DOI：10.11830/ISSN.1000-5013.202504041

密钥驱动下多语义维度结构化成绩表生成式隐写方法

卢璟[1,2]，伏容[1,2]，田晖[1,2]

（1. 华侨大学 计算机科学与技术学院，福建 厦门 361021；
2. 厦门市数据安全与区块链技术重点实验室，福建 厦门 361021）

摘要： 针对教育结构化数据隐写中修改痕迹显著、多语义维度协同编码不足及安全可控性薄弱的问题，提出一种基于密钥驱动与多字段协同生成的无载体隐写方法。该方法以结构化成绩表为隐写对象，融合姓氏、名字、性别、成绩等多语义维度信息，通过主密钥分层派生子密钥，实现高保真度的隐写成绩表生成与可控嵌入。实验结果表明：隐写生成的姓名符合自然命名规范，课程成绩分布特征与基础表高度一致，统计特性稳定且无系统性偏差。与传统方法相比，文中方法在不可感知性、数据保真度与隐蔽性方面更具优势，为教育数据隐写提供了一种兼顾容量与安全性的可靠方案。

关键词： 生成式隐写；结构化成绩表；多语义维度；密钥驱动嵌入

中图分类号： TP 309.2　　**文献标志码：** A　　**文章编号：** 1000-5013(2025)03-0296-12

Key-Driven Multi-Semantic Dimensional Generative Steganography for Structured Academic Transcripts

LU Jing[1,2], FU Rong[1,2], TIAN Hui[1,2]

(1. College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China;
2. Xiamen Key Laboratory of Data Security and Blockchain Technology, Xiamen 361021, China)

Abstract: Aiming at the issues of significant modification traces, insufficient collaborative encoding of multi-semantic dimensional fields, and weak security controllability in educational structured data steganography, this paper presents a key-driven and multi-semantic dimensional generation-based carrierless steganography method. Using a structured academic transcript as the object for steganography, the method integrates multiple semantic dimensions, including surnames, given names, gender, and grades. It achieves high-fidelity generation and controllable embedding of steganographic academic transcripts by hierarchically deriving sub-keys from the master key. Experimental results show that the generated names conform to natural naming conventions, the grade distribution characteristics remain highly consistent with the original sheet, and the statistical properties are stable with no systematic bias. Compared with traditional methods, this approach demonstrates significant advantages in imperceptibility, data fidelity, and concealment, providing a reliable solution for educational data steganography that balances capacity and security.

Keywords: generative steganography; structured academic transcript; multidimensional semantics; key-driven embedding

收稿日期： 2025-04-22

通信作者： 田晖(1982－)，男，教授，博士，博士生导师，主要从事网络与信息安全、信息隐藏及其检测和数据安全等的研究。E-mail:htian@hqu.edu.cn。

基金项目： 国家自然科学基金资助项目(61972168，U1536115)

在当今信息数字化飞速发展的时代，高对抗环境下的隐蔽通信需求（如军事指令传递、跨境情报交互）对数据安全技术提出了更高要求[1]。信息隐藏技术作为一种可实现“通信行为隐匿”的安全手段，其重要性在深度监控场景中愈发凸显。从早期在图像[2-3]、音频[4-5]、视频[6-7]等传统媒体中的信息嵌入，到近年来对具有天然伪装优势的结构化文档[8-9]、自然语言文本[10-11]等领域的研究拓展，信息隐藏技术正从理论验证迈向对抗性实战应用阶段。特别是结构化文档因其格式规范、流通高频的特点，能够完美融入日常业务流程，已成为解决“囚徒问题”中载体合法性的关键突破方向。当前针对 HTML 网页[12]、XML 文档[13]或 Excel 数据表[14]等结构化文本的隐写研究，虽能实现基础信息嵌入，但在军事级隐蔽通信场景中仍存在较大缺陷：一是忽略字段语义关联性导致逻辑冲突，易被看守方统计检测[15]；二是隐写规则固定且集中秘密信息集中“存储”，一旦部分规则泄露则整个系统被破解。

在教育行业中，成绩表以其格式规范统一和信息结构明晰的特点，成为了一种极具价值且广泛流通的结构化文本。更重要的是，其天然包含的姓名、性别、课程、成绩等多维异构信息，为隐蔽通信提供了理想的载体条件。然而，目前针对成绩表生成式隐写（即根据秘密信息嵌入需要生成载密对象过程）的研究仍为空白。尽管部分生成式文本隐写[10-11]通过直接合成含密数据可为结构化数据隐写提供了有益借鉴，但在教育结构化数据中仍存在显著局限：一是现有编码策略仅针对单一类型字段，难以协调多语义维度的隐蔽性与载体合法性；二是生成过程缺乏严格的密码学约束，安全可控性不足，存在密钥泄露与模型逆向攻击风险。

鉴于此，本文提出一种基于密钥驱动的多语义维度结构化成绩表生成式隐写方法，旨在构建一个“字段级异构控制，多字段协同生成，密钥感知可逆提取”的统一框架。该方法以结构化成绩表为载体，融合姓氏、名字、性别、成绩等多个语义维度，通过主密钥分层派生控制信息分段与字段生成路径，从而实现具有高保真度的隐写成绩表生成。

1　多语义维度结构化成绩表生成式隐写方法

1.1　研究框架

针对教育结构化数据（如成绩表）隐写中存在的篡改痕迹显著、多语义维度字段协同编码不足、安全可控性薄弱等问题，提出一种基于密钥驱动与多字段协同生成的无载体隐写方法，其算法流程如图 1 所示。算法的核心应用场景聚焦于高隐蔽性、高安全性的秘密信息传递。文中所提方法的无载体特性使其特别适用于需要隐蔽嵌入且不引入统计异常的敏感数据传递。

