doi:10.11959/j.issn.1000-436x.2015273

基于主体行为的多方安全协议会话识别方法

朱玉娜，韩继红，袁霖，陈韩托，范钰丹
（解放军信息工程大学 三院，河南 郑州 450004）

摘 要：针对分布在多个相关流中的多方安全协议会话问题，提出了多方安全协议会话的 3 个启发式的主体行为特征——邻接主机行为、主体角色行为以及主机消息行为，给出了主体行为特征检测原理，提出了多方安全协议会话识别方法。针对 3 个典型的多方安全协议，分别在 3 种会话运行场景下进行实验，结果表明该方法识别率在 90%以上，误报率和漏报率在 6%以下，能够有效地识别协议会话。
关键词：安全协议；协议识别；会话识别；主体行为
中图分类号：TP393.08 **文献标识码**：A

Towards session identification using principal behavior for multi-party secure protocol

ZHU Yu-na, HAN Ji-hong, YUAN Lin, CHEN Han-tuo, FAN Yu-dan
(The Third College, PLA Information Engineering University, Zhengzhou 450004, China)

Abstract: Aiming at the problem of session identification for muti-party secure protocol, three characters were presented, i.e., neighboring-host-behavior(NHB), host-role-behavior(HRB) and principal-message-behavior (PMB), to explore the correlation among multiple flows employed in a same session. Then a session identification approach was proposed using these features. Finally, the approach was evaluated on three classical multi-party secure protocols in three scenes. The experimental results indicate the identification precision is above 90%, and the false negatives rate and false positives rate are below 6%.
Key words: security protocol; protocol identification; session identification; principal behavior

1 引言

网络协议会话识别有助于更细粒度、更精确地控制、管理流量，对入侵检测、流量监控、协议安全性在线分析、用户行为分析等网络安全关键应用都具有重要现实意义。随着密码技术的广泛应用，包含大量密文数据的安全协议被大量应用在互联网各种核心、关键服务中。与安全协议相关的各种数据在网络流量中比重日益增加。如何对安全协议会话进行有效识别已成为当前网络安全技术中亟待研究的关键技术之一。

网络流是指具有相同五元组（源 IP、源端口、目的 IP、目的端口、传输层协议）取值的报文序列[1]。假定主机 A 和主机 B 进行通信，将 A、B 通信过程中同一传输方向（$A\rightarrow B$ 或 $B\rightarrow A$）的报文序列定义为单向流，如图 1 所示。

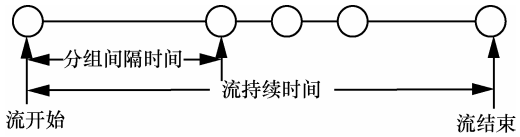


图 1 网络单向流

若 A 和 B 相互之间均有流量传输，将 A、B 之间不同传输方向（$A\rightarrow B$ 和 $B\rightarrow A$）的报文序列定义为双向流，如图 2 所示。

协议会话是指协议参与方之间的一次协议完整交互过程，包括协议一次通信建立和结束之间的

收稿日期：2015-03-16；**修回日期**：2015-05-11
基金项目：国家自然科学基金资助项目（61309018）
Foundation Item: The National Natural Science Foundation of China (61309018)

所有报文。按照参与方数目，可将安全协议划分为两方安全协议和多方安全协议。对两方安全协议而言，当主机 *A* 向主机 *B* 发起协议时，*A* 通过“IP 地址+端口号”，与 *B* 的协议应用程序进行通信。对 TCP 流，*A* 和 *B* 之间建立 TCP 连接；对 UDP 流，*A* 直接向 *B* 传送报文，建立 *A* 和 *B* 之间的通信信道，随后 *B* 对 *A* 在信道上反方向进行响应。因此通常情况下，双方协议一次会话过程包含在一个 TCP 连接或一个 UDP 双向流中，为 2 个源宿地址对调、方向不同的单向流。

图 2　网络双向流

对多方安全协议而言，一次会话过程分布在多个单向流中，例如 Kerberos 协议。由于安全协议本身具有“高并发性”的特点，在 Kerberos 协议实际运行中，还存在大量协议会话实例并发运行的情况。

现有协议识别方法主要是标识网络流量所使用的应用层协议。普通通信协议的报文为明文数据，在协议识别的基础上，通过分析报文数据，即可恢复协议会话。而安全协议则采用密码技术，协议报文包含大量密文数据，无法通过检测报文内容恢复会话。对两方安全协议而言，只需识别流所使用的应用层协议类型，即可分析包含在一个 TCP 连接或者 UDP 双向流中的协议会话过程。对分布在多个流的多方安全协议会话而言，现有方法不能确定哪些流的报文具有关联关系，无法恢复多方安全协议会话过程。例如，假定在某个时间窗口内，存在 *n* 次 Kerberos 并发会话，如图 3 所示。现有方法[2]可以识别某个流为 Kerberos 协议类型，但无法确定哪几个流中的报文属于同一次 Kerberos 协议会话。

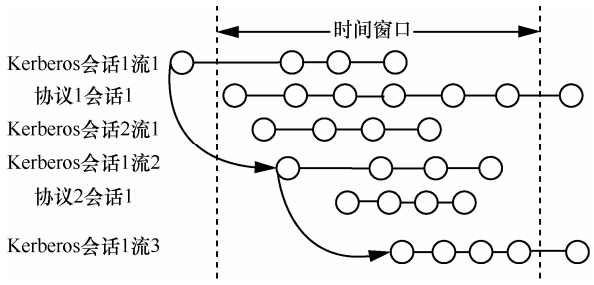


图 3　某个时间窗口内的网络流

针对这一问题，本文重点研究多方安全协议会话识别问题，提取属于同一次会话的流之间的关联特征，恢复协议会话过程。本文主要贡献有：1)给出了多方安全协议会话 3 个启发式的主体行为特征——邻接主机行为（NHB, neighboring-host-behavior）、主机角色行为（HRB, host-role-behavior）以及主体消息行为（PMB, principal-message-behavior）；2)基于会话主体行为特征给出相应的特征检测原理；3)提出了多方协议会话识别方法，并通过实验验证了该方法的有效性。

2　相关工作

协议识别是会话识别的基础，会话识别是协议识别的细化和延伸，两者的应用领域和应用场景各有侧重。对带宽管理、网络 QoS 等业务而言，没有必要精确跟踪和记录协议每次会话的各个状态，只需识别协议类型即可。而对入侵检测、协议安全性在线检测、用户行为分析等网络安全业务而言，对网络流量进一步识别到协议会话级别，可以更细粒度、更精确地控制、管理流量，具有重要现实意义。

