计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20160483
53(10):2173-2188,2016

口令安全研究进展

王 平[1,3] 汪 定[1] 黄欣沂[2]
[1](北京大学信息科学技术学院 北京 100871)
[2](福建师范大学数学与计算机科学学院 福州 350117)
[3](北京大学软件与微电子学院 北京 102600)
(wangdingg@pku.edu.cn)

Advances in Password Security

Wang Ping[1,3], Wang Ding[1], and Huang Xinyi[2]
[1](*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)
[2](*School of Mathematics and Computer Science, Fujian Normal University, Fuzhou* 350117)
[3](*School of Software and Microelectronics, Peking University, Beijing* 102600)

Abstract Identity authentication is the first line of defense for information systems, and passwords are the most widely used authentication method. Though there are a number of issues in passwords regarding security and usability, and various alternative authentication methods have also been successively proposed, password-based authentication will remain the dominant method in the foreseeable future due to its simplicity, low cost and easiness to change. Thus, this topic has attracted extensive interests from worldwide researchers, and many important results have been revealed. This work begins with the introduction of users' vulnerable behaviors and details the password characteristics, distribution and reuse rate. Next we summarize the primary cracking algorithms that have appeared in the past 30 years, and classify them into groups in terms of the difference in dependence on what information is exploited by the attacker. Then, we revisit the various statistical-based evaluation metrics for measuring the strength of password distributions. Further, we compare the state-of-the-art password strength meters. Finally, we summarize our results and outline some future research trends.

Key words identity authentication; password security; vulnerable behavior; guessing attack; strength evaluation

摘 要 身份认证是确保信息系统安全的第一道防线,口令是应用最为广泛的身份认证方法.尽管口令存在众多的安全性和可用性缺陷,大量的新型认证技术陆续被提出,但由于口令具有简单易用、成本低廉、容易更改等特性,在可预见的未来仍将是最主要的认证方法.因此,口令近年来引起了国内外学者的广泛关注,涌现出了一大批关于口令安全性的研究成果.从用户生成口令时的脆弱行为入手,介绍了中英文用户口令的特征、分布和重用程度;总结了近30年来提出的几个主流口令猜测算法,并根据它们所依赖的攻击对象的信息不同进行了分类;然后,回顾了当前广泛使用的基于统计学的口令策略强度评价标准;此外,对比了当前主流的几个口令强度评价器.最后,对当前研究现状进行了总结,并对未来研究方向进行了展望.

收稿日期:2016-06-15;修回日期:2016-09-07
基金项目:国家重点研发计划项目(2016YFB0800603);国家自然科学基金项目(61472016,61472083)
This work was supported by the National Key Research Program of China (2016YFB0800603) and the National Natural Science Foundation of China (61472016,61472083).

关键词 身份认证;口令安全;脆弱行为;猜测攻击;强度评价

中图法分类号 TP391

随着信息化进程的不断推进,人们的日常生活不断网络化,资产不断数字化,身份认证逐渐成为保障用户信息安全的基本手段.基于口令的认证伴随着大型机的问世而诞生,在20世纪60年代起被广泛用于大型机的访问控制[1],避免分时操作系统的时间片被滥用.20世纪90年代互联网进入千家万户以来,互联网服务(如邮件、电子商务、社交网络)蓬勃发展,口令成为互联网世界里保护用户信息安全的最主要手段之一[2].

随着互联网的发展,一方面越来越多的服务需要口令保护,另一方面人类大脑能力有限,只能记忆5~7个口令[3],导致用户不可避免地使用低信息熵的弱口令[4],在多个网站中重用同一口令[5],在纸上记口令[6],带来严重的安全危胁.2004年,比尔·盖茨对外宣告口令将消亡,微软公司将使用多因子认证替代纯口令认证[7].后续一系列学术研究也分别从口令无法抵抗离线猜测攻击[8-9]、口令过期策略(password expiration policy)无法保证更新后的口令的不可预测性[10]等方面论证了口令认证技术的不可持续性.与时同时,大量的替代口令的认证方案不断被提出,比如多因子认证[11]、图形口令[12]、生物认证[13]、行为认证[14]等.相比之下,研究口令的相关工作却较少.

出乎意料的是,时至今日,口令的地位在工业界不仅丝毫没有被撼动,反而在越来越多的信息系统应用中得到加强.这一现象吸引了越来越多的学者关注,开始引起学术界的反思.研究发现,虽然这些替代型方案有的在安全性方面优于口令,有的在可用性方面胜过口令,但几乎都在可部署性(deployability)方面劣于口令,并且各自存在一些固有的缺陷[15].比如,基于硬件(如智能卡、USB key)的认证技术成本高昂,使用不方便;基于生物(如指纹、虹膜)的认证技术不具有可撤销性,且存在隐私泄露问题.因此,学术界逐渐开始形成一个共识[2,16-18]:在可预见的未来,口令仍将是最主要的身份认证方式.

既然口令不可替代,只有深入理解口令的安全性和可用性,人类才能更好地与之共存(living with passwords)[16-17].自2012年以来,口令研究逐渐成为一个热点,涌现出了一大批关于口令的研究成果,本文主要关注安全性方面的研究进展.关于口令的可用性,读者可关注人机交互方面的刊物,它已成为该领域一个重要研究分支[19].需要指出的是,与口令强度无关的攻击(如社会工程学[20]、恶意口令捕获软件[21])也不是本文关注点.

口令安全性研究的难点在于,口令是人生成的,与人的行为直接相关,而每个人的行为因内在或外在环境的不同而千差万别.比如说,同样是注册一个163邮箱帐户,有的人觉得这个帐户不重要,会使用"123456"作口令.有的人后面会经常使用这一邮箱,因此采用精心构造的一个字符串(比如"brysjhhrhl",一句诗的首字母)作口令.众所周知,"123456"是弱口令;但是,如果很多用户也使用诗词的首字母作口令,那么攻击者 $\mathcal{A}$ 很可能了解这一用户行为,进而"brysjhhrhl"也可能是弱口令.至于这2个口令谁更安全,需要具体考察它们针对4种口令猜测攻击(见第2节)的抵抗能力.再比如,给定一个口令"Wang.123",该口令如果是由"Li"姓用户产生,那么该口令可很好抵御定向在线口令猜测攻击(targeted online password guessing attack).如果该口令如果是由"Wang"姓用户产生,显然它是一个弱口令[22],无法抵御定向在线口令猜测攻击;无论对于任何用户,这一口令都无法抵抗离线口令猜测攻击[22-23].在用户脆弱口令行为研究方面,焦点主要集中在用户的倾向性构造模式选择[24-26]、口令重用[27-28]、基于个人信息构造口令[22,29]3个方面.

基于对用户脆弱口令行为的更深入理解,一方面攻击者会不断改进其口令猜测算法,另一方面系统管理员也可以阻止弱口令的使用.近年来一个突出变化是,口令攻击算法逐渐摆脱了依靠"奇思妙想"启发式方法,进入了依赖可靠的概率模型科学化算法的新阶段,如基于概率上下文无关方法(probabilistic context-free grammars, PCFG)[30],基于马尔可夫链(Markov-Chain)的方法[23],基于自然语言处理技术(natural language processing, NLP)的方法[31].与此同时,管理员也可以:1)设计更准确的口令强度评测器(password strength meter, PSM)[28,32-33],以便用户注册、更新口令时对用户提交的口令的强度进行及时反馈;2)设计更安全可用的口令生成策略,比如研究发现策略"口令长度12位以上,包含2类字符"要比策略"必须8位以上,包括字母、数字和特殊字符"更可用、更安全[34].

口令安全研究根据其研究方法大致可分为3个阶段.第1阶段为1999年以前,主要采用启发式方式,没有理论体系,口令安全研究更多是一门艺术,

欧美少数几个研究机构零星地发表一些成果(如文献[35-37]);第2阶段为2000年到2008年,口令理论体系初现端倪,但主基调与微软的"口令替代计划"类似,这一阶段研究大多集中于揭示口令的弱点,表明口令在身份认证领域将无法担当主要角色(如文献[8,38-39]);第3阶段为2009年以来,口令安全理论体系逐渐完善,形成了以Markov[23]、PCFG[30]为代表的概率攻击理论模型,以Zipf原理为基础的口令分布理论模型[25],以α-guesswork为代表的口令分布强度评价理论模型[40],使口令安全研究摆脱了传统的依赖简单统计方法和启发式"奇思妙想",进入了以严密理论体系为支撑的科学轨道.值得一提的是,口令的Zipf分布[25]由我国学者发现.

综上,本文主要从用户脆弱口令行为、口令攻击算法、口令分布强度评价指标和口令强度评价方法4个方面,对国内外最新研究进展进行综述.

1　用户的脆弱口令行为

用户的不安全口令行为是造成口令无法达到理想强度的直接原因[41],因此理解用户的脆弱口令行为成为研究口令安全性的基础前提.一方面用户往往需要管理几十上百个口令帐户[24],并且这一数字在不断增长,此外各个网站的口令设置要求往往差异很大[22,42];另一方面,用户用于处理信息安全事务的精力十分有限且保持稳定[43],并不会随着时间的推移而有较大幅度提高.这一根本矛盾导致了用户的一系列脆弱行为.近期研究表明,现实中用户的口令行为往往是理智的[44],并且只有这样,普通用户才能可持续地管理不断增多的口令帐户[45].

当前,广泛采用的口令脆弱行为挖掘方法既有实证分析(如文献[23-26]),也有用户调查(如文献[27-28]);实证分析的数据既有来自于黑客泄露(如文献[23-25,46]),也有来自于企业合作(如文献[40,47]);用户调查既有小规模的传统调查(如文献[27-28]),也有基于外包服务的新型大规模网络调查(如文献[34]).为更好显示实证分析结果,本文使用了8个知名真实口令集,如表1所示.总的来说,已发现用户的脆弱口令行为主要可以归为以下3类.