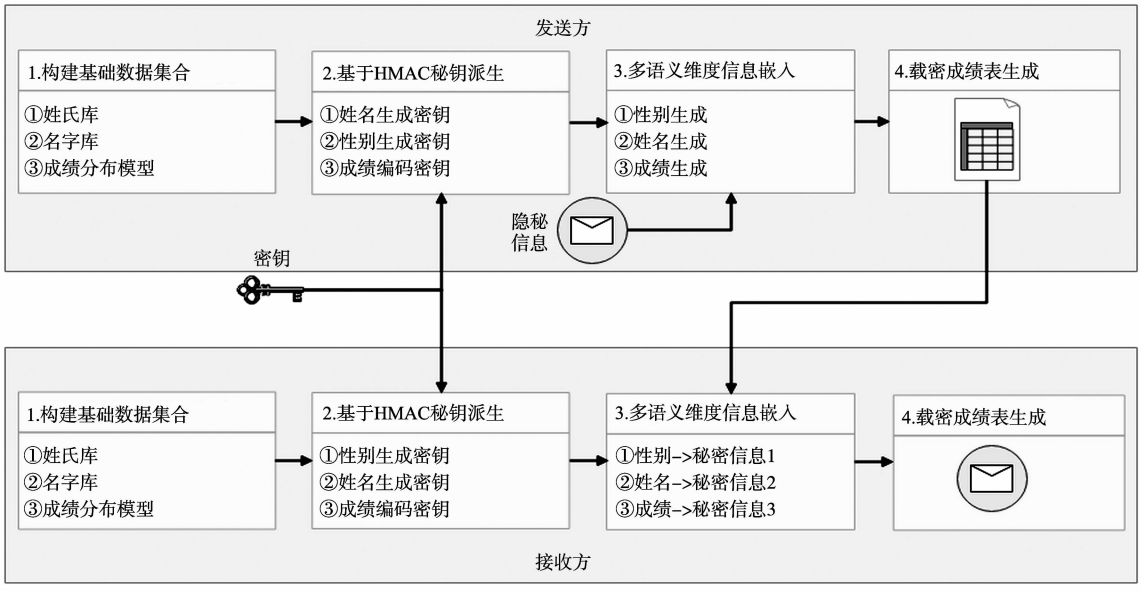


图 1　无载体隐写方法的算法流程

Fig. 1　Algorithm flow of carrierless steganography method

首先，构建基础数据集合，包括姓氏库、名字库、课程列表以及成绩分布模型。这些信息由隐蔽通信

的双方共享。其次,引入基于哈希消息验证码(Hash-based message authentication code, HMAC)的密钥派生算法,生成字段级子密钥(包括姓名生成密钥、性别生成密钥和成绩编码密钥)。由于收发双方共享通信密钥,双方能够生成相同的字段级子密钥。随后,发送方根据待嵌入的秘密信息,调用多语义维度信息嵌入模块,分别完成姓名、性别以及科目-成绩的自适应生成。最后,发送方将姓名、性别、科目-成绩等数据整合为符合教务规范的结构化成绩表并输出。接收方在收到载密成绩表后,可通过嵌入逆操作完成秘密信息的提取。

不难看出,在通信过程中,收发双方仅需共享基础数据(如姓氏库、名字库和课程列表)和通信密钥,无需传输额外的载体文件(如加密证书),从而显著降低了信息泄露的风险。此外,接收方若无正确的密钥,仅能获取表层的成绩数据,而无法感知到隐写信息的存在,这完全符合 Kerckhoffs 准则。因此,文中提出的方法是一种具有高安全性的隐蔽通信技术。

1.2 基础数据集合构建

通信双方预先生成并同步以下多语义维度数据资源。

1) 姓氏集合 $\mathscr{F}=\{f_i\}_{i=0}^{N_s-1}$,其中 N_s 为姓氏的总个数。本研究参考《二〇二一年全国姓名报告》,取人口数量排序前 256 个姓氏,构成姓氏库,如图 2 所示。因而,每个姓氏可编码为 8 bit,即可实现 8 bit 秘密信息的隐藏。

王	李	张	刘	陈	杨	赵	黄	周	吴	徐	孙	胡	朱	高	林	何	郭	马	罗	梁	宋	郑	谢
韩	唐	冯	于	董	萧	程	曹	袁	邓	许	傅	沈	曾	彭	吕	苏	卢	蒋	蔡	贾	丁	魏	薛
叶	阎	余	潘	杜	戴	夏	钟	汪	田	任	姜	范	方	石	姚	谭	廖	邹	熊	金	陆	郝	孔
白	崔	康	毛	邱	秦	江	史	顾	侯	邵	孟	龙	万	段	雷	钱	汤	尹	黎	易	常	武	乔
贺	赖	龚	文	庞	樊	兰	殷	施	陶	洪	翟	安	颜	倪	严	牛	温	芦	季	俞	章	鲁	葛
伍	韦	申	尤	毕	聂	丛	焦	向	柳	邢	路	岳	齐	梅	莫	庄	辛	管	祝	左	涂	谷	祁
时	舒	耿	牟	卜	詹	关	苗	凌	费	纪	靳	盛	童	欧	甄	项	曲	成	游	阳	裴	席	卫
查	屈	鲍	位	覃	霍	翁	隋	植	甘	景	薄	单	包	司	柏	宁	柯	阮	桂	闵	欧阳	解	强
柴	华	车	冉	房	边	辜	吉	饶	刁	瞿	戚	丘	古	米	池	滕	晋	苑	邬	臧	畅	宫	来
嵺	苟	全	褚	廉	简	娄	盖	符	奚	木	穆	党	燕	郎	邸	冀	谈	姬	屠	连	郜	晏	栾
郁	商	蒙	计	喻	揭	窦	迟	宇	敖	糜	鄢	冷	卓	花	仇								

图 2 姓氏库

Fig. 2 Surname database

2) 名字集合 $\mathscr{G}=\{\mathscr{G}_m,\mathscr{G}_f\}=\{\{g_{m,i}\}_{i=0}^{N_m-1},\{g_{f,i}\}_{i=0}^{N_f-1},\}$,其中 $\mathscr{G}_m\bigcap\mathscr{G}_f=\varnothing$,$\mathscr{G}_m$ 为男性名字库,N_m 为男性名字的总个数,$\mathscr{G}_f$ 为女性名字库,N_f 为女性名字的总个数。参考历史户籍数据,分别选取 128 个男性名字和 128 个女性名字构成男性名字库和女性名字库,分别如图 3,4 所示。因而,每个名字可编码为

伟	杰	勇	涛	强	浩宇	沐宸	浩然	宇轩	建华	俊杰	子轩	浩轩	宇航	一鸣	天宇
俊豪	博文	明轩	子豪	浩辰	梓豪	宇琛	文博	志强	嘉豪	俊宇	昊然	子涵	天佑	睿	梓轩
宇豪	家豪	文轩	明杰	鑫	俊熙	皓轩	瑞	泽宇	博涛	浩南	思远	晨阳	逸轩	明哲	俊凯
子墨	宇航	建国	建军	国庆	志伟	海涛	振华	志明	俊林	文强	宏伟	俊义	志勇	海波	晓东
晓明	志军	文涛	永强	志刚	文杰	海峰	立伟	立邦	国华	振宇	志远	永健	俊峰	建平	建辉
建龙	国栋	志斌	国伟	建邦	志名	国邦	国正	志骏	志平	建强	国富	志新	建业	志俊	国平
建明	建新	家军	志勤	家明	家辉	家伟	志文	文彬	文昊	天翔	天瑞	天浩	天翊	宇林	泽洋
子恒	子安	子骏	子辰	子诺	子瑜	子谦	子睿	宇凡	宇翔	宇恒	宇哲	宇鹏	宇成	宇飞	宇辰

图 3 男性名字库

Fig. 3 Male name database

敏	静	丽	艳	若汐	一诺	梓涵	雨桐	欣怡	诗涵	依诺	梓萱	语桐	可欣	欣妍	梦瑶
雨萱	晨曦	思涵	紫萱	诗琪	雨欣	梦琪	梓晴	思彤	佳琪	雨彤	雅婷	雪儿	婉婷	晓萱	欣悦
思雨	梦洁	雨嫣	雨薇	美琳	诗雨	慧妍	雅静	欣宜	子涵	若曦	艺涵	思颖	嘉怡	雨萌	佳怡
雨菲	心怡	秀英	桂英	玉兰	秀兰	红梅	丽华	丽娟	芳	霞	秀珍	玉珍	淑芬	春梅	秀云
丽萍	美玲	雪梅	丽娜	淑英	秀芳	美珍	春兰	玉梅	秀荣	玉芳	美芳	丽芬	秀华	秋菊	玉霞
小芳	玉琴	丽珍	秀玲	秀琴	美娟	美云	秀娟	美华	淑华	桂芳	桂兰	玉英	凤英	秀莲	玉华
玉香	翠花	秀芬	春燕	雅雯	晓玲	慧敏	嘉欣	思琪	佳慧	佳雯	梦婷	晓彤	雅琪	诗悦	雨珊
雨晴	若萱	思悦	梓琪	馨怡	嘉琳	诗颖	雅萱	瑾萱	欣瑶	悦然	诗妍	佳莹	若晴	雪瑶	婉清

