or manipulated by an attacker.

JavaScript code

PDF files may contain embedded JavaScript code or code retrieved from URIs [5], including 3D content, form validation, and calculations. Typically, a malicious JavaScript code in a PDF file attempts to exploit a vulnerability in the PDF viewer in order to divert the normal execution flow to the embedded malicious code. This is achieved by a heap spraying[9] attack. JavaScript also allows the download of an executable file that may contain malicious content. Alternatively, JavaScript code can access Websites, whether malicious or benign.

Code obfuscation and filters

Code obfuscation is used legitimately to prevent reverse engineering of proprietary applications. However, it can be also used by attackers to hide malicious content. Filters are used in PDFs to compress data for encoding and reduce file size and are frequently used by attackers to conceal malicious content. Available filters and their primary purposes are discussed by Baccas and Kittilsen [6], [7].

Embedded Files

A PDF file can contain other file types, such as HTML, JavaScript, SWF, XLSX, EXE, or even another PDF file, which can be used to embed malicious files that are frequently obfuscated. When special techniques are applied, the embedded file can be opened without alerting the user. Recently, Maiorca et al. [3] presented a novel evasion technique called "reverse mimicry," which was designed to evade state-of-the-art malicious PDF detectors based on their logical structure[10] [4]. Mimicry attacks inject malicious content into a benign PDF while maintaining its benign structure. This method can be automated

[7] http://www.israeldefense.co.il/?CategoryID=512&ArticleID=5766.
http://www.ynet.co.il/articles/0,7340,L-4481380,00.html.
http://www.israelhayom.co.il/article/152741

[8] http://www.seculert.com/

[9] **Heap Spraying** - A technique used in exploits to assist random code execution.

[10] PDF logical structure is a hierarchy of structural elements, each represented by a dictionary. See the PDF file structure section.

easily and does not require knowledge of the structural features used in the maliciousness detector.

Form submission and URI attacks

Hamon [8] presented practical techniques that can be used by attackers to execute malicious code from a PDF file. The author showed that security mechanisms, such as the Protected Mode of Adobe Reader X or the URL Security Zone Manager of Internet Explorer, can be easily disabled by changing the corresponding registry key. Moreover, a URI[11] address can be used (instead of a URL), directing the user to any type of file located remotely, including executables. It should be noted that Adobe Reader version X, released in 2011, included a new sandbox isolated environment, Protected Mode Adobe Reader (PMAR), that ensures that malicious code operations cannot affect the operating system. Nevertheless, most organizations (including universities) do not always keep up with the newest versions of PDF readers, and thus, are exposed to many of the well-known attacks.

3. Analyzing Vulnerabilities of Popular Scholarly Digital Libraries

Now, we briefly present the most popular libraries, their market share, and their uniqueness, and then explain what vulnerabilities exist within them. In addition, we present new vulnerabilities that we utilized. We first present three libraries in which we found a vulnerability, and then, we briefly present additional scholarly digital libraries that should be further checked for vulnerabilities, and finally, for the reader's convenience, we provide a summary table of the different scholarly digital libraries and the manner in which they work.

3.1. Google Scholar

Google Scholar[12] is a free public Web search engine for scholarly literature. It consists of nearly 100 million scholarly documents and is considered the largest scholarly digital library, encompassing 87% of these documents [9]. It indexes scholarly literature across publishing sources. Current articles are indexed and can be found when searched. A user clicking on an article that appears on the results page of Google Scholar is usually directed to the article's Web page on the publisher's official Website. In addition to articles on the publisher's Website, other versions of the papers, from other places on the Web are also indexed (e.g., papers from a researcher's Web page on an academic institution's Website).

In order to demonstrate contamination of a digital library such as Google Scholar, we used the Web page of a researcher at a known university (we do not give details for privacy reasons). The articles on the researcher's Web page were indexed by Google Scholar previously and can be accessed by clicking the "All X versions" link under the relevant article in Google Scholar, as shown in Figure 1. With no connection to the researchers' names appearing in Figure 1, after we had obtained another researcher's

[11] **URI** – "a compact string of characters for identifying an abstract or physical resource," RFC2396. It is an extension of URL used for identifying any Web resource (not limited to Web pages).

[12] https://scholar.google.co.il/

permission, we downloaded the most popular paper (a PDF file) from his Web page and injected a malicious JavaScript into it using a PDF editing program called *PDFFill*[13] (such that the malicious JavaScript code is launched when the new article's file is opened). Then, we replaced the benign paper with this new malicious version of the paper on the researcher's website. Now, the malicious paper is available for downloading through Google Scholar using the original indexing information that was neither changed nor updated toward the replacement of the paper behind the published URL. The vulnerability in Google Scholar lies in the indexing mechanism, which checks only the title and author's name and pays no attention to whether a new file was uploaded with the same title and author's name.

As far as we could determine, Google Scholar does not verify that the uploaded paper is related to the researcher's homepage. Thus, a malicious PDF file that carries the same title and authors of a popular paper can easily be created and placed on other Web pages unconnected to a researcher's home page within a university. These malicious papers can be easily promoted with an acceptable payment to Google for a promoted link. Thereby, the attacker uses several elements to launch his attack. First, he takes advantage of the popularity of a particular paper, second he uses the fact that Google Scholar is a trusted source of information, and third he exploits a vulnerability in the Google Scholar indexing mechanism. Consequently, the attacker achieves his attack goals by redirecting the download traffic to his malicious version.

3.2. CiteseerX

CiteSeerX[14] is a growing scientific literature digital library and search engine that focuses primarily on literature in the areas of computer and information science. It is unique in that it collects papers solely from researchers' homepages from the domains of universities and physically stores the papers themselves, in addition to linking to them. The result is that the library contains over four million academic papers in PDF format, and its total size is estimated at about 3.8 terabyte.