Table 1　Basic Information about the Password Datasets Used

表1　本文使用口令集的基本信息

Password Dataset	Service Type	Language	Leaked Time	Total Passwords	Unique Password	Personal Info	Typical Reference
Dodonew	Gaming,Ecommerce	Chinese	2011-12	16 258 891	10 135 260		Ref[23-26]
CSDN	Programmer Forum	Chinese	2011-12	6 428 277	4 037 605		Ref[23-26]
126	Email	Chinese	2011-12	6 392 568	3 778 168		Ref[28]
12306	Train Ticketing	Chinese	2014-12	129 303	117 808	√*	Ref[29]
Rockyou	Social Networks	English	2009-12	32 581 870	14 326 970		Ref[23-24,30]
000webhost	Web Hosting	English	2015-10	15 251 073	10 583 709		Ref[28]
Yahoo	Web Portal	English	2012-07	442 834	342 510		Ref[23-25]
Rootkit	Hacker Forum	English	2011-02	69 419	56 900	√**	Ref[27]

Notes: * The 12306 dataset includes five types of personal information: name, birthday, email, phone number and national identity card number.

** The Rootkit dataset includes four types of personal information: name, birthday, user name and email.

1.1　口令构造的偏好性选择

1.1.1　国民口令

1979年,Morris和Thompson在他们的开创性论文里分析了3 289个真实用户口令,发现86%落入普通字典,33%可以在5 min内搜索出来.后续大量研究(如文献[23-28])表明,除了选择单词作口令,用户常常将单词进行简单变换,以满足网站口令设置策略的要求.

比如"123456a"可以满足"字母+数字"的策略要求.这些最流行的单词及其变换就形成了国民口令,如表2所示.中文国民口令多为纯数字,而英文国民口令多含字母,这体现了语言对口令行为的影响.有趣的是,爱情这一主题在国民口令中占据了重要地位.高达1.01%~10.44%的用户选择最流行的10个口令,这意味着攻击者$\mathcal{A}$只要尝试10个最流行的口令,其成功率就会达到1.01%~10.44%.同时,这也预示着人类生成的口令远不是均匀分布,那到底是什么分布呢?

Table 2 Top-10 Most Popular Passwords of Each Service

表 2 各个网络服务中最流行的 10 个口令

Rank	Dodonew	CSDN	126	12306	Rockyou	000webhost	Yahoo	Rootkit
1	123456	123456789	123456	123456	123456	abc123	123456	123456
2	a123456	12345678	123456789	a123456	12345	123456a	password	password
3	123456789	11111111	111111	5201314	123456789	12qw23we	welcome	rootkit
4	111111	dearbook	password	123456a	password	123abc	ninja	111111
5	**5201314**	00000000	000000	111111	**iloveyou**	a123456	abc123	12345678
6	123123	123123123	123123	**woaini1314**	**princess**	123qwe	123456789	qwerty
7	a321654	1234567890	12345678	123123	1234567	secret666	12345678	123456789
8	12345	88888888	**5201314**	000000	rockyou	YfDbUfNjH10305070	sunshine	123123
9	000000	111111111	18881888	qq123456	12345678	asd123	**princess**	qwertyui
10	123456a	147258369	1234567	1qaz2wsx	abc123	qwerty123	qwerty	12345
Percetage/%	3.28	10.44	3.52	1.28	2.05	0.79	1.01	3.94

1.1.2 Zipf 分布

在 2012 年以前,学术界普遍假设口令满足均匀分布(如文献[48-49]),这有 2 方面的原因:1)缺少大规模真实口令数据,口令具体是什么分布难以实证;2)在均匀分布的假设下,分析问题最为方便简单. 自 2009 年第 1 个千万级口令集 Rockyou 泄露以来,如表 1 所示,数以百计的知名网站被攻陷[50],这为研究口令分布提供了充足原始数据. 关于口令 Zipf 分布的发现经历了一个"否定—肯定"的曲折过程. 因为人类自然语言满足 Zipf 分布[51],很自然的一个想法是,人类生成的口令也可能满足 Zipf 分布. 2012 年,Malone 和 Maher[52] 分析了 3 200 万条 Rockyou 数据和另外 3 个小于 10 万条的数据集,将整个口令集输入 Zipf 模型,发现拟合出来的参数通不过 Kolmogorov-Smirnov(KS)检验. 因此,他们得到结论:口令不服从 Zipf 分布. 同年,Bonneau[40] 使用类似方法分析了 7 000 万条 Yahoo 口令,也否定了口令服从 Zipf 分布的可能性.

2014 年,Wang 等人[25] 根据大数定律,指出那些低频次口令天然无法反映其真实频率,因此只有将那些高频次口令(如出现频次不小于 4 的口令)输入 Zipf 模型才有意义. 基于这一新方法,Wang 等发现在 10 万抽样样本下,通过 Zipf 模型拟合的参数可通过 KS 检验. 这意味着,Zipf 模型能够很好地刻画口令分布(双对数坐标下为直线,如图 1 所示):

$$f_r=\frac{C}{r^s},$$

其中,r 表示排名,f_r 表示排名为 r 的口令的频率,C 和 s 为常数,由具体的分布(即数据集)决定.

当前,这一发现已被广泛应用于多个场合,如精确刻画可证明安全协议中攻击者优势[53-54]、评估基因保护系统的抗攻击能力[55]、评估口令 Hash 函数的强健性[56]. 同时,这一规律表明,口令频次呈多项式下降,高频的口令和低频的口令都会占据整个口令集的重要部分. 这也从根本上说明了为什么漫步猜测攻击(见 2.1 节)会如此有效.

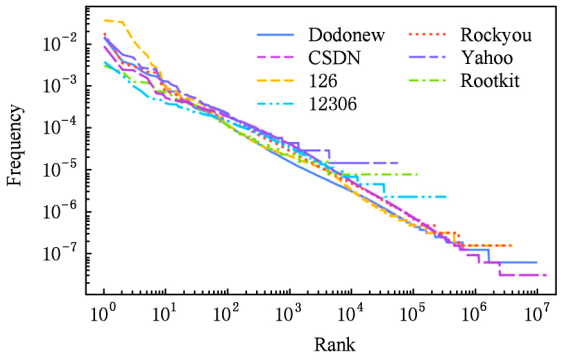


Fig. 1 Human-chosen passwords follow Zipf's law.

图 1 人类生成的口令服从 Zipf 分布

1.1.3 字符组成结构

当网站设置了口令生成策略时,口令的字符组成很大程度上由口令策略所决定. 当网站未设置口令构成策略时,用户口令的结构直接体现了用户的偏好[22-24].

表 3 中最突出的现象是,绝大多数中文口令包含数字,并且 27%~45%仅由数字构成;英文口令喜欢包含字母,低于 16%的口令仅由数字构成,有相当一部分由一串小写字母后面跟 1 组成. 由于高达 99.57%的 000webhost 口令由字母和数字共同

构成,这意味着该网站在运行不久后就执行了“字母+数字”的口令策略.在这种情况下,用户也显示了偏好:54.42%的000webhost口令符合“一串小写字母+一串数字”的结构.用户的这些偏好正是攻击者$\mathcal{A}$所努力挖掘的对象.

1.1.4　口令长度

用户口令长度也直接受网站策略影响.当网站未设置长度限制时,口令的长度分布由受网站服务类型(重要程度)的影响.比如,000webhost提供建站服务,口令具有管理员权限,该网站34.7%的口令长度不低于11,这一比例是其他任意网站的2倍以上.表4显示,对于普通网站来说,90%以上口令的长度介于6-11之间,这一信息对攻击者$\mathcal{A}$减少猜测空间具有重要作用.

Table 3　Composition Structures of Passwords Chosen by Chinese and English Users

表3　中英文用户口令的字符组成结构　　%

Datasets	^[a-z]+$	[a-z]	^[A-Za-z]+$	[a-zA-Z]	^[0-9]+$	[0-9]	^[a-zA-Z0-9]+$	^[a-z]+[0-9]+$	^[a-zA-Z]+[0-9]+$	^[a-z]+1$
Dodonew	10.30	66.32	10.92	69.05	**30.76**	**88.52**	98.33	43.50	45.74	1.40
CSDN	11.64	51.39	12.35	54.33	**45.01**	**87.10**	96.31	26.14	28.45	0.24
126	32.66	66.63	34.86	68.87	**30.66**	**63.24**	95.92	21.99	23.15	2.35
12306	5.26	72.52	5.42	72.94	**27.03**	**94.56**	99.87	50.85	51.50	0.93
Rockyou	41.71	**80.58**	44.07	83.89	15.94	54.04	96.25	27.70	30.18	**4.55**
000webhost	0.04	**98.04**	0.26	99.57	0.02	**98.41**	93.08	**54.42**	60.95	**4.66**
Yahoo	33.09	**92.83**	34.64	94.06	5.89	64.74	97.15	38.27	41.85	**4.80**
Rootkit	41.60	**84.64**	43.84	85.84	13.88	53.97	93.90	19.19	21.55	1.81

Note: The first line of the table are presented in regular expressions. For instance, ^[a-z]+$ stands for the passwords that are only composed of lower-case letters, [a-z] stands for the passwords that include lower-case letters, and ^[a-z]+[0-9]+$ stands for the passwords that are beginning with lower-case letters and ending up with digits.

Table 4　Length Distribution of Passwords Chosen by Chinese and English Users

表4　中英文用户口令的长度分布　　%

Datasets	1~5	6	7	8	9	10	11	12	13	14	≥15
Dodonew	2.46	12.31	15.87	20.86	22.89	16.37	5.21	1.76	0.89	0.56	0.83
CSDN	**0.63**	**1.29**	**0.26**	36.38	24.15	14.48	9.78	5.75	2.61	2.41	2.26
126	**0.00**	26.16	19.33	22.67	11.26	8.17	4.60	1.76	0.90	0.68	0.12
12306	3.58	11.21	15.08	26.32	23.35	18.13	3.43	1.51	0.55	0.31	0.88
Rockyou	1.93	26.05	19.29	19.98	12.12	9.06	3.57	2.10	1.32	0.86	0.47
000webhost	**0.02**	5.70	7.92	21.81	15.41	14.51	**10.49**	**7.67**	**4.14**	**3.14**	**9.20**
Rahoo	6.39	17.98	14.82	26.90	14.90	12.37	4.79	4.91	0.60	0.34	0.80
Rootkit	**0.00**	24.37	16.84	25.80	11.01	7.39	3.50	2.25	1.02	0.62	0.00

1.2　口令重用

一直以来,用户的口令重用行为被认为是不安全的,所以用户应当避免重用[57-59].但是,近期研究发现,面对如此多的需要管理的帐号,重用口令是用户理智的做法[44,60],关键在于如何重用[45].只有跨不同安全级(或重要程度)帐户重用口令,才是应努力避免的[61].表5中总结了用户口令直接重用和间接(修改后)重用的一些数据.

