基于双重存证的跨域流转多副本发现机制

罗海洋[1,2,3]，邝彬[1,2,3]，郭守坤[1,3]，张玲翠[1,3]，牛犇[1,3]，李凤华[1,2,3]

（1. 中国科学院信息工程研究所，北京 100085；2. 中国科学院大学网络空间安全学院，北京 100049；
3. 网络空间安全防御重点实验室，北京 100085）

摘 要：针对泛在共享环境下信息频繁跨节点、跨系统交换时有意或无意留存而导致的隐私信息泄露问题，提出了一种基于双重存证的跨域流转多副本发现机制，可实现对流转信息的传播路径、传播方式进行溯源，并构造信息的多副本传播图。根据存证时机和方式的不同，双重存证包括流转主动存证和操作被动存证，信息在被分享前，由信息分享者主动记录传播路径和传播方式，生成流转主动存证记录；信息在被操作前，由系统自动记录传播路径，生成操作被动存证记录；相比单一存证，双重存证能够提高构造的信息多副本传播图的完整性和真实性，能够发现存证行为异常的节点并进行处置；基于社会惩戒理论，证明了存证行为异常发现与处置的有效性。实验开发了针对 OFD 的双重存证多副本发现原型系统，验证了所提机制对信息传播图构造完整性的提升。

关键词：多副本发现；存证系统；跨域流转；社会惩戒；传播图

中图分类号：TN92

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2024023

Cross-domain multi-copy of flow discovery mechanism based on dual certificate storage

LUO Haiyang[1,2,3], KUANG Bin[1,2,3], GUO Shoukun[1,3], ZHANG Lingcui[1,3], NIU Ben[1,3], LI Fenghua[1,2,3]

1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
3. Key Laboratory of Cyberspace Security Defense, Beijing 100085, China

Abstract: To solve the problems of the privacy information leakage caused by the deliberate or inadvertent retention of information when information was frequently exchanged across nodes and systems in a ubiquitous sharing environment, a cross-domain multi-copy of flow discovery mechanism based on dual certificate storage was proposed, which could trace the propagation path and channel, and construct a multi-copy propagation graph of the information. Depending on the timing and method of certification, the dual certification was comprised active circulation certification and passive operation certification. Before the information was shared, the information sharer actively recorded the propagation path and method to generate active circulation certification records. Before the information was operated, the system automatically recorded the propagation path to generate passive operational certification records. Compared with single certificate storage, the dual certificate storage could improve the integrity and authenticity of the constructed multi-copy propagation graph of information, and could detect nodes with abnormal certificate storage behavior and provide disposals. Based on the theory of social punishment, the effectiveness of abnormal certificate storage behavior detection and handling was demonstrated. A prototype system for multi-copy discovery of OFD with dual certificate storage is developed, the improvement of information dissemination graph construction integrity by the proposed mechanism is verified.

Keywords: multi-copy discovery, certificate storage system, cross-domain flow, social punishment, propagation graph

收稿日期：2023−11−02；**修回日期：**2023−12−13

通信作者：郭守坤，guoshoukun@iie.ac.cn

基金项目：国家重点研发计划基金资助项目（No.2021YFB3101301）；国家自然科学基金资助项目（No.62332018, No.61932015）；国家社科基金重大项目（No.22&ZD147）

Foundation Items: The National Key Research and Development Program of China (No.2021YFB3101301), The National Natural Science Foundation of China (No.62332018, No.61932015), Major Programs of the National Social Science Foundation of China (No.22&ZD147)

0　引言

信息技术的不断发展和广泛应用，促进了“万物智慧互联、信息泛在共享”。借助因特网架构以及第五代移动通信网络、物联网、云计算、卫星通信等信息服务和基础设施的不断升级与改进，信息传输的速度更快、覆盖更广、互联更强、渠道更宽，使信息的传播和分享变得更加便捷高效，信息在不同系统、不同设备和不同主体之间的流转和留存也越来越频繁。信息技术的升级不仅提高了信息传播的效率，也拓展了信息传播的范围，信息通过各种渠道，如社交媒体、电子邮件、U 盘和蓝牙等，以前所未有的速度和广度传播。

然而，一旦信息脱离信息主体的管控范围，信息主体便失去了对信息的控制能力，信息可以通过各种渠道进行流转。通常情况下，信息主体很难有效地控制、追踪和溯源信息的流转轨迹。随着信息的频繁流转，信息会在多个信息系统或者节点间有意或者无意的留存，这进一步增大了隐私泄露的潜在风险，一旦这些信息被非法滥用或者泄露，就会严重威胁信息主体的隐私安全。近年来，数据留存滥用事件频繁发生，隐私信息泄露的案例也不断出现。因此，在加强个人隐私保护意识、制定相关隐私保护政策和法律法规的同时，还需要依赖先进的隐私保护技术作为支撑，确保信息在传输和存储过程中的隐私安全得到充分保障。

针对泛在共享环境下信息的隐私保护，近年来，有学者提出隐私计算的概念[1]，从计算角度给出了信息全生命周期权益控制和隐私保护的体系框架，将隐私保护的研究推向一个新高度。隐私计算通过信息流转前的传播控制和流转后的延伸控制，在信息流转的全流程中为信息保驾护航。

信息流转多副本发现是信息延伸控制的重要基础和关键环节。通过信息多副本发现，能够溯源信息的流转轨迹，发现保存有信息或者信息副本的节点。基于信息多副本发现的结果，能够在信息被分享后，为信息追加延伸控制策略，实现信息全生命流程的受控共享；能够对信息流转节点的安全性进行分析，及时发现隐私泄露风险，并采取相应保护措施。此外，信息流转多副本发现在传播分析、舆情评估、个性化推荐等方面也能发挥积极作用。

信息传播渠道的多样性和信息系统的多元性，给信息多副本发现带来了严峻的挑战。用户从一个平台下载信息，并通过在线传输（社交网络、电子邮件等）或离线传输的方式（U 盘、蓝牙等）将信息传输给另一个用户，已逐渐成为互联网中的一种常态。信息传输渠道的多样性也增大了对信息的传输路径进行监管、溯源的难度。目前对于信息多副本的研究主要基于单个系统内副本的高效管理与检索[2-5]，并不适用于动态流转场景下的信息多副本查找。

本文针对泛在共享环境下信息的跨域流转，提出了一种基于双重存证的信息多副本发现机制。具体地，在信息分享前，信息分享者通过主动填报信息的流转路径和传播渠道，生成信息的流转主动存证记录；在信息分享后，在信息接收者操作信息前，系统自动获取信息的流转路径，生成信息的操作被动存证记录。通过对流转主动存证记录和操作被动存证记录进行融合分析，得到信息的多副本传播图，并发现主动存证行为异常的节点。基于社会惩戒理论，对发现的异常节点进行处置，并证明了处置过程的有效性。本文主要的贡献点如下。

1) 提出了一种基于双重存证的信息多副本发现机制，适用于信息跨系统多渠道流转的复杂场景。该机制结合流转主动存证和操作被动存证进行信息流转多副本查找，提高了信息多副本传播图构造的完整性、丰富性和真实性。该机制还能够快速识别存证异常行为并定位存证异常节点。

2) 提出了基于社会惩戒理论异常节点处置方法。将社会工程学领域的社会惩戒原理应用于信息流转的异常节点处置，通过反馈机制，提高异常节点的主动存证意识，进一步提高了信息多副本图构造的完整性和真实性。

3) 开发了针对 OFD（open fixed-layout document）的双重存证多副本发现原型系统，并在不同规模的数据集上测试了所提方案的有效性和效率。实验表明，相比于单一存证，双重存证结合社会惩戒措施，能够大幅提高获得的信息多副本传播图的完整性。

