第 45 卷 第 11 期 计 算 机 工 程 2019 年 11 月
Vol. 45 No. 11 Computer Engineering November 2019

·人工智能及识别技术· 文章编号：1000-3428(2019)11-0234-09 文献标志码：A 中图分类号：TP18

基于流知识图谱的通信网络流连接行为分析

胡航宇，翟学萌，胡光岷

（电子科技大学 宽带光纤传输与通信网技术教育部重点实验室，成都 611731）

摘 要： 图模型能够直观、完整地刻画网络流的连接模式，在网络流行为分析中具有独特的优势，但现有图模型方法存在构图方式单一、信息包含不完整、分析手段不够丰富等问题，通过借鉴知识图谱的概念，提出一种基于流知识图谱的网络流行为分析模型——网络流连接图。通过收集网络流信息构造网络流连接关系的基本模型，基于网络流属性信息设定图节点等级和边权值，在此基础上，利用节点与边的筛选规则提取网络应用行为的核心连接方式和简化网络规模，采用复杂网络特征分析方法提取网络流行为特征参数。实验结果表明，网络流连接图能够充分利用网络流行为测量数据中的可用信息，准确刻画网络应用流连接关系的固有特征，有效地检测与识别 DDoS 攻击、蠕虫传播以及端口扫描等网络异常行为，同时网络流连接图表现出良好的可扩展性，适合多种图挖掘算法的应用。

关键词： 网络流行为分析；网络流；知识图谱；特征参数提取；异常检测

开放科学（资源服务）标志码（OSID）：

中文引用格式： 胡航宇，翟学萌，胡光岷. 基于流知识图谱的通信网络流连接行为分析[J]. 计算机工程，2019，45(11)：234-242.

英文引用格式： HU Hangyu，ZHAI Xuemeng，HU Guangmin. Analysis of communication network flow connection behavior based on flow knowledge graph[J]. Computer Engineering，2019，45(11)：234-242.

Analysis of Communication Network Flow Connection Behavior Based on Flow Knowledge Graph

HU Hangyu，ZHAI Xuemeng，HU Guangmin

（Key Laboratory of Optical Fiber Sensing and Communications，Ministry of Education，University of Electric Science and Technology of China，Chengdu 611731，China）

【Abstract】 The graph model method has unique advantages in network flow behavior analysis，because it can intuitively and completely describe the connection mode of network flow. However，the current methods have many problems，such as single composition mode，incomplete information and insufficient analysis means etc. Therefore，by referring to the concept of knowledge graph，this paper proposes a network flow behavior analysis model based on flow knowledge graph，namely，the network flow connection graph. We first build the basic model of the network flow connection relationship by collecting the network flow information. Then we set the graph node level and the edge weight value based on the network flow attribute information. According to the filtering rules of the node and edge，we extract the core connection mode of the network application behavior and simplify the network scale. Finally，we adopt the complex network feature analysis method to extract the network flow feature parameters. Experimental results show that network flow connection graph can fully utilize the available information in the network flow behavior measurement data，accurately characterize the inherent characteristics of the network application flow connection relationship，and effectively detect and identify network abnormal behaviors such as DDoS attacks，worm propagation and port scanning. Besides，the network flow connection graph shows good scalability，making it suitable for the application of multiple graph mining algorithms.

【Key words】 network flow behavior analysis；network flow；knowledge graph；feature parameter extraction；anomaly detection

DOI： 10.19678/j.issn.1000-3428.0052745

基金项目： 国家自然科学基金(61471101，61571094)。

作者简介： 胡航宇(1988—)，男，博士研究生，主研方向为网络行为分析；翟学萌，博士研究生；胡光岷，教授、博士生导师。

收稿日期： 2018-09-25 **修回日期：** 2018-11-12 **E-mail：** huhangyuuestc@gmail.com

0　概述

目前信息技术和互联网应用服务发展迅速,多样化的网络用户和终端导致了网络流和流量向高速化以及大数据量方向发展。理解和掌握网络流行为有助于预测和识别网络空间各种事件的发生,因此,全面准确地分析与挖掘网络流行为特征是构建安全可靠网络环境的前提条件,也是目前学术界和工业界共同关注的前沿科学问题之一。

网络流行为分析是通过对不同时间段、不同地理位置产生的网络流行为信息进行整合和提取,然后建立流行为的多种表征模型,运用一系列科学分析方法对模型进行分析,获得充分刻画网络流行为的特征参数,挖掘隐藏在多条网络流之间的关联关系,及时发现网络异常行为与正常行为的区别,总结网络流行为的演化规律。传统的网络流行为分析方法主要通过对网络数据本身采用精细分析的手段获得相应特征,如:采用深度数据包检测(DPI)方法[1]对网络数据报文的负载进行提取与分析;将网络流量看成是一种随时间变化而变化的信号,采用时间序列分析方法[2-3]提取其时域特征,从而对网络流行为进行建模、预测与识别;利用网络流量数据的基本特征(数据包长度、流持续时间等)及其高阶统计量特征(数据包长度小波分析特征、包长度高阶矩、到达时间间隔高阶矩、传输速率高阶累积量等)采用机器学习的方法[4]对网络流进行分类,从而挖掘具有相似特征的网络流行为集合。然而,持续增大的通信网络规模以及海量网络主机交互数据使得传统精细分析的方法开销越来越大,并且无法应对网络行为分析的实时性要求,同时考虑到各种安全因素的影响,越来越多的网络应用业务在网络层的数据包分组中采用流量加密和流量混淆等技术对负载信息进行处理,使得精细网络流行为分析方法的难度日趋增大。