According to the way in which CiteseerX collects academic documents, we identified several methods by which a malicious PDF paper could be indexed by a popular digital library. A malicious paper could be uploaded to a researcher's Website directly. This can happen unintentionally if the paper was infected by a malware resident on the computer before it was placed on the Website. Alternatively, the paper could be contaminated using a free, malicious PDF creator that injects malicious code into the edited papers. Another likely scenario is that the researcher's page could be hacked, with the attacker replacing a benign paper with a malicious one. In each of these examples, a malicious paper finds its way to the researcher's homepage within an academic institution's trusted domain, making it available for uploading by CiteseerX as well as to the general public worldwide.

3.3. Social Network Based Scholarly Digital Libraries

Research-Gate[15] (founded in 2008) is a social networking site for scientists and researchers, enabling them to share papers, communicate, and find collaborators. Today,

[13] http://www.pdfill.com/
[14] http://csxstatic.ist.psu.edu/about
[15] http://www.researchgate.net/

it has more than six million members. Research-Gate is also considered an academic digital library as its members can upload and share papers with other members. Academia.edu (launched in September 2008) is a platform for academics for sharing research papers, monitoring their impact, and following researchers in a particular field. A total of 17,896,413 academics have signed up to Academia.edu, adding 5,089,710 papers and 1,450,657 research interests. Academia.edu attracts over 15.7 million unique visitors a month[16]. Research-Gate and Academia.edu are examples of scholarly academic digital libraries affiliated with social networks for researchers whose purpose it is to share data, papers, and knowledge with other researchers.

We created a fictitious profile of a famous researcher through Academia.edu, a process during which we were asked many questions about the researcher and were even asked to upload some of his papers. We uploaded several of his well-known and published papers in order to boost the profile's credibility and gain the trust of colleagues. After several weeks, when the profile was active and papers had been downloaded from the profile, we were able to easily upload a malicious version of the same papers in order to perform an attack. The uploading of an existing malicious PDF file (a non 0-day malicious PDF file) that should have been recognized by an antivirus tool was not stopped by any security mechanisms of the library. Thus, we also show here that social relationships and trust can be sufficient for leveraging a social network-based library for the distribution of a malicious PDF based attack.

3.4. Additional Existing Scholarly Digital Libraries

The following libraries are additional existing scholarly digital libraries that we have not yet checked for vulnerabilities; however, we assume that vulnerabilities exist and should have been further investigated.

Microsoft Academic Search[17] is a free public Web search engine for academic papers and literature, developed by Microsoft Research for the purpose of algorithm research on object-level vertical search, data mining, entity linking, and data visualization. Microsoft Academic Search consists of almost 50 million scholarly documents and is considered one of the top alternatives to Google-Scholar [9].

Web of Science[18] is an online subscription-based scientific citation indexing service maintained by Thomson Reuters that provides comprehensive citation search. It consists of nearly 50 million scholarly documents and is considered, together with MAS, one of the largest academic digital libraries after Google-Scholar [9]. One should note that Web of Science does not index the PDF files, as Google-Scholar does.

PubMed[19] is a free search engine that primarily accesses the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine at the National Institutes of Health maintains the database as part of the Entrez system of information retrieval. PubMed comprises over 24 million citations of biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publishers' Websites.

arXiv is an automated electronic repository and distribution server for research articles, consisting of electronic preprints of scientific papers in the fields of mathematics,

[16] https://www.academia.edu/about
[17] http://academic.research.microsoft.com/
[18] https://apps.webofknowledge.com/
[19] http://www.ncbi.nlm.nih.gov/pubmed

physics, astronomy, computer science, quantitative biology, statistics, and finance, which can be accessed online. Almost all scientific papers within arXiv are self-archived, meaning that they were uploaded by the users themselves.

Table 1 summarizes the details of the interesting aspects mentioned in this section. The largest libraries, Google-Scholar, MAS, and Web of Science, do not rely on papers uploaded by users as they crawl papers from the publishers as well and do not store them. Note that there are several closed group\ libraries within the Darknet, such as Libgen, Sci-hub and Booksc, and we assume that specifically in these not wide-open libraries the probability and percentage of malicious papers is higher than in the known and wide-open libraries. This assumption should be scrutinized in future research.

Scholarly digital libraries	Upload by User (Preprint)	Crawling Publisher	Crawling Authors Homepage (Preprint)	Indexing the PDF content	Store the PDF	Link to the original PDF	Number of Scholarly documents (In Millions)
Google Scholar	No	Yes	Yes	Yes	No	Yes	99.3
Microsoft Academic	No	Yes	Yes	Yes	No	Yes	50
Web of Science	No	Yes	No	No	No	Yes	50
CiteseerX	Yes	No	Yes	Yes	Yes	Yes	4.2
PubMed	No	Yes	No	Yes	No	Yes	24
Research Gate	Yes	No	No	Yes	Yes	No	Unknown
Academia.edu	Yes	No	No	Yes	Yes	No	5
arXiv	Yes	No	No	Yes	Yes	No	1
http://libgen.org/scimag (Darknet)	Yes	Yes	No	No	Yes	No	36
http://sci-hub.org/ (DarkNet)	No	Yes	No	No	Yes	No	Unknown
http://booksc.org/ (DarkNet)	No	Yes	No	No	Yes	No	18

Table 1. Summary of Scholarly digital libraries' details regarding to their crawling, indexing and redirecting approaches to the scholarly documents.

4. Methods

On this section we present the Dataset scanning tools and technical details that allowed us to provide the results and insights of this study.

4.1. Dataset

As part of this collaborative study with the CiteseerX team, we scanned and analyzed the CiteseerX digital library as our dataset. Our goal was to determine whether this platform had already been used, either intentionally by an attacker or unintentionally by an innocent researcher, to distribute malicious PDF files, and in so doing, to measure the extent of harm that can be caused by such a scenario. When we began scanning, the CiteseerX library contained 4,044,118 academic papers in PDF format that were collected up to the end of 2014, from more than 188 different countries over most of the continents, written by 1.3 million disambiguated authors from 4963 different universities.

