一种基于条件随机场的击键特征身份鉴别方法*

李 晨, 张功萱, 岳宝玲, 贺定龙
(南京理工大学 计算机科学与工程学院, 南京 210094)

摘 要: 为了提高网络系统中常用的用户名/密码身份鉴别系统的安全性,提出基于用户击键特征的二次身份鉴别方案。该方案在现有的用户名/密码认证方案上加入基于条件随机场的击键特征建模和识别步骤,并在理论上对该方案作了可行性、安全性、效率和准确率分析。通过使用公开数据的实验发现,该方案在提高原方案安全性的同时,具有效率高、准确率高的特点。通过与其他基于相同公开数据的实验结果对比发现,条件随机场模型在击键特征识别领域具有很好的识别效果。

关键词: 击键特征; 身份鉴别; 条件随机场; 行为特征
中图分类号: TP393.08 **文献标志码**: A **文章编号**: 1001-3695(2014)07-2112-04
doi:10.3969/j.issn.1001-3695.2014.07.046

Keystroke dynamics based user authentication using conditional random fields

LI Chen, ZHANG Gong-xuan, YUE Bao-ling, HE Ding-long
(*School of Computer Science & Engineering*, *Nanjing University of Science & Technology*, *Nanjing* 210094, *China*)

Abstract: Based on the widely used username/password authentication system, this paper proposed a two-factor authentication system using keystroke dynamics. The scheme embedded keystroke dynamics modeling and identification into the username/password authentication system. Its feasibility, security, efficiency and accuracy were theoretically analyzed. Experiments using publicly available keystroke dynamics datasets show that it has great efficiecy and accuracy with improved security. Performance evaluation with different models using the same public datasets indicates that conditional random field is an effective model for keystroke dynamics.

Key words: keystroke dynamics; identification; conditional random fields; behavior characteristics

传统的身份认证方案是基于用户名/密码实现的,这种方案存在严重的安全缺陷,但它依然是当前最常用的身份认证方案。研究表明,每个人都有独一无二的击键特征,通过分析用户的击键特征就可以鉴别其身份[1]。击键特征的采集是非侵入式的,可以很方便地在已有的用户名/密码身份认证系统中加入基于击键特征分析的二次认证模块,以此提高系统的安全性。本文首次将条件随机场模型应用到击键特征分析,并提出基于击键特征识别的二次身份认证方案。该方案可以在不影响用户使用的情况下,增强现有密码认证方案的安全性。

绝大多数人在用键盘输入时使用双手,因此任一击键与其相邻的击键有物理联系。由于个人使用习惯不同,敲击键盘的力度和熟练程度都因人而异。这些击键特征类似于人的笔迹,本文通过分析这些击键特征,实现对用户身份的鉴别[2]。

1 条件随机场模型

条件随机场(conditional random fields, CRFs)[3]由 Lafferty 等人在 2001 年提出,该模型是在给定需要标记的观察序列的条件下,计算整个标记序列的联合概率。条件随机场是一种无向图模型,在给定的观测序列之上定义了一个对数分布的标记序列。

条件随机场的定义:令 $G=(V,E)$ 表示一个无向图,任取 $y_v \in Y$,存在 $v \in V$ 与之一一对应;当在条件 X 下,随机变量 y_v 服从图 G 的马尔可夫属性,称 (Y,X) 是一个条件随机场。

理论上,只要图 G 能够满足标记序列间的条件独立性,图 G 的结构可以是任意的。在对序列建模时,最常用并且最重要的是一阶链式模型,如图 1 所示。这个链式结构在标记序列上定义了一个联合概率分布 $p(y|x)$。

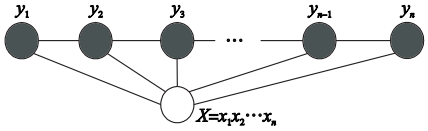

图1 一阶条件随机场模型

在给定观测序列 X 的条件下,条件随机场概率模型的数学公式定义为

$$\begin{cases} p(y|x,\lambda)=\dfrac{1}{Z(x)}\exp\left(\sum_j \lambda_j F_j(y,x)\right) \\ Z(x)=\sum_y \exp\left(\sum_k \lambda_k F_k(y,x)\right) \\ F_j(y,x)=\sum_{i=1}^{n} f_j(y_{i-1},y_i,x,i) \end{cases}$$

