doi:10.11959/j.issn.1000-436x.2016243

基于相似度的伪随机序列中超长稀疏特征分析

曹春杰[1,2]，孙敬张[1]，张智强[1]，王隆娟[1]，黄梦醒[1,2]
（1. 海南大学南海海洋资源利用国家重点实验室，海南 海口 570228；
2. 海南大学信息科学技术学院，海南 海口 570228）

摘 要：无线通信网络中的伪随机序列相似性分析是信息对抗领域中的研究热点。针对无线网络序列相似度分析中存在的序列超长、特征极其稀疏、无法在工程应用中实时处理等难点问题，提出了在一定可容忍误判概率下的序列相似度分析方法。首先对真随机序列相似度概率分布进行了理论分析；然后根据 NIST SP 800-22 评估标准对伪随机比特流进行了随机性分析，并对其随机性进行了有效性判定；最后结合实际无线通信网络中的超长伪随机序列进行了相似度分析和验证。结果表明：在误判概率约为 1%时，相似度下界为 0.62。上述结果对于协议分析、流量分析、入侵检测等网络安全领域有重要的借鉴意义和理论价值。

关键词：伪随机序列；相似度；稀疏特征；超长特征
中图分类号：TP391 **文献标识码**：A

Analysis of super-long and sparse feature in pseudo-random sequence based on similarity

CAO Chun-jie[1,2], SUN Jing-zhang[1], ZHANG Zhi-qiang[1], WANG Long-juan[1], HUANG Meng-xing[1,2]
(1. State Key Laboratory of Marine Resource Utilization in the South China Sea, Hainan University, Haikou 570228, China;
2. College of Information Science & Technology, Hainan University, Haikou 570228, China)

Abstract: Similarity analysis of pseudo-random sequence in wireless communication networks is a research hotspot problem in the domain of information warfare. Based on the difficulties in super-long sequence, extremely sparse feature, and futilities in engineering application for real-time processing exist in similarity analysis of sequence in wireless network, a method of similarity analysis of sequence in a certain margin of misacceptance probability was proposed. Firstly, the similarity probability distribution of real-random sequence was theoretically analyzed. Secondly, according to the standard of NIST SP 800-22, the randomness of pseudo-bitstream was analyzed and the validity of pseudo-bitstream was judged. Finally, similarity was analyzed and verified by combining super-long pseudo-random sequence in real wireless communication networks. The results indicate that the lower bound of similarity value is 0.62 when misacceptance probability uncertainty at about 1%. Above conclusion is considerable importance from the significance and theoretical values in network security domains, such as protocol analysis, traffic analysis, intrusion detection and others.

Key words: pseudo-random sequence, similarity, sparse feature, super-long feature

1 引言

在非合作信息对抗领域中，信号捕获解调之后的比特流是经过加扰也就是随机化处理[1,2]的序列，并且实际获取的数据序列中不只包含加扰数据，而且还有同步、控制等非加扰信息[3]。这就需要从获取的数据序列中剥离出其他信息，而仅仅保留加扰的数据后才能对未知信息做进一步

收稿日期：2016-09-15
基金项目：国家自然科学基金资助项目（No.61661019）；海南省重大科技计划基金资助项目（No.ZDKJ2016015）；海南省自然科学基金资助项目（No.20156217）
Foundation Items: The National Natural Science Foundation of China (No.61661019), The Major Science and Technology Project of Hainan Province (No.ZDKJ2016015), The Natural Science Foundation of Hainan Province (No.20156217)

处理。因此，从信息获取角度来看讲，该问题的解决也就是如何识别伪随机序列的特征，在信息截获、信息对抗和智能通信等领域[4]具有广泛的应用前景。

伪随机序列广泛应用在无线通信网络系统中[5~7]，其中，传输的数据流大部分都经过加密[8,9]，然后以各种帧或者流的形式进行传输。这些帧通过帧头来进行帧的同步，甚至是位同步。并且帧头可以在帧的起始，也可以隐含在帧内。因此，这些加密的伪随机数据流整体表现出较好的随机性，但局部存在某些结构特征。正是这些结构特征使其区别于真正的随机序列，并且使伪随机序列具有一定程度的相似性[10]。为了提高安全性，有些数据流的帧同步字段极其稀疏，给序列分析带来极大的挑战。同时伪随机序列数据的规模巨大，且在持续不断地增长，所以对于序列数据的实时处理和挖掘变得异常困难。因此，如何有效地找出伪随机序列中存在的结构特征是协议分析、流量分析[11]、入侵检测[12]等网络安全领域的重要问题。

