基于马尔可夫链的生成式区块链隐蔽通信模型

佘维[1,2,3]，荣欣鹏[1,3]，刘炜[1,2,3]，田钊[1,3]
（1. 郑州大学网络空间安全学院，河南 郑州 450001；2. 河南省网络密码技术重点实验室，河南 郑州 450001；
3. 郑州市区块链与数据智能重点实验室，河南 郑州 450001）

摘 要：为了解决目前区块链隐蔽通信中信道构建风险高、信息交叉、隐蔽性不足等问题，提出了一种基于马尔可夫链的生成式区块链隐蔽通信模型。首先，发送方使用文本数据集获取候选单词集并进行马尔可夫模型训练，获得转移概率矩阵，并生成哈夫曼树集合；随后，对需要传输的秘密信息二进制流进行迭代式哈夫曼解码，以获得一组符合正常语言与语义特征、可读性强的载密信息语句，利用生成式隐写方法完成秘密信息嵌入；然后，将该载密信息进行环签名后，作为正常交易发布到区块链网络中并完成打包和出块；最后，接收方利用相同的文本数据集获取转移概率权值哈夫曼树，逆向操作获得秘密信息二进制流。实验结果表明，相较于目前的同类模型，所提模型可进一步提高嵌入强度和时间效率，降低隐蔽信道构建风险，避免信息交叉，提升隐蔽性。

关键词：隐蔽通信；区块链；马尔可夫链；生成式文本隐写；环签名
中图分类号：TP309.2
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022194

Generative blockchain-based covert communication model based on Markov chain

SHE Wei[1,2,3], RONG Xinpeng[1,3], LIU Wei[1,2,3], TIAN Zhao[1,3]
1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China
2. Henan Key Laboratory of Network Cryptography Technology, Zhengzhou 450001, China
3. Zhengzhou Key Laboratory of Blockchain and Data Intelligence, Zhengzhou 450001, China

Abstract: To solve the problems of high channel construction risk, information crossover, and insufficient concealment in the blockchain covert communication, a generative blockchain-based covert communication model based on Markov chain was proposed. First, the text data set was used by sender to obtain the candidate words set and trained the Markov model to obtain the transition probability matrix, generated the Huffman tree set. Secret message to be transmitted was performed iterative Huffman decoding on the binary stream to obtain a set of highly readable carring-secret message statements that conformed to normal language and semantic characteristics, a generative steganography was used to complete secret message embedding. Then, the carring-secret message was ring-signed and published to the blockchain as a normal transaction packing and block generation were completed in the network. Finally, the same text data set was used by the receiver to obtain the Huffman tree of transition probability weights, the binary stream of secret message was obtained by reverse operation. Simulation results demonstrate that, compares with the current similar models, the proposed model can further improve the embedding strength and time efficiency, reduce the risk of covert channel construction, avoid information crossover, and improve the concealment.

Keywords: covert communication, blockchain, Markov chain, generative text steganography, ring signature

收稿日期：2022–06–20；**修回日期**：2022–09–08
通信作者：田钊，tianzhao@zzu.edu.cn
基金项目：河南省高校科技创新人才支持计划基金资助项目（No.21HASTIT031）；河南省重点研发与推广专项基金资助项目（No.212102310039, No.212102310554）；河南省重大公益专项基金资助项目（No.201300210300）；河南省网络密码技术重点实验室研究课题基金资助项目（No.LNCT2022-A04）
Foundation Items: Program for Science&Technology Innovation Talents in Universities of Henan Province (No.21HASTIT031), Key R&D and Promotion Project in Henan Province (No.212102310039, No.212102310554), Major Public Welfare Project of Henan Province (No.201300210300), Program for Henan Key Laboratory of Network Cryptography Technology (No.LNCT2022-A04)

0 引言

隐蔽通信通常被认为是在网络环境中满足现行协议或标准的通信双方制定相应规则进行不引起第三方注意的隐蔽信息传输行为[1]。随着个人计算机计算能力的大幅度提升与计算架构的快速发展，传统的保护隐私安全的方法受到了前所未有的挑战[2-3]。隐蔽通信是一种非常规的通信方法，近年来越来越多的研究者提出可将其作为对传统加密通信的有力补充手段之一。

研究者通常认为现代的隐蔽通信起源于 Simmons 提出的囚徒模型[4]。有别于常用的隐蔽信息传输媒介，区块链作为一种具备去中心化、去信任、不可篡改、开放共识等特点的分布式技术平台，其特性契合隐蔽通信的需求，可解决传统隐蔽通信自身的诸多痛点[5]。因此自 2018 年以来，研究者开始探索利用区块链构造隐蔽通信的方法。

Partala[6]首次尝试利用区块链作为媒介构建隐蔽通信信道，并提出区块链隐蔽信道（BLOCCE, blockchain covert channel）模型，该模型将信息隐藏至交易地址的最后一位，并顺序使用对应循环生成的交易地址使其保证秘密信息的顺序性。此后，研究者针对该模型的不足进行改进，提出改良的区块链隐蔽通信方法，并尝试降低其通信成本[7-8]。Guo 等[9]与蓝怡琴等[10]通过结合多层可链接自发匿名群签名实现混合，并引入新的椭圆曲线算法以及其他密码学技术的门罗币，在其区块链应用中构建隐蔽通信信道，利用了门罗币自身具备的高安全性以提高隐蔽信道的隐蔽性。She 等[11]提出了一种结合区块链和星际文件系统（IPFS, interplanetary file system）的双隐写模型，这种协作模型有效解决了区块链上构建隐蔽通信信道的隐藏容量问题，且由于链上存储的特殊信息较少，因此大大提高了信道隐蔽性。余维等[12]提出了一种面向纯文本信息隐藏的区块链隐蔽通信模型，该模型相比于 BLOCCE 在一定程度上提高了嵌入强度；随后，余维等[13]结合属性基加密（CP-ABE, ciphertext-policy attribute-based encryption）技术与基于生成式对抗网络（GAN, generative adversarial networks）的图像隐写术提出了一种可隐藏敏感文档和发送者身份的区块链隐蔽通信模型。姜鹏坤等[14]提出了一种利用哈希算法构建免传输密码表匹配二进制位的区块链隐蔽通信方案。近几年，基于区块链构建隐蔽通信信道的研究不断深入，区块链 3.0 时代应用落地的需求日益强烈，但现有的研究中仍存在以下尚未完全解决的问题。

问题 1 预处理或信道构建过程中通常需要人工干预，为信道的隐蔽增添了不确定性；同时需要通过预先协商的方式确定开始标识符、结束标识符、信息接收对象或密钥等关键信息以达到通信同步，增加了隐蔽信道的构建成本和构建风险。

问题 2 一些基于区块链交易地址进行信息隐写的方案较简易，隐蔽性较差，难以抵抗常规的隐蔽信道检测。且由于区块链的透明性，所有节点都可查看区块信息，若该区块链网络中存在其他隐蔽通信对象，容易出现“信息交叉”的现象，即互相可探测其他隐蔽信道传输的秘密信息。