4.2. Scanning Tool for Malicious files

We used the VirusTotal[20] service to scan the entire CiteseerX library for malicious PDF files. VirusTotal, a subsidiary of Google, is a free online service that provides comprehensive analysis of files and Websites (URLs) by a set of ~57 antivirus engines and Website scanners. VirusTotal allows a user to submit suspicious files for analysis. After the analysis, VirusTotal provides a report that specifies the identification of suspicious files for each of the antivirus engines. When a file is about to be scanned, VirusTotal calculates its hash to determine whether if it was previously scanned. If so, the stored report is provided to the user; otherwise, the file is then uploaded and scanned, and a report is generated when the process is complete. VirusTotal also provides a rich and public API[21] for the submission of files and URL addresses and retrieval of the analysis reports. The public API can be used through several programming languages that assist with automating the submission and report retrieval procedures. Note that we considered a PDF file as a malicious, only if at least 5 different anti-viruses detect it as a malicious file. In addition we emphasize that rather than presenting a novel technique of malicious PDF files detection, the goal of this study is revealing a simple yet very dangerous way by which the scholarly digital libraries can be utilized as a platform for malware distribution.

4.3. Scanning Technical Details

Since only a small percentage of CiteseerX's papers had been previously scanned, we uploaded all of its content, file by file, to VirusTotal, in order to scan the whole library. Three interrelated problems with this approach were encountered: 1) the enormous size of the library; 2) the length of time it would take to upload the entire library to VirusTotal; and 3) VirusTotal's submission limit for regular users of four per minute. A quick calculation showed that the scanning process would take approximately 700 days. To cope with the issue of data size, we took a different approach and used VirusTotal's option to analyze URL addresses of files (URI[22]). When a URI address is submitted to VirusTotal for analysis, it downloads the file that stands behind the address and analyzes it too. However, there is no guarantee that this operation is actually done. The submission of the URI addresses of CiteseerX's articles to VirusTotal for analysis (versus submitting the actual PDF files) facilitated the process and shortened the time it took to upload the entire library (~3.8 TB). To circumvent the limitation of four files per minute, we requested special privileges (a private API key) from VirusTotal that allowed us to perform many more submissions per minute. We used VirusTotal's private API to scan the entire CiteseerX library consisting of 4,044,118 academic papers in PDF format. Initially, we submitted the URI addresses of the PDF files iteratively for analysis. Then, we submitted a request for the analysis reports. The scan of the CiteseerX scholarly digital library was completed in five months.

[20] https://www.virustotal.com/
[21] https://www.virustotal.com/en/documentation/public-api/
[22] http://searchsoa.techtarget.com/definition/URI

5. Scanning results

On this section we present the results and their analysis regarding to the scanning process of the PDF files within CiteseerX library. We provide analysis both in the aspects of crawling and downloading the malicious papers, on the basis of worldwide breakdown.

5.1. Crawled Malicious Papers

Of the 4,044,118 URI addresses of PDF files that were submitted for analysis from the CiteseerX library, only 2,586,820 were actually scanned (the process that was previously described). Of these files, 753 (~0.3%) were found and classified as malicious by VirusTotal's antivirus engines. Figure 3 present the breakdown of the threats identified.

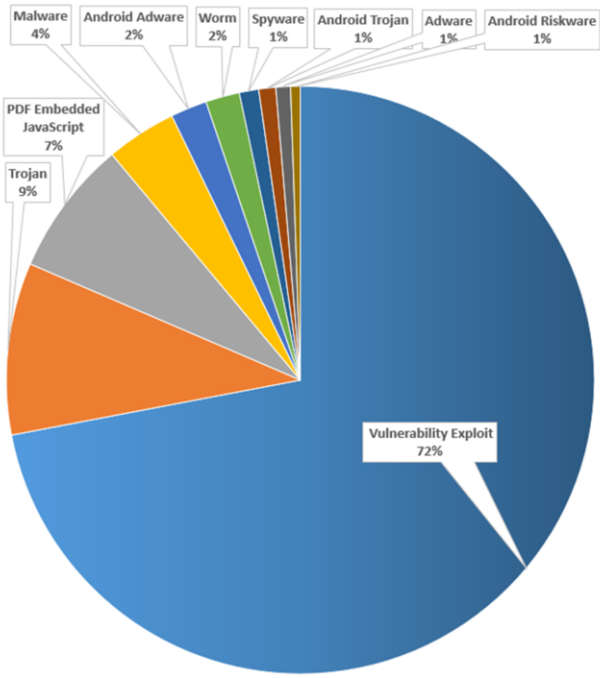

Figure 3. Breakdown of the threats identified among the 753 malicious PDF files found by VirusTotal on the CiteseerX library.

The threats' categories were provided by the identifying antivirus engine. As can be seen in Figure 3, 72% of the malicious files were identified based on vulnerability exploitation[23]. Usually, vulnerability in the PDF file format is exploited utilizing an embedded JavaScript code[24]. 9.5% were classified as a Trojan, a malicious program that when executed performs covert actions that have not been permitted by the user. 7.5% of the malicious files contained JavaScript code that was recognized as malicious.

[23] http://searchsecurity.techtarget.com/definition/exploit
[24] http://blogs.technet.com/b/mmpc/archive/2013/04/29/the-rise-in-the-exploitation-of-old-pdf-vulnerabilities.aspx

JavaScript code can be identified as malicious although it does not exploit any vulnerability and is considered malicious when the code signature is known to represent a malicious code. 3.9% of the malicious files were classified as malware, which means that malicious software (e.g., Exe, PDF, etc.) was found embedded in them. 3.4% of the malicious files contained a threat (Adware[25], Trojan, or Riskware[26]) targeting the Android operating system widely used on mobile devices. 1.9% of the malicious files contained a computer worm[27], which is a malicious program that can propagate by autonomously copying itself from one machine to another. A small percentage of files (1.1%) were classified as Spyware[28], which is a malicious computer program aimed at collecting personal information from the victim's computer. Although it does not damage the victim's computer, it can cause damage to the victim by stealing sensitive information. An Adware is a program that aims to support advertising and operates without the user's permission. An additional 5,775 files were identified as malicious by the Fortinet antivirus, because they contained a suspicious threat called "HTML/Redirector.BK!tr". These files might be malicious since they may direct the user to malicious destinations such as Websites, IP-addresses, and servers. A deeper analysis is required to reach a final decision; however, when the percentage of malicious PDF files in the CiteseerX library rises from 0.3% to 2%, this strengthens even more the phenomenon we are presenting in this paper.