其中:$Z(x)$ 是归一化因子;λ_j 是需要从训练数据中估计的参数;$f_j(y_{i-1},y_i,x,i)$ 是一个状态函数 $s(y_{j-1},y_j,x,i)$ 或者转移函数 $t(y_{j-1},y_j,x,i)$。

2 基于条件随机场的击键特征识别方法

2.1 特征提取

击键特征的特征向量由于没有统一的命名,各个学者可能使用不同的名字指代相同的特征[4],本文采用如下命名标准。

收稿日期: 2013-08-22; **修回日期**: 2013-09-25 **基金项目**: 国家自然科学基金资助项目(61272420)

作者简介:李晨(1989-),男,江苏江阴人,硕士,主要研究方向为计算机网络(cpylua@163.com);张功萱(1961-),男,教授,博士,主要研究方向为网络与分布式计算;岳宝玲(1989-),女,硕士,主要研究方向为可信计算;贺定龙(1989-),男,硕士,主要研究方向为可信计算.

原始特征通常包括(P 表示按下键,R 表示释放键):a)保持时间(key hold),$KH_i = R_i - P_i$;b)间隔时间(key interval),$KI_{ij} = P_j - R_i$;c)按键时间(key press),$KP_{ij} = P_j - P_i$;d)扫描码。其他的特征一般而言都是从这些原始特征计算而来,叫做扩展特征。常见的扩展特征有 n 元组和欧氏距离等。两类特征间没有绝对的好坏之分,依据算法模型的不同,选择合适的特征向量可以提高模型的识别能力[5]。本文仅使用原始特征来构成特征向量。

条件随机场模型中的观察序列 X_i 是一系列随机变量,给每次击键建立一个特征向量,将连续击键对应的特征向量作为一次完整的观察序列。每个击键特征是一个键值对 key = value。一个特征向量包括 12 个击键特征,H、DD 和 HD 分别表示当前键的保持时间、按键时间和间隔时间;H[-1]、DD[-1]和 UD[-1]以及 H[1]、DD[1]和 UD[1]分别表示前一个键和后一个键的对应时间;W[-1]、W[0]和 W[1]分别表示前一个键、当前键和后一个键的扫描码。需要说明的是,第一个键没有 H[-1]、DD[-1]、UD[-1]和 W[-1],而最后一个键没有 H[1]、DD[1]、UD[1]和 W[1]。

标记序列 y_i 是一组 positive 和 negative 值,前者用来标记真实用户的特征向量,后者用来标记入侵者的特征向量。由于条件随机场模型的训练是有监督的,用于模型训练的数据需要预先做好标记,识别时根据训练好的模型预测观察序列对应的标记序列。

2.2　模型的训练和识别

本文使用一阶条件随机场建立用户模型,该模型不对数据作任何假设,可以较好地处理数据的随机特征。一阶条件随机场模型形式简单,训练速度较快,能满足识别的要求。条件随机场的参数 λ 可以用最大熵模型的参数估计方法,即对概率的对数最大似然函数求最大值。训练过程中的观察序列是击键特征向量序列,标记序列是一系列 positive 和 negative 标记。给定训练集 $D = \{(x^{(k)}, y^{(k)})\}$,根据最大熵模型对参数 $\lambda = \lambda_1, \lambda_2, \cdots, \lambda_n$ 估计采用最大似然估计法。条件概率 $p(y|x,\lambda)$ 的对数似然函数形式为

$$L(\lambda) = \sum_k [\log \frac{1}{Z(x^{(k)})} + \sum_j \lambda_j F_j(y^{(k)}, x^{(k)})]$$

对 λ_j 求对数似然函数的偏导,可以得到

$$\frac{\partial L(\lambda)}{\partial \lambda_j} = E_{\tilde{p}(Y,X)}[F_j(Y,X)] - \sum_k E_{p(Y|x^{(k)},\lambda)}[F_j(Y,x^{(k)})]$$

其中:$\tilde{p}(Y,X)$ 是训练数据的经验分布;$E_p[\cdots]$ 是分布 p 的数学期望。训练过程就是在训练数据的条件约束下,求上式等于零时的参数。

