基于网络模块化结构的异常发现

王 娟* 秦志光 刘 峤 钱伟中
(电子科技大学计算机科学与工程学院 成都 611731)

摘 要：该文针对大规模高速网络海量数据和异常检测率较低的问题，将复杂网络的模块概念引入网络异常检测领域，化网络检测为数个网络模块检测的综合。首先通过建立网络划分策略与网络检测率关系模型，理论地证明按照网络本身所具有的模块结构划分网络有利于网络总体的检测。其次在真实网络采集的数据集上用并行处理技术进行实验，结果表明基于网络模块的检测比基于网络的检测能提供更加准确和高效的检测结果。

关键词：网络异常检测；网络模块度；分流
中图分类号：TP393 **文献标识码**： A **文章编号**：1009-5896(2011)01-0180-05
DOI: 10.3724/SP.J.1146.2010.00204

Anomaly Detection Based on Network Module Structure

Wang Juan Qin Zhi-guang Liu Qiao Qian Wei-zhong
(*School of Computer Science and Engineering*, *University of Electronic Science and Technology*, *Chengdu* 611731, *China*)

Abstract: The large scale and high speed networks create massive data and have low detection accuracy. To address the problems, the idea "module" is brought from complex network into anomaly detection area. Firstly, the relations between network partition strategy and network detection accuracy are modeled, and a theoretically proof is given that partition strategy which based on network modularity is favorable for anomaly detection. Secondly, the module-based detection is proved that has higher detection rate and efficiency than network-based detection by theoretical analysis and experiments. Finally, by using flow-splitting and parallel processing technologies this approach can improve efficiency obviously.

Key words: Network anomaly detection; Network modularity; Flow-split

1 引言

在对大规模网络进行异常检测的实践研究中，我们发现现有的方法[1–6]离实际应用还有一定差距。存在的问题主要有两个：其一是面对海量的网络数据，检测效率不能满足实时性的要求；其二是检测结果的漏报误报率较高。对于第一个问题，目前常用的应对方法之一是对原始数据抽样，但是这样处理对检测准确率往往有反效果；应对方法之二是对原始数据分流，该方法对提高效率有一定帮助，但是对检测率的提高也没有贡献。而后一个问题，除了各异常检测方法本身固有的缺陷外，影响检测率的主要因素是监控范围的选择。目前的方法基本都是以能监控到的整个网络为对象。而在大规模网络中影响全网的异常其实是很少出现的。常见的异常都是局部性质的，从全网角度观察势必受到无关网络活动的干扰而产生漏报误报。

因此，从检测的效率和准确率两方面考虑，本文将网络划分成范围较小的子网的集合，按照子网分流，并分别进行异常检测。理论分析显示网络的划分策略对网络整体的检测准确率和效率都有重要影响。而目前分流中采用的平均划分则是最差的策略。进一步分析揭示一个具有模块化特征的子网会比不具有模块特征的子网有更好的检测效果。复杂网络相关实证研究发现，现实网络中普遍存在模块化特性。于是，本文的网络划分依据网络本身具有的模块特性进行，将网络划分为具有模块化特征的网络模块的集合，并设计模块检测特征，比较网络特征，可以避免不相关的干扰。在中国教育网西南节点采集的数据上的实验显示，中国教育网西南地区网络确实存在模块化结构。而基于网络模块结构的异常检测比基于网络的检测，检测效率和准确率都有显著提高。

2 网络检测率与划分策略关系分析

本节对网络整体的检测率与网络划分策略之间的关系做理论分析，作为选择划分策略的依据。需

2010-03-09 收到，2010-06-10 改回
国家自然科学基金(60903157)和国家信息安全计划(2006C27)资助课题
*通信作者：王娟 wangjuan@uestc.edu.cn

