(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113381996 B
(45)授权公告日 2023.04.28

(21)申请号 202110637965.1

(22)申请日 2021.06.08

(65)同一申请的已公布的文献号
申请公布号 CN 113381996 A

(43)申请公布日 2021.09.10

(73)专利权人 中电福富信息科技有限公司
地址 350000 福建省福州市鼓楼区五凤街道软件大道89号12号楼

(72)发明人 黄丽荣 陈耿生 蔡悦贞 戴宏鹏 黄嘉诚

(74)专利代理机构 福州君诚知识产权代理有限公司 35211
专利代理师 彭东

(51)Int.Cl.
H04L 9/40(2022.01)
G06N 20/00(2019.01)
G06F 21/55(2013.01)
G06F 18/214(2023.01)

(56)对比文件
US 2010138919 A1,2010.06.03
US 2013174256 A1,2013.07.04
CN 104683346 A,2015.06.03
牛伟纳;张小松;孙恩博;杨国武;赵凌园;.基于流相似性的两阶段P2P僵尸网络检测方法.电子科技大学学报.2017,(06),全文.
苏欣;张大方;罗章琪;曾彬;黎文伟;.基于Command and Control通信信道流量属性聚类的僵尸网络检测方法.电子与信息学报.2012,(08),全文.

审查员 张丽萍

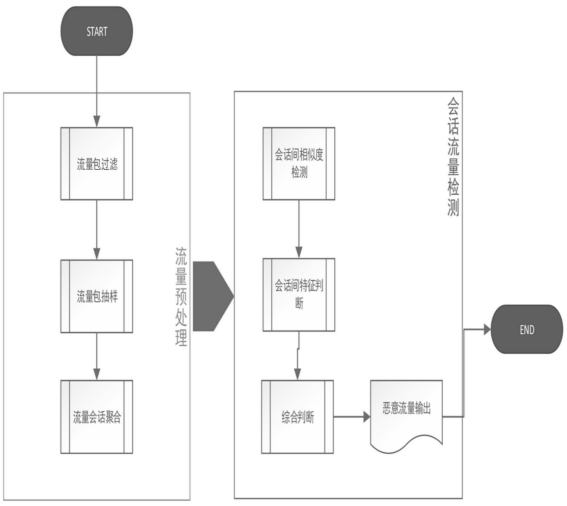
权利要求书1页 说明书3页 附图1页

(54)发明名称

基于机器学习的C&C通讯攻击检测方法

(57)摘要

本发明公开基于机器学习的C&C通讯攻击检测方法,其包括以下步骤:获取连续下行流量包并对流量包进行过滤,使得流量包长度的分布呈正态分布,根据指定条件对流量包进行会话聚合;利用随即簇抽样和Apriori算法提取会话流量特征;利用对序列相似度检测采用编辑距离与求最长公共子序列(LCS)相结合的方式,对聚合之后的流量上下文数据进行相似度计算。本发明不需要依赖特征库,可以检测未被发现的恶意软件通讯;对大量攻击流量样本进行检测时,检测时间复杂度较低,检测时间更短。



CN 113381996 B

1.基于机器学习的C&C通讯攻击检测方法,其特征在于:其包括以下步骤:

步骤1,流量包过滤:获取连续下行流量包并对流量包进行过滤,使得流量包长度的分布呈正态分布,

步骤2,流量会话聚合:根据指定条件对流量包进行会话聚合;

步骤3,利用随机簇抽样和Apriori算法提取会话流量特征;

步骤4,利用对序列相似度检测采用编辑距离与求最长公共子序列相结合的方式,对聚合之后的流量上下文数据进行相似度计算;

步骤5,根据会话的下行流量上下文相似度是否超过设定值来判断是否异常C&C通讯。

2.根据权利要求1所述的基于机器学习的C&C通讯攻击检测方法,其特征在于:步骤1根据流量包长度的正太分布设置过滤阈值,将部分无相关流量过滤;通过设定包过滤率计算得小流量包的包长度临界值,最终过滤包长度采用正态分布估计和设定阈值方式综合计算来确定。

