面向上下文图形可视化挖掘企业网络行为①

陶维成

(南京航空航天大学 计算机科学与技术学院，南京 210026)
(芜湖职业技术学院 信息工程系，芜湖 241006)

摘 要：在企业分布式网络系统中，精确、动态地发现用户的网络行为日益具有挑战性。目前网络管理系统主要依赖于对用户身份的推理来进行网络行为管理，此类方法由于所收集到的数据缩放比粗糙，从而在大规模的网络环境下不能精确地对网络行为进行挖掘、发现和管理。对主机、用户、应用程序等网络上下文内容进行可视化挖掘，为动态地进行网络管理提供了重要帮助。

关键词：网络行为；可视化；上下文图形；可视化挖掘；距离方差

Finding Enterprise Network Behavior Through the Context of Graph Visual Mining

TAO Wei-Cheng

(School of Information Science and Technology, University of Aeronautics and Astronautics, Nanjing 210026, China)
(Department of Information Engineering, Wuhu Institute of Technology, Wuhu 241006, China)

Abstract: In the enterprise distributed network system, it is becoming increasingly challenging to accurately and dynamically identify who is doing what. The current network management system mainly relies on the identity of the users reasoning. This method cannot precisely mine, find and manage the network behavior in the large scale of the network because of the rough scale of collected data. This paper discovers the content and the dynamic process for network management of host, users and applications, which provides an important help to network context visual mining.

Key words: network behavior; visualization; context graph; visual mining; distance variance

在企业分布式大规模网络系统中，网络行为管理非常困难。如何精确跟踪网络运行状况，以及主机、用户、应用程序及文件等使用情况？目前主要是基于静态的管理方法，如分析 IP、端口、主机等网络日志信息，难以实时、精确、动态地知道用户正在做什么[1]。本文提出了一个动态的，可视化的方法，知道用户何时何地使用网络在做什么。为此，应实现如下目标：

* 何人，何事，何时，何处（4W）：知道网络上正在发生什么事情，也就是何人(who)在何处(where)何时(when)正在运行何种程序(what)，将与其连接相关的上下文信息记录下来。

* 智能化：将可视化与数据挖掘结合起来进行日常网络监控和管理实务。如构建决策树来对网络事件分类，或通过用户/应用程序的相似行为进行集群计算，识别潜在的问题。

1 网络上下文图形

本文将网络上下文定义成详细的网络连接信息(协议，源/目标 IP/端口)，时间，用户，应用程序，应用程序内容摘要，及其相关的网络文件访问。

用常见的网络命令或操作系统提供的网络工具收集网络上下文内容数据，如 netstat、PS、lsof 等。一个应用视图存在三种形式，本地用户(本地目录或用户路径)，本地机器，以及企业服务，形成网络连接上下文信息，如图 1 所示。

2 在上下文图形中可视化挖掘

将网络上下文进行可视化和探索对于网络监控和

① 收稿时间:2011-11-02;收到修改稿时间:2011-12-27

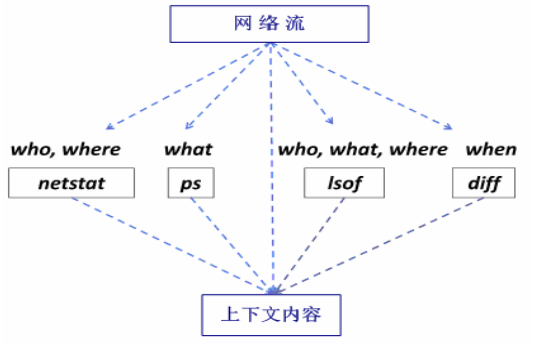

图 1 4W 网络上下文数据流

管理是一种有效的途径。可视化和分析必须有一个开始点[2]。纯自动计算分析不能替代可视化，因为通过传统自动数据挖掘方法可能错过一些特定的模式或知识[3]。理想情况下，在可视化过程期间，一些统计图形和数据挖掘或机器学习算法能增强管理人员的领域知识，以及只有对那些重要的事情才进行可视地探究。