问题 3 传统隐蔽通信通常需要引入第三方公共媒介，若通过数学方法统计分析则可能存在一定身份暴露的风险。而完全严格基于区块链平台的隐蔽信道构建方法的隐蔽容量过小，导致完成一次通信过程的时延较长，或对出块交易顺序等条件要求苛刻，很难满足实际应用场景需求。

为解决以上问题，本文提出了一种基于马尔可夫链的生成式区块链隐蔽通信（MC-GBCC, generative blockchain-based covert communication based on Markov chain）模型。首先，发送方将秘密信息转换为二进制流，同时使用与接收方相同的文本数据集进行马尔可夫训练[15]，再使用训练得到的条件转移概率矩阵构建哈夫曼树；然后，针对二进制流进行哈夫曼解码过程得到可读性较强的文字信息；最后，通过环签名进行签名后向区块链中发布带有该文字信息的交易[16]，接收方查询到该信息后通过逆向操作便可获得该秘密信息。

本文的贡献主要有以下几个方面。

1) 本文在区块链隐蔽通信信道构建中使用生成式文本隐写技术，大大降低了人工干预可能带来的信道不稳定的风险，满足区块链网络合约执行自动化这一设计思想，更有利于隐蔽信道在区块链网络中的构建，同时为其带来更强的可拓展性与灵活性，减少了信道暴露的风险，一定程度上解决了问题 1。

2) 本文在隐写过程中引入马尔可夫链与哈夫曼编码结合的形式，在保证其信道隐蔽性的前提下，免除了预协商过程，实现了完全的异步式通信，不需要协商隐蔽通信位置以及通信的开始与结束，降低了隐蔽传输的成本，进一步解决了问题 1。

3) 不同“通信对”只需维持不同的模型训练文本数据集即可保证该通信双方信道的完全隐蔽，并将不同隐蔽通信对隔离开，不会存在其他用户“误读”的情况，解决了问题 2。

4) 相较于以往基于区块链区块结构的隐写方法，在不接入第三方平台的情况下大大提高了信道隐藏容量，使其在隐藏容量与潜在风险两者中达到相对平衡；此外，严格将区块链作为通信媒介并引入环签名技术，实现了现实对象与网络节点的剥离，打破了可能存在的映射关系，使实际隐蔽通信对象完全实现现实身份的隐藏，区块链的透明性使通信发送方不需要将携带秘密信息的交易直接发给信息接收方，解决了问题 3。

1　相关技术

本节主要介绍区块链隐蔽通信、马尔可夫链、生成式文本隐写与环签名以及哈夫曼编码的相关知识。

1.1　区块链隐蔽通信

比特币的诞生将区块链带入人们视野之中，智能合约的引入大大推动了区块链的爆发式发展，近年来，区块链已经开始在能源、医疗、教育、交通等领域崭露头角[17]。

区块链作为一种结合了密码学、数学、博弈论等多学科的分布式技术平台，其自身具备匿名性、不可篡改、可追溯、合约执行自动化、去信任等多种特性[18-19]。传统的隐蔽通信通常利用第三方媒体的单一信道定向发送方式，由于系统中心化以及极易受到网络环境的影响，可能降低隐蔽信道的可靠性与隐蔽性，而区块链技术为隐蔽通信信道的构建提供了更多的可能性，其特性能够弥补传统隐蔽通信方法易受篡改、可靠性不稳定、信道单一等不足。区块链隐蔽通信模型如图 1 所示。

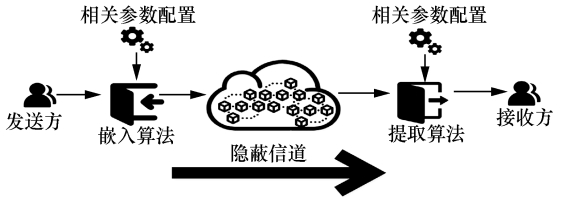

图 1　区块链隐蔽通信模型

区块链隐蔽通信按照构建原理可分为基于区块结构、基于外部载体、基于业务操作时间和基于系统机制 4 类[20]。本文所提 MC-GBCC 模型可归类至基于结构的区块链隐蔽通信。

1.2　马尔可夫链

马尔可夫链是一组离散随机变量的集合，在状态空间中随机地从一个状态转移到另一个状态，且具备无记忆性，即在给定当前知识或信息的前提下，下一步的状态只与当前状态有关，与过去状态（即以前的状态）无关。马尔可夫链根据时间状态是否离散可分为连续时间马尔可夫链与离散时间马尔可夫链。本文使用的是离散时间马尔可夫链，即时间和状态都是离散取值的[21-22]。马尔可夫链时间与状态空间对应关系如图 2 所示。

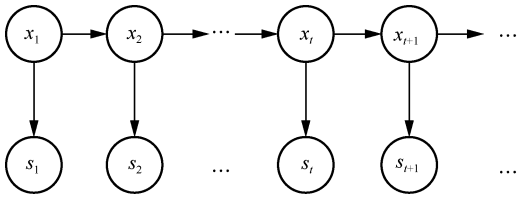

图 2　马尔可夫链时间与状态空间对应关系

给定随机变量集合 $X=\{x_n, n>0\}$，且该随机变量可能的取值都在有限状态集 $S=\{s_n, n>0\}$ 内，即 $x_i=s_i, s_i\in S$，若符合以下条件

$$p(x_{t+1}\mid x_t, x_{t-1}, \cdots, x_1\}=p(x_{t+1}\mid x_t) \tag{1}$$

则集合 X 被称为马尔可夫链；集合 X 的取值被称为状态空间；马尔可夫链在状态空间内的取值被称为状态，其性质被称为马尔可夫性质，即“无记忆性”。通过马尔可夫链的模型转换，便可以得到各状态之间的转换概率，从而得到条件转移概率矩阵（也称状态分布矩阵），例如

$$\boldsymbol{P}_{mn}=\begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{pmatrix} \tag{2}$$

其中，有

$$\sum_{j=1}^{n} p_{ij}=1, i\in\{1,2,\cdots,m\} \tag{3}$$

1.3　生成式文本隐写与环签名

现有的文本隐写算法通常可分为两大类：修改式隐写算法与生成式隐写算法[23]。近年来，随着自然语言处理技术的发展，以文本生成技术为基础的生成式隐写算法的发展愈加成熟。该类算法有别于传统文本隐写，其不需要人为选择文本载体，可自动生成载体并进行秘密信息嵌入，大大降低了人为因素所带来的不确定性和算法应用落地的门槛。

环签名是一种数字签名方案，目的在于允许用户在环内保持匿名的前提下代表环进行数字签名，其环成员中没有管理者和可信中心，成员权限地位平等，同时签名过程不需要成员合作，只需利用自身私钥以及集合中其他成员的公钥即可独立完成签名过程，其他成员并不知道自己已被包含其中。环签名满足如下性质。