Figure 4 presents the distribution of the malicious scholarly documents according to the geographical location from which they were crawled by CiteseerX scholarly digital libraries. More than 55% of the malicious papers in CiteSeerX were crawled from IP's belonging to USA universities, whereas about 33% were crawled from IP's belonging to European universities.

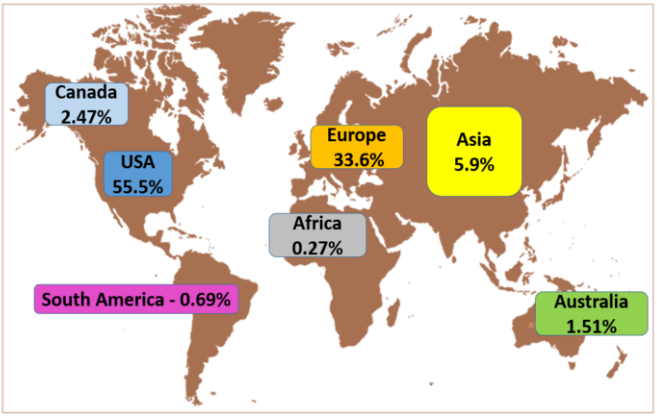

Figure 4. Distribution of the malicious scholarly documents according to the geographical location from which they were crawled by CiteseerX scholarly digital library.

In Table 2, we can see the top 11 European countries in terms of the percentage of malicious scholarly documents crawled from their IP's. Germany was the origin of 10.7% of the malicious papers in CiteSeerX out of the total world's malicious papers,

[25] http://www.pctools.com/security-news/what-is-adware-and-spyware/
[26] http://usa.kaspersky.com/internet-security-center/threats/riskware
[27] http://www.pctools.com/security-news/what-is-a-computer-worm/
[28] http://www.microsoft.com/security/pc-security/spyware-whatis.aspx

and is the origin of more than 31% of the malicious papers in CiteSeerX out of Europe's malicious papers share. It was followed by United Kingdom (6.04%), Holland (2.74%), and France (2.61%). Each of the other European countries not presented here were the origin of less than 0.41% of malicious scholarly documents out of the total world's malicious papers in CiteSeerX.

Country	Percentage
Germany	10.70%
United Kingdom	6.04%
Holland	2.74%
France	2.61%
Austria	2.33%
Luxembourg	2.06%
Sweden	0.82%
Switzerland	0.82%
Denmark	0.69%
Italy	0.55%
Turkey	0.55%

Table 2. Breakdown of the distribution of the malicious scholarly documents according to universities in countries within Europe from which they were crawled by CiteseerX out of total world's malicious papers.

Asia includes several countries, e.g., China, Russia and Korea, which on the one hand are known to have a large population of researchers and on the other were found to be the origin of many malwares. We were surprised to find that only 5.9% of the malicious papers were crawled from IP's belonging to an Asian institution.

In Table 3, we can see some interesting statics regarding Asian countries. Israel is quite a small country with a population constituting 0.1% of the world's population and naturally thus has also a small research community as compared to other countries in the world. Nevertheless, Israel is the origin of 1.51% of the malicious papers in CiteSeerX out the total world's malicious papers, whereas China, that has 20% of the worlds' population is the origin of only 0.55% of the malicious papers in CiteSeerX out the total world's malicious papers. Note that we did not find any malicious paper crawled from Russia or Korea which are the origin of many malicious Android applications found in applications' markets [31].

Country	Percentage
Israel	1.51%
Japan	1.23%
India	0.69%
Republic of Korea	0.69%
China	0.55%
Singapore	0.55%
Taiwan	0.27%
Hong Kong	0.14%
Iran	0.14%
Pakistan	0.14%

Table 3. Breakdown distribution of the malicious scholarly documents according to universities in countries within Asia from which they were crawled by CiteseerX scholarly digital libraries out of the total world's malicious papers.

5.2. Downloaded Malicious Papers

We now present the impact and power attack of malicious papers published in a scholarly digital library. Using CiteeSeerX's database and its Website historic log files, we extracted and aggregated the information regarding the download data of the malicious papers we found. We faced a big-data scale problem due to the enormous amount of data we needed to extract and process, and therefore, we extracted the downloading information for only the top 31 malicious papers identified by a larger number of antivirus engines out of the total 723 that were found. We also focused on download statistics for the five preceding years and therefore we can provide conclusions regarding updated download trends. In addition, we also used GNU Parallel[29] to boost the speed and reduce the very long running time. These data comprised 5197 successful downloads of malicious papers (during 2009-2014) that resulted from only 31 malicious papers crawled by CiteeSeerX's, meaning that scholarly digital libraries have an average 'damage coefficient' of 167 in the last 5 years. The average number of different countries that downloaded malicious papers was 16 over most of the continents (apart from Antarctica), which constitutes a very wide coverage of the worlds' research population within universities and other institutions. Table 4 presents information regarding the top 20 most downloaded malicious papers during the last 5 years. The most downloaded malicious paper is on the topic of Computer Forensics and apparently was a malicious version of a very popular paper; it was downloaded 2213 times in 108 different countries on all continents (apart from Antarctica). The popular topics among malicious papers were related to computers, such as cyber security and computer sciences.

Paper's Topic	Origin Country	Total Number of Downloads	Number of Countries
Computer Forensics	USA	2213	108
Network Security	France	860	77
Computer Hardware	Germany	633	59
Computer Networks	Germany	480	52
Social Networks	USA	235	51
Learning	Germany	123	20
Mathematics	Germany	90	12
Computer Science	USA	79	11
Software Engineering	BVI	77	4
Mathematics	USA	57	7
Computer Science	USA	57	4
Sociology	Brazil	46	10
Economics	USA	42	4
Computer Science	China	35	7
Computer Science	France	23	7
Computer Science	Holland	20	3
Astronomy	USA	20	3
Medical	USA	17	5
Physics	USA	13	4
Meteorology	Canada	13	3

Table 4. Top 20 most downloaded malicious scholarly documents during the last 5 years, their origin country, and the number of countries in which they were downloaded.

