网络协同取证计算研究

张有东[1),2),3)] 曾庆凯[1),2)] 王建东[4)]

[1)](南京大学计算机软件新技术国家重点实验室 南京 210093)

[2)](南京大学计算机科学与技术系 南京 210093)

[3)](淮阴工学院计算机工程系 江苏 淮安 223003)

[4)](南京航空航天大学信息科学与技术学院 南京 210016)

摘 要 网络取证面临着复杂多样的网络入侵环境,尤其是对于复合攻击的取证,为此提出了网络协同取证计算新概念.通过对传统的函数依赖关系理论的扩展,提出了以一定概率相依赖的概率函数依赖关系及其分析方法与算法,进而结合贝叶斯网络理论、报警关联分析技术和对 K2 算法的改进,提出了一种网络协同取证分析算法,算法通过对元报警事件的聚类和综合不同的网络取证数据源,能够直观地再现复杂网络攻击的犯罪场景,有效地实现网络取证分析;而且即使在部分数据缺失情况下,算法也可推理攻击的发生过程.

关键词 协同取证;复合攻击;概率函数依赖;贝叶斯网络;犯罪场景

中图法分类号 TP393 **DOI 号**: 10.3724/SP.J.1016.2010.00504

Studies of Network Coordinative Forensics Computing

ZHANG You-Dong[1),2),3)] ZENG Qing-Kai[1),2)] WANG Jian-Dong[4)]

[1)](*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093)

[2)](*Department of Computer Science and Technology, Nanjing University, Nanjing* 210093)

[3)](*Department of Computer Engineering, Huaiyin Institute of Technology, Huaian, Jiangsu* 223003)

[4)](*Institute of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016)

Abstract The network forensic is faced with the question of the complex network intrusion analysis especially to forensic to the multi-step attack. So a new concept of coordinative forensics computing is defined. Through to extend the theory of function dependency, a new analysis method and algorithms called probability function dependency relations, which the dependency is in a probability, are proposed. Combined this method with the Bayesian network, alert correlation technology and the algorithm of K2 that is improved, the algorithm of network forensic computing is proposed also. The algorithm is able to synthesize the various forensic data resource and meta-alert events to reappearance the crime scenario of the complex network attack intuitively, and to realize the network forensic analysis effectively. Even if in the case of some data lost, the algorithm is able to infer the attacked process too.

Keywords coordinative forensics; multi-step attack; probability function dependency; Bayesian network; crime scenario

---

收稿日期:2007-07-12;最终修改稿收到日期:2009-03-11.本课题得到国家自然科学基金(60773170,60721002,90818022)、国家"八六三"高技术研究发展计划项目基金(2006AA01Z432)、高等学校博士学科点专项科研基金(200802840002)、江苏省高校自然科学研究项目(06KJD520019)资助.**张有东**,男,1967 年生,博士,教授,主要研究领域为信息安全、数据挖掘.E-mail: z.yd@163.com. **曾庆凯**,男,1963 年生,博士,教授,博士生导师,主要研究领域为信息安全、分布计算. **王建东**,男,1945 年生,教授,博士生导师,主要研究领域为人工智能、知识工程、机器学习和数据挖掘等.

1 引　言

对于复杂的网络入侵,其攻击行为往往是分步、多变或综合的,对其入侵行为的认定需要对不同网络安全设施获取的信息进行关联分析,才可能重构入侵者的入侵过程,获得其犯罪证据.事实上,对于复杂的案件,法律上更关注证据之间的关联性,对于计算机网络犯罪而言,有因果关系的、相互确证的多个独立证据之间的关联也有利于重建攻击过程.

定义 1.　协同取证(coordinative forensics)计算是指从目标系统所有可利用资源中发现、关联、解释、分析信息,以确定证据因果关系、再现网络犯罪场景的过程并形成证据链,从而支持法庭举证.

网络取证系统需要保存大量的不同来源的数据,这些保存的数据并不直接等同于举证证据,一般地可以认为是一种疑似证据,协同取证计算是一种事后取证分析,是通过对保存的疑似证据的分析,再现(重构)攻击的场景(或犯罪场景),从中发现真正的犯罪证据.

在入侵检测技术的研究中,已经提出了对攻击场景的构建问题及报警关联分析技术.本文从对报警关联分析技术研究着手,提出了一种基于贝叶斯网络(Bayesian Network,BN)的协同取证计算算法CFA,CFA 的核心思想是通过概率函数依赖关系分析,对 K2 算法的先验假设进行改进,并将无丢失数据处理的 K2 算法扩展到对不完整数据的处理,使之适用于网络电子证据的分析.CFA 可以综合多方数据源,实现犯罪场景(crime scenario)再现.

2 报警关联分析技术

2000 年第一次大规模的 DDoS 攻击引发了对报警关联(alert correlation)分析技术的研究,并受到越来越多的关注.报警关联指对入侵报警信息进行组合、解释和分析,以识别攻击和进行攻击场景重构.入侵报警信息可以是来自同一个 IDS 的,也可以是来自不同的 IDS 甚至异构的 IDS.

报警关联分析起初主要是对来自于同一个 IDS 产生的报警(消息),通过报警属性值之间的相似性对报警进行聚类,使得每个聚类的报警集合具有某些相同的特性[1].这种方法不能完全揭示相关报警之间的因果联系,无法进行入侵的意图识别和入侵行为预测.为此,研究者提出了复合攻击的概念,认为一些复杂的攻击是由若干单步攻击按照一定的逻辑关系、在特定的时间和空间形成的一个攻击序列,这里的每个单步攻击是指一个独立的、不可分割的攻击行为,而综合多个原始报警或其它元报警所产生的报警,称为超报警事件或元报警事件(Hyper Alert Event 或 Meta Alert Event)[2],超报警概念的提出有利于构造层次化的报警聚合与关联.