1) 无条件匿名性。攻击者确定签名是由环中哪个成员生成的概率极低，即使攻击者获取了所有环成员的私钥，能确定真实签名者的概率也不超过 $\frac{1}{r}$，其中 r 为环成员个数。

2) 正确性。签名能被所有人验证。

3) 不可伪造性。环中其他成员与外部攻击者均无法伪造真实签名者签名。

1.4 哈夫曼编码

哈夫曼编码是一种在压缩字符编码的同时不需要添加额外分隔符，且能保证解码唯一性的不定长编码[24]。该编码方法完全依照字符出现概率来构造平均字长最短的编码，权值越大的节点越接近根节点，也被称为最佳编码方式，其构建过程如图 3 所示。

2 MC-GBCC 模型说明

MC-GBCC 模型如图 4 所示。本节分为预处理、嵌入、传输、提取 4 个过程对 MC-GBCC 模型进行说明。

2.1 预处理过程

秘密信息发送方维持一份与接收方相同的文本数据集 A，并且双方均使用数据集 A 进行马尔可夫模型训练，因此双方可得到相同的模型及其条件转移概率矩阵。一般希望数据集 A 足够大，通常数据量越大的文本数据集进行马尔可夫模型训练之后得到的生成式文本可读性越强，其对应的隐蔽信道的隐蔽性也越强。

在嵌入过程之前要先对所持有的文本数据集 A 进行文本预处理，操作如下

$$L(s_i)=u(h(g(f(s_i)))),\mathrm{se}_i\in\mathrm{SE} \tag{4}$$

其中，$f(x)$ 表示删除 x 中的特殊字符，$g(x)$ 表示删除 x 中的网页链接，$h(x)$ 表示 x 全部转换为小写，$u(x)$ 表示删除 x 中的表情符号，SE 表示文本数据集 A 中以结束符为标志的句子集合，即 $\mathrm{SE}=\{\mathrm{se}_1,\mathrm{se}_2,\cdots,\mathrm{se}_n\}$。

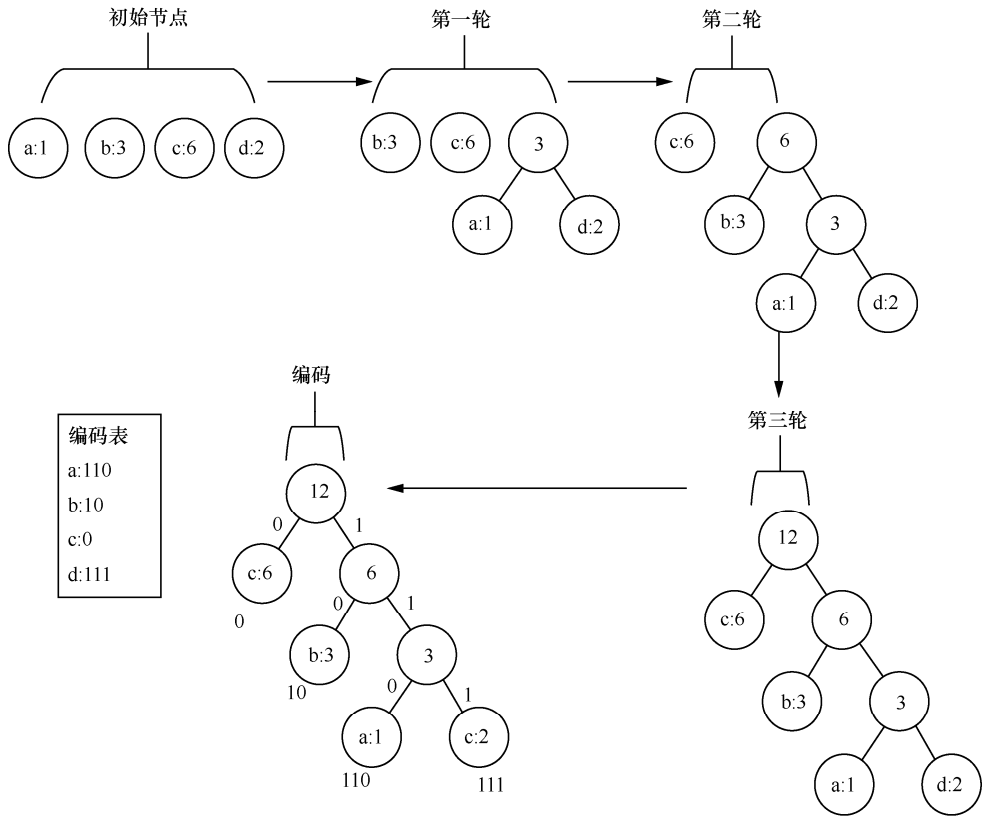

图 3 哈夫曼编码构建过程

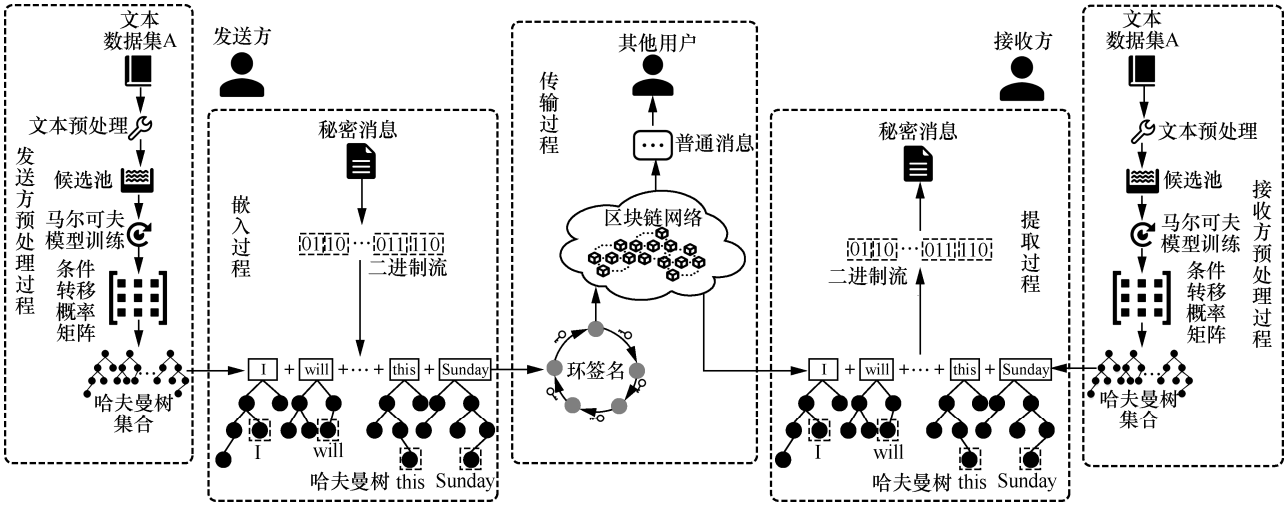

图 4　MC-GBCC 模型

接着，对预处理后的文本数据集 A 进行马尔可夫模型训练，训练步骤如下。

1) 建立一个包含数据集内全部单词的集合 $D=\{d_1,d_2,\cdots,d_{\text{num}}\}$，即给定的随机变量有限集合，将该字典中的所有单词放入候选池中作为备选。

