doi:10.3969/j.issn.1000-436x.2012.z1.029

# 面向多级安全的结构化文档描述模型

苏锰[1]，李凤华[2]，史国振[3]，李莉[3]

（1. 西安电子科技大学 综合业务网理论及关键技术国家重点实验室，陕西 西安 710071；
2. 中国科学院 信息工程研究所，北京 100093；3. 北京电子科技学院 电子信息工程系，北京 100070）

**摘 要**：面向多要素访问控制和多级安全需求，为了解决网络环境的开放性、多样化所带来的安全问题，基于现有的结构化文档描述模型及访问控制模型，提出了一种面向多级安全的结构化文档描述模型和描述方法，并给出安全属性的描述结构及其对应的可扩展标识语言（XML）实例，最后对提出的模型进行了安全性分析。

**关键词**：结构化文档；访问控制；多级安全；可扩展标识语言；描述方法

**中图分类号**：TP 309.2 **文献标识码**：B **文章编号**：1000-436X(2012)Z1-0222-06

# Representation model of structured document for multilevel security

SU Mang[1], LI Feng-hua[2], SHI Guo-zhen[3], LI Li[3]

(1. National Key Laboratory of Integrated Services Network, Xidian University, Xi'an 710071, China;
2. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;
3. Department of Electronic Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China)

**Abstract:** In order to solve the problem caused by the variety and openness of the network, the research for the models of structured document and access control were taken. A new structured document representation model and method for the security requirements of the multi-element access control and multi-level security was proposed, and corresponding structure of security attribute and example of extensible markup language (XML) was given. Finally, analyzed the security performance.

**Key words:** structured document; access control; multilevel security; XML; representation method

## 1 引言

随着网络、数字出版等技术的进步，阅读终端的飞速发展，使文档阅读的需求发生了巨大的变化，要求文档能够面向多样化、普及化的终端，既有版式的清晰性和条理性，也要具备流式的内容可变性，并能够自适应终端屏幕大小。结构化文档融合了流式和版式描述信息，面向多样化的阅读、显示需求，已经逐步成为互联网信息传播的重要载体。针对结构化文档格式的研究一直是文档描述的重点。一个文档可以采用层次化组织的物理和逻辑结构进行描述，物理结构反映文档的布局，逻辑结构反映文档的组织。文档的物理结构和逻辑结构的整体构成了文档模型[1]。

访问控制最初面向大型机资源共享的需求，传统的访问控制研究经历了自主访问控制、强制访问

**收稿日期**：2012-06-28

**基金项目**：国家自然科学基金资助项目（61170251）；教育部重点项目基金资助项目（209156）；北京市自然科学基金资助项目（4102056）；新闻出版重大科技工程项目基金资助项目（GXTC-CZ-1015004/05）

**Foundation Items:** The National Natural Science Foundation of China (61170251); The Key Program of Scientific and Technology Research of Ministry of Education (209156); The Natural Science Foundation of Beijing (4102056); The Major Science and Technology Project of Press and Publication-Research and Development (GXTC-CZ-1015004/05)

控制、基于角色的访问控制等模型。为了适应分布式网络环境的特点，出现了基于任务的访问控制、面向分布式和跨域的访问控制、与时空相关的访问控制等模型。云计算、移动计算等的出现，使得访问控制的研究向细粒度、多要素的方向发展，基于属性的访问控制、基于行为的访问控制等模型相继出现。目前如何针对网络环境下信息的传输进行对象化、细粒度的访问控制，满足用户个性化需求的同时，保证信息资源合理、合法使用成为了访问控制研究面临的新挑战。

多级安全[2]主要关注信息的分级管理和访问授权，保证不同安全级别的信息只能被享有相应权限的用户访问，BLP[3]、Biba[4]等模型通过实施严格的强制访问控制策略，在一定程度上保护了信息的机密性和完整性。

目前，泛在网络环境下的信息多以结构化文档的方式进行交互和传播，而且随着在线交互设备的多样化，结构化文档的访问控制及安全属性描述已经逐渐走向对象级、细粒度，即文档包含子文档，子文档包含对象，客体的访问控制以对象为单位。现有的结构化文档描述模型中缺少针对访问控制和多级安全的支持，导致在多级安全环境下，结构化文档的机密性、完整性受到威胁，基于结构化文档的访问控制不能迎合多级安全的需求。因此本文提出一种面向多级安全的结构化文档描述模型，能够保证文档流式和版式信息完备，并解决结构化文档在日趋复杂的网络环境下机密性、完整性、访问控制等问题。

