计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20210456
59(12):2708-2722,2022

参数化混合口令猜测方法

韩伟力[1] 张俊杰[1] 徐 铭[1] 王传旺[1] 张浩东[1] 何震瀛[1] 陈 虎[2]
[1](复旦大学数据分析与安全实验室 上海 200438)
[2](华南理工大学软件学院 广州 510006)
(wlhan@fudan.edu.cn)

Parameterized Hybrid Password Guessing Method

Han Weili[1], Zhang Junjie[1], Xu Ming[1], Wang Chuanwang[1], Zhang Haodong[1], He Zhenying[1], and Chen Hu[2]
[1](*Laboratory of Data Analytics and Security*, *Fudan University*, *Shanghai* 200438)
[2](*School of Software Engineering*, *South China University of Technology*, *Guangzhou* 510006)

Abstract The textual password based authentication method is still the mainstream for users to authenticate their identities. To better study password security, researchers propose many data-driven password guessing methods, such as probabilistic context-free grammars (PCFG) and Markov model-based methods. These methods have unique advantages in guessing passwords, i.e., they can guess specific types of passwords with a smaller number of guesses. To make full use of these advantages for better guessing efficiency, we propose a general practical framework of parameterized hybrid guessing. The framework consists of a model pruning method and an allocation strategy for guesses which is theoretically proved optimal. It can mix the guess advantages of different data-driven methods to generate more efficient guessing sets. To verify the generality and efficiency of the framework, we analyze and mix the different advantages of existing data-driven guessing methods, then design multiple parameterized hybrid guessing methods (collectively referred to as hyPassGu) composed of multiple models based on the framework for guessing practice. We also evaluate these hybrid guessing methods using four large-scale password datasets (more than 150 million passwords in total) leaked from real websites. The experimental results show that hyPassGu constructed by different methods surpass the guessing efficiency of the single method, and surpass the best efficiency of the single method by 1.52%~35.49% under 10^{10} guesses. Finally, the comparative experimental results under different guesses show that the performance of the optimal allocation strategy proposed in this paper is stably better than the average allocation strategy and random allocation strategy, and has a relative improvement of 16.87% on the password dataset with the largest dispersion. Moreover, the more diversified hybrid method shows better guessing efficiency as a whole.

Key words password security; data-driven guessing; probabilistic context-free grammar; Markov model; hybrid model

摘 要 基于文本口令的认证方法仍是当前用户身份认证的主流方式.为更好地研究口令安全性,研究人员提出了多种数据驱动的口令猜测方法,如概率上下文无关文法(probabilistic context-free

收稿日期:2021-05-11;**修回日期**:2022-01-29
基金项目:国家自然科学基金项目(U1836207);上海市科委基金项目(21511101600)
This work was supported by the National Natural Science Foundation of China (U1836207) and Shanghai Science and Technology Fund (21511101600).

grammars, PCFG)和马尔可夫(Markov)方法等.这些方法在猜测口令时有其独特的猜测优势,即能够以更小的猜测数猜中特定类型的口令.为充分利用这些优势以实现更优的猜测效率,提出了一个通用的参数化混合猜测框架.该框架由模型剪枝方法和理论证明最优的猜测数分配策略构成,能够混合不同数据驱动方法的猜测优势以生成更高效的猜测集.为了验证框架的通用性和最优性,通过分析并混合现有数据驱动猜测方法的不同优势,基于该框架设计了多个混合多元模型的参数化混合猜测方法(统称为hyPassGu)用于猜测实践.并且,还利用从真实网站泄露的4个大规模口令数据集(总共超过1.5亿条口令)对这些混合猜测方法进行了评估实验.实验结果表明,由不同方法组合构建的hyPassGu均表现出超越单一方法的猜测效率,且在10^{10}猜测数下超越了单一方法最优效率的1.52%~35.49%.此外,不同猜测数下的对比实验结果表明,提出的最优分配策略的猜测表现稳定,优于平均分配策略和随机分配策略,并在分布离散程度最大的口令数据集上有16.87%的相对提升,同时更多元的混合方法整体上也表现出更好的猜测效率.

关键词 口令安全;数据驱动猜测;概率上下文无关文法;马尔可夫模型;混合模型

中图法分类号 TP309

迄今为止,基于文本口令的身份认证仍是保护用户个人信息和在线财产的最流行方式之一[1].尽管近年来研究人员提出了各种身份认证方法,例如基于指纹或手势的身份认证[2],但因口令易用和低成本的显著优势[3-4],互联网用户仍倾向于使用文本口令.

为更好地研究口令的安全性,研究人员提出了多种数据驱动的口令猜测方法,如概率上下文无关文法(probabilistic context-free grammars, PCFG)[5]、马尔可夫(Markov)方法[6]和长短期记忆(long short-term memory, LSTM)[7]神经网络方法,并致力于优化这些猜测方法[8-17].由于这些方法具备不同的猜测优势,即能够以更小的猜测数猜中特定类型的口令,Ur等人[18]提出使用每条口令被不同方法猜中所需要的最小猜测数作为评估该口令安全性的保守指标,称为"Min_auto".这样的保守指标完美地结合了不同方法的猜测优势,可以看作是多方法混合猜测的理想上界.但在实际的猜测场景中,实现Min_auto的猜测效果需要的猜测数是单一方法的多倍,且不同方法会生成大量重复的猜测口令,严重浪费了计算资源.

为了在实际猜测的场景中减少计算资源的浪费,并在有限的猜测数内最大化地利用不同方法的猜测优势,本文提出了一个通用的参数化混合猜测的框架.该框架可以混合不同数据驱动方法的猜测优势以生成更高效的猜测集.该框架由2部分构成:模型剪枝和最优猜测数分配.模型剪枝可以确保框架中的每个方法仅生成自身擅长猜测的口令,从而避免与其他方法生成重复口令;而理论证明最优的猜测数分配方案可以确保不同方法生成指定猜测数的口令时,框架的整体猜测效率将达到最优.

为了验证框架的通用性和最优性,本文也基于该框架进行了猜测实践.本文分析了现有数据驱动的口令猜测方法的猜测优势,并根据分析结果将具备不同猜测优势的方法作为框架的组成模块,设计了多个混合不同猜测模型的参数化混合口令猜测方法(parameterized hybrid password guessing Method, hyPassGu).实验评估结果显示,不同方法组合构建的hyPassGu均超越了单一方法的猜测效率,验证了框架的通用性.而和其他猜测数分配方案的比较结果也证实了框架中猜测数分配方案的最优性.

本文的主要贡献有2个方面:

1) 提出了一个能够结合不同数据驱动方法猜测优势的通用参数化混合猜测框架.在该框架中,本文通过分析数据驱动方法的猜测模型,提出了模型剪枝的通用方案.此外,为了保证框架整体猜测效率的最优,本文提出并在框架中应用了理论证明最优且适用多个方法的猜测数分配策略.

2) 基于框架和对现有数据驱动方法的优势分析,提出了组合不同方法的多个参数化混合猜测方法(统称为hyPassGu),并通过对这些方法猜测效率的评估实验验证了框架的通用性和最优性.实验结果表明,不同方法组合构建的hyPassGu均表现出超越单一方法的猜测效率,在10^{10}猜测数下超越了单一方法最优效率的1.52%~35.49%.并且,相较于单一方法的最优效率,hyPassGu与Min_auto的差距相对缩短了16.78%~63.99%.此外,实验结果还表明,最优分配策略的猜测表现稳定优于平均分配

策略和随机分配策略,并在 10^8 猜测数下在离散程度最大的数据集上有 16.87%的相对提升,同时更多元的混合方法整体表现出更好的猜测效率.这些结果进一步验证了框架的通用性和最优性.

1 相关工作

1.1 数据驱动猜测方法的优化

在过去的数十年里,研究人员致力于改进现有数据驱动猜测方法在猜测某些口令时的不足,以此来提高方法的猜测效率.

对于 PCFG 方法,Veras 等人[13]利用自然语言处理算法提取分析口令中的语义模式,增强了模型对于字母字符串的泛化能力,使得模型可以生成更多可能的字母字符串.Houshmand 等人[8]通过引入更多的口令结构模式来更细粒度地定义口令中的结构,增强了模型对于结构的泛化能力.Li 等人[11]引入了拼音字典来弥补模型在中文口令破解方面的不足.韩伟力等人[15]通过裁剪 PCFG 模型的搜索空间并引入变形的方式,来生成与样本口令高度相似的模拟口令.邹静等人[16]则是根据用户习惯来 2 次划分高概率结构,以解决高概率结构只能生成少量口令的问题.

对于 Markov 方法,Ma 等人[12]提出了 Markov 模型的变种,即 n 阶 Markov 模型,来增强 Markov 模型对上下文信息的利用能力.这里的 n 指的是用来推测下一个字符出现概率的连续字符数量.此外,他们还提出了用变阶 Markov 模型,即 Backoff 来解决高阶 Markov 模型存在的数据稀疏问题.与 n 阶 Markov 模型不同,Backoff 会选择出现频率达到预设阈值的尽可能长的字符串来推测下一个出现的字符,以此在尽可能利用上文信息的同时避免数据稀疏的问题.

不同于已有研究对单一方法的改进思路,本文提出的参数化混合猜测旨在最大化利用多个数据驱动猜测方法各自不同的猜测优势,即利用不同方法的优势来弥补各自的不足.

1.2 现有猜测方法中的混合猜测实践

现有的猜测方法中,一些方法已经实现了混合猜测的简单实践,如 Hashcat[19]和 PCFGv4.1[20].