1　相关工作

本节从信息副本管理与信息多副本发现 2 个方面对现有相关解决方案进行梳理。

1.1　信息副本管理

在信息的本地副本管理方面，Curtmola 等[2]首

先提出了一种可证明安全的、用于记录多副本信息的目录结构。熊金波等[3]基于密码学技术，将数据封装成副本关联对象，并利用副本关联对象建立副本关联模型管理副本并实现关联删除。Du 等[4]通过将盲 RSA 加密的数据密钥外包给第三方，提出了一种基于预删除序列和 Merkle 哈希树的多副本关联数据删除方案。Zhang 等[5]引入副本关联模型，提出了一种云环境下的多副本关联删除方案（MADS, multi-replica associated deleting scheme），MADS 由数据存储算法、副本生成算法、副本删除和反馈算法组成。为了高效安全地对数据副本进行动态管理，Miao 等[6]设计了多副本关联树，并提出了一种多副本场景下的云数据完备删除方案。

但是，上述研究方案都基于密码学相关技术，只适用于单一系统的信息副本发现与管理，无法满足泛在共享环境下的信息多副本发现需求。

1.2 信息多副本发现

Phan 等[7]提出了一种能够跨社交平台进行图片传播路径还原的方案。该方案利用不同社交平台图片压缩算法的差异性，并借助深度学习技术，能够高准确率地还原图片的历史传播平台和路径。You 等[8]将图片在社交平台上的传播路径视为共享历史的时间序列，并采用了基于序列到序列的模型，在性能方面取得了显著提升。类似地，Siddiqui 等[9]利用基于深度学习技术检测社交网络中的图像来源。然而，该方案虽然能定位到社交平台，但无法精准定位到具体的用户，目前也仅能识别最多在社交平台之间传播三次的传播链，并且只能应用于特定格式的信息。

李凤华等[10]提出了一种利用加密方式将传播记录写入图片的可交换图像文件格式（EXIF）字段的方案，以实现对图片在跨社交网络分享时传播路径的溯源；采用了嵌套签名算法以确保传播记录的真实性和不可篡改性。然而，该方案虽然能还原特定图片的传播链，但无法构造完整的图片传播图。另外，很多压缩算法会抹掉图片的 EXIF 信息，这也是该方案的局限之一。

此外，Zhang 等[11]提出了一种基于区块链的跨社交网络图片分享隐私保护框架，基于智能合约，生成分享图片的跨平台传播树。Tang 等[12]通过使用社交特征和内容特征，提出了一种在博客上进行信息流检测和跟踪的方法。Patsakis 等[13]提出了一种基于数字水印的方案，能够在多个在线社交网络之间执行用户的隐私策略，用户还能够追踪其媒体文件的使用情况。

上述信息流转多副本发现方案局限于特定的信息格式或传播平台，无法满足信息的多模态、传播的多渠道、平台的多元化时的应用需求。

还有一些研究通过信息传播建模的方式对信息的传播路径进行预测，从而发现可能保存有信息的节点，例如基于传染病模型[14-15]的信息传播预测。Guille 等[16]提出了 T-BaSIC 模型，该模型考虑了用户之间的社交关系、话题内容和时间 3 个维度，估计信息在不同节点间传播的概率。Dickens 等[17]基于独立级联模型预测 Twitter 中信息的转发路径。Liu 等[18]考虑网络动态演化与信息传播之间的关系，并对传播过程进行了理论与仿真分析。

但是，此类方案更多应用于谣言溯源、信息影响力评估等领域，只能从概率角度给出可能保存有信息多副本的节点，结果具有可抵赖性。

2 预备知识

2.1 社会惩戒

社会惩戒是社会心理学领域的一种行为调整理论，是指行为方在受到负面评价、警告或惩罚后，会自觉对相关行为进行调整从而避免再次受到负面反馈的现象。这种现象通常发生在个体的行为违反了法律法规、社会规范、道德标准或群体共识并被发现的情况下。社会惩戒的最终目的是促使行为方改善其不良行为，共同维护良好秩序。本文对社会惩戒理论进行简单建模，如图 1 所示。

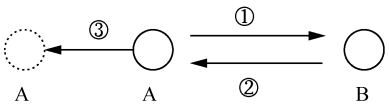

图 1 社会惩戒模型

在图 1 中，节点 A 和 B 可以表示个人、群体或者组织等，社会惩戒过程主要分为 3 个阶段。

① 行为违规。A 的行为违反了法律法规、社会规则或者行业规范等。

② 负面反馈启示。当 A 的异常行为被 B 发现并被警告或处置时，构成了负面反馈。负面反馈会激发 A 的负面情绪，使 A 认识到自己的行为不被接纳或不符合社会规范。

③ 自我调整和学习。作为回应，A 可能会进行自我调整，改善自己的行为，试图避免再次因其

行为得到负面反馈。

在整个过程中，A 获得了一个经验教训，即异常行为可能会带来负面后果，从而减少了异常行为出现的频率；或者，A 会以更小的风险来执行某种行为。

社会惩戒是建立在社会共识基础上的一种机制，通过负面反馈、自我调整和学习等方式，激励个体自觉规范其行为，以避免再次得到负面反馈。这种现象反映了社会反应对个体行为的影响。

2.2　信息多副本传播图

信息多副本传播图是指信息在设备、网络或其他媒介中传播的具体路径、步骤和方式。在信息多副本传播图中，主要包括以下 3 种节点。

1) 信息所有节点：第一个发布该信息的节点，是信息多副本传播图的起始节点。

2) 信息转发节点：接收到信息并将其转发的节点，是信息多副本传播图的中继节点。

3) 静默的信息接收节点：接收到信息且无转发操作的节点，是信息多副本传播图的终止节点。

将以上 3 种角色结合传播媒介、传播时间等其他因素，可以构成一个完整的信息多副本传播图。信息多副本传播图示例如图 2 所示。在信息的流转过程中，对信息的流转方式不进行限制，信息可以通过蓝牙、微信和电子邮件等方式，在不同应用域的不同节点中流转和留存。

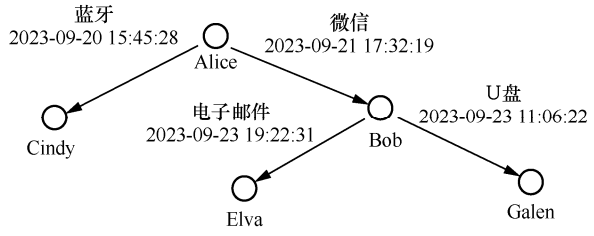

图 2　信息多副本传播图示例

2.3　存证系统

存证系统是一种基于技术手段来确保数据或记录的真实性、完整性和不可篡改性的证据存储系统。它通过密码学、时间戳、区块链和数字签名等技术，记录和存储特定数据或事件在某个时间点的具体状态，以便在后续进行审计和追溯。

存证系统的主要作用是提供可信、不可变的证据，用于验证或还原特定信息或事件的历史状态、演变趋势等关键信息。

按照存证生成主体的不同，存证可以分为主动存证和被动存证。主动存证指的是用户或者节点主动去记录关键信息（如标识码、时间戳等）从而生成存证记录，主要依赖于存证生成主体的主动存证意识，用于确保存证记录的完整性和真实性；被动存证指的是信息系统通过特定应用程序，对数据的状态进行监控，在满足特定状态时自动生成存证记录，被动存证又称强制性存证，主要用于确保存证记录的可信度和不可篡改性。

3　基于双重存证的跨域流转多副本发现

3.1　系统架构和流程设计

本节介绍了基于双重存证的跨域流转多副本发现机制的系统架构和工作流程设计，其系统架构如图 3 所示，按照功能的不同，可以分为应用域和监管域。系统参数如表 1 所示。