对双方安全协议流量而言，协议会话包含在 2 个源宿地址对调、方向不同的单向流中，只需识别协议类型，即可分析协议会话过程。下面对现有安全协议识别方法进行介绍。

现有方法[3]主要包括基于端口的方法、基于 DPI 的方法和基于流统计特征的方法。随着动态分配端口技术的广泛使用，基于端口的协议识别方法不再适用。基于 DPI 的识别方法[4,5]主要利用协议在握手及密钥协商阶段进行明文通信的特性，通过载荷特征匹配进而识别安全协议类型，但该类方法无法处理协议载荷完全加密的情况。

为解决这一问题，人们提出了基于流统计特征的识别方法。Wright 等[6]对各种协议和加密技术进行分析，结果表明，对于特定的协议，分组大小、分组到达时间以及到达方向等统计特征不受加密影响。相同协议的网络流具有相似统计特征，因此依据网络报文流量统计特征可以识别协议。由于同一协议的流量特征并非严格一致，需要采用模糊判断的方法完成识别，因此该类识别方法大都借助机器学习方法。2005 年，Zuev[7]等首次提出了基于概率模型的朴素贝叶斯方法，但是其流识别率仅为 67%。Moore 等[8]在此基础上采用基于关联的快速过滤算法，识别效率显著提高，达到 90%以上。随

后，基于流统计特征的识别方法受到关注，成为研究热点[9~11]。但该类方法仅考虑单个流的协议类型识别，没有考虑流之间的关联性。Wang 等[12]针对分布式应用系统，使用分层机制和触发条件，识别与具体分布式应用程序相关的所有相关联流。该方法主要针对 P2P 协议，且假设一个主机仅运行一个应用程序，不适用于多方安全协议会话识别问题。

综上所述，由于现有方法没有充分考虑分布在多个相关流中的多方安全协议会话过程，多方安全协议会话识别问题仍有待解决。

3 多方安全协议会话识别问题描述

在协议执行过程中，协议参与方之间需要进行信息交互。记主机 *A* 和主机 *B* 为协议的 2 个参与方。*A* 向 *B* 发送信息，*B* 在接收到期望的数据后进行响应。本文将 *B* 进行响应前，*A* 向 *B* 发送数据的过程作为一个交互步骤，将该步骤中的所有数据作为一个消息。

一个消息可能包含在一个或多个数据分组中。对捕获的协议数据分组进行处理，参照 TCP/IP 体系结构，依次剥离数据分组头中的每一层首部内容，对报文载荷进行重组，获取消息数据。

安全协议会话采用下述四元组表示

(*protocol ID*, *STime*,{*ip*, *role*}, {*message*})

其中，*protocol ID* 为识别的安全协议 ID；*STime* 为协议会话开始时间；{*ip, role*}为参与协议会话的所有主机 IP 以及主机对应角色的集合，其中主机所担任的角色可以为发起者、响应者或者可信中心；{*message*}为该会话的消息集合，每一个消息采用三元组（发送者、接收者、消息序号）表示，其中发送者和接收者采用(*ip*, *port*)表示，消息序号表示该消息为协议会话中的第几步交互。

当前通信模式包括 Client/Server（C/S）模式、Browser/Server（B/S）模式和 Peer to Peer（P2P）模式等。其中，服务提供者（例如服务器）可以设置指定端口；用户（例如客户端）则大都由系统随机指定端口，协议不同会话中用户的端口是不同的。因此大多数情况下，一个流对应一次会话。少数情况下，用户也设置指定端口，一个流中可能存在多次会话。

本文重点研究多方安全协议一个流对应一次会话的情况，分析同一次会话的主体行为特征，并据此恢复协议会话过程。

4 会话主体行为特征分析

在协议实际执行中，由协议参与方、新鲜数以及消息中含有的其他基本信息唯一标识一次会话。安全协议采用密码技术，这些信息大都包含在密文中。由于无法对消息进行解密，不能获得被加密的内容，因此不能通过检测消息内容确定会话。

为此，本文从多方协议同一次会话的主体行为出发，给出 3 个启发式的行为特征——邻接主机行为（NHB）、主体角色行为（HRB）以及主体消息行为（PMB）。下面进行详细阐述。

4.1 邻接主机行为

对于多方安全协议而言，属于同一次协议会话的多个流之间通常存在相同的主机。为便于描述，给出下述定义。

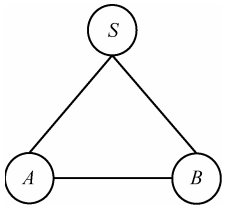
邻接主机流：假定流 f_1=(*SrcIP*$_1$, *SrcPort*$_1$, *DstIP*$_1$, *DstPort*$_1$, *transport protocol*)，流 f_2=(*SrcIP*$_2$, *SrcPort*$_2$, *DstIP*$_2$, *DstPort*$_2$, *transport protocol*)，如果 $\{SrcIP_1, DstIP_1\} \cap \{SrcIP_2, DstIP_2\} \neq \varnothing$，则称 f_1 和 f_2 为邻接主机流。

邻接主机行为特征：多方协议一次会话不存在孤立边。假定多方协议一次会话分布在 n 个流$\{f_1, f_2, \cdots, f_n\}$中，对任一个流，在会话中都至少存在一个流与其构成邻接主机流。

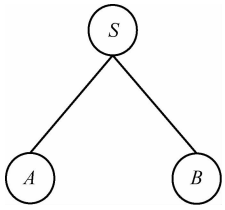
下面对该特征进行阐述。在一次协议会话中，对参与方之间的交互行为采用图的形式进行抽象。其中，参与方主机采用图的节点表示，参与方之间的流采用图的边表示。为便于描述，2 个相同主机之间的流用一条边表示。

在三方安全协议中，记协议参与方为 *A*、*B*、*S*。三方协议会话结构有 3 种，如图 4(a)~图 4(c)所示。由图可知，在三方安全协议的一次会话中，不同流之间都存在相同主机，任一流与其邻接主机流集合包含该会话的所有流。