3.根据权利要求1所述的基于机器学习的C&C通讯攻击检测方法,其特征在于:步骤2中根据源地址、源端口、目的地址或目的端口进行会话聚合。

4.根据权利要求1所述的基于机器学习的C&C通讯攻击检测方法,其特征在于:步骤3中处理数据量过大时,采用reservoir sampling算法进行概率抽样。

基于机器学习的C&C通讯攻击检测方法

技术领域

[0001] 本发明涉及通信安全技术领域,尤其涉及基于机器学习的C&C通讯攻击检测方法。

背景技术

[0002] 目前对C＆C通信检测有三个方面,分别是基于流量包的统计特征检测、基于流量payload中的特征码检测、基于现有恶意软件有监督机器学习方法检测。

[0003] 现有针对C＆C通信攻击检测一定的不足之处。首先,现有方法对未公布的或者未发现的恶意软件的检测有一定的缺陷。其次,现有方法的检测效果更多的依赖特征库是不是全面。最后,由于正常用户使用网络场景比较多样化,很容易造成正常用户流量属性特征与恶意流量属性特征相似的情况,如根据数据包的大小及到达时间隔来判断,现有的部分聊天软件的通信过程就有可能出现和恶意软件有类似的特征。所以现有方法对C＆C通信的检测精度以及检测效果有一定的局限性。对C＆C通信检测的方面一定的不足之处。基于流量包的统计特征检测,由于恶意软件本身的通信会随网络拥塞的变化而变化,且随着现正常网络应用场景越来越多,容易造成正常用户流量与恶意用户流量的统计特征相似,从而有较高的误报率。基于流量payload中的特征码检测,对现有已知的恶意软件有较高的检测效果,但是如果恶意软件发生变异导特征码发生变化则会造成检测失效。基于现有恶意软件有监督机器学习方法检测主要是基于现有恶意软件的流量特征进行有监督学习,其检测效果更多的依赖于机器学习的训练集的覆盖广度以及学习方法的科学性。

发明内容

[0004] 本发明的目的在于提供基于机器学习的C&C通讯攻击检测方法。

[0005] 本发明采用的技术方案是:

[0006] 基于机器学习的C&C通讯攻击检测方法,其包括以下步骤:

[0007] 步骤1,流量包过滤:获取连续下行流量包并对流量包进行过滤,使得流量包长度的分布呈正态分布,

[0008] 步骤2,流量会话聚合:根据指定条件对流量包进行会话聚合;

[0009] 步骤3,利用随即簇抽样和Apriori算法提取会话流量特征;

[0010] 步骤4,利用对序列相似度检测采用编辑距离与求最长公共子序列(LCS)相结合的方式,对聚合之后的流量上下文数据进行相似度计算。

[0011] 步骤5,根据会话的下行流量上下文相似度是否超过设定值来判断是否异常C&C通讯。

[0012] 进一步地,作为一种较优实施方式,步骤1根据流量包长度的正太分布设置过滤阈值,将部分无相关流量过滤,

[0013] 进一步地,作为一种较优实施方式,步骤1通过设定包过滤率计算得小流量包的包长度临界值,最终过滤包长度采用正态分布估计和设定阈值方式综合计算来确定。

[0014] 进一步地,作为一种较优实施方式,步骤2中根据源地址、源端口、目的地址或目的

端口进行会话聚合。

[0015] 进一步地，作为一种较优实施方式，步骤3中处理数据量过大时，采用reservoir sampling算法进行概率抽样。

[0016] 进一步地，作为一种较优实施方式，步骤4中先对序列对进行编辑距离计算，根据计算结果进行筛选去掉距离值较大的序列对，后对序列对进行ＬＣＳ计算。

[0017] 本发明采用以上技术方案，对于网络流量中根据流量过滤，抽样，聚合之后，对聚合之后的会话流量数据进行上下文相似度检测，进而检测是否存在恶意软件通讯。本发明具有如下优点：1、不需要依赖特征库，可以检测未被发现的恶意软件通讯。2、不同于现有恶意软件有监督机器学习方法检测主要是基于现有恶意软件的流量特征进行有监督学习，其检测效果更多的依赖于机器学习的训练集的覆盖广度以及学习方法的科学性。3、对Ｃ＆Ｃ通信检测上，基于下行payload相似度检测算法相对于流量包检测算法和payload特征码检测有较高的准确率和召回率，同时在检测时间也有一定的优势，尤其对大量攻击流量样本进行检测时，检测时间复杂度较低，检测时间更短。