Table 5　Statistics About Password Reuse

表5　口令重用统计数据

References	Year	Research Method	Direct Reuse/%	In-direct Reuse/%
Ref[27]	2014	User survey	51	26
Ref[27]	2014	Empirical	43	30
Ref[46]	2014	Empirical	21	26
Ref[28]	2016	User survey	45	33
Our statistics	2016	Empirical	34	31

2014年,Das等人[27]进一步研究了口令间接重用时新口令和原口令间的相似度,并使用了一系列文本相似度算法(如最大共同长度LCS、文氏距离Levenshtien、曼哈顿距离Manhattan、重合度Overlap)进行测量.结果表明,只有约30%的用户重用口令

时简单修改(即新旧口令相似度在[0.8,1]),绝大多数用户的新旧口令相似度小于0.8,表明修改幅度较大.图2基于4种相似度算法,对12306与CSDN间接口令重用相似度进行了测量,结果与Das等人[27]的发现基本一致.

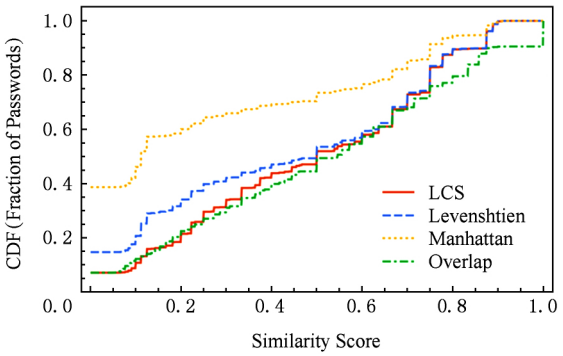


Fig. 2 Similarity scores of PW reuse between 12306 and CSDN.

图2 12306与CSDN间接口令重用相似度

为更好理解不同网站间用户口令间接重用的程度,图3基于Levenshtien相似度算法,对表1中口令集进行了测量.结果表明,中文用户口令重用的问题要更严重:约有40%以上间接重用的中文口令相似度在[0.7,1],而英文口令仅有20%.

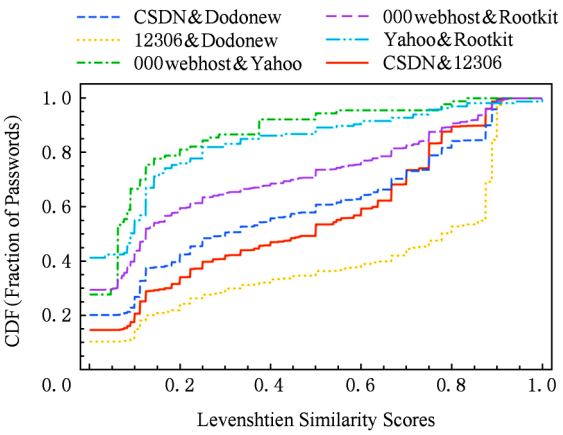


Fig. 3 Levenshtien similarity scores of indirect password reuse.

图3 不同网站口令间接重用的文氏相似度

1.3 基于个人信息构造口令

早在1979年,Morris和Thompson[35]就发现用户构造口令时喜欢使用姓名等个人信息,在猜测字典中加入姓名库可以显著提高口令猜测成功率.除了姓名、生日、用户名、Email前缀、身份证号、电话号码[24,26,35],甚至地名这些个人相关信息都可能被用户使用[62].表6给出了6类常用个人信息在中英文口令中的含有率.其中,生日的含有率最高,其次为用户名、姓名、Email前缀,也有少量用户使用身份证号和手机号作口令.Wang等人[24]发现,在4.36%含有长为11位数字串的中文用户口令中,66.74%包含手机号.

Table 6 Personal Information Usages in 12306

表6 12306网站口令中个人信息使用频率 %

Types of Personal Info	12306[29]	Rootkit
Name	22.35	3.12
Birthdate	24.10	1.19
Account Name	23.60	1.59
Email Prefix	12.66	0.77
ID Number	3.00	
Phone Number	2.73	

表7给出了4种不同类型姓名在口令中的构成比例.其中"Abbr. full name"表示"名的缩写+姓氏",如"wang ping"的缩写为"pwang".可以看出,相当比例的用户使用姓氏或"名的缩写+姓氏"作为口令的构成部分.需要注意的是,英文网站Rootkit用户使用个人信息相较中文网站12306较少,但这并不能得到"中文用户更倾向于在口令中使用个人信息"的结论,这是因为Rootkit网站是黑客论坛,其中的用户具有比普通用户更高的安全意识和更丰富的安全技术知识.总的来说,正如第2节中所证实的,用户使用个人信息构造口令的习惯严重降低了口令强度,定向攻击者可依此大大增强其效率.

Table 7 Various Name Usages in Passwords

表7 各类姓名的使用频率 %

Types of Name Usages	12306	Rootkit
Full Name	4.68	1.38
Family Name	11.15	2.28
Given Name	6.49	0.49
Abbreviate Full Name	13.64	0.15

2 口令猜测攻击

一般来说,口令的安全性可分为2类[23-24]:1)整个口令集的安全性(即口令分布的安全性);2)单个口令的安全性.第1类安全性的评价既可以采用攻击算法进行实际攻击,也可以采用基于统计学的评价指标;第2类安全性的评价只能采用攻击算法进行实际攻击,然后根据攻击结果来衡量,当前广泛采用的衡

量指标是成功攻击该口令所需要的猜测次数[63-64].本节关注攻击算法,基于统计学的评价指标见第3节.如表8所示,根据攻击过程中是否利用用户个人信息,口令猜测算法可分为漫步攻击和定向攻击;依据攻击是否需要与服务器交互,可分为在线攻击和离线攻击.基于PCFG的算法[30]和Markov算法[23]是当前主流的2个漫步攻击算法,也是其他算法的基础,因此本节详细介绍这2个算法,其他算法简要概述.

Table 8　A Taxonomy of Password (PW) Guessing Attacks

表8　口令猜测攻击的分类

Types of Guessing		Exploit Per-Sonal Info	Interact with Server	Necessities for Attacks	Main Counter-Measures	Guess # Allowed	Main Constraints	Typical Reference
Trawling	Online		√	Connection to network	Detection, Lockout	$\leqslant 10^4$	Guess number allowed	Ref[27,80]
	Offline			With Hashed PW dataset	Iterished Hash, salt	$>10^9$	Computational power	Ref[23, 30-31]
Targeted	Online	√	√	Connection to network	Detection, Lockout	$\leqslant 10^4$	Guess number allowed	Ref[22,29]
	Offline	√		With Hashed PW dataset	Iterished Hash, salt	$>10^9$	Computational power	Ref[29]

2.1　漫步攻击算法

所谓漫步攻击(trawling attacking)是指攻击者$\mathcal{A}$不关心具体的攻击对象是谁,其唯一目标是在允许的猜测次数下,猜测出越多的口令越好.这意味着,最优的攻击者会首先猜测排名$r=1$的口令,接着猜测排名$r=2$的口令,依次类推.

2.1.1　启发式算法

早期的口令猜测算法基本都是漫步攻击算法,且没有严密的理论体系,很大程度上依靠零散的"奇思妙想".比如,构造独特的猜测字典[35,37],采用"精心设计"的猜测顺序[65],基于开源软件(如John the Ripper)[65-66].如Bonneau[40]所指出的,这些启发式算法难以重现,难以相互公平对比.因此,这里我们仅概述它们的攻击效果.如图4所示,绝大多数(如文献[35,65])启发式算法的攻破率在30%以下,猜测字典大小为$2^{12}\sim 2^{20}$.虽然少数算法(如文献[35,68])能够达到60%以上,但其测试集都较小,都小于10万,并且往往仅在一个测试集上进行评估.

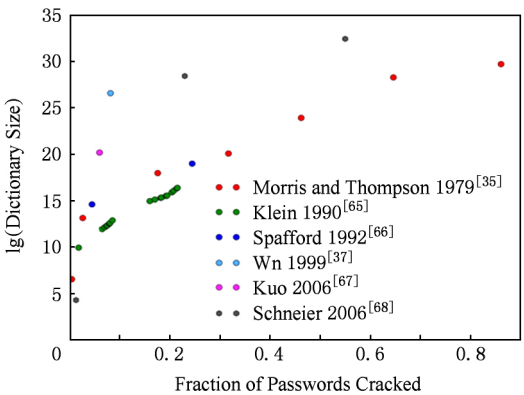


Fig. 4　Effectiveness of various heuristic guessing algorithms[40].

图4　基于启发式方法的攻击效果[40]

2.1.2　PCFG

2009年,Weir等人[30]提出了第1个完全自动化的、建立在严密的概率上下文无关文法(PCFG)基础之上的漫步口令猜测算法.该算法的核心假设是口令的字母段L、数字段D和特殊字符段S是相互独立的.它首先将口令根据前述3种字符类型进行切分,比如"wang123!"被切分为L_4:wang,D_3:123和S_1:!,$L_4D_3S_1$被称为该口令的结构(模式).该算法主要分为训练和猜测集生成2个阶段.