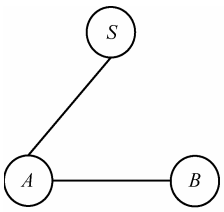
当协议参与方数目大于 3 时，每个参与方主机与一个或多个主机相连，且会话结构中不存在孤立边。下面采用反证法进行说明。假定协议中 A_i 和 A_j 构成孤立边，其会话结构如图 4(d)所示。协议需要完成某项任务，而 A_i 和 A_j 没有与其他主体进行交互，也无法和其他主体一起实现协议目标，因此协议会话中不可能出现孤立边，任一个流都存在邻接主机流。



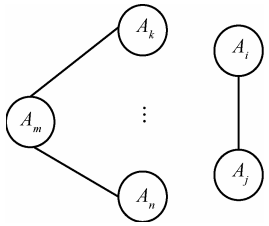
(a) 三方安全协议会话结构 1



(b) 三方安全协议会话结构 2



(c) 三方安全协议会话结构 3



(d) 协议参与方大于 3 且存在孤立边时会话结构

图 4　多方安全协议会话结构

据此，在确定某个消息为多方安全协议类型后，以该消息所属流 f_1 的源主机和目的主机 IP 为中心，建立与 f_1 具有相同协议类型的邻接主机流的集合。对多方安全协议而言，在该集合中至少存在一个流 f_2 与 f_1 属于同一会话。同理，在 f_2 的邻接主机流集合至少存在流 f_3 与 f_2 属于同一会话。由此可以得到 f_1、f_2 和 f_3 属于同一会话。由于会话不存在孤立边，通过邻接主机流集合可以查找到会话的所有流。

4.2　主机角色行为

协议参与者主机遵循协议规范，按照自己的角色（发起者、响应者或可信中心）与其他主机进行交互，并通过消息的收发来确认是在与自己所期望的主体进行交互。

主机角色行为特征：在协议一次会话中，担任不同角色的主机是确定的。角色相同，其对应的主机也相同。

据此，给出下述命题。其中命题 2 是命题 1 的逆否命题。

对多方协议消息 p_i 和 p_j 而言，记 p_i、p_j 的发送方和接收方主机分别为 A_i、B_i、A_j、B_j，主机对应角色分别为 $role_A^i$、$role_B^i$、$role_A^j$、$role_B^j$。

命题 1　若 p_i 和 p_j 属于同一次会话，$role_\alpha=role_\beta$，$role_\alpha \in \{role_A^i, role_B^i\}$，$role_\beta \in \{role_A^j, role_B^j\}$，则担任角色 $role_\alpha$ 的主机和担任角色 $role_\beta$ 的主机也相同。

命题 2　若 $role_\alpha=role_\beta$，$role_\alpha \in \{role_A^i, role_B^i\}$，$role_\beta \in \{role_A^j, role_B^j\}$，且担任角色 $role_\alpha$ 的主机和担任角色 $role_\beta$ 的主机不同，则 p_i 和 p_j 不属于同一次会话。

主机担任的角色不同，执行的收发行为不同。对多方协议消息 p_i，可以根据消息的通信双方确定参与协议的主机 A_i、B_i 和主机对应的角色 $role_A^i$、$role_B^i$，进而从主机担任的角色出发，查找与该主机角色相关的消息。

4.3　主体消息行为

主体消息行为特征：多方安全协议一次会话中，前后交互步骤的消息之间存在具有相关性的字段。

以认证协议为例，协议主体需要确认是与自己所期望的主体进行交互，因此，在协议的设计中，为防止攻击，收发的消息中都存在具有相关性的字段，这些字段能为接收方/发送方识别以认证消息来源。为描述该特征，本文给出下述定义。

原子字段：对安全协议消息进行解析时，如果一段连续字节序列是最小不可分割的（如版本号、命令、长度指示字段、主体标识、随机数、密文数据等），则称该段字节序列为原子字段。

变换 $f: v \to t$ 表示将一个变量 v 变换为项 t。

在协议中可以进行加密、异或、散列或进行其他变换（例如，协议中经常出现随机数的加减 1 变换）。

新鲜数据：协议中具有新鲜性的数据，例如随机数、时间戳。

报文相关性：对 2 个消息 p_1 和 p_2 而言，假定 p_1 包含原子字段 $f_1(m)$，p_2 包含原子字段 $f_2(m)$。若 $f_1(m)$和 $f_2(m)$都为对同一新鲜数据 m 进行变换得到的数据，则称 p_1 和 p_2 存在相关性，记为 $p_1 \mathrm{R} p_2$，$f_1(m)$和 $f_2(m)$为消息相关项。

在安全协议中，消息相关项通常为具有新鲜性的随机数，也可能为相关的密文等。由于安全协议采用密码技术，消息载荷可能为明文、密文或者明密文的混合。在同一次会话中，消息相关存在以下 3 种情况。

1) 明文相关：$f_1(m)$和 $f_2(m)$都为明文原子字段。多数情况下 $f_1(m)=f_2(m)=m$；少数情况下，$f_1(m)\neq f_2(m)$，例如，$f_1(m)=m$，$f_2(m)=m+1$。

2) 明密文相关：$f_1(m)$为明文原子字段，$f_2(m)$为密文原子字段，该情况下 $f_1(m)\neq f_2(m)$。

3) 密文相关：$f_1(m)$为密文原子字段，$f_2(m)$为密文原子字段，多数情况下，$f_1(m)\neq f_2(m)$；少数情况下，$f_1(m)= f_2(m)$（例如，Ottway-Rees 协议中，主体在接收到包含密文的消息后，向另一主体发送包含

该密文的另一消息）。

在安全协议中，新鲜数据 m 由协议某个参与方生成，随后该参与方发送包含 m 的消息，其中 m 可能为明文数据，也可能在加密形式中出现。根据安全协议的数据准则，为防止 m 被篡改，保证协议的安全性，在协议后续执行过程中 m 一般在加密形式中出现。因此安全协议中消息相关基本都为明密文相关或者密文相关。

若 $f_1(m)=f_2(m)$，通过检测消息载荷内容即可确定报文相关性。若 $f_1(m)\neq f_2(m)$，由于无法解密协议中的密文数据，因此不能通过消息载荷内容检测消息相关性。为解决这一问题，本文将在 5.3 节给出主体消息相关性的检测方法。

5 会话主体行为特征检测

本节基于第 4 节的会话主体行为特征，给出相应的特征检测方法。

协议相同位置的消息具有相似的统计特征和负载内容特征。由于安全协议包含大量密文信息，与普通网络协议相比，负载内容特征串较为稀少。在识别协议的基础上，可根据消息大小等统计特征进一步确定消息对应协议的第几步交互步骤。