Hashcat 中的混合攻击模式实际上是一种组合攻击.它将字典攻击和其他诸如暴力攻击、掩码攻击的攻击方式相结合,既发挥了字典中单词常被用户使用的优势,又利用其他攻击方式来扩展字典攻击的搜索空间,提高了猜测效率.以字典攻击和暴力攻击相结合的方式为例,暴力攻击所能生成的字符串可以作为字典中每一个单词的前缀或者后缀.换句话说,如果使用暴力攻击来为单词添加 1 个数字作为后缀,那么根据单词“hello”可以得到“hello0”到“hello9”的猜测口令.

PCFGv4.1 则是设计了一个可选的参数“coverage”来控制基于 Markov 模型的暴力生成方法的使用.这个想法尝试使用暴力生成方法来扩展模型的搜索空间,从而提高猜测效率.但是,该工作缺乏对 PCFGv4.1 和 Markov 模型的深入分析,因而不能最大程度地利用每个方法的独特优势.

此外,也有一些研究工作对猜测集的混合进行了相关探索.Parish 等人[21]通过计算不同猜测方法生成的口令之间的相似性及互补性、利用不同方法的互补性进行混合猜测实验,进而提升混合猜测集的破解效率.汪定等人[22]则通过利用循环神经网络(recurrent neural network,RNN)、LSTM 分别和 PCFG,Markov,PR 模型组合生成猜测集的方法,首次证实了在相同猜测数下,组合不同类型模型所产生口令猜测集的破解率通常高于单一猜测集.这些研究工作对猜测集的混合做了一定探索,但缺乏针对猜测集的高效混合策略的系统分析和考虑,如利用不同方法的优势互补而非简单的互补来提高猜测集的有效性.

相较于已有工作在猜测集混合方面的探索,本文从优势互补的角度出发,通过理论分析和证明提出了针对数据驱动猜测方法的通用的参数化混合猜测框架,并基于该框架和现有的数据驱动猜测方法进行了猜测实践,实现了高效的参数化混合猜测.

2 参数化混合猜测框架

为了在有限的猜测数内最大化地利用不同方法的猜测优势,本文提出了如图 1 所示的参数化混合猜测的框架.该框架具体分为 2 个部分:模型剪枝和最优猜测数分配.其中模型剪枝是为了让不同猜测方法在发挥自身猜测优势的同时避免生成彼此重复的猜测口令,从而提高猜测数的利用率;而最优猜测数分配是为了在总猜测集的数目确定的情况下,通过调整不同猜测方法的可用猜测数(即图 1 中的 k_1~k_n),使得总猜测集的猜测效率最优.下面将介绍这 2 个部分的做法.

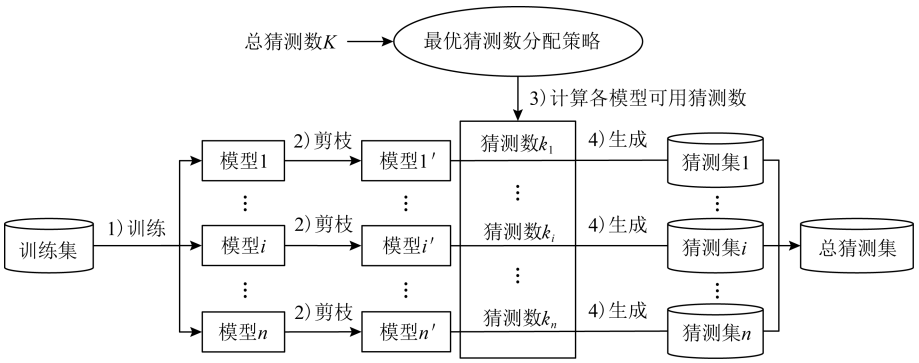


Fig. 1 Framework of parameterized hybrid password guessing

图 1 参数化混合口令猜测的框架

2.1 模型剪枝

数据驱动的猜测方法是基于训练集训练出的生成模型,一般可以看作 1 棵树,其中每个叶子节点对应的是 1 条猜测口令,而从根节点到叶子节点的路径则可以看作该猜测口令的生成路径.模型生成 1 条口令的过程就是从根节点查找到该口令对应的叶子节点的路径,而这条路径上的中间节点表示的是组成该口令的字符串片段以及对应的概率.如果某条口令对应的生成路径缺失了部分节点,那么模型从根节点将无法找到该口令对应的叶子节点,也因此不会生成出这条口令.

如图 2 所示,将数据驱动的生成方法抽象成 1 棵生成树,其中 S_i 表示生成过程中的 1 种状态,不同的下标数字表示不同的状态.在数据驱动的生成方法中,S_i 可以指代口令的上下文信息(如 Markov 模型使用的上下文字符)、口令的结构或者口令结构的内容实例(如 PCFG 模型使用的口令结构和结构对应的具体实例).

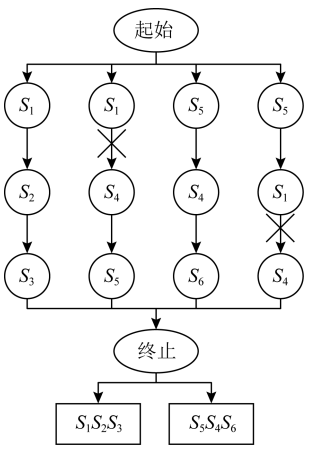


Fig. 2 Example of model pruning of data-driven password guessing methods

图 2 数据驱动的口令猜测方法的模型剪枝示例

从起始符开始,生成模型将根据生成过程的上一状态选择下一状态,直到遇到终止符为止.图 2 中删除了从状态 S_1 到状态 S_4 的转移路径,有 2 条生成路径将受此影响而无法正常生成猜测口令,因此图 2 中的生成方法最终只能生成由状态 $S_1S_2S_3$ 和 $S_5S_4S_6$ 构成的 2 条猜测口令.

通过删除一些不必要的生成路径,模型将只保留其擅长猜测的口令所对应的生成路径,而不同模型擅长猜测的口令互不相同,这就可以保证不同模型将不会生成重复的猜测.删除生成路径的做法就是模型剪枝,具体的实现方式是把生成路径中的某个不影响其他路径的节点删除,即将该节点对应的字符串片段和相应的转移概率从模型中移除.

2.2 最优猜测数分配

通过建立概率猜测方法猜测口令,破解数与猜测数之间关系的数学模型,本节提出了 2 个引理和 1 个定理来探寻最优的猜测数分配方案.该方案是基于理想攻击场景提出的,即训练集与测试集具有相同的分布.这类攻击场景也被称为站内猜测(intra-site guessing),被研究者们广泛采用[8,11,22-25].此外,考虑到大猜测数的情况下训练集的限制会导致概率猜测方法出现瓶颈,即破解口令数不再增长,本文在本节中仅分析概率猜测方法尚未达到破解瓶颈时的情况,并将在 4.5 节的实验评估中讨论大猜测数对分配方案的影响.

随着猜测数的增长,一个概率猜测方法破解出的口令数也会增长,也就是说它的“破解数-猜测数”曲线是一条单调递增的曲线(曲线上的点的横坐标表示使用的猜测数,纵坐标表示破解口令数).由于概率猜测方法会优先猜测概率更高的口令,而更高的概率往往意味着猜测方法认为该口令在测试集中出现的频率也更高.这带来的结果就是小猜测数下

单次命中能破解的口令数要多于大猜测数下单次命中能破解的口令数.换句话说,单位猜测数里被猜中的口令数会随着猜测数的增长而减小,即“破解数-猜测数”曲线的一阶导数逐渐变小,最后趋近于0.因此,一个概率猜测方法的“破解数-猜测数”曲线可以看作一个单调递增的凹函数(凹函数的二阶导数小于0,即一阶导数单调递减).

基于上述数学模型,最优的猜测数分配策略的目标是在不同概率方法的“破解数-猜测数”曲线上分别寻找1个点,使得所有点的横坐标的和为定值时,纵坐标的和是所有组合情况中的最大值,此时每个概率方法曲线上的点对应的横坐标就是被分配的可用猜测数.引理1证明了上述目标在概率方法仅有2个时的等价条件,定理1则是将等价条件拓展到 n 个方法的情况.