要强调的是虽然我们将网络划分为一个个子网分别进行检测，但是最终的目的还是要获得比较高的网络整体检测率，对于单独子网的检测率的高低并无特别要求。

2.1 子网检测指标

检测离不开检测指标。子网的检测指标与网络指标类似，也是单位时间某网络特性的度量，例如：总数据包数，总流量数。但是，由于子网仅仅是网络的一部分，因此它的检测指标有自己的特点：具有地域性和方向性。地域性指是子网内部指标(标记为 f^i)还是子网外部指标(f^e)；进一步而言，外部指标是流入子网的还是流出子网的，这就是方向性。这种对指标的定义有利于避免无关子网活动和子网内部外部活动的互相干扰。

2.2 检测率与划分策略关系建模

设网络总设备数为常数 C，某种度量的网络值在特定时刻为常数 F，一般来说 F 正比于 C。设网络被划分为 n 个子网，C_i 为子网 i 包含的设备数，f_i 为子网 i 的度量，f_i 与 C_i 成正比。则有

$$f_1 + f_2 + \cdots + f_n = F \tag{1}$$

$$C_1 + C_2 + \cdots + C_n = C \tag{2}$$

设攻击度量为 f^t，异常检测的基本假设是异常值与正常值是有区别的，区别越大则检测效果越好。因此定义某子网 i 的攻击影响 ∂_i 为 $\partial_i = f^t/f_i$。该值越大攻击越明显，越容易被检测到。对特定强度的攻击，检测的效果取决于 f_i，即只与 $1/f_i$ 有关。而网络的检测效果是每个模块的检测结果的综合，可以表达为

$$f^t\left(\frac{1}{f_1}+\frac{1}{f_1}+\cdots+\frac{1}{f_n}\right) \tag{3}$$

如前所述我们的最终目的是为了使得整个网络的检测率提高。即在攻击强度一定且满足式(1)的条件下，使得式(4)的值尽可能地大。

$$\frac{1}{f_1}+\frac{1}{f_2}+\cdots+\frac{1}{f_n} \tag{4}$$

而该式实际上取决于各个 f_i 的大小，而 f_i 又正比于子网所包含的设备数 C_i，即与整个网络的划分策略有关。可以看到网络划分策略对网络检测率有决定性影响。

2.3 划分策略选择

实践中由于对每个子网都要进行特征提取，异常检测，因此时空消耗基本正比于子网的个数。根据时空资源，设定监控的子网个数为 n。下面分析子网数确定条件下的划分策略选择。首先，平均划分是最坏的划分策略。以下证明显示当n一定时，平均划分使得式(4)达到最小值，即网络整体的检测率最低。设 $\overline{\alpha} = (\sqrt{f_1}, \sqrt{f_2}, \cdots, \sqrt{f_n})$，$\overline{\beta} = (1/\sqrt{f_1}, 1/\sqrt{f_2}, \cdots, 1/\sqrt{f_n})$，则根据柯西不等式有

$$\left|\overline{\alpha}\cdot\overline{\beta}\right| = 1+1+\cdots+1 = n \le |\overline{\alpha}|\left|\overline{\beta}\right| = \sqrt{f_1+f_2+\cdots+f_n}\sqrt{\frac{1}{f_1}+\frac{1}{f_2}+\cdots+\frac{1}{f_n}}$$

得到 $\frac{1}{f_1}+\frac{1}{f_2}+\cdots+\frac{1}{f_n} \ge \frac{n^2}{F}$，等号成立的条件为 $\overline{\alpha} = \lambda\overline{\beta}$，即 $f_i = \lambda$，结合式(1)，有 $\lambda = F/n$，即 $f_1 = f_2 = \cdots = f_n = F/n$，$\sum\frac{1}{f_i}$取得最小值。度量与所含设备数成正比，因此平均划分网络模块是最不理想的划分方式。不均匀划分势在必行。

对于不均匀划分后的子网，在用子网指标检测时，以下分析揭示具有模块性质的子网会比不具有模块性质的子网有更好的检测效果。如前所述，子网的检测指标分为内部指标(f^i)和外部指标(f^e)，则子网的检测率由这两部分之和构成：

$$\partial_i = \frac{f^t}{f_i} = f^t\left(\frac{1}{f_i^i}+\frac{1}{f_i^e}\right) \tag{5}$$