5.1 邻接主机行为检测原理

在确定某个消息 p_i 为多方安全协议类型后，建立与该消息 p_i 所属流 f_i 具有相同协议类型的邻接主机流的集合。随后在邻接主机流集合中进一步确定与 f_i 属于同一次会话的流。

由于安全协议“高并发性”的特点，在 f_i 的邻接主机流集合中，可能存在协议的多次并发会话。

例如，在多方协议存在可信中心主机 T 的情况下，主机在传输数据之前通常与 T 的指定端口建立连接，如图 5 所示。

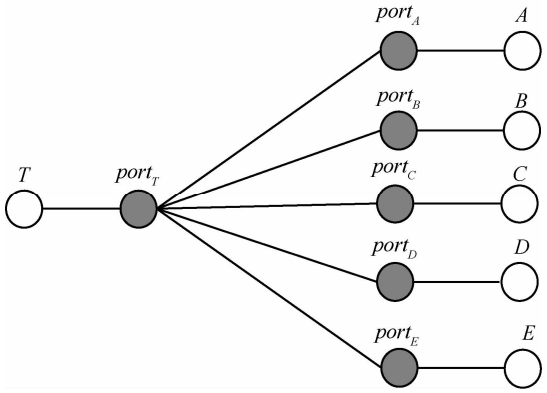


图 5 多个主机与可信中心相连

根据三元组性质[2,13,14]，与同一主机指定端口相连的多个流属于同一类型协议，因此与 T 相连的多个流互为邻接主机流，并且具有相同协议类型。这多个流可能为协议的并发会话，也可能属于同一次会话。

由于协议遵循特定规范，同一协议相同位置的消息具有相似的统计特征（如消息大小、方向等）和消息内容特征（特征串）。对不同位置的消息而言，消息内容特征和消息统计特征也不相同。如图 6 所示。

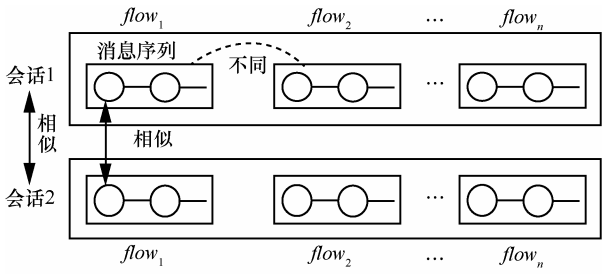


图 6 区分并发会话

对邻接主机流中消息特征进行比较，特征不同，则属于协议不同位置的消息；特征相似，则属于协议相同位置消息，存在并发会话。

假定同一次会话中 p_i 所属流 f_i 与 p_j 所属流 f_j 为邻接主机流。在 f_i 的邻接主机流集合中，若仅存在一个第 j 个位置消息 p_j^1，则 p_j^1 与 p_i 属于同一次会话，p_j^1 所属流与 f_i 也属于一次会话；若存在多个第 j 个位置消息 $p_j^1, p_j^2, \cdots p_j^n$，则存在并发会话，需要进一步判定与 p_i 属于同一次会话的 p_j^k（$1 \leqslant k \leqslant n$）。

5.2 主机角色行为检测原理

对多方协议第 i 个消息 p_i 和第 j 个消息 p_j，根据消息的通信双方确定参与协议的主机和主机对应的角色。记 p_i 发送方主机为 A_i，角色为 $role_A^i$；接收方主机为 B_i，角色为 $role_B^i$。由命题 1 可知，在一次会话中，相同角色对应的主机相同。通过 A_i、B_i 角色确定第 j 个位置消息对应的主机和角色，发送者主机记为 A_j，角色记为 $role_A^j$；接收者主机记为 B_j，角色记为 $role_B^j$。随后与 p_j 对应的主机、角色进行比较。由命题 2 可知，若 p_j 对应的主机、角色与 A_j、B_j、$role_A^j$、$role_B^j$ 不一致时，则 p_j 与 p_i 不属于同一次会话。若一致，则需要进一步进行判定。

5.3 主体消息行为检测原理

由 4.3 节可知，由于无法解密协议的密文数据，因此不能通过消息载荷内容检测主机消息行为特

征。为此本文建立消息相关性与协议主机交互行为之间的关系，从而将检测消息之间的相关性转化为检测协议的交互行为。

假定同一次会话中 p_i 和 p_j 具有相关性（$j>i$），p_i 包含 $f_1(m)$对应的网络数据（m 为新鲜数据），p_j 包含 $f_2(m)$对应的网络数据。

记 p_i 的发送者和接收者分别为主机 A_i、A_{i+1}；p_j（$1 \leqslant k \leqslant n$）的发送者和接收者分别为主机 A_j、A_{j+1}。主机 A_{i+1} 在接收包含 $f_1(m)$的消息 p_i 后，验证 p_i 是否符合协议规范。其中，当 $f_1(m)$为密文数据时，A_{i+1} 需要对 $f_1(m)$进行解密，获取其中的数据 m。若 p_i 符合协议规范，A_{i+1} 进行响应，向另一主机 A_{i+2} 发送第 i+1 条消息 p_{i+1}。随后，与之相似，主机 A_{i+2} 向另一主机 A_{i+3} 发送第 i+1 条消息 p_{i+2}；…；主机 A_j 向另一主机 A_{j+1} 发送第 j 条消息 p_j。

根据安全协议的设计准则，后续执行过程中新鲜数据 m 一般在加密形式中出现，$f_2(m)$通常为密文形式。对 A_j 而言，只有 A_j 接收到包含 m 的消息，A_j 才可以发送包含 $f_2(m)$的消息。

可达：在协议执行过程中，若主机 A 向主机 B 发送消息，则称 A 和 B 之间是可达的。

可达具有传递性，若主机 A 和 B 之间可达，B 和 C 之间可达，则 A 和 C 之间也是可达的。

在恢复协议会话过程中，根据消息特征，构建会话可能的消息序列 $p=\{p_1,p_2,\cdots,p_n\}$。在 p 序列中，根据 $p_i,p_{i+1},\cdots,p_j$ 确定每个消息的发送者主机和接收者主机。若 p_i 和 p_j 具有相关性，则 A_i 和 A_j 之间是可达的。