在此基础上,研究者又提出了入侵事件关联方法[3],将每个入侵检测传感器产生的、由于入侵行为而引发的报警称为一个事件.入侵事件关联是对单个或多个 IDS 的传感器产生的事件进行再组织和再分析以发现攻击行为的过程.报警事件关联扩大了 IDS 的检测范围,尤其是在大型的交换式网络系统中.

目前已经提出了许多报警关联算法,对于大型网络上的复杂的攻击,研究者提出了用图示的方法来表示一个攻击的过程从而再现攻击,一些研究论文称之为攻击场景(attack scenario)[4]或攻击图[5]的重构.但实际上这类算法主要是基于融合技术产生超报警,并构造超报警的关联图,而所谓的场景也就是指此关联图,它并不等同于网络取证所要求的原始的攻击现场.

文献[6]从网络取证的需求出发,提出了通过证据图进行网络取证分析的方法,该方法将整个系统分为攻击者(attacker)、受害者(victim)、跳板主机(stepping stone)和后台攻击者(background attacker)4 类角色,首先对报警事件进行时序聚类,然后将聚类结果用图示的方法表示,即表示为证据图,最后在专家干预下,在证据图上进行推理取证,这实际上是一种将攻击事件聚类后可视化的方法.

3 协同取证计算基本思想

如前所说,现代网络攻击大多数不是孤立的行为,往往是经过多个单步攻击后才能完成的一种复合攻击.如图 1 表示的一个实际攻击过程[7],该过程分为 5 个步骤:地址探测、端口扫描、获取口令文件、口令破解和登录系统.图中每个节点代表某种攻击方式,攻击的每一步都有多种方式可供选择,攻击者只要成功地使用其中一种即可实施该步攻击.对这种复合攻击的取证是对网络取证提出的新的挑战,现有网络取证技术还没有涉及对该问题的研究.

图 1 反映的是一个因果关系图,但是要发现这个因果关系图,需要在海量的信息中进行因果数据

挖掘.然而数据挖掘算法大都是对事物间的统计关联关系的挖掘,如关联规则挖掘、聚类、分类等,都没有涉及到事物之间的底层因果结构.

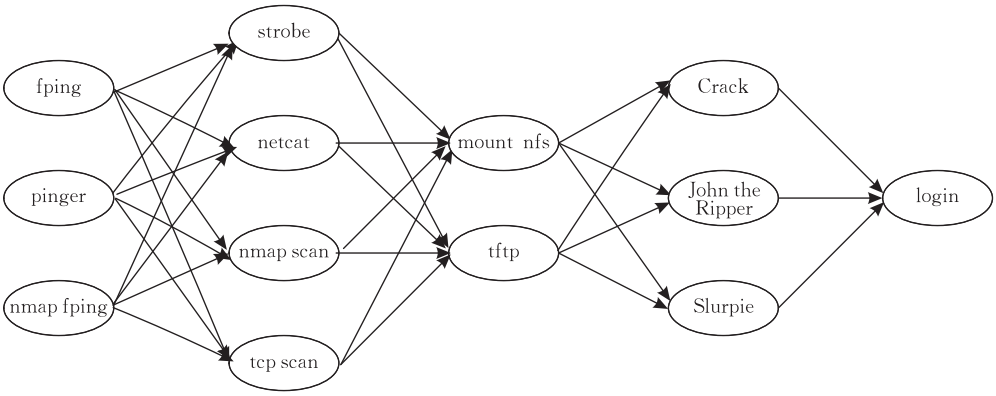

图 1 一个复合攻击实例

因果贝叶斯网络学习理论的发展为因果知识发现的研究提供了契机,基于因果贝叶斯网络学习理论的知识发现能够进一步获得对事物的因果本质的认识,是对统计数据的更高层次的因果知识的发现.分析图 1,它实际上是一个典型的包含隐藏变量的因果贝叶斯网络,在前 3 步攻击之间都可以认为是隐藏了隐藏变量 L_1、L_2 和 L_3(图略).

本文所提出的 CFA 算法是基于贝叶斯网络进行因果分析的协同取证计算方法,它通过对综合多个原始报警或其它元报警所产生的元报警事件的分析,生成元报警贝叶斯网络,从而反映更高层次的因果知识,实现犯罪场景的重构与事后取证.算法的基本步骤为

1. 数据预处理,对不同证据源的原始事件经过预处理转换为统一的标准报警数据格式 IDMEF,形成标准原始事件;

2. 对标准原始事件聚类形成元事件;

3. 用贝叶斯网络对元事件进行因果分析;

4. 重构攻击场景.

下面将分析 CFA 算法的理论基础与核心组成.

4 贝叶斯网络基本原理

4.1 贝叶斯网络模型

贝叶斯网络最早是由 Pearl 提出的,它模拟人的认知思维推理模式,用一组条件概率函数以有向无环图(directed acyclic graph)形式表示因果推理模型.它由两部分组成:(1)贝叶斯网络结构(G).G 的每个节点代表对象的一个属性变量,变量可以是离散的或连续的;G 的每条弧代表一个概率依赖关系.(2)每个变量对应的条件概率分布(CPD).当变量为离散变量时,常用条件概率表(CPT)来表示.