网络流是网络运行过程的信息载体,记录了源目的主机之间的通信交互信息。具体地说,网络流描述的是在一个时间窗口内一对 IP 地址的通信数据(时间戳、IP 地址、端口信息、协议等)。由于每一条网络流都连接至少两台主机和端口,网络流的连接行为表示了服务器、终端用户、交换设备及其所承载的各种应用服务之间的连接模式,因此以深度流分析为研究手段,通过建立网络应用、用户以及其他网络实体之间产生的网络流连接关系模型,对不同时间段、不同地理位置产生的网络流行为特征进行测量与分析,尤其是对网络应用流行为的建档与网络异常流行为的发现与预警成为目前网络行为分析的热点研究话题。

通过借鉴知识图谱的概念,本文提出一种基于流知识图谱的网络流行为分析模型——网络流连接图(Network Flow Connectivity Graphs,NFCCs)。该模型将网络主机抽象为图中节点,将主机间通信行为抽象为边,基于网络流属性信息分别为节点和边设定相应等级和权值,并且用不同颜色对赋值结果进行可视化处理。以 NFCGs 为基础,根据不同研究目的可以设计不同过滤规则对初始模型进行处理和简化,从而提取网络应用核心连接关系,同时根据复杂网络特征分析的方法计算多种网络流连接图特征,通过比较网络正常流与异常流在连接行为特征以及子图结构方面的差异,挖掘现有方法难以获得的疑似异常行为子图结构。

1　相关研究

基于图模型的流连接行为分析方法[5]能够准确地捕获数据个体之间的连接关系,因此,把网络流行为信息与图进行融合得到新的连接图模型是网络流连接行为分析的一种趋势。

文献[6]提出采用有监督的盲分类 BLINC 方法通过提取网络流传输层的特征建立单个主机的连接模式,分别从网络行为的社交、功能及应用 3 个层面分析其相应的连接模式,采用图模式匹配的方法进行网络应用流分类。文献[7-9]提出了网络流量传播图(Traffic Dispersion Graphs,TDGs)的概念来描述网络交互之间的连接关系。文献[10-11]则在 TDGs 的研究基础上提出了流量活动图(Traffic Activity Graphs,TAGs),并采用改进后的正交非负矩阵分解方法对 TAGs 图进行降维处理,从而提取主机交互的核心连接方式,通过分析不同应用行为 TAG 图的连通度、演化过程以及子图相似性来推断特定新应用类型的发生以及识别蠕虫传播行为。文献[12-14]利用二部图建立网络源目的主机间的连接模式,进而通过一模投影的方式将二部图划分为源主机组与目的主机组两类,两类主机组内均包含多个子图结构,采用无监督聚类方法将具有相似聚类系数的主机聚为一类,采用该方法能够准确地区分 C/S 和 P2P 网络应用行为,且不需要训练样本,对于新出现的业务也能够准确划分到具有相似行为的类中。文献[15]则针对分类算法中的“概念漂移”现象,提出了用户连接图(Host Connectivity Graphs,HCGs),应用“图挖掘”的理论将用户连接图划分为互不相交的行为子簇,计算行为子簇内网络流量属性的信息熵区分用户行为,以准确地进行网络应用流分类。文献[16]提出将网络流行为信息和基于图的分数传播方法相结合以检测通信网络 HTTP 流中恶意客户机行为。

上述图模型技术及其分析方法在网络流行为分析中都取得了很好的效果,然而它们大多只考虑主机之间是否存在连接,没有考虑主机之间连接的详细情况和紧密程度,也即上述图模型中的主机之间

存在多条连接与一条连接等价的情况,因而不能充分利用连接关系的可用信息,难以全面挖掘网络流行为,不能有效地运用于大规模网络流连接行为的分析。

随着网络结构越来越复杂,仅依靠连接关系是否存在,无法准确描述和刻画网络流连接行为。为了能够全面刻画骨干通信网络流行为的连接特性,本文认为图模型应满足以下 4 项条件:

1)图包含尽可能多的网络流连接行为的信息量,使得大部分与网络流连接行为相关的特征可以从图中轻易提取得到。

2)图不仅能直观展现连接关系,还应该足够简洁清晰,有较好的可视化效果。

3)能够通过多种简化方案获得简化的子图结构,通过简化子图反映某种特点或是某种特定的网络流行为。

4)图模型应与复杂网络具有某些相似的特征,便于结合现有网络科学研究思路研究其特征分析与演化规律挖掘方法等。

本文通过借鉴知识图谱的概念,基于网络流行为特征建立一种网络流知识图谱,进而通过复杂网络与大数据图分析的方法对网络流行为进行发现与分析,挖掘通信网络流之间的潜在关系,讨论网络流连接行为的发展趋势。

2 基于流知识图谱的网络流连接图构建

随着骨干通信网络的快速发展,对海量网络数据的处理与分析是影响网络行为分析准确性的重要元素。本节在宏观层面对网络流行为进行分析,主要介绍基于流知识图谱的网络流连接图成图方法与依据。知识图谱的概念最早由美国谷歌公司提出[17],是谷歌公司用于增强其搜索引擎功能的辅助知识库。

定义 1(知识图谱) 知识图谱是结构化的语义知识库,用以符号形式描述物理世界中的概念及其相互关系。其基本组成单位是“实体-关系-实体”三元组,以及实体及其相关属性值,实体间通过关系相互联结,构成网状的知识结构[17-18]。