## 2　结构化文档

结构化文档同时描述了文档的版式信息和流式信息，能够更好的适用于自适应显示。在众多的结构化文档描述模型中，PDF、XPS 和 CEBX 较为成熟。其中，Adobe 推出的 PDF 1.3 规范引入了 logical structure，PDF 1.4 规范引入了 tagged PDF 来完善流式信息的表达；其后又将 XML 引入，用于对 MARS 文档格式中信息进行结构化的描述。李宁等人针对“标文通”与 Tagged PDF 的信息交换进行了实验，为减少办公文档的跑版问题提供了积极的借鉴意义[5]。微软公司也在其固定版式文件 XPS（XML paper specification)中采用类似的方式对逻辑结构信息进行了兼容[6]，但是以上研究并没有完全解决信息数据的结构化问题。Bloechle 等人基于 Dori 模型开展了一系列的研究工作，于 2006 年提出了 XCDF[7]格式，XCDF 文档与 Tagged PDF 相比，版式信息与流式信息的结合更为紧密合理，并且采用了 XML 来描述相关信息，使得其构造、使用更为方便，基于上述研究，文献[8]提出了一种从已有固定版式文档中重新构造文档逻辑结构的方法——Dolores。为了缩小文档体积、便于使用，Bloechle 对 XCDF 格式进行了优化[9]。

北大方正公司 2005 年在原来 CEB 版式结构文档的基础上启动了 CEBX 计划，并吸收 Tagged-PDF、MARS 流式特征，推出了 CEBX 1.1 版本，能够较好的解决版式和流式文档的融合问题，并分别针对移动设备和文档存储，提出了 CEBX 1.2-M 和 CEBX 1.2-A 版本。CEBX 采用了打包的形式，将文档整体描述、安全描述、版式信息、流式信息以及资源和物理层信息进行整合。CEBX 添加了文档整体安全描述[10]，能够实现整个文档及其包含文件的加密、签名以及整体使用权限的定义，并且支持 DRM 解决方案，初步解决了结构化文档在网络传输和使用过程中的机密性、完整性等问题。

但是，随着分布式计算、移动计算、云计算以及泛在计算的出现，网络环境日趋复杂，如何对结构化文档进行多级安全管理，并满足用户随时、随地访问结构化文档的控制需求，成为结构化文档描述的未来的研究方向。

## 3　面向多级安全的结构化文档的描述模型

### 3.1　面向多级安全的结构化文档描述

针对上述结构化文档在泛在网络环境中面临的访问控制和多级安全管理问题，本文将基于 CEBX 等结构化文档描述方法，提出一种如图 1 所示的新型结构化文档描述模型。该模型分为 2 个层次，第 1 层包含了文档入口、文档安全属性描述、文档根节点、页面信息、文档逻辑结构描述、文档样式结构描述。其中，文档入口描述了文档的安全属性、基础信息、文档根节点等内容及其相互关联关系；文档安全属性描述了对文档信息进行加密和签名所使用的算法、密钥以及初始向量等信息；文档根节点的定义主要用于实现文档的嵌套和包含，描述了文档及其子文档之间的逻辑关系，子文档同样包含了文档入口、安全属性描述等信息；文档逻辑结构描述与文档样式结构描述对文档的元素组织形式、显示方式进行了描述，包含了文档章、节

等的组织结构和样式表等信息；页面信息描述了页面的逻辑组成、关联关系、数量等信息。为了进一步描述结构化文档所包含资源及其物理数据，定义了模型的第 2 个层次，包含页面，每个页面由资源目录、资源描述和物理数据组成。资源是对一组图元或其他数据描述的集合。在页面中出现的图元、使用的数据或者结构都保存在资源中，在需要使用时从相应的资源中读取。一个文档可以包含一个或多个资源。

在图 1 所示的结构化文档模型中，文档逻辑结构描述、文档样式结构描述需要在网络传输和使用中保证其完整性，从而保证文件格式和版式的正常显示。并且需要保证文档所包含资源的合法使用，因此需要结合目前网络环境的多样性和用户访问个性化的需求，为资源描述添加安全属性描述，包含该资源的域安全属性、时态属性、环境属性，为了能够满足多级安全管理的需求，为安全属性描述添加了安全级别和访问范畴的定义。