附图说明

[0018] 以下结合附图和具体实施方式对本发明做进一步详细说明；

[0019] 图1为本发明基于机器学习的C&C通讯攻击检测方法的流程示意图。

具体实施方式

[0020] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图对本申请实施例中的技术方案进行清楚、完整地描述。

[0021] 如图1所示，本发明公开了基于机器学习的C&C通讯攻击检测方法，其包括以下步骤：

[0022] 步骤1，流量包过滤：获取连续下行流量包；目前现有网络环境中流量越来越大，而恶意软件下行流量包大部分较小，为了避免非相关流量没有意义的分析检测导致资源浪费对流量包进行过滤，使得流量包长度的分布呈正态分布，

[0023] 进一步地，作为一种较优实施方式，步骤1根据流量包长度的正太分布设置过滤阈值，将部分无相关流量过滤。具体的，通过设定包过滤率计算得小流量包的包长度临界值，最终过滤包长度采用正态分布估计和设定阈值方式综合计算来确定。

[0024] 步骤2，流量会话聚合：根据指定条件对流量包进行会话聚合；

[0025] 步骤3，利用随即簇抽样和Apriori算法提取会话流量特征；

[0026] 步骤4，利用对序列相似度检测采用编辑距离与求最长公共子序列(LCS)相结合的方式，对聚合之后的流量上下文数据进行相似度计算。

[0027] 步骤5，根据会话的下行流量上下文相似度是否超过设定值来判断是否异常C&C通讯。

[0028] 进一步地，作为一种较优实施方式，步骤2中根据源地址、源端口、目的地址或目的端口进行会话聚合。

[0029] 进一步地，作为一种较优实施方式，步骤3中抽样是指从总体中通过一定的抽样算法抽取出能代表总体的样本。通过对抽取样本的特征检测来预测总体特征，本发明检测连

续下行流量的payload中内容相似性，考虑到实际攻击过程中相同名利可能出现连续性的情况，所以采用随机簇抽样算法，如果处理数据量过大时，可以采用reservoir sampling算法经洗概率抽样。

[0030] 进一步地，作为一种较优实施方式，步骤4中先对序列对进行编辑距离计算，根据计算结果进行筛选去掉距离值较大的序列对，后对序列对进行ＬＣＳ计算。

[0031] 具体地，对下行流量包序列相似度的检测主要是基于求最长公共子序列(LCS)以及计算两序列的编辑距离的值算法结合。其中ＬＣＳ即最长公共子序列，通过求两个序列的最大公共子序长度从而求出两序列的相似度。一般都是采用动态规划的算法求最长公共子序列。其中编辑距离，又称为Levenshtein距离，表示从一个字符串转化为另一个字符串所需要的最少编辑次数，这里的编辑是指将字符串中的一个字符替换成另一个字符，或者插入删除字符。

[0032] 由于编辑距离的计算时间复杂度较低可先去掉一些无关序列对，又由于ＬＣＳ计算相似度更准确，从而使用检测结果更有可信度。

[0033] 本发明采用以上技术方案，对于网络流量中根据流量过滤，抽样，聚合之后，对聚合之后的会话流量数据进行上下文相似度检测，进而检测是否存在恶意软件通讯。本发明具有如下优点：1、不需要依赖特征库，可以检测未被发现的恶意软件通讯。2、不同于现有恶意软件有监督机器学习方法检测主要是基于现有恶意软件的流量特征进行有监督学习，其检测效果更多的依赖于机器学习的训练集的覆盖广度以及学习方法的科学性。3、对Ｃ＆Ｃ通信检测上，基于下行payload相似度检测算法相对于流量包检测算法和payload特征码检测有较高的准确率和召回率，同时在检测时间也有一定的优势，尤其对大量攻击流量样本进行检测时，检测时间复杂度较低，检测时间更短。