针对网络流行为信息,如何将流信息与图中的元素相结合,以期获得全面表征网络流连接关系的图结构,是本文需要解决的关键问题,同时也是构图的重要思路。本文采用基于流知识图谱的方法,对通信网络的主机进行连接,进而分析与挖掘网络流之间的连接关系。

定义 2(流知识图谱) 流知识图谱是刻画网络流传播过程中网络实体连接关系的一种知识图谱组织形式,借鉴知识图谱的基本概念,通过网络流连接关系将各种网络实体关联起来,进而通过对网络流量数据进行抽取、融合,构成结构网络。

2.1 网络流连接图的生成

首先对网络流数据进行处理以获得网络流信息序列。传统网络流通常使用五元组的形式进行定义,即拥有相同 < srcIP, dst IP, srcPort, dstPort, protocol > 序列信息的数据包可以合并成为一条网络流。然而,采用五元组网络流定义的方式会造成图中两点之间存在多条连接的情况,这不仅会增加分析的难度,也同时影响了可视化的效果,为此本文提出采用聚合流的概念生成网络流信息序列。

定义 3(聚合流) 聚合流是满足以下 2 个条件的最大流集合:1)聚合流都是由同一组源主机与目的主机对组成,不考虑它们之间存在的端口号或者承载的协议;2)2 条网络流之间的时间间隔应该比较小,隔了较长时间的网络流,它们之间往往不具备任何关联关系。

聚合流集合描述的是主机对之间所有连接信息,其他流连接的信息可以用特征向量表示,以此建立网络流信息序列 $N = <H, F>$。为全面描述网络流行为特征,将网络中拥有 IP 地址的主机或网关等实体对象抽象为图中的节点 $v_i \in V$,若节点 i,j 之间有通信交互则将对应的节点连成一条边 $e_{i,j} \in E$,再结合网络流信息多个属性特征,从而构造出的图模型结构 $G = <V, E, A>$,具体算法如下:

算法 网络流连接图 NFCGs 生成算法

输入 网络流信息序列 $N = <H, F>$

输出 网络流连接图的模型 $G = <V, E, A>$

1)收集网络流行为特征信息,提取网络流中的主机连接关系以及统计相应属性的量值大小。

2)根据研究目的的不同生成图中最基本的节点集 $v_i \in V$ 与边集 $e_{i,j} \in E$。

3)根据不同研究目的,所需网络规模以及节点和边属性统计值大小设定节点和边阈值,进行初步筛选。

4)根据节点属性统计量值大小对节点进行分级处理,并将节点设置为不同颜色去区分等级高低,同时根据量值大小给边设定权值。

5)利用可视化软件后生成网络流连接图。

图 1 所示为基于 NFCGs 的网络流连接行为分析的处理过程。该过程包括图模型的建立、网络核心连接关系的提取、网络流连接行为特征参数提取和流行为的定量分析 4 个模块。图 2 则分别展示网络流量传播图和网络流连接图的模型示意图。可以看出,NFCGs 可以看作是带有权值的网络流量传播图,这些信息的嵌入有利于更加清晰地掌握和理解网络流连接的详细情况,因此,在网络流连接行为分析中,网络流连接图的构建是很重要的,而且必要的。

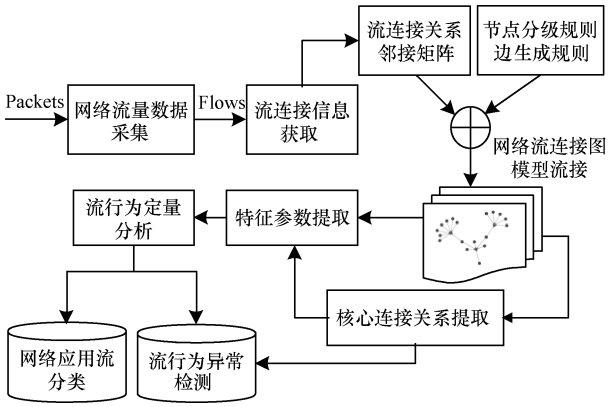
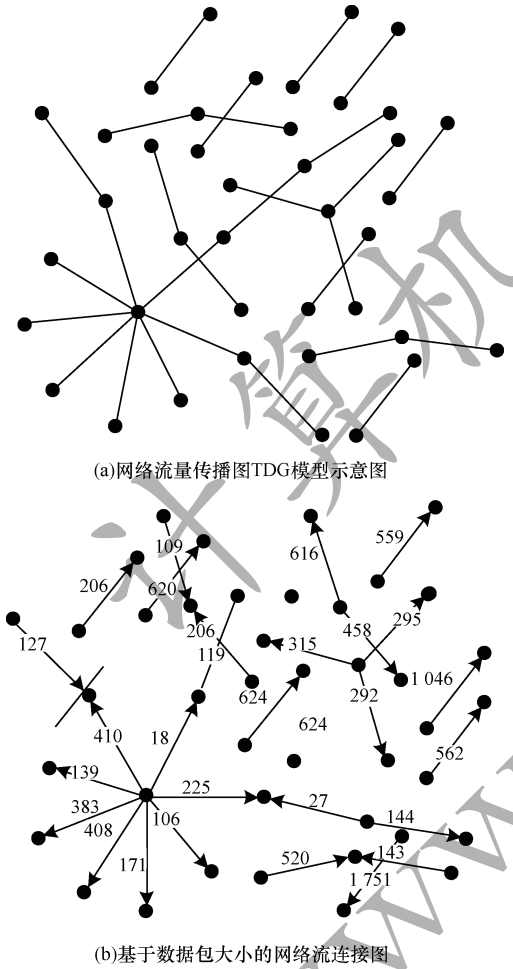