文档逻辑结构和样式结构描述的完整性标识保证了结构化文档在网络传输过程中文档格式、显示形式等描述的完整、不可篡改；资源安全属性描述的添加能够为用户提供在任意时间、任意地点对任意资源合法访问的控制以及满足资源多级管理的需求。

### 3.2　安全属性描述

安全属性描述包含了文档整体的安全属性描述、针对逻辑结构描述和样式结构描述的完整性标识以及针对资源访问控制和多级安全管理的环境、时态、安全等级、访问范畴和域安全属性的描述。综合各类不同安全属性描述的特点，为图 1 中的描述模型添加安全属性描述定义，说明如图 2 和表 1 所示。

访问控制标签（access control label）主要包含了权限描述、权限对象、用户信息、管理员信息、域安全属性、时态属性、环境属性、安全级别和访问范畴。其中，权限定义了 Read、Write、Create、Modify 4 类，并且可以依据需要将其具体化，例如：针对多媒体文件，可以定义为 View（查看）、Play（播放）等。为了保证权限信息的完整性，为该项内容定义了签名标签。为了支持对结构化文档跨域流通时的控制，定义了域安全属性，主要描述在传播过程中所经由域的约束信息。时态、环境属性的定义用于对用户访问进行控制，结合基于行为的访问控制模型[11]，时态和环境属性分别标识了可以对文档及其对象进行访问的时间区段和环境要求。安全等级和访问范畴的定义为多级安全管理提供支