同上面的分析过程，当 $f^i = f^e$ 时式(5)到达最小值。于是我们希望划分能保证每个子网的 $f^i > f^e$ (如果是 $f^i < f^e$，则将定义的内部和外部对调，得到同样结果)，这个恰恰是复杂网络中关于模块化特性的定义，即一个网络模块内部节点之间活动紧密而与外部联系较为松散，其模块化程度定义为 $f^i/f^e > 1$。大量复杂网络的实证研究表明，现实网络无论是社会网络还是其他类型网络也呈现明显的模块化结构[7−11]。因此本文选择复杂网络的模块发现方法作为网络划分策略，并给基于这种划分的检测命名为基于网络模块化结构的异常检测。

3　基于网络模块化结构的异常检测

总体来说基于网络模块化结构的异常检测分为离线生成模块部分和在线检测部分。

3.1 离线部分 — 生成网络模块

首先，目前大规模网络的异常检测基本是基于Netflow[12]数据。一条 Netflow 记录一个 OD (Origin-Destination) 流的字节数，数据包数，协议等信息。但是在将整个网络抽象成点(具有独立 IP 的网络设备)和边(设备间的流)来发现模块时只需要其中的源、目的 IP 对信息作为边的集合输入后面的模块发现算法：{<源 IP，目的 IP，时间戳>}。

其次，在实施检测之前需要一段时间的网络历史数据来生成检测中需要的模块结构—相关 IP 地址的集合。这个时间应该说越长越有利于发现稳定

的模块。但是由于大规模网络生成的流量记录是海量的，保存长期数据的代价比较大。这里，我们推荐的最少历史时间是 24 h。原因在于根据网络异常检测的已有经验显示，入侵与攻击都偏向长期在线的 IP。这很容易理解，长期在线 IP 多为服务器，时而在线时而不在线的基本是个人主机。前者是攻击的主要目标，也是异常监控的重要对象。所以本文模块发现的对象是长期在线 IP。根据采集的真实数据发现 24 h 的数据足以发现 99.99%以上的长期在线 IP。因为 24 h 开机的个人用户仅占极少数。

最后，依据长期在线 IP 集合提取出长期在线 IP 对作为下面模块发现的数据，即
{<源 IP，目的 IP >}，源 IP，目的 IP 都属于长期在线 IP 集合。

复杂网络的研究结果中，已经有很多经典的网络模块发现方法。本文根据网络异常检测领域的具体情况，并参考文献[13]，其中对现有方法效率和适用范围的比较，选择了CPM(Clique Percolation Method)作为模块发现算法。该算法是第一个能够处理重叠网络簇结构的模块发现算法。在实际网络中某个IP属于两个以上模块的情况是很可能存在的，特别是某些核心服务器。加之CPM相比其他算法，其模块发现精度和效率都较高适宜处理中、大规模网络。

3.2 在线异常检测部分

在获得模块结构后就可以在线检测异常，流程如图 1。

(1)分流　获得模块的 IP 集合后，在进行在线检测时首先对单位时间收集的数据进行基于模块结构的分流，即依据模块结构将网络流量分为与模块相关的模块流量。进一步，如果流的源目的地址都在某模块中则被划分为该模块内部流量；若只有源或目的地址在模块中则被划分为外部流量。

(2)特征提取　分别从模块内部流量和模块外部流量中，提取出模块内部特征和模块外部特征。分别用于检测模块内部和模块与外部之间的异常。这里需要说明的是，由于监控对象是模块因此模块外部特征就有方向性即该特征是流入模块还是流出模块。这是基于模块结构的检测区别于网络检测的特性之一。这些特征中有经典的统计特征如“总数据包数目”，也有目前的研究热点“熵”。不过最终都汇聚成特征时间序列。

(3)异常检测　本文的检测是指对特征值构成的时间序列的分析，即输入是：$f_{t1},\cdots,f_{ti}$ (ti 是时刻)，如果有攻击发生，造成特征值变动 Δf ，则分析就是要找出 Δf 产生的具体时间和来源。