2.1 可视的变化性与网络图的不变性

为了帮助管理员/研究者理解他们的网络，由此对非正常活动进行探测。网络图形集合的决定因素是其变化性和不变性，用可视的最大公共子图(MCS)和最小的公共子图(MCP)表示。图的最大公共子图(MCS)定义为出现在所有超图中的最大子图。所有 n 个网络图的 MCS 在整个监控周期内具有不变性。即主机，用户和应用程序结点，以及它们之间总是出现连接的边。最大公共子图同构是一个最优化问题，即著名的 NP 难题[4]。然而，在一个企业网络集合中，由于每一个结点被唯一地确定(IP 地址，用户 ID 或进程二元路径)，实际上，许多 NP 难题能被有效地解决[5]。

网络图的最小公共子图(MCP)定义为包括所有图的最小图的子图。网络图在 MCP 和 MCS 之间具有不同的变化性，即：$VAR_n = MCP_n - MCS_n$。

MCS 和 MCP 在网络监控和管理中是重要的，当 MCPs 在网络中生长为最大可能活动时，MCSs 作为不变式蕴含强连接和在构成稳定的长期存活的网络结点中保持一致关系。然而，上述子图和超图表现出离散属性(即：0 表示不出现，或 1 表示出现)，用概率表示结点/边，最小公共子图的概率(MCPP)是一个扩展的 MCP 作为边权计算的概率，即：

$$W(u,v)=\frac{F(u,v)}{|G|},\forall u,v$$

式中，$F(u,v)$ 是 $edge(u,v)$ 的出现频率，$|G|$ 表示快照图数目。

结点/边出现概率的关系是 $P_{MCS}=1>P_{\{i,j\}}>P_i \wedge P_j>0$。构造 MCPP 的概率可用于预测网络链接和探测异常现象的可能性。例如，假设一个用户 U 对应用程序 A1 和主机 H1 的连接概率为 0.9，并且对应用程序 A2 和主机 H2 的概率是 0。任何在图 U，A1 和 H1 中边的丢失，或在 U，A2 和 H2 中突然出现新的连接，这些情况便值得怀疑且需要深入跟踪调查。

2.2 图的距离方差

对一个只有数十个结点组成的网络可能较容易进行可视化检查，当网络有数千个结点时， 进行手工可视地比较几乎不可能，如图 2 所示。用可改变的视图来测量每个来自于期望图的快照的不同（或相似)，如图 3 所示。为了实现这些，需建立三个图，即： MCS，MCP 和 MCPP，在所有超图的距离方差中生成一个统计图表。一般而言，方差得分越高，图就越“异常”。随着图的可能生长，MCS 越小，MCP 作为最大图，距离方差在它们之间有一相对位置。在图 3 中，我们发现 MCS 本质上是一个阈值为 1 的 MCCP，类似地，MCP 是一个趋近于 0 的无限小的 MCPP。当阈值设为 0.5 时，MCPP 的曲线更加平滑，在时间轴上的 19-22 和 26 处(图 3 中红色的高亮部分)更加精确地指出了需要进一步跟踪研究的网络图。

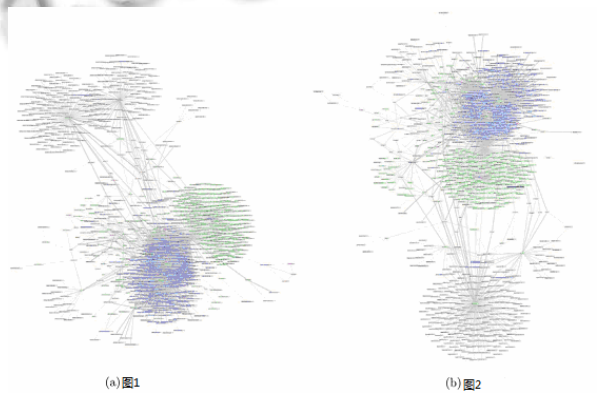

图 2 大规模网络连接图

为了从预期的图中计算距离，采用基于编辑距离的思想进行距离度量。在信息理论上，编辑距离是将

其中之一转换成另一个操作数。为了量化网络图的相似度，图编辑距离(GED)[5]建议用来测量拓扑的变化。图编辑距离的基本思想是修改图的相关代价以至于使它变成与其它图同构。通常有三种转换操作：插入，删除及置换。由于顶点标签置换不是有效的编辑操作，因为每个结点在企业网络中表示一个特定的主机，用户或应用程序，标签置换是一个基本的二步操作，如：移除旧的结点插入新的结点之后执行从一个图到另一个更大距离的置换。计算编辑距离的一个途径是从 g1 到 $MCS(g1,g2)$ 计算删除代价，以及加上从 $MCS(g1,g2)$ 到 g2 的插入代价。使用式（1）计算两个图之间的图编辑距离：