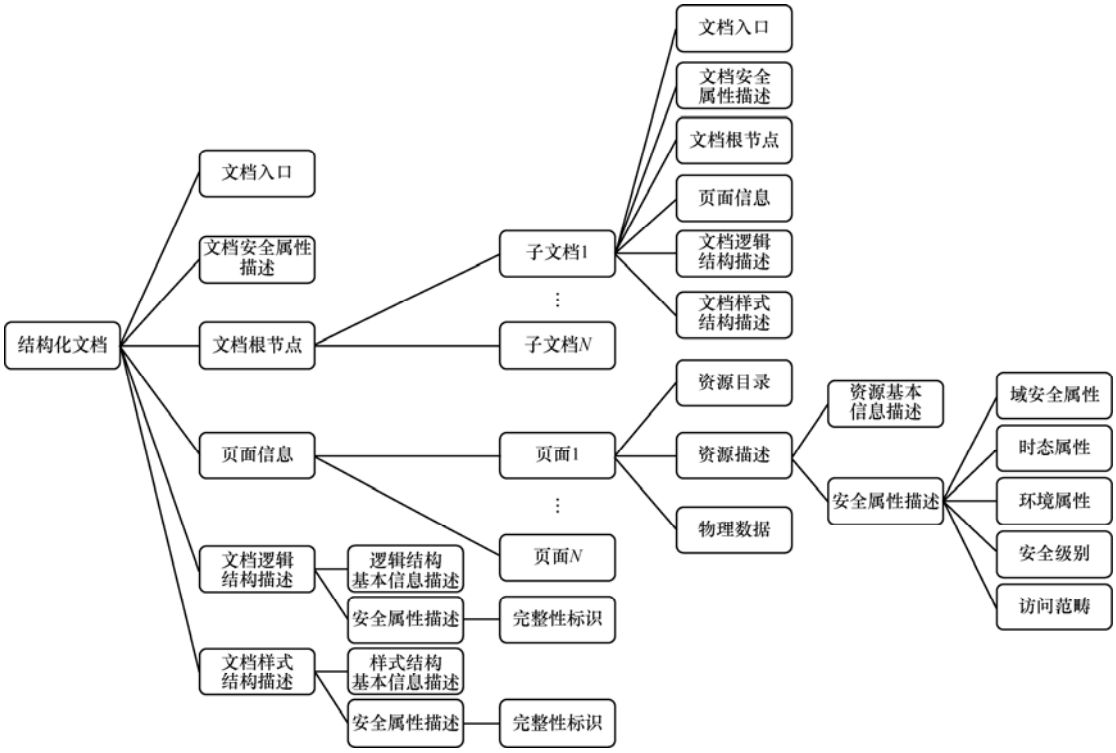



图 1　泛在网络环境下结构化文档描述模型

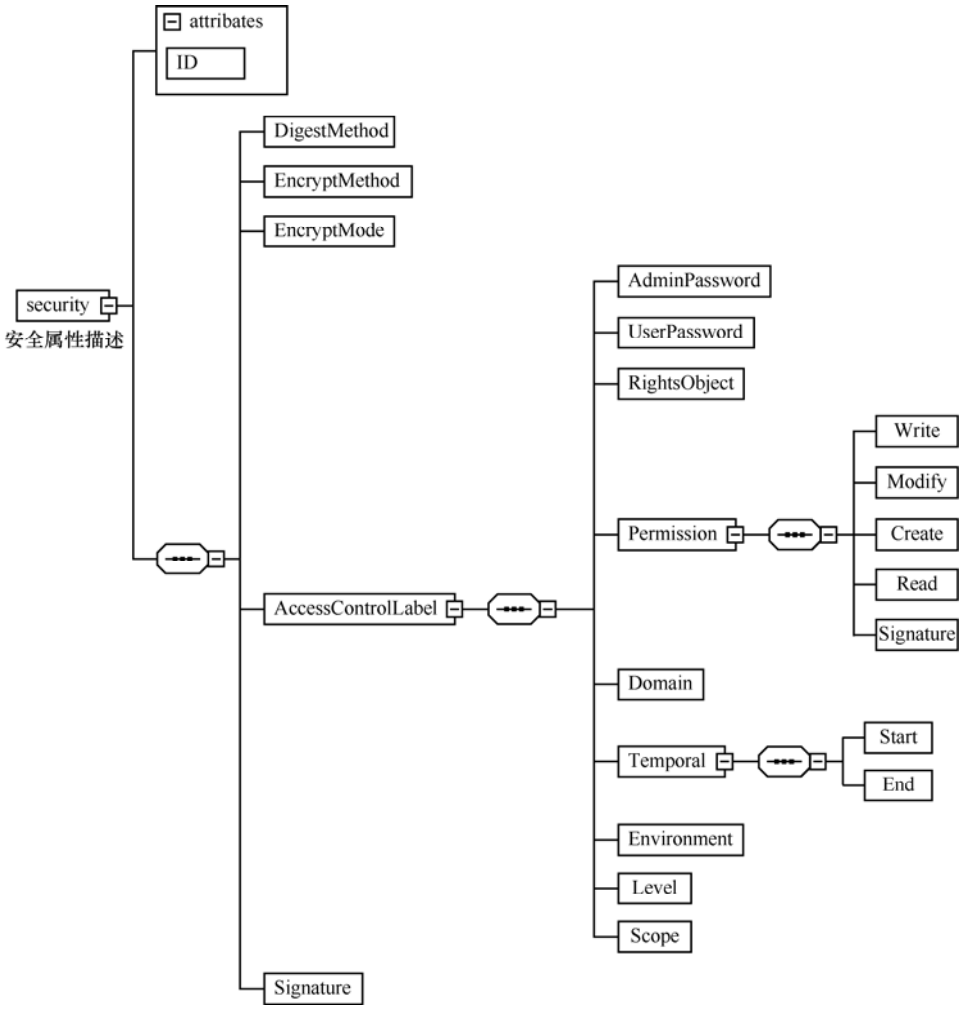



图 2　安全属性描述结构定义

持，安全级别标识了能够访问该文档或者资源对象主体的最低安全级别，访问范畴则标识了访问主体所处的组信息，例如：部门、系部等。

签名标签的定义主要用于保证文档及其相关信息的完整性，该标签中定义了签名所使用的算法、签名的有效期以及签名生成的数据即完整性标识信息，如图 3 所示。其中，ID 为数字签名的唯一标识，TimeStamp 为时间戳，用于记录签名时间和数字签名的有效期。由于结构化文档描述文件包含信息较多，因此在进行数字签名前，需要生成摘要数据。DigestMethod 和 DigestValue 分别表示了摘要算法和摘要数据。SignatureMethod 和 Signature Value 分别对应签名算法和签名数据。CertificationType 和 CertificationData 分别描述用于验证签名的证书类型和证书数据。在网络数据的传输过程中，接收方将依据接收到文档的 Signature 中摘要算法、签名算法、证书数据中的公钥信息生成验签数据，并与摘要数据对比以确认结构化文档该部分信息的完整性。