(4)检测结果整合　把各个模块内外检测的结果整合为网络的异常检测结果。整合操作是“或(|)”运算。即设有 n 个模块，t 时刻模块 C_i 的检测结果为 R_i (0：正常，1：异常)，则网络异常检验结果为 $R_1 \mid R_2 \mid \cdots \mid R_n$ ，即只要有一个模块检测出异常，则认为网络中有异常。

4 实验和分析

本文采用的数据是从中国教育网西南节点收集的从 2008-11-17 到 2008-11-23 之间的 Netflow 记录。这是一个中型的主干网络，链接了中国西南地区的所有大学和实验室。其中心路由器 5 min 产生约 100 万条 Netflow 记录。包含不同 IP 数约 200 万个。采用的实验工具是 CFinder[14]和 Matlab7.0。

(1)发现长期在线 IP 738 个，仅占总 IP 数的 0.038%。

(2)根据长期在线 IP 集提取相应 IP 对记录作为 CPM 算法的输入，发现如图 2 所示模块。

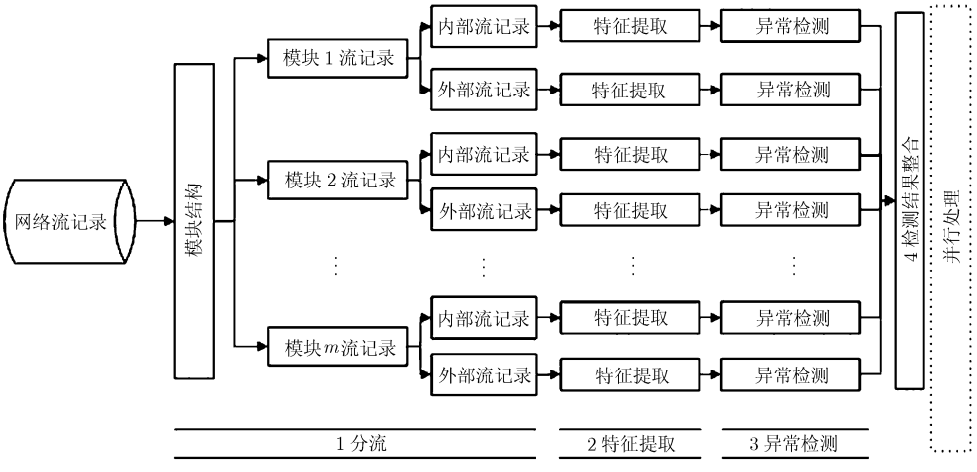

图 1 基于网络模块化结构的异常检测在线检测部分

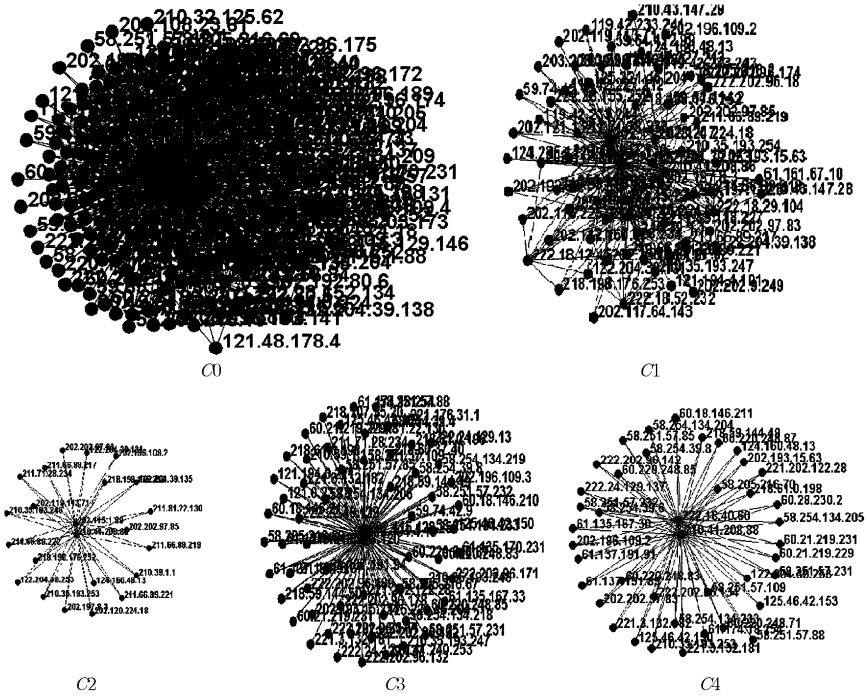

图 2 从中国教育科研网西南地区节点数据中生成的模块

模块特征总结如表 1。

表 1 模块性质汇总

	包含节点数	内部边数	外部边数	模块度
模块 C0	197	635	104	6.12
模块 C1	69	226	54	4.19
模块 C2	23	43	33	1.3
模块 C3	64	125	84	1.49
模块 C4	41	79	65	1.22

可见中国教育网西南地区节点之间确实存在模块结构。各个模块的模块度也远远大于 1。可以用模块作为检测对象。

我们选择了两种异常检测方法，一种是经典的基于正态分布的异常检测[4]; 另一种是当前比较流行的基于熵分析的异常检测[3]。用这两种方法分别以网络为对象和以模块为对象来考察基于模块检测的优缺点。检测结果的对比如表 2 和表 3。

可以看到各个模块的检测率都比较高。综合各个模块检测结果形成的网络检测结果也比以网络为对像的检测结果要高得多。综合方法如 3.2 节所述，是将各个模块的检测结果进行的一个或运算，运算结果作为基于模块的网络异常检测结果，其检测率如表 3 所示是 90.9%，检测到了绝大部分异常。而以网络为监控对象进行检测的检测率仅仅为 13.6%。相当多的异常被掩盖了。如前所述是由于基于模块的检测摈弃了无关的网络活动，而网络检测率低则是各个模块活动互相干扰的结果。

表 2 基于正态分布的模块和网络检测率对比(%)

	检测率	误报率	漏报率
模块 0	82.89	11.84	18.18
模块 1	100	8.6	0
模块 2	100	9.68	0
模块 3	100	10.75	0