[29] http://www.gnu.org/software/parallel/

Figure 5 presents the distribution of the malicious scholarly documents according to the geographical location from which they were downloaded from CiteseerX scholarly digital libraries. More than 41% of the malicious papers in CiteSeerX were downloaded from IP's belonging to USA, whereas about 28% were downloaded from IP's belonging to Asia.

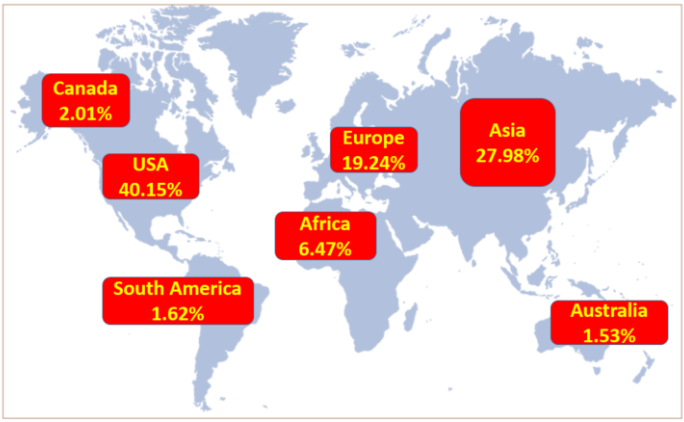

Figure 5. Distribution of the malicious scholarly documents according to the geographical location from which they were downloaded from CiteseerX scholarly digital library.

Figure 4 shows that the USA was the origin of more than 55% of the malicious papers in CiteseerX, while Table 5 shows that the USA was also the most popular destination, where more than 40% of the malicious papers were downloaded, followed by India (9.52%), China (5.04%), and the UK (3.77%). As can be seen, using a scholarly digital library as a platform, an attacker can easily distribute a worldwide attack through a malicious scholarly document.

Country	Percent of Downloads
United States	40.15%
India	9.52%
China	5.04%
United Kingdom	3.77%
Germany	2.87%
Philippines	2.77%
Spain	2.16%
Canada	2.01%
Iran	1.67%
Malaysia	1.23%
Italy	1.21%
France	1.15%
Australia	1.14%
Egypt	1.10%
Ethiopia	0.98%
Russia	0.89%
Korea	0.87%

Table 5. Top countries downloaded most of the malicious scholarly documents from CiteseerX during the last five years.

6. Directions for Security Enhancements

Several steps can be taken to mitigate and improve the detection of malicious PDF files within scholarly digital libraries. We will elaborate on several, beginning with the simplest and easiest to apply.

6.1. Compatibility Check of PDF Files

We found that many of the malicious files are not compatible with the PDF file format specifications according to the Adobe PDF Reference[30] and cannot in fact be opened by the PDF reader and viewed by the user. In cases involving malicious PDF files, the malicious operations will be executed anyway. We suggest using these observations to flag files that can initially be blocked from publication by the digital libraries. In order to empirically support our claim, we collected and created a dataset of malicious and benign PDF files. We acquired a total of 50,908 PDF files, including 45,763 malicious and 5,145 benign files, from 4 sources, as presented in Table 6 below. Note that this collection is not related to the CiteseerX collection that we analyzed on the previous sections. The benign files were reported to be virus-free by Kaspersky antivirus software. The malicious PDF files contain several types of malware families, such as viruses, Trojans, and backdoors. We also included obfuscated PDF files.

The analysis of our large dataset of 50,908 files by the parser (PdfFileAnalyzer[31]) shows that most of the malicious files (96.5%) are not compatible with the PDF file format specifications according to the Adobe PDF Reference. When the user tries to open an incompatible file (malicious or benign), the PDF reader is not able to open it and provides an error message. If it is a malicious PDF file, the malicious operation is executed; if it is a benign file, nothing occurs. However, in both cases the file remains unopened and cannot be viewed by an innocent user. Thus, it is clear that there is no reason to deliver an incompatible file to the user, and this observation should be taken into account in academic digital libraries, which can easily identify such files and mark them as suspicious, or even block them from being published before they are ever opened by an innocent user.

The incompatibility observed was located at the end of the file between the "startxref" and "%%EOF" lines. This line should contain a number serving as a reference (offset) to where the last cross reference table section is located in the file. In cases of incompatibility, the number that appears is incorrect. Table 6 includes the number of compatible files (bracketed) in each of our collected datasets. Note that while incompatible benign files were not present in our dataset, this does not necessarily mean that there were no such files. It might, however, suggest the very low probability of incompatibility among benign files and it provides support of our observation mentioned above.

[30] http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf
[31] http://www.codeproject.com/Articles/450254/PDF-File-Analyzer-With-Csharp-Parsing-Classes-Vers

Dataset Source	Year	Malicious Files	Benign Files
VirusTotal Repository	2012-2014	17,596 (1,017)	-
Srndic and Laskov [4]	2012	27,757 (437)	-
Contagio Project	2010	410 (175)	-
Internet and BGU Univ. (random selection)	2013-2014	0	5,145
Total		**45,763 (1,629)**	**5,145**

Table 6. Our collected dataset categorized as malicious, benign and the rate of incompatibles PDF files among the categories.

6.2. Update Check of Re-uploaded PDF Files

One of the vulnerabilities we found in academic digital libraries (particularly Google Scholar) relies on the fact that once a new paper is initially uploaded and indexed, it is then assumed to be scanned to verify that it is virus-free. However, in cases in which the clean paper file behind the indexed link was later replaced by a malicious version, the file was not rescanned and is now the paper's version available as a malicious file through these libraries. We suggest applying a simple check of the hash function behind each indexed file after it is first uploaded. The original hash function is compared to a daily hash function of each indexed file; thus a mismatch between the daily hash of the file and the original version acquired on the initial upload serves as an indication that the file should be further scanned to verify that it is virus free.