2) 每个句子 se_i 的相应位置和单词可对应至马尔可夫链中的离散取值的时间与状态，即

$$\text{ST}=\{\text{word}_{\text{se}_i,1},\text{word}_{\text{se}_i,2}\cdots,\text{word}_{\text{se}_i,l_i}\},\text{word}_{\text{se}_i,j}\in D \tag{5}$$

其中，$i\in\{1,2,\cdots,n\}$，$j\in\{1,2,\cdots l_i\}$，se_i 为句子编号，l_i 为该句的单词个数。

3) 利用一阶马尔可夫链（实际使用中可根据需求采取二阶或三阶马尔可夫链使之生成更加具备可读性的文本）计算每个单词的转移概率为

$$p(\text{word}_{\text{se}_i,r}=d_j\mid\text{word}_{\text{se}_i,r-1})\approx\frac{\text{count}(\text{word}_{\text{se}_i,r-1},d_j)}{\text{count}(\text{word}_{\text{se}_i,r-1})} \tag{6}$$

同时满足

$$\sum_{i=1}^{sn}p(\text{word}_{\text{se}_i,r}=d_j\mid\text{word}_{\text{se}_i,r-1})=1 \tag{7}$$

其中，SN 表示集合 S 的元素个数。

通过以上三步即可得到所需的条件转移概率矩阵为

$$\boldsymbol{P}_{nn}=\begin{pmatrix}p_{11}&\cdots&p_{1n}\\ \vdots&\ddots&\vdots\\ p_{n1}&\cdots&p_{nn}\end{pmatrix} \tag{8}$$

4) 初始哈夫曼树 t_1 表示所有可能位于初始位置的单词，利用出现概率作为权值构成初始哈夫曼树 t_1；除初始位置外，以某个单词 word_i 为基准转移到下一个位置 word_{i+1} 的所有可能的每个单词转移概率 $p(\text{word}_i)$ 为权值构成哈夫曼树，即对应矩阵 $\boldsymbol{P}_{nn}$ 中的某一行构成哈夫曼树（舍弃概率为 0 的情况）。这样的情况共有 $n-1$ 种，即除初始哈夫曼树 t_1 以外还可构造 $n-1$ 个状态转移哈夫曼树。定义以下函数

$$t_i=o(\text{word}_{i-1},p(\text{word}_{i-1})) \tag{9}$$

其中，$i\in\{2,3,\cdots,n\}$，$o(x,y)$ 表示哈夫曼树构造算法。利用该函数构造各状态哈夫曼树并与 t_1 共同得到哈夫曼树集合 $T=\{t_1,t_2,\cdots,t_n\}$。

将训练得到的马尔可夫模型 M 与哈夫曼树集合 T 保存，以便未来使用该数据集与对方通信。以上预处理过程只需在双方第一次通信时进行即可。

2.2　嵌入过程

发送方嵌入秘密信息步骤如下。

1) 发送方将需要传输的秘密信息文本 secret_message 转换为二进制流 bin_1，即

$$\text{bin}_1=z(\text{secret\_message}) \tag{10}$$

其中，$z(x)$ 表示将文本信息 x 转换为二进制流。

2) 从初始哈夫曼树 t_1 中查找与二进制流开始阶段的对应单词，并将其放入初始位置，即

$$\text{word}_1=c(\text{bin}_1,t_1) \tag{11}$$

$$\text{bin}_2=q(\text{bin}_1,\text{word}_1) \tag{12}$$

其中，$c(x,y)$ 表示哈夫曼解码函数；$q(x,y)$ 表示每完成一次解码过程则将 y 对应的二进制流，即解码读取的前 n 位二进制流从长二进制流 x 中删去。

3) 完成初始二进制流信息嵌入之后，继续对剩

余信息进行嵌入，即

$$\text{word}_i = c(\text{bin}_i, r(\text{word}_{i-1})) \tag{13}$$

$$\text{bin}_{i+1} = q(\text{bin}_i, \text{word}_i) \tag{14}$$

其中，$i \in \{2,3,4,\cdots\}$，$r(x)$ 表示查找单词 x 在集合 T 中对应的状态哈夫曼树。

4) 重复步骤 3)直至二进制流信息嵌入完毕，若最后位置的二进制流无法顺利解码，则在末尾自动补 0 直至完全解码完毕。例如，二进制流若为 0110…01111，二进制流解码至最后的 11 无法在哈夫曼树的叶子节点找到（无法解码），则在末尾自动补 0 直至能够完成解码。

5) 解码完毕后，得到 $\text{len}-1$ 个 word_i，其中，$i \in \{2,3,\cdots,\text{len}\}$，len 为解码得到的载密语句单词个数。并得到由 word_1 与 word_i 组成的向量 $\mathbf{SEN} = (\text{word}_1, \text{word}_2, \cdots, \text{word}_{\text{len}})$，将该向量按照其排列顺序组成语句作为载密语句输出。

2.3 传输过程

秘密信息嵌入过程完成后，得到载密信息 cover-message，该载密信息是可读性较强的普通文本信息。传输过程中，秘密信息发送方创建一笔交易 Transaction(cover-message)。其中，载密信息 cover-message 存入交易的 data 字段，而由于区块链的公开透明性，任何本区块链网络中的用户均可访问链上的所有信息，因此为提高信道隐蔽性，这里交易接收方可任意选择区块链网络中的用户，并不要求交易接收方为本次隐蔽通信的实际秘密信息接收方。

为隐藏实际信息发送方身份，需要进行环签名操作，这里将简要描述环签名算法步骤。首先定义以下函数

$$C_{k,v}(y_1,\cdots,y_n) = E_k(y_n \oplus E_k(\cdots \oplus E_k(y_1 \oplus v)\cdots)) = v \tag{15}$$

其中，E_k 为对称加密算法，k 为 E_k 对应的密钥。对包含该秘密信息的区块链交易 Transaction (cover-message) 进行环签名，环签名中各成员公钥分别为 $P_1, P_2, \cdots, P_n$，消息发送方拥有公钥 P_s 及其对应的私钥。令 $k = \text{hash}(\text{Transaction(cover-message)})$，即交易的 hash 值作为环签名递推公式（对称加密）的密钥，发送方随机选取一个值 v。然后，除发送方 x_s 以外随机选取 $n-1$ 个值 $\{x_1, x_2, \cdots, x_{s-1}, x_{s+1}, \cdots, x_n\}$，并利用对应的公钥 P_i 通过 $y_i = g_i(x_i)$ 计算得到相应的 $n-1$ 个值 $\{y_1, y_2, \cdots, y_{s-1}, y_{s+1}, \cdots, y_n\}$。接下来，令 $C_{k,v}(y_1, y_2, \cdots, y_s, \cdots, y_n) = v$，计算令等式成立的 y_s。可以把 y_s 看作通过公钥 P_s 加密得到，而发送方拥有 P_s 对应的私钥，因此可通过 $x_i = g_i^{-1}(y_i)$ 解密 y_s 得到 x_s。最后，利用递推式得到关于发送方创建的交易消息 Transaction(cover-message) 的环签名，该签名是一个 $2n+1$ 元组 $(P_1, P_2, \cdots, P_n; v; x_1, x_2, \cdots, x_n)$，区块链网络用户通过该签名无法获知真正的信息发送方。