条件随机场模型中的状态函数和转移函数,即 $f_j(y_{i-1}, y_i, x, i)$ 可以根据需要自由设置,通常这些函数满足如下形式(其中 I 是恒等函数,函数 g 是模型相关的任意函数):

$$f = I(y_i = i) \times g;\ f = I(y_{i-1} = i) \times I(y_i = i) \times g$$

完成模型训练后,将得到的模型参数作为用户信息存储起来。在用户鉴别阶段,给定一个特征向量序列,要找到观察序列在用户模型下最可能的标记序列。这个问题可以通过 Forward-Backward 算法和 Viterbi 算法解决。如果产生标记序列的概率 $P > \theta$,就接受模型的预测;否则拒绝模型的预测。θ 是一个可调节的阈值。

3　基于击键特征识别的二次身份鉴别方案

击键事件在本质上是随机事件,连续的击键事件相互存在着一定的联系。条件随机场模型可以很好地描述这类问题。将密码认证系统和击键特征识别相结合,如图 2、3 所示,虚线框标出的步骤是新方案加入的。与传统密码系统一样,整体分为两大部分,即用户注册和用户认证。在注册中都加入了击键特征模型训练的步骤,训练得到的模型参数作为用户信息的一部分存储起来。认证用户时,在传统密码验证通过后,增加了用户击键特征识别的步骤,只有击键特征匹配的请求才能通过认证。

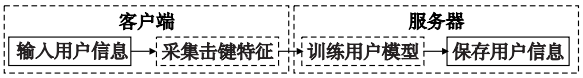

图2　用户注册流程

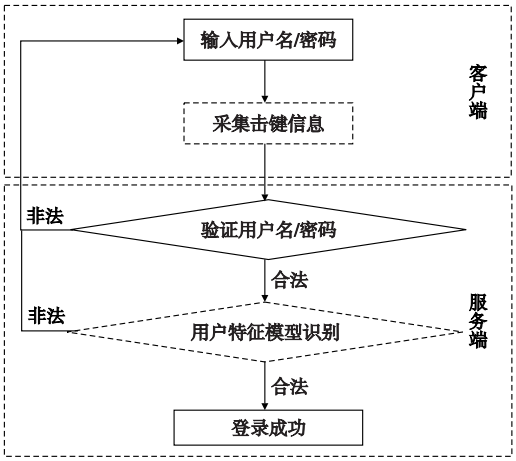

图3　用户认证流程

当前的用户名/密码身份认证系统为了抵御机器攻击,通常采用验证码作为辅助手段增强系统安全性,随着计算机性能的提升,这种基于图形的验证码越来越容易被机器通过算法识别。本文提出的方案是对现有用户名/密码身份认证系统的一个改进,击键特征作为一种人体生物特征,理论上不可能被机器模拟,安全性要高于基于验证码的方案。新方案在注册阶段需要一定的计算资源训练用户模型,但之后认证过程中的额外开销非常小,用户察觉不到,从而保证了完好的用户体验[6]。

4　方案分析

本文提出的基于条件随机场模型的击键特征二次身份认证方案是对传统用户名/密码方案的一个扩展下面从四个角度对新方案进行分析。

4.1　可行性分析

新方案是对原有用户名/密码方案的一个非侵入式扩展,仅在原方案的流程中加入了特征采集和模型训练识别的步骤。新方案的优势是可以在现有用户名/密码系统上修改实现,不对系统其他部分产生可见影响。另外,击键特征作为一种生物识别信息已经有许多学者证明是可行的。因此,新方案具有非常好的可行性。

4.2　安全性分析

身份认证方案的安全性是整个方案的核心,业界为了提高用户名/密码方案的安全性,往往采用加密信道传输信息,如 SSL/TLS。这种方式可以解决用户名/密码方案存在的很大一部分不足,如中间人攻击、网络嗅探、重放攻击等。对于这类网络攻击,用户名/密码方案配合加密信道具有比较好的抗攻击性。由于新方案是原始方案的一个扩展,所以新方案继承了原方案的这类抗攻击性。

如果攻击者通过某种途径获得了用户的登录凭证,那么原方案将无法区分攻击者和合法用户。这是现有互联网体系下用户名/密码方案面临的最严重的安全问题,而新方案则可以提高这种情况下认证系统的安全性。下面的分析假设系统使用了SSL/TLS之类的加密信道,这在当今的互联网领域是非常合理的假设。