由于伪随机序列来源复杂，有大量不可预测的噪声干扰存在，相似性分析问题变得比较复杂。之前大多数的相似性分析基于欧氏距离以及各种改进形式，这些方法简单、直接。但对随机序列数据属性比较敏感。因此前有提出了各种改进方法。Agrawal 等[13]将时间扭曲距离(DTW)的概念应用到分析中，改变了传统距离度量方式抗噪能力较差的缺点，但该方法的时间复杂度较高，接近于 $O(n^2)$，无法进行实时分析处理。Keogh 等[14,15]将序列进行符号化，提出了分段累积近似(PAA)、分段线性表示(PLR)和符号集合近似(SAX)等方法，并定义了相似度，该相似度是原欧氏距离相似度的下界，也就是最短边界距离。Perng 等[16]提出了界标模型(LM)，定义一些能反应序列变化特征的点，如极值点、拐点等重新描述原序列，提出了新的相似性分析方法。董晓莉等[17]通过分段线性近似方法，用斜率来描述原序列，提出了七元模式集合。但在实际应用中，随机序列变幻莫测，特征极其稀疏，且对实时性要求高，现有的方法无法满足以上要求。

综上所述，针对无线网络序列相似度分析中存在的序列超长，特征极其稀疏，无法在工程应用中实时处理等难点问题，本文提出了在一定可容忍误判概率下的序列相似度分析方法。本文从伪随机序列相似度入手，首先对真随机序列相似度概率分布进行了理论分析；然后根据 NIST SP 800-22 评估标准[18]对伪随机比特流进行了随机性分析，并对其随机性进行了有效性判定；最后结合实际无线通信网络中的超长伪随机序列进行相似度分析和验证。结果表明：在误判概率约为 1%时，相似度下界为 0.62。上述结果在实际序列分析中极大地缩减超长序列的留存时间，实现了实时性分析，且不受误码的影响。

2 相关概念

定义 1 随机序列[19]。随机序列是指序列尽可能近于均匀分布、各相继码元统计独立、完全不可预测。一个性能良好的随机序列应满足以下随机假设条件，简称为游程定理。

1) 平衡性。序列中 0 或 1 的个数相差至多为 1，也就是 0 与 1 出现的概率基本相等。

2) 游程特性。长为 1 的游程占游程总数的 $\frac{1}{2}$，长为 2 的游程占游程总数的 $\frac{1}{4}$，长为 i 的游程占游程总数的 $\frac{1}{2^i}$，且在等长的游程中，0 的游程个数和 1 的游程个数相等[10]。

3) 自相关性。异自相关函数为

$$R(\tau)=\frac{1}{T}\sum_{k=1}^{T}(-1)^{\alpha_k}(-1)^{\alpha_{k+\tau}},0<\tau\leqslant T-1 \tag{1}$$

定义 2 汉明距离[20]。描述 2 个长为 n 的码字 $x=(x_1x_2\cdots x_n)$，$y=(y_1y_2\cdots y_n)$ 之间的距离。

$$D(x,y)=\sum_{k=1}^{n}(x_k\oplus y_k) \tag{2}$$

其中，$D(x,y)$是 2 个码字在相同位置上不同码符号的数目的总和，它能够反映 2 个码字之间的差异。进而提供码字之间的相似程度的客观依据，$x_k\in\{0,1\}$，$y_k\in\{0,1\}$。

定义 3 序列相似度。设线性有序集合 $x=(x_1x_2\cdots x_n)$，$y=(y_1y_2\cdots y_n)$，$L=x\cap y=(l_1l_2\cdots l_k)$，其中，$1\leqslant k\leqslant\min(m,n)$，且 $l_1l_2\cdots l_k$ 排列次序同其在 x、y 中出现的次序。已知 $Len(L)=k$，称 $D(x,y)=\frac{\max(Len(L))}{n}=\frac{\max(k)}{n}$ 为序列 x 和 y 的相似度。因此，2 个序列的相似度可通过 2 个序列的汉明距离来计算。

3　随机序列相似度分布函数

从随机序列的定义可以看出，序列中出现 0 和 1 的概率相等，并且前后码元之间是统计独立的。那么，2 个随机序列相应位置比特是否相同的概率也相等，并且前一组比特的比对结果和后一组比特的比对结果是独立的。因此，2 个长为 n 随机序列的相似性可通过 n 重伯努利实验来分析。