考虑离散变量的情况,设有 n 维训练数据集 Ξ,$V=\{V_1,V_2,\cdots,V_k\}$ 是贝叶斯网络中的节点,离散变量 V_i 的 v_i 个状态对应于 v_i 种可能的模型结构,用 $p(v_i)$ 表示其概率分布,变量 V_i 的 CPT 说明条件分布 $p(V_i|\Xi)$.对于每一种模型结构 v_i,存在一个连续向量值变量 θ_{vi},其中 θ_{vi} 值对应的是可能模型的真实参数.对于 θ_{vi},使用概率密度函数 $p(\theta_{vi}|v_i)$ 进行编码.

则对每个 v_i 和 θ_{vi} 使用贝叶斯规则计算其后验概率分布如下:

$$p(v_i|\Xi)=\frac{p(v_i)p(\Xi|v_i)}{\sum_{v_i'}p(v_i')p(\Xi|v_i')} \tag{1}$$

$$p(\theta_{v_i}|\Xi,v_i)=\frac{p(\theta_{v_i}|v_i)p(\Xi|\theta_{v_i},v_i)}{p(\Xi|v_i)} \tag{2}$$

定义 2. 如果一条弧由节点 V_i 到 V_j,则 V_i 是 V_j 的双亲(父节点),V_j 是 V_i 的后继.

假设 1. 条件独立性假设:给定双亲,假设贝叶斯网络的每个变量条件独立于图中的非后继,即图中的每个节点 V_i 条件独立于由 V_i 的双亲节点给定的 V_i 的非后代节点构成的任何节点子集.

设 $A(V_i)$ 是图中非 V_i 后代节点的任何节点集合,设 $Pa(V_i)$ 是图中 V_i 的直接双亲,则

$$p(v_i|A(V_i),\ Pa(V_i))=p(v_i|Pa(V_i)).$$

定义 3. 一个贝叶斯网络定义为一个三元组 (G,Ξ,P),这里 $G=(V,E)$ 是一个具有节点 $V=\{V_1,V_2,\cdots,V_k\}$ 和弧 E 的有向无环图,P 表示概率分布,Ξ 为实例空间.

应用链规则以及 Markov 条件,根据假设 1,得到贝叶斯网络中所有节点的联合概率如下:

$$p(V_1,V_2,\cdots,V_k)=\prod_{i=1}^{k}p(V_i\mid Pa(V_i)) \quad (3)$$

贝叶斯网络可表示事件的因果关系,其连接节点的弧表达了两个节点间的直接的因果影响,它也可被看作是拥有许多不同组合的一个抽象知识库.所谓因果贝叶斯网络就是指具有因果关系的贝叶斯网络,网络中的每个节点的父节点被解释为该节点相对于模型中其它节点的直接原因[8].

贝叶斯网络模型的假定避免了搜索贝叶斯网络结构的问题,尽管这个强限制假定不是很现实,然而大量的实验表明,即使在违背这种独立性假定的条件下,它仍能表现出很好的健壮性.

定义 4.　没有双亲节点的节点 V,概率不以其它节点为条件,$p(V)$称为该节点的先验概率.

为了计算给定因果贝叶斯网络的联合概率,需要知道先验概率和以双亲节点为条件的每个节点的条件概率函数.因此,一个随机变量集合的概率的一个完整说明涉及到这些变量的一个贝叶斯网络及网络中每个变量的 CPT.

4.2　贝叶斯网络学习

学习一个贝叶斯网络的问题是寻找一个网络,包括 DAG 结构和 DAG 中每个节点的概率分布或 CPT,它能最好地匹配一个数据训练集 Ξ,Ξ 是所有变量值的实例集合.

在学习和训练贝叶斯网络时,网络结构可以由领域专家预先给定,这种情况下贝叶斯网络的学习问题变得比较容易,否则需要由训练数据导出,理论上讲,训练得出的模型是实际贝叶斯网络的一个同构图,因此,训练本身是一个不断渐进的学习过程.另外,网络变量可能是可见的,也可能隐藏在所有或某些训练样本中,还可能是缺失的.根据统计学领域对多变量联合概率分布的近似分解方法,贝叶斯网络的学习可分为基于评价与搜索和基于独立性检验两大类算法[9].基于评价与搜索的典型算法是 Cooper 等提出的 K2 算法[10],K2 算法在给定节点顺序先验信息的情况下,利用贝叶斯概率作为标准来评价模型与数据的符合程度,通过不断向网络中增加边的贪婪搜索方法找到最佳网络结构.

5　CFA 核心算法分析

5.1　网络结构学习的 K2 算法[10]

设 Z 是数据库 D 的变量集,B_S表示仅包含在 Z 中的变量的任意一个贝叶斯网络结构,B_{S_i} 和 B_{S_j} 是两个包含 Z 中变量的贝叶斯网络结构.用 π_i表示 x_i 的双亲,用 ω_{ij} 指明在 π_i中变量的值的第 j 个唯一的实例(相对于 D 中案例的顺序),则有

$$P(B_S,D)=\int_{B_P}P(D\mid B_S,B_P)f(B_P\mid B_S)P(B_S)\mathrm{d}B_P \quad (4)$$

这里 B_P的值表示与贝叶斯网络结构 B_S相关联的条件概率分配,f 是给定 B_S后作用于 B_P的条件概率密度函数.

假设 2.　给定贝叶斯网络模型,数据库中的一条记录独立地发生.

根据此假设,有

$$P(B_S,D)=\int_{B_P}\left[\prod_{h=1}^{m}P(C_h\mid B_S,B_P)\right]f(B_P\mid B_S)P(B_S)\mathrm{d}B_P \quad (5)$$

这里的 m 是 D 中的案例数,C_h 是 D 中的第 h 个案例.