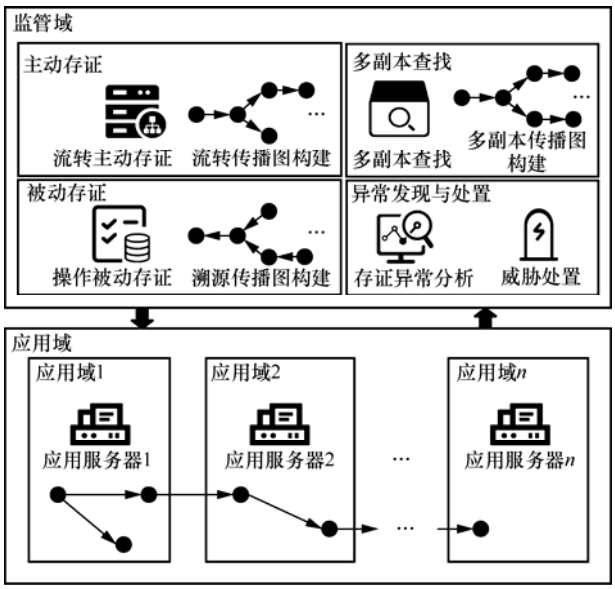

图 3　系统架构

应用域主要负责数据或信息生成、流通和使用，可以根据行政位置、服务厂商和服务类型等因素，将应用域划分为不同的应用子域，每个应用子域都有自己独立的数据服务器，并且应用子域之间服务数据不共享，但是信息可以通过多种途径在不同的应用子域之间进行流传和留存。

以社交软件为例，微信、微博和 QQ 都可以看作一个社交软件域下不同的应用子域，每个子域有其特定的功能、用户群体和运营模式，应用数据相互独立。但是各个应用子域里的图片、文本、文件等信息，可以通过用户在不同的应用子域间进行流转。例如，用户在微信上下载文档保存到本地后，

再将其上传至微博。信息在不同应用子域之间流转，构成了信息的跨域流转多副本传播图。

表 1 系统参数

参数	含义
X	信息
S	信息分享者
R	信息接收者
C	信息传播途径
A	流转主动存证记录
O	操作被动存证记录
G_{Prop}	基于存证系统构建的信息多副本传播图
G'_{Prop}	完整的信息多副本传播图
G_{Flow}	基于流转主动存证构建的流转传播图
G_{Trace}	基于操作被动存证构建的溯源传播图
$G.V=\{v_1,v_2,\cdots,v_n\}$	图 G 的节点集合 $G.V$ 及其节点 v
$G.E=\{e_1,e_2,\cdots,e_m\}$	图 G 的边集合 $G.E$ 及其边 e
V'	异常节点集合
E'	异常路径集合
$X_{\text{ID}}=\text{GenerateID}(X)$	信息标识符生成函数，生成 X 的信息标识 X_{ID}
$(X_{\text{ID}}, S_{\text{ID}})=\text{Extract}(X)$	元数据提取函数，提取 X 的信息标识 X_{ID} 和最近操作用户标识 S_{ID}

监管域主要负责对应用域行为进行监督、管理和规范，可以由监管机构、网络服务提供商、行业或者用户自行建立。在本文中，监管域主要负责存证记录的存储、检索以及信息多副本传播图的构建，确保存证记录的完备性和信息传播的可追溯性。监管域主要由以下 4 个部分组成。

1) 主动存证：通过用户主动存证，记录信息的传播路径与传播方式，并根据主动存证记录构建信息的流转传播图。

2) 被动存证：通过系统被动存证，记录信息的溯源路径，并根据被动存证记录构建信息的溯源传播图。

3) 多副本查找：通过对存证记录进行融合分析和交叉验证，维护信息的多副本传播图，确保信息的可追溯性和完整性。

4) 异常发现与处置：监控和检测存证数据的异常情况，及时定位到异常节点并采取相应的处置措施。

应用域中的节点通过主动或者被动的方式向监管域上报存证记录，监管域通过存证系统对信息的传播路径进行还原，构建对应的信息多副本传播图，并对应用域中节点的存证行为进行监督和反馈。

3.2 双重存证算法

本节对双重存证算法的细节进行详细的介绍。

在本文系统中，双重存证算法包括流转主动存证和操作被动存证，如图 4 所示。流转主动存证是信息分享者在分享信息前，主动填报上传的存证记录；操作被动存证是信息接收者在打开、浏览、使用信息前，由系统自动生成并上传的存证记录。

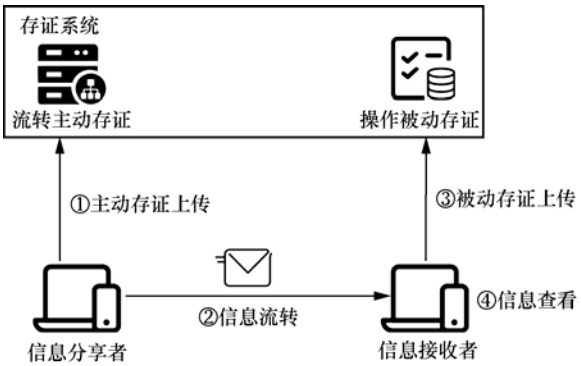

图 4 双重存证算法

流转主动存证从信息分享者的角度出发，在分享信息前，由信息分享者主动记录并上报信息的流转状态。流转主动存证记录 A 的形式化描述为

$$A=(X_{\text{ID}}, S_{\text{ID}}, R_{\text{ID}}, C) \tag{1}$$

其中，X_{ID} 表示信息标识，是信息在特定场景下具有唯一性的标识符或标记，信息标识能够在系统中快速准确地识别和区分不同的信息实体，以便对其进行操作、管理或追踪；S_{ID} 和 R_{ID} 分别表示信息分享者和信息接收者的用户标识，用户标识用来唯一表征用户的身份，由应用域编码和用户编码构成，通过对用户标识的解析，能够定位到特定应用子域中的特定用户；C 表示信息传播途径，在本文系统中，对信息的传播途径不进行限制，信息可以通过蓝牙、微信、电子邮件和 U 盘复制等方式在单一域内或者不同域间进行传播，信息分享者在进行主动存证时，需要对信息的传播途径进行记录。

信息标识依据信息的模态、存储形式等，以不同的方式嵌入信息中。例如，针对图片信息，将信息标识嵌入 EXIF 元数据中的 IFD 结构中，或者通过图片隐写、图片水印等方式嵌入；针对

文档信息（PDF、Word、OFD 等），将信息标识嵌入文档的元数据字段中，或者通过批注、注释等方式嵌入。

用户在填报流转主动存证记录时，系统根据用户选中的待分享信息，自动提取并识别对应的信息标识。当系统无法在信息中提取到信息标识时，表明该信息在系统中首次出现，系统自动为该信息生成信息标识并嵌入。信息标识字段如图 5 所示。

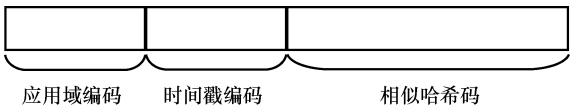

图 5　信息标识字段

应用域编码指的是信息在系统中首次出现时，对应的应用域以及应用子域的编码；时间戳编码指的是由信息在系统中首次出现的时间生成的编码；相似哈希码由相似哈希算法产生，相似哈希算法通过对相似输入产生相似输出，能够用于对信息相似性进行快速检测。在信息的传播图构造中，使用相似哈希码能够对不同源但相似的信息进行检测，并根据应用需求以及相似程度决定是否需要将其传播图合并。

相似哈希算法主要由 2 个步骤构成，其形式化描述如下。

1) 将输入信息 X 映射为特征向量 $\boldsymbol{V}$

$$\boldsymbol{V} = F(X) \tag{2}$$

2) 将特征向量 $\boldsymbol{V}$ 映射为哈希值 h_x

$$h_x = H(\boldsymbol{V}) \tag{3}$$

相似哈希算法中的特征提取函数 F 和相似哈希函数 H 可以由使用者根据实际需求自行选择。对于如何设计函数 F 和 H，当前已有大量研究[19-21]，不是本文的重点讨论内容。

操作被动存证从信息接收者的角度出发，在接收信息后，在应用子域中打开信息前，由系统自动记录并上报信息的溯源状态。操作被动存证记录 O 的形式化描述为

$$O = (X_{\mathrm{ID}}, S_{\mathrm{ID}}, R_{\mathrm{ID}}) \tag{4}$$

其中，X_{ID} 表示信息标识；S_{ID} 表示信息分享者的用户标识，也就是信息来源节点，系统会将最近对信息进行操作用户的标识嵌入信息中，跟随信息一起流转，通过这种方式，在生成操作被动存证时，系统能够获取信息的上一跳节点，并在当前用户打开信息时，更新信息中嵌入的最近操作用户标识；R_{ID} 表示信息接收者的用户标识。

