A Structural and Content-based Approach for a Precise and Robust Detection of Malicious PDF Files

作者：

Davide Maiorca, Davide Ariu, Igino Corona and Giorgio Giacinto

会议：

in Proceedings of the 1st International Conference on Information Systems Security and Privacy. ScitePress Digital Library,2

## 中文标题：

一种基于结构和内容的方法，用于精确和可靠地检测恶意PDF文件

## Questions:

### What is the problem?

提出一种机器学习的系统，可以自动的检测恶意恶意PDF文件，使用PeePDF和Origami 两个解析工具，根据文件结构和文件内容的分析量化，然后选择合适算法（decision treein）来有效的对付一些逃逸攻击

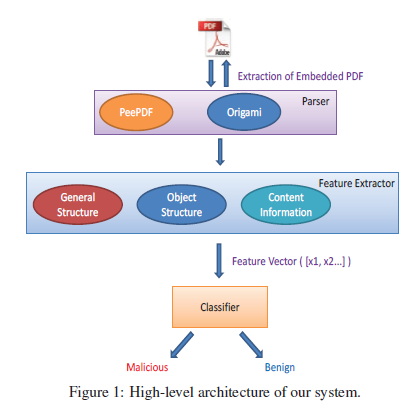
### Why is the problem important?

在过去的几年中，恶意PDF文件已经成为现代计算机系统安全性的严重威胁。

### What is the old technique?

基于正则匹配的检测不能很好的对付那些新奇的攻击方法

### What is the new technique?



### Why is the problem in the futuer?

## Abstract：

` During the past years, malicious PDF files have become a serious threat for the security of modern computer systems. They are characterized by a complex structure and their variety is considerably high. Several solutions have been academically developed to mitigate such attacks. However, they leveraged on information that were extracted from either only the structure or the content of the PDF file. This creates problems when trying to detect non-Javascript or targeted attacks. In this paper, we present a novel machine learning system for the automatic detection of malicious PDF documents. It extracts information from both the structure and the content of the PDF file, and it features an advanced parsing mechanism. In this way, it is possible to detect a wide variety of attacks, including non-Javascript and parsing-based ones. Moreover, with a careful choice of the learning algorithm, our approach provides a significantly higher accuracy compared to other static analysis techniques, especially in the presence of adversarial malware manipulation.

摘要 –

在过去的几年中，恶意PDF文件已经成为现代计算机系统安全性的严重威胁。它们的特点是结构复杂，种类繁多。学术上已经开发了几种解决方案来缓解这种攻击。但是，他们利用从PDF文件的结构或内容中提取的信息。当检测non-Javascript或有针对性的攻击时会产生问题。在本文中，我们提出了一种用于自动检测恶意PDF文档的新型机器学习系统。它从PDF文件的结构和内容中提取信息，并且这些特征是用先进的解析器来完成的。通过这种方式，可以检测各种各样的攻击，包括non-Javascript和基于分析的攻击。此外，通过仔细选择学习算法，与其他静态分析技术相比，我们的方法提供了明显更高的准确性，特别是在存在敌对恶意软件操作的情况下。

## Contibutions:

*•* We develop a novel, machine learning based approach to the detection of malicious PDF files that leverages on information extracted from both the *structure* and the *content* of a PDF file;

*•* We experimentally evaluate the performances of our system on a dataset that contains a wide number of PDF-related vulnerabilities. We compare such performances to the ones of the most important publicly available tools;

*•* We evaluate the robustness of our system against automatic attacks and evasion attempts that have proved to be extremely effective against public available tools;

*•* We discuss the limits of our system and the role of the *learning algorithm* in assessing its robustness. In relation to that, we provide research guidelines for future work.

•我们开发了一种新颖的基于机器学习的方法来检测恶意PDF文件,利用从PDF文件的结构和内容中提取的信息

•我们评估系统的性能是建立在有价值的pdf相关漏洞的数据集上，我们与其中一些比较重要公开可用的工具进行性能对比

•我们评估了系统的鲁棒性，我们的系统在自动攻击和逃逸的一些实验中进行抵制，这些实验已证明对公共工具非常有效;

•我们讨论系统的局限性以及学习算法在评估其鲁棒性方面的作用。 就此而言，我们为未来的工作提供研究指导。

## Conclution

Malicious PDF files have become a well-known threat in the past years. PDF documents constitute a very effective attack vector for cyber-criminals. In spite of the efforts of software vendors such as Adobe, PDF software is often vulnerable to zero-day attacks. In this work, we presented a new approach that leveraged on both structural and content-based information to provide a very accurate detection of PDF malware. We also showed that our approach, with a careful choice of the learning algorithm, is also able to cope with evasion attacks. Finally, our work also clearly pointed out the need of secure learning techniques for malware detection , in order to cope with deliberate adversarial attacks.

恶意PDF文件在过去几年已成为众所周知的威胁。 PDF文件构成网络犯罪非常有效的攻击媒介。 尽管Adobe等软件供应商付出了努力，但PDF软件经常容易遭受零日攻击。 在这项工作中，我们提出了一种利用结构和基于内容的信息来提供更加准确的检测PDF恶意软件的新方法。 我们还表明，我们的方法通过仔细选择学习算法，也能够应对逃逸攻击。 最后，我们的工作也明确指出了需要安全的学习技术来检测恶意软件，以应对故意的对抗性攻击。