定理 1.　设 Z 为 n 个离散变量集,Z 中的一个变量 x_i有 r_i个可能的值$(v_{i1},\cdots,v_{ir_i})$.设 D 为具有 m 个案例的数据库,每个案例包含对 Z 中每个变量的一个赋值.设 B_S表示仅包含在 Z 中变量的一个贝叶斯网络结构,B_S中的每个变量 x_i有一个双亲用 π_i表示,用 ω_{ij} 指明在 π_i中相对于 D 的变量的值的第 j 个唯一的实例.假设 π_i中有 q_i个这样的唯一实例,定义 N_{ijk}为 D 中案例的数量,变量 x_i有值 v_{ik},并且 π_i被实例化为 ω_{ij},设

$$N_{ij}=\sum_{k=1}^{r_i}N_{ijk},$$

则

$$P(B_S,D)=P(B_S)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \quad (6)$$

K2 算法寻求最大 $P(B_S,D)$的网络结构,但作为节点数目的函数,可能的结构呈指数增长[11],显然,穷举所有可能的结构在大多数情况下是不可行的.为此,下文将提出一种基于概率函数依赖关系分析的方法,对变量集进行约简,然后通过启发式方法,搜索适合的贝叶斯网络结构.

5.2　概率函数依赖关系分析

通常,人们对数据集中属性的函数依赖关系会有一些先验的认识,函数依赖关系是数据库规范化理论的重要内容,已形成比较成熟的理论体系,应用规范化理论,可以极大地提高对数据库数据分析的效率.

下面我们将提出一种概率函数依赖的方法，相对于概率函数依赖而言，上述函数依赖可以认为是分布概率为1的概率函数依赖关系.

定义5. 设有关系模式$R(U)$，X和Y是属性变量集U的子集，概率函数依赖是形为$X \xrightarrow{P} Y$的一个命题，只要r是R的当前关系，对r中任意两个实例t和s，都有$t[X]=s[X]$以概率P蕴涵$t[Y]=s[Y]$，那么，概率函数依赖$X \xrightarrow{P} Y$在关系模式$R(U)$中成立.

其中，$t[X]$表示实例t在属性变量集X上的值；实例概念等同于数据库理论中的元组，另外，为叙说方便，前文中的变量在本节用属性变量表示.

定义6. 设F是概率函数依赖集，被F逻辑蕴涵的概率函数依赖全体构成的集合，称为概率函数依赖集F的闭包(clouser)，记为F^+.

$$F^+=\{X \xrightarrow{P} Y \mid F \mapsto X \xrightarrow{P} Y\}.$$

定理2. 如果$A_1,\cdots,A_n$是关系模式R的属性变量集合，那么$X \xrightarrow{P} A_1,\cdots,A_n$成立的充分必要条件是$X \xrightarrow{P} A_i\ (i=1,2,\cdots,n)$成立.

定义7. 设F是属性变量集U上的概率函数依赖集，X是U的子集，那么(相对于F)属性变量集X的闭包X^+表示从F集使用概率函数依赖推理规则推出的所有满足$X \xrightarrow{P} A_i$的属性变量A_i的集合：

$$X^+=\{A_i \mid X \xrightarrow{P} A_i \text{在} F^+ \text{中}\}.$$

算法1. 求属性变量集合X关于F的属性变量闭包X^+.

输入：关系模式R的全部属性变量集合U，在U上的概率函数依赖F，U的子集X

输出：关于F的属性变量闭包X^+

1. $i=0$，$X(i)=X$；

2. 在F中寻找尚未用过的左边是$X(i)$的子集的概率函数依赖：

$Y_j \to Z_j\ (j=1,2,\cdots,k)$，其中$Y_j \subset X(i)$；

在Z_j中寻找$X(i)$中未出现过的属性变量集合A，令$X(i+1)=X(i)A$，若无这样的A，则转步4.

3. 判断是否有$X(i+1)=X(i)$，若是，则转步4；否则转步2.

4. 输出$X(i)$，即为X^+.

算法中，对于步3的计算停止条件，有以下4种等价情况：

(1) $X(i+1)=X(i)$；

(2) 当发现$X(i)$包含了全部属性变量时；

(3) 在F中的概率函数依赖的右边属性中再也找不到$X(i)$中未出现过的属性变量；

(4) 在F中未用过的概率函数依赖的左边属性变量已没有$X(i)$的子集；

定义8. 如果关系模式$R(U)$上的两个概率函数依赖集F和G，有$F^+=G^+$，则称F和G是等价的概率函数依赖集，记作$F \equiv G$.

定义9. 设F是属性变量集U上的概率函数依赖集，如果$F_{\min}$是F的一个最小的概率函数依赖集，那么$F_{\min}$应满足下列4个条件：

(1) $F^+_{\min}=F^+$；

(2) 每个概率函数依赖的右边都是单属性变量；

(3) 每个概率函数依赖的左边没有冗余的属性变量(即F中不存在这样的概率函数依赖$X \xrightarrow{P_1} Y$，X有真子集W使得$F-\{X \xrightarrow{P_1} Y\} \cup \{W \xrightarrow{P_2} Y\}$与$F$等价).

(4) $F_{\min}$中没有冗余的概率函数依赖(即F中不存在这样的函数依赖$X \xrightarrow{P_1} Y$，使得F与$\{F-\{X \xrightarrow{P_1} Y\}\}$等价).

显然，每个概率函数依赖集F至少存在一个最小的概率依赖集$F_{\min}$，且$F \equiv F_{\min}$.

算法2. 求概率函数依赖集的最小的概率函数依赖集.

输入：一个概率函数依赖集F

输出：F的一个等价最小的概率依赖集$F_{\min}$

1. 应用分解规则，使F中每一个依赖的右部属性变量单一化.

2. 逐个检查F中左部是非单属性变量的依赖，去掉各依赖左部冗余的属性变量.