图 1　基于 NFCGs 网络流连接行为分析的处理过程



(a)网络流量传播图TDG模型示意图

(b)基于数据包大小的网络流连接图

图 2　NFCGs 模型示意图

2.2　网络核心连接关系的提取

由于骨干通信网络流量规模巨大,网络结构复杂,因此如何选择合适的时间间隔范围构建 NFCGs 图模型是一项难题。如果时间间隔比较大,则节点需要可视化的数量有很多,最终形成的图会变得特别复杂,增加了流连接行为的分析难度;而时间间隔比较小,则难以全面展现网络流行为的信息。此外,在网络流信息提取过程中会捕获得到一些弱连接结构,这些弱连接结构通常是一些不相关或是不重要的连接关系,节点的等级也比较低,并且往往是随机产生的,它们可能在一个时间间隔内生成,而在下一个时间间隔内就停止,这些弱连接结构相当于网络流连接的冗余结构,若冗余结构大量存在会影响网络流行为分析的准确程度,因此,需要提取反映网络流的主要连接关系。

为解决以上问题,本文采用节点与边的过滤规则选择适合的属性特征向量,例如包含 2 台主机连接的容量(如数据包数量、字节数)、连接的数量、端口开放的种类和连接存活时间等,从而删去图中的不重要或者无关联关系的节点与边,最终获得网络流连接行为的核心连接关系。通过引入节点的属性和等级以及赋予边权值,NFCGs 图能够扩展得到多种属性和等级机制定义下的加权图,与其他图模型方法相比能够包含和展现更多网络流连接信息。如图 3所示,2 台主机流交互过程端口开放种类大于 1 000的网络流连接图,其中边权值表示的是端口开放的种类数目。

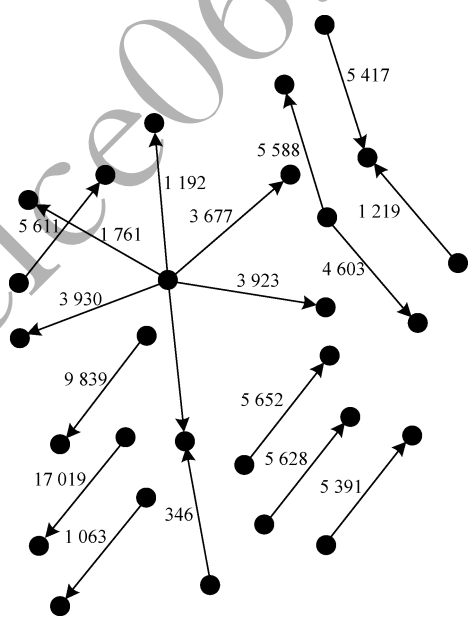


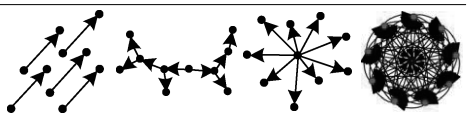
图 3　端口开放种类大于 1 000 的网络流连接图

3　网络流连接行为特征参数

网络流连接行为图模式特征是指通过构建网络连接图,提取能够对相应连接图进行描述的特征。这类特征主要描述的是网络主机通信交互模式的结构信息,解决了网络主机之间“和谁进行了通信”“同时和多少主机通信”“与谁交互行为相似”等结构性问题。

以网络流连接图为基础,本文将网络流连接图看成一种加权的复杂网络:每个节点具有不同的等级或属性,每条边都具有相应的权值。通过借鉴复杂网络特征参数分析研究思路[19],采用不同的特征值描述流连接图,提取适用于不同规模网络行为特征分析的多种参数,从而对不同网络应用流行为进行定量分析。表 1 所示为 4 种基本连接方式的图特征参数示例。

表 1 4 种连接关系的图特征参数

连接方式	点-点连接	树型连接	星型连接	全连接
示意图模型				
平均度	1	1.8	1.8	9
RCD 最大值	0.11	0.33	1	1
入度节点比例/%	50	50	90	0
出度节点比例/%	50	10	10	0
出入度比例/%	0	40	0	100
最大深度	1	3	1	8

3.1 基本特征

该类特征描述的是网络拓扑结构的基本参数,如节点数量、边数量和节点平均度等。该类特征的变化反映了网络中节点与边的增加或减少,网络空间事件的发生可能会引起这类特征的突然变化,例如蠕虫病毒的传播会导致这类特征的大量增加,FTP 服务器的突然关闭则会导致这类特征减少。

平均度:节点度的k_i定义是与节点直接相连边的条数,所有节点度的平均值称为网络流连接图的平均度。该项参数是度量网络流连接关系复杂程度的最基本量化参数,反映了目标节点的活跃程度,即在观测时间间隔内不同网络主机与目标节点有交互行为的数量。网络流连接关系越稠密,图中节点的平均度越大;而相反地图中孤立点越多,连接关系越稀疏,则节点平均度越小。给定网络流连接图 G 的邻接矩阵 $\boldsymbol{A}=(a_{ij})N\times N$,则有:

$$k_i=\sum_{j=1}^{N}a_{ij}=\sum_{i=1}^{N}a_{ji} \tag{1}$$

$$<k>=\frac{1}{N}\sum_{i=1}^{N}k_i=\frac{1}{N}\sum_{i,j=1}^{N}a_{ij}=\frac{2E}{N} \tag{2}$$