在训练阶段,最关键的是统计出口令模式频率表Σ_1和字符组件(语义)频率表Σ_2.基于CFG方法,对泄漏口令进行统计,得到各种模式的频率和模式中数字组件、特殊字符组件的频率,获得Σ_1和表Σ_2.比如针对$L_4D_3S_1$,统计在全部口令中以$L_4D_3S_1$为模式的口令频率,以及"wang"在长为4的字母段的频率,"123"在长为3的数字串中的频率和"!"在长为1的特殊字符串中的频率.整个过程如图5所示:

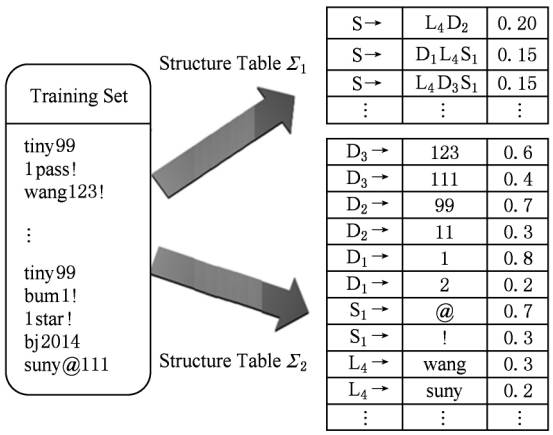


Fig. 5　Illustration of the training phase of PCFG[30].

图5　PCFG算法[30]的训练过程

在猜测集生成阶段,依据上面获得的模式频率表 Σ_1 和语义频率表 Σ_2,生成一个带频率猜测的集合,以模拟现实中口令的概率分布. 比如,猜测"wang123 !"的概率计算为 $P(\text{wang123 !})=P(S\rightarrow L_4D_3S_1)\times P(L_4\rightarrow \text{wang})\times P(D_3\rightarrow 123)\times P(S_1\rightarrow !)=0.15\times 0.3\times 0.6\times 0.3=0.0081$. 这表明,"wang123 !"的可猜测度为 0.0081. 整个过程如图 6 所示:

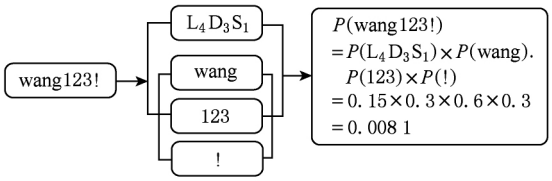


Fig. 6 Illustration of the guess generation phase of PCFG[30].

图 6 PCFG 算法[30]的猜测集生成过程

这样,就能获得每个字符串(猜测)的概率,按照概率递减排序即可获得一个猜测集.

2.1.3 Markov

2005 年, Narayanan 和 Shmatikov[38] 首次将 Markov 链技术引入到口令猜测中来. 该算法的核心假设是:用户构造口令从前向后依次进行. 它不像 PCFG 那样对口令进行分割,而是对整个口令进行训练,通过从左到右的字符之间的联系来计算口令的概率. 类似 PCFG 模型,Markov 模型也分为训练和猜测集生成 2 个阶段.

在训练阶段,统计口令中每个子串后面跟的一个字符的频数. Markov 模型有阶的概念,n 阶 Markov 模型就需要记录长度为 n 的字符串后面跟的一个字母频数. 例如,在 4 阶 Markov 中,口令 abc123 需要记录的有:开头是 a 的频数,a 后面是 b 的频数,ab 后面是 c 的频数,abc 后面是 1 的频数,abc1 后面是 2 的频数,bc12 后面是 3 的频数. 这样,每个字符串在训练之后都能得到一个概率,即从左到右,将长度为 n 的子串在训练结果中进行查询,将所有的概率相乘得到该字符串的概率. 在 4 阶 Markov 模型下,口令 abc123 的概率计算如下:

$$P(\text{abc123})=P(\text{a})\times P(\text{b}|\text{a})\times P(\text{c}|\text{ab})\times P(1|\text{abc})\times P(2|\text{abc1})\times P(3|\text{bc12}),$$

其中,$P(3|\text{bc12})=\dfrac{\text{bc12 后是 3 的频数}}{\text{bc12 后有字符的频数}}$. 其他概率部分也以相同的方式进行计算. 这样就能获得每个字符串的概率,按照概率递减排序即可获得一个猜测集.

在 2014 年,Ma 等人[23]首次将平滑技术和正规化技术运用到 Markov 模型中. 平滑技术是为了消除数据集中过拟合(overfitting)问题,主要有 Laplace 平滑和 Simple Good-Turing 平滑 2 种;正规化技术是为了使得攻击算法所生成的猜测的概率总和始终为 1,形成一个概率模型,主要有 End-Symbol 正规化和长度分布正规化 2 种. 下面以 Laplace 平滑技术和 End-Symbol 正规化技术为例说明.

Laplace 平滑是在训练完毕之后,对于每个字符串的频数都加 0.01,再去计算字符串的概率. 例如,对于字符串 abc123,其中的概率计算如下:

$$P(3|\text{bc12})=\frac{\text{bc12 后是 3 的频数}+0.01}{\text{bc12 后有字符的频数}+0.01\times\text{字符总数}},$$

其中字符总数一般是在 95 个可打印 ASCII 字符 0x20~0x7E 中增加一个结尾符号,共 96 个字符.

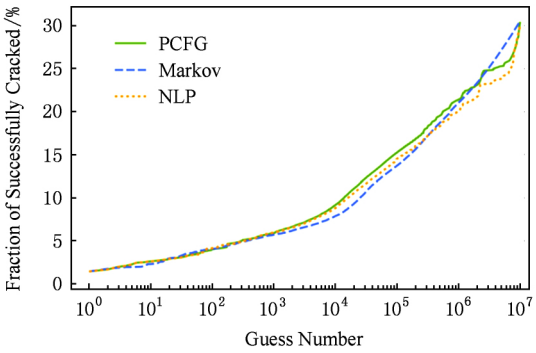
End-Symbol 正规化在每个口令的最后加了一个结尾符号(假设记为⊥),把⊥当作一个字符,训练时将⊥一起加入频数统计,除了口令开头不能出现⊥以外,在口令其他地方都当作正常字符. 生成猜测集时,只有以⊥结尾的符串才能够作为猜测,其他字符串都不能作为猜测输出.

2.1.4 NLP

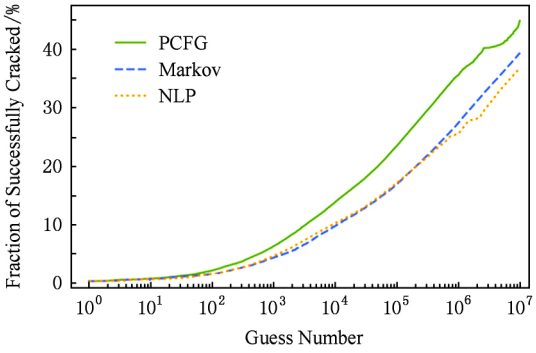
2014 年,Veras 等人[69]指出口令中包含大量深层次语义信息,比如一个口令以"ilove…"起始,它后面接男女姓名的可能性远大于"123"或"asd". 但是,当前的 PCFG 算法和和 Markov 算法都没有利用这些深层次语义信息. 于是,Veras 等人在 PCFG 算法将口令进行 L,D,S 分段的框架内,进一步对 L 段进行语义挖掘,提出了融合语义的 NLP 算法. 该算法 2 个核心点:1)分词(segmentation),因为口令的词段间没有明确的分割符;2)词性标注(part-of-speech tagging 或 POS tagging),即对词语标注一个合适的词性,也就是确定这个词是名词、动词、形容词或其他词性. 为了有效分词和词性标注,Veras 等人收集整理了一个英语语料库集合,主体为当代美国英语语料库(COCA),另加英文人名、城市地点名等语料库. 基于这些该语料库,他们使用自然语言处理的方法对口令进行切词、词性标注,并在此基础上对英文用户口令所具有的语义含义进行分析、抽象和实例化. NLP 算法的猜测集生成阶段与 PCFG 算法相同.

根据图 7 和文献[23-24]的结果,PCFG 算法在小猜测次数下(即在线猜测攻击)最优,Markov 算法在大猜测次数下(即离线猜测攻击)开始显示优势,NLP 攻击效果介于 PCFG 和 Markov 之间. NLP

算法为口令猜测提供了一个新的思路,尚有继续改进的空间.比如利用隐语义模型(latent Dirichlet allocation, LDA)[70]来更好地控制语义标注和抽象化的粒度.



(a) Training on 126, testing on Dodonew



(b) Training on Rockyou, testing on Yahoo

Fig. 7　Comparison of trawling guessing algorithms.

图 7　漫步猜测攻击算法对比

2.2　定向攻击算法

定向口令猜测攻击的目标是尽可能以最快速度猜测出所给定用户在给定服务(如网站、个人电脑)的口令[22,29].因此,攻击者会利用与攻击对象相关的个人信息(personal information, PI),以增强猜测的针对性.用户的个人信息有很多种,比如人口学相关信息(姓名、生日、年龄、职业、学历、性别等),用户在该网站的过期口令(旧口令),用户在其他网站(泄露)的口令.当前定向口令猜测研究尚在起步阶段,主要集中在如何利用人口学相关信息方面.

2.2.1　Targeted-Markov

2016 年,Wang 等人[71]首次提出了基于 Markov 链的定向攻击猜测算法.该算法基于漫步 Markov 攻击模型[23],其基本思想是:人群中有多少比例使用某种 PI,那么攻击对象也将有同样可能性(比例)使用该 PI.为实现这一思想,文献[71]首先将 PI 分为 6 大类(即用户名 A、邮箱前缀 E、姓名 N、生日 B、手机号 P 和身份证 G),并且对每一大类根据需要的粒度进一步细分.比如,N 可分为 N_1(姓名全称),N_2(姓氏),N_3(名),…,N_7(首字母大写的姓氏)共 7 小类.接着,假设集合 $\{N_1, N_2, N_3 \cdots, N_7; B_1, B_2, \cdots; \cdots; G_1, G_2 \cdots\}$ 中共有 k 个元素,将每一个元素视为与 95 个 ASCII 可打印字符同等地位的基本字符,这样 Markov 模型中将有 $95+k$ 个基本字符.然后,将训练集每个口令中所有的 PI 信息替换成对应的 PI 类型基本字符,训练阶段的剩余步骤与漫步 Markov 模型[23]相同.