$$\frac{C_{del}^{v}*|V_{g_1}-V_{MCS(g_1,g_2)}|+C_{del}^{E}*|E_{g_1}-E_{MCS(g_1,g_2)}|+C_{add}^{V}*|V_{g_2}-V_{MCS(g_1,g_2)}|+C_{add}^{E}*|E_{g_2}-E_{MCS(g_1,g_2)}|}{C_{del}^{v}*V_{g_1}+C_{del}^{v}*V_{g_2}+C_{del}^{E}*E_{g_1}+C_{add}^{E}*E_{g_2}} \quad (1)$$

式(1)中 C_{del}^{v} 和 C_{del}^{E} 是相应顶点和边的删除代价，和是相应顶点 C_{add}^{V} 和 C_{add}^{E} 边的插入代价。如果所有的代价函数都是相同的，那么式（1）可简化成式（2）：

$$(g_1,g_2)=\frac{|g_1|+|g_2|-2|mcs(g_1,g_2)|}{|g_1|+|g_2|} \quad (2)$$

如果两个图完全一样，那么分子将为 0，其结果是距离为 0。另一方面，如果两个图不共享一个结点或边，那么结果是其中之一的距离值。

一旦对所有成对的图计算了距离矩阵，需要绘制和可视化图的相关位置。就可视化目的而言，维度通常仅有 2D 或 3D。多维度尺寸(MDS)[6,7]提出了将高维度数据通过映射它们到低维度空间进行可视化。通过映射网络图到 2D 空间对相关的位置进行可视化，每个结点表示一个具体日期内的一个网络图。期望图(EG)是一个 MCPP，其连接概率的阈值设为 0.5。在图 4 中，我们不但能看到来自预期网络图的距离（已高亮显示），而且所有网络图自身的距离变化也同样显示了出来。尽管 EG 位于所有图的中央，它被定位在更接近大多数图的右边， 那些异常点清楚地孤立在图的左边。

3　可视化过程的应用示例

应用上下文图形可视化挖掘企业网络行为，生成聚集视图，从而可视化地理解诸如 IE 用户和相关的 Web 流量，使用相同应用的相似集合，识别潜在的异常用户行为等。网络连接的本地上下文可扩展成主机、用户、应用程序及文件来识别正常的和蓄意的用户。