相对连接密度(Relatively Connectivity Density,RCD):该项参数计算的是图中节点度数与度数最大可能值的比值,反映了网络流连接图中节点的相对集中程度。当 RCD 值越趋近于 1 时,表示图中有可能出现“主导”节点的存在;而当 RCD 值等于 1 时,网络结构将会呈现如表 1 中星型连接和全连接两种连接关系模式。同时 RCD 的值还能够帮助检测某些网络恶意行为的存在,如当 DDoS 攻击、恶意扫描等行为发生时,RCD 值会突然升高。给定节点数为 N 的网络流连接图,其中最大相对连接密度描述了网络流连接图中度数最大节点的度中心性,其定义为:

$$RCD_{\max}=\frac{k_{\max}}{N_{\text{total}}-1}\times 100\% \tag{3}$$

出入度占比:该项参数描述的是网络流连接图中只有入度的节点、只有出度的节点、出入度都有的节点分别占全部节点数量的比例。

3.2 状态特征

该类特征表示 NFCGs 图的节点和边在网络全局整体结构中的状态和紧密程度。这些特征包括节点强度、图的连接密度、节点聚类系数、图的聚类系数等。某些网络空间事件的发生不会引起网络拓扑结构的改变,但会对节点和边的状态产生影响,例如 Alpha flow 这类异常行为并不会改变节点的连接关系,但会使得边权值向量中的流连接数量增大从而使得节点强度发生改变。

最大连通片大小:指图中最大连通子图所包含节点数与图总节点数的比值。在网络流连接图中,连通片满足 2 个条件:1)连通性,该子图中的任意 2 个顶点之间都存在路径;2)孤立性,网络中不属于该子图的任一顶点与该子图中的任一顶点之间不存在路径。图中包含顶点数最多的连通片就称为最大连通片,最大连通片越大的图,节点间连通性就越好。在节点数为 N 的网络流连接图中,假设最大连通片中的节点数为 N_{GCC},则最大连通片大小定义为:

$$GCC=\frac{N_{GCC}}{N}\times 100\% \tag{4}$$

最大深度:指网络流连接图中最大单个方向链路的条数,该参数能够直观地帮助区分网络应用客户机/服务器 C/S 结构和 P2P 应用,从而准确地实现应用流分类。

3.3 统计特征

该类特征描述了节点和边的概率分布、不确定程度以及稳定性度量等,包括节点度分布、富人俱乐部连接特性和联合度分布。在大多数情况下,网络流行为的统计特征随着时间的变化而表现出一种近乎随机的规律,网络空间事件的发生则会影响这种规律,例如 DDoS 攻击不一定会对图基本特征和状态特征造成改变,但是统计特征方面则会有明显的变化。因此,根据网络事件的不同特点,及早地发现

相应规律的变化能够帮助识别网络事件的发生。

节点度分布:该项参数计算的是图中节点度数的概率分布情况。在网络流连接图中,定义度为 k 的节点数占整个网络节点数的比例为 P_k,从概率统计的角度看,P_k 为网络中随机选择节点的度为 K 的概率,该概率分布就是节点度分布。

富人俱乐部连接特性:将网络流连接图中的节点度数按降序排列,计算第 1 名到第 K 名的节点之间的实际边数 $E_{1,k}$ 以及这些节点的理论最大边数 $E_{k\max}$,将两者相比得到网络流连接图的富人俱乐部连接程度,即定义为:

$$RCC_k = \frac{N_{1,k}}{N_{k\max}} \times 100\% \tag{5}$$

联合度分布:随机在图中取一条边,记边 2 个端点度数 k_1、k_2 的概率为 P_{k_1,k_2},该概率分布即为联合度分布。

4　实验结果与分析

本文在实验中使用了 2 个不同的骨干通信网络数据集:一个是 CAIDA OC-48,采集于美国某个大型节点的骨干通信网络链路 5 min 流信息文件,共 1 h 数据;另外一个是日本 MAWI 工作组主导的 WIDE-Project,为 2017 年 4 月 7 日采集的 1 h 流信息文件。图 4 为 CAIDA OC-48 链路 1 000 条流的网络流连接图。

