**王凤娇 研究工作进展与规划**

目录

[一、2018年Q2工作概述 1](#_Toc521574276)

[1.1 PDF恶意文件静态检测技术研究成果 1](#_Toc521574277)

[1.2 Q2季度KR与自我评定 2](#_Toc521574278)

[二、2018年Q3工作OKR 2](#_Toc521574279)

[2.1恶意文档检测引擎 2](#_Toc521574280)

[目标 O 2](#_Toc521574281)

[时间范围 2](#_Toc521574282)

[关键成果KR 2](#_Toc521574283)

# 一、2018年Q2工作概述

此研究项目的目标旨在为工业界（如蓝盾）和学术界提供一个基于AI的文档（PDF，word等）分类器，此分类器具有 准确度极高，误报率极低和抗逃逸能力强三大特点。通过对文档领域特征的挖掘，并结合数据科学中的复杂建模，模型成功被构建。此模型可广泛应用于蓝盾多条产品线上，如在终端安全产品上加入此模组，又或者在邮件服务器上加入此模组，均是非常有意义的应用场景，AI文档分类器不仅限于pdf，如可扩展到word，execl，ppt等有结构的文档。

## 1.1 PDF恶意文件静态检测技术研究成果

1）基于人工智能的恶意文档检测引擎，在**二十万测试数据**中，识别率高达**99.8%；**

2）构建鲁棒性好的AI分类模型，并在**4种逃逸的情景**中有抗逃逸的能力；

3）对于AI对抗性机器学习，提出了**5 种抗逃逸的方法；**

4）在“全国计算机安全学术交流会”中投稿**中文论文一篇**；

5）加入对抗性学习与特征有效性的研究，并投稿**英文论文一篇；**

6）收集与阅读国内外安全前沿技术论文 21 篇；

## 1.2 Q2季度KR与评定

1）（基础）**加固 分类器** 的方法（参考**Feature Squeezing**一文）---DONE（7分）

2）（基础）阅读并收集至少**5份顶会论文**，每周 **1-2 个读书笔记**分享 ---DONE（9分）

3）（中级）根据现有技术，对已经选取的**特征进行计算合并**，并且在新的模型中 优化现有特征 ---DONE（8分）

4）（中级）完成**中文论文**的修稿工作，以及**图表制作**，数据整理，模型训练等---DONE（9分）

5）（中级）研究现有攻击逃逸的技术，并针对现有的逃逸技术，训练**一个 具有抗逃逸** 的**分类器 一个**---DONE（9分）

6）（高级）完成**一篇**论文并**提交 到合适的会议** 中---DONE（9分）

# 二、2018年Q3工作OKR

## 2.1恶意文档检测引擎

### 目标 O

* 在PDF文件检测的AI模型中添加对抗性学习，与动态分析

### 时间范围

* 2018 Q3-Q4

### 关键成果KR

1. （中级）完成 **1篇** 恶意文档检测以及对抗性学习的英文论文，并**投稿**到对应的**安全会议**中；
2. （中级）在对抗性学习中，添加**抗逃逸的方法 2 个**；
3. （初级）研读最新顶级会议中的研究方法并做阅读笔记 **至少5 篇**；
4. （高级）搭建恶意样本逃逸的**沙箱环境1 个**，添加**动态检测实验1 个**；
5. （高级）研究并实现NDSS会议中Weilin Xu自动逃逸分类器的相关算法及过程；
6. （中级）输出模型的**训练与预测API**，构建**在线恶意文件检测系统**；
7. （高级）结合产品部署于公司防火墙，邮件服务器等；