6.3. New PDF Malware Backward Check

While the vulnerabilities of new PDF files are identified from time to time by virus experts, the duration of the discovery period might be quite long. The new vulnerability is meanwhile being used and distributed in additional PDF files. Considering a 0-day malware contains such new vulnerabilities, it will probably evade the widely used antivirus tools. Therefore, as new vulnerabilities are discovered and antivirus tools are updated accordingly, we suggest a periodic re-check to provide a comprehensive review of a process that could easily be automated for all the files in the scholarly digital library.

6.4. Machine Learning Algorithms for Unknown Malware Detection

To date, antivirus packages are not sufficiently effective at intercepting malicious PDF files, even in the case of highly prominent PDF threats (Tzermias et al. [13]). On the other hand, according to studies such as [13], [4], [5], [14], [15], [16], [17], [18], [19], [20], machine learning (ML) methods can effectively distinguish between malicious and benign PDF files.

In this section, we explain comprehensively and in depth which and how existing high performance detection methods based on machine learning should be applied by scholarly digital libraries. These solutions should be applied offline after a new scholarly PDF document is found and before it is published and indexed. We propose using a machine learning-based detection model that includes a hybrid detection approach that conducts both static and dynamic analysis, as suggested in [13], [3], [15], and [19]. Thus, the chance of an attack evading the detection mechanism is significantly reduced,

because most attacks can be determined by dynamic analysis. Still, several techniques may evade detection, including those that perform the malicious actions of the PDF file only when specific conditions are met (e.g., time, date, IP, and specific user intervention). In these instances, dynamic analysis will be ineffective since it will not encounter and detect the malicious behavior through its analysis. Static analysis scrutinizes the file's genes, content, and structure, which are usually constant; consequently, static analysis will not be affected by these techniques and will therefore be more effective than dynamic analysis. Because of these advantages of static analysis, we suggest an initial static analysis stage for unknown PDF files. A file that is not identified as malicious after this initial stage and is also not detected by antivirus tools would merit further dynamic analysis.

For the static analysis phase, the key to precise and sensitive detection is preliminary knowledge of the primary attack and evasion techniques that could be used by a PDF file, as described in Section 2-b. The first mission is therefore to find and extract the indicators that assist and support the determination of these attacks. A prerequisite for a comprehensive analysis of a given PDF file is the development of a sophisticated and robust parser that is able to extract all the relevant information from the analyzed file (including corrupted PDF files, embedded EXE, PDF, and SWF files).

According to Vatamanu et al.'s study [14] in which the largest PDF file corpus was used, about 93% of the one million malicious PDF files (out of a corpus of 2.2 million) contained JavaScript, whereas only 5% of the benign PDF files contained JavaScript code. Therefore, as a mitigation strategy for malicious JavaScript code, all the JavaScript code should be extracted using a robust parser (including an unrelated object of JavaScript code as presented in [3]). The JavaScript code should be analyzed using two different direct representations that provide high TPR, the lexical analysis of JavaScript code [5], and tokenization of the embedded Java Script [14]. Direct representation means analyzing the code itself, while indirect representation means analyzing meta-features related to the entire content of the file. The JavaScript will also be dynamically analyzed during the dynamic analysis phase. We also suggest conducting an indirect static analysis, which analyzes the general descriptive content in the PDF file rather than directly analyzing the JavaScript code. This can be achieved by an approach that utilizes the meta-features of the content and structure of the PDF file, such as structural paths [4], summarized meta-features [16], and frequency of keywords [17], which also provided satisfactory results. The advantage of using meta-features such as structural paths [4] is that they are not affected by code obfuscation. It was shown to be a very effective method to discriminate malicious PDFs from benign PDFs, even in malicious files created two months after the classification model was created.

As a solution to embedded malicious files (reverse mimicry attacks [3]), the parser should also indicate whenever this scenario (a file embedded inside the PDF) exists in the suspicious PDF file. Generally speaking, there are few benign reasons to embed a file inside a PDF file. In addition, the parser should recursively extract every embedded file inside the PDF and analyze it using the static analysis methods suggested above. One of the reverse mimicry attacks [3] that embeds malicious EXE files in the PDF and auto-executes it when the PDF file is opened is based on a well-known legitimate feature that has been blocked in Adobe Reader X (version 10). Many organizations, however, do not update their installed software, and thus, are exposed to EXE running (such as in Adobe Reader MS Office). Regardless, when another feature or vulnerability has been found that allows the operation of running EXE files embedded in PDF files, it can be detected

with a variety of advanced techniques aimed at the detection of malicious executables using static and dynamic analysis.

All the extracted features mentioned in this section can be leveraged by an ensemble of classifiers such that each classifier will be induced from different sets of features. Menahem et al. [12] showed that applying an ensemble of classifiers using different features can significantly improve detection capabilities.

The attacks that were presented by Hamon et al. [8] dynamically load malicious code from a remote source as well as URI resolving (executing external malicious file). These attacks usually rely on clicking on a link; however, it is possible to open the link when the PDF file is opened, and therefore the PDF file becomes the link. Consequently, as indicated by Hamon et al. [8], the /OpenAction command is considered dangerous and can be detected by simple static analysis. Restricted use of this command will help prevent this kind of attack.

In the dynamic analysis phase, it is more effective to rely on hooking the Adobe Reader or using hardware virtualization to execute the JavaScript code embedded in the PDF file rather than to run it in an emulator, as presented in [19] and [20]. The malicious JavaScript code inside the PDF, however, can recognize that it is being executed in an emulated environment, and therefore, it might refrain from performing its malicious behavior. This will, however, probably provide a solution for malicious obfuscated JavaScript code that was not detected by the static analysis.

As far as we could identify, no product or academic solution actually analyzes the URLs inside the links in a PDF file. A link, having been clicked, can refer the user to a malicious Website that, when loaded, initiates an attack on the user's computer. An attacker can place a malicious link inside a benign file and persuade the user to click it. Dynamic analysis methods will not be able to detect this kind of attack, since user intervention is needed to click on the link. However, static analysis methods can easily extract and analyze the links that may be malicious. Thus, we recommend the addition to the detection model of a module that checks the links inside the PDF file for maliciousness, as this module can integrate many of the academic solutions designed for analyzing links (URLs) or Websites for maliciousness [21][25-30].