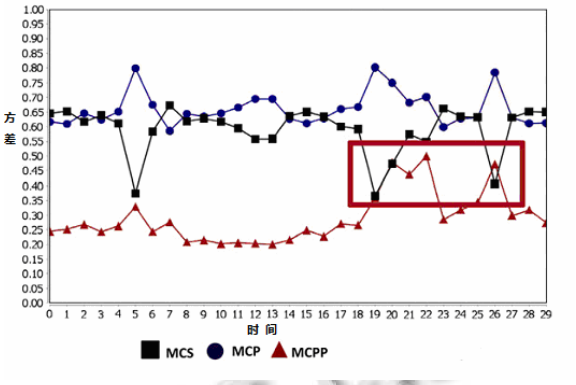

图 3　MCS、MCP 及 MCPP 的距离方差

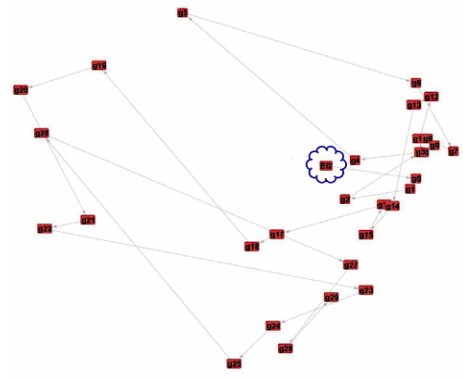

图 4　一个可变的 MDS 视图的演化和相关关系

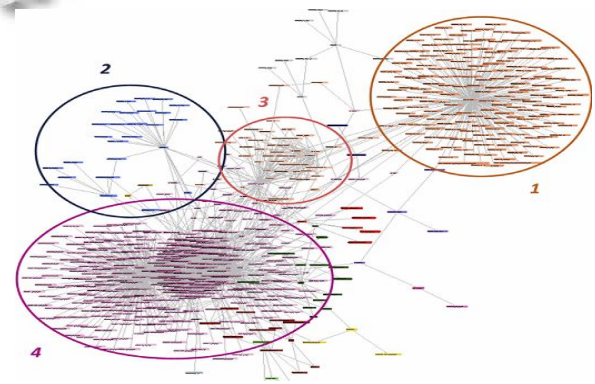

图 5　一个聚集视图的示例

在一个企业网络环境中，通过智能数据收集体获取主机、用户和应用程序结点数据。根据应用的相似

性，将具有相似应用的用户聚集成簇，并用不同颜色表示，如图 5 和表 1 所示。

表 1 一个聚集视图簇和用户数

聚集	用户数	描述
簇1	542	通过IE连接的所有外部域名
簇2	55	通过主机访问内部的Web服务
簇3	218	用户共享目录服务产生怀疑的集合
簇4	1334	本地用户形成的结构趋于合理的社区

4 结语

在上下文图形中进行可视化挖掘，识别何人在何时何处做何事，突破了网络行为管理的局限性，智能化使得日常网络行为管理变得轻松起来。通过对可视的变化性与网络图的不变性，图的距离方差的研究，为网络图的可视化奠定了理论基础。目前主要集中在结合可视化技术和图形挖掘来最大化地进行过程知识的获取和异常探测，下一步将进一步优化算法，提高响应速度和挖掘结果的精确性。

参考文献

1 陈友,程学旗,杨森.面向网络论坛的高质量主题发现.软件学报,2011,22(8):1785−1804.

2 李伟,罗军舟.面向网络管理知识获取的一种序列模式挖掘新算法.解放军理工大学学报(自然科学版),2008,19(5):445−449.

3 阳万安,胡其华.一种网络数据包的智能分类方法.科学技术与工程,2009,9(12):3301−3305.

4 Garey MR, Johnson DS, Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Company, 1979.

5 Bunke H, Dickinson PJ, Kraetzl M, Wallis WD. A Graph-Theoretic Approach to Enterprise Network Dynamics. Progress in Computer Science and Applied Logic (PCS), vol. 24, BirkhSuser, Boston, 2007.

6 Cox TF, Cox M, Multidimensional Scaling, 2nd ed. Chapman & Hall/CRC, 2000.

7 Bunke H, Dickinson P, Humm A, Irniger C, Kraetzl M. Applied graph theory in computer vision and pattern recognition. Ch Graph Sequence Visualisation and its Application to Computer Network Monitoring and Abnormal Event Detection, vol. 52, Springer, Berlin, Heidelberg, 2007.

(上接第 113 页)

参考文献

1 3GPP. TS 26.346 V9.4.0_2010, Multimedia broadcast /multi-cast service (MBMS); protocols and codecs.

2 3GPP. TS 26.234 V9.3.0_2010, Transparent end-to-end packet-switched streaming service (PSS); protocols and codecs.

3 ITU-T.G.1030_2005, Estimating end-to-end performance in IP networks for data applications.

4 ITU-T.G.1080_2008, Quality of experience requirements for IPTV services.

5 ITU-T SG12.2009-2012 Work Program. ITU, 2010.http://www.itu.int/ITU-T/workprog/wp_search.aspx?isn_sp=545&isn_sg=551.

6 ITU-T.P.862_2001, Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

7 赵川斌,张骥,任义.基于用户感知的网络优化体系建设探讨.移动通信,2011,6:17−22.

8 赵飞龙,梅杓春,余轮.移动通信网的 QoE 测量及其量化方法.电子测量与仪器学报,2010,24(3):230−236.