模块 4	88.89	5.38	11.1
模块综合	**95.46**	26.89	**4.55**
网络	27.27	6.45	72.73

表 3 基于熵分析的模块和网络检测率对比(%)

	检测率	误报率	漏报率
模块 0	72.3	4.3	27.3
模块 1	100	12.9	0
模块 2	66.7	9.68	33.3
模块 3	100	9.7	0
模块 4	66.7	6.5	33.33
模块综合	**90.9**	26.88	9.1
网络	13.6	7.5	86.4

表 4 比较了不同划分策略对检测率的影响。表中的检测率同表 2 和表 3 中的模块综合检测率。平

表 4 不同划分策略的检测率对比(%)

	检测率	误报率	漏报率
平均划分	77.27	30.11	22.73
任意不均匀划分	84.76	28.34	15.24
基于模块划分	**95.46**	26.89	**4.55**

均划分是将网络节点平均划分为 5 个子网。任意不均匀划分是按照表 1 模块所包含的节点数划分子网，但是节点选择是随意的使得子网没有模块性质。检测方法就是表 2 对应的正态分布。可以看到平均划分的检测率最低，基于模块的划分无论从准确率还是误报/漏报率来说都比另外两种划分方法拥有优势。任意不均匀划分检测率居中，但是划分方法随意性太强，在实际应用中没有指导意义。因此再次证明具有复杂网络理论依据和最好实验效果的模块划分是一个理想的选择。

最后，以往分析单位时间(5 min)近百万条 Netflow 记录需要 6.5 min 的时间，现在在同样单机设备同样数据库条件下，由于分流和并行处理方法的采用，仅仅需要约 25 s，效率提高了 93.6%。

5 结束语

本文从实际应用的角度力图提高对大规模网络检测的效率和准确率。面对海量数据和网络检测率较低的问题，将大规模的网络划分为较小规模网络的集合分别检测是经典分流方法的一个自然启发。通过对划分策略和网络检测率的建模理论分析显示划分应该是不均匀的，具有模块性的划分对检测率有利。加上复杂网络研究揭示的现实网络本身具有模块性的结果，决定了划分基于网络模块结构。在此基础上设计了一个基于网络模块化结构的异常检测方法。主要包括了离线的模块结构生成部分和在线异常检测部分。由于缩小监控范围和集中监控对象，屏蔽了无关活动的干扰，实验表明基于模块的检测比基于网络的检测能提供更加准确的检测结果和更高的分析效率。

参 考 文 献

[1] Paul B, Jeffery K, David P, and Amos R. A signal analysis of network traffic Anomalies[C]. Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement, Marseille, France, 2002: 71–82.