因此，对 p 序列，检测 A_i 和 A_j 之间是否可达。若不可达，则 p_i 和 p_j 不具有相关性，不属于同一次会话。若可达，进一步判定 p_i 和 p_j 是否属于同一次会话。

6　多方安全协议会话识别方法

结合第 5 节的主体行为特征检测原理，本节给出会话识别方法，其框架如图 7 所示。

6.1　数据预处理

Wireshark 软件内嵌一个 Lua 语言执行引擎，并提供一系列 Lua 脚本函数接口。基于 Lua 脚本可以编写 Wireshark 插件，实现协议识别、协议报文解析，也可以获取 Wireshark 提供的与协议相关的信息（如报文捕获时间、发送双方 IP、端口等信息）。

在 Wireshark 捕获报文后，基于 Lua 脚本获取报文原始数据，对报文进行重组，得到消息数据，并据此建立重组后的 flow 表。随后对消息进行协议识别。当消息识别为多方协议类型时（例如 Kerberos 协议），由于协议相同位置的消息特征相似，进一步识别为协议的第几条消息。

若该消息为协议第一个消息 p_1，对 p_1 构建邻接主机流，并在待识别会话表中增加一个新的会话，存储协议 ID、会话开始时间、该消息发送者和接收者的 IP、port、角色以及消息序号。若消息未识别，进一步等待后续消息。

6.2　会话识别

当 p_j 识别为协议第 j 个（$j \geqslant 2$）消息时，为确定 p_j 所属的会话，执行如下步骤。

Step1　在待识别会话表中，检测 p_j 所属流（或者反向流）的其他报文是否已经存在。若存在，将该消息并入相应的会话，更新会话信息；若不存在，转至 Step2。

Step2　根据时序关系确定 p_j 所属会话。在待识别会话表中，检测协议第 $j-1$ 个位置消息的数目 n_1。若 $n_1=1$，p_j 与 $j-1$ 个位置消息属于同一次会

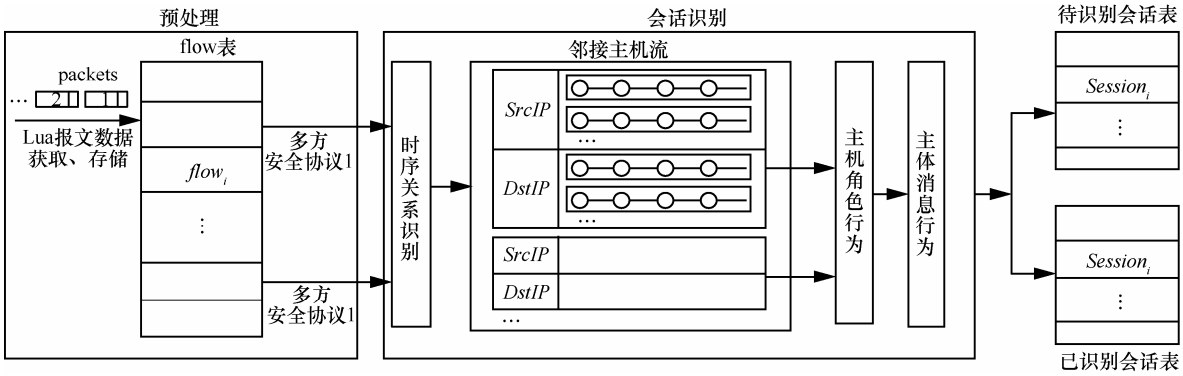


图 7　多方安全协议会话识别方法框架

话。若 $n_1>1$，转至 Step 3。

Step3 构建邻接主机流，根据 NHB 确定 p_j 所属会话。

由 4.1 节可知，对三方安全协议，只需对协议第一个消息 p_1 所属流建立邻接主机流集合即可。对多方安全协议，首先根据 p_1 建立邻接主机流表；对后续消息 p_i，检测是否存在以 p_i 源主机（或目的主机）为中心的邻接主机流集合。若不存在，建立 p_i 的邻接主机流集合。本文将与 p_i 所属流 f_i 具有相同协议 ID 的邻接主机流集合记为 NH_{p_i}。

随后判定 p_j 属于邻接主机流集合 NH_{p_i} 的数目 n_2。若 $n_2=1$，则 p_j 与 p_i 属于同一次会话。若 $n_2>1$，转至 Step 4。

Step4 根据 PMB 和 HRB 确定 p_j 所属会话。对 p_j 所属的多个邻接主机流集合依次进行处理。记第 t 个（$1\leqslant t\leqslant n_2$）邻接主机流集合为 $NH^t_{p_i}$。

记 p_i 发送者和接受者主机为 A、S，主机对应角色分别为 $role_A$、$role_s$。根据 A、S 的角色行为确定第 j 个位置消息的主机及角色，并与 p_j 对应的主机及角色进行比较。

1) 若两者不一致，则对下一邻接主机流集合进行处理。

2) 若两者一致，根据 PMB 确定 p_j 所属会话。若同一次会话中 p_j 与第 l 个消息 p_l 具有相关性（$l<j$），根据交互行为检测 p_j 与第 l 个位置消息 p_l 的相关性，即检测 p_l 与 p_j 的发送者主机之间是否可达。若不可达，则对下一邻接主机流集合进行处理。

3) 若可达，根据 $NH^t_{p_i}$ 中后续消息进行判定。若可以确定与 p_i 属于同一次会话的第 i' 位置的消息（$i'>j$），在 $NH^t_{p_i}$ 中进一步确定第 j 个位置消息的数目 n_3。

①若 $n_3=1$，则 p_j 与该 p_i 属于同一次会话。

②若 $n_3>1$，则 $NH^t_{p_i}$ 中存在并发会话，根据 6.3 节进一步进行识别。

Step5 若协议会话相关消息全部识别，则更新已识别会话表、未识别会话表、邻接主机流表。将协议会话消息序列存储至已识别会话表中，并将该会话从未识别会话表删除；同时在邻接主机流表中删除相关流。若协议会话构建未完成，继续对捕获的后续数据分组进行处理，直到会话相关消息全部识别。