若实验 E 只有 2 个可能结果：A、$\overline{A}$，则称 E 为伯努利实验。设

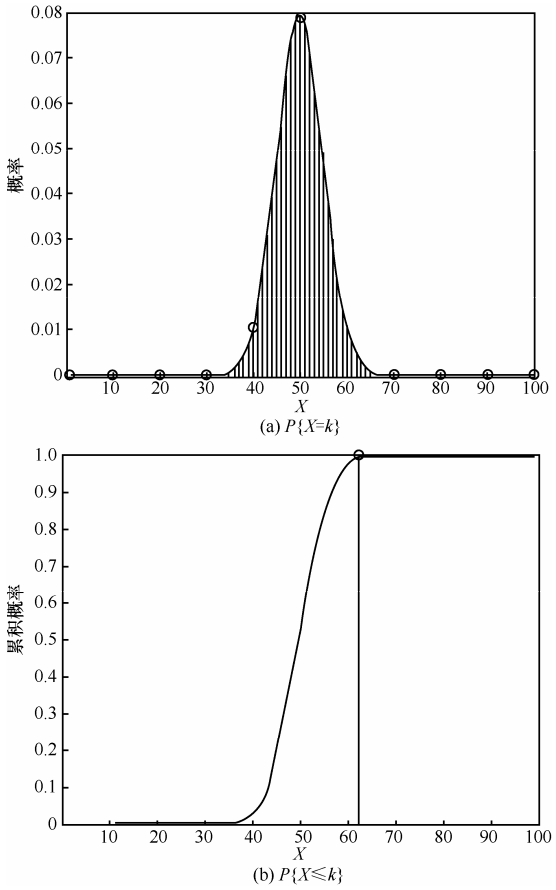
$$P(A)=p,\ 0<p<1，\ 则\ P(\overline{A})=1-p \tag{3}$$

将实验 E 独立重复地进行 n 次，则称这一串独立重复的实验为 n 重伯努利实验。

设 X 表示 n 重伯努利实验中事件 A 发生的次数，则 X 是一个随机变量。n 次实验中 A 发生 k 次的概率为

$$P\{X=k\}=\binom{n}{k}p^k q^{n-k}，\quad k=0,1,2,\cdots,n \tag{4}$$

称随机变量 X 服从参数为 n、p 的二项分布[21]，记为 $X\sim B(n,p)$。图 1(a)是在 n=100 的条件下，$P\{X=k\}$ 的概率分布，图 1(b)是 $P\{X\leqslant k\}$ 的概率分布。



(a) $P\{X=k\}$　(b) $P\{X\leqslant k\}$

图 1　随机序列相似度概率分布

由图 1 可知，在 n=100，k=60 的情况下，即当 2 个随机序列相似度 $D(x,y)$=0.6 时，$P\{X=k\}$=0.010 8，$P\{X\leqslant k\}$= 0.982 4。则 2 个随机序列相似度超过 0.6 的概率为 1−0.982 4=0.017 6。在 n=100，k=62 的情况下，即当 2 个随机序列相似度 $D(x,\ y)$=0.62 时，$P\{X=k\}$=0.004 5，$P\{X\leqslant k\}$= 0.994 0，2 个随机序列相似度超过 0.62 的概率为 1−0.994 0=0.006。因此，对于相似度超过 0.62 的 2 个随机序列，可判别为具有相同数据源，误判概率为 0.006。

4　伪随机序列性质分析

由于实际加密的网电空间数据流都不是真随机序列，而是采用加密算法产生的伪随机序列。本文通过对 AES 加密后的伪随机序列的随机性进行了分析，采用了美国国家标准技术研究所（NIST, National Institute of Standards and Technology）公布的测试方法，NIST 测试包[22]中包含以下 16 个指标组成的基本测试程序集，从不同角度检验被测序列在统计特性上相对于理想随机序列的偏离程度，是目前使用最为广泛的随机性测试指标的基本标准。接着对伪随机序列的性质进行分析。