当系统提取不到信息中的 X_{ID} 和 S_{ID} 时，表明信息在该系统中首次出现，系统生成该信息的标识 X_{ID} 并嵌入信息中，此时构造的操作被动存证记录 O 可以实例化为

$$O' = (X_{\mathrm{ID}}, \mathrm{null}, R_{\mathrm{ID}}) \tag{5}$$

其中，S_{ID} 置为 null 表示当前节点为信息的起始节点，即信息多副本传播图中的起始节点。

将主动存证和被动存证进行结合，得到双重存证算法的形式化描述，如算法 1 所示。

算法 1　双重存证算法

输入　当前用户标识 U_{ID}，信息 X，信息接收者标识 R_{ID}，信息传播途径 C

输出　主动存证记录 A 或被动存证记录 O

1) $(X_{\mathrm{ID}}, S_{\mathrm{ID}}) = \mathrm{Extract}(X)$
2) if $(S_{\mathrm{ID}} \neq U_{\mathrm{ID}})$
3) $\quad X \leftarrow X - S_{\mathrm{ID}}$
4) $\quad X \leftarrow X + U_{\mathrm{ID}}$
5) end if
6) if $(X_{\mathrm{ID}} == \mathrm{null})$
7) $\quad X_{\mathrm{ID}} = \mathrm{GenerateID}(X)$
8) $\quad X \leftarrow X + X_{\mathrm{ID}}$
9) end if
10) if $(U_{\mathrm{ID}}$ share $X)$ then
11) $\quad A \leftarrow (X_{\mathrm{ID}}, U_{\mathrm{ID}}, R_{\mathrm{ID}}, C)$
12) $\quad$ return A
13) end if
14) if $(U_{\mathrm{ID}}$ operate $X)$ then
15) $\quad O = (X_{\mathrm{ID}}, S_{\mathrm{ID}}, U_{\mathrm{ID}})$
16) $\quad$ return O
17) end if

在算法 1 中，函数 $\mathrm{Extract}(X)$ 表示提取信息 X 中嵌入的信息标识 X_{ID} 和上一跳节点标识 S_{ID}；函数 $\mathrm{GenerateID}(X)$ 表示根据信息 X，构造特定格式的信息标识 X_{ID}。

3.3　信息多副本发现算法

本节介绍基于双重存证还原信息多副本传播图的具体细节。

图 G 的表示方式如式(6)所示。

$$G=(V,\ E,\ W,\ \text{Attr}) \tag{6}$$

其中，V 表示图 G 的顶点集合，如式(7)所示；E 表示图 G 的边集合，如式(8)所示；W 表示图 G 的权重集合；Attr 表示图 G 的属性集合。

$$V=\{v_1,v_2,\cdots,v_n\} \tag{7}$$

其中，n 为顶点的数量，每个顶点 v_i 代表一个用户，可以用二元组 $v_i=(u_i,\ \text{attr}_i)$ 表示，u_i 是顶点的唯一标识，attr_i 是顶点的属性信息集合。

$$E=\{e_1,e_2,\cdots,e_m\} \tag{8}$$

其中，m 为边的数量，每条边 e_i 代表信息的一条传播路径，用三元组 $e_i=(v_{\text{start}},\ v_{\text{end}},\ \text{attr}_i)$ 表示，v_{start} 表示边的起始节点，v_{end} 表示边的结束节点，attr_i 是边的属性信息。

基于本文的实际应用场景和需求，将图 G 实例化为信息 X_{ID} 的流转传播图 $G_{\text{Flow}}(X_{\text{ID}})$ 和溯源传播图 $G_{\text{Trace}}(X_{\text{ID}})$，节点集合 V 可以表示为边中出现的用户标识 U_{ID} 的集合，也就是保存有信息多副本的用户的集合；边集合 E 的具体定义如式(9)和式(10)所示。

$$G_{\text{Flow}}(X_{\text{ID}}).E=\{(S_{\text{ID}},\ R_{\text{ID}},\ \langle X_{\text{ID}},\ C\rangle),\ \cdots\} \tag{9}$$

$$G_{\text{Trace}}(X_{\text{ID}}).E=\{(S_{\text{ID}},\ R_{\text{ID}},\ \langle X_{\text{ID}}\rangle),\ \cdots\} \tag{10}$$

构建信息标识为 X_{ID} 的信息的多副本传播图 $G_{\text{Prop}}(X_{\text{ID}})$ 的步骤如下。

1) 通过查询主动存证层的流转主动存证记录，构建 X_{ID} 的流转传播图 $G_{\text{Flow}}(X_{\text{ID}})$。

$$G_{\text{Flow}}(X_{\text{ID}})=\text{Query}(A,\ X_{\text{ID}}) \tag{11}$$

2) 通过查询被动存证层的操作被动存证记录，构建 X_{ID} 的溯源传播图 $G_{\text{Trace}}(X_{\text{ID}})$。

$$G_{\text{Trace}}(X_{\text{ID}})=\text{Query}(O,\ X_{\text{ID}}) \tag{12}$$

3) 合并顶点集合，得到完整的保存有信息多副本的用户集合。

$$G_{\text{Prop}}(X_{\text{ID}}).V=G_{\text{Flow}}(X_{\text{ID}}).V\cup G_{\text{Trace}}(X_{\text{ID}}).V \tag{13}$$

4) 合并边集合，得到信息多副本传播路径。

$$G_{\text{Prop}}(X_{\text{ID}}).E=G_{\text{Flow}}(X_{\text{ID}}).E\bigcup G_{\text{Trace}}(X_{\text{ID}}).E \tag{14}$$

5) 得到 X_{ID} 的多副本传播图 $G_{\text{Prop}}(X_{\text{ID}})$。

相比于单一存证系统，基于双重存证构造的 $G_{\text{Prop}}(X_{\text{ID}})$ 能够更加完整、详细地记录信息多副本的流转信息。

3.4 存证异常行为发现与处置算法

本节介绍基于双重存证的存证异常行为发现与处置算法的具体细节。

存证异常行为发现的主要目的是发现不主动或者不诚实上报流转主动存证记录的节点。

为了更好地描述异常节点的发现过程，定义命题 P 与命题 Q。

$$\text{P}:\ \forall v_i\in V,\ v_i\ \text{is honest} \tag{15}$$

$$\text{Q}:\ G_{\text{Trace}}(X_{ID}).V\subseteq G_{\text{Flow}}(X_{\text{ID}}).V\ \&\&$$
$$G_{\text{Trace}}(X_{\text{ID}}).E\subseteq G_{\text{Flow}}(X_{\text{ID}}).E \tag{16}$$

在双重存证中，所有存证节点都诚实的情况下，信息流转传播图即完整的信息多副本传播图，而信息溯源传播图为信息流转传播图的子图。因为可能存在一些信息接收者没有对文档进行操作，或者文档在传输、存储过程中发生了丢失，产生的被动存证只记录了部分传播路径。但是被动存证记录中包含的每条路径，诚实的信息分享者在分享信息前都通过主动存证记录上报过，所以生成的信息溯源流转图为信息传播流转图的子图。命题 P 与 Q 的关系如式(17)所示。

$$\text{P}\Rightarrow\text{Q} \tag{17}$$

其逆否命题为

$$\neg\text{Q}\Rightarrow\neg\text{P} \tag{18}$$

基于式(18)，得到节点存证异常行为发现算法，如算法 2 所示。

算法 2 存证异常行为发现算法

输入 流转传播图 $G_{\text{Flow}}(X_{\text{ID}})$，溯源传播图 $G_{\text{Trace}}(X_{\text{ID}})$