发送方将通过环签名后的交易发布到区块链网络中，并在各节点中进行广播，被验证交易正确有效之后打包至区块内经过共识更新该新生成的区块，至此载密信息完成上链，包括隐蔽通信信息接收者在内的所有区块链用户都可收到并查看链上信息。

2.4 提取过程

秘密信息接收方拥有与发送方相同的文本数据集 A，首先使用与预处理过程同样的方法构造哈夫曼树，如 2.1 节所述，由于使用了完全相同的文本数据集与处理算法，因此可生成完全相同的集合 $T = \{t_1, t_2, \cdots, t_n\}$。同样地，处理一次之后可将模型 M 与哈夫曼树集合 T 保存以便以后的双方通信，降低其通信成本。接收方提取步骤如下。

1) 接收方遍历区块链网络中新生成的交易信息，即

$$\text{message} = \gamma(\text{time}, \text{Transaction(text)}) \tag{16}$$

其中，time 表示设定的区块链交易读取时间间隔，函数 $\gamma(x,y)$ 表示每隔 x 时间间隔遍历 y 中新生成的交易字段。

2) 将 message 设定为向量 **MES**，即

$$\mathbf{MES} = (\text{word}'_1, \text{word}'_2, \cdots, \text{word}'_{\text{len}}) \tag{17}$$

其中，word'_i 表示载体信息中的单词，i 的取值为 $\{i \mid i \in N^*, 2 \leqslant i \leqslant \text{len}\}$，len 表示单词个数。

3) 利用训练得到的哈夫曼树集合 T 对向量 **MES** 中的元素 word'_i 逐一解码，并定义以下函数

$$\text{code}_i = v(\text{word}'_i, T), i \in N^*, i \in [1, \text{len}] \tag{18}$$

其中，$v(x,y)$ 表示在集合 y 中寻找合适的元素对单词 x 进行解码得到二进制流，并将解码得到的二进制 code_i 放入向量 $\boldsymbol{C}$ 中，即

$$\boldsymbol{C} = (\text{code}_1, \text{code}_2, \cdots, \text{code}_{\text{len}}) \tag{19}$$

4) 若不能完全执行步骤 3)，则证明该消息并不是发给自身的秘密信息；若能得到完整有序集合 C，则该有序集合元素组成的二进制流为发送给自身的秘密信息，即

$$\text{bin}_1 = \varphi(\text{code}_1, \text{code}_2, \cdots, \text{code}_{\text{len}}) \tag{20}$$

其中，$h(x_1,x_2,\cdots,x_n)$ 表示将二进制 $x_1,x_2,\cdots,x_n$ 进行拼接，拼接得到的秘密二进制流 bin_1 进行逆操作，即

$$\text{secret\_message}=z^{-1}(\text{bin}) \tag{21}$$

其中，$z(x)$ 表示将秘密信息文本 x 转化为二进制流。接收方通过上述步骤得到秘密信息文本 secret_message。

由于区块链网络的透明性，网络中其他用户可任意查看区块交易信息，但无法辨别并区分含有秘密信息的交易信息，只可查看到合法交易内容，且由于该交易信息是可读性较强的文本信息，进一步增强了信道的隐蔽性，使其他用户无法察觉该信道的存在性。

3　仿真实验与性能分析

本节介绍了仿真实验，并结合实验结果对 MC-GBCC 模型进行分析。实验采用的硬件平台为 PC，处理器为 AMD 5800H Ryzen 7@3.20 GHz，内存为 16.0 GB。

3.1　仿真实验

仿真实验在 Ubuntu 系统下使用 FISCO BCOS 模拟区块链网络，在区块链网络中创建 20 个账户，各账户信息如表 1 所示。

假设账户 1 是隐蔽通信发送方在 FISCO BCOS 区块链网络中的账户，账户 2 是接收方账户。在构建隐蔽通信信道之前，双方约定用于马尔可夫链训练的文本数据集，通常马尔可夫模型训练要求该文本数据集由人书写并且具备足够数量，本文实验中选用 Go 等[25]公开的用于情感分析的从 Twitter 中提取的 1 600 000 条评论，该类型文本更加符合人类日常语言习惯。在实际应用场景之下，为便于普通用户使用也可选择较大单词量的电子书。

预处理过程中，首先利用式(4)对文本数据集进行文本预处理，部分预处理后文本内容如表 2 所示。然后，利用式(5)～式(9)进行马尔可夫链训练，训练得到的模型可存储至本地用于未来与该接收方的通信。表 3 为通过该模型随机生成的文本。

表 2　部分预处理后文本内容

行号	内容
1	were you drinking out of the forgotten table drinks
2	i baked you a cake but i ated it
3	i hate when i have to call and wake people up
⋮	⋮

表 3　随机生成的文本

行号	内容
1	it is almost uninterrupted summer
2	it should have triumphed over the red flower
3	it is said we proceeded to show him for this part of course

根据训练结果可知，该模型生成文本基本符合人类语言习惯，具有较强隐蔽性。

发送方希望传输的秘密信息为“we will do it tomorrow”。在嵌入过程中，发送方利用式(10)将文本信息转换为二进制流“1110111110010110000011101111101001110110011011001000000110010011011111000001101001111010010000011101001101111”

表 1　FISCO BCOS 模拟区块链网络账户信息

账户	公钥	地址
1	MFYwEAYHKoZIzj0CAQYFK4EEAAoDQgAE1AmpGkJIvbwXh0W5LoSz/3Dz4NgWyCaj2zvJzQ0oeVHKd3RZ2LsCxI9cL1oJ6EeC4nFb4K6e/hqOcQrBkEyQfQ==	0x4bfb4f1264cbb514efe0e79a46b20414923f8793
2	MFYwEAYHKoZIzj0CAQYFK4EEAAoDQgAE610GR0AyCCw78xnzXwIg5JeV2AqolzE2sH11pIfyJ6DJ0SP231rUlhsBrPJ/GXIqwMWdpbftWc57sQ+XSSwnIg==	0x30a92712e37b6073be81c1e34f723a9087d7fe44
3	MFYwEAYHKoZIzj0CAQYFK4EEAAoDQgAEf+aglmFKjTAqjQ6CLC/nqt5GL0lU1Xvfbc2RfDOabl3hGN5+wB6Z4qRaDZ8ipZ23cEjlVJlZIIv2zCCeJFwjg==	0x1851cddaad12516966af0dff981948b6651b3676
4	MFYwEAYHKoZIzj0CAQYFK4EEAAoDQgAEEyyhr/81h6ddsRFfIBsWXmq/M2SSyC/xR8UfbXKrlUtLod6GdOkM1OkHHP5H1TZYpxw2pV3MsZOCaDzvoINi8g==	0xa12c5e342abe3fac5a085d62db38371d507d23c7
5	MFYwEAYHKoZIzj0CAQYFK4EEAAoDQgAEg+5PW6D7sbBXsS0PkuAKUf9uyimKPuvtFAu/+Oa8h7uGHRCYgSYaR2dgzP2VstgUX2UHOBKdslSYmyLAUegbDg==	0x9b89659737c90a697459fb3e70759774559b34ce
⋮	⋮	⋮
20	MFYwEAYHKoZIzj0CAQYFK4EEAAoDQgAEC7xmrxLLJESpjah0Gn1XJwh5g2sww5nx00w4C+loMEE0E4E6ogAtHCp+7LLkuSsH7fw/WY/QOQacuay9IjZ9Ow==	0x5f761a44f9d99eda0e3dca67bdd10638771eed59