猜测集生成阶段分 2 步:1)运行漫步 Markov 模型[23]的猜测集生成过程,产生中间猜测集,该猜测集既包含“123456”这样的可以直接使用的猜测,也会包含带 PI 类型基本字符的“中间猜测”(如 N_1, $N_2$123);2)将“中间猜测”里的 PI 基本字符替换为攻击对象的相应 PI 信息,如将 $N_2$123 替换为 wang123(假设攻击对象姓名为“wang ping”)①.

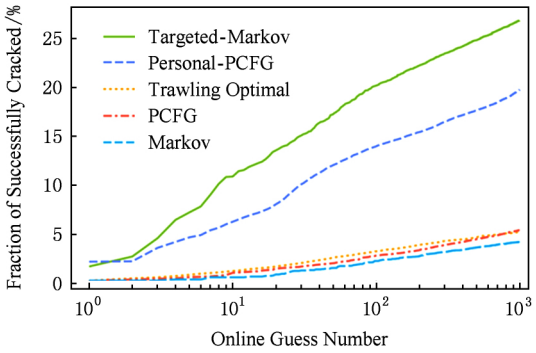
2.2.2　Personal-PCFG

2016 年,Li 等人[29]首次提出了基于概率上下文无关文法(PCFG)的定向攻击猜测算法,称为 Personal-PCFG.该算法基于漫步 PCFG 攻击模型[30],基本思想与 PCFG 攻击模型完全相同:将口令按字符类型按长度进行切分.为实现这一思想,文献[29]首先将 PI 分为 6 大类(即用户名 A、邮箱前缀 E、姓名 N、生日 B、手机号 P 和身份证 G),并将这 6 种 PI 字符类型视为与漫步 PCFG 模型里的 L, D, S 同等地位,这样 Personal-PCFG 中有 9 种类型字符;接着,在训练过程中,将训练集中每个口令如同漫步 PCFG 攻击模型[30]那样,按相应字符类型及其长度进行分段,比如“wang123!”被切分为 $N_4D_3S_1$ 分段,因为“wang”属于姓名 N 这一大类,长度为 4.剩余训练过程与漫步 PCFG 模型[30]类似.

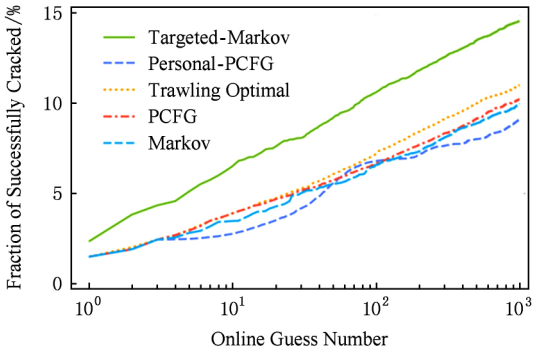
猜测集生成阶段分 2 步:1)运行漫步 PCFG 模型[30]的猜测集生成过程,产生中间猜测集,该猜测集既包含“123456”这样的可以直接使用的猜测,也会包含带 PI 类型字符的“中间猜测”(如 N_1, $N_2$123);2)将“中间猜测”里的 PI 类型字符替换为攻击对象的相应长度的 PI 信息,如将 $N_4$123 替换为 wang123(假设攻击对象姓名为“wang ping”)①.

图 8 对 Targeted-Markov[71]和 Personal-PCFG[29]这 2 个定向口令猜测攻击算法进行了对比.同时,为了更好显示定向攻击算法在在线猜测方面的优势,

① 文献[29]只考虑长度大于等于 3 的 PI 信息.



(a) 50% 12306 training vs 50% 12306 testing



(b) 50% Rootkit training vs 50% Rootkit testing

Fig. 8 Comparison of targeted guessing algorithms.

图 8 定向猜测攻击算法对比

我们将定向攻击算法与漫步攻击算法(Markov[23]、PCFG[30]和 Trawling-optimal[40])也进行了对比. 结果显示,在猜测数为 10～1 000 时,Targeted-Markov[71]的攻击效率比 Personal-PCFG[29]高出 37.18%～73.29%,比理想漫步算法(Trawling-optimal[40])高出 412%～740%.

需要指出的是,除了用户的人口学相关信息(姓名、生日等),用户在其他网站的泄露的口令也可以被攻击者利用来进行定向攻击. 可以预见,这种利用用户口令重用脆弱行为的定向攻击,其危害可能要比基于人口学相关信息的攻击更为严重. 但除了文献[27]给出的一个简单的启发式的攻击方法外,当前尚未见基于严格理论模型的公开成果.

3 口令分布安全性评价指标

2012 年,Bonneau[40]指出,采用攻击算法来评估口令分布的安全性可能带来不确定性,因为攻击算法的效率严重依赖于攻击模型以及模型参数(如训练集、平滑方法、归一化方法)的选择. 相比之下,基于统计学的评价指标则避免了这一难题. 当前,广泛采用的统计学指标主要有 6 类,下面分别介绍,最后利用它们来测量本文所使用的真实口令集.

不失一般性,设分布为 $\mathcal{X}$,样本空间大小为 N,事件概率从大到小依次为 $p_1, p_2, \cdots, p_N$,满足 $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_N$.

3.1 信息熵

Shannon 信息熵(Shannon entropy)[72]被广泛用来测量一个分布的不确定性:

$$H_1(\mathcal{X}) = \sum_{i=1}^{N} -p_i \times \log p_i.$$

3.2 最小熵

最小熵(min entropy)[72]被用来测量一个分布中概率最大事件出现的不确定性:

$$H_\infty(\mathcal{X}) = -\log p_1.$$

3.3 猜测熵

猜测熵(guessing entropy)[73]被用来测量当一个漫步攻击者按最优方式攻击时,猜测出 $\mathcal{X}$ 中任一元素需要的平均猜测次数:

$$G(\mathcal{X}) = \sum_{i=1}^{N} i \times p_i,$$

其等价的比特表现形式为 $\widetilde{G(\mathcal{X})} = \mathrm{lb}(G(\mathcal{X}) - 1)$.

3.4 β-success-rate

β-success-rate[74]被用来测量当一个漫步攻击者被限制至多猜测 β 次时,其平均成功率:

$$\lambda_\beta(\mathcal{X}) = \sum_{i=1}^{N} i \times p_i,$$

其等价的比特表现形式为 $\widetilde{\lambda_\beta(\mathcal{X})} = \mathrm{lb}\left(\frac{\beta}{\lambda_\beta(\mathcal{X})}\right)$.

3.5 α-work-factor

α-work-factor[74]被用来测量当一个漫步攻击者想要达到至少 α 的成功率时,它需要对每个帐户至少发起的猜测次数为:

$$\mu_\alpha(\mathcal{X}) = \min\left\{j \,\middle|\, \sum_{i=1}^{j} p_i \geqslant \alpha\right\},$$

其等价的比特表现形式为 $\widetilde{\mu_\alpha(\mathcal{X})} = \mathrm{lb}\left(\frac{\mu_\alpha(\mathcal{X})}{\lambda_{\mu_\alpha}}\right)$.

3.6 α-guesswork

2012 年,Bonneau[40]指出,虽然对每个帐户发起 $\mu_\alpha(\mathcal{X})$ 次猜测确实能够保证 α 的成功率,但对于有些帐户,攻击者并不需要 $\mu_\alpha(\mathcal{X})$ 次猜测就能攻破. α-work-factor 指标不能刻画这一情形,这是 α-work-factor 的一个固有缺陷. 于是,Bonneau 指出了 α-guesswork[40]这一指标. 它用来测量当一个漫步攻击者想要达到 α 的成功率时,它需要对每个帐户平均发起的猜测次数:

$$G_\alpha(\mathcal{X}) = (1-\lambda_{\mu_\alpha})\times\mu_\alpha + \sum_{i=1}^{\mu_\alpha} p_i \times i,$$

其等价的比特表现形式为

$$\widetilde{G_\alpha(\mathcal{X})} = \mathrm{lb}\left(\frac{2G_\alpha(\mathcal{X})}{\lambda_{\mu_\alpha}} - 1\right) + \mathrm{lb}\left(\frac{1}{2-\lambda_{\mu_\alpha}}\right).$$

3.7　指标的应用示例

上述6个指标中,信息熵和最小熵天然以bit为单位,而其他指标有的是百分比,有的是猜测次数. Bonneau[40]巧妙地使用"有效密钥长度"的思想,将它们都等价转换为以bit单位,以便进行对比. 如表9所示,最小熵、β-success-rate[74]和小成功率的α-guesswork适于用来衡量一个口令分布抵抗在线猜测攻击的能力,而信息熵和大成功率的α-guesswork适于用来衡量一个口令分布抵抗离线猜测攻击的能力.

表9所示为在8个口令分布中,CSDN抵抗在线猜测攻击能力最差,000webhost最优;Rookit抵抗离线猜测攻击能力最差,Dodonew最优. 如第2节所言,在这个8个网站中,000webhost提供的是建站服务,其上口令具有管理员权限,并且执行了最严格的口令生成策略(即"字母+数字"),因此不难理解000webhost抵抗在线猜测攻击能力最优. 出乎意料的是,CSDN执行了最长的口令长度要求(即"口令长度不小于8"),但其抵抗在线猜测攻击能力最差,这可能因其仅是一个技术交流论坛. 这再一次证实:网站服务类型是影响口令强度的第一因素[24],其次为口令生成策略[28].

值得注意的是,本文的6个指标,都只能用来衡量口令分布在漫步攻击者模型下的安全强度,无法衡量口令分布在定向猜测模型下的安全强度. 如何设计在定向猜测模型下的、多维度的口令分布安全指标,是值得研究的一个课题.