输出 异常节点集合 V'，异常路径集合 E'

```
1)  V' ← ∅, E' ← ∅
2)  for each e in G_Trace(X_ID).E
3)      if(e ∉ G_Flow(X_ID).E)
4)          V' ← V' ⋃ {e.v_start}
5)          E' ← E' ⋃ {e}
6)      end if
7)  end for
8)  return V', E'
```

在得到异常节点集合 V' 后，向 V' 中的每一个异常节点发送警告信息，告知节点其存证异常行为已经被系统发现，并要求其对存证行为进行自我调

整。对于多次出现异常行为且拒绝调整的用户，系统将进行广播通报，最极端的情况下，系统将强制注销该用户。

3.5 算法有效性分析

3.5.1 信息多副本发现算法有效性分析

流转主动存证主要依赖用户的主动存证意识，但不是所有用户都会意识到存证的重要性或愿意主动存证，存在用户遗忘、疏忽或错误的可能性，导致存证的不及时、不全面或不准确。若用户未及时进行主动存证，后续多副本查找时无法获得完整的信息传播路径。若存在恶意用户，故意上传虚假、伪造的存证记录，后续多副本查找时会出现虚假路径，破坏存证的真实性和可信度。

操作被动存证只在信息被操作时触发记录，所以任何未触发操作的信息流转都不会被记录，这可能导致记录的延迟，从而难以完整地追溯信息的历史传播记录，构造出完整的信息多副本传播图。当信息接收者收到信息后，在多副本查找时还未对信息进行操作时，该传播路径和信息副本留存节点无法被及时发现。并且，由于信息传播方式的多样性和复杂性，仅依靠操作被动存证，无法获取信息的来源途径。

双重存证通过将操作被动存证和流转主动存证结合，克服了单一存证的缺点。通过流转主动存证，能够得到信息较完整的传播路径和具体的传播方式；通过操作被动存证，能够对流转主动存证记录的真实性进行验证。通过双重存证，能够提高还原出的信息多副本传播图的完整性、真实性和可信度。

以图 $G^{'}$ 为参照，图 G 的完整性可以表示为

$$C(G|G')(0<C(G|G')<1) \tag{19}$$

由式(19)可知，相对于完整的信息多副本传播图 $G_{\text{Prop}}^{'}$，流转主动存证构造的流转传播图 G_{Flow} 的完整性表示为 $C(G_{\text{Flow}}|G_{\text{Prop}}^{'})$，操作被动存证构造的溯源传播图 G_{Trace} 的完整性表示为 $C(G_{\text{Trace}}|G_{\text{Prop}}^{'})$，由存证记录构造的信息多副本传播图 G_{Prop} 的完整性表示为 $C(G_{\text{Prop}}|G_{\text{Prop}}^{'})$。参考信息论中互信息的定义，可以得到式(20)。

$$\begin{aligned} &C(G_{\text{Prop}}|G_{\text{Prop}}^{'})= \\ &C(G_{\text{Trace}}|G_{\text{Prop}}^{'})+C(G_{\text{Flow}}|G_{\text{Prop}}^{'})- \\ &I(G_{\text{Trace}}|G_{\text{Prop}}^{'},\ G_{\text{Flow}}|G_{\text{Prop}}^{'}) \end{aligned} \tag{20}$$

其中，$I(G_{\text{Trace}}|G_{\text{Prop}}^{'},\ G_{\text{Flow}}|G_{\text{Prop}}^{'})$ 表示 G_{Flow} 与 G_{Trace} 之间基于 $G_{\text{Prop}}^{'}$ 的相关程度，其取值范围为 $[0,\ \min(C(G_{\text{Flow}}|G_{\text{Prop}}^{'}), C(G_{\text{Trace}}|G_{\text{Prop}}^{'}))]$。由式(20)可知，通过结合 G_{Flow} 和 G_{Trace} 构造出的信息多副本传播图 G_{Prop} 的完整性不小于 G_{Flow} 或 G_{Trace}。

3.5.2 存证异常行为发现与处置算法有效性分析

基于社会惩戒理论，向存证异常行为节点发送系统通知对其存证行为进行警告后，节点会对自身的存证行为进行调整，这种调整会在一定程度上提高节点存证行为的诚实度，从而提高得到的信息多副本传播图的完整性。接下来基于社会惩戒理论，对存证异常行为发现与构造的信息多副本传播图的完整性之间的关系进行具体分析。

在信息多副本传播图 G_{Prop} 中，用节点 v_i 的诚实度 $v_i.h$ 表示 v_i 及时、正确地进行主动存证的概率，并且每一个节点的诚实度是相互独立的，不会受到其他节点诚实度的影响。

当需要还原出完整的 G_{Prop} 时，G_{Prop} 中每一条边 e_i 被主动存证的概率如式(21)所示。

$$P(e_i)=e_i.v_{\text{start}}.h \tag{21}$$

而用户的每一次主动存证行为，不受到其他用户主动存证行为或者自身之前主动存证行为的影响，因此，主动存证行为之间是相互独立的。由此得到通过主动存证还原出完整的 G_{Prop} 的概率 $P(e_1,\ e_2,\cdots,\ e_m)$ 如式(22)所示。

$$P(e_1,\ e_2,\cdots,\ e_m)=\prod_{i=1}^{m}P(e_i) \tag{22}$$

结合式(21)和式(22)，可以得到式(23)。

$$P(e_1,\ e_2,\cdots,\ e_m)=\prod_{i=1}^{m}e_i.v_{\text{start}}.h \tag{23}$$

利用图论相关知识，可以得到式(24)。

$$P(e_1,\ e_2,\cdots,\ e_m)=\prod_{i=1}^{n}v_i.h^{d_{\text{out}}(v_i)} \tag{24}$$

其中，$d_{\text{out}}(v_i)$ 表示节点 v_i 的出度，即从该节点出发的边的数量，在 G_{Prop} 中表示节点上报主动存证记录的数量。

当某个节点 v_k 由于不诚实的存证行为收到系统警告时，基于社会惩戒理论，v_k 会在一定程度上改善自己的行为，提高其诚实度，假设 v_k 诚实度的

变化为 $\Delta k(\Delta k>0)$。

由于 v_k 收到系统警告，v_k 必定有存证行为，否则 v_k 不会成为存证异常节点，此时 $d_{\text{out}}(v_i)>0$。在此条件下，通过主动存证还原出完整的 G_{Prop} 的概率表示为 $P^{'}(e_1, e_2,\cdots, e_m)$。

$$\frac{P^{'}(e_1, e_2,\cdots, e_m)}{P(e_1, e_2,\cdots, e_m)}=\frac{v_1.h^{d_{\text{out}}(v_1)}\cdots v_k.h+\Delta k^{d_{\text{out}}(v_k)}\cdots v_n.h^{d_{\text{out}}(v_n)}}{v_1.h^{d_{\text{out}}(v_1)}\cdots v_k.h^{d_{\text{out}}(v_k)}\cdots v_n.h^{d_{\text{out}}(v_n)}}=\frac{(v_k.h+\Delta k)^{d_{\text{out}}(v_k)}}{v_k.h^{d_{\text{out}}(v_k)}}=\left(1+\frac{\Delta k}{v_k.h}\right)^{d_{\text{out}}(v_k)}>1 \tag{25}$$

由式(25)可知，基于社会惩戒理论，系统在发现异常节点后，通过向异常节点发送警告信息，能够增大得到完整的 G_{Prop} 的概率。

在一些应用场景下，并不需要得到完整的 G_{Prop}，只需要得到保存有信息多副本的完整的节点集合 $G_{\text{Prop}}.V$，而对于信息如何传递并不关心。在该情况下，当需要还原出完整的 $G_{\text{Prop}}.V$ 时，$G_{\text{Prop}}.V$ 中每一个节点（起始节点除外）v_j 被主动存证的概率如式(26)所示。

$$P(v_j)=1-\prod_{i=1,e_i.v_{\text{end}}\text{ is }v_j}^{m}(1-e_i.v_{\text{start}}.h) \tag{26}$$