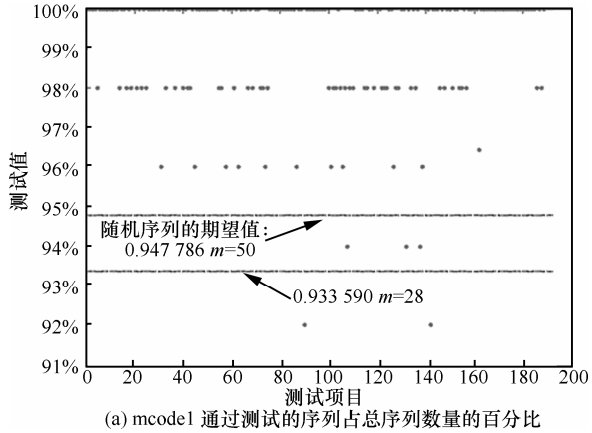
4.1　伪随机序列随机性分析

根据 NIST SP 800-22 标准对 AES128 加密后的序列进行了随机性分析，NIST 测试包中包含的 16 个指标：单比特频数测试、分块块内频数测试、游程测试、块内长游程测试、二进制矩阵秩测试、离散傅里叶变换测试、非重叠块匹配测试、重叠块匹配测试、Maurer 的通用统计测试、Lempel-Ziv 压缩测试、线性复杂度测试、串行检验测试、近似熵测试、累加和测试、随机游动测试、随机游动状态频数测试。

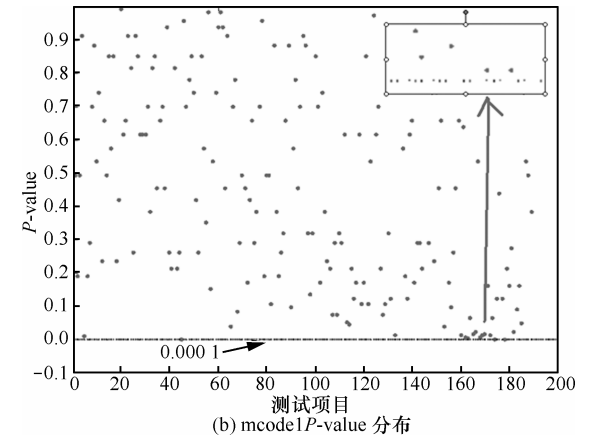
一般对序列进行随机性指标的检测，可以选择其中的几个常用指标进行测试。接着采用 2 种方法对测试结果进行评估：通过序列占总序列数量的比例评估方法和 P-value 的均匀分布评估方法，其中，P-value 是一个真随机序列比特检验序列随机性差的概率，结果如图 2 所示。

由图 2(a)可知，根据第一种评估方法，该序列通过了 0-1 频率、游程检验、块内最大游程、离散傅里叶变换、Maurer 通用统计、线性复杂度等所有 16 项检验。由图 2(b)可知，对于随机游动状态频率测试，由于逐比特累加和为 0 的次数大于 500 才能做有效分析，且对于 m=28，P>0.933 590 即认为是通过了检验。该序列的 50 个子序列中有 28 个有效