[2] Anukool L, Mark C, and Christophe D. Mining anomalies using traffic feature distributions [J]. *SIGCOMM Computer Communications Review*, 2005, 35(4): 217–228.

[3] George N, Vyas S, David G A, Hyong K, and Hui Z. An empirical evaluation of entropy-based traffic anomaly detection[C]. Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, New York, NY, USA 2008: 151–156.

[4] Soo K S and Reddy A L N. An Evaluation of the Effectiveness of Measurement-Based Anomaly Detection Techniques[C]. Proceedings of the 26th IEEE International Conference Workshops on Distributed Computing Systems, Lisboa, Portugal, 2006: 1–6.

[5] Xin L, Fang N, and Crovella M, *et al*.. Detection and identification of network anomalies using sketch subspaces[C]. Proceedings of the 6th ACM SIGCOMM conference on Internet measurement.,Rio de Janeriro, Brazil, 2006: 147–152.

[6] Lin L Z, Min H G, and Miao Y X, *et al*.. Detecting distributed network traffic anomaly with network-wide correlation analysis [J]. *EURASIP Journal on Advances in Signal Processing*, 2009, doi: 10.1155/2009/752818.

[7] Newman M E J. The structure and function of complex networks. *SIAM Review*, 2003, 45(2): 167–256.

[8] 马卫东, 李幼平, 马建国, 周明天. 面向Web网页的区域用户行为实证研究. 计算机学报, 2008, 31(6): 960–967.
Ma Wei-dong, Li You-ping, Ma Jian-guo, and Zhou Ming-tian. Empirical study of region user behaviors for web pages[J]. *Chinese Journal of Computers*, 2008, 31(6): 960–967.

[9] 王科, 胡海波, 汪小帆. 中国高校电子邮件网络实证研究. 复杂系统与复杂性科学, 2008, 5(4): 66–74.
Wang Ke, Hu Hai-bo, and Wang Xiao-fan. Empirical analysis of an email network in a Chinese university[J]. *Complex Systems and Complexity Science*, 2008, 5(4): 66–74.

[10] 徐玲, 胡海波, 汪小帆. 一个中国科学家合作网的实证分析. 复杂系统与复杂性科学, 2009, 6(1): 20–28.
Xu Ling, Hu Hai-bo, and Wang Xiao-fan. Empirical analysis of a China scientists collaboration network[J]. *Complex Systems and Complexity Science*, 2009, 6(1): 20–28.

[11] 刘建香. 复杂网络及其在国内研究进展的综述. 系统科学学报, 2009, 17(4): 31–37.
Liu Jian-xiang. Complex network and review of domestic research[J]. *Journal of Systems Science*, 2009, 17(4): 31–37.

[12] Cisco NetFlow, http://www.cisco.com/en/US/tech/tk812/tsd_technology_support_protocol_home.html, 2005.

[13] 杨博, 刘大有, Liu Jiming等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54–66.
Yang Bo, Liu Da-you, and Liu J, *et al*.. Complex network clustering algorithms [J]. *Journal of Software*, 2009, 20(1): 54–66.

[14] Gergely P, Imre D, Illés F, and Tamás V, *et al*.. CFineder, http://angel.elte.hu/cfinder/, 2009, 9.

王 娟： 女，1981年生，博士生，研究方向为网络安全.

秦志光： 男，1956年生，教授，博士生导师，研究方向为密码学、信息安全.

刘 峤： 男，1974年生，博士生，研究方向为机器学习.

钱伟中： 男，1976年生，讲师，研究方向为网络安全.