Table 9　Evaluation Results of Various Statistical Strength Metrics for Each Password Distribution

表9　真实口令分布安全性的统计学指标评价结果　　b

Datasets	Statistics about Resistance Against Online Guessing					Statistics about Resistance Against Offline Guessing				
	H_∞	$\widetilde{\lambda_{10}}$	$\widetilde{\lambda_{100}}$	$\widetilde{\lambda_{1000}}$	$\widetilde{G_{0.1}}$	$\widetilde{G_{0.2}}$	$\widetilde{G_{0.3}}$	$\widetilde{G_{0.5}}$	H_1	$\widetilde{G_{0.3}}$
Dodonew	6.11	8.25	10.80	13.51	14.42	**17.97**	**19.97**	**21.65**	**21.76**	**22.64**
CSDN	**4.77**	**6.58**	**9.56**	**12.56**	**6.09**	13.94	17.41	20.30	19.47	21.30
126	5.76	8.10	10.68	13.15	12.71	15.68	17.51	19.82	20.16	21.13
12306	8.37	9.61	11.52	13.83	14.75	16.07	16.40	16.61	16.63	16.71
Rockyou	6.81	8.93	11.10	13.11	12.77	14.88	16.77	19.79	21.07	22.65
000webhost	**9.26**	**10.21**	**11.92**	**14.33**	**17.03**	**18.76**	19.52	20.41	20.60	20.79
Yahoo	8.05	9.95	11.76	13.55	13.93	15.72	16.63	17.68	17.88	18.03
Rootkit	6.08	7.99	10.39	12.80	11.52	**13.73**	**14.64**	**15.27**	**15.29**	**15.52**

4　口令强度评测

造成当前弱口令频繁出现的一个直接原因在于普通用户口令安全意识不足,甚至是不正确的[75],不知道如何正确构造(设置)与给定网络服务重要程度相匹配的口令. 造成用户安全意识缺陷的原因有2个:1)口令是私密数据,普通用户往往不知道其他用户的口令是什么样,当选择大众化的口令或口令构造模式时却并不知情,甚至自认为是"独特的","巧妙的"[76];2)绝大多数主流网站的口令强度评测器(password strength meter, PSM)设计是启发式的,没有投入足够的努力("bear no indication of any serious efforts"[77]),向用户反馈的口令强度结果不准确,且常常与其他网站相互冲突,不可避免造成用户的困惑、挫败感和误解[22]. 这2条原因都突出了在用户生成口令时,网站向用户提供及时、准确、一致的口令强度反馈结果的必要性. 并且,近期研究表明,表现形式醒目、反馈结果准确的PSM确实能够显著提高口令的安全性[78-79].

鉴于此,近年来学术界、工业界和标准组织在PSM设计方面投入了大量的努力,提出了一系列PSM. 根据文献[2,80],标准组织界影响力最大的PSM当属NIST entropy[81];文献[77]调研了当前工业界22个主流PSM,表现最突出的为Zxcvbn[82]和KeePSM[83];学术界最先进的3个算法为fuzzyPSM[28]、PCFG-based PSM[32]和Markov-based PSM[33]. 根据底层设计思想的不同,可以将上述这些PSM分为

3 类:基于规则(如文献[80])、基于模式检测(如文献[82-83])和基于攻击算法(如文献[28,32-33]).

4.1 基于规则的口令强度评价方法

影响最大的基于规则的方法当属 NIST PSM[80]. 当前在主流网站上应用的口令强度评价方法绝大多数为基于规则的方法,沿用了 NIST PSM 的思想:口令强度依据长度和所包含的字符类型而定. 典型的代表就是腾讯网站[84]:1)当口令长度 $len<6$ 或 $len\leqslant 8$ 且仅由数字组成时,只进行警告,不输出强度值;2)当口令长度 $len\geqslant 6$,且仅由一种字符组成时,评价为"弱";3)当口令长度 $len\geqslant 8$,包含 2 类字符为"中",包含 3 类或 4 类字符为"强".

这类方法的最大优点是简单,比如腾讯涉及 PSM 的代码只有 12 行;缺陷也是非常明显的:评价结果不准确,容易误判——低估强口令和高估弱口令都会发生.

4.2 基于模式检测的口令强度评价方法

此类方法(如文献[82,84])主要目标是检测口令的各个子段(sub-parts of a password)所属的构造模式(如键盘模式、顺序字符模式、首字母大写等等). 接着,对发现的各个模式赋予相应的分数(score). 然后,将口令的所有模式的分数加和,得到该口令的总得分,即为口令强度值. 比如,Zxcvbn[82]主要考虑了 3 类模式:1)键盘模式,如 qwert,asd,zxcvbn;2)常见语义模式,如日期、姓名;3)顺序字符模式,如 123456,gfedcba. 这些模式被视为弱口令的标志,会给一个较低的分数. 此外,Zxcvbn[82]还考虑了字典模式:将口令的子段与构造的一系列字典(如 top 10000 口令,1 003 个男性姓名,3 814 个女性姓名)进行匹配,根据查找比对的次数进行相应赋值. 如果口令的一个子段不属于上述 4 类模式,则被视为随机字符段,该子段被赋予一个较高的分数.

这类方法比基于规则的方法更科学,但仍属于启发式方法,没有自适应性;一些弱模式(比如字符跳变 Leet:password→p@ssword)如果没有被考虑到,就会漏检,造成弱口令被高估为强口令.

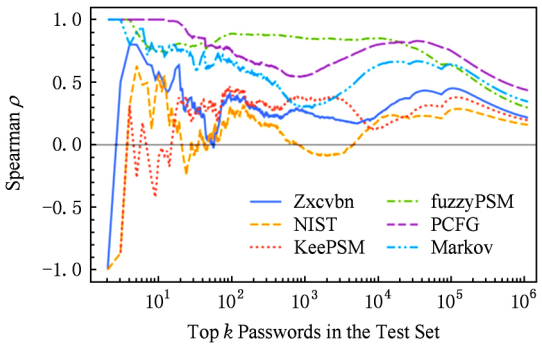
4.3 基于攻击算法的口令强度评价方法

安全永远是相对的. 相对于某种攻击者模型而言,一个口令可能是安全的;相对于另外一种攻击者模型而言,可能是非常脆弱的. 比如,"Wangping.123"可以很好抵抗漫步在线猜测攻击(如 Markov[23],PCFG[30]),但相对定向在线猜测攻击(如 Targeted Markov[71],Personal-PCFG[29])而言,却明显是弱口令.

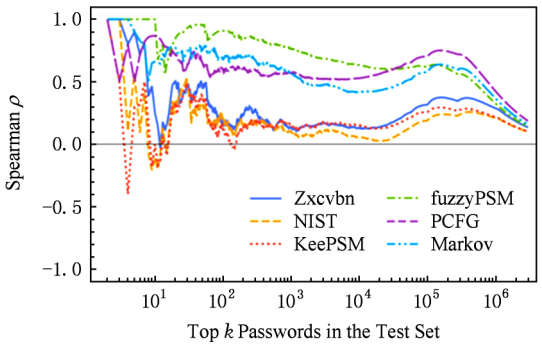
因此,评价口令强弱的一个自然途径就是,使用攻击算法对给定口令进行攻击,依据攻击的难易程度进行强弱判定. 基于这一思想,近期提出了几个基于漫步攻击算法的 PSM. 一个广泛被接受的直观感觉是,攻击算法越高效,基于其上的 PSM 将更准确(见文献[32]). 文献[28]指出事实并非如此:虽然 Markov 算法[23]攻击效率要比 PCFG 算法[30]好,但大规模真实口令集对比实验显示,PCFG-based PSM[32]却比 Markov-based PSM[33]更准确. 一个可能的原因是,Markov 算法[23]使用了平滑技术(如 Laplace 和 Good-Turning),导致其可以破解出更多口令,但同时也带来了数据稀疏性问题.

2016 年,Wang 等人[28]通过用户调查证实,用户在向一个新网站注册口令时,往往是重用(44.8%)一个现有的口令,或者修改(32.6%)一个现有口令,而仅有 14.5%的用户从无到有构造一个全新的口令. 基于这一发现,他们设计了一个基于模糊概率上下文无关文法的口令强度评价算法. 该算法使用一个弱口令集 A(如来自社交网站)作为基础口令集,使用另外一个强口令集 B(如来自电子商务网站)作为修改口令集,然后学习(Learn)用户从口令集 A 取口令然后修改为口令集 B 中相似口令的方法和相应频率,得到一个概率模型 fuzzy PCFG. 在口令评价阶段,基于 fuzzy PCFG 对给定口令计算一个概率值,该概率值作为此口令的强度.

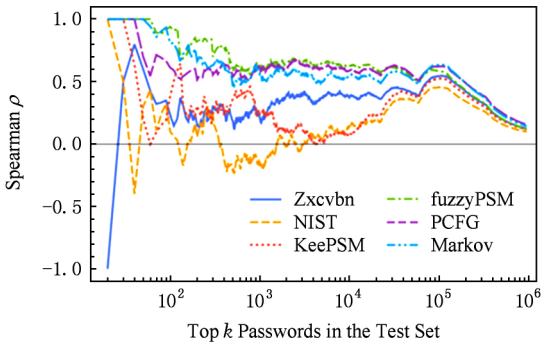
图 9 采用 Spearman 系数对上述 6 个主流 PSM 进行了对比. 对于一个给定的测试口令集,PSM 会对其中每个口令输出一个强度值(概率值);如果将这些口令将其强度值排序,会产生一个口令强度序列. Spearman 系数用以衡量一个 PSM 输出的强度序列与理想 PSM 输出的强度序列间的相关关系:2 个序列完全相同则 Spearman 系数为 1,倒序则 Spearman 系数为−1,Spearman 系数越大表明 2 个序列越接近. 因此,Spearman 系数越大表明其越接近理想 PSM. 图 9 显示,在识别弱口令方面,fuzzyPSM>PCFG-based PSM>Markov-based PSM>Zxcvbn>KeePSM>NIST PSM;在识别强口令方面,PCFG-based PSM>Markov-based PSM>fuzzyPSM>Zxcvbn>KeePSM>NIST PSM. 由于 PSM 最核心功能是防止用户选择弱口令:准确向用户反馈口令的强度,当发现用户选择了弱口令时,进行重点提示或阻止. 因此,总体来说,fuzzyPSM 表现最优,学术界的 PSM 要优于工业界 PSM,标准组织 NIST 的 PSM 的评价结果准确性最低.



(a) 50% CSDN training vs 50% CSDN testing



(b) 50% Dodonew training vs 50% Dodonew testing



(c) 50% Yahoo training vs 50% Yahoo testing

Fig. 9　Comparison of six leading password strength meters.