11101101110111111100101110010110111111110111”。这里使用 ASCII 码将字符转换为整数，再将整数转换为二进制。

接下来，利用式(11)～式(14)进行秘密信息嵌入，得到的载密文本为“i hate when i have so much school work to do and does anyone know if there is a book for student”。

在传输过程中发送方不需要直接选择接收方账户，而可任意选择账户发送交易信息。发送方将生成的文本放入交易 data 字段中，并利用式(15)进行环签名，环签名后向区块链网络中发布交易，交易经过验证、共识后打包成块，为便于观察，将每个区块打包交易数设置为 1。区块部分字段结构和交易部分字段结构分别如表 4 和表 5 所示。

由表 5 可知，载密信息已成功上链。提取过程中，接收方利用式(16)～式(20)进行接收秘密信息的尝试，每隔一段时间主动利用通过文本数据集 A 产生的哈夫曼森林对交易的 data 字段进行编码，在对表 5 中的交易 data 字段编码时，能顺利得到一串匹配的二进制流“11101111100101100000111011111010011101100110110010000011001001101111100000110100111101001000001110100110111111011011101111111001011100101101111111110111”，接着利用式(15)相关函数验证环签名，验证通过后利用式(21)进行该二进制流的逆向操作即可得到秘密信息“we will do it tomorrow”。至此，一次完整的隐蔽通信流程结束，此后双方通信中不需要预处理过程，只需发送方嵌入过程和传输过程、接收方提取过程即可。

3.2 效率分析

本节主要从嵌入强度、时间效率两方面对 MC-GBCC 模型性能通过实验进行分析。

1) 嵌入强度

Houmansadr 等[26]认为隐蔽通信的传输效率可定义为每个隐蔽数据流所传递的隐蔽信息的比特数，对应本文模型为秘密信息发送方所提交的每笔交易信息（即生成的可读性较强的文本）中所包含的秘密信息的比特数，因此本文模型与其他区块链隐蔽信道嵌入强度可表示为

$$r = \lim_{N\to\infty} \frac{K}{N} \tag{22}$$

其中，r 为每个隐蔽通信数据流所携带的秘密信息的比特数，K 为使用 $N+1$ 个隐蔽信息数据流发送的隐蔽信息比特数。

为保持对比的相对合理性，本节选择具有较强通用性的区块链隐蔽通信信道构建方法[6-10,12]与本文模型对比，条件设定为每个区块内打包一笔交易。本文利用模型生成 1 000 轮并编码计算其可携

表 4 区块部分字段结构

字段	内容
hash	0x8035ddf8cbaccde09e6c8e8f8ca9c234891592152c08be6776c3e1e958cf6b43
TransactionHash	0xd4ebcc34d085c4486313f727aa2b05eb69bcca402b1c905f1b8caffcb7e64028
parentHash	0xce5bbcbd5ff2445780f19ba8e14b6ac04da26ede9dfb70f7885a8bfb54acff3f
transactionsRoot	0x82e86fa5a9cccdb1aaffd9ef75ff64746cb50282b7a3ab9c253d70edc4734078
receiptsRoot	0x31fedd393d3f418ebb75f6c3b8320c71c34effbf326d629535dbef54051d7b86
stateRoot	0x490d0bed041454dcd8b314c22e71d234a11d8b987df62cbbbf8625be5117267f
sealer	156fa8be62dda0bc318f71c713e8fd50365269cd8f1018abd4977d1c7fb7017bfd1c479555063891730ffb48a547ea19a534b2f782893847b20e5c124c6aa5fd
timestamp	0x18109da9d51

表 5 交易部分字段结构

字段	内容
hash	0xd4ebcc34d085c4486313f727aa2b05eb69bcca402b1c905f1b8caffcb7e64028
blockHash	0x8035ddf8cbaccde09e6c8e8f8ca9c234891592152c08be6776c3e1e958cf6b43
from	0x4bfb4f1264cbb514efe0e79a46b20414923f8793
gas	0x419ce0
data	i hate when i have so much school work to do and does anyone know if there is a book for student
to	0xe03cd2fde34edb14424271781f7e7b4ad2c58a53
gasPrice	0x51f4d5c00

带最大秘密信息的平均值。

嵌入强度对比如表 6 所示。相比于其他基于区块字段结构进行隐蔽信道构建的模型，本文模型的嵌入强度具有较大的优势。文献[7]提出对地址空间的低 α 位进行秘密信息嵌入，但实际操作中 α 的增大会大大增加哈希函数计算的时间开销，严重降低系统效率，因此通常 α 取值较小。

表 6　　嵌入强度对比

模型	嵌入强度/bit
文献[6]	1
文献[7]	α
文献[8]	16～32
文献[9]	256
文献[10]	256
文献[12]	2
本文模型	449

2) 时间效率

时间效率同样是判断隐蔽通信信道性能的重要指标之一。除区块链网络交易发布时延外，影响本文模型时间效率的关键因素在于秘密信息的嵌入过程与提取过程。本文利用上述预处理后的文本数据集进行马尔可夫模型训练，并利用训练后的模型进行秘密信息嵌入与提取。

为了便于比较相关区块链隐蔽通信模型的时间效率，统一嵌入相同秘密信息量为 8 bit，本文模型模拟嵌入过程与提取过程各 50 次，嵌入过程与提取过程时间如图 5 所示。从图 5 可以看出，嵌入过程与提取过程时间为 1.40～1.75 s，并无较大波动。