图 4　女性名字库

Fig. 4　Female name database

7 bit，即可实现 7 bit 秘密信息的隐藏。进一步可知，如加上 1 bit 的性别字段，姓名和性别共可实现 16 bit(2 B)的信息隐藏。注意，对高频姓名库的严格筛选，其根本目标并非追求户籍人口的全覆盖，而是基于隐蔽通信的对抗性需求，即让生成的姓名更加的“自然”和“真实”，从而避免引起攻击者的“怀疑”。

3) 成绩分布模型 $\mathscr{D}_i \simeq \mathscr{N}(\mu \cdot (1+a_i), \sigma^2 \cdot (1+b_i)^2)$，定义各课程难度系数与相关性矩阵。其中，$\mu$ 为分布的均值，对应基准平均分；σ^2 为分布的方差，反映成绩的离散程度；$A=\{a_1, a_2, \cdots, a_p\}$ 为难度系数集合，$a_i \in [0, \lambda]$ 为第 i 门课程的难度系数；λ 为最大难度系数；$B=\{b_1, b_2, \cdots, b_p\}$ 标准差波动系数集合，b_i 为标准差波动系数。

1.3　基于 HMAC 的密钥派生

为提高算法安全性，引入 HMAC 密钥派生(key derivation)机制。它是一种基于 HMAC 算法的密码学方法，用于从一个主密钥(Key$\in\{0,1\}^{256}$)安全地派生出多个子密钥。其核心原理是通过分层哈希运算和上下文绑定，生成具备独立性的密钥，避免主密钥的直接暴露。定义 HMAC( )为安全哈希函数(如 SHA-256)，根据性别、姓名和分数等 3 种字段隐写的需要，定义三级初始密钥派生过程如下：

1) 64 bit 性别生成初始密钥 K_{gender}，即

$$K_{\text{gender}} \leftarrow \text{HMAC}(\text{Key}, \text{"gender"})[0:63]。\tag{1}$$

其含义是利用主密钥 Key 对“gender”执行安全哈希函数运算(如 SHA-256)，并截取前 64 位作为 K_{gender}；

2) 64 bit 姓名生成初始密钥 K_{name}，即

$$K_{\text{name}} \leftarrow \text{HMAC}(\text{Key}, \text{"name"})[0:63]。\tag{2}$$

其含义是利用主密钥 Key 对“name”执行安全哈希函数运算(如 SHA-256)，并截取前 64 位作为 K_{name}；

3) 128 bit 分数生成初始密钥 K_{score}，即

$$K_{\text{score}} \leftarrow \text{HMAC}(\text{Key}, \text{"score"})[0:127]。\tag{3}$$

其含义是利用主密钥 Key 对“score”执行安全哈希函数运算(如 SHA-256)，并截取前 128 位作为 K_{score}。

1.4　多语义维度信息嵌入

假设待嵌入的二进制秘密信息流为 $M\in\{0,1\}^*$，课程集合为 $C=\{c_1, c_2, \cdots, c_p\}$，对应的课程难度系数集合为 $A=\{a_1, a_2, \cdots, a_p\}$，标准差波动系数集合为 $B=\{b_1, b_2, \cdots, b_p\}$，单条记录的嵌入容量记为 n，则其计算公式为

$$n = n_{\text{name}} + \sum_{j=1}^{p} b_j。\tag{4}$$

其中：n_{name} 为每个姓名的嵌入容量，如前分析，结合性别和姓名的隐藏容量为 16 bit；b_j 为每门课程分数的嵌入容量。进一步，可得到嵌入所有秘密信息所需的记录数为

$$R = \left\lceil \frac{|M|}{n} \right\rceil。\tag{5}$$

定义 S_1、S_2、$S_3 \in \{0,1\}^{\infty}$ 分别是 K_{gender}、K_{name} 和 K_{score} 派生的伪随机二进制密钥流。以下为秘密信息的嵌入过程。

1) 头部构造。为便于通信双方同步,构造16 bit头部 $H\in\{0,1\}^{16}$。即

$$H=\underbrace{h_0h_1}_{\Delta\text{size}}\parallel\underbrace{h_2\cdots h_{15}}_{\text{Payload length}}。\tag{6}$$

其中:h_0h 为课程分数修改集合标识;$h_2\cdots h_{15}$表示嵌入字节数,最大可表示4 095 B。完整的嵌入信息为 $\widetilde{M}=H\parallel M$。

2) 基础成绩表生成。根据给定的基础正态分布函数,结合难度系数集合(A)和标准差波动系数集合(B),生成成绩矩阵 $\boldsymbol{\Phi}\in\mathbb{R}^{R\times p}$,且第 i 行第 j 列的元素(成绩)$\varphi_{i,j}$满足

$$\varphi_{i,j}\simeq\mathcal{N}(\mu\cdot(1+\alpha_j),\sigma^2\cdot(1+b_j)^2)。\tag{7}$$

其中:μ 为基准平均分;σ 为标准差调节因子;$a_j\in[0,\lambda]$为第 j 门课程的难度系数;b_j 为标准差波动系数;λ 为最大难度系数。本模型基于经典测试理论(classical test theory, CTT)的核心假设构建,即大规模学生群体的课程成绩应服从正态分布。相关参数的具体解释及取值范围,如表1所示。

表1 基础成绩表生成模型相关参数的定义

Tab.1 Definition of parameters related to basic academic transcript generation model

参数	定义解释	取值范围与约束
基准平均分 μ	反映课程基础难度(如 $\mu=60$ 表示中等难度课程)	$\mu\in[40,85]$,上限受防溢出条件约束,即 $\mu(1+\lambda)\leqslant100$
难度系数 a_j	动态调节课程难度等级,a_j 越大,平均分越高(课程越简单)	$a_j\in[0,\lambda]$, $\lambda=\min(0.4,\frac{100}{\mu}-1)$
波动系数 b_j	控制成绩离散程度,b_j 越大,学生能力差异表现越显著	$b_j\in[-0.3,0.3]$,确保标准差 $\sigma(1+b_j)$ 非负且合理
标准差调节因子 σ	初始标准差(反映整体区分度)	$\sigma\in[10,20]$

3) 基于姓名-性别的信息嵌入。对于成绩表中的第 $i\in[0,R-1]$条记录,性别索引、姓氏索引和名字索引分别为

$$\mathrm{idx}_i^{\mathrm{g}}=\widetilde{M}[16i]\oplus S_1[i]\in\{0,1\},\tag{8}$$

$$\mathrm{idx}_i^{\mathrm{sn}}=(\textstyle\bigoplus_{k=0}^{7}(\widetilde{M}[16i+k],S_2[15i+k]))\bmod N_{\mathrm{s}},\tag{9}$$

$$\mathrm{idx}_i^{\mathrm{gn}}=(\textstyle\bigoplus_{k=8}^{15}(\widetilde{M}[16i+k],S_2[15i+k]))\bmod N_{\mathrm{gn}}。\tag{10}$$

其中:$N_{\mathrm{gn}}=N_{\mathrm{m}}=N_{\mathrm{f}}$ 为名字库的大小。从而,可得到生成的名字为

$$\mathrm{Name}_i=\begin{cases}(\mathscr{F}[\mathrm{idx}_i^{\mathrm{sn}}],\mathscr{G}_{\mathrm{f}}[\mathrm{idx}_i^{\mathrm{gn}}]), & \text{If } \mathrm{idx}_i^{\mathrm{g}}=0\\(\mathscr{F}[\mathrm{idx}_i^{\mathrm{sn}}],\mathscr{G}_{\mathrm{m}}[\mathrm{idx}_i^{\mathrm{gn}}]), & \text{其他。}\end{cases}\tag{11}$$