[0034] 显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。通常在此处附图中描述和示出的本申请实施例的组件可以以各种不同的配置来布置和设计。因此，本申请的实施例的详细描述并非旨在限制要求保护的本申请的范围，而是仅仅表示本申请的选定实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

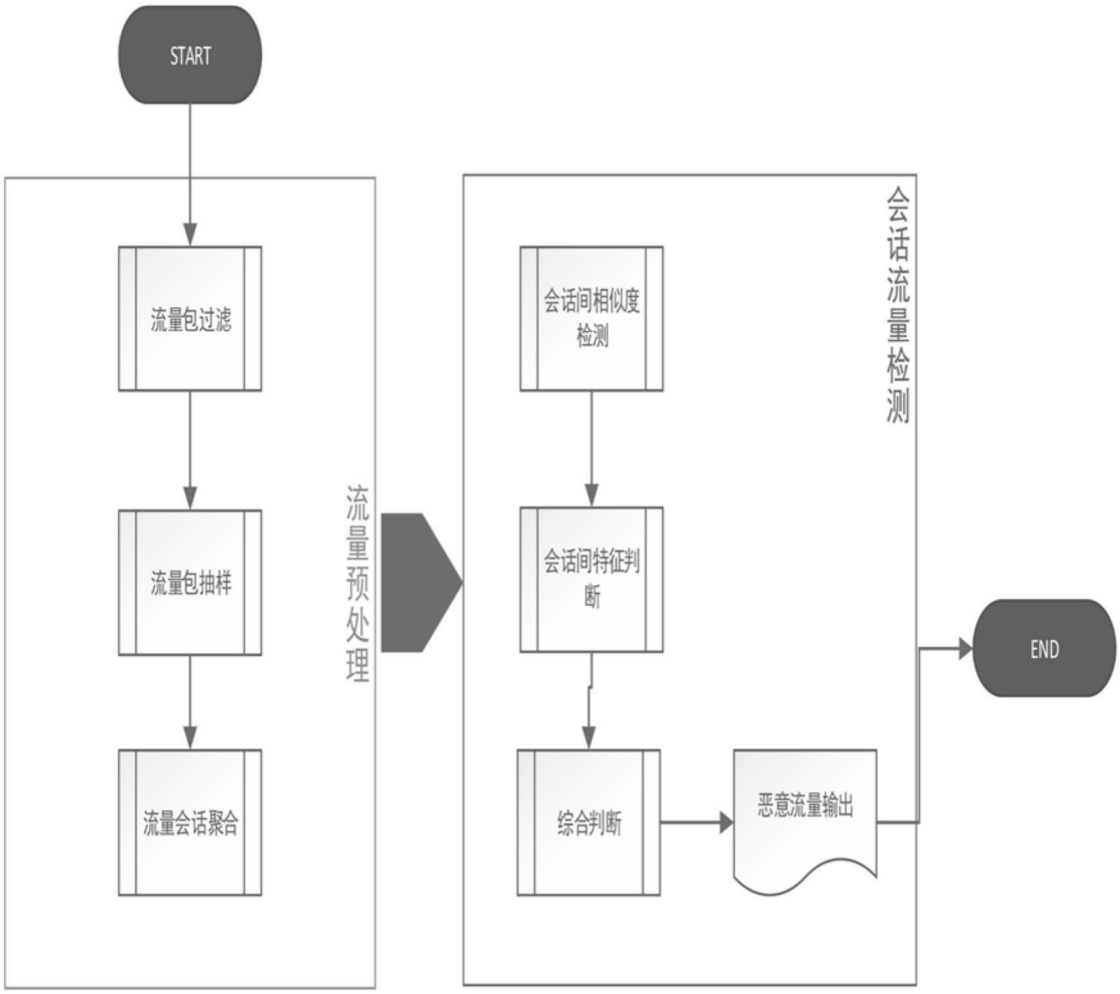

START
流量包过滤
流量包抽样
流量会话聚合
流量预处理
会话流量检测
会话间相似度检测
会话间特征判断
综合判断
恶意流量输出
END

图1