序列，且 P>0.933 590，认定其通过该项检验。根据第 2 种评估方法，该序列完全通过了检验，具有较好的随机性能。



(a) mcode1 通过测试的序列占总序列数量的百分比



(b) mcode1P-value 分布

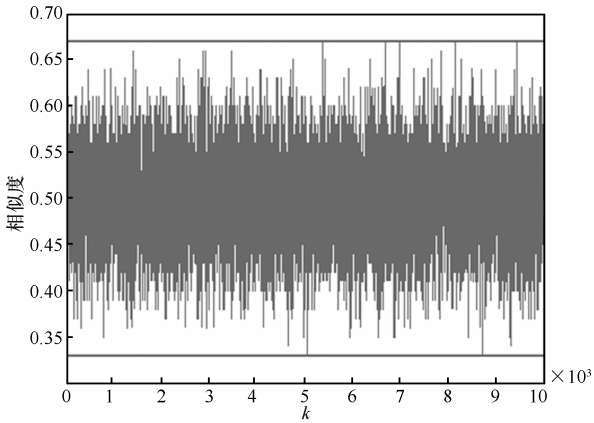
图 2　伪随机序列随机性分析

综上所述，AES128 加密后的伪随机序列可认定为随机序列。因此，伪随机序列可以采用随机序列的相似性分析方法进行研究，并且具有和随机序列相同的概率分布。

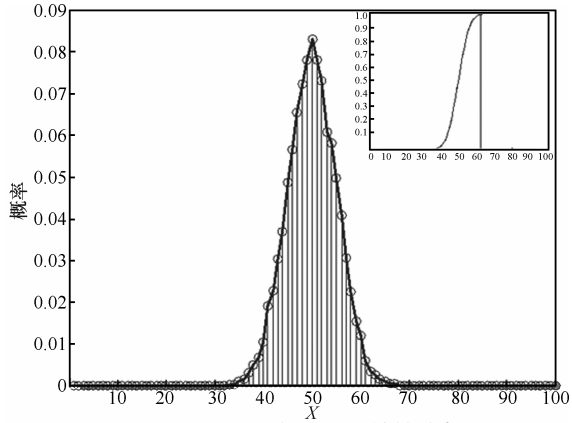
4.2　伪随机序列相似性分析

根据上述对真随机序列相似度概率分布的理论分析，通过 n 重伯努利实验对伪随机序列进行相似性分析。采用的参数：实验次数 M=10 000，序列长度 n=100，比特相等概率 $p=\frac{1}{2}$，明文由 Matlab 产生，密文由 AES128 产生。实验结果如图 3 所示，其中，图 3(a)为 M=10 000 时实验的相似度分布情况，图 3(b)为 $P\{X=k\}$的概率分布和 $P\{X\leqslant k\}$的概率分布。

由图 3 可知，对于伪随机序列，其相似度主要分布于[0.33 0.67]之内，其概率分布与随机序列的概率分布吻合。在 n=100，k=60 的情况下，即当 2 个伪随机序列相似度 $D(x, y)$=0.6 时，$P\{X=k\}$=0.011 9，$P\{X\leqslant k\}$=0.984 7。则 2 个伪随机序列相似度超过 0.6 的概率为 1−0.984 7=0.015 3。在 n=100，k=62 的情况下，即当 2 个伪随机序列相似度 $D(x, y)$=0.62 时，$P\{X=k\}$=0.003 4，$P\{X\leqslant k\}$= 0.994 1，2 个伪随机序列相似度超过 0.62 的概率为 1−0.994 1=0.005 9。因此，伪随机序列相似度分布也服从参数为 n、p 的二项分布，并且，对于相似度超过 0.62 的 2 个伪随机序列，可判别为具有相似的协议结构，误判概率为 0.005 9。即当 2 个伪随机序列的相似度大于或等于 0.62 时，可判定为同一特征序列，正确率为 1−0.005 9=0.994 1（如图 3(b)右上角小图所示）。



(a) M=10 000时验的相似度分布



(b) $P\{X=k\}$和$P\{X\leqslant k\}$的概率分布

图 3　伪随机序列相似度概率分布

5　结束语

针对无线网络序列相似度分析中存在的序列超长，特征极其稀疏，无法在工程应用中实时处理等难点问题，提出了在一定可容忍误判概率下的序列相似度分析方法。结果表明：在误判概率约为 1% 时，相似度下界为 0.62。上述结果对于协议分析、流量分析、入侵检测等网络安全领域有着重要的借鉴意义和理论价值。

参考文献：

[1] WANG F, HUANG Z T, ZHOU Y. A new method for m-sequence and gold-sequence generator polynomial estimation[C]//International Symposium on Microwave, Antenna, Propagation and Emc Technologies for Wireless Communications. 2007: 1039-1044.

[2] 杨忠立, 刘玉君. 自同步扰乱序列的综合算法研究[J]. 信息技术, 2005, (2):30-32.
YANG Z L, LIU Y J. Algorithm research of self-synchronizing scrambler sequence[J]. Information Technology, 2005, (2): 20-32.

[3] 廖红舒, 袁叶, 甘露. 自同步扰码的盲识别方法[J]. 通信学报, 2013(1):136-143.
LIAO H S, YUAN Y, GAN L. Novel blinal recognition method for self-synchronized scrambler[J]. Journal on Communications, 2013(1): 136-143.

[4] 杨学军, 苏金树. 关于我国网络电磁空间安全战略的思考[J]. 国防科技, 2010, 31(4): 1-3.
YANG X J, SU J S. Security strategy of network and electromagnetism in China: perspectives and suggestions[J]. National Defense Science and Technology, 2010, 31(4): 1-3.

[5] DONG G, JIAN P. Sequence data mining[J]. Advances in Database Systems, 2007, 33(2):800

[6] 肖国震. 伪随机序列及其应用[M]. 北京: 国防工业出版社, 1985.
XIAO G Z. Pseudo random sequence and its applications[M]. Beijing: National Defence Industry Press, 1985.