4) 基于课程分数的信息嵌入。根据 $H[0:1]$确定修改的集合 Δ,计算每个课程所需比特数 $b=\lceil\log_2|\Delta|\rceil$。对于待嵌入秘密信息流中第 $k\in[16R,|\widetilde{M}|-1]$比特开始的 b bit。计算课程分数修改量为

$$\delta_k=\Delta[(\textstyle\bigoplus_{i=k}^{k+b}(\widetilde{M}[i],S_3[i-16R]))],\bmod|\Delta|]。\tag{12}$$

进一步修改相应的课程分数为

$$\widetilde{\Phi}_{i,j}=\Phi_{i,j}+\delta_k。\tag{13}$$

其中:$i=\lfloor(k-16R)/p\rfloor$,$j=(k-16R)\bmod p$,$p=bN_c$, N_c 为总的课程数目。

为便于理解上述算法流程,举例说明如下。

假设 $i\in[0,R-1]$条记录对应需嵌入的秘密信息为01101100 11100011,$S_1[i]=1$, $S_2[15i:15i+14]=$10110111 0110110,则 $\mathrm{idx}_i^{\mathrm{g}}=0\oplus1=1$,表示是男性;$\mathrm{idx}_i^{\mathrm{sn}}=11011001\oplus10110111=01101110$,其十进制表示为110,对应姓氏库中的“倪”;$\mathrm{idx}_i^{\mathrm{gn}}=1100011\oplus0110110=1010101$,其十进制表示为85,对应男性名字库中的“志名”。由此,可知生成的姓名为“倪志名”。

此外,假设 $\Delta=(-1.5,-1,-0.5,0,0.5,1,1.5,2)$,则 $b=3$;待嵌入的二进制秘密信息 $\widetilde{M}[k:k+b]$为010,$S_3[k-16R:k+b-16R]=111$,则 $\delta_k=\Delta[(010\oplus111)\bmod8]=1$;如 $\Phi_{i,j}$ 为85,则嵌入后的课程分数为86。

1.5　多语义维度信息提取

多语义维度信息提取是上述嵌入操作的逆过程，有 5 个主要步骤。

1）头部信息解析。头部信息隐藏在第 1 条记录的姓名中，为此首先提取第 1 条记录性别字段中隐藏的信息。即

$$\widetilde{M}[0]=\mathrm{idx}_0^{\mathrm{g}}\oplus S_1[0]。\tag{14}$$

进而，从姓名中获取姓氏 $\mathscr{F}[\mathrm{idx}_0^{\mathrm{sn}}]$，并从中可以提取秘密信息段。即

$$\widetilde{M}[1:8]=(\bigoplus_{k=0}^{7}(B(\mathrm{idx}_0^{\mathrm{sn}}),S_2[k]))。\tag{15}$$

其中：$B(*)$为将十进制数转二进制序列的函数。根据性别选择名字库 G_{m} 或 G_{f}，获取名字索引$\mathrm{idx}_0^{\mathrm{gn}}$，并从中提取秘密信息段。即

$$\widetilde{M}[9:15]=(\bigoplus_{k=0}^{6}(B(\mathrm{idx}_0^{\mathrm{gn}}),S_2[k+8]))。\tag{16}$$

根据头部信息确定课程分数修改集合 Δ 和每门课程分数的嵌入比特数 $b=\lceil\log_2|\Delta|\rceil$，以及负载长度 $h_2\cdots h_{15}=\widetilde{M}[2:15]$和嵌入信息的字节数 $L=\sum_{k=2}^{15}(h_k\cdot 2^{14-k})$。

2）基础成绩表重构。按式(7)重构原始成绩矩阵 $\boldsymbol{\Phi}\in\mathbb{R}^{R\times p}$。

3）基于姓名-性别的信息提取。对第 $i\in[1,R-1]$条记录，首先提取性别字段中隐藏的信息，即

$$\widetilde{M}[16i]=\mathrm{idx}_i^{\mathrm{g}}\oplus S_1[i]。\tag{17}$$

进而，从姓名中获取姓氏 $\mathscr{F}[\mathrm{idx}_0^{\mathrm{sn}}]$，并从中可以提取秘密信息段。即

$$\widetilde{M}[16i+1:16i+8]=(\bigoplus_{k=0}^{7}(B(\mathrm{idx}_i^{\mathrm{sn}}),S_2[15i+k])。\tag{18}$$

根据性别选择名字库 $\mathscr{G}_{\mathrm{m}}$ 或 $\mathscr{G}_{\mathrm{f}}$，获取名字索引$\mathrm{idx}_0^{\mathrm{gn}}$，并从中提取秘密信息段。即

$$\widetilde{M}[16i+9:16i+15]=(\bigoplus_{k=0}^{6}(B(\mathrm{idx}_i^{\mathrm{gn}}),S_2[15i+k+8]))。\tag{19}$$

4）基于课程分数的信息提取。对于 $k\in[16R,|\widetilde{M}|-1]$比特开始的 b bit，计算课程分数修改量为

$$\delta_{i,j}=\widetilde{\Phi}_{i,j}-\Phi_{i,j}。\tag{20}$$

其中：$i=\lfloor(k-16R)/p\rfloor$，$j=(k-16R)\bmod p$。进一步可从中提取秘密信息段，即

$$\widetilde{M}[k:k+b-1]=\bigoplus_{i=k}^{k+b-1}(B(\Delta^{-1}[\delta_{i,j}]),S_3[i-16R])。\tag{21}$$

其中：$\Delta^{-1}[\cdot]$为 $\Delta[\cdot]$的逆映射，即找到修改量对应的元素索引。令 $k=k+b$，重复本步骤，直至所有秘密信息被提取完毕。

5）秘密信息重组。将提取的比特流按顺序拼接，去除头部 H，得到原始秘密信息 M，可满足

$$M=\widetilde{M}[16:8L+16]。\tag{22}$$

2　性能对比与分析

2.1　与传统方法的对比分析

文中提出的无载体生成式隐写方法与传统隐写技术（例如基于 LSB 的数据隐写）存在根本性差异。传统方法通过修改载体数据嵌入秘密信息，而文中所提方法通过密钥驱动直接生成符合规范的数据。因此，二者在容量、隐蔽性等关键指标上缺乏直接可比性。借鉴已有无载体隐写研究，采用理论对比方式，文中所提方法与传统方法的功能对比矩阵如表 2 所示。

综上所述，文中所提方法在不可感知性、数据保真度、隐藏容量与隐蔽性方面显著优于传统方法，为教育数据隐写提供了一种兼顾容量与安全性的可靠方案。

2.2　隐藏容量

对于所提出的结构化成绩表生成式隐写方法，每条记录的嵌入容量如式(4)所示，取决于课程数量和每门课程分数能够嵌入的比特数（$b=\lceil\log_2|\Delta|\rceil$）。假设选取四种修改集合 $\Delta_0=(0,1)$、$\Delta_1=(-0.5,0,0.5,1)$，$\Delta_2=(-1.5,-1,-0.5,0,0.5,1,1.5,2)$、$\Delta_3=(-1.75,-1.5,-1.25,-1,-0.75,-0.5,$

$-0.25,0,0.25,0.5,0.75,1,1.25,1.5,1.75,2$),不难得出结论:成绩数据表的隐写容量取决于课程数、学生数以及修改集合。对于给定的修改集合 $\Delta_i(i=0,1,2,3)$,隐藏容量(SC_i)的计算式为

$$SC_i = (sc_{name} + sc_{gender} + \lceil \log_2 | \Delta_i | \rceil \cdot N_c) \cdot N_s。 \tag{23}$$

其中:sc_{name} 和 sc_{gender} 分别为每个姓名和性别能够嵌入的比特数,分别为 15 和 1 bit;N_c 和 N_s 分别为课程数和学生数。