引理1. 给定2个光滑且单调递增的凹函数(concave function) $f(x)$ 和 $g(x)$,对于这2个函数上任意2个横坐标 (x_1,x_2) 和为 C(C 是常数)的点,存在 x_1^*,x_2^* 满足下列等价条件:

$$f(x_1^*)+g(x_2^*)=\max(f(x_1)+g(x_2))\Leftrightarrow f'(x_1^*)=g'(x_2^*).$$

证明. 首先,假设对于这2个函数上任意2个横坐标 (x_1,x_2) 和为 C(C 是常数)的点,存在 x_1^*,x_2^* 满足 $f(x_1^*)+g(x_2^*)=\max(f(x_1)+g(x_2))$.然后,将点 $(x_1^*,f(x_1^*))$ 向右稍稍移动到点 $(x_1^*+\Delta x,f(x_1^*+\Delta x))$,将点 $(x_2^*,g(x_2^*))$ 向左稍稍移动到点 $(x_2^*-\Delta x,g(x_2^*-\Delta x))$ $(\Delta x\to 0)$.根据假设,可以得到 $f(x_1^*)+g(x_2^*)\geqslant f(x_1^*+\Delta x)+g(x_2^*-\Delta x)$.同理,可以得到 $f(x_1^*)+g(x_2^*)\geqslant f(x_1^*-\Delta x)+g(x_2^*+\Delta x)$.然后开始下面的推导:

$$f(x_1^*)+g(x_2^*)=\max(f(x_1)+g(x_2)),\tag{1}$$

$$\Leftrightarrow\begin{cases}f(x_1^*)+g(x_2^*)\geqslant f(x_1^*-\Delta x)+g(x_2^*+\Delta x),\\ f(x_1^*)+g(x_2^*)\geqslant f(x_1^*+\Delta x)+g(x_2^*-\Delta x),\end{cases}\tag{2}$$

$$\Leftrightarrow\begin{cases}\lim\limits_{\Delta x\to 0}\dfrac{f(x_1^*)-f(x_1^*-\Delta x)}{\Delta x}\geqslant \lim\limits_{\Delta x\to 0}\dfrac{g(x_2^*+\Delta x)-g(x_2^*)}{\Delta x},\\ \lim\limits_{\Delta x\to 0}\dfrac{g(x_2^*)-g(x_2^*-\Delta x)}{\Delta x}\geqslant \lim\limits_{\Delta x\to 0}\dfrac{f(x_1^*+\Delta x)-f(x_1^*)}{\Delta x},\end{cases}\tag{3}$$

$$\Leftrightarrow\begin{cases}f'_-(x_1^*)\geqslant g'_+(x_2^*),\\ g'_-(x_2^*)\geqslant f'_+(x_1^*),\end{cases}\tag{4}$$

$$\Leftrightarrow f'(x_1^*)=g'(x_2^*).\tag{5}$$

式(1)到式(2)的正向推导在上面已经提过.式(2)意味着 $f(x_1^*)+g(x_2^*)$ 为 $f(x_1)+g(x_2)$ 的局部最大值,考虑到 $x_1+x_2=C$,$f(x_1)+g(x_2)$ 可以看作 $f(x)+g(C-x)$,而它的二阶导数为 $f''(x)+g''(C-x)$.由于 $f(x)$ 和 $g(x)$ 均为凹函数,它们的二阶导数小于0,因此 $f''(x)+g''(C-x)$ 也小于0,由此可得 $f(x)+g(C-x)$ 也是凹函数,因此局部最大值就是它的全局最大值,这意味从式(2)到式(1)也成立,因此式(1)和式(2)等价.而式(2)到式(3)是移位操作加上不等式的2边同时除以一个正数的操作,式(4)则是对式(3)的化简.式(4)中,$f'_-(x_1^*)$ 表示的是点 $(x_1^*,f(x_1^*))$ 处一阶导数的左极限值,$f'_+(x_1^*)$ 表示的是点 $(x_1^*,f(x_1^*))$ 处一阶导数的右极限值.$g'_-(x_2^*)$ 和 $g'_+(x_2^*)$ 的含义也类似.由于函数都是光滑的,因此每个点上的一阶导数值可以等价于一阶导数值的左极限值和右极限值,并且由于凹函数的性质决定了函数的一阶导数是单调递减的,因此可以得到 $f'_-(x_1^*)\geqslant f'_+(x_1^*)$ 和 $g'_-(x_2^*)\geqslant g'_+(x_2^*)$.结合式(4)我们可以推导出式(5):$f'(x_1^*)=g'(x_2^*)$.上述公式之间都具备等价性.

证毕.

定理1. 给定 n 个光滑且单调递增的凹函数 $f_1(x),f_2(x),\cdots,f_n(x)$,对于这 n 个函数上任意 n 个横坐标 $(x_1,x_2,\cdots,x_n)$ 和为 C(C 是常数)的点,存在 $x_1^*,x_2^*,\cdots,x_n^*$ 满足下列等价条件:

$$\sum_{i=1}^{n}f_i(x_i^*)=\max\sum_{i=1}^{n}f_i(x_i)\Leftrightarrow f'_1(x_1^*)=f'_2(x_2^*)=\cdots=f'_n(x_n^*),$$

$n\geqslant 2$ 且 n 为整数.

证明. 可以利用数学归纳法来证明该定理.首先,根据引理1,定理1在 $n=2$ 的时候成立.接着,假设定理1在 $n=k$($k\geqslant 2$ 且 k 是个整数)的时候成立.那么当 $n=k+1$ 的时候,对于 $k+1$ 个点中的任意 k 个点,定理1成立.换句话说,当固定 $k+1$ 个点中的任意1个点,然后寻找其他 k 个点的纵坐标和为最大值的情况时,这 k 个点的一阶导数一定相同.更进一步,固定 $k+1$ 个点中的其他任意点,再重复上述操作,可以发现 $k+1$ 个点的一阶导数的值会趋于相同,最终可以得到 $k+1$ 个点的纵坐标和为最大值的条件是这 $k+1$ 个点的一阶导数都相同.与引理1中式(2)到式(1)的推导类似,$k+1$ 个凹函数组成的函数的二阶导数也小于0,这意味组成的函数也是凹函数,所以找到的局部最大值也是全局最大值.

证毕.

由引理 1 和定理 1 可知,最优猜测数分配策略应当找到不同概率方法的“破解数-猜测数”曲线上一阶导数相等的点,即单位猜测数内破解口令数相等的点.文献[26]表明口令数据集的分布符合 Zipf 定律,即将口令按照在数据集中出现的频率降序排序之后,口令出现频率的 lg 值和口令排名的 lg 值满足式(6).基于该定律,对于理想的猜测方法(每次猜测能破解出剩余口令中概率最高的口令)而言,其每次猜测破解的口令数目的 lg 值(即式(6)中的 $\lg y$)和已经使用的猜测数的 lg 值(即式(6)中的 $\lg x$)满足:

$$\lg y=\lg C-\alpha\times\lg x. \tag{6}$$

尽管实际的概率猜测方法不能保证每次猜测都能够破解出 1 条口令,但在总体趋势上满足了优先猜测概率更高(出现频率也更高)的口令,因此可以假设概率猜测方法在单位猜测数内所能破解的口令数目的 lg 值和已经使用的猜测数的 lg 值也满足式(6).此外,同一数据集内不同猜测方法的 α 值会在该数据集 Zipf 分布的 α 值周围扰动,因此可以看作近似相同.本文也将在第 3 节的应用实践中通过实证实验来验证该假设的相关性质.

基于上述假设,引理 2 证明了不同猜测方法曲线上一阶导数相等的点的横坐标与数据集中口令分布的等价关系.

引理 2. 对于 2 个符合式(6)并且有着同样的 α 值的猜测方法,即 $\lg y_1=\lg C_1-\alpha\times\lg x_1$ 和 $\lg y_2=\lg C_2-\alpha\times\lg x_2$,对 $\forall y_1=y_2$,$\exists x_1/x_2\approx\sqrt[\alpha]{test_1/test_2}=\sqrt[\alpha]{train_1/train_2}$.其中,$test_1$ 和 $test_2$ 表示测试集中这 2 个猜测方法分别针对破解的类别的口令数量,而 $train_1$ 和 $train_2$ 表示训练集中和这 2 个猜测方法针对破解的类别相同的口令数量.

证明. 这 2 条单位猜测数内破解口令数曲线可表示为 $\lg y_1=\lg C_1-\alpha\times\lg x_1$ 和 $\lg y_2=\lg C_2-\alpha\times\lg x_2$.

首先,当它们的纵坐标相同时,即 $y_1=y_2$,可以得到 $x_1/x_2=\sqrt[\alpha]{C_1/C_2}$.

其次,当它们的横坐标相同时,即 $x_1=x_2$,可以得到 $y_1/y_2=C_1/C_2$.这意味着,对于任意的 $x_1=x_2$,y_1/y_2 是一个定值,因此可以得到 $C_1/C_2=y_1/y_2=\sum_{i=1}^{n}y_{1i}\Big/\sum_{i=1}^{n}y_{2i}=\int y_{1i}\mathrm{d}x\Big/\int y_{2i}\mathrm{d}x$.而事实上,$\int y_{1i}\mathrm{d}x$ 等于单位猜测数内破解口令数曲线与坐标轴形成的区域面积,也就是被破解的总口令数.当对猜测方法应用了恰当的平滑方法且使用的猜测数足够大时,可以相信测试集中所有的口令都能够被破解出来.由于不同的方法针对破解的口令类别不同,并且概率猜测方法假设训练集和测试集同分布,因此,当 $y_1=y_2$ 时,可以得到:

$$x_1/x_2=\sqrt[\alpha]{C_1/C_2}=\sqrt[\alpha]{\int y_{1i}\mathrm{d}x\Big/\int y_{2i}\mathrm{d}x}\approx\sqrt[\alpha]{test_1/test_2}=\sqrt[\alpha]{train_1/train_2},$$

这里引入了口令类别的概念来泛化各个方法的猜测优势,这样做一方面可以避免过拟合,使得方法的猜测优势不局限于特定数据集的特定口令,另一方面也可以利用训练集和测试集的同分布特性,基于训练集中的口令分布提前为框架中的各个方法分配猜测数.

综合上述引理 1,2 可知,当不同猜测方法使用的猜测数和各自的目标口令数量的比值正相关(即 $x_1:x_2:\cdots:x_n=\sqrt[\alpha]{test_1:test_2:\cdots:test_n}$,其中 $x_1,x_2,\cdots,x_n$ 表示不同方法各自使用的猜测数,α 是式(6)中的系数)时,框架可以达到总猜测效率的近似最优值.而由于训练集和测试集同分布,因此可以根据训练集的口令分布来为不同的猜测方法分配猜测数.