6.3 邻接主机流存在并发会话时识别方法

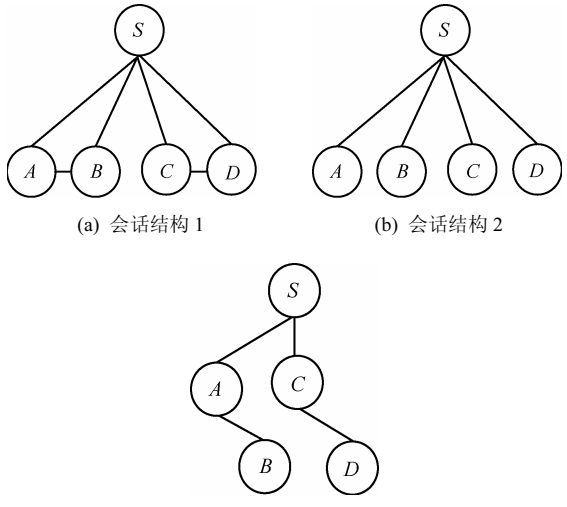
记 $NH^t_{p_i}$ 中第 j 个位置消息为 p_j^1，p_j^2，…，p_j^m。由于 p_j^1，p_j^2，…，p_j^m 所属流与 p_i 所属流为邻接主机流，因此 p_j^1，p_j^2，…，p_j^m 各自的主机集合一定或者包含主机 A 或者包含主机 S。而 p_j^1，p_j^2，…，p_j^m 通信双方角色符合根据 p_i 确定的角色。分 2 种情况讨论。

1) p_j^1，p_j^2，…，p_j^m 对应的主机存在一个相同主机 A 或者 S，且该主机在 p_j^1，p_j^2，…，p_j^m 中担任的角色相同。

由 p_j^1，p_j^2，…，p_j^m 可以确定参与并发会话的主机。对没有参与并发会话且属于 $\{A,S\}$ 的主机而言，在执行其角色过程中，其相应的交互行为处于同一会话中，因此由该主机的角色行为，可以确定与该会话相关的消息和主机，进而确定与 p_i 属于同一次会话的消息。

例如，假定主机 S 同时参与 2 个会话 A、B、S 和 C、D、S。2 个相同主机之间的流用一条边表示，其会话结构如图 8 所示。

图 8(a)中 AS 的邻接主机流为 $\{AB,SB,SC,SD\}$。假定 SB 和 SD 中同时存在第 j 个位置消息，比较该消息的通信双方，可以确定主机 S 参与了并发会话。主机 $A\in\{A,S\}$ 且没有参与并发会话。从 A 的角色行为出发，在 $\{AB,SB,SC,SD\}$ 中与 A 相连的流 AB 与 AS 属于同一会话，会话用户集合为 $\{A,B,S\}$。因此 SB、AS 属于同一会话。同理，图 8(c)会话结构中，也可以区分不同会话。



(a) 会话结构 1　(b) 会话结构 2

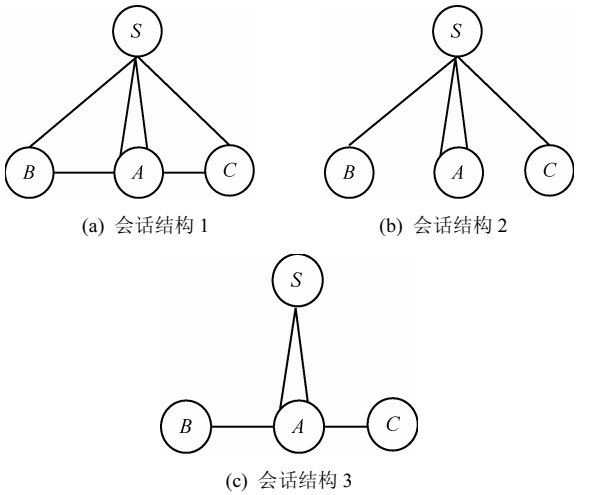
(c) 会话结构 3

图 8 主机 S 同时参与 2 个会话情况下的会话结构

图 8(b)中 AS 的邻接主机流为 $\{SB,SC,SD\}$，无法区分不同会话。由于大多数安全协议目的是保证后续传输过程的安全性，因此参与协议同一会话的主机一般进行后续通信。由此可以从主机后续传输行为出发，确定属于同一次会话的协议主机集合。

2) p_j^1，p_j^2，…，p_j^m 对应的主机集合存在 2 个相同主机 A 和 S，且 A、S 在 p_j^1，p_j^2，…，p_j^m 中担任的角色相同。

例如，假定主机 A、S 同时参与 2 个会话 A、B、S 和 A、C、S，并且 2 个相同主机之间的流用一条边表示，其会话结构如图 9 所示。



(a) 会话结构 1　　(b) 会话结构 2

(c) 会话结构 3

图 9　主机 A、S 同时参与 2 个会话情况下的会话结构

在图 9(a)中，2 个 AS 边分别记为 AS_1 和 AS_2。SB 的邻接主机流为 $\{AS_1,AS_2,AB,SC\}$。

假定 AS_1 和 AS_2 中同时存在第 j 个位置消息，比较该消息的发送双方，可以确定主机 A、S 参与并发会话。主机 $B\in\{B,S\}$ 且没有参与并发会话。从 B 的角色行为出发，在 $\{AS_1,AS_2,AB,SC\}$ 中与 B 相连的流 AB 与 SB 属于同一次会话。

为进一步确定 AS_1 和 AS_2 中哪一个和 SB 属于同一次会话，执行下述步骤。

Step1　由于协议一般包含挑战和响应，可根据 AS_1 和 AS_2 中后续消息进行识别。例如，假定某时刻 AB 和 SB 已执行 k 个消息，$k>j$，根据角色 A 的行为，$k+1$ 位置消息属于 AS。随后根据 AS_1 和 AS_2 到达的消息进行判定。

Step2　若 AS_1 和 AS_2 中相同协议位置的消息总是同时出现，这时无法根据后续消息识别会话。由于主机优先响应先到达它的消息，根据先进先出原则识别会话。

根据消息特征，SB 和 SC 中消息具有相似的特征。若某一时刻 B 向 S 发送第 k 位置消息，随后 C 向 S 发送第 k 位置消息，S 接收到消息后进行响应，向 A 发送第 $k+1$ 位置消息。本文认为 S 会优先响应 B 发送的消息。对捕获的 2 个由 S 响应的 $k+1$ 位置消息，将第 1 个 $k+1$ 位置消息与 SB 判定为属于同一次会话。这可能存在一定误差，在分布式网络中，先发送报文不一定先到达，这将在以后工作中进一步优化。