表 2 两种隐写方法的功能对比

Tab. 2 Functional comparison of proposed and traditional methods

对比维度	传统隐写方法	文中方法
技术基础	修改数据的二进制冗余位	密钥驱动生成合规结构化数据
载体依赖性	必须存在可修改的载体文件	无需任何载体,直接生成目标数据
隐蔽性机制	依赖冗余空间隐蔽性	依赖生成数据的统计真实性
多字段协同	仅支持单一模态(成绩)	跨姓名/性别/成绩编码
数据完整性	可能破坏数据完整性	完全保持数据规范完整性

上述 4 种修改集合对应的隐藏容量与课程数目及学生数目的关系,如图 5 所示。

从图 5 中可以清晰地看出以下规律:1) 当课程数量较少(即 $N_c < sc_{name}/\lceil \log_2 | \Delta_i | \rceil$)时,姓名和性别能够嵌入的比特数在每个元组条目中的占比相对较大,因此学生数目对隐藏容量的影响更为显著。2) 在上述四种修改集合中,Δ_0 为最严格的修改方式,导致每门课程分数的嵌入容量最低(仅为 1 比特);相反,Δ_3 允许最宽松的修改,使得每门课程分数的嵌入容量达到最高(4 *bit*)。然而,宽松的修改集合可能会引入更多的数据扰动,因此需要根据实际需求进行权衡。总体而言,当课程数较少但学生数较多时,Δ_0 表现出更好的性能,建议在小规模课程场景下优先选择 Δ_0;反之,Δ_3 在大规模数据表中具有更高的容量,更适合需要高隐写负载的场景。

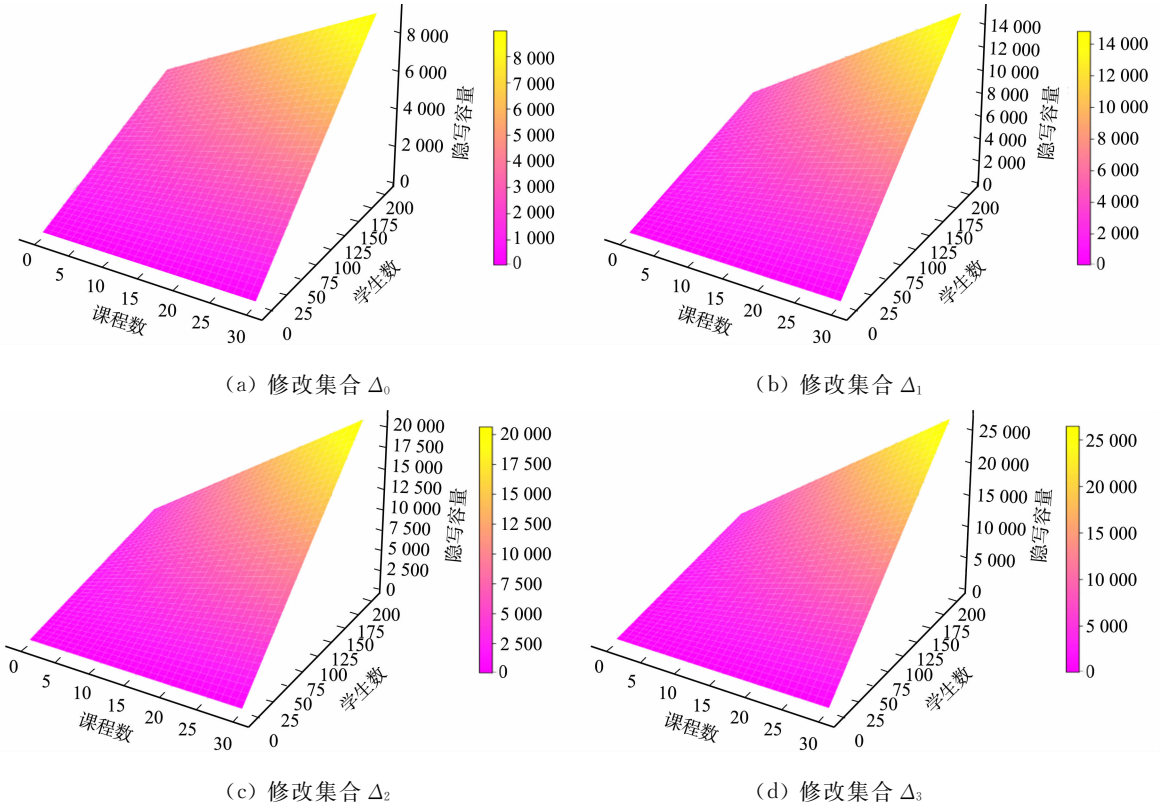


图 5 采用不同修改集合时隐藏容量与课程数目及学生数目的关系

Fig. 5 Relationship between hidden capacity and number of courses and students when using different modification sets

2.3 隐写安全性

为了验证文中所提方法的隐写透明性(隐写的不可见性),取字符串“HQU”的 SHA-256 哈希值作

为初始密钥，假设待嵌入的秘密信息为“欢迎报考华侨大学”，并随机生成基础成绩表，共包含语文、数学、英语、物理、化学、生物、历史及地理 8 门课程，如表 3 所示。进而，分别采用以上四种修改集合，得到的隐写成绩表分别如表 4～7 所示。

表 3　随机生成的基础成绩表

Tab. 3　Randomly generated basic academic transcript

姓名	性别	语文	数学	英语	物理	化学	生物	历史	地理
待嵌入信息时自适应生成	待嵌入信息时自适应生成	78.75	77.50	69.75	68.75	69.50	74.25	75.25	78.00
		76.25	81.50	77.25	60.00	77.25	76.00	79.50	76.50
		72.50	77.75	72.50	62.50	73.50	77.75	77.00	78.00
		75.50	76.25	74.00	68.75	63.00	79.75	73.00	69.75
		72.50	85.00	72.50	71.00	73.50	75.25	77.00	77.00
		72.50	79.75	84.75	68.50	70.00	76.25	81.75	79.25
		74.75	80.75	75.50	67.50	68.50	73.50	75.50	83.75
		68.25	76.25	70.25	66.75	75.00	73.25	81.75	75.75
		68.75	78.75	79.50	62.00	77.00	79.25	65.00	74.50
		72.25	80.75	69.25	65.00	76.50	81.00	78.75	75.75

表 4　采用修改集合 Δ_0 生成的隐写成绩表

Tab. 4　Steganographic academic transcript generated using modification set Δ_0

姓名	性别	语文	数学	英语	物理	化学	生物	历史	地理
熊丽萍	女	78.75	77.50	69.75	68.75	69.50	75.25	76.25	79.00
袁志远	男	77.25	82.50	77.25	60.00	78.25	76.00	80.50	77.50
郑国平	男	72.50	77.75	73.50	62.50	73.50	77.75	78.00	79.00
柏雨欣	女	75.50	76.25	74.00	69.75	63.00	80.75	74.00	69.75
谈文涛	男	73.50	85.00	72.50	72.00	74.50	76.25	78.00	78.00
龚静	女	73.50	79.75	84.75	69.50	71.00	76.25	81.75	80.25
岳玉梅	女	75.75	80.75	76.50	67.50	68.50	74.50	76.50	83.75
秦沐宸	男	68.25	76.25	70.25	66.75	76.00	74.25	81.75	75.75
方诗妍	女	68.75	78.75	79.50	62.00	77.00	79.25	65.00	74.50
宇雨彤	女	72.25	80.75	69.25	65.00	76.50	81.00	78.75	75.75