图 9　当前 6 种主流口令强度评价算法对比

需要指出的是,本文所讨论的 6 个主流 PSM 都基于漫步猜测攻击,未考虑用户个人信息对口令安全性的影响. 尽作者所知,当前尚无基于定向猜测攻击 PSM 的公开研究成果出现. 随着用户越来越多的个人相关信息开始上网,越来越多的网站被攻破、口令被泄露,设计基于定向攻击者模型的 PSM 是一个具有重要现实意义的研究方向.

5　总结与展望

口令由于使用简单、成本低廉、容易修改,克服了基于物理硬件的认证技术和生物特征认证技术的缺陷,学术界逐渐认识到:在可预见的未来,口令仍无可替代. 因此,理解口令的安全性显得十分必要,近年来也涌现出了一大批相关研究成果. 本文系统性地总结了当前国内外在用户脆弱口令行为、口令猜测算法、口令分布强度评价、口令强度评价 4 个领域的研究进展,概述了研究思路和主要研究方法,并指出了存在的不足和值得进一步研究的方向. 当前,关于口令安全的研究整体尚处于起步阶段,国内外的相关研究力量匮乏,在很多重要方向上亟待深入研究,存在大量机遇. 下面探讨口令安全未来的研究需求:

1)口令的 Zipf 分布模型带来的启示

当前,几乎所有的基于口令的安全协议(如文献[85-87])都假设"口令满足均匀". 但事实上,口令是严重偏态的 Zipf 分布,意味着这些安全协议不可避免低估了攻击者的优势,亦即低估了协议存在的安全风险. 因此,在更符合实际的 Zipf 口令分布假设下,如何精确刻画攻击者优势是值得进一步探讨的问题.

2)基于隐语义 LDA 模型的漫步猜测攻击

基于 NLP 的漫步猜测攻击算法效果不佳的一个重要原因是,NLP 方法在语义标注和抽象化的粒度难以调控,而隐语义 LDA 模型[70]正好可以弥补这一不足. 如何将 LDA 模型引入口令猜测,以实现细粒度的语义标注和抽象是一个有前景的研究方向.

3)基于泄露口令的定向猜测攻击

一方面,用户重用口令的趋势越来越严重,另一方面被攻陷的网站越来越多,相当比例的用户口令曾被泄露. 攻击者必然会利用用户曾经泄露的口令来攻击用户当前的口令. 如何使用概率模型准确模拟这一攻击行为?除了泄露的口令,如果再利用个人信息(如姓名、生日),攻击者的定向猜测攻击成功率会有多高?各类个人信息对安全性的影响有何不同?文献[71]首次对这些问题进行了一些探讨,提出了一系列攻击算法,期待更多相关研究.

4)基于定向攻击模型的口令强度评测

当前口令强度评测研究主要集中在,如何基于漫步猜测攻击模型,设计更逼近理想漫步猜测攻击者的 PSM. 随着用户越来越多的个人相关信息被公开,在其他网站使用的口令被泄露,设计基于定向攻击者模型的 PSM 是一个具有重要现实意义的研究方向.

5)服务器端口令集泄露的检测

当前很多网站(如 Myspace, Dropbox, Last. fr[88])都是口令泄露多年后才觉察到,然后通知用户更新口令,往往为时已晚. 如何在服务器端及时检测到口

令集的泄露,是一项亟待解决的课题. 2013 年,Juels 和 Revist 提出了 honeywords 的思想,并给出了几个启发式生成 honeywords 的方法[89]. 作者将"如何对这些 honeywords 的生成方法进行实证评估"遗留为公开问题. 实际上,这一问题的解决有赖于对另一更基本问题的深刻认识:攻击者如何进行最优攻击. 这可能涉及到复杂的统计学模型.

总之,口令安全是一个理论性与实践性都很强的多学科交叉(如密码学、统计学、自然语言处理、机器学习)研究课题,充满机遇与挑战,相信必定会吸引更多的学者的关注和研究.

参 考 文 献

[1] Axel B, Boudhayan C, Lennie D, et al. Security on the IBM Mainframe [EB/OL]. [2014-12-01]. http://ibm.co/1UMpdH7

[2] Bonneau J, Herley C, Van Oorschot P C, et al. Passwords and the evolution of imperfect authentication [J]. Communications of the ACM, 2015, 58(7): 78-87

[3] Keith M, Shao B, Steinbart P. The usability of passphrases for authentication: An empirical field study [J]. International Journal of Human-Computer Studies, 2007, 65(1): 17-28

[4] Yan J, Blackwell A, Anderson R, et al. Password memorability and security: Empirical results [J]. IEEE Security & Privacy, 2004, 2(5): 25-31

[5] Florencio D, Herley C. A large-scale study of web password Habits [C] //Proc of WWW 2007. New York: ACM, 2007: 657-666

[6] Adams A, Sasse M A. Users are not the enemy [J]. Communications of the ACM, 1999, 42(12): 40-46

[7] Munir K. Gates predicts death of the password [EB/OL]. [2004-02-25]. http://www.cnet.com/news/gates-predicts-death-of-the-password/

[8] Clair L, Johansen L, Enck W, et al. Password exhaustion: Predicting the end of password usefulness [G] //LNCS 4332: Proc of ICISS 2006. Berlin: Springer, 2006: 37-55

[9] Brumen B, Taneski V. Moore's curse on textual passwords [C] //Proc of MIPRO 2015. Piscataway, NJ: IEEE, 2015: 1360-1365

[10] Zhang Y, Monrose F, Reiter M. The security of modern password expiration: An algorithmic framework and empirical analysis [C] //Proc of ACM CCS 2010. New York: ACM, 2010: 176-186

[11] Wang D, Wang N, Wang P, et al. Preserving privacy for free: Efficient and provably secure two-factor authentication scheme with user anonymity [J]. Information Sciences, 2015, 321: 162-178

[12] Biddle R, Chiasson S, Van Oorschot P C. Graphical passwords: Learning from the first twelve years [J]. ACM Computing Surveys, 2012, 44(4): No. 19

[13] Yang Y, Lu H, Liu J K, et al. Credential wrapping: From anonymous password authentication to anonymous biometric authentication [C] //Proc of ASIACCS 2016. New York: ACM, 2016: 141-151

[14] Zheng N, Paloski A, Wang H. An efficient user verification system using angle-based mouse movement biometrics [J]. ACM Trans on Information and System Security, 2016, 18(3): 1-27

[15] Bonneau J, Herley C, Oorschot P, et al. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes [C] //Proc of IEEE S&P 2012. Piscataway, NJ: IEEE, 2012: 553-567

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

[17] Lindell Y. A Post-Password World? [EB/OL]. [2016-04-06]. https://www.dyadicsec.com/post-password_world/

[18] Freeman D, Dürmuth M, Biggio B. Who are you? A statistical approach to measuring user authenticity [C] //Proc of NDSS 2016. San Diego, CA: Internet Society, 2016: 1-15

[19] Garfinkel S, Lipford H R. Usable security: History, themes, and challenges [J]. Synthesis Lectures on Information Security, Privacy, and Trust, 2014, 5(2): 1-124

[20] Song Y, Yang C, Gu G. Who is peeping at your passwords at starbucks? to catch an evil twin access point [C] //Proc of IEEE/IFIP DSN 2010. Piscataway, NJ: IEEE, 2010: 323-332

[21] Zhang F, Leach K, Wang H, et al. Trustlogin: Securing password-login on commodity operating systems [C] //Proc of ASIACCS 2015. New York: ACM, 2015: 333-344

[22] Wang D, Wang P. The emperor's new password creation policies [G] //LNCS 9327: Proc of ESORICS 2015. Berlin: Springer, 2015: 456-477

[23] Ma J, Yang W, Luo M, et al. A study of probabilistic password models [C] //Proc of IEEE S&P 2014. Piscataway, NJ: IEEE, 2014: 689-704

[24] Wang D, Cheng H, Wang P, et al. Understanding passwords of Chinese users: Characteristics, security and implications, CACR Report [EB/OL]. ChinaCrypt 2015, [2004-02-25]. http://t.cn/RG8RacH

[25] Wang, D, Jian G, Huang X, et al. Zipf's law in passwords [J/OL]. IEEE Trans on Information Forensics and Security, 2016. [2016-06-15]. http://eprint.iacr.org/2014/631.pdf

[26] Liu Gongshen, Qiu Weidong, Meng Kui, et al. Password vulnerability assessment and recovery based on rules mined from large-scale real data [J]. Chinese Journal of Computers, 39(3): 454-467 (in Chinese)
(刘功申, 邱卫东, 孟魁, 等. 基于真实数据挖掘的口令脆弱性评估及恢复[J]. 计算机学报, 2016, 39(3): 454-467)

[27] Das A, Bonneau J, Caesar M, et al. The tangled web of password reuse [C] //Proc of NDSS 2014. San Diego, CA: Internet Society, 2014: 1-15

[28] Wang D, He D, Cheng H, et al. fuzzyPSM: A new password strength meter using fuzzy probabilistic context-free grammars [C] //Proc of IEEE/IFIP DSN 2016. Piscataway, NJ: IEEE, 2016: 595-606

[29] Li Y, Wang H, Sun K. A study of personal information in human-chosen passwords and its security implications [C] //Proc of INFOCOM 2016. Piscataway, NJ: IEEE, 2016: 1-9

[30] Weir W, Aggarwal S, Medeiros B, et al. Password cracking using probabilistic context-free grammars [C] //Proc of IEEE S&P 2009. Piscataway, NJ: IEEE, 2009: 391-405

[31] Veras R, Collins C, Thorpe J on the semantic patterns of passwords and their security impact [C] //Proc of NDSS 2014. San Diego, CA: Internet Society, 2014: 1-15

[32] Castelluccia C, Dürmuth M, Perito D. Adaptive password strength meters from markov models [C] //Proc of NDSS 2012. San Diego, CA: Internet Society, 2012: 1-15

[33] Houshmand S, Aggarwal S. Building better passwords using probabilistic techniques [C] //Proc of ACSAC 2012. New York: ACM, 2012: 109-118