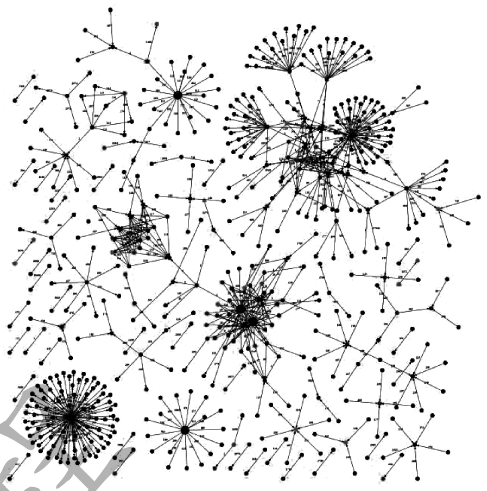


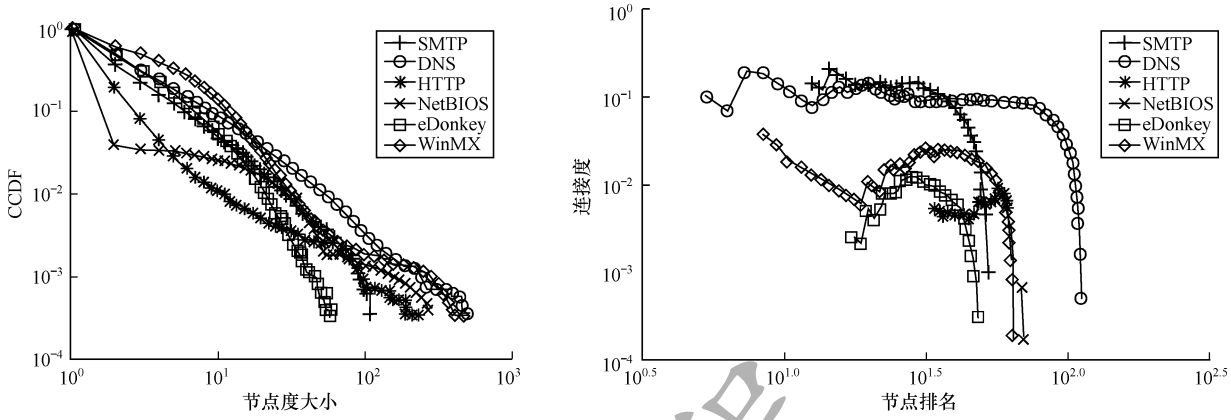
图 4　CAIDA OC-48 链路 1 000 条流的网络流连接图

4.1　网络应用流连接行为分析

为进一步研究网络流连接图的基本特征,本文对图的节点度分布、富连接性和联合度分布 3 个图分布参数进行了计算,表 2 是根据 8 种已知端口提取得到应用行为网络流连接图的基本特征参数;图 5 是其中 6 种网络应用节点度分布图和富人俱乐部的连接特性图;图 6 则以等值线图的形式表示 SMTP、DNS、HTTP、NetBIOS、eDonkey 以及 WinMX 应用的联合度分布,其中,横纵坐标分别表示边的两端点的度数取对数,图中颜色是该边的概率取对数后取绝对值,颜色越深表示概率越大,白色表示概率为 0。

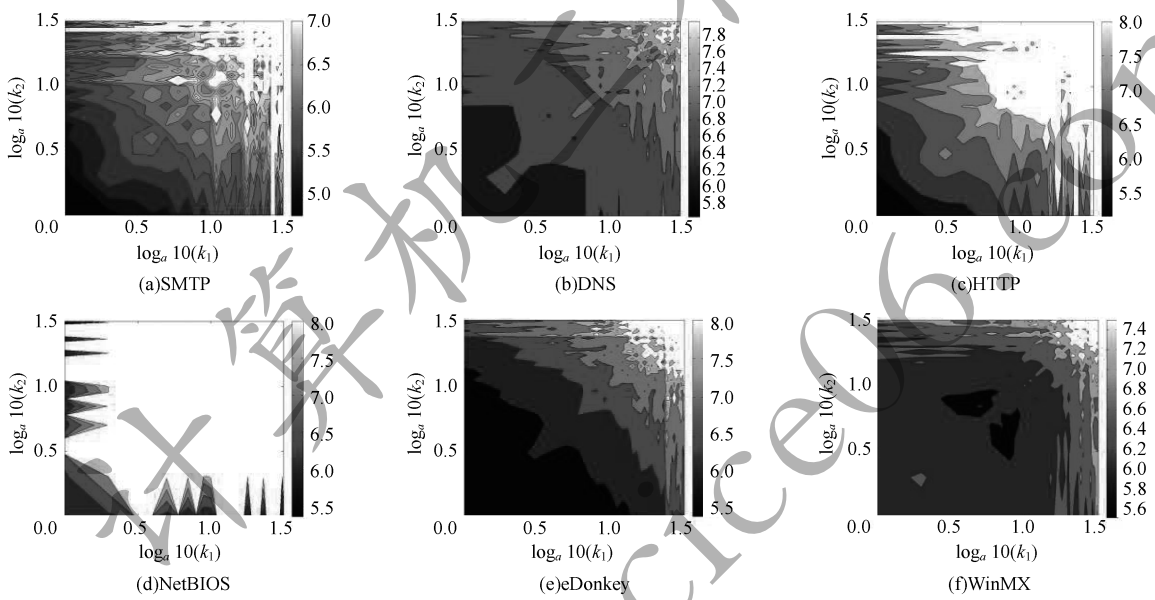
表 2　多种网络应用流行为图特征参数与经过节点和边过滤后图特征参数对比提取结果

应用类型	节点阈值	边阈值	节点数	边数	平均度	最大连接密度/%	入度节点比例/%	出度节点比例/%	最大连通片/%	最大深度
SMTP(25)	0	0	3 146	4 345	2.76	3.66	47.97	52.10	32.25	45
	100	20	73	110	3.01	41.67	68.49	31.51	42.47	9
DNS(53)	0	0	9 155	20 265	4.43	7.14	36.70	63.62	92.61	337
	100	20	79	158	4.00	17.95	46.84	53.16	70.89	15
HTTP(80)	0	0	12 889	13 185	2.05	6.39	25.39	74.61	65.40	4
	500	40	304	312	2.05	33.33	17.43	82.57	59.54	4
NetBIOS(137)	0	0	10 969	10 523	1.92	4.46	95.54	4.47	4.54	3
	1	1	1 161	2 003	3.45	56.21	89.75	10.25	56.33	5
eDonkey(4662)	0	0	10 161	14 355	2.83	0.94	36.55	63.47	85.86	5
	10	5	480	991	4.13	4.59	55.42	44.58	99.17	3
WinMX(6257)	0	0	5 966	14 015	4.70	5.52	38.55	61.46	98.27	523
	20	5	127	215	3.39	7.94	51.18	48.82	98.43	22
FTP(21)	0	0	6 682	6 071	1.82	4.83	65.42	34.58	35.91	3
	100	20	103	87	1.69	43.14	15.53	84.47	43.69	2
Telnet(23)	0	0	4 013	3 893	1.94	4.61	96.94	3.06	4.63	2
	20	5	30	29	1.93	100.00	96.67	3.33	100.00	2