表 5　采用修改集合 Δ_1 生成的隐写成绩表

Tab. 5　Steganographic academic transcript generated using modification set Δ_1

姓名	性别	语文	数学	英语	物理	化学	生物	历史	地理
冉丽萍	女	78.75	77.50	70.75	68.25	70.50	74.75	76.25	78.50
袁志远	男	76.25	81.00	77.75	59.50	77.25	76.00	79.50	77.00
郑国平	男	73.00	77.25	72.50	63.00	74.50	77.75	77.00	77.50
柏雨欣	女	76.50	76.75	75.00	69.25	62.50	79.25	74.00	70.75
谈文涛	男	73.00	86.00	72.00	70.50	74.00	75.75	77.00	76.50
龚静	女	73.00	80.25	85.75	68.50	69.50	77.25	82.75	80.25
岳玉梅	女	74.75	80.75	75.50	67.50	68.50	73.50	75.50	83.75
秦沐宸	男	68.25	76.25	70.25	66.75	75.00	73.25	81.75	75.75
方诗妍	女	68.75	78.75	79.50	62.00	77.00	79.25	65.00	74.50
宇雨彤	女	72.25	80.75	69.25	65.00	76.50	81.00	78.75	75.75

表 6　采用修改集合 Δ_2 生成的隐写成绩表

Tab. 6　Steganographic academic transcript generated using modification set Δ_2

姓名	性别	语文	数学	英语	物理	化学	生物	历史	地理
熊晓明	男	77.75	79.00	69.75	70.25	68.00	72.75	75.75	78.00
袁志远	男	75.25	83.50	79.25	60.00	77.25	75.50	79.50	75.00
郑国平	男	73.50	78.75	72.00	61.00	72.50	79.75	78.50	77.50
柏雨欣	女	77.50	75.25	74.00	68.25	63.00	80.75	75.00	70.25
谈文涛	男	72.50	84.50	72.50	72.50	75.00	75.25	77.00	77.00

续表

Continue table

姓名	性别	语文	数学	英语	物理	化学	生物	历史	地理
龚静	女	72.50	79.75	84.75	68.50	70.00	76.25	81.75	79.25
岳玉梅	女	74.75	80.75	75.50	67.50	68.50	73.50	75.50	83.75
秦沐宸	男	68.25	76.25	70.25	66.75	75.00	73.25	81.75	75.75
方诗妍	女	68.75	78.75	79.50	62.00	77.00	79.25	65.00	74.50
宇雨彤	女	72.25	80.75	69.25	65.00	76.50	81.00	78.75	75.75

表7 采用修改集合 Δ_3 生成的隐写成绩表

Tab.7 Steganographic academic transcript generated using modification set Δ_3

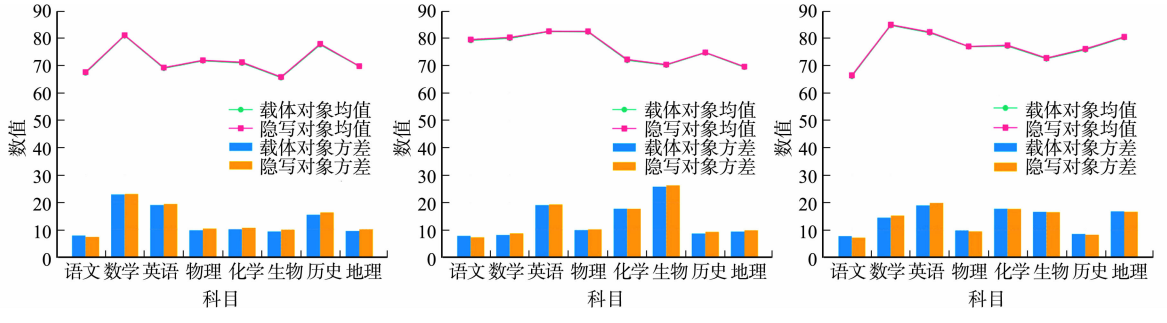
姓名	性别	语文	数学	英语	物理	化学	生物	历史	地理
冉晓明	男	77.25	78.75	70.00	69.75	68.75	74.00	74.25	76.50
袁志远	男	77.75	83.00	79.25	58.50	77.50	75.50	78.75	78.50
郑国平	男	73.25	78.25	74.25	63.00	75.25	77.25	76.00	79.50
柏雨欣	女	75.25	75.25	74.75	70.75	65.00	79.50	72.25	68.25
谈文涛	男	72.50	85.00	72.50	71.00	73.50	75.25	77.00	77.00
龚静	女	72.50	79.75	84.75	68.50	70.00	76.25	81.75	79.25
岳玉梅	女	74.75	80.75	75.50	67.50	68.50	73.50	75.50	83.75
秦沐宸	男	68.25	76.25	70.25	66.75	75.00	73.25	81.75	75.75
方诗妍	女	68.75	78.75	79.50	62.00	77.00	79.25	65.00	74.50
宇雨彤	女	72.25	80.75	69.25	65.00	76.50	81.00	78.75	75.75

在上述实验中,秘密信息均能完整嵌入并成功提取。通过分析这些数据可以发现:1) 隐写表中生成的所有姓名均严格符合常见中文命名习惯,有效避免了生僻字和怪异组合的出现,从而确保了姓名的自然性和合理性;2) 隐写表的成绩分布特征(包括均值、方差以及分数段分布)与基础成绩表高度一致,说明嵌入修改并未破坏成绩的自然性。综上所述,文中所提出的隐写方法在不可感知性方面表现出色,能够有效隐藏秘密信息,同时保持数据的自然性和合理性。

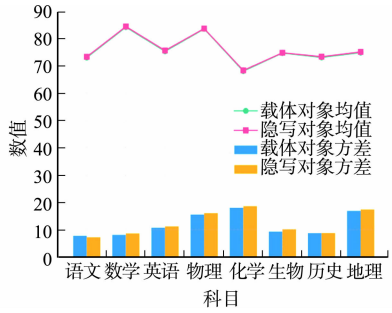
为了进一步验证文中所提方法的隐蔽性,尤其是隐写前后各科成绩的变化情况,随机选取20组长度约为360 B的秘密信息,并采用修改集合 Δ_1 将其分别嵌入包含语文、数学、英语、物理、化学、生物、历史和地理8门课程的成绩表中。通过隐写前后各门课程的均值和方差的对比,结果如图6所示。

从图6中可以看出:在所有实验组中,隐写前后各科成绩的均值变化均不超过0.5分,方差变化控制在5%左右。这些变化符合自然评分波动的规律,满足了不可见性的要求。此外,修改集合 Δ_1 在不同科目和不同秘密信息下表现稳定,未出现系统性偏差或规律性特征。这说明,本文方法在保持成绩表统计特征稳定的基础上,成功实现了高隐蔽性的信息嵌入。

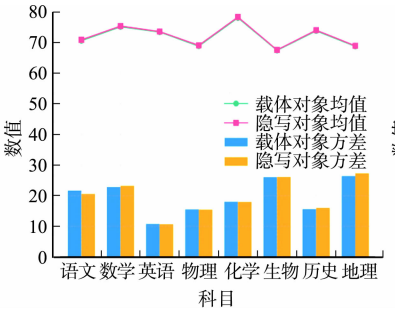
综上,可从形式化安全模型与实验验证两个维度系统评估所提出方法的安全性。基于Real-or-Random(RoR)模型,能够证明该方案满足IND-CPA安全性(indistinguishability under chosen-plaintext attack),即对于任意概率多项式时间攻击者,其区分真实成绩表与隐写生成成绩表的优势可忽略不计($Adv \leqslant negl(\lambda)$)。具体而言:1) 密钥安全性。本方案采用HMAC-SHA256算法派生字段级子密钥,充分满足伪随机函数(PRF)的特性。2)前向安全性。主密钥到子密钥的不可逆性由HMAC的抗