当攻击者获得用户登录凭证并尝试登录系统时,在传统方案用户名/密码认证通过后,新方案增加了对用户击键特征的识别,只有当三者全部匹配,此次认证才算通过。击键特征作为生物信息的一种,有着与实体身份不可分离的联系,因此,攻击者的击键特征理论上无法匹配真实用户的击键特征。假设原方案的危险系数是 k,击键特征识别算法的失误率是 p,那么新方案的危险系数可以表示为 $k' = kp$,由于 $0 < p < 1$,所以 $k' < k$,这就证明了当攻击者已经获取用户登录凭证后,新方案的安全性要好于原方案。

4.3 效率分析

新方案的扩展主要在两个方面,即客户端的击键特征采集、服务器端的模型训练和特征识别。下面从这两个方面分析新方案的效率。

1)客户端击键特征采集 用户的登录凭证一般而言不超过30个字符,而原始击键特征主要是按键的时间特征和扫描码,这些信息都可以在用户输入登录凭证时通过软件实时获得。所以在客户端采集击键特征的开销可以忽略不计。

2)服务器端的训练和识别 用户模型的训练效率与训练算法有关,条件随机场模型常用的训练算法包括L-BFGS和随机梯度下降。两种算法每次迭代的时间复杂度都是 $O(n)$,其中 n 是训练集的大小。对于一般训练集,迭代100次左右后就能收敛到一个较好的值。新方案中训练集大小是可调的,可以根据识别率的需要增大或减少训练集,通常训练集大小在500左右。通过以上分析,同时考虑到模型训练过程中涉及较多复杂的数学运算,单个用户模型训练所消耗的时间应该在几秒钟以内。用户模型是一次训练多次使用,建立模型后的识别过程通常使用Forward-Backward算法或者Viterbi算法,两者的时间复杂度都是 $O(N^2T)$,其中 N 是状态空间的大小,T 是观察序列大小。新方案中 N 和 T 都很小,所以一次识别的耗时在毫秒级别,非常高效。

4.4 准确率分析

条件随机场模型在结构化序列标记领域有着非常广泛的应用,该模型非常适合击键特征这种具有上下文关系的序列,通常条件随机场模型对于这类问题有较高的准确率。另外,一个模型的假报警率(false alarm rate)和失误率(miss rate)是可调节的,通过调校模型参数可以取得满意的准确率和较好的用户体验。

5 实验及结果

5.1 实验方法

在击键特征分析的研究中,研究者们提出了各自的鉴别方案,并独立进行了实验。由于实验环境和实验方式的不同,实验结果往往不具可比性[7]。CMU的Killourhy等人[8]在2009年意识到了这个问题,他们着手收集整理了一批击键特征数据,并将其公布在互联网上,方便该领域的其他研究者使用。

在实验中,本文使用了Killourhy和Maxion收集整理的公共数据库。数据库中包含了51位参与者每人400次的采样。参与者的年龄分布覆盖了各个年龄段,每位参与者的密码都是相同的:.tie5Roanl,这个密码具有相当的典型性。每次采样11次按键,包括10个密码字面值和最后的回车键。每次采样从这11次按键中提取了31个特征,其中包括11个保持时间、10个间隔时间以及10个按键时间。Killourhy的数据库中每行除了包含31个击键特征外,还包含了参与者编号、会话编号和输入编号,这三个信息对于训练模型没有用处,因此将它们删除了。由于原始数据中的最后一个回车键只包含保持时间,在本文的特征向量中没有包含回车键。

本文用Python脚本从数据库中为每个参与者生成一份训练数据、一份真实用户测试数据和一份入侵者测试数据,总共有51个参与者。训练数据中包含有对应参与者的300个经过标记的观察序列,这300个观察序列是从该用户所有的观察序列中随机抽样得到的;另外还包含250个经过标记的入侵者观察序列,先从剩余的50位参与者中随机抽样25位作为入侵者,从这25位入侵者的所有观察序列中每人随机抽样10个观察序列。真实用户测试数据中包含的是对应参与者所有观察序列和训练数据中300个观察序列的差集,共100个。入侵者测试数据是从剩余25位参与者中每人随机抽样10个观察序列得到的,总共100个,与真实用户测试数据个数相同。