(a)节点度的互补累积分布 (b)富人俱乐部特性

图 5 SMTP、DNS、HTTP、NetBIOS、eDonkey、WinMX 的节点度分布以及富人俱乐部连接特性



(a)SMTP (b)DNS (c)HTTP

(d)NetBIOS (e)eDonkey (f)WinMX

图 6 SMTP、DNS、HTTP、NetBIOS、eDonkey 以及 WinMX 的联合度分布等值线分布

上述实验结果验证了不同的网络应用具有差异很大的特征参数,通过计算相应特征参数可以建立网络应用的特征参数库,对于采用隐匿端口或是加密端口技术进行网络数据传输的应用流行为,能够通过计算其特征参数以进行参数的模糊匹配,进而实现未知网络应用流行为的识别。此外,图主要连接关系结构中节点的平均度会更接近理论值,最大连接密度升高,深度降低,而对于非 P2P 应用,最大连通片大小会降低,这是因为阈值的筛选会将次要节点过滤掉以凸显网络应用的核心结构,且能够扩大应用行为基本特征参数之间的差异,从而提高流行为识别的准确性。值得注意的一点是 NetBIOS 应用行为在统计量筛选阈值为 1 时,节点数量和边数量急剧减少,说明 NetBIOS 应用行为有很多流数量为 1 的节点和边,体现了 NetBIOS 应用行为的独有特征。

4.2 网络异常流行为的检测与识别

当不同网络异常行为发生时,根据其特点所对应连接行为图的变化也是不同的,从而造成各个应用行为的特征参数发生不同变化,表 3 给出常见网络异常及其行为描述。

表 3 通信网络中常见异常及其行为描述

网络异常行为	描述
DoS/DDoS 攻击	攻击者以极大的通信量冲击目标网络或用极大量的连接请求冲击目标主机,以消耗可用的网络资源或系统资源,致使目标网络或主机瘫痪
Alpha 攻击	点到点的高比特率传输,如点到点之间大文件的传输
FlashCrowd	对单一目标的资源和服务突发性地大量请求
端口扫描	对目标主机所有端口进行扫描,以找出主机安全隐患
网络扫描	对目标网络所有主机的一组端口进行扫描,以找出存在隐患的主机
蠕虫传播	一种特殊的网络扫描,通过扫描网络中存在隐患漏洞的主机,利用安全漏洞进行传播的自我复制传播模式感染网络上的主机,会破坏蚕食系统使其完全瘫痪

表 4 给出当 DDoS 和蠕虫传播异常行为分别通过熟知端口(25、80 以及 4662)扩散以及同种应用行为下网络流连接图的基本特征参数,通过比较网络正常应用行为与网络异常行为在流连接行为特征参数的差异性,能够有效地检测与识别网络异常流行为。

表 4　正常网络流行为与注入异常流行为之间的差异性

应用类型	节点数	边数	平均度	最大连接密度/%	入度节点/%	出度节点/%	最大连通片/%	最大深度
SMTP(含 DDoS)	800	785	1.96	32.17	97.88	2.13	75.43	2
SMTP(正常)	3 146	4 345	2.76	3.66	47.97	52.10	32.25	45
HTTP(含 DDoS)	2 344	2 272	1.94	78.79	94.54	5.46	78.80	3
HTTP(正常)	12 889	13 185	2.05	6.39	25.39	74.61	65.40	4
eDonkey(含蠕虫)	5 769	5 040	1.75	1.06	20.96	79.04	40.80	295
eDonkey(正常)	10 161	14 355	2.83	0.94	36.55	63.47	85.86	5

通过表 3、表 4 可以得出以下结论:

1)通过 25 端口进行 DDoS 传播的主机占用了大量 SMTP 邮件服务器的资源,与大量用户产生连接关系,因此,网络流连接行为特征会表现出节点数量和边数量减小,最大连接密度的增大,只有入度的节点比例爆发性增长,图深度大幅降低。

2)当 DDoS 攻击发生时,网络流连接行为特征参数表现出连接密度增大,入度的节点比例爆发性增长,最大连接子图结构所占比例大幅增大。

3)蠕虫传播通常起始于一台或少量主机,然后快速地扩展到整个网络主机中,这样的行为会导致短时间内被感染的网络连接主机数量突然增大,使得整个网络结构出现扩散的特性,因此,网络流连接行为表现出节点数量和边数量增多,节点平均度降低,最大连接子图比例增大,图深度大幅增大。

图 7 所示为 DDoS 与蠕虫传播异常行为的联合度分布,DDoS 异常行为的联合度分布图与正常的分布图差别巨大,只有少数种类度数的节点相连,说明了 DDoS 单一分布的特性,而蠕虫传播的分布较为均匀,体现了其均匀散播的特性。网络异常行为的发生使得网络主机连接行为发生相应变化,造成各个应用行为的特征参数发生不同变化,因此,可以通过分析特征参数的突变检测识别网络异常流行为。