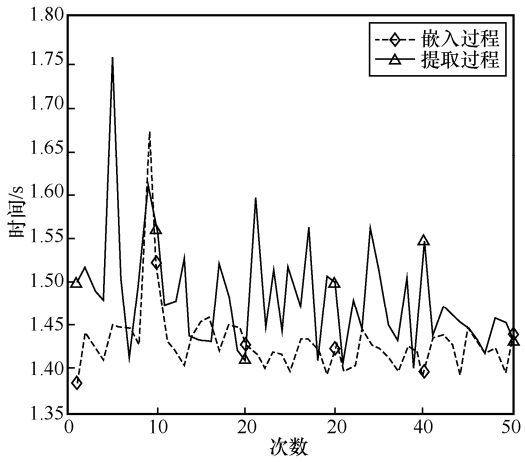

图 5　嵌入过程与提取过程时间

若考虑到区块链网络自身每秒处理事务数（TPS, transaction per second），本文模型时间效率仍能达到秒级。文献[6]的隐蔽信道构建模型发送 8 bit 信息的时间大于 30 min；文献[8]多次重用 Vanitygen（比特币地址生成软件）生成地址嵌入信息，但随着每个地址嵌入位数的增加其难度呈指数级增长，若每个地址中直接嵌入 8 bit 信息则地址生成时间将达到小时级，若采取降低每个地址嵌入信息比特数但提升发送交易数的方法，其时间效率过于依赖区块链网络 TPS 且对交易顺序有严格要求；文献[11,13]由于引入了图像隐写术与 IPFS，隐蔽通信时间为分钟级；文献[12]使用了基于空格法的信息隐藏技术，其每笔交易内所含的信息比特数较低，因此在传输与本文模型相同信息量的秘密信息的情况下传输时间为分钟级。

本文模型与文献[6,8,11-13]模型的传输时间对比如表 7 所示，在相同嵌入量下，本文模型所使用的嵌入方法在时间效率上具有一定优势。

表 7　　传输时间对比

模型	传输时间
文献[6]	>30 min
文献[8]	小时级
文献[11]	分钟级
文献[12]	分钟级
文献[13]	分钟级
本文模型	秒级

3.3　安全性分析

本文所提 MC-GBCC 模型降低了隐蔽信道构建风险、避免了信息交叉、提升了隐蔽性，本节从上述三方面进行安全性分析。

3.3.1　降低隐蔽信道构建风险

MC-GBCC 模型基于马尔可夫链进行生成式文本隐写，一方面不需要人为操作即可实现载体文本与载密文本的自动生成，降低了人为误操作或选择载体不恰当所带来的信道暴露风险；另一方面通信双方不需要在区块链网络中交换信道关键参数，只需双方持有同一份文本数据集即可实现完全异步式通信，在缩减信道构建成本的同时降低了构建风险。

3.3.2　避免信息交叉

每对隐蔽通信方采取不同的文本数据集，能够保证即使区块链网络中存在其他使用同一方法构

建隐蔽信道的通信对，仍不暴露己方信道，且不存在误读其他信道数据的风险，避免了其他区块链隐蔽通信模型可能出现的信息交叉现象。

3.3.3 提升隐蔽性

1) 本文模型生成的载密信息仍具备较强可读性，不易引起怀疑。在信息论领域中，perplexity（困惑度）通常用于度量概率分布或概率模型预测样本的好坏，同样地，在自然语言处理领域中，困惑度可用于度量句子的质量，其计算式为

$$\text{perplexity} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log p(s_i)} \tag{23}$$

其中，$p(s_i)$为句子s_i中单词的分布概率，N为句中单词总数。困惑度越小，生成文本与训练文本统计分布越一致，则模型越好，越不容易引起怀疑，但实际上困惑度很难界定一个安全区间。针对MC-GBCC模型，本节完全随机生成了100个已嵌入秘密信息的生成语句，并根据式(23)计算困惑度，结果如图6所示。

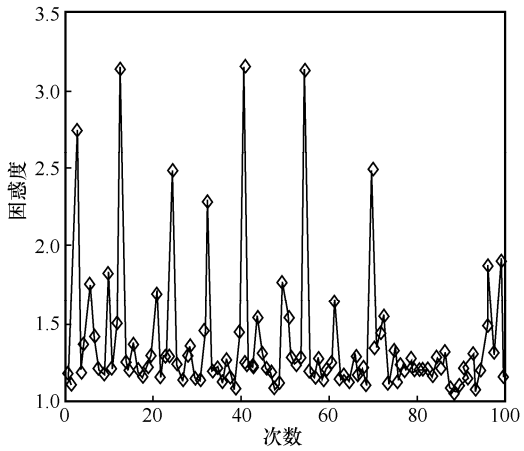

图 6 困惑度计算结果

由图6可知，MC-GBCC模型困惑度主要集中在1～3。而在文献[15]中，相同嵌入率下的困惑度可能超过200，与之相比， MC-GBCC模型困惑度较低且更稳定，因此，本文模型在信息嵌入过程中自动生成的语句不易引起怀疑，具有较强的隐蔽性，在未知足够信息下证明该隐蔽信道存在基本不可行。

2) 区块链网络打破了现实身份与虚拟身份的映射关系，且包含载密信息的交易接收方并不一定与实际秘密信息接收方相同。

3) 环签名技术的无条件匿名性融入区块链交易发布之中进一步混淆了发送方身份，即使攻击者获得所有环成员的私钥，能确定真实发送方的概率也不超过$\frac{1}{r}$，其中r为环成员个数。

以上三点共同保证了本文模型具备足够高的隐蔽性。

综上，本文所提 MC-GBCC 模型与文献[6-13]模型相比，在嵌入强度与时间效率方面具有明显优势，解决了其他区块链隐蔽通信模型所未解决的通信对隔离问题，并且通信过程中免除了预协商过程，大大降低了信道构建风险，实现了异步式通信，并提升了隐蔽性。不同模型的综合对比如表8所示。

4 结束语

本文提出了一种基于马尔可夫链的生成式区块链隐蔽通信模型，一定程度上解决了目前区块链隐蔽通信中信道构建风险高、信息交叉、隐蔽性不足等问题。所提模型结合马尔可夫链与哈夫曼编码技术实现了隐蔽通信的完全异步式，并降低了信道构建的潜在风险，通过约定文本数据集将不同通信

表 8 不同模型的综合对比

模型	嵌入强度/bit	传输时间	特征
本文模型	449	秒级	异步式、降低构建风险、隔离通信对、提升隐蔽性
文献[6]	1	>30 min	第一个可证明安全区块链隐蔽信道
文献[7]	α	—	基于文献[6]改进型
文献[8]	16～32	小时级	基于文献[6]改进型
文献[9]	256	—	可拓展隐藏容量
文献[10]	256	—	安全性取决于门罗币的安全性
文献[11]	—	分钟级	引入外部载体
文献[12]	2	分钟级	文本型秘密信息嵌入
文献[13]	—	分钟级	结合外部载体与生成对抗网络思想

双方隔离开，避免了信息交叉，进一步提高了隐蔽性，并引入环签名隐藏发送方身份。实验结果表明，对比同类区块链隐蔽通信模型，所提模型具备更优秀的嵌入强度与时间效率以及足够的安全性。

参考文献：

[1] 李彦峰, 丁丽萍, 吴敬征, 等. 网络隐蔽信道关键技术研究综述[J]. 软件学报, 2019, 30(8): 2470-2490.
LI Y F, DING L P, WU J Z, et al. Survey on key issues in networks covert channel[J]. Journal of Software, 2019, 30(8): 2470-2490.