用户可以根据需求的不同而选取不同的字段，针对文档逻辑结构描述和样式结构描述需要选取 Signature 标签；针对资源的安全属性描述则需要选取访问控制标签，Signature 标签可以按照需要取舍。

**表 1　　安全属性描述标签说明**

| 名称 | 说明 |
|---|---|
| ID | 该安全性描述的唯一标识 |
| DigestMethod | 为支持文档及其对象的签名需求，为文档定义使用的摘要算法 |
| EncryptMethod | 为了支持文档及其对象的加解密，为其定义加解密算法 |
| EncryptMode | 加密模式 |
| AccessControlLabel | 定义了该结构化文档访问控制相关的信息，包含了管理员、用户的密码、权利描述对象、权限描述、时态、环境、安全级别以及范畴等 |
| Signature | 数字签名，用于文档及其对象完整性描述 |

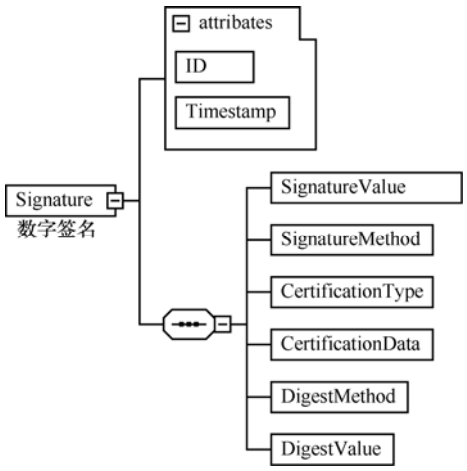



图 3　Signature 描述结构定义

### 3.3　安全属性描述实例

为了进一步说明图 1 所示模型以及图 2、图 3 所描述结构的使用方法，本节将给出一个针对性的实例。定义结构化文档的逻辑结构和样式结构描述的完整性标签，采用 MD5 算法计算消息摘要，RSA 算法生成签名，证书采用 X.509 格式，签名生成时间为当前系统时间。对应的安全属性描述文件 Security_1.xml 如下。

```
<?xml version="1.0" encoding="UTF-8"?>
<Security ID="001">
…
<Signature ID="01" TimeStamp="2010-05-08 08:08:08">
<DigestMethod>MD5</DigestMethod>
<DigestValue>…</DigestValue>
<SignatureMethod>RSA</SignatureMethod>
<SignatureValue>…</SignatureValue>
<CertificationType>X.509</CertificationType>
<CertificationData>…</CertificationData>
</Signature>
</Security>
```

针对该结构化文档的访问控制需求，例如，该文档的访问时间是上午 8 点到下午 5 点，地点为公司内部，可以被安全级别 3 及以上级别部门 A 的人员进行修改操作。Domain 标签将记录该文档在跨域传递过程中经由安全域的信息，如 ID、网络位置等内容。具体描述文件 Security_2.xml 如下。

```
<?xml version="1.0" encoding="UTF-8"?>
<Security ID="001">
…
<AccessControlLabel>
<AdminPassword>…</AdminPassword>
<UserPassword>…</UserPassword>
<RightsObject>
    <Permission>
        Modify
    </Permission>
</RightsObject>
<Domain>
    …
</Domain>
<Temporal>
    <Start>08:00:00</Start>
    <End>17:00:00</End>
</Temporal>
<Environment>In Company</Environment>
<Level>3</Level>
<Scope>Department A</Scope>
</AccessControlLabel>
…
</Security>
```

## 4　安全性分析

### 4.1　完整性

结构化文档安全属性描述模型为结构化文档、子文档及其对象定义了安全属性标签，包含了完整性标记，能够保证逻辑结构描述、样式结构描述以及资源和数据在网络传输过程中的完整性和不可篡改性。

### 4.2　机密性

该模型支持为文档及其描述文件和资源的加密，可以定义对应的加解密算法、工作模式、密钥以及初始化向量。能够保证在文档传输和使用过程中，数据信息的机密性。

### 4.3　访问控制

安全属性描述中包含域属性、时态、环境属性，为用户描述访问时所处的物理环境、软硬件平台、时间状态等信息，并对结构化文档进行对象级的环境、时态约束。文档管理系统通过定义用户与结构化文档，添加主客体环境、时态标签，实现结构化文档的多要素访问控制，进一步适用于分布式计算、云计算、泛在计算等复杂网络环境。