节点 v_k 的父节点集合表示方式如式(27)所示。

$$\text{In}(v_j)=\{u\mid\exists e,\ v_j.\text{equal}(e.v_{\text{end}})\wedge u.\text{equal}(e.v_{\text{start}})\} \tag{27}$$

结合式(26)和式(27)，可以得到式(28)。

$$P(v_j)=1-\prod_{i=1}^{d_{\text{in}}(v_j)}(1-\text{In}(v_j).u_i.h) \tag{28}$$

其中，$d_{\text{in}}(v_j)$ 表示节点 v_j 的入度。由于用户主动存证行为的独立性，通过主动存证还原出完整的 $G_{\text{Prop}}.V$ 的概率 $P(v_1, v_2,\cdots, v_n)$ 如式(29)所示。

$$P(v_1, v_2,\cdots, v_n)=\prod_{j=1}^{n}P(v_j) \tag{29}$$

同样，假设通过社会惩戒措施，v_k 诚实度的变化为 $\Delta k(\Delta k>0)$，此时，通过主动存证还原出完整的 $G_{\text{Prop}}.V$ 的概率表示为 $P^{'}(v_1, v_2,\cdots, v_m)$。

$$\frac{P^{'}(v_1, v_2,\cdots, v_n)}{P(v_1, v_2,\cdots, v_n)}=\frac{\prod_{i=1,v_k\in\text{In}(v_i)}^{n}\left(1-\prod_{j=1,\text{In}(v_i).u_j\neq v_k}^{d_{\text{in}}(v_i)}(1-\text{In}(v_i).u_j.h)(1-(v_k.h+\Delta k))\right)}{\prod_{i=1,v_k\in\text{In}(v_i)}^{n}\left(1-\prod_{j=1}^{d_{\text{in}}(v_i)}(1-\text{In}(v_i).u_j.h)\right)}=\frac{\prod_{i=1,v_k\in\text{In}(v_i)}^{n}\left(1-\prod_{j=1}^{d_{\text{in}}(v_i)}(1-\text{In}(v_i).u_j.h)+\Delta k\prod_{j=1,\text{In}(v_i).u_j\neq v_k}^{d_{\text{in}}(v_i)}(1-\text{In}(v_i).u_j.h)\right)}{\prod_{i=1,v_k\in\text{In}(v_i)}^{n}(1-\prod_{j=1}^{d_{\text{in}}(v_i)}(1-\text{In}(v_i).u_j.h))}=1+\frac{\Delta k\prod_{i=1,v_k\in\text{In}(v_i)}^{n}\prod_{j=1,\text{In}(v_i).u_j\neq v_k}^{d_{\text{in}}(v_i)}(1-\text{In}(v_i).u_j.h))}{\prod_{i=1,v_k\in\text{In}(v_i)}^{n}\left(1-\prod_{j=1}^{d_{\text{in}}(v_i)}(1-\text{In}(v_i).u_j.h)\right)}>1 \tag{30}$$

由式(30)可知，基于社会惩戒理论，系统在发现异常节点后，通过向异常节点发送警告信息，能够增大得到完整的 $G_{\text{Prop}}.V$ 的概率。

通过对信息多副本发现算法以及存证异常行为发现与处置算法的有效性进行分析，说明了基于流转主动存证与操作被动存证的交叉验证，以及对存证异常节点的发现与处置，能够增大多副本查找结果的完整性和真实性。

4 实验及分析

为验证所提出的机制，本文开发了针对 OFD 的双重存证多副本发现原型系统，包括 Android 插件和服务器端，其中，Android 插件在文档分享过程中扮演第三方的角色，不受限于任何特定的 OFD 文件阅读器，可以集成到任何现有的 OFD 文件阅读器中。Android 插件负责流转主动存证记录和操作被动存证记录的上报，服务器端负责存证信息的接收、存储、管理和检索，以及信息流转多副本传播图的构造、存证异常节点发现以及对异常节点采取相应处置措施。

4.1 实验环境

实验架构如图 6 所示，基于泛在共享环境的特点，不同的用户域模拟不同的应用系统，存证服务器则模拟信息监管部门的信息设施。

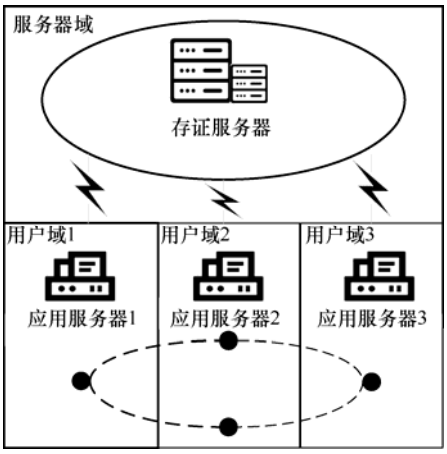

图 6　实验架构

本文实验使用多台 Android 手机以及基于 Android 操作系统的平板电脑作为客户端设备，使用 3 台服务器（2 GB RAM，单核，CentOS 7）作为应用服务器，使用一台服务器（2 GB RAM，单核，CentOS 7）作为存证服务器，使用关系型数据库（MySQL 5.7.37）存储用户信息，使用图数据库（Neo4j 3.5.13）存储流转主动存证记录和操作被动存证记录。基于该实验环境，测试本文所提方案的效果。

为测试所提方案的实际效果，将该原型系统在小范围内进行推广试用。目前，已有 17 个用户注册使用，并生成流转主动存证记录和操作被动存证记录共 923 条。

4.2　双重存证效果评估

4.2.1　功能正确性评估

当用户想要将一个 OFD 文件分享给另一个用户时，需要主动填写并生成流转主动存证记录，在选择待分享文档后，系统自动提取该文档中的文档标识，当文档标识不存在时，表明该文档第一次在系统中出现，系统为该文档生成文档标识。构造的文档流转主动存证记录实例如图 7 所示。


```
"SysInfo": {
      "SysType":0x01,
      "data": [{
            "infoID": "BA4A7F24-ACA7-4844-98A5-464786DF5C09",
            "Sender":"User2",
            "Recipient":"User1",
            "FileID":"IAQ5Cd41jyFhGtOr",
            "Channel": "Bluetooth"
      }]
}"hashSign": "2Y4dEOdbTb4nspfFreQIAVHxd6sJ+e+3utp7nCA680o="
```


图 7　文档流转主动存证记录实例

当用户想要操作或者打开一个文档时，系统会自动从文档中提取文档标识和文档的上层节点，构造文档操作被动存证并上传。构造的文档操作被动存证记录实例如图 8 所示。在功能方面，该原型系统能够及时、正确生成并上报流转主动存证记录和操作被动存证记录。


```
"SysInfo": {
      "SysType":0x02,
      "data": [{
            "infoID": "80442A35-CCBD-2416-3902-915428B4580C",
            "Sender":"User3",
            "Operator":"User4",
            "FileID":"n7sI4ACZeNfXrLYS",
            "Device":"Phone",
            " IP":"159.226.94.153",
            "Time":"2023-03-21 17:18:43"
      }]
}"hashSign": "2Y4dEOdbTb4nspfFreQIAVHxd6sJ+e+3utp7nCA680o="
```


图 8　文档操作被动存证记录实例

4.2.2　性能评估

在实现的原型系统中设置时间获取模块，对从用户打开主动存证页面，到生成流转主动存证记录的耗时，以及用户打开文件，到生成操作被动存证记录的耗时进行记录。

在 401 次有效的主动存证行为中，整个主动存证流程的平均耗时为 10.5 s，其中选择文档的平均耗时为 8 s。

在 522 次有效的被动存证行为中，整个被动存证流程的平均耗时为 189 ms。

在性能方面，由于流转主动存证需要用户手动选择相关文档，耗时相对较高，可以通过设置最近打开文件列表等方式，优化用户的主动存证时间，而被动存证流程几乎不会影响用户的使用体验。