3. 去掉冗余的依赖. 即从第一个依赖开始，从F中去掉它(假设该依赖为$X \xrightarrow{P_1} Y$)，然后在剩下的依赖中求X^+，看X^+是否包含Y，若是，则去掉$X \xrightarrow{P_1} Y$；若不包含Y，则不能去掉$X \xrightarrow{P_1} Y$.

根据以上的分析，当已知数据库D的先验概率函数依赖关系F时，就可以根据算法1和2得到最小的概率函数依赖集F^+，这样，非集合中的属性变量不可能作为集合中属性变量的双亲节点，这就大大降低了贝叶斯网络学习的计算复杂性. 事实上，先验概率函数依赖关系在很多情况下是存在的，如通过数据挖掘方法得到的关联规则集合就蕴含了一个概率函数依赖关系集合.

5.3 启发式搜索方法

通过对 D 的概率函数依赖关系分析,能够对所有的 n 个变量指定一个顺序,这样,如果在顺序中 x_i 先于 x_j,则不允许结构中有一条弧从 x_j 到 x_i,非集合 F^+ 中的变量不可能是集合中变量的双亲节点.给定这样一个顺序作为约束,如果顺序变量包括了所有 n 个变量,则有 $2^{\binom{n}{2}}=2^{n(n-1)/2}$ 个可能的贝叶斯网络结构.当 n 较大时,应用等式(6)对 $2^{n(n-1)/2}$ 个中的每个可能的结构进行计算还是很难实现的,因此,除节点顺序外,假设在观察到数据 D 之前,B_S 等优先级,则可得到

$$P(B_S,D)=c\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \quad (7)$$

对于每个 B_S,这里 c 是先验概率 $P(B_S)$ 常量,最大化等式(7),只需要发现每个变量的双亲集合,最大化乘积的第 2 项,有

$$\max_{B_S}[P(B_S,D)]=$$

$$c\prod_{i=1}^{n}\max_{\pi_i}\left[\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}!\right] \quad (8)$$

这里在等式(8)右边,最大化发生在 x_i 的双亲 π_i 的每个实例.

为了最大化 $P(B_S,D)$,可用一种贪婪搜索(greedy-search)方法修改等式(8)右边的最大化操作,算法开始时假设节点没有双亲节点,每次加入对结构的概率增长最大的节点作为双亲,当增加单个双亲不能增加概率时,停止增加双亲到该节点.实现贪婪搜索使用的函数如下:

$$g(i,\pi_i)=\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \quad (9)$$

这里的 N_{ijk} 相对于 x_i 的双亲 π_i 和数据库 D 被计算,r_i、q_i 的定义同定理 1,则结合算法 1 和 2 以及 K2 算法得到的启发式搜索算法如算法 3 所示.

算法 3.　启发式搜索算法 K2_C.

输入:数据库 D,最小概率函数依赖集(算法 2 的结果)
　　(节点数 $l\leq n$)
输出:每个节点的双亲节点
Procedure K2_C;
For $i=1$ to l
　$\pi_i=\varnothing$;
　$P_{old}=f(i,\pi_i)$;　//该函数由等式(9)计算
　$OKToProceed=$true;
　While $OKToProceed$ and $|\pi_i|<u$ do
　　Let z be the node in $Pred(x_i)-\pi_i$ that
　　maximizes $f(i,\pi_i\cup\{Z\})$;
　　$P_{new}=g(i,\pi_i\cup\{Z\})$;
　　If $P_{new}>P_{old}$ Then
　　$P_{new}=P_{old}$;
　　$\pi_i=\pi_i\cup\{Z\}$;
　　Else $OKToProceed=$false;
　End While;
　Write (x_i,π_i);　//输出 x_i 及其双亲集合 π_i;
End For;

假设等式(9)的所有阶乘都已计算并存入一个数组,因为 N_{ij} 不可能大于 m,因此等式(9)中没有因子大于 $(m+r-1)!$,所以可以在 $O(m+r-1)$ 时间内把 $1\sim(m+r-1)$ 的整数阶乘计算并存储到数组中.算法中的函数 g 最多被调用 $l-1$ 次,因为 x_i 最多有 $l-1$ 个双亲节点,因此 P_{new} 最多需要 $O(murl)$ 的时间就可以完成所有的操作.While 循环中的其它语句是以 $O(l)$ 时间执行的,每次进行循环时,需要循环的次数是 $O(u)$ 次,For 语句需要循环 l 次.综合以上情况,算法 3 的时间复杂度为 $O(m+r-1)+O(marl)O(u)l=O(mu^2r^2l)$,在最坏的情况下,当 $u=n$,$l=n$ 时,其时间复杂度为 $O(mn^4r)$.

5.4 CPT 学习

5.4.1　无缺失数据学习

知道贝叶斯网络结构后,要得到某个节点 x_i 的 CPT,如果有充足的样本,只要计算每个节点和它的双亲的采样统计信息即可.

遵从前面的约定,设变量 x_i 有 r_i 个可能的值 $(v_{i1},\cdots,v_{ir_i})$,其双亲节点为 π_i,用 ω_{ij} 指明在 π_i 中相对于 D 的变量的值的第 j 个唯一的实例,采样统计结果表示为

$$\hat{p}(x_i=r_{ik}\,|\,\pi_i=w_{ijk})=\frac{x_i=r_{ik}\text{ 和 }\pi_i=w_{ij}\text{ 的采样统计数}}{\pi_i=w_{ij}\text{ 的采样统计数}} \quad (10)$$

5.4.2　不完整数据学习

当贝叶斯网络结构给定而某变量有空缺或不全时,根据具体情况,可运用贝叶斯分类的链规则技巧避开空缺项进行计算.一般地,我们引入梯度下降方法学习贝叶斯网络的 CPT[12].