3 参数化混合猜测方法

基于第 2 节提出的框架,本节将现有的数据驱动猜测方法应用到框架中来设计实际可用的参数化混合猜测方法,具体步骤分为:优势分析、模型剪枝和猜测数分配.为方便表示,这些参数化混合猜测方法统称为 hyPassGu.

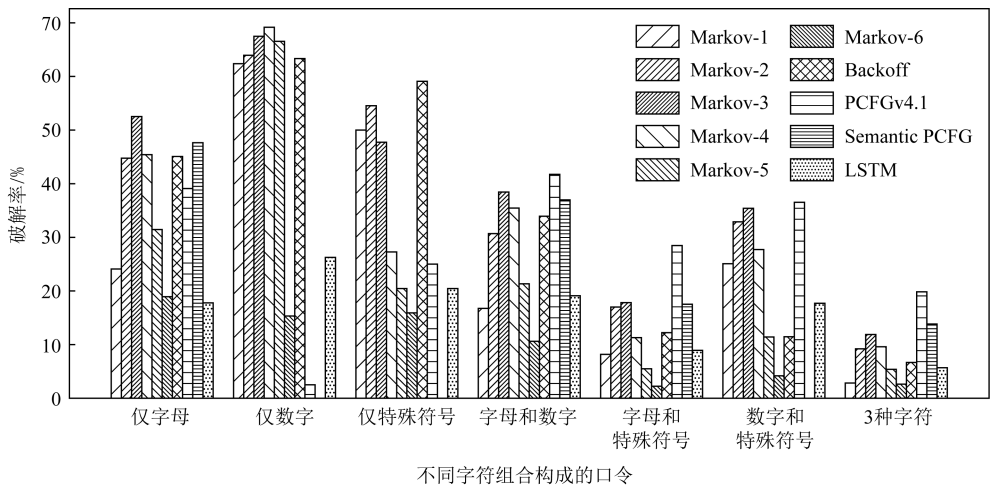
3.1 数据驱动方法的优势分析

本节将口令根据其构成字符种类分为 7 类:仅由字母构成、仅由数字构成、仅由特殊符号构成、由字母和数字构成、由字母和特殊符号构成、由数字和特殊符号构成、由 3 种字符一起构成.接着,本节量化比较了现有的概率猜测模型在不同类别的口令上的猜测效率,并分析了其优势背后的方法原理.为了衡量模型生成新的有效口令的能力,实验将数据集按 3:1的比例拆分成训练集和测试集,并从测试集中去除了在训练集中出现的口令来计算猜测效率.本节选取了中文数据集 CSDN 和英文数据集 Rockyou 来进行分析实验,以此探究各模型在这两种数据集上的猜测优势规律.

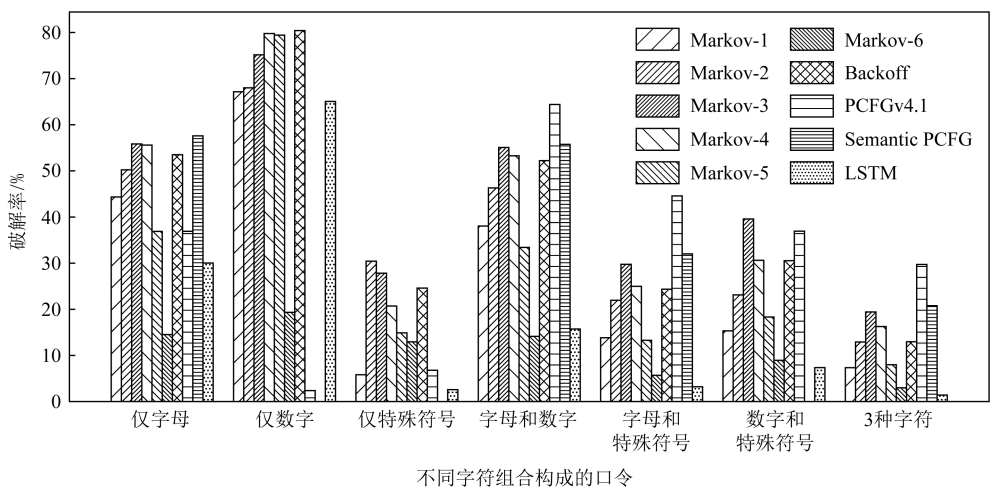
根据图 3 展示的结果,这些概率方法在不同类别的口令上表现差别很大.

3 阶 Markov(Markov-3)在猜测仅由字母构成

的口令上整体表现最优,在 CSDN 数据集上优于其他方法,在 Rockyou 数据集上略逊于 Semantic PCFG;4 阶 Markov(Markov-4)在猜测仅由数字构成的口令上整体表现最优,在 CSDN 数据集上优于其他方法,在 Rockyou 数据集上略逊于 Backoff;而在猜测仅由特殊符号构成的口令上则是 Backoff 和 2 阶 Markov(Markov-2)分别在 CSDN 和 Rockyou 数据集上表现最优.Markov 模型在 1 类字符构成的口令上的优异猜测效果表明,口令中使用的连续同类字符之间的关联会更加紧密,因此使用多个连续同类字符来预测下一个同类字符会很有效.并且,相较于使用连续出现的字符预测下一个出现的不同类的字符,预测同类的字符需要的搜索空间更小,这也会帮助模型达到更准确的预测效果.一般而言,更高的阶数会使得 Markov 模型的猜测效率更好,但过高的阶数也会造成过拟合的问题.例如,图 3 中的 6 阶 Markov 模型就表现出了严重的过拟合问题,导致其在各类口令上的猜测效果都不佳.



(a) CSDN



(b) Rockyou

Fig. 3 Guessing efficiency of single method on different types of passwords under 10^{10} guesses

图 3 10^{10}猜测数下单一方法在不同类别口令上的猜测效率

PCFGv4.1 在猜测 2 类及以上字符构成的口令方面表现几乎均优于其他方法,除了在 Rockyou 数据集的数字和特殊符号构成的口令上表现略逊于 3 阶 Markov,这是因为 PCFGv4.1 引入了多词模式、键盘模式、上下文模式和年份模式来更细粒度地区分口令的结构,从而使得模型在猜测 2 类及以上字符构成的口令时,一方面能生成更多可能的组合,另一方面也能减少一些不必要的搜索空间来提高猜测效率.其中,多词模式根据单词出现的频率将连续的长字母串分成多个单独单词来表示结构,可以丰富长字母串的组合情况;键盘模式将识别出的键盘上的连续输入看作特定的键盘结构,可以减少模型

针对数字、字母、特殊符号混杂的键盘串的不必要搜索;上下文模式识别常见的特殊符号和字母数字的组合来作为特定结,如“<3”,“:p”这些类似表情的输入,也是对不同类字符混杂字符串搜索空间的减少;而年份模式则是把单独出现的连续 4 个数字中符合要求的数字识别为年份,这种更细粒度的类别标签可以减少年份串和其他字符串组合时的搜索空间.

总的来说,现有的概率猜测模型中,Markov 模型(1~5 阶、Backoff)在猜测仅由 1 类字符构成的口令方面表现明显优势,而 PCFGv4.1 在猜测 2 类及以上字符构成的口令方面表现优于其他方法,两者互补的优势为后续参数化混合猜测方法的设计提供了便利.而 Markov-6、Semantic PCFG 和 LSTM 则没有表现出明显的猜测优势,因此不在混合猜测方法设计的考虑范围之内.

3.2 hyPassGu 的模型剪枝

了解到不同方法的猜测优势后,需要通过模型剪枝来限制各方法生成的口令.剪枝方法参照了 2.1 节的介绍,本文将以 Markov 模型和 PCFGv4.1 模型为例,结合 2.1 节的样例介绍具体的做法.

根据 3.1 节的优势分析结果,Markov 模型应当仅生成 1 类字符构成的口令,比如“password”这种仅由小写字母构成的口令,因此需要对其模型剪除 2 类及以上字符构成的口令的生成路径;PCFGv4.1 模型则与之相反,需要剪除 1 类字符构成的口令的生成路径.

Markov 模型的生成方法依赖上文的字符内容来预测下一个要生成的字符,对照图 2 中的示例可知,状态 S_i 在 Markov 模型中表示上下文的字符.以 1 阶 Markov 模型为例,口令“abc”的生成路径为“起始”-“起始 a”-“ab”-“bc”-“c 终止”.通过观察可以发现,在 1 类字符构成的口令所对应的生成路径中,所有的状态都仅包含 1 类字符,而 2 类及以上字符构成的口令中至少存在 1 段路径的状态包含多类字符,因此只需要从 Markov 模型中删除包含多类字符的状态转移路径,就可以剪除 2 类及以上字符构成的口令的生成路径.在实际的实践中,通过限制训练集为仅包含 1 类字符构成的口令,来使得 Markov 模型统计的状态转移只包含同类字符之间的情况,从而保证 Markov 模型仅能生成 1 类字符构成的口令.

PCFG 模型的生成方法与 Markov 模型略有不同,它的状态分为 2 类:一类是口令结构、一类是结构的具体内容实例.举例来说,口令“abc123”的生成路径为“起始”-“L_3D_3”-“$abcD_3$”-“abc123”-“终止”.通过观察可以发现,在起始符到口令结构的选择这一步,PCFG 模型通过口令结构就可以知道最终生成的口令将包含几类字符,因此通过删除 1 类字符构成的口令所对应的口令结构,模型将只能生成 2 类及以上字符构成的口令.在实际的实践中,对于 PCFGv4.1 训练得到的口令结构文件进行分析,去除了其中仅由 1 类字符构成的口令结构,并对剩余的口令结构概率做了归一化处理.具体来说,本文去除了仅由“A”(英文字母)、“D”(数字)、“O”(特殊符号)、“Y”(年份,纯数字)中的某一类标签组成的口令结构.