### 4.4　多级安全管理

安全属性描述中包含的安全级别和访问范畴

能够约束主客体的安全级别及所属范围，针对不同的安全级别设置不同的访问规则及其操作类型，从而对结构化文档实现多级安全管理。

## 5　结束语

分布式计算、移动计算、云计算以及泛在计算的出现推动了信息化社会的发展，结构化文档作为一种融合了版式和流式信息的表现形式，在网络信息的传播中扮演了重要的角色。但是，网络环境的复杂特性为结构化文档的访问控制带来了新的挑战，不同的网络环境、物理位置、用户角色、时间状态等使得传统的访问控制方式不能够适用于多样化环境下的结构化文档管理。而且，多级安全的出现使得结构化文档的描述日趋复杂。因此，需要一种结合多种访问要素、具有多级安全特征的结构化文档描述方法。本文通过对传统结构化文档描述模型的研究，结合访问控制和多级安全需求，提出了一种面向多级安全的结构化文档描述模型定义和描述方法，定义了安全属性的描述结构，并给出了相应的 XML 描述实例。该模型能够解决结构化文档在网络跨域流转过程中逻辑结构描述、样式结构描述以及资源数据的完整性和机密性问题，保证结构化文档的合理、合法使用。

## 参考文献：


[1] KLINK S, DENGEL A, KIENINGER T. Document structure analysis based on layout and textual features[A]. Proceedings of the 4th IAPR International Workshop on Document Analysis Systems[C]. Rio de Janeiro, Brazil. 2000. 99 - 111.

[2] The future of multi-level secure (MLS) information systems[EB/OL]. http://csrc.nist.gov/nissc/1998/ proceedings /panelF3.pdf, 1998.

[3] BELL D E. Looking Back at the Bell-LaPadula model[A]. Proceedings of the 21st Conference On Annual Computer Security Applications[C]. Washington, DC, USA, 200.337-351.

[4] BIBA K J. Integrity Considerations for Secure Computer Systems[R]. MTR-3153, The Mitre Corporation, 1977, 04.

[5] 李宁, 田英爱, 侯霞等. 办公文档与固定版式文档格式关系探讨[J]. 电子学报, 2008, 36(B12): 128-132.
LI N, TIAN A Y, HOU X, *et al*. A discussion on relationship between revisable and non-revisable document formats[J]. Acta Electronica Sinica, 2008, 36(B12): 128-132.

[6] Microsoft Corporation. XPS Specification and Reference Guide[S]. 2010, 06, 30.

[7] BLOECHLE J L, RIGAMONTI M, HADJAR K, *et al*. Xcdf: a canonical and structured document format[A]. Proceedings of the 7th International Workshop on Document Analysis Systems[C]. Nelson, New Zealand, 2006. 141 - 152.

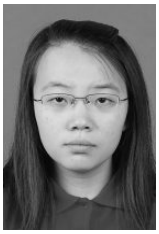
[8] BLOECHLE J L, PUGIN C, INGOLD R. Dolores: an interactive and class-free approach for document logical restructuring[A]. Proceedings of the 8th International Workshop on Document Analysis Systems[C]. Nara, Japan, 2008. 644 - 652.

[9] BLOECHLE J L, LALANNE D, INGOLD R. OCD: an optimized and canonical document format[A]. Proceedings of the 10th International Conference on Document Analysis and Recognition[C]. Barcelona, USA, 2009. 236 - 240.

[10] CEBX/Mv1.2 Standard Manual[S]. 2011.8.

[11] 李凤华, 王巍, 马建峰等. 基于行为的访问控制模型及其行为管理[J]. 电子学报, 2008, 10, 36(10): 1881-1890.
LI F H, WANG W, MA J F, *et al*. Access control model and administration of action[J].Acta Electronica Sinica, 2008, 10, 36(10): 1881-1890.


## 作者简介：

**苏铓**（1987-），女，内蒙古赤峰人，西安电子科技大学博士生，主要研究方向为访问控制与网络安全。

**李凤华**（1966-），男，湖北浠水人，博士，中国科学院信息工程研究所研究员、博士生导师，科技处处长，主要研究方向为网络安全与可信计算。

**史国振**（1974-），男，河南济源人，博士，北京电子科技学院副教授，主要研究方向为信息安全和系统软件设计。

**李莉**（1974-），女，山东青岛人，硕士，北京电子科技学院副教授，主要研究方向为嵌入式系统。