在图 9(b)、图 9(c)会话结构中，识别会话原则与图 9(a)相似，本文不再进行说明。

7　实验结果与分析

Kerberos 协议是一种广泛应用的认证协议，提供了一种具有较高安全性的用户身份认证和资源访问控制的机制。Ban-Yahalom（BY）协议和 Needham-Schroeder（NS）共享密钥协议属于经典基础安全协议，通过可信第三方分发新鲜的对称共享密钥并实现互认证。为评估安全协议会话识别方法的有效性，本文选取这 3 个经典的多方安全协议进行实验，结果表明该方法具有较高的识别精度。

7.1　实验环境和实验数据

Kerberos 数据集、BY 数据集以及 NS 数据集都由实验室局域网产生。每个数据集包含以下 3 种场景：1)运行单个会话；2)同一时间运行多个会话，且参与会话的主机不同；3)同一时间运行多个并行会话。

以 Kerberos 协议为例，其实验环境如图 10 所示。Windows Server 2003 中的 Active Directory 支持 Kerberos V5 协议，实验室在 Windows 域环境中运行 Kerberos，构建 2 个不同域，每个域中包含一个域服务器、2 台普通客户端、2 台应用程序服务器。

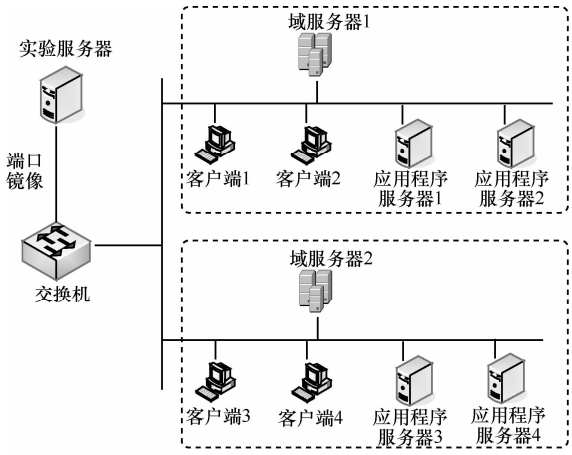


图 10　Kerberos 实验环境

Kerberos 协议规范将 AS 和 TGS 描述为不同的服务器。在实际运行中，它们是一个 Kerberos 服务器内不同的协议入口点。本文将 Kerberos 协议作为三方协议进行分析。

由于 Kerberos 协议有不同的工作模式，而一个完整的认证过程包括 6 条消息，第 6 条为可选消息，该种情况下会话数目是可控的。本文对该模式下 Kerberos 协议进行分析。

BY 协议和 NS 共享密钥协议实验中，网络拓扑结构与之类似，构建 2 个小型局域网，每个局域网内包含 1 台服务器、4 台客户端主机，由局域网服务器担任可信第三方。实验室采用 Spi2Java 工具分别生成 BY 协议和 NS 协议的应用程序，并在各个主机上运行。

实验数据如表 1 所示。

表 1 实验数据集

数据集	场景	会话数目	流数目
Kerberos 协议	单个会话	400	2 400
	多个会话	400	2 400
	并行会话	400	2 400
BY 协议	单个会话	400	1 200
	多个会话	400	1 200
	并行会话	400	1 200
NS 共享密钥协议	单个会话	400	1 600
	多个会话	400	1 600
	并行会话	400	1 600

对数据集每个消息标识其所属的会话，确定会话基准 Groudtruth。随后对数据集利用本文提出的会话识别算法进行测试。

在数据预处理阶段，为识别 BY 协议、NS 共享密钥协议以及 Kerberos 协议前 4 条消息，当协议采用固定端口时，采用端口识别协议；当协议采用动态端口时，鉴于负载内容特征和数据分组大小特征是比较稳定的识别特征，本文采用这 2 个特征识别协议。Kerberos 协议第 5、6 条消息 AP-REQ 和 AP-REP 则主要采用负载内容特征进行识别。这 2 条消息是由客户端访问应用服务器产生的，其内容包含在应用服务器消息中。例如客户端访问 IIS 服务器的 WWW 服务，则 AP-REQ 和 AP-REP 所属流识别为 HTTP 协议。在本实验室捕获的 Kerberos 流量中，Kerberos 通过 GSS-API（generic security service application program interface）实现，AP-REQ 和 AP-REP 消息中包含特定字符串，其中，GSS-API OID（对象标识）为固定值“2b 06 01 05 05 02”，Kerberos V5 OID 为“2a 86 48 86 f7 12 01 02 02 02”，而 AP-REQ 消息类型值为“0e”，AP-REP 消息类型值为“0f”。对不同的应用流量，本文采用关键字匹配将包含这 2 个消息的报文识别为 Kerberos 协议消息。

7.2 实验结果

本文采用如下性能指标对识别算法的识别能力进行评估。记样本中流数目为 N，流 F 所属会话为 S。N_1 表示流 F 所属会话 S 被正确识别的样本数，N_2 表示流 F 所属会话 S 没有别识别的样本数，N_2 表示流 F 所属会话 S 没有被识别的样本数，N_3 表示流 F 所属会话 S 被识别为非 S 的样本数。识别正确率=$\frac{N_1}{N}$，漏报率=$\frac{N_2}{N}$，误报率=$\frac{N_3}{N}$。识别正确率越高，误报率和漏报率越低，相应的识别效果越好。在不同场景下识别率总体效果如图 11 所示。

完整实验结果如表 2 所示。

表 2 实验识别结果

数据集	场景	识别率/%	漏报率/%	误报率/%
Kerberos 协议	单个会话	97.5	2.5	0
	多个会话	96	4	0
	并行会话	89.5	3.5	7
BY 协议	单个会话	94.5	5.5	0
	多个会话	94.25	5.75	0
	并行会话	89.75	4.75	5.5
NS 共享密钥协议	单个会话	95.5	4.5	0
	多个会话	94.25	5.75	0
	并行会话	88.375	6.5	5.125