设 D 是 n 个训练样本的集合 $\{C_1,C_2,\cdots,C_n\}$,p_{ijk} 是具有双亲 π_i 的变量 x_i 的 CPT 项,用 ω_{ij} 指明 π_i 的第 j 个唯一的实例,p_{ijk} 可以看作权,类似于神经网络中隐藏单元的权,权的集合总称为 p,这些权被初始化为随机概率值.梯度下降策略采用贪心爬山法,在每次迭代中,修改这些权,并最终收敛到一个局部最优解.

基于 p 的每个可能设置都等可能地假定，梯度下降方法能搜索最好地对数据建模的 p_{ijk} 值，其目标是通过按 $\ln P_p(D)$ 梯度来最大化 $P_p(D)=\prod_{d=1}^{n}P_p(C_d)$. 给定贝叶斯网络结构和 p_{ijk} 的初值，该算法表示如下.

算法 4. 梯度下降方法训练贝叶斯网络.

1. 计算梯度. 对于每个 i,j,k，计算：

$$\frac{\partial \ln P_p(D)}{\partial p_{ijk}}=\sum_{d=1}^{n}\frac{p(x_i=v_{ij},\pi_i=w_{ik}\mid C_d)}{w_{ijk}} \tag{11}$$

式(11)右端的概率要对 D 中的每个样本 C_d 计算，为方便起见，称此概率为 p^1. 当 x_i 和 π_i 表示的变量对某个 C_d 隐藏时，对应的概率 p^1 可以使用贝叶斯网络推理的标准算法(如一些商用软件包提供的标准功能)由样本的观察变量计算.

2. 沿梯度方向前进一小步：用

$$p_{ijk}\leftarrow p_{ijk}+(l)\frac{\partial \ln P_p(D)}{\partial p_{ijk}} \tag{12}$$

更新权值，其中 l 是表示步长的学习率，而 $\frac{\partial \ln P_p(D)}{\partial p_{ijk}}$ 由式(11)计算. 学习率设置为一个小常数.

3. 重新格式化权值：

由于权值 p_{ijk} 是概率值，它们必须在 0.0 和 1.0 之间，并且对于所有的 i,k，$\sum_j p_{ijk}$ 必须等于 1. 在权值被式(12)更新后，可以对它们进行归一化处理来保证这一条件.

6 实验与分析

6.1 元报警事件聚类

我们使用基于属性相似性和上下文要求的报警聚合来合并原始报警为元报警，原始报警的格式为(AlertID, Classificaton, SrcIP, DseIP, detectime, hyperID)，元报警的格式为(HyperID, Classification, SrcIP, DesIP, StartTime, EndTime, Count). 报警聚合过程将去除重复报警并产生保留重要信息粒度的易于分析的元报警. 每个元报警与原始报警之间具有一对多的关系，在元报警中，Count 域记录合并入元报警的原始报警的数目，原始报警中的 HyperID 域记录着它并入的元报警的唯一标识码.

元报警的聚合使用基于 Leader-Follower 模型的聚合算法[13]，其聚合准则是把具有相同源地址和目的地址的报警组合为相同类，其时间戳落入一个自扩展的时间窗口. 当原始报警的时间戳超出元报警的时间窗口时，元报警的时间窗口在预定义的界限 T 内自动扩展. 这意味着能够把连续重复的原始报警合并为一个具有适当 T 时间窗口的元报警.

实际上，影响聚合结果的一个重要因素是攻击类的划分，对于针对相同的系统弱点或具有相似的结果的攻击类型，可以在一个更高的层次上划分攻击类，例如：由 Snort 产生的“SCAN Nmap TCP”和“SCAN Nmap XMAS”报警可以合并入“SCAN Attack”类，当然，这种报警抽象及定义合适的攻击类需要人工干预.

6.2 事例分析

图 1 是一个每个单步攻击的方法可选的复合攻击，而有些攻击的攻击步骤也是可选的，例如图 2 所示的 IIS 攻击，攻击者启动攻击时首先对攻击主机进行 ping 操作，然后对其 80 端口进行扫描，第 3 步则有两个选择性操作，即 Scan Port139 或 Scan IIS Server，第 4 步也存在一个选择，既可在 Scan Port139 后直接进行攻击，也可回到步骤 3 后继续攻击. 也就是说第 4、5 步存在两个选择性攻击步骤，整个攻击过程可以 4 步完成，也可以 5 步完成，但这并不影响攻击的结果.

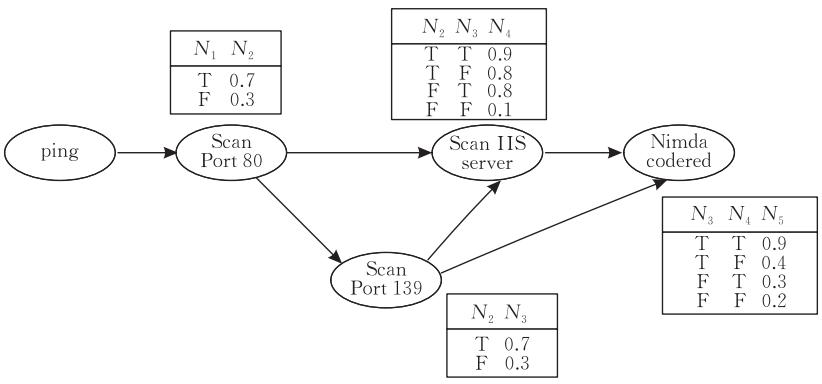

图 2 IIS Attack 的贝叶斯网络

图中的 CPT 是用 Snort 得到的对一个实际攻击数据学习的概率分布，CPT 中的 $N_1\sim N_5$ 对应攻击步，$P(N_1)$ 可以学习得到，也可根据经验直接设置为 0.5.