本文对每个参与者进行模型训练和测试,测试结果包含对每个测试观察序列的预测标记、真实标记以及产生预测标记的概率。计算所有用户在不同阈值下的假报警率和失误率的平均值,最后求得模型的等错误率(equal error rate,EER)。

5.2 结果分析

在实验结果中本文没有直接使用假报警率和失误率作为评价性能的依据,而是使用等错误率。一般而言,假报警率和失误率间有一个类似反比的关系,其中一个的减小往往意味着另一个的增大,无法直观地分析识别性能。等错误率是指假报警率和失误率相等时的数值,它是描述识别错误率最好的单一数值,等错误率越低意味着识别算法的错误率越低。

实验中本文测试了五种训练算法,五种算法的等错误率如表1所示,按等错误率从小到大排序。通过分析表中数据发现,梯度下降算法是性能最好的,lbfgs和l2sgd算法的等错误率非常接近,都在0.110以下。被动主动算法和平均感知算法的等错误率和梯度下降算法相比要差3~4个百分点。自适应权重向量正规化算法是表现最差的训练算法,等错误率几乎达到了最好的lbfgs算法的两倍。本文得出的结论是训练算法在条件随机场模型的识别性能上有很大的影响,其中梯度下降一类算法是最好的训练算法。

表1 五种条件随机场训练算法的等错误率

序号	模型/训练算法	等错误率(EER)
1	L-BFGS梯度下降(lbfgs)	0.108
2	随机梯度下降(l2sgd)	0.110
3	被动主动(pa)	0.132
4	平均感知(ap)	0.141
5	自适应权重向量正规化(arow)	0.212

在Killourhy的实验结果中,性能最好的是缩放曼哈顿距离模型,它的等错误率是0.096。表现最好的前四个模型等错误率相差很小,可以认为,性能最好的模型等错误率应该在0.100上下。本文使用条件随机场模型得到的最好等错误率

是 0.108(lbfgs),和 Killourhy 的最好结果 0.096 仅相差 1%。由此可以证明,条件随机场在击键特征分析中是一个可行的模型,具有较好的准确率。

本文分析了实验中训练和识别操作所用时间,测试环境是 2 GHz 的四核 Intel Xeon CPU。使用训练好的模型对用户进行识别,平均可以达到约 9214 次/s 的速度,如图 4 所示。用户模型只需要训练一次,在实验中使用性能最好的 lbfgs 算法训练一个用户模型大约需 1 s,这个一次性的额外开销是完全可以接受的。因此,基于击键特征分析的二次身份鉴别方案具有非常好的性能,通过消耗很少的额外服务器资源将生物信息识别整合到身份鉴别过程中,极大地提高了身份鉴别系统的安全性。

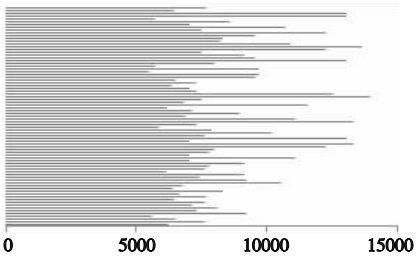

图4　模型的识别速度/次/s

值得指出的是,Killourhy 在实验的训练部分没有加入入侵者数据,本文为了建立正确的条件随机场模型,在训练数据中加入了入侵者数据。Uzun 等人[9] 使用 Killourhy 的公开数据,针对神经网络模型进行过独立的实验。他们在训练过程中加入了入侵者数据,得到的最好 EER 达到了 0.080 7,这一结果比 Killourhy 等人的最好结果 0.096 要好。但是在 Uzun 的实验中,测试和训练过程使用了同一入侵者的数据,在实际应用中入侵者数据是不可能预先收集的。因此,他们的实验结果有待进一步确认。本文虽然在训练数据中也加入了入侵者数据,但是测试和训练时使用的入侵者数据来自完全不相交的参与者集合,所以不存在 Uzun 等人的结果偏差问题。