由上可知，在 3 个场景下都可以较好地识别协议会话。

7.3 分析与讨论

本文方法与现有工作侧重点不同。现有方法主要是进行协议识别，在此基础上可以分析两方协议会话，而本文主要是进行多方协议会话识别。

对实验中存在误报率和漏报率流量进行分析，有如下结论。

1) 漏报率发生，由于任何一种协议识别方法都不能保证百分之百的识别率，而协议类型无法识别，会进而影响会话识别的效果，因此漏报率无法完全避免。

2) 误报率发生，主要发生在协议的并行会话场

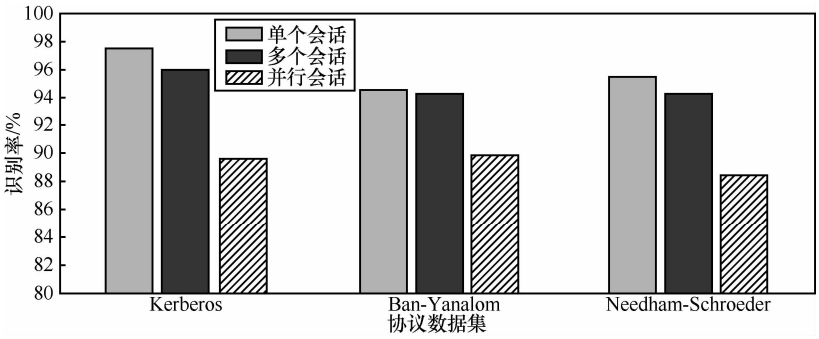


图 11　协议数据集识别率

景下。对 Kerberos 协议而言，当客户端 C 通过同一域服务器主机同时访问多个应用程序服务器 V_1 和 V_2 时，根据先进先出机制识别协议第 5、6 条报文所属会话时，造成误识别。BY 协议和 NS 共享密钥协议与之相似，同样是由于先进先出机制造成误报。由 6.3 节可知，只有在以下情况采用先进先出机制识别会话，2 个相同主机参与并行会话，且对这 2 个主机之间的流而言，相同协议位置的消息总是同时出现。在实际网络中，这种情况发生的概率不高，因此误报率也不高。在今后工作将对此进行改进，进一步降低误报率。

8　结束语

本文提出了多方安全协议会话识别方法，识别分布在多个相关流中的多方安全协议会话过程。该方法首先研究在协议同一次会话中主体的行为特征，结果表明多方安全协议会话具有 NHB 特征、HRB 特征以及 PMB 特征。随后给出了主体行为特征检测原理，其中建立了消息相关性与协议交互行为之间的关系，将检测消息相关性转化为检测协议的交互行为。最后，本文给出了多方安全协议会话识别方法，并通过实验验证了该方法的有效性。结合现有协议识别方法和本文方法，可以实现对安全协议的会话识别，从而更细粒度、更精确地控制、管理流量。

但本文方法还存在一定的局限性：1)仅针对 Kerberos 协议、Ban-Yahalom 协议以及 Needham-Schroeder 共享密钥协议进行实验，需对更多协议进行实验以验证本文方法有效性；2)本文捕获流量为协议完整的运行过程，需要考虑当捕获流量不完整时的会话识别；3)在对协议并行会话时识别时，需要对先进先出机制进行改进，以进一步降低误报率。

参考文献：

[1] CLAFFY K C. Internet Traffic Characterization [D]. SanDiego: University of California, 1994.

[2] MA J, LEVCHENKO K, KREIBICH C, *et al*. Unexpected means of protocol inference[A]. ACM SIGCOMM Conference on Internet Measurement[C]. Rio de Janeriro, Brazil, 2006. 313-326.

[3] DAINOTTI A, PESCAPÉ A, CLAFFY K C. Issues and future directions in classification[J]. IEEE Network, 2012, 26(1): 35-40.

[4] BERNAILLE L, TEIXEIRA R. Early identification of encrypted applications[A]. Proceedings of PAM[C]. Belgium, 2007.

[5] BUJLOW T, VALENTIN C E, PERE B R. Comparison of deep packet inspection (DPI) tools for traffic classification[R]. Polytechnic University of Catalonia, 2013.

[6] WRIGHT C, MONROSE F, MASSON G M. HMM profiles for network traffic classification (extended abstract)[A]. Proceedings of the ACM Workshop on Visualization and Data Mining for Computer Security[C]. New York, 2004.9-15.

[7] ZUEV D, MOORE A W. Traffic classification using statistical approach[A]. Proceedings of the 6th International Workshop on Passive and Active Network Measurement[C]. Berlin, Germany, 2005. 321-324.

[8] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[A]. Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems[C]. Banff, Alberta, Canada, 2005. 50-60.

[9] NGUYEN T T T, ARMITAGE G , A survey of techniques for Internet Traffic classification using machine learning[J]. IEEE Commun. Surveys & Tutorials, 2008, 10(4):56-76.

[10] XIE G W, ILIOFOTOU M, KERALAPURA R, *et al*. SubFlow: towards practical flow-level traffic classification[A]. The 31st Annual IEEE International Conference on Computer Communications: Mini-Conference[C]. Orlando, Florida, USA , 2012. 2541-2545.

[11] GRIMAUDO L, MELLIA M, BARALIS E, *et al*. SeLeCT: self-learning classifier for Internet traffic[J]. IEEE Transactions on Network and Service Management, 2014, 11(2): 144-157.

[12] WANG D, ZHANG L, YUAN Z, *et al*. Characterizing application behaviors for classifying P2P traffic[A]. International Conference on Computing, Networking and Communications(ICNS)[C]. Honolulu, HI, 2014.21-25.

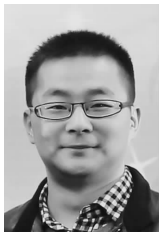
[13] CANINI M, LI W, ZADNIK M, *et al*. Experience with highspeed automated application-identification for network-management[A]. Proceeding of ACM/IEEE Symposium on Architectures for Networking and Communications Systems[C]. Princeton, New Jersey, USA, 2009.209-218.

[14] ZHANG J, XIANG Y, WANG Y, *et al*, Network traffic classification using correlation information[J]. IEEE Transactions on Parallel Distrib System, 2013, 24(1): 104-117.

作者简介：

朱玉娜（1985-），女，山东菏泽人，解放军信息工程大学博士生，主要研究方向为安全协议逆向与识别。

韩继红（1966-），女，山西定襄人，博士，解放军信息工程大学教授、博士生导师，主要研究方向为网络与信息安全、安全协议形式化分析与自动化验证。



袁霖（1981-），男，河南商丘人，博士，解放军信息工程大学副教授，主要研究方向为安全协议形式化分析与自动化验证、软件可信性分析。

陈韩托（1990-），男，浙江奉化人，解放军信息工程大学硕士生，主要研究方向为协议在线安全性分析。

范钰丹（1981-），女，河南邓州人，解放军信息工程大学讲师，主要研究方向为安全协议形式化分析与自动化验证。