Full dynamic analysis of PDF files is a costly approach. For instance, Checkpoint Threat Emulation32 and SourceFire FireAMP[33] execute the entire PDF file in an isolated environment (sandbox) and examine the effect of the behavior and actions on the system during runtime. Nevertheless, this detection approach provides a comprehensive indication of the file's purposes and is robust against many evasion techniques, such as code obfuscation and URI resolving. Therefore, we suggest the integration of a full dynamic analysis module that might detect malicious behavior or determine the intention of PDF files in cases where the static or dynamic analyses (based on analysis of specific components of the PDF file) are unable to provide the comprehensive inspection provided by full dynamic analysis.

We also suggest running each suspicious PDF file through several versions of Adobe Reader (or any PDF reader) in order to compare its behavior. Some malicious PDF files will behave differently depending on the version of Adobe Reader used, because vulnerabilities are treated differently from one version to another. The differing behavior might provide an indication of a file's maliciousness.

[32] https://threatemulation.checkpoint.com/teb/

[33] http://www.sourcefire.com/security-technologies/advanced-malware-protection/fireamp

Moreover, one should remember that many organizations currently rely on outdated versions of PDF readers due to financial constraints and lack of proper administrative controls. The fact that many organizations do not update their installed software (including their PDF readers) exposes their computers and users to many known exploitations and bugs associated with PDF readers, such as the JBIG2Decode algorithm and util.printf Java function, as was discussed by Stevens [24]. New readers take these exploits into account; however, the exploits and bugs remain relevant in older versions of software. As a rule of thumb, we therefore recommend that organizations strive to equip themselves with the latest version of PDF readers as a standard security policy.

We also suggest applying an active learning framework for enhancing the detection of malicious PDF files that was recently presented by Nissim et al. [2], [36], which addresses an important issue that none of the above mentioned papers considered: the detection model's lack of updatability. It is not adequate to construct and calibrate a preliminary detection model based on sophisticated feature extraction techniques, but rather the model should be constantly updated in light of the daily creation of new malicious PDF files. While machine learning has been successfully used to induce malicious PDF detection models, all methods utilizing this approach focus on passive learning. Alternatively, we suggest focusing on active learning [23] and the use of the active learning methods that have been specially designed to enrich the detection model with new malicious PDF files over the course of several days, thus ensuring that the detection model is up to date. This notion was successfully used to enhance the detection of variety of malware including executable malwares in the Microsoft Windows OS [22], malicious documents of MS office word [34], Android malware [35], and is expected to enhance the detection of malicious PDF files as well.

7. Discussion and Conclusion

This study revealed the phenomenon of the contamination of scholarly digital libraries with malicious PDF documents and showed how these libraries can be easily used for launching and distributing targeted cyber-attacks aimed at a specific group of researchers, universities, institutions, and countries. As far as we know, there are no reliable reports of the accurate percentage of malicious PDF files on the Web, and therefore, we cannot determine whether scholarly digital libraries are more or less contaminated than the Web itself. In addition, as we found these malicious documents on CiteSeerX, we will have to remove these papers from it and also update other scholarly digital libraries regarding these malicious papers in order to prevent the attacks they are carrying from being further distributed. This process of removal should be done through cooperation with the authors of these papers, as the authors might discover the existence of resident malware in their computers that caused the infection of their paper, in the case that their paper was not contaminated intentionally.

In this study, we evaluated more than two million scholarly papers in the CiteSeerX library and found it to be contaminated with a surprisingly large number (0.3%-2%) of malicious PDF files belonging to a variety of different virus families, 72% of which exploit vulnerabilities in PDF readers. These malicious documents were uploaded from 46 different countries covering most of the continents. The USA's universities were found to be the origin of more than 55% of the malicious papers in CiteSeerX. The USA also downloaded more than 41% of these malicious scholarly papers during the last five years. On average, a malicious paper was downloaded 167 times in 5 years by researchers from many different countries worldwide. The most popular malicious scholarly document is a malicious version of a famous paper in the computer forensics domain,

crawled from the USA, and downloaded 2213 times in 108 different countries. Therefore, as we indicated, several vulnerabilities exist in scholarly digital libraries, and an attacker needs only to have a malicious version of a popular paper on an attractive topic (e.g., cyber-security) in a scholarly digital library to utilize the high damage coefficient we found and thus cover most of countries in the world. We also suggested several solutions for mitigating such attacks, including simple deterministic solutions and also advanced machine learning-based frameworks that should both be integrated in scholarly digital libraries.

In future work, we suggest that the other digital libraries for which we presented vulnerabilities be scanned further, and also that additional scholarly digital libraries be investigated for vulnerabilities, such as MAS, Web of Science, and Pub-Med. We also suggest investigating the rate of contamination of digital libraries within the Darknet, such as Libgen, Sci-hub, and Booksc, which we presented as well.

References

[1] Nir Nissim, Aviad Cohen, Chanan Glezer, Yuval Elovici, Detection of malicious PDF files and directions for enhancements: A state-of-the art survey, Computers & Security, Volume 48, February 2015, Pages 246-266, ISSN 0167-4048, http://dx.doi.org/10.1016/j.cose.2014.10.014.

[2] Nir Nissim, Aviad Cohen, Robert Moskovitch, Oren Barad, Mattan Edry, Assaf Shabatai and Yuval Elovici, " ALPD: Active Learning framework for Enhancing the Detection of Malicious PDF Files aimed at Organizations.", JISIC (2014)

[3] D. Maiorca, I. Corona and G. Giacinto. Looking at the bag is not enough to find the bomb: An evasion of structural methods for malicious PDF files detection. Presented at Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security. 2013, .

[4] N. Šrndic and P. Laskov. Detection of malicious pdf files based on hierarchical document structure. Presented at Proceedings of the 20th Annual Network & Distributed System Security Symposium. 2013.

[5] P. Laskov and N. Šrndić. Static detection of malicious JavaScript-bearing PDF documents. Presented at Proceedings of the 27th Annual Computer Security Applications Conference. 2011, .

[6] Baccas, Paul. "Finding rules for heuristic detection of malicious pdfs: With analysis of embedded exploit code." Virus Bulletin Conference. 2010.