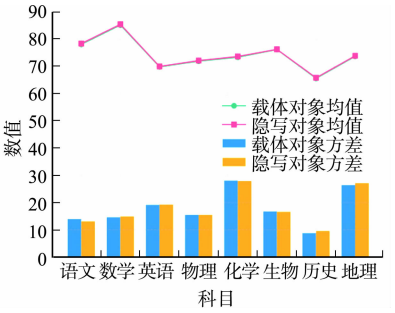
(a) 隐写实验1 (b) 隐写实验2 (c) 隐写实验3



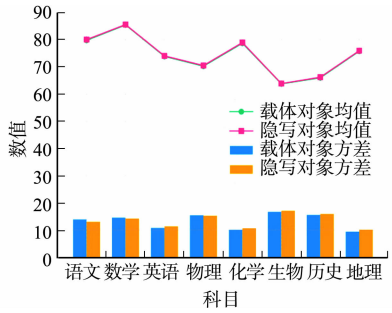
（d）隐写实验 4



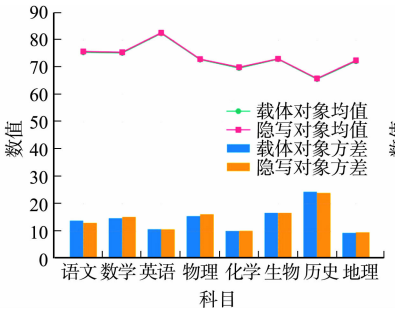
（e）隐写实验 5



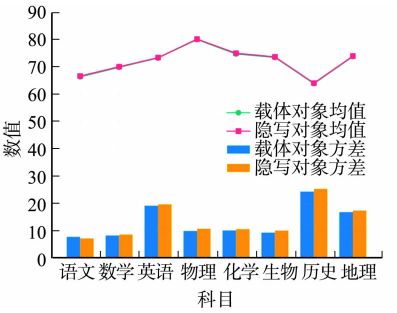
（f）隐写实验 6



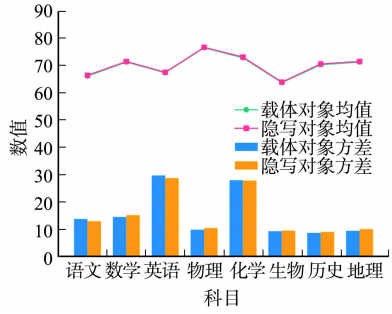
（g）隐写实验 7



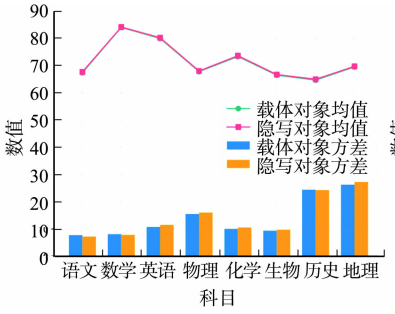
（h）隐写实验 8



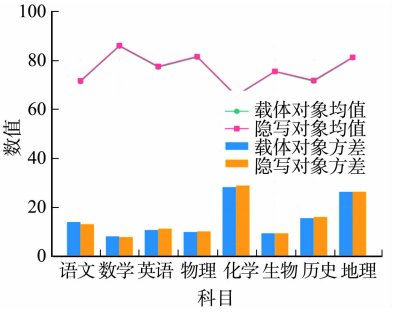
（i）隐写实验 9



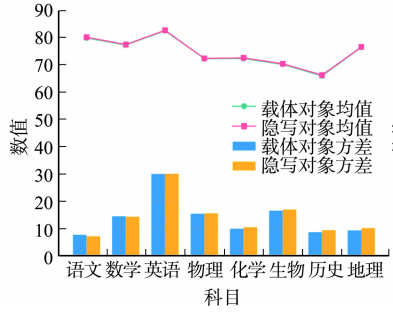
（j）隐写实验 10



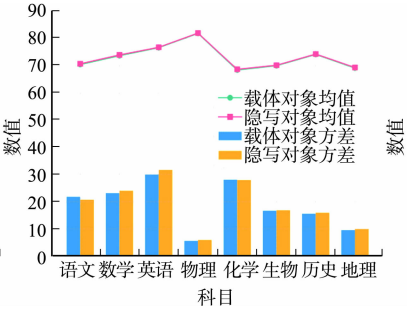
（k）隐写实验 11



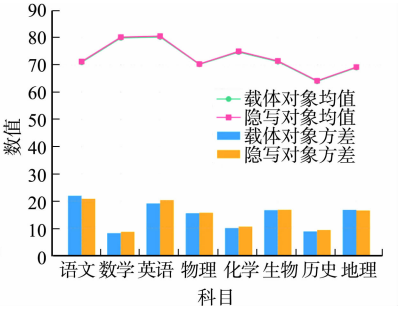
（l）隐写实验 12



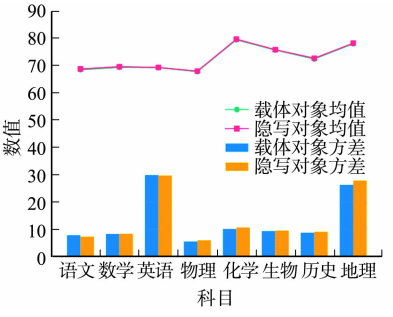
（m）隐写实验 13



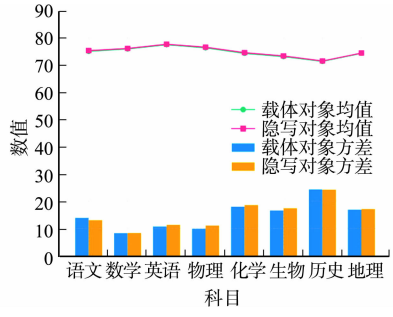
（n）隐写实验 14



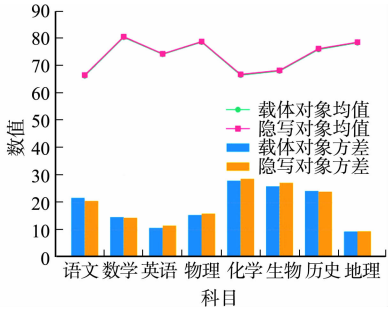
（o）隐写实验 15



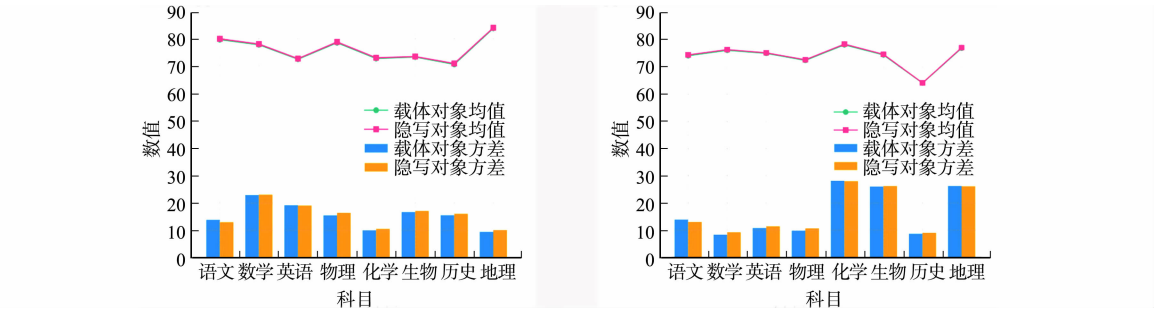
（p）隐写实验 16



（q）隐写实验 17



（r）隐写实验 18



(s) 隐写实验19 (t) 隐写实验20

图6 随机秘密信息隐写实验中课程成绩统计特征对比

Fig. 6 Comparison of statistical features of course grades in steganographic experiments with randomly generated secret messages