3.3 hyPassGu 的猜测数分配

在对 hyPassGu 中的模型剪枝后,参照 2.2 节中的最优猜测数分配策略,本节将验证式(6),并探究在实际数据集上 α 取值的情况.以 10^8 作为单位猜测数,本节使用式(6)来拟合实际的猜测方法.

表 1 展示了 CSDN 数据集上的拟合结果.PCFGv4.1 方法破解情况线性拟合的决定系数 $R^2=0.99$,Markov 方法破解情况线性拟合的决定系数平均为 0.75 左右,这意味着大部分数据点符合回归曲线.

Table 1 Fitting Results of Zipf's Law Based on CSDN

表 1 基于 CSDN 数据集 Zipf 定律的拟合结果

猜测方法	Zipf 拟合曲线方程	决定系数(R^2)	p
PCFGv4.1	$\lg y=4.653\,539-1.227\,910\times\lg x$	0.985 044	$<10^{-3}$
Markov-1	$\lg y=4.773\,878-1.175\,560\times\lg x$	0.796 625	$<10^{-3}$
Markov-2	$\lg y=4.880\,280-1.254\,109\times\lg x$	0.751 652	$<10^{-3}$
Markov-3	$\lg y=4.793\,465-1.182\,219\times\lg x$	0.746 041	$<10^{-3}$
Markov-4	$\lg y=4.916\,158-1.306\,722\times\lg x$	0.805 097	$<10^{-3}$
Markov-5	$\lg y=4.782\,584-1.262\,160\times\lg x$	0.729 870	$<10^{-3}$
Backoff	$\lg y=4.854\,809-1.247\,217\times\lg x$	0.678 522	$<10^{-3}$

此外,p 值检验的结果显示所有的 $p<10^{-3}$,这意味着回归曲线的系数可以很好地表示数据点.不同方法在 CSDN 数据集上拟合结果的系数 α 值都很接近,集中在 1.2 和 1.3 左右.此外,本文基于 Rockyou 数据集的拟合结果与基于 CSDN 数据集的拟合结果相似,系数 α 值也集中在 1.2 和 1.3 左右.基于这些结果,本文推荐使用 1.2 和 1.3 作为最优猜测数分配策略中 α 的值.

4 实验与评估

基于第 3 节的方法设计,本节通过实验评估不同配置下的 hyPassGu 的猜测效率,以此验证框架的通用性以及猜测数分配策略的最优性.此外,还探究了猜测数大小和口令数据集离散程度对分配策略的影响,并讨论了混合方法种类数对猜测效率的影响.

4.1 数据集

如表 2 所示,本文在实验中使用了 4 个从真实网站泄露的大规模口令数据集:CSDN[27],Youku[28],Rockyou[29],Neopets[30].其中,CSDN 和 Youku 的数据集泄露自中文网站,而 Rockyou 和 Neopets 的数据集泄露自英文网站.此外,CSDN 和 Rockyou 数据集是 2011 年之前泄露的,Youku 和 Neopets 数据集则是近 6 年泄露的.这些数据集在区域跨度和时间跨度上都颇具代表性,并且在先前的研究[11-12,14,23-25]中被研究人员广泛使用.

为了尽可能保证本文实验中使用的数据是真实用户创建的口令,本文参照先前的研究[12],从数据集中清理了包含非可打印 ASCII 码的口令和长度超过 32 的口令,最终本文从 4 个数据集中获取了共计超过 1.5 亿条的口令数据.

Table 2 Basic Information of Password Datasets

表 2 口令数据集的基本信息

数据集	语言	泄露年份	口令总数	清理后总数
CSDN	中文	2011	6 425 650	6 378 862
Youku	中文	2016	47 593 747	47 412 795
Rockyou	英文	2009	32 603 387	31 108 427
Neopets	英文	2016	68 743 269	67 672 205

本节实验中的数据集均按 3∶1随机拆分为训练集和测试集,并且为了衡量模型生成新的有效口令的能力,测试集中去除了训练集中出现过的口令后来计算猜测效率.

4.2 模型剪枝后的优势评估实验

根据 3.1 节分析得到的结论,本节按照 3.2 节介绍的方法对 Markov 模型(1～5 阶、Backoff)和 PCFGv4.1 这 7 个模型进行剪枝,并评估了各模型在剪枝后对其擅长猜测口令的破解率相较于剪枝前的提升幅度.评估结果如表 3 所示,各模型在剪枝后,在其擅长猜测口令上的破解率相较于之前均有不同程度的相对提升(0.24%～54.45%),验证了剪枝方案的有效性.

Table 3 Relative Improvement of Guessing Advantages After Model Pruning Under 10^{10} Guesses

表 3 10^{10} 猜测数下模型剪枝后的猜测优势的相对提升 %

数据集	Markov-1	Markov-2	Markov-3	Markov-4	Markov-5	Backoff	PCFGv4.1
CSDN	7.56	7.17	9.66	8.28	3.21	12.20	0.75
Youku	15.20	21.56	18.53	17.62	15.38	3.58	0.53
Rockyou	12.35	13.16	9.30	4.75	0.59	54.45	1.41
Neopets	15.96	12.81	8.51	10.59	3.91	45.05	0.24

4.3 二元 hyPassGu 的评估实验

1) 实验配置

本节选取了 6 对存在不同优势的概率模型,即 PCFGv4.1 和 6 种不同的 Markov 模型(Backoff 和 1～5 阶 Markov 模型).模型剪枝后,PCFGv4.1 将只能生成 2 类及以上字符构成的口令,而 Markov 模型将只能生成 1 类字符构成的口令.图 4 展示了组合不同猜测方法的二元 hyPassGu 在 CSDN 数据集上的猜测效果,其中每张子图里的 Min_auto 曲线使用的模型为该子图中对应的 2 个单一方法.

2) 实验结果

如图 4 所示,二元 hyPassGu 的猜测效率相较于单一方法提升了 1.52%～35.49%.其中,优势差异越明显的 2 个方法,在参数化混合之后的猜测效率提升就越大,但想要获得更高的总破解效率,则需要结合优势更强的方法.

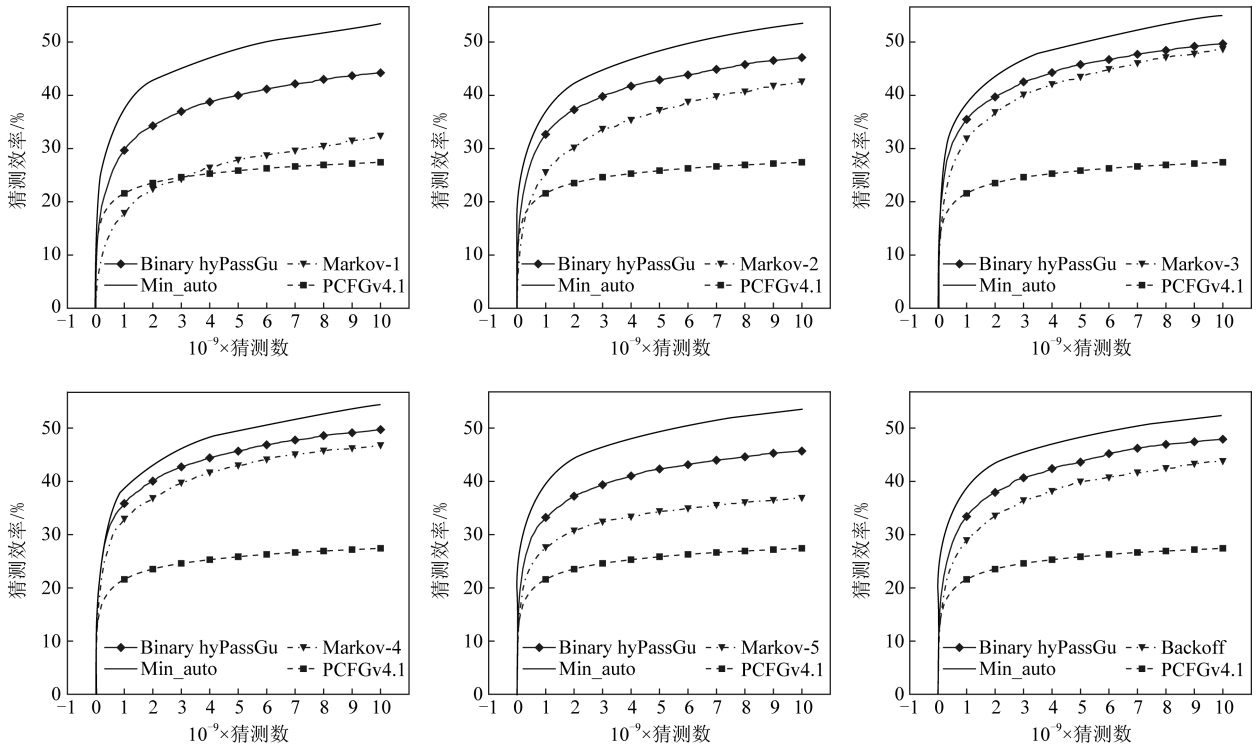


Fig. 4 Guessing efficiency of binary hyPassGu compared with that of other methods based on CSDN

图 4 基于 CSDN 数据集二元 hyPassGu 与其他方法的猜测效率

具体来说,1 阶 Markov 在猜测仅由数字构成的口令上表现远优于 PCFGv4.1,而 PCFGv4.1 在猜测 2 类及以上的字符构成的口令上表现远优于 1 阶 Markov,因此两者的优势结合后猜测效率的相对提升达到了 35.49%;而对于 3 阶 Markov 和 4 阶 Markov 而言,它们在猜测 2 类及以上字符构成的口令上不如 PCFGv4.1 表现优异,但与 PCFGv4.1 的差距相较于 1 阶 Markov 来说要小很多,因此优势结合后猜测效率的相对提升也要少一些.但是,鉴于 3 阶 Markov 和 4 阶 Markov 在猜测 1 类字符构成的口令上更强的优势,将它们与 PCFGv4.1 结合可以在总的破解率上达到更高,即达到 49.88%和 49.92%的破解率.