据此,就可以对捕获的网络数据和日志数据进行推理,分析得到所需的网络电子证据,另外,从上述过程也可以发现,即使在部分证据丢失的情况下,也可以很好地推理攻击过程的发生.

6.3　比较分析

我们用两个数据集对算法进行了测试,一个是 Cooper 使用的[10] ALARM 网络产生的 10000 个案例的数据集[14],用于与 K2 算法的对比实验.另一个是 DARPA2000 入侵场景关联评测数据集①,用于取证分析对比实验.测试用计算机为 DELL D620,系统环境为 Windows Vista.

ALARM 包含了 37 个网络节点和 46 条边,为了对比与验证算法的有效性,实验中训练集的大小分别使用了 100、500、1000、2000 和 3000 几种尺度,对每个数据集比较了边的增加、边的减少及网络结构计算时间 3 种性能,并以 ALARM 的结构为标准,分别用均值和标准差来度量,实验结果如表 1 所示.实验中未考虑联合概率分布计算情况的比较,显然这种比较很难选择一种标准来衡量.从实验结果可以看到,K2_C 算法的准确率和计算时间都有明显改善.但是,正如前面分析,计算时间的提高与先验有关,显然当先验的准确性较高时,计算时间将大大减少.

表 1　对比分析结果

算法	数据集大小	增加边结果		减少边结果		时间/s
		均值	标准差	均值	标准差	
K2	100	0.75	1.28	0.22	0.48	321.00
K2_C		0.19	0.40	0.62	0.86	130.00
K2	500	0.22	0.42	0.11	0.31	2213.00
K2_C		0.19	0.40	0.22	0.48	1077.00
K2	1000	0.11	0.31	0.03	0.16	6783.00
K2_C		0.24	0.49	0.22	0.48	4909.00
K2	2000	0.05	0.23	0.03	0.16	9147.00
K2_C		0.19	0.40	0.11	0.31	6658.00
K2	3000	0.00	0.00	0.03	0.16	18848.00
K2_C		0.16	0.37	0.05	0.23	11249.00

我们还用 DARPA 2000 数据集进行了取证计算实验.DARPA 2000 数据集是 DARPA 资助 MIT 林肯实验室构造的入侵场景关联评测数据集,已成为入侵报警关联算法、场景构建算法的有效性验证的标准数据集.

DARPA 数据集包括用 Tcpdump 分别在非军事区(DMZ)和内部网(Inside)中监听到的全部数据包.该数据集包括 LLDOS 1.0 和 LLDOS2.0.2 两个攻击场景实例.在 LLDOS1.0 攻击场景中,攻击者通过 Solaris sadmind 服务漏洞攻陷并控制了 Eyrie空军基地网络中的 3 台主机,上传了 Mstream 工具,并对一个政府网站发动了 DDoS 攻击.

LLDOS2.0.2 攻击场景与 LLDOS1.0 相似,所不同的是攻击者对漏洞主机的发现以及 Mstream 分布式拒绝服务攻击的上传使用了更加隐蔽的方法,从而能够验证关联算法在底层入侵检测系统存在漏报情况下的关联效果.

我们在 LLDOS1.0 数据集上进行了实验,首先使用 Leader-Follower 算法进行元报警事件聚类,得到元报警事件集合,因元报警事件较少,直接使用算法 3 进行贝叶斯网络结构学习,得到的结构如图 3 所示.

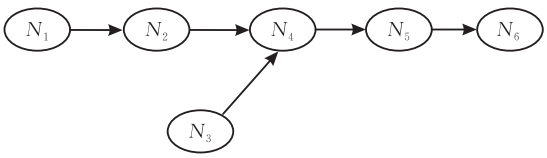

图 3　LLDOS1.0 的元报警贝叶斯网络

图中节点与元报警的对应关系如表 2 所示.

表 2　节点与元报警的对应关系

节点	元报警
N_1	Sadmind_Ping
N_2	Sadmind_Amslverify_Overflow
N_3	Email_Almail_Overflow
N_4	Rsh
N_5	Mstream_Zombie
N_6	Strem_DoS

事实上,LLDOS1.0 中包含了一个复合攻击的过程,其完整的攻击序列分为 5 个攻击阶段:(a)通过 IPSweep 进行主机探测;(b)使用 Sandmin_Ping 进行 Sadmind daemon 服务端口扫描,探测可能存在 Sadmind 漏洞的主机;(c)利用主机漏洞进行系统入侵,获得 3 台主机的 root 控制权限;(d)在被攻破的主机上安装可用于 DDoS 攻击的木马;(e)利用被控制主机发起 DDoS 攻击.图 3 的贝叶斯网络完全重构了这个攻击过程.但图中的 Email_Almail_Overflow 并不是一个单步攻击,分析发现仅与 Rsh 具有较高的概率函数依赖关系,如何去除这样的噪音节点是需要进一步研究的问题.

除了上述攻击场景重构功能之外,算法在对报警事件的检测方面也表现了较好的性能,我们对两个数据集进行检测并在相似环境下与 Ning[2] 的报

① MIT Lincoln Laboratory, 2000 DARPA Intrusion Detection Scenario Specific Data Sets. http://www.ll.mit.edu/IST/ideval/data/2000/2000_data_index.html

警关联算法进行了相同的实验比较，用 Real Secure Network Sensor 6.0 对数据集进行比对分析，实验结果显示在误警率和报准率方面的性能都有所提高，如表 3 所示.