4.3　多副本查找效果评估

4.3.1　功能正确性评估

在进行多副本查找时，根据文档 ID 在存证服务器中进行检索，结合对应文档的流转主动存证记录和操作被动存证记录，还原出文档的传播路径。在本例中，选取系统中传播范围最广的文档，进行实验功能效果的展示。

根据文档流转主动存证记录，构造文档流转传播图，如图 9 所示。

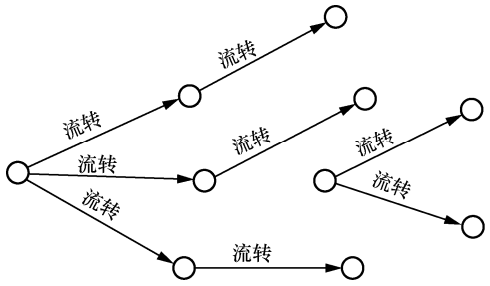

图 9　流转传播图

根据文档操作被动存证记录，构造文档溯源传播图，如图 10 所示。

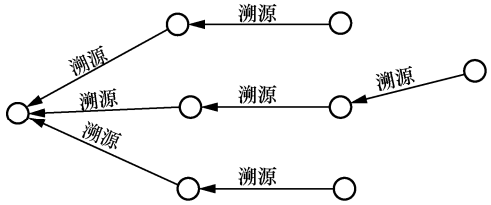

图 10　溯源传播图

结合文档传播流转图和文档溯源流转图，本文系统构造的文档多副本传播图如图 11 所示。

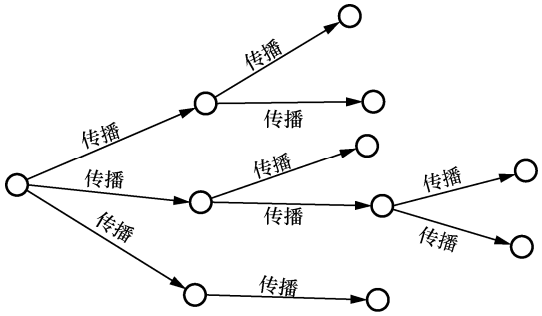

图 11　多副本传播图

通过对用户进行调查，得到该文档的完整传播图如图 12 所示。结合图 9~图 12，基于此实例，以该文档的完整传播图为基准，对多副本查找各个阶段生成的传播图的完整性进行分析并对比，结果如表 2 所示。

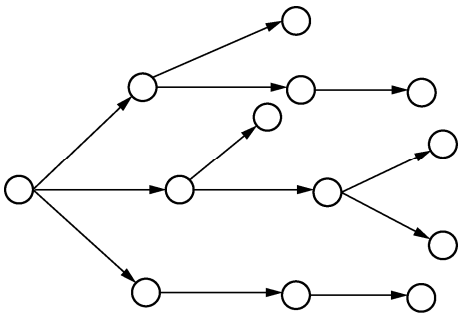

图 12　完整传播图

表 2　　传播图的完整性对比

传播图类型	节点数量/个	边数量/条	节点完整性	边完整性
流转传播图	10	8	76.9%	66.7%
溯源传播图	8	7	61.5%	58.3%
多副本传播图	11	10	84.6%	83.3%
完整传播图	13	12	—	—

由表 2 的对比结果可知，相比于单一存证机制，通过结合流转传播图与溯源传播图的信息，生成的多副本传播图的节点完整性和边完整性都得到了提升。

4.3.2　性能评估

从小范围推广使用得到的数据集中，对涉及的全部文档进行多副本查找，并计算各个步骤的平均时间消耗。在实现的原型系统中，文档流转传播图生成的平均耗时为 7.2 ms，文档溯源传播图生成的平均耗时为 6.9 ms，得到文档多副本传播图的平均耗时为 18.6 ms。

在性能方面，在该数据集中信息多副本查找不会造成太大的时间开销。

4.4　存证异常行为发现与处置效果评估

4.4.1　功能正确性评估

通过对流转主动存证信息和操作被动存证信息进行融合分析，来发现存证异常行为，并对存证异常行为涉及的节点进行警告，节点接收到的系统警告如图 13 所示。

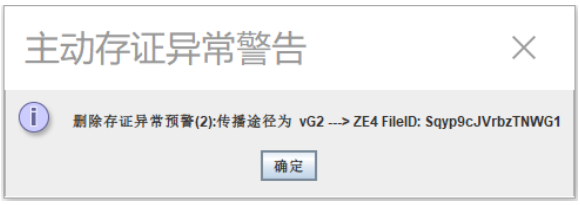

图 13　主动存证异常警告

为了对节点存证行为的变化进行观察，以验证存证异常行为发现与处置的效果，在系统运行一段时间后，每隔 2 天对全部节点的存证异常率进行计算，得到的结果如表 3 所示，其中存证异常率等于操作被动存证记录数除以存证异常行为数。

表 3　　平均异常率对比

时间状态	操作被动存证记录数量/条	存证异常行为数量/条	存证异常率
初始状态	243	46	18.9%
第 2 天	302	52	17.2%
第 4 天	356	60	16.9%
第 6 天	429	66	15.4%

由表 3 可知，通过双重存证的交叉验证，并对存证行为异常节点进行发现和处置，能够在一定程度上改善节点的主动存证行为，减少由于节点存证不主动、不及时和不正确导致的存证异常率。

4.4.2 性能评估

从小范围推广使用得到的数据集中，在 20 次重复实验下，存证异常行为发现的平均耗时为 882 ms。

在性能方面，在该数据集中存证异常行为发现不会造成太大的时间开销。

4.5 基于模拟数据的实验及分析

4.5.1 效果评估

为了评估本文所提方法的有效性，通过分析实际数据集的数据特征，构造了包含 1 000 个节点的网络，每个节点被赋予初始诚实度（0 到 1 之间）表示节点主动正确进行流转主动存证的概率以及活跃度（0 到 1 之间）表示节点接收到信息后会打开的概率。其中，节点的诚实度会慢慢调整，但节点的活跃度保持不变。

采用等概率随机扩散的方式，从不同节点出发，一共模拟 100 条信息的传播路径并生成 100 张完整的信息传播图。在信息到达每一个节点时，根据节点的诚实度和活跃度，模拟节点的行为并在节点执行对应行为时获取相应的存证信息。

在第 1 轮迭代结束后，一共获得了 1 863 条信息传播路径，100 条信息共涉及 1 815 个保存有信息或者信息副本的节点。然后，通过存证异常行为发现得到主动存证行为异常的节点，基于社会惩戒理论，在被发现后，随机增大异常节点的诚实度，之后进行第 2 轮迭代。

为了保证传播路径的一致性，第 2 轮迭代信息从相同的节点出发，并通过相同的传播路径进行传播，并记录此次迭代获得的存证信息。

同理，进行第 3 轮、第 4 轮、第 5 轮迭代，分别记录每次迭代过程中获得的存证信息，并基于每次迭代获得的存证信息进行多副本查找，查找结果如表 4 所示。

由表 4 的对比结果可以看到，随着迭代轮次的增加，节点的诚实度越来越高，还原出的流转传播图的完整性越来越高，由此构造的多副本传播图的完整性也越来越高。

4.5.2 性能评估

为了研究本文所提出的方法在不同数据规模下的效率，基于实际数据集，构造了 3 个不同规模的模拟数据集，并在不同数据集下测试并记录了信息多副本发现和存证异常行为发现的响应时间，测试过程中，每个环节重复 20 次，并计算时间的平均值，如表 5 所示。

在性能方面，通过在 4 个不同规模的模拟数据集中进行测试并计算平均响应时间，可以看到响应时间与数据集的规模基本上呈线性相关。

5 结束语

本文针对泛在共享环境下信息的跨域流转，提出了一种基于双重存证的信息流转多副本发现机制。该机制主要包括双重存证算法、信息多副本发现算法和存证异常行为发现与处置算法。双重存证算法通过结合流转主动存证与操作被动存证，对信