[7] SCHNEIER B. Secrets and lies: digital security in a networked world[J]. Info, 2003, 5(1): 163-165.

[8] 翁贻方, 郑德玲, 王云雄. 基于混沌序列密码的网络信息加密系统[J]. 微计算机信息, 2007, 23(30): 94-95.
WENG Y F, ZHENG D L, WANG Y X. Research and implementation of the chaotic stream cipher-based network information cryptography system[J]. Microcomputer Information, 2007, 23(30): 94-95.

[9] SPRING M, WESTERMEYER J, HALCON L, et al. Technique for processing encoded information in a wireless communication nerwork[J]. Journal of Nervous & Mental Disease, 2015, 191(12): 813-819.

[10] MCELIECE R J. The theory of information and coding[J]. Mathematical Gazette, 2002.

[11] 张进, 黄清杉, 赵文栋,等. 面向骨干网流量分析与管理的计数器结构[J]. 软件学报, 2013, 24(9): 2165-2181.
ZHANG J, HUANG Q S, ZHAO W D, et al. Statistics counter architecture for backbone network traffic analysis and management[J]. Journal of Software, 2013, 24(9): 2165-2181.

[12] 蒋建春, 马恒太. 网络安全入侵检测:研究综述[J]. 软件学报, 2000, 11(11): 1460-1466.
JIANG J C, MA H T. A survey of intrusion detection research on network security[J]. Journal of Software, 2000, 11(11): 1460-1466.

[13] BERNDT D J, CLIFFORD J. Using dynamic time warping to find patterns in time series[C]// Working Notes of the Knowledge Discovery in Databases Workshop. 1994:359-370.

[14] KEOGH E, LIN J, FU A. HOT SAX: efficiently finding the most unusual time series subsequence[C]//5th IEEE International Conference on Data Mining. 2005.

[15] KEOGH E, CHAKRABARTI K, PAZZANI M, et al. Dimensionality reduction for fast similarity search in large time series databases[J]. Knowledge & Information Systems, 2001, 3(3): 263-286.

[16] PERNG C S, WANG H, ZHANG S R, et al. Landmarks: a new model for similarity-based pattern querying in time series databases[J]. Icde, 2000, 93(7): 33-42.

[17] 董晓莉, 顾成奎, 王正欧. 基于形态的时间序列相似性度量研究[J]. 电子与信息学报, 2007, 29(5): 1228-1231.
DONG X L, GU C K, WANG Z O. Research on shape-based time series similarity measure[J]. Journal of Electoral of Electronics Information Technology, 2007, 29(5):1228-1231.

[18] RUKHIN A, SOTO J, NECHVATAL J. A statistical test suite for random and pseudorandom number generators for cryptographic applications[S]. NIST Special Publication 800-22, 2013.

[19] SCHNEIER B P. Applied cryptography: protocols, algorithms, and source code in C[J]. Government Information Quarterly, 1994, 13(3): 336.

[20] 王新梅, 肖国镇. 纠错码原理与方法[M]. 西安:西安电子科技大学出版社, 1996.
WANG X M, XIAO G Z. Theory and method of error-correcting codes[M]. Xi'an: Xidian University Press, 1996.

[21] 盛骤. 概率论与数理统计(第四版) [M]. 北京: 高等教育出版社, 2008.
SHENG Z. Probability and mathematical statistics (4thed)[M]. Beijing: Higher Education Press, 2008.

[22] RUKHIN A, SOTO J, NECHVATAL J, et al. A statistical test suite for random and pseudorandom number generators for cryptographic applications[J]. Research Gate, 2010.

作者简介：

曹春杰（1977-），男，河北衡水人，海南大学信息科学技术学院教授、硕士生导师，主要研究方向为无线网络安全、信息对抗。

孙敬张（1993-），男，海南海口人，海南大学硕士生，主要研究方向为数据库安全、搜索加密。

张智强（1992-），男，安徽铜陵人，海南大学硕士生，主要研究方向为视频监控系统漏洞分析。

王隆娟（1976-），女，海南东方人，海南大学讲师，主要研究方向为信息安全、计算机应用。

黄梦醒（1973-），男，河南信阳人，海南大学教授、博士生导师，主要研究方向为数据与知识工程、个性化服务、电子商务与电子政务、云计算、智能信息处理。