6　结束语

基于公开数据的可重现实验结果对于击键特征分析领域是非常重要的,Killourhy 等人发布的公开数据和评估结果是对该领域的重要贡献。条件随机场的训练过程中训练算法的选择对识别性能影响非常大,不当的训练算法会导致模型识别性能急剧下降。实验结果证明,条件随机场模型在击键特征分析领域是可行和有效的。结合本文提出的二次身份鉴别方案,可以将击键特征分析整合到现有的密码认证系统中,提高系统的安全性。实验数据表明,新方案额外引入的训练阶段和识别阶段的开销都很小。

基于击键特征的身份鉴别系统的准确率远未到达可以独立应用的要求,目前来看,只能作为一种辅助身份鉴别系统和其他系统一起使用,这类身份鉴别系统还有很大的进步空间。

参考文献:

[1] CRAWFORD H. Keystroke dynamics: characteristics and opportunities[C]//Proc of the 8th Annual International Conference on Privacy Security and Trust. [S.l.]: IEEE Press, 2010: 205-212.

[2] JENKINS J, NGUYEN Q, REYNOLDS J, *et al*. The physiology of keystroke dynamics[C]//Proc of SPIE: Independent Component Analyses, Wavelets, Neural Networks, Biosystems, and Nanoengineering IX. 2011.

[3] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[4] BALAGANI K, PHOHA V, RAY A, *et al*. On the discriminability of keystroke feature vectors used in fixed text keystroke authentication [J]. Pattern Recognition Letters, 2011, 32(7): 1070-1080.

[5] SHANMUGAPRIYA D, PADMAVATHI G. An efficient feature selection technique for user authentication using keystroke dynamics[J]. International Journal of Computer Science and Network Security, 2011, 11(10): 191-195.

[6] GIROUX S, WACHOWIAK R, WACHOWIAK M. Keystroke-based authentication by key press intervals as a complementary behavioral biometric[C]//Proc of IEEE International Conference on Systems, Man and Cybernetics. [S.l.]: IEEE Press, 2009: 80-85.

[7] BHATT S, SANTHANAM T. Keystroke dynamics for biometric authentication: a survey[C]//Proc of International Conference on Pattern Recognition, Informatics and Medical Engineering. [S.l.]: IEEE Press, 2013: 17-23.

[8] KILLOURHY K, MAXION R. Comparing anomaly-detection algorithms for keystroke dynamics[C]//Proc of the 39th Annual International Conference on Dependable Systems and Networks. [S.l.]: IEEE Press, 2009: 125-134.

[9] UZUN Y, BICAKCI A. A second look at the performance of neural networks for keystroke dynamics using a publicly available dataset [J]. Computers & Security, 2012, 31(5): 717-726.

(上接第 2111 页)

[6] LIAO Y, WANG S. A secure dynamic ID based remote user authentication scheme for multi-server environment [J]. Computer Standards & Interfaces, 2009, 31(1): 24-29.

[7] HSIANG H, SHIH W. Improvement of the secure dynamic ID based remote user authentication scheme for multi-server environment [J]. Computer Standards & Interfaces, 2009, 31(6): 1118-1123.

[8] SOOD S K, SARJE A K, SINGH K. A secure dynamic identity based authentication protocol for multi-server architecture[J]. Journal of Network and Computer Applications, 2011, 34(2): 609-618.

[9] LEE C, LIN T, CHANG R. A secure dynamic ID based remote user authentication scheme for multi-server environment using smart cards [J]. Expert Systems with Applications, 2011, 38(11): 13863-13870.

[10] 杨力,马建峰. 可信的智能卡口令双向认证方案[J]. 电子科技大学学报, 2011, 40(1): 128-133.

[11] 王亚飞. 一种基于智能卡口令认证方案的研究[J]. 计算机应用与软件, 2011, 28(9): 295-297.

[12] 杜世琼,王世卿. 基于属性的 Chameleon 远程认证[J]. 计算机工程与设计, 2012, 33(11): 4081-4085.

[13] LIAO Y P, HSIAO C M. A novel multi-server remote user authentication scheme using self-certified public keys for mobile clients [J]. Future Generation Computer Systems, 2013, 29(3): 886-900.

[14] TSENG Y M, WU T Y, WU J D. A pairing-based user authentication scheme for wireless clients with smart card[J]. Informatics, 2008, 19(2): 285-302.

[15] KOBLITZ N, MENEZES A, VANSTONE S. The state of elliptic curve cryptography[J]. Design, Codes and Cryptography, 2000, 19(2-3): 173-193.