表 4　不同迭代次数下的多副本查找结果

迭代次数	平均节点诚实度	流转传播图		溯源传播图		多副本传播图	
		节点完整性	边完整性	节点完整性	边完整性	节点完整性	边完整性
1	0.3	28.21%	30.54%	34.27%	38.22%	38.02%	48.26%
2	0.5	37.63%	43.10%	34.93%	41.98%	41.60%	53.62%
3	0.7	50.25%	64.09%	36.80%	41.12%	54.10%	72.09%
4	0.8	71.57%	76.76%	33.61%	38.75%	76.36%	83.84%
5	0.9	86.83%	91.20%	36.36%	37.57%	92.29%	93.34%

表 5　多副本查找及主动存证异常发现效率

节点数量/个	关系数量/条	主动存证异常数量/条	信息多副本发现平均响应时间/ms	存证异常行为发现响应时间/ms
30	150	4	5	899
300	1 500	27	22	1 376
3 000	15 000	310	70	1 632
30 000	150 000	3 700	800	5 843

息的分享和操作行为进行记录；信息多副本发现算法通过对流转主动存证与操作被动存证进行融合分析，以相互补充的方式构造信息多副本传播图；存证异常行为发现与处置算法通过交叉验证发现存证异常行为，并对存证异常行为关联的节点进行处置。结合图论、互信息量和社交惩戒理论等，对信息多副本发现算法和存证异常行为发现与处置算法的有效性进行了分析，证明了所提机制能够提高信息多副本传播图构造的完整性和真实性。实验开发了针对 OFD 的双重存证多副本发现原型系统，在真实数据集和模拟数据集上，验证了所提机制相比于单一存证机制在构造信息多副本传播图完整性上的提升，并基于社会惩戒理论，在模拟数据集上通过多轮迭代模拟，进一步验证了存证异常行为发现与处置算法的有效性。

未来工作中，考虑将所提机制应用于更广泛的实际场景，以适应不同领域和应用场景的信息流转需求，并抽象归纳成具备普适性的框架或模型。另外，针对双重存证算法，通过对流转主动存证和操作被动存证的存证流程和存证格式进行优化，进一步提高存证过程的效率和可靠性；针对信息多副本发现算法，通过引入相似性度量方法、数据标记技术或深度学习模型等，将算法从精准检索扩展到更复杂的相似性分析，使其更好地适应泛在共享环境下信息流转的复杂性和多变性；针对存证异常行为发现与处置算法，引入更智能的异常节点检测方法，进一步研究异常节点的行为特征，以提高对异常节点的检测精度，并扩展到对不同类型异常的区分。

参考文献：

[1] 李凤华, 李晖, 牛犇. 隐私计算理论与技术[M]. 北京: 人民邮电出版社, 2021.
LI F H, LI H, NIU B. Privacy computing theory and technology[M]. Beijing: Posts & Telecom Press, 2021.

[2] CURTMOLA R, KHAN O, BURNS R, et al. MR-PDP: multiple-replica provable data possession[C]//Proceedings of 28th International Conference on Distributed Computing Systems. Piscataway: IEEE Press, 2008: 411-420.

[3] 熊金波, 沈薇薇, 黄阳群, 等. 云环境下的数据多副本安全共享与关联删除方案[J]. 通信学报, 2015, 36(S1): 136-140.
XIONG J B, SHEN W W, HUANG Y Q, et al. Security sharing and associated deleting scheme for multi-replica in cloud[J]. Journal on Communications, 2015, 36(S1): 136-140.

[4] DU L, ZHANG Z W, TAN S C, et al. An associated deletion scheme for multi-copy in cloud storage[C]//International Conference on Algorithms and Architectures for Parallel Processing. Berlin: Springer, 2018: 511-526.

[5] ZHANG Y Y, XIONG J B, LI X, et al. A multi-replica associated deleting scheme in cloud[C]//Proceedings of 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS). Piscataway: IEEE Press, 2016: 444-448.

[6] MIAO Y, HUANG Q, XIAO M Y, et al. Blockchain assisted multi-copy provable data possession with faults localization in multi-cloud storage[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 3663-3676.

[7] PHAN Q T, BOATO G, CALDELLI R, et al. Tracking multiple image sharing on social networks[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 8266-8270.

[8] YOU J X, LI Y M, LIANG R Q, et al. Image sharing chain detection VIA sequence-to-sequence model[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[9] SIDDIQUI N, ANJUM A, SALEEM M, et al. Social media origin based image tracing using deep CNN[C]//Proceedings of Fifth International Conference on Image Information Processing (ICIIP). Piscataway: IEEE Press, 2019: 97-101.

[10] 李凤华, 孙哲, 牛犇, 等. 跨社交网络的隐私图片分享框架[J]. 通信学报, 2019, 40(7): 1-13.
LI F H, SUN Z, NIU B, et al. Privacy-preserving photo sharing framework cross different social network[J]. Journal on Communications, 2019, 40(7): 1-13.

[11] ZHANG M, SUN Z, LI H, et al. Go-sharing: a blockchain-based privacy-preserving framework for cross-social network photo sharing[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(5): 3572-3587.

[12] TANG J T, WANG T, WANG J. Information flow detection and tracking on Web2.0 BLOGS based on social networks[C]//Proceedings of 9th International Conference for Young Computer Scientists. Piscataway: IEEE Press, 2008: 1664-1670.

[13] PATSAKIS C, ZIGOMITROS A, PAPAGEORGIOU A, et al. Distributing privacy policies over multimedia content across multiple online social networks[J]. Computer Networks, 2014, 75: 531-543.

[14] XIONG F, LIU Y, ZHANG Z J, et al. An information diffusion model based on retweeting mechanism for online social media[J]. Physics Letters A, 2012, 376(30/31): 2103-2108.

[15] 王超, 杨旭颖, 徐珂, 等. 基于 SEIR 的社交网络信息传播模型[J]. 电子学报, 2014, 42(11): 2325-2330.
WANG C, YANG X Y, XU K, et al. SEIR-based model for the information spreading over SNS[J]. Acta Electonica Sinica, 2014, 42(11): 2325-2330.

[16] GUILLE A, HACID H. A predictive model for the temporal dynamics of information diffusion in online social networks[C]//Proceedings of the 21st International Conference on World Wide Web. New York: ACM Press, 2012: 1145-1152.

[17] DICKENS L, MOLLOY I, LOBO J, et al. Learning stochastic models of information flow[C]//Proceedings of IEEE 28th International Con-

ference on Data Engineering. Piscataway: IEEE Press, 2012: 570-581.

[18] LIU C, ZHOU N, ZHAN X X, et al. Markov-based solution for information diffusion on adaptive social networks[J]. Applied Mathematics and Computation, 2020, 380: 125286.

[19] CHARIKAR M S. Similarity estimation techniques from rounding algorithms[C]//Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing. New York: ACM Press, 2002: 380-388.

[20] DATAR M, IMMORLICA N, INDYK P, et al. Locality-sensitive hashing scheme based on p-stable distributions[C]//Proceedings of the Twentieth Annual Symposium on Computational Geometry. New York: ACM Press, 2004: 253-262.

[21] IRIE G, LI Z G, WU X M, et al. Locally linear hashing for extracting non-linear manifolds[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 2123-2130.

[作者简介]

罗海洋（1997− ），男，湖南娄底人，中国科学院信息工程研究所博士生，主要研究方向为隐私计算、隐私保护。

邝彬（2000− ），男，湖南永州人，中国科学院信息工程研究所博士生，主要研究方向为隐私计算、隐私保护。

郭守坤（1994− ），男，河南周口人，中国科学院信息工程研究所工程师，主要研究方向为隐私计算、数据安全。

张玲翠（1986− ），女，河北故城人，博士，中国科学院信息工程研究所高级工程师、硕士生导师，主要研究方向为网络与系统安全、数据安全。

牛犇（1984− ），男，陕西西安人，博士，中国科学院信息工程研究所研究员、博士生导师，主要研究方向为数据安全、隐私计算。

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，主要研究方向为网络与系统安全、信息保护、隐私计算。