[7] Kittilsen, Jarle. "Detecting malicious PDF documents." (2011).

[8] Hamon, Valentin. "Malicious URI resolving in PDF documents." Journal of Computer Virology and Hacking Techniques 9.2 (2013): 65-76.

[9] Khabsa, Madian, and C. Lee Giles. "The number of scholarly documents on the public web." PLOS one 9.5 (2014): e93949.

[10] Gargouri, Y., et al., Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research, PLOS ONE, 5(10): e13636, October 18, 2010 GS Biblio

[11] Lawrence, S., Free online availability substantially increases a paper's impact, Nature, 31 May 2001 GS Biblio

[12] E. Menahem, A. Shabtai, L. Rokach and Y. Elovici. Improving malware detection by applying multi-inducer ensemble. Comput. Stat. Data Anal. 53(4), pp. 1483-1494. 2009.

[13] Z. Tzermias, G. Sykiotakis, M. Polychronakis and E. P. Markatos. Combining static and dynamic analysis for the detection of malicious documents. Presented at Proceedings of the Fourth European Workshop on System Security. 2011.

[14] C. Vatamanu, D. Gavriluţ and R. Benchea. A practical approach on clustering malicious PDF documents. Journal in Computer Virology 8(4), pp. 151-163. 2012.

[15] F. Schmitt, J. Gassen and E. Gerhards-Padilla. PDF scrutinizer: Detecting JavaScript-based attacks in PDF documents. Presented at Privacy, Security and Trust (PST), 2012 Tenth Annual International Conference On. 2012.

[16] C. Smutz and A. Stavrou. Malicious PDF detection using metadata and structural features. Presented at Proceedings of the 28th Annual Computer Security Applications Conference. 2012.

[17] D. Maiorca, G. Giacinto and I. Corona. "A pattern recognition system for malicious pdf files detection," in Machine Learning and Data Mining in Pattern RecognitionAnonymous 2012.

[18] H. Pareek, P. Eswari, N. S. C. Babu and C. Bangalore. Entropy and n-gram analysis of malicious PDF documents. Int. J. Eng. 2(2), 2013.

[19] X. Lu, J. Zhuge, R. Wang, Y. Cao and Y. Chen. De-obfuscation and detection of malicious PDF files with high accuracy. Presented at System Sciences (HICSS), 2013 46th Hawaii International Conference On. 2013.

[20] K. Z. Snow, S. Krishnan, F. Monrose and N. Provos. SHELLOS: Enabling fast detection and forensic analysis of code injection attacks. Presented at USENIX Security Symposium. 2011.

[21] J. Ma, L. K. Saul, S. Savage and G. M. Voelker. Learning to detect malicious URLs. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), pp. 30. 2011.

[22] Nir Nissim, Robert Moskovitch, Lior Rokach, Yuval Elovici, Novel Active Learning Methods for Enhanced PC Malware Detection in Windows OS, Expert Systems with Applications, Available online 19 March 2014, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2014.02.053.

[23] Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison 52 (2010): 55-66.

[24] Stevens, D., "Malicious PDF Documents Explained," Security & Privacy, IEEE, vol.9, no.1, pp.80,82, Jan.-Feb. 2011. doi: 10.1109/MSP.2011.14

[25] Xuewen, Zhu, Wan Xinochuan, and Ye Hua. "Detection of malicious URLs in a web page." U.S. Patent No. 8,505,094. 6 Aug. 2013.

[26] B. Eshete. Effective analysis, characterization, and detection of malicious web pages. Presented at Proceedings of the 22nd International Conference on World Wide Web Companion. 2013.

[27] H. Zhou, J. Sun and H. Chen. Malicious websites detection and search engine protection. Journal of Advances in Computer Network 1(3), 2013.

[28] K. Su, K. Wu, H. Lee and T. Wei. Suspicious URL filtering based on logistic regression with multi-view analysis. Presented at Information Security (Asia JCIS), 2013 Eighth Asia Joint Conference On. 2013.

[29] S. Chitra, K. Jayanthan, S. Preetha and R. U. Shankar. Predicate based algorithm for malicious web page detection using genetic fuzzy systems and support vector machine. International Journal of Computer Applications 40(10), pp. 13-19. 2012.

[30] D. Ranganayakulu and C. Chellappan. Detecting malicious URLs in E-mail–An implementation. AASRI Procedia 4pp. 125-131. 2013.

[31] Axelle Apvrille, Tim Strazzere: Reducing the window of opportunity for Android malware Gotta catch 'em all. Journal in Computer Virology 8(1-2): 61-71 (2012).

[32] Nissim, N., Boland, M. R., Moskovitch, R., Tatonetti, N. P., Elovici, Y., Shahar, Y., & Hripcsak, G. (2015). An Active Learning Framework for Efficient Condition Severity Classification. In Artificial Intelligence in Medicine (pp. 13-24). Springer International Publishing. (AIME-15).

[33] Nir Nissim, Mary Regina Boland, Nicholas P. Tatonetti, Yuval Elovici, George Hripcsak, Yuval Shahar, Robert Moskovitch, Improving condition severity classification with an efficient active learning based framework, Journal of Biomedical Informatics, Volume 61, June 2016, Pages 44-54, ISSN 1532-0464.

[34] Nir Nissim, Aviad Cohen, Yuval Elovici. ALDOCX: Detection of Unknown Malicious Microsoft Office Documents using Designated Active Learning Methods Based on New Structural Feature Extraction Methodology. IEEE Transactions on Information Forensics and Security, 15.1 (2017): 40-55.

[35] Nissim, N., Moskovitch, R., BarAd, O., Rokach, L., & Elovici, Y. (2016). ALDROID: efficient update of Android anti-virus software using designated active learning methods. Knowledge and Information Systems (2016), 1-39.

[36] Nissim, N., Cohen, A., Moskovitch, R., Shabtai, A., Edri, M., BarAd, O., & Elovici, Y. (2016). Keeping pace with the creation of new malicious PDF files using an active-learning based detection framework. *Security Informatics*,*5*(1), 1-20.