恶意程序分类识别

1 恶意程序类别

恶意程序是指在未经授权的情况下，在信息系统中安装、执行以达到不正当目的的程序。恶意程序分类说明如下。

1.1 特洛伊木马

特洛伊木马（简称木马）是以**盗取用户个人信息、远程控制用户计算机**为主要目的的恶意程序，通常由控制端和被控端组成。由于它像间谍一样潜入用户的计算机，与战争中的“木马”战术十分相似，因而得名木马。按照功能，木马程序可进一步分为7类。

- 盗号木马是用于窃取用户电子邮箱、网络游戏等账号的木马。
- 网银木马是用于窃取用户网银、证券等账号的木马。
- 窃密木马是用于窃取用户主机中敏感文件或数据的木马。
- 远程控制木马是以不正当手段获得主机管理员权限，并能够通过网络操控用户主机的木马。
- 流量劫持木马是用于劫持用户网络浏览的流量到攻击者指定站点的木马。
- 下载者木马是用于下载更多恶意代码到用户主机并运行，以进一步操控用户主机的木马。

1.2 僵尸程序

僵尸程序是**用于构建大规模攻击平台**的恶意程序。按照使用的通信协议，僵尸程序可进一步分为IRC僵尸程序、HTTP僵尸程序、P2P僵尸程序和其他僵尸程序4类。

1.3 蠕虫

蠕虫是指能自我复制和广泛传播，以占用系统和网络资源为主要目的的恶意程序。按照传播途径，蠕虫可进一步分为邮件蠕虫、即时消息蠕虫、U盘蠕虫、漏洞利用蠕虫和其他蠕虫5类。

1.4 病毒

病毒的主要特征是感染正常文件。病毒是**通过感染计算机文件进行传播，以破坏或篡改用户数据，影响信息系统正常运行**为主要目的的恶意程序。编制者在计算机程序中插入的破坏计算机功能或者数据的代码，能影响计算机使用，能自我复制的一组计算机指令或者程序代码。其具有传播性、隐蔽性、感染性、潜伏性、可激发性、表现性或破坏性。

一般病毒的生命周期：开发期→传染期→潜伏期→发作期→发现期→消化期→消亡期。

1.5 勒索软件

勒索软件是**黑客用来劫持用户资产或资源并以此为条件向用户勒索钱财的一种恶意软件。**勒索软件通常会将用户数据或用户设备进行加密操作或更改配置，使之不可用，然后向用户发出勒索通知，要求用户支付费用以获得解密密码或者获得恢复系统正常运行的方法。

1.6 移动互联网恶意程序

移动互联网恶意程序是指在用户不知情或未授权的情况下，在移动终端系统中安装、运行以达到不正当的目的，或具有违反国家相关法律法规行为的可执行文件、程序模块或程序片段。按照行为属性分类，移动互联网恶意程序包括恶意扣费、信息窃取、远程控制、恶意传播、资费消耗、系统破坏、诱骗欺诈和流氓行为8种类型。

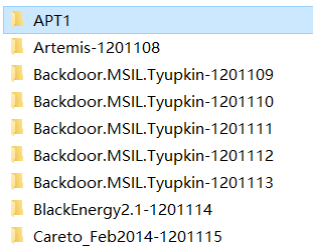
必须强调的是，现在很少有木马单打独斗，大部分都是“多毒种作战”。下载者木马体积小巧，不易被察觉，再由它下载其他木马到用户计算机。释放器木马负责安装复杂木马，一旦运行了释放器木马就很难手动清除。 还有的木马选择和蠕虫病毒搭伙，前阵子通过 NSA “永恒之蓝” 漏洞传播的“WannaMine”，就是配合蠕虫病毒来进行传播的挖矿木马。

2 恶意程序分类识别

随着反恶意代码技术的逐步发展，主动防御技术、云查杀技术已越来越多的被安全厂商使用，但恶意代码静态检测的方法仍是效率最高，被运用最广泛的恶意代码查杀技术。

2.1 数据集

MIST数据集，包含了多种恶意程序。代码中只用到其中四种恶意程序。



{"entityId": 652, "entityType": "content", "event": "malware", "eventTime": "2016-12-15T09:01:34.646+0000", '
d5d38a3 c2f5b9f8 afa0ff8b 16799c50 256261b4 34d1c350 78bace59 07cc694b 4b99ff73 1b0bdf03 eebc733
0774c03f 8fc9485d 330cfab4 8fc9485d 330cfab4 8fc9485d 330cfab4 8fc9485d 9327ee8d 8fc9485d 9327ee8
3 3e427332 2e5d8aa3 aebc6894 2e5d8aa3 aebc6894 2e5d8aa3 aebc6894 2e5d8aa3 8525de72 2e5d8aa3 867

2.2 特征提取

2.2.1 2-gram+TF-IDF

实现

```
vectorizer = CountVectorizer(
    decode_error='ignore',
    ngram_range=(2, 2),
    token_pattern=r'\b\w+\b',#按单词切分
    strip_accents='ascii',
    max_features=max_words,
    stop_words='english',
    max_df=1.0,#作为一个阈值，词是否当作关键词。表示词出现的次数与语料库文档数的百分比
    min_df=1)
print(vectorizer)
x = vectorizer.fit_transform(x)
transformer = TfidfTransformer(smooth_idf=False)
x = transformer.fit_transform(x)
```

2.3 模型训练与验证

SVM

```
s = SVC(C=1.0, kernel='linear', decision_function_shape='ovr')
```

XGoosting

```
model = XGBClassifier(n_estimate=100, max_depth=8, n_jobs=1)#n_estimate:决策树个数
```

MLP

```
model = MLPClassifier(hidden_layer_sizes=(10, 4), activation='relu',
                      solver='adam',shuffle=True, verbose=1)
```

https://www.kaggle.com/c/malware-classification/

https://github.com/daxiongshu/kaggle_Microsoft_Malware