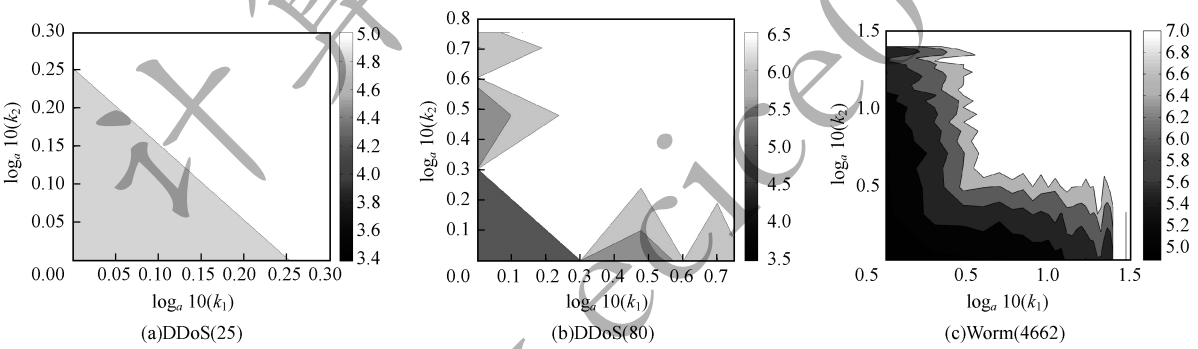


图 7　DDoS(25)、DDoS(80)、Worm(4662)异常行为的联合度分布

5　结束语

本文针对现有图模型研究方法构图方式单一、信息包含不完整、分析手段不够丰富等问题,提出一种新的网络流连接行为分析模型——网络流连接图。通过对节点赋予不同的属性和等级以及对边赋予不同的权值,直观地将网络流行为信息展现在图中,并且根据不同研究目的对图进行处理与简化得到核心连接关系结构,通过比较网络正常流与异常流在图特征参数上的差异性,挖掘现有方法难以获得的疑似异常行为。实验结果表明,网络流连接图能够有效地检测 DDoS 攻击、蠕虫传播以及端口扫描等网络异常流行为,从而更全面准确地理解与管理骨干通信网络。下一步将研究网络流连接行为的动态演化特征参数提取方法以及网络流连接图的动态网络子图挖掘方法,在构建网络流连接图的基础上,提取符合异常事件连接关系模式的模体(Motif)结构,进而开展基于模体挖掘的大规模网络异常事件识别的研究。

参考文献

[1] BOUKHYOUTA A,MOKHOV S A,LAKHDARI N E,et al. Network malware classification comparison using DPI and flow packet headers[J]. Journal of Computer Virology and Hacking Techniques,2016,12(2):69-100.

[2] BARFORD P,KLINE J,PLONKA D,et al. A signal analysis of network traffic anomalies[C]//Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment. New York,USA:ACM Press,2002:71-82.

[3] ZHOU Yingjie,HU Guangming. GNAED:a data mining framework for network-wide abnormal event detection in

backbone networks[C]//Proceedings of IPCCC'11. Washington D. C., USA: IEEE Press, 2011: 1-2.

[4] NGUYEN T T T, ARMITAGE G. A survey of techniques for internet traffic classification using machine learning[J]. IEEE Communications Surveys and Tutorials, 2008, 10(4): 56-76.

[5] AKOGLU L, TONG H, KOUTRA D. Graph based anomaly detection and description: a survey[J]. Data Mining and Knowledge Discovery, 2015, 29(3): 626-688.

[6] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark[J]. Computer Communications, 2005, 35(4): 229-240.

[7] ILIOFOYOU M, PAPPU P, FALOUTSOS M. et al. Network monitoring using traffic dispersion graphs [C]// Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. New York, USA: ACM Press, 2007: 315-320.

[8] ILIOFOYOU M, PAPPU P, FALOUTSOS M, et al. Network traffic analysis using traffic dispersion graphs: techniques and hardware implementation[EB/OL]. [2018-08-10]. https://www.docin.com/p-1647584729.

[9] ILIOFOYOU M, KIM H C, FALOUTSOS M, et al. Graph-based p2p traffic classification at the internet backbone [C]//Proceedings of IEEE INFOCOM Workshops. Washington D. C., USA: IEEE Press, 2009: 1-6.

[10] JIN Yu, SHARAFUDDIN E, ZHANG Zhili. Unveiling core network-wide communication patterns through application traffic activity graph decomposition[C]// Proceedings of the 11th International Joint Conference on Measurement and Modeling of Computer Systems. New York, USA: ACM Press, 2015: 49-60.

[11] JIANG N, CAO J, JIN Y, et al. Identifying suspicious activities through DNS failure graph analysis[C]// Proceedings of the 18th IEEE International Conference on Network Protocols. Kyoto, Japan: [s. n.], 2010: 5-8.

[12] XU Kuai, WANG Feng. Behavioral craph analysis of internet applications [C]//Proceedings of IEEE GLOBECOM'11. Houston, USA: [s. n.], 2011: 123-135.

[13] XU Kuai, WANG Feng, GU Lin. Network-aware behavior clustering of internet end hosts[C]//Proceedings of INFOCOM'11. Shanghai, China: [s. n.], 2011: 257-268.

[14] XU Kuai, WANG Feng, GU Lin. Behavior analysis of internet traffic via bipartite graphs and one-mode projections[J]. IEEE/ACM Transactions on Networking, 2014, 22(3): 236-245.

[15] 张震, 汪斌强, 陈鸿昶, 等. 互联网中基于用户连接图的流量分类机制[J]. 电子与信息学报, 2013, 35(4): 958-964.

[16] LIU L, SAHA S, TORRES R, et al. Detecting malicious clients in ISP networks using http connectivity graph and flow information[C]//Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Washington D. C., USA: IEEE Press, 2014: 150-157.

[17] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.

[18] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908.

[19] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社. 2012.

编辑 索书志