此外,与多模型结合的理想上界 Min_auto 相比,二元 hyPassGu 与它在猜测效率上的差距控制在 5.22%~6.68%.相较于单一方法的最优效率与 Min_auto 的差距,二元 hyPassGu 将差距缩短了 16.78%~63.99%.

4.4 三元 hyPassGu 的评估实验

1) 实验配置

本节选取 3 阶 Markov、4 阶 Markov 和 PCFGv4.1 来设计三元 hyPassGu.考虑到仅由特殊符号构成的口令在现有的真实口令集中的占比不到 0.1%,难以针对这类口令单独分配猜测数,因此由 3 阶 Markov 来猜测仅由特殊符号构成的口令.各模型的猜测任务具体为:3 阶 Markov 猜测仅由字母构成和仅由特殊符号构成的口令;4 阶 Markov 猜测仅由数字构成的口令;PCFGv4.1 猜测 2 类及以上字符构成的口令.在 4 个数据集上的评估实验结果如图 5 所示,其中 Min_auto 曲线中使用的模型为 3 阶 Markov、4 阶 Markov 和 PCFGv4.1.

2) 相较于单一方法的猜测效率提升

如图 5 所示,在 4 个数据集上的评估实验表明,三元 hyPassGu 在 10^{10} 猜测数下的猜测效率超越单一猜测方法的最优猜测效率 4.55%~11.58%.不同数据集上单一方法的表现各有优劣,其中在 CSDN 数据集上 3 阶 Markov 表现最优,在 Youku 数据集上 4 阶 Markov 表现最优,在 Rockyou 数据集上,3 阶 Markov 和 4 阶 Markov 表现最优,而在 Neopets 数据集上,PCFGv4.1 和 4 阶 Markov 表现最优,而相比之下,结合了不同方法猜测优势的三元 hyPassGu 在 4 个数据集上均表现出相较于单一猜测方法猜测效率的稳定提升.

此外,相较于单一方法的最优效率,三元 hyPassGu 与 Min_auto 猜测效率的差距缩短了 22.54%~43.19%.

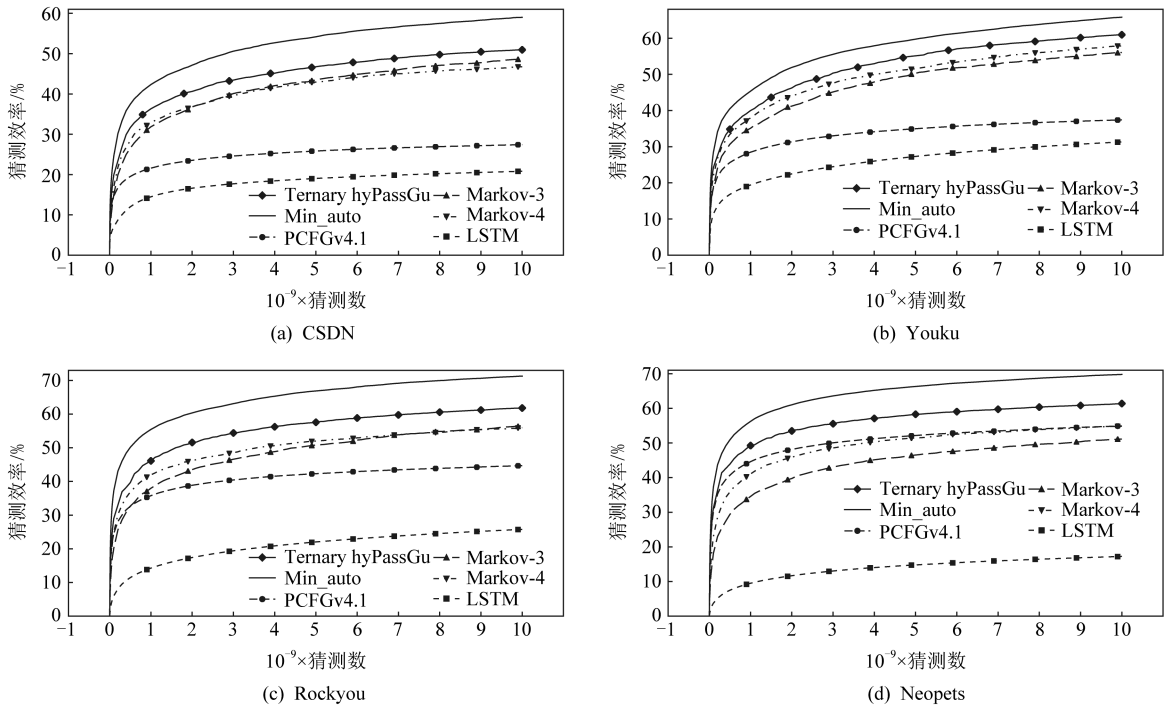


Fig. 5 Guessing efficiency of ternary hyPassGu compared with that of other methods

图 5 三元 hyPassGu 与其他方法的猜测效率的比较

3) 不同猜测数分配策略的比较

本节利用 4 个数据集上不同策略的比较实验来验证 2.2 节提出的最优猜测数分配策略的最优性.

如表 4 所示,第 1 列表示数据集,第 2 列和第 3 列表示不同 α 取值的最优猜测数分配策略下的三元 hyPassGu 的猜测效率,第 4 列表示平均分配策略(即混合方法中每个方法使用的猜测数一样)下的三元 hyPassGu 的猜测效率,第 5 列表示 1 000 组随机分配策略下的三元 hyPassGu 的平均猜测效率,第 6 列表示在实际的破解实验中三元 hyPassGu 能达到的最优性能.为了找到三元 hyPassGu 的实际最优性能,需要比较模型在所有可能配置下的性能.以 10^{10} 猜测数为例,需要使用 PCFGv4.1、3 阶 Markov 和 4 阶 Markov 各自生成 10^{10} 条猜测口令,然后遍历总数为 10^{10} 的所有可能的组合,最终找到能猜测出最多数量口令的组合并把它的表现作为 hyPassGu 的最优性能.

从表 4 中可以看出,不同 α 值对猜测效率的影响很小,整体来说 $\alpha=1.2$ 时猜测效率更好.最优分配策略下的三元 hyPassGu 表现远优于随机分配策略下的三元 hyPassGu,超出了 7.60%～17.05%.此外,最优分配策略下的三元 hyPassGu 的猜测效率与实际最优性能的差值在 0.07%～0.42%之间,而平均分配策略下的三元 hyPassGu 的猜测效率与实际最优性能的差值是上述差值的 3.6～8.7 倍.上述差值表明最优分配策略的可提升空间很有限,因此可以认为本文提出的猜测数分配策略是最优的.

Table 4 Guessing Efficiency of Ternary hyPassGu with Different Allocation Strategies Under 10^{10} Guesses

表 4 10^{10} 猜测数下不同分配策略的三元 hyPassGu 的猜测效率 %

数据集	最优		平均	随机	实际最优性能
	$\alpha=1.2$	$\alpha=1.3$			
CSDN	50.96	50.96	50.15	45.08	51.27
Youku	60.95	60.94	59.25	52.07	61.37
Rockyou	61.86	61.86	61.52	57.49	61.93
Neopets	61.29	61.23	58.98	53.90	61.59

4.5 最优猜测数分配策略的探究实验

鉴于表 4 中最优分配策略在 10^{10} 猜测数下相较于平均分配策略并没有明显的猜测效率优势(尤其在 CSDN 和 Rockyou 数据集上),本节利用蒙特卡洛方法[31]进一步探究了猜测数大小和口令数据集离散程度这 2 个因素对最优猜测数分配策略的影响.其中,根据已有研究[8,11,13,18,21-25]常用生成离线口令猜测集的大小,本节选取了 10^8～10^{14} 作为猜测数

的范围;另一方面,本节使用各特征的口令占比计算得到的标准差(standard deviation, SD)作为口令数据集离散程度的度量指标.

实验结果如表 5 所示,第 1 列表示数据集,第 2 列表示各数据集中不同特征口令占比计算得出的标准差(该值的大小反应了数据集中不同特征口令分布的离散程度,值越大,数据集口令分布越不均匀),第 3,6,9,12 列表示各猜测数下平均分配策略的猜测效率,第 4,7,10,13 列表示各猜测数下最优分配策略的猜测效率(根据 4.4 节的结果,$\alpha=1.2$),第 5,8,11,14 列则表示最优分配策略相较于平均分配策略的相对提升.