[34] Shay R, Komanduri S, Durity A L, et al. Designing Password Policies for Strength and Usability [J]. ACM Trans on Information and System Security, 2016, 18(4): No. 13

[35] Morris R, Thompson K. Password security: A case history [J]. Communications of the ACM, 1979, 22(11): 594-597

[36] Riddle B, Miron M, Semo J. Passwords in use in a university timesharing environment [J]. Computer & Security, 1989, 8(7): 569-579

[37] Wu T. A real-world analysis of kerberos password security [C] //Proc of NDSS 1999. San Diego, CA: Internet Society, 1999: 1-10

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

[39] Ives B, Walsh K R, Schneider H. The domino effect of password reuse [J]. Communications of the ACM, 2004, 47(4): 75-78

[40] Bonneau J. The science of guessing: Analyzing an anonymized corpus of 70 million passwords [C] //Proc of IEEE S&P 2012. Piscataway, NJ: IEEE, 2012: 538-552

[41] Adams A, Sasse M A. Users are not the enemy [J]. Communications of the ACM, 1999, 42(12): 40-46

[42] Florêncio D, Herley C. Where do security policies come from? [C] //Proc of SOUPS 2010. New York: ACM, 2010: 1-14

[43] Beautement A, Sasse M A, Wonham M. The compliance budget: Managing security behaviour in organizations [C] //Proc of NSPW 2009. New York: ACM, 2009: 47-58

[44] Stobert E, Biddle R. The password life cycle: User behaviour in managing passwords [C] //Proc of SOUPS 2014. New York: ACM, 2014: 243-255

[45] Florêncio D, Herley C, Van Oorschot P C. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts [C] //Proc of SEC 2014. Berkeley, CA: USENIX Association, 2014: 575-590

[46] Ji S, Yang S, Hu X, et al. Zero-sum password cracking game: A large-scale empirical study on the crackability, correlation, and security of passwords [J]. IEEE Trans on Dependable and Secure Computing, 2016, Doi: 10.1109/TDSC.2015.2481884

[47] Bailey D, Dürmuth M, Paar C. Statistics on password re-use and adaptive strength for financial accounts [C] //Proc of SCN 2014. Berlin: Springer, 2014: 218-235

[48] Halevi S, Krawczyk H. Public-key cryptography and password protocols [J]. ACM Trans on Information System Security, 1999, 2(3), 230-268

[49] Katz J, Ostrovsky R, Yung M. Efficient and secure authenticated key exchange using weak passwords [J]. Journal of the ACM, 2009, 57(1): 1-41

[50] Troy Hunt. Pwned websites [EB/OL]. [2016-06-15]. https://haveibeenpwned.com/PwnedWebsites

[51] Zipf G. Human behavior and the principle of least effort [M]. Reading, MA: Addison-Wesley, 1949

[52] Malone D, Maher K. Investigating the distribution of password choices [C] //Proc of WWW 2012. New York: ACM, 2012: 301-310

[53] Zhang L, Zhang Z, Hu X. UC-secure two-server password-based authentication protocol and its applications [C] //Proc of ASIACCS 2016. New York: ACM, 2016: 153-164

[54] Wang D, Wang P. Two birds with one stone: Two-factor authentication with security beyond conventional bound [J]. IEEE Trans on Depend Security Computer, 2016, Doi: 10.1109/TDSC.2016.2605087

[55] Huang Z, Ayday E, Hubaux J, et al. Genoguard: Protecting genomic data against brute-force attacks [C] //Proc of IEEE S&P 2015. Piscataway, NJ: IEEE, 2015: 447-462

[56] Blocki J, Datta A. CASH: A cost asymmetric secure hash algorithm for optimal password protection [C/OL] //Proc of IEEE CSF 2016. Piscataway, NJ: IEEE, 2016 [2016-06-15]. http://arxiv.org/pdf/1509.00239v1.pdf

[57] McDowell M, Hernan S, Rafail J. Choosing and protecting passwords [EB/OL]. [2013-02-06]. https://www.us-cert.gov/ncas/tips/ST04-002

[58] Jacoby D. False Perceptions of IT Security: Passwords [EB/OL]. [2014-12-16]. https://blog.kaspersky.com/false-perception-of-it-security-passwords/7036/

[59] Singer A, Anderson W, Farrow R. Rethinking password policies [J]. Usenix Login, 2013, 38(4): 14-19

[60] Blocki J, Blum M, Datta A. Naturally rehearsing passwords [C] //Proc of ASIACRYPT 2013. Berlin: Springer, 2013: 361-380

[61] Centre for the Protection of National Infrastructure. Password guidance: Executive summary [EB/OL]. [2015-09-08]. https://www.gov.uk/government/publications/password-policy-simplifyig-your-approach/password-policy-executive-summary

[62] Yampolskiy R V. Analyzing user password selection behavior for reduction of password space [C] //Proc of IEEE CCST 2006. Piscataway, NJ: IEEE, 2006: 109-115

[63] Dell'Amico M, Michiardi P, Roudier Y. Password strength: An empirical analysis [C] //Proc of INFOCOM 2010. Piscataway, NJ: IEEE, 2010: 1-9

[64] Dell'Amico M, Filippone M. Monte carlo strength evaluation: Fast and reliable password checking [C] //Proc of ACM CCS 2015. New York: ACM, 2015: 158-169

[65] Klein D. Foiling the cracker: A survey of, and improvements to, password security [C] //Proc of USENIX SEC 1990. Berkeley, CA: USENIX Association, 1990: 5-14

[66] Spafford E. Observations on reusable password choices [C] //Proc of USENIX SEC 1992. Berkeley, CA: USENIX Association, 1992: 1-16

[67] Kuo C, Romanosky S, Cranor L F. Human selection of mnemonic phrase-based passwords [C] //Proc of SOUPS 2006. New York: ACM, 2006: 67-78

[68] Schneier B. Real-World Passwords [EB/OL]. [2006-12-14]. https://www.schneier.com/blog/archives/2006/12/realworld_passw.html

[69] Veras R, Collins C, Thorpe J. On semantic patterns of passwords and their security impact [C] //Proc of NDSS 2014. San Diego, CA: Internet Society, 2014: 1-16

[70] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[71] Wang D, Zhang Z, Wang P, et al. Targeted online password guessing: An underestimated threat [C] //Proc of ACM CCS 2016. New York: ACM, 2016: 1-13

[72] Cachin C. Entropy measures and unconditional security in cryptography [D]. Zürich: ETH, 1997

[73] Pliam J. On the incomparability of entropy and marginal guesswork in brute-force attacks [G] //LNCS 1977: Proc of INDOCRYPT 2000. Berlin: Springer, 2000: 67-79

[74] Boztas S. Entropies, guessing, and cryptography, 6 [R]. Melbourne: Department of Mathematics, Royal Melbourne Institute of Technology, 1999

[75] Ur B, Noma F, Bees J, et al. I added '!' at the end to make it secure: Observing password creation in the lab [C] //Proc of SOUPS 2015. New York: ACM, 2015: 123-140

[76] Ur B, Bees J, Segreti S, et al. Do users' perceptions of password security match reality [C] //Proc of ACM CHI 2016. New York: ACM, 2016: 3748-3760

[77] Carnavalet X, Mannan M. A large-scale evaluation of high-impact password strength meters [J]. ACM Trans on Information and System Security, 2015, 18(1): 1-32

[78] Ur B, Kelley P, Komanduri S, et al. How does your password measure up? The effect of strength meters on password creation [C] //Proc of SEC 2012. Berkeley, CA: USENIX Association, 2012: 65-80

[79] Egelman S, Sotirakopoulos A, Beznosov K, et al. Does my password go up to eleven? the impact of password meters on password selection [C] //Proc of CHI 2013. New York: ACM, 2013: 2379-2388

[80] Weir M, Aggarwal S, Collins M, et al. Testing metrics for password creation policies by attacking large sets of revealed passwords [C] //Proc of ACM CCS 2010. New York: ACM, 2010: 162-175

[81] Burr W, Dodson D, Perlner R, et al. Electronic authentication guideline, NIST SP800-63-2 [R]. Reston, VA: National Institute of Standards and Technology, 2013

[82] Wheeler D. Zxcvbn: Realistic password strength estimation [EB/OL]. [2012-04-10]. https://blogs.dropbox.com/tech/2012/04/zxcvbn-realistic-password-strength-estimation/

[83] Reichl D. Details on the quality/strength estimations in KeePass [EB/OL]. [2012-04-10]. http://keepass.info/help/kb/pw quality est.html

[84] Tencent. Password strength meter of Tecent [EB/OL]. [2016-06-10]. http://zc.qq.com/chs/v2/

[85] Halevi S, Krawczyk H. Public-key cryptography and password protocols [J]. ACM Trans on Information and System Security, 1999, 2(3): 230-268

[86] Yi X, Rao F, Tari Z, et al. ID2S password-authenticated key exchange protocols [J]. IEEE Trans on Computers, 2016, doi:10.1109/TC.2016.2553031

[87] Jarecki S, Krawczyk H, Shirvanian M, et al. Device-enhanced password protocols with optimal online-offline protection. [C] //Proc of ASIACCS 2016. New York: ACM, 2016: 177-188

[88] Wilson M. 43 million Last.fm account details leaked after 2012 hack [EB/OL]. [2016-09-04]. http://betanews.com/2016/09/04/last-fm-password-leak/

[89] Juels A, Rivest R. Honeywords: Making password-cracking detectable [C] //Proc of ACM CCS 2013. New York: ACM, 2013: 145-160

Wang Ping, born in 1961. PhD, professor in Peking University. His main research interests include system security and distributed computing.

Wang Ding, born in 1985. PhD candidate in the School of Electrical Engineering and Computer Science, Peking University. His main research interests include password cryptography, cryptographic protocols and provable security (wangdingg@pku.edu.cn).

Huang Xinyi, born in 1981. PhD, professor at Fujian Normal University and the Co-Director of Fujian Provincial Key Laboratory of Network Security and Cryptology. His main research interests include applied cryptography and network security (xyhuang81@gmail.com).