碰撞性提供保障。3)抗统计检测。从上述实验可知,生成的姓名、性别和成绩字段均通过统计检验,与真实教育数据无显著差异。

2.4 算法的时空复杂度分析

所提出方法的时间复杂度主要取决于结构化成绩表的元组条目数(记为 n),即。

1) 密钥派生阶段。采用基于 HMAC 的安全哈希函数(如 SHA-256)进行密钥派生。为满足性别、姓名和分数三种字段的隐写需求,共执行三级初始密钥派生操作,时间复杂度为 $O(1)$(三次固定计算,与条目数 n 无关)。

2) 字段生成阶段。姓名-性别生成:基于预加载的姓氏/名字库,结合性别进行索引查询,单条目时间复杂度为 $O(1)$;成绩生成:通过分布保持的方式直接修改各课程成绩,单条目时间复杂度为 $O(1)$。

3) 总时间复杂度($T(n)$)为

$$T(n) = T_{\text{秘钥派生}} + n \times (T_{\text{姓名-性别}} + T_{\text{成绩}}) = O(1) + n \times O(1) = O(n)。 \tag{24}$$

由此可以得出结论:所提出方法的时间复杂度与元组条目数成线性关系,每个条目的生成均为常数时间操作。

所提出方法的存储开销由基础数据预存和运行时内存占用两部分构成。

1) 基础数据预存部分。包括姓氏库(256 条)、名字库(128×2 条)和固定大小的课程列表,总空间复杂度为 $O(1)$,与元组条目数 n 无关。

2) 运行时内存占用部分。每个元组存储姓名、性别、成绩等字段,与传统成绩表完全一致,空间复杂度为 $O(n)$ 。

3) 总空间复杂度为

$$S(n) = S_{\text{基础库}} + S_{\text{元组数据}} = O(1) + O(n) = O(n)。 \tag{25}$$

由此可得结论:所提方法的空间复杂度与元组条目数呈线性关系,与传统成绩表的存储需求一致。

综上,文中所提出方法在保持与传统成绩表相同存储效率($O(n)$)的前提下,通过常数时间的字段生成策略,实现了与元组条目数严格线性相关的时间复杂度($O(n)$)。这一特性使其能够支持教育大数据场景下的高吞吐量生成需求。

3 结论

文中针对教育结构化数据隐写中存在的修改痕迹显著、多语义维度字段协同性不足及安全可控性薄弱等问题,提出了一种基于密钥驱动与多字段协同生成的无载体隐写方法。通过主密钥分层派生子密钥(姓名生成、性别生成及成绩编码密钥),构建多语义维度字段的自适应生成路径,实现了高保真隐写成绩表的可控嵌入与隐蔽通信。理论模型表明,隐写容量由课程数、学生数及修改集合的宽松度联合决定,其中 Δ_3(16 值修改)在课程数较多时单科容量可达 4 bit,Δ_0(2 值修改)则适用于小规模场景的隐蔽性优先需求。

实验验证中,隐写表生成的姓名严格符合中文命名规范,各科成绩均值波动和方差变化均非常小,

且分数段分布与原始数据一致；通过 20 组随机秘密信息隐写测试，各科目统计特性稳定，无系统性偏差或规律性特征。此外，本研究从形式化安全模型和实验验证两个维度，系统性地评估了所提方法的安全性。与现有方法相比，文中所提方法在不可感知性、数据保真度及容量-隐蔽性均衡方面表现更优，为教育数据安全通信提供了兼顾实用性与安全性的创新解决方案。

参考文献：

[1] 彭飞，肖荻昱. 大数据时代国内外个人信息保护研究热点和演化趋势：基于科学知识图谱分析的文献计量方法[J]. 情报科学，2024，42(5)：102-112. DOI：10.3969/j.issn.1005-8095.2024.05.002.

[2] CHEN Y，WANG H，LI W，*et al*. A steganography immunoprocessing framework against CNN-based and handcrafted steganalysis[J]. IEEE Transactions on Information Forensics and Security，2024，19：6055-6069. DOI：10.1109/TIFS.2024.3409075.

[3] 付章杰，王帆，孙星明，等. 基于深度学习的图像隐写方法研究[J]. 计算机学报，2020，43(9)：1656-1672. 1DOI：10.11897/SP.J.1016.2020.01656.

[4] SU W，NI J，HU X，*et al*. Efficient audio steganography using generalized audio intrinsic energy with micro-amplitude modification suppression[J]. IEEE Transactions on Information Forensics and Security，2024，19：6559-6572. DOI：10.1109/TIFS.2024.3417268.

[5] 李敬轩，胡润文，阮观奇，等. 基于手工特征提取与结果融合的 CNN 音频隐写分析算法[J]. 计算机学报，2021，44(10)：2061-2075. DOI：10.11897/SP.J.1016.2021.02061.

[6] SHENG Qingxin，FU C，LIN Zhaonan，*et al*. Content-aware tunable selective encryption for HEVC using sine-modular chaotification model[J]. IEEE Transactions on Multimedia，2025，27：41-55. DOI：10.1109/TMM.2024.3521724.

[7] 李林聪，姚远志，张晓雅，等. 基于修改概率转换和非加性嵌入失真的视频隐写方法[J]. 电子与信息学报，2020，42(10)：2357-2364. DOI：10.11999/JEIT200001

[8] KOPTYRA K，OGIELA M R. Distributed steganography in PDF files：Secrets hidden in modified pages[J]. Entropy，2020，22(6)：600. DOI：10.3390/e22060600.

[9] 郝宇，施勇，薛质，等. Office XML 文档信息隐藏方法[J]. 计算机技术与发展，2017，27(10)：96-100. DOI：10.3969/j.issn.1673-629X.2017.10.021.

[10] XIANG Lingyun，OU Chengfu，ZENG Daojian. Linguistic steganography：Hiding information in syntax space[J]. IEEE Signal Processing Letters，2024，31：261-265.

[11] 秦川，王萌，司广文，等. 基于绝句生成的构造式信息隐藏算法[J]. 计算机学报，2021，44(4)：773-785. DOI：10.11897/SP.J.1016.2021.00773.

[12] KAUSHIK K，SHARMA G，NAROOKA P，*et al*. HTML smuggling：Attack and mitigation[C]// 16th International Conference on Security of Information and Networks (SIN). Jaipur，India：IEEE Press，2023：1-5. DOI：10.1109/SIN60469.2023.10474837.

[13] FENO H R. Unveiling hidden messages：A robust approach to detecting structural text steganography in office open XML documents[C]// 5th International Conference on Emerging Trends in Electrical，Electronic and Communications Engineering (ELECOM). Balaclava，Mauritius：IEEE Press，2024：1-6. DOI：10.1109/ELECOM63163.2024.10892148.

[14] 杨文秀，吴建荣，常潇，等. 基于 Excel 文件的信息隐藏方法，装置，设备及存储介质：202210749023.7[P]. 2025-04-05.

[15] WENDZEL S，CAVIGLIONE L，MAZURCZYK W，*et al*. A revised taxonomy of steganography embedding patterns[C]// Proceedings of the 16th International Conference on Availability，Reliability and Security. New York，USA：Association for Computing Machinery，2021：1-12. DOI：10.1145/3465481.347006.

（责任编辑：黄仲一　　英文审校：陈婧）