上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

PHP-WEB漏洞挖掘
需求分析

小组成员：金子本、王新忠、洪逸杰、张泽挥、杨景犀、钟一鸣

项目名称：PHP-WEB漏洞挖掘

指导老师：黄征

日期：2022年5月27日

目录

第一章 项目背景和意义

随着互联网的持续发展与广泛应用，网络已经成为人们生产生活中密不可分的一部分，但随着互联网行业竞争愈发激烈，软件的迭代也变得越来越快，但这同样也带来了日益突出的安全问题。

根据奇安信等机构发布的《2021中国网站安全报告》，2021年全年，奇安信“补天漏洞响应平台”共收录全国各类网站安全漏洞146293个，共涉及网站115243个。从漏洞的危险程度来看，其中高危漏洞占比21.9%，中危漏洞占比72.1%，低危漏洞占比6.0%；从漏洞的技术类型来看，如图1所示，2021年全年，在补天平台收录的网站安全漏洞中，信息泄露漏洞占比最高，达36.0%，其次是SQL注入漏洞，占比18.4%，弱口令占比12.9%。

如图2所示，在OWASP 2021年 TOP10报告中，注入漏洞（包括SQL注入、跨站脚本等）依旧高居第三位。

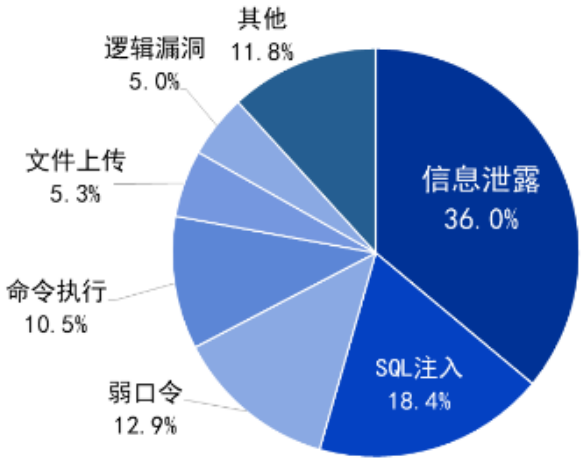

图1.2021年奇安信补天平台收录网站安全漏洞技术类型分布

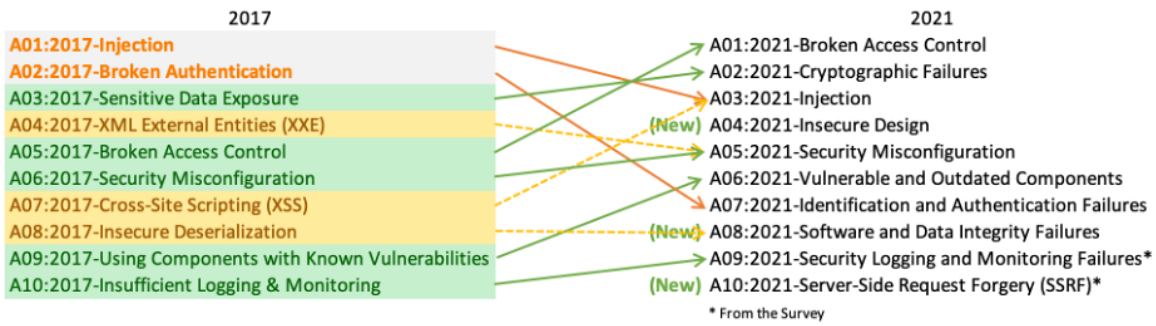

图2. OWASP TOP 10

PHP目前依旧是世界上市场占有率最大的Web程序语言，如图3所示，截止2022年5月，服务端的编程语言中，PHP以77.5的占有率高居第一，许多高流量的Web应用程序及Web开发框架都是由PHP编写的，包括WordPress、Drupal、Laravel等。如此高的市场比重，意味着PHP安全面临着严峻的考验，因此对源代码进行快速准确的漏洞挖掘，从根源上消除安全漏洞，存在着重要意义。

Server-side Programming Languages

Most popular server-side programming languages

© W3Techs.com	usage	change since 1 May 2022
1. PHP	77.5%	
2. ASP.NET	7.8%	+0.1%
3. Ruby	5.8%	-0.1%
4. Java	4.1%	+0.1%
5. Scala	2.7%	-0.1%

percentages of sites

第二章 相关工作

目前国内外针对PHP的漏洞检测提出了很多的技术，本节将对现今主要的检测技术做一个简单的阐述。

2.1静态分析

静态分析是对程序的源代码进行分析，它不需要实际执行应用程序，而是通过构建程序状态的模型，然后确定程序如何对此状态做出反应来进行操作。然后静态分析对应用程序源代码进行编译，将其转换为由控制数据和程序路径的控制流程图，通过数据流分析或控制流分析等方法对其进行分析，以此来判断程序中每个节点中变量可能的取值情况，来发现有可能存在的安全漏洞。

但这样的一个方法存在的一个主要问题就在于其准确率非常依赖于预定义的规则，就目前而言其准确率在各方法中并不算理想。

2.2 动态分析

动态分析主要是通过模拟用户来执行程序，从而来观察程序真实的执行操作，由于其不需要进行抽象，动态分析相比静态分析来说更加精确，更贴近程序实际运行时的行为。

但动态分析的方法存在的主要问题在于其时间复杂度过高，对一个大型项目想要获得准确的结果需要大量的计算。

2.3 污点分析

污点分析主要通过对应用程序输入接口的数据进行标记，并跟踪其在程序中的执行路径来确定漏洞是否存在。应用程序中与该数据通过算术或逻辑运算等方式建立关系的数据都将被判定是否被污染，然后通过一定的检测手段来对污染数据传播的路径进行分析，通过污染数据是否被消除来确定应用程序是否存在漏洞。

但目前而言，污点分析存在一个很严重的问题就是其误报率很高，很容易将良性的代码识别为恶意的，导致会浪费很多的人力在对漏洞复核上。

2.4 图神经网络

在上述几种方法中，程序代码常被编译为程序控制流图，代码属性图等，这些图中存储着大量原始代码的信息，并保存了原始代码的结构特征，相比于直接将原始代码提取向量作为神经网络的输入，将这些图作为图神经网络的输入，由于其更高的信息量，训练出来模型的预测精度也会得到明显的提高。

但基于图神经网络的办法有一个缺点在于其计算开销巨大，且定位粒度不够，现有的模型通常只能定位到文件级的漏洞，而不能精确定位到某一行。

第三章 项目目标

由于现在主流的Web漏洞检测技术存在着计算复杂度高、识别误报率高、检测粒度大等问题，我们小组就想要通过图神经网络与污点分析相结合的方法，在减少误报率的同时使检测粒度更低，实现一个高效自动化的PHP Web漏洞的挖掘系统，并将其部署至网页端供用户访问。