[2] BERNSTEIN D J, HENINGER N, LOU P, et al. Post-quantum RSA[C]//Proceedings of 2017 International Workshop on Post-Quantum Cryptography. Berlin: Springer, 2017: 311-329.

[3] BERNSTEIN D J, BREITNER J, GENKIN D, et al. Sliding right into disaster: left-to-right sliding windows leak[C]//Proceedings of International Conference on Cryptographic Hardware and Embedded Systems. Berlin: Springer, 2017: 555-576.

[4] PETITCOLAS F A P, ANDERSON R J, KUHN M G. Information hiding-a survey[J]. Proceedings of the IEEE, 1999, 87(7): 1062-1078.

[5] 刘敖迪, 杜学绘, 王娜, 等. 区块链技术及其在信息安全领域的研究进展[J]. 软件学报, 2018, 29(7): 2092-2115.
LIU A D, DU X H, WANG N, et al. Research progress of blockchain technology and its application in information security[J]. Journal of Software, 2018, 29(7): 2092-2115.

[6] PARTALA J. Provably secure covert communication on blockchain[J]. Cryptography, 2018, 2(3): 18.

[7] 宋上, 彭伟. BLOCCE+: 一种改进的基于区块链的隐蔽通信方法[J]. 重庆理工大学学报(自然科学), 2020, 34(9): 238-244.
SONG S, PENG W. BLOCCE +: an improved blockchain-based covert communication approach[J]. Journal of Chongqing University of Technology (Natural Science), 2020, 34(9): 238-244.

[8] ZHANG L J, ZHANG Z J, WANG W Z, et al. A covert communication method using special bitcoin addresses generated by vanitygen[J]. Computers, Materials & Continua, 2020, 65(1): 597-616.

[9] GUO Z Z, SHI L C, XU M Z, et al. MRCC: a practical covert channel over monero with provable security[J]. IEEE Access, 2021, 9: 31816-31825.

[10] 蓝怡琴, 张方国, 田海博. 利用门罗币实现隐蔽通信[J]. 西安电子科技大学学报, 2020, 47(5): 19-27.
LAN Y Q, ZHANG F G, TIAN H B. Using Monero to realize covert communication[J]. Journal of Xidian University, 2020, 47(5): 19-27.

[11] SHE W, HUO L J, TIAN Z, et al. A double steganography model combining blockchain and interplanetary file system[J]. Peer-to-Peer Networking and Applications, 2021, 14(5): 3029-3042.

[12] 佘维, 霍丽娟, 田钊, 等. 面向纯文本信息隐藏的区块链隐蔽通信模型[J]. 计算机科学, 2022, 49(1): 345-352.
SHE W, HUO L J, TIAN Z, et al. Blockchain covert communication model for plain text information hiding[J]. Computer Science, 2022, 49(1): 345-352.

[13] 佘维, 霍丽娟, 刘炜, 等. 一种可隐藏敏感文档和发送者身份的区块链隐蔽通信模型[J]. 电子学报, 2022, 50(4): 1002-1013.
SHE W, HUO L J, LIU W, et al. A blockchain-based covert communication model for hiding sensitive documents and sender identity[J]. Acta Electronica Sinica, 2022, 50(4): 1002-1013.

[14] 姜鹏坤, 张问银, 王九如, 等. 基于正常交易掩盖下的区块链隐蔽通信方案[J]. 网络与信息安全学报, 2022, 8(4): 77-86.
JIANG P K, ZHANG W Y, WANG J R, et al. Blockchain covert communication scheme based on the cover of normal transactions[J]. Chinese Journal of Network and Information Security, 2022, 8(4): 77-86.

[15] DAI W H, YU Y, DAI Y H, et al. Text steganography system using Markov chain source model and DES algorithm[J]. Journal of Software, 2010, 5(7): 785-792.

[16] RIVEST R L, SHAMIR A, TAUMAN Y. How to leak a secret[M]. Berlin: Springer, 2001.

[17] 马春光, 安婧, 毕伟, 等. 区块链中的智能合约[J]. 信息网络安全, 2018(11): 8-17.
MA C G, AN J, BI W, et al. Smart contract in blockchain[J]. Netinfo Security, 2018(11): 8-17.

[18] KUZLU M, PIPATTANASOMPORN M, GURSES L, et al. Performance analysis of a hyperledger fabric blockchain framework: throughput, latency and scalability[C]//Proceedings of IEEE International Conference on Blockchain (Blockchain). Piscataway: IEEE Press, 2019: 536-540.

[19] 袁勇, 王飞跃. 区块链技术发展现状与展望[J]. 自动化学报, 2016, 42(4): 481-494.
YUAN Y, WANG F Y. Blockchain: the state of the art and future trends[J]. Acta Automatica Sinica, 2016, 42(4): 481-494.

[20] 佘维, 荣欣鹏, 贾骏, 等. 区块链隐蔽通信的构建技术及检测方法研究综述[J]. 郑州大学学报(理学版), 2022, 54(6): 1-11.
SHE W, RONG X P, JIA J, et al. Technology development and research status of blockchain covert communication and detection methods[J]. Journal of Zhengzhou University (Natural Science Edition), 2022, 54(6): 1-11.

[21] OLEKH T, GOGUNSKII V. Use of discrete and continuous Markov chains for system absorbing states[C]//Proceedings of IEEE International Conference on Advanced Trends in Information Theory. Piscataway: IEEE Press, 2019: 518-521.

[22] LI F, XU S Y, SHEN H, et al. Extended dissipativity-based control for hidden Markov jump singularly perturbed systems subject to general probabilities[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(9): 5752-5761.

[23] 康慧娴, 易标, 吴汉舟. 文本隐写及隐写分析综述[J]. 应用科学学报, 2021, 39(6): 923-938.
KANG H X, YI B, WU H Z. Recent advances in text steganography and steganalysis[J]. Journal of Applied Sciences, 2021, 39(6): 923-938.

[24] ARSHAD R, SALEEM A, KHAN D. Performance comparison of huffman coding and double huffman coding[C]//Proceedings of Sixth International Conference on Innovative Computing Technology (INTECH). Piscataway: IEEE Press, 2016: 361-364.

[25] GO A, BHAYANI R, HUANG L. Twitter sentiment classification using distant supervision[R]. CS224N Project Report, 2009.

[26] HOUMANSADR A, BORISOV N. CoCo: coding-based covert timing channels for network flows[C]//Proceedings of the Information Hiding. Berlin: Springer, 2011: 314-328.

[作者简介]

佘维（1977− ），男，湖南常德人，博士，郑州大学教授、博士生导师，主要研究方向为区块链技术、信息安全、智能系统。

荣欣鹏（1998− ），男，山东东营人，郑州大学硕士生，主要研究方向为区块链技术、信息安全。

刘炜（1981− ），男，河南安阳人，博士，郑州大学副教授、博士生导师，主要研究方向为区块链技术、信息安全、智慧医疗。

田钊（1985− ），男，河南荥阳人，博士，郑州大学讲师、硕士生导师，主要研究方向为区块链技术、信息安全、智能交通。