表 3 入侵检测结果比较

数据集		攻击数	工具	报警数	检测攻击数	报准率/%	正确报警	误报率/%
LLDOS1.0	DMZ	89	Real Secure	891	51	57.30	57	93.60
		89	CFA	56	52	58.43	55	1.79
	Inside	60	Real Secure	922	37	61.67	44	95.23
		60	CFA	42	38	63.33	41	2.38
LLDS2.0.2	DMZ	7	Real Secure	425	4	57.14	6	98.59
		7	CFA	47	5	71.43	45	4.26
	Inside	15	Real Secure	489	12	80.00	16	96.73
		15	CFA	59	13	86.67	56	5.08

7 结束语

本文将报警关联技术、贝叶斯网络学习方法、概率函数依赖理论相结合，提出了一种能够对来自于不同数据源的数据进行网络协同取证计算的算法. 计算机取证是一项专业性很强的工作，通过贝叶斯网络可以直观地再现攻击场景，便于进行网络取证分析. 因此下一步的工作主要是实现自动生成可视化的贝叶斯网络图，用图示的形式直观地反映取证计算的结果，以便于广泛使用.

参 考 文 献

[1] Debarh H, Wespi A. Aggregation and correlation of intrusion detection alerts//Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection (RAID). Davis, CA, USA, 2001: 85-103

[2] Ning P, Cui Y, Reeves D S. Constructing attack scenarios through correlation of intrusion alerts//Proceedings of the 9th ACM Conference on Computer and Communications Security. Washington, DC, USA, 2002: 245-254

[3] Andersson D, Fong M, Valdes A. Heterogeneous sensor correlation: A case study of live traffic analysis//Proceedings of the 2002 IEEE Information Assurance Workshop. West Point, NY, USA, 2002: 1-12

[4] Dain O M, Cuningham R K. Building scenarios from a heterogeneous alert stream//Proceedings of the 2001 IEEE Workshop on Information Assurance and Security. West Point, NY, 2001: 231-235

[5] Ning P, Xu D B, Healey C G, Amant R S. Building attack scenarios through integration of complementary alert correlation methods//Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS). San Diego, CA, 2004: 97-111

[6] Wang W, Daniels T E. Building evidence graphs for network forensics analysis//Proceedings of the 21st Annual Computer Security Applications Conference (ACSAC'05). Tucson, AZ, 2005: 254-266

[7] Bao Xuhua, Dai Yingxia, Feng Pinghui, Zhu Pengfei, Wei Jun. A detection and forecast algorithm for multi-step attack based on intrusion intention. Journal of Software, 2005, 16(12): 2132-2138(in Chinese)
(鲍旭华，戴英侠，冯萍慧，朱鹏飞，魏军. 基于入侵意图的复合攻击检测和预测算法. 软件学报，2005，16(12)：2132-2138)

[8] Pearl J. Graphical models for probabilistic and causal reasoning//Tucker Allen B ed. Computer Science and Engineering Handbook. CRC Press, 1997: 697-714

[9] Cheng J, Greiner R, Kelly J, Bell D, Liu W. Learning Bayesian networks from data: An information theory based approach. Artificial Intelligence, 2002, 137(1-2): 43-90

[10] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 1992, 9(4): 309-347

[11] Robinson R W. Counting unlabeled acyclic digraphs//Little C H C ed. Combinatorial mathematics V. Lecture Notes in Mathematics 622. Australia: Springer-Verlag, 1997: 239-273

[12] Russell S, Binder J, Koller D, Kanazawa K. Local learning in probabilistic networks with hidden variables//Proceedings of the 14th Joint International Conference on Artificial Intelligence (IJCAI'95). Montreal, Canada, 1995, 2: 1146-1152

[13] Valdes A, Skinner K. Probabilistic alert correlation//Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection (RAID). Davis, CA, USA, 2001: 54-68

[14] Beiulich I A, Suermondt H J, Chavez R M, Cooper G E. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks//Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine. London, England, 1989, 38: 247-256

ZHANG You-Dong, born in 1967, Ph. D., professor. His main research interests include information security and data mining.

ZENG Qing-Kai, born in 1963, Ph. D., professor, Ph. D. supervisor. His main research interests include information security and distributed computing.

WANG Jian-Dong, born in 1945, professor, Ph. D. supervisor. His main research interests include artificial intelligence, knowledge engineering, machine learning and data mining.

Background

This work is supported by the National Natural Science Foundation of China under grant Nos. 60773170, 60721002, and 90818022, the National High Technology Research and Development Program (863 Program) of China under grant No. 2006AA01Z432, and Specialized Research Fund for the Doctoral Program of Higher Education of China under grant No. 200802840002. The project aims to construct the method and technology of the security system, and the authors' work for audit and forensics computer is one of the important services of the projects task.

The network forensic is faced with the question of the complex network intrusion analysis especially to forensic to the multi-step attack, but the current technology can not forensic to these kind of attack still.

The authors have done research on forensics computer analysis, especially for networks intrusion electric evidence analysis. Currently, they have made some progress in analyzing the massive or heterogeneity network dubious evidence information. Through to extend the theory of function dependency, a new analysis method and algorithms called probability function dependency relations are proposed. Combined these methods with the Bayesian network, alert correlation technology and the improved K2 algorithm, the algorithm of network forensic computing is proposed also. The algorithms introduced in this paper have been integrated into the Network Forensics Analysis System developed by the project team. The system can produce a meta-alert Bayesian network to indicate the high level knowledge and reappearance the crime scenario. The work introduced in this paper is very useful and practical in the network crime and security system research areas.