Table 5 Performance Improvement of Optimal Allocation Strategy Under Different Guesses when $\alpha=1.2$

表 5 $\alpha=1.2$ 时不同猜测数下最优分配策略的性能提升

数据集	标准差	10^8 猜测数下性能/%			10^{10} 猜测数下性能/%			10^{12} 猜测数下性能/%			10^{14} 猜测数下性能/%		
		平均	最优	提升	平均	最优	提升	平均	最优	提升	平均	最优	提升
CSDN	0.18	19.13	19.91	4.08	50.15	50.96	1.62	66.18	66.25	0.11	70.35	70.42	0.10
Youku	0.26	23.15	24.23	4.67	59.25	60.95	2.87	78.03	78.58	0.70	83.41	83.61	0.24
Rockyou	0.15	28.82	29.03	0.73	61.52	61.80	0.46	77.06	77.39	0.43	83.79	83.95	0.19
Neopets	0.38	26.74	31.25	16.87	58.98	61.38	4.07	74.85	76.34	1.99	83.22	83.98	0.91

1) 猜测数大小对最优分配策略的影响

如表 5 所示,在 10^8 猜测数下,最优分配策略相较于平均分配策略的提升效果最好,平均相对提升 6.59%.而当猜测数在 $10^8\sim10^{14}$ 之间时,最优分配策略相较于平均分配策略的提升效果逐渐降低,原因是随着猜测数的增长,概率方法由于训练集的限制逐渐达到瓶颈,破解数将不再增长,不同分配策略的效率差距也因此减小.

2) 数据集的口令离散程度对最优分配策略的影响

如表 5 所示,通过对比可以发现,最优分配策略相较于平均分配策略的提升效果与标准差值呈正相关,即标准差值越大,数据集口令分布越不均匀、最优分配策略的猜测优势越明显.其中,Neopets 数据集的离散程度最高,最优分配策略带来的提升也最明显,在 10^8 猜测数下提升可以达到 16.87%.这是

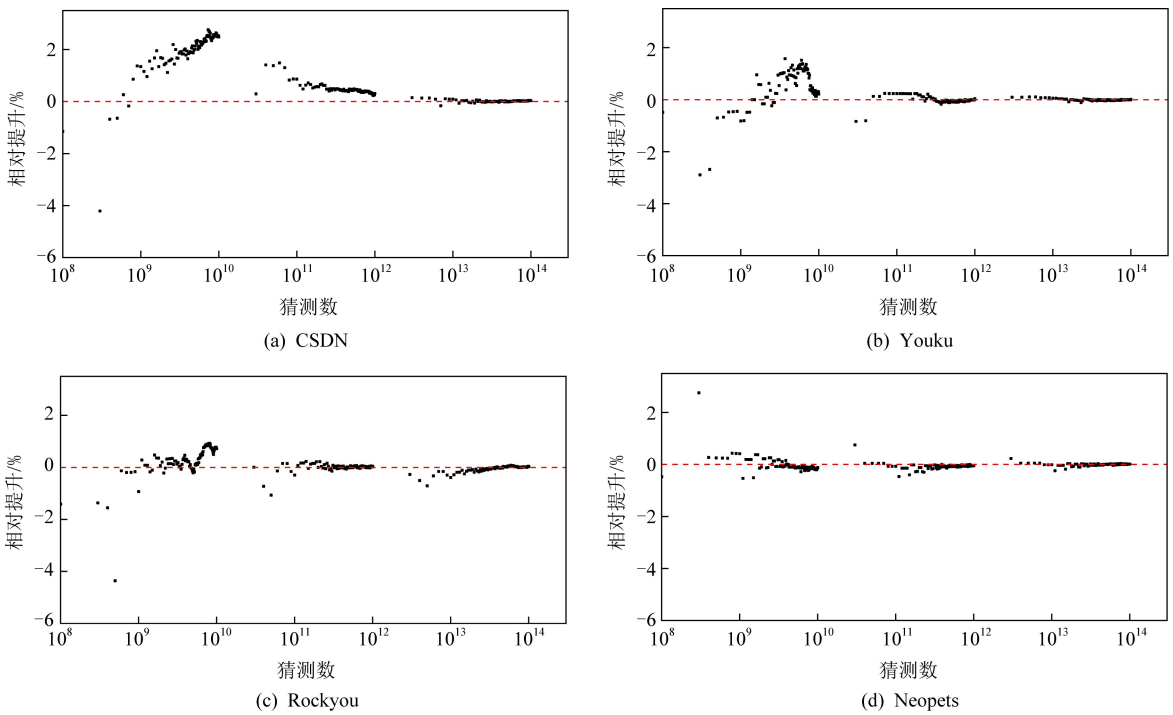


Fig. 6 Improvement of the best performance of ternary hyPassGu compared with that of binary hyPassGu

图 6 三元 hyPassGu 相较于二元 hyPassGu 最优效率的提升

因为当标准差值较小、数据集中对应特征的口令分布趋于平均时,最优分配策略将采取和平均分配策略基本一致的分配方案,两者猜测效率也将相似.

4.6 混合方法种类(元数)对破解效果的影响

本节通过对比分析三元 hyPassGu 和二元 hyPassGu 的猜测效率,讨论混合方法种类数(元数)对破解效果的影响.图 6 中的横坐标表示使用的猜测数(考虑到最优分配策略发挥作用的猜测数区间,这里也选取了 $10^8 \sim 10^{14}$ 作为猜测数的范围),纵坐标表示的是三元 hyPassGu 相较于二元 hyPassGu 中最好的猜测效率的相对提升.其中,三元 hyPassGu 混合的是 PCFGv4.1、3 阶 Markov 和 4 阶 Markov,而二元 hyPassGu 分别混合的是 PCFGv4.1 和 3 阶 Markov,以及 PCFGv4.1 和 4 阶 Markov.

从图 6 可以看出,三元 hyPassGu 相较于二元 hyPassGu 整体表现更好,有一定的猜测效率提升,而在较大猜测数下($10^{12} \sim 10^{14}$),三元 hyPassGu 和二元 hyPassGu 的猜测效率趋于相同.这一点与最优分配策略的提升表现一致,意味着概率方法在大猜测数下的破解瓶颈也导致了更多元的混合方法在大猜测数下并未有明显的猜测效率提升.

5 讨 论

1) 更多元的混合猜测方法

随着口令猜测方面研究的发展,优化的猜测方法甚至全新的方法,例如基于表征学习的方法[22]正不断被提出.将一个新的口令猜测方法结合到本文提出的框架中有以下 2 种情况:①如果它根据概率降序生成候选口令,参考本文第 3 节的步骤就可以尝试将它结合到框架中.②如果它随机地生成候选口令需要先引入口令的概率计算和排序,例如,表征学习的采样方法[22],而这是我们未来的工作方向.

此外,本文在分析基于深度学习技术的口令猜测方法 LSTM 时发现其没有相较于其他方法独有的猜测优势,因此在实现混合方法时并未结合基于深度学习的口令猜测方法.在未来的研究中,可以考虑利用迁移学习等技术来提高深度学习方法在特定类型口令上的猜测效率,并且在改进 PCFG 等传统方法时也可从强化已有猜测优势的角度入手,进一步提高这些概率方法的优势猜测能力,从而使得多元混合方法在大猜测数下(10^{12} 以上)也能有更好的猜测表现.

2) 最优猜测数分配策略的可能偏差

猜测数的最优分配策略可能会由于以下一些步骤导致一些偏差:首先,当使用 Zipf 定律去拟合不同猜测方法单位猜测数内破解的口令数时,Markov 方法的决定系数 $R^2 < 0.9$,这意味着拟合结果不能很好地解释所有的原始数据.其次,单位猜测数内破解的口令数的积分值被近似为目标集的大小是很理想的近似,而实际的曲线积分值会小于目标集的大小.尽管这些可能的偏差会导致最优分配策略和实际最优性能的差距,但是第 4 节的实验评估结果显示两者的表现非常接近,并且可供继续优化的空间也很有限.此外,分配策略中 α 值的选择依赖于实证分析的结果,但是 4 个数据集上的良好表现也表明 α 值的选择可以看作是通用的.

6 总 结

本文提出了一个能够在有限猜测数内充分利用不同方法的猜测优势来高效地生成口令的参数化混合猜测的框架.基于现有的数据驱动猜测方法,本文通过分析它们的猜测优势并组合不同方法来构建实际可用的参数化混合猜测方法.评估结果表明,不同方法组合构建的参数化混合猜测方法均表现出超越单一方法的猜测效率,验证了框架的通用性.在 10^{10} 猜测数下,参数化混合猜测方法超越了单一方法的最优效率 1.52%~35.49%,且最优猜测数策略配置下的参数化混合猜测方法表现均优于其他策略配置下的参数化混合猜测方法.此外,相较于单一方法的最优效率,参数化混合猜测方法与 Min_auto 理想值的差距相对缩短了 16.78%~63.99%.不同猜测数下的实验结果还表明,最优分配策略的猜测表现优于平均分配策略和随机分配策略,并在 10^8 猜测数下,在离散程度最大的数据集上有 16.87%的相对提升.这些结果进一步验证了框架的最优性.

本文的实验结果与分析充分说明了结合不同概率模型的猜测是比单一猜测方法更高效的生成口令猜测集的方式.在接下来的工作中,我们将尝试将更多的猜测方法结合到本文提出的混合猜测框架中.我们还将引入不同的口令分类方案,例如根据口令长度分类,更加深入挖掘数据驱动方法的猜测优势.此外,我们还将研究动态调整框架中不同方法可用猜测数的方案.

作者贡献声明:韩伟力负责提出混合不同猜测方法优势的研究方向和算法思路,以及文章整体架

构的设计和修改;张俊杰负责参数化混合猜测方法hyPassGu的设计和改进,以及全文内容的撰写;徐铭负责全文内容表达的修改,以及hyPassGu改进思路的讨论和补充;王传旺和张浩东负责完成hyPassGu和其他方法的对比实验,以及文章定理的验证;何震瀛和陈虎负责文章整体思路的指导,以及文章内容的修改.

参 考 文 献

[1] Wang Ping, Wang Ding, Huang Xinyi. Advances in password security [J]. Computer Research and Development, 2016, 53(10): 2173-2188 (in Chinese)
(王平, 汪定, 黄欣沂. 口令安全研究进展[J]. 计算机研究与发展, 2016, 53(10): 2173-2188)

[2] Cheon E, Shin Y, Huh J H, et al. Gesture authentication for smartphones: Evaluation of gesture password selection policies [C] //Proc of the 41st IEEE Symp on Security and Privacy. Los Alamitos, CA: IEEE Computer Society, 2020: 249-267

[3] Bonneau J, Herley C, Oorschot P C, et al. The quest to replace passwords: A framework for comparative evaluation of Web authentication schemes [C] //Proc of the 33rd IEEE Symp on Security and Privacy. Los Alamitos, CA: IEEE Computer Society, 2012: 553-567

[4] Herley C, Oorschot P C. A research agenda acknowledging the persistence of passwords [J]. IEEE Security & Privacy, 2012, 10(1): 28-36

[5] Weir M, Aggarwal S, Medeiros B, et al. Password cracking using probabilistic context-free grammars [C] //Proc of the 30th IEEE Symp on Security and Privacy. Los Alamitos, CA: IEEE Computer Society, 2009: 391-405

[6] Narayanan A, Shmatikov V. Fast dictionary attacks on passwords using time-space tradeoff [C] //Proc of the 12th CCS. New York: ACM, 2005: 364-372

[7] Melicher W, Ur B, Segreti S M, et al. Fast, lean, and accurate: Modeling password guessability using neural networks [C] //Proc of the 25th USENIX Security Symp. Berkeley, CA: USENIX Association, 2016: 175-191

[8] Houshmand S, Aggarwal S, Flood R. Next gen PCFG password cracking [J]. IEEE Transactions on Information Forensics and Security, 2015, 10(8): 1776-1791

[9] Hranický R, Zobal L, Rysavý O, et al. Distributed PCFG password cracking [G] //LNCS 12308: Proc of the 25th ESORICS. Berlin: Springer, 2020: 701-719

[10] Li Yue, Wang Haining, Sun Kun. A study of personal information in human-chosen passwords and its security implications [C/OL] //Proc of the 35th INFOCOM. Piscataway, NJ: IEEE, 2016 [2022-01-20]. https://ieeexplore.ieee.org/document/7524583

[11] Li Zhigong, Han Weili, Xu Wenyuan. A large-scale empirical analysis of Chinese Web passwords [C] //Proc of the 23rd USENIX Security Symp. Berkeley, CA: USENIX Association, 2014: 559-574

[12] Ma J, Yang Weining, Luo Min, et al. A study of probabilistic password models [C] //Proc of the 35th IEEE Symp on Security and Privacy. Los Alamitos, CA: IEEE Computer Society, 2014: 689-704

[13] Veras R, Collins C, Thorpe J. On semantic patterns of passwords and their security impact [C/OL] //Proc of the 21st NDSS. Reston, VA: The Internet Society, 2014 [2022-01-20]. https://www.ndss-symposium.org/ndss2014/programme/semantic-patterns-passwords-and-their-security-impact/

[14] Wang Ding, Zhang Zijian, Wang Ping, et al. Targeted online password guessing: An underestimated threat [C] //Proc of the 23rd CCS. New York: ACM, 2016: 1242-1254

[15] Han Weili, Yuan Lang, Li Sisi, et al. An efficient algorithm to generate password sets based on samples [J]. Chinese Journal of Computers, 2017, 40(5): 1151-1167 (in Chinese)
(韩伟力, 袁琅, 李思斯, 等. 一种基于样本的模拟口令集生成算法[J]. 计算机学报, 2017, 40(5): 1151-1167)

[16] Zou Jing, Lin Dongdai, Hao Chunhui. A password cracking method based on structure division probability [J]. Chinese Journal of Computers, 2014, 37(5): 1206-1215 (in Chinese)
(邹静, 林东岱, 郝春辉. 一种基于结构划分概率的口令攻击方法[J]. 计算机学报, 2014, 37(5): 1206-1215)

[17] Xie Zhijie, Zhang Min, Guo Yuqi, et al. Modified password guessing methods based on TarGuess-I [J/OL]. Wireless Communication and Mobile Computing, 2020 [2022-01-20]. https://www.researchgate.net/publication/347323082_Modified_Password_Guessing_Methods_Based_on_TarGuess-I

[18] Ur B, Segreti S M, Bauer L, et al. Measuring real-world accuracies and biases in modeling password guessability [C] //Proc of the 24th USENIX Security Symp. Berkeley, CA: USENIX Association, 2015: 463-481

[19] Steube J, Gristina G. Hashcat [EB/OL]. (2015-12-05) [2022-01-20]. https://hashcat.net/hashcat/

[20] Weir M. Practical PCFG password cracking [EB/OL]. (2019-12-01) [2022-01-20]. https://www.youtube.com/hashtag/passwordscon

[21] Parish Z, Cushing C, Aggarwal S, et al. Password guessers under a microscope: An in-depth analysis to inform deployment [J]. arXiv preprint, arXiv: 2008.07986, 2020

[22] Wang Ding, Zou Yunkai, Tao Yi, et al. Password guessing based on recurrent neural networks and generative adversarial networks [J]. Chinese Journal of Computers, 2021, 44(8): 1519-1534 (in Chinese)
(汪定, 邹云开, 陶义, 等. 基于循环神经网络和生成式对抗网络的口令猜测模型研究[J]. 计算机学报, 2021, 44(8): 1519-1534)

[23] Ji Shouling, Yang Shukun, Hu Xin, et al. Zero-sum password cracking game: A large-scale empirical study on the crackability, correlation, and security of passwords [J]. IEEE Transactions Dependable and Secure Computing, 2017, 14(5): 550-564

[24] Pasquini D, Gangwal A, Ateniese G, et al. Improving password guessing via representation learning [C] //Proc of the 42nd IEEE Symp on Security and Privacy. Los Alamitos, CA: IEEE Computer Society, 2021: 1382-1399

[25] Wang Ding, Wang Ping, He Debiao, et al. Birthday, name and bifacial-security: Understanding passwords of Chinese Web users [C] //Proc of the 28th USENIX Security Symp. Berkeley, CA: USENIX Association, 2019: 1537-1555

[26] Wang Ding, Cheng Haibo, Wang Ping, et al. Zipf's law in passwords [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2776-2791

[27] Mandalia R. Hackers strike Chinese cyberspace, leak data of 6 million CSDN users [EB/OL]. (2011-12-29) [2022-01-20]. https://www.itproportal.com/2011/12/29/hackers-strike-chinese-cyberspace-leak-data-6-million-csdn-users/

[28] Waqas. Chinese video service giant Youku hacked: 100M accounts sold on dark Web [EB/OL]. (2017-04-16) [2022-01-20]. https://www.hackread.com/chinese-video-service-youku-hacked-accounts-sold-on-darkweb/

[29] Cubrilovic N. Rockyou hack: From bad to worse [EB/OL]. (2009-12-14) [2022-01-20]. https://techcrunch.com/2009/12/14/rockyou-hack-security-myspace-facebook-passwords/

[30] Cox J. Another day, another hack: Tens of millions of Neopets accounts [EB/OL]. (2016-05-06) [2022-01-20]. https://www.vice.com/en/article/ezpvw7/neopets-hack-another-day-another-hack-tens-of-millions-of-neopets-accounts

[31] Amico M D, Filippone M. Monte Carlo strength evaluation: Fast and reliable password checking [C] //Proc of the 22nd CCS. New York: ACM, 2015: 158-169

Han Weili, born in 1975. PhD, professor. Distinguished member of CCF, member of the ACM. His main research interests include data security, access control, and security of artificial intelligence systems.

韩伟力, 1975 年生. 博士, 教授. CCF 杰出会员, ACM 会员. 主要研究方向为数据安全、访问控制和人工智能系统安全.

Zhang Junjie, born in 1997. Master candidate. His main research interests include password security and system security.

张俊杰, 1997 年生. 硕士研究生. 主要研究方向为口令安全和系统安全.

Xu Ming, born in 1995. PhD candidate. Her main research interests include password security and system security.

徐　铭, 1995 年生. 博士研究生. 主要研究方向为口令安全和系统安全.

Wang Chuanwang, born in 1997. Master candidate. His main research interests include password security and system security.

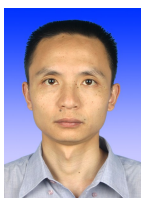
王传旺, 1997 年生. 硕士研究生. 主要研究方向为口令安全和系统安全.

Zhang Haodong, born in 1997. Master candidate. His main research interests include data and information security.

张浩东, 1997 年生. 硕士研究生. 主要研究方向为数据与信息安全.

He Zhenying, born in 1977. PhD, associate professor. His main research interests include DB4AI, database management system and data analytical system.

何震瀛, 1977 年生. 博士, 副教授. 主要研究方向是 DB4AI、数据库管理系统和数据分析系统.



Chen Hu, born in 1974. PhD, associate professor. His main research interests include high performance computing and information security.

陈　虎, 1974 年生. 博士, 副教授. 主要研究方向是高性能计算和信息安全.