一种基于模糊概率上下文自由语法的新型密码强度测量方法

王丁
北京大学
wangdingg@pku.edu.cn

何德彪
武汉大学
hedebiao@163.com

程海波、王平
北京大学
{chenghaibo, pwang}@pku.edu.cn

摘要：为了给用户提供及时的反馈，几乎每一个受人尊敬的互联网服务都在用户注册或更改密码时使用密码强度计（PSM）。在密码研究中，精心设计良好的pms确实有助于提高用户选择的密码的强度，这是一个罕见的好消息。然而，在工业世界中领先的psms(e。g.，Zxcvbn，KeePSM和NIST PSM)主要由简单的启发式规则组成，被发现高度不准确，而学术界最先进的psm(e。g.，基于概率上下文无关的语法和基于马尔可夫的语法)仍然远远不能令人满意，特别是在测量弱密码方面不能。由于防止弱密码是任何PSM的主要目标，这意味着现有的PSM在很大程度上不能达到其目的。

为了填补这一空白，本文提出了一种基于真实用户行为的新PSM。我们的用户调查显示，当为一个新的网络服务选择密码时，大多数用户(77。38%)只需从内存中检索一个现有的密码，然后重用（或稍微修改）它。这与大多数现有psm中看似直观但不现实的假设（通常是隐含的）形成了鲜明的对比，即当用户注册时，一个全新的密码是通过混合字母、数字和/或符号段或组合n个克来构建的。*为了模拟用户的现实行为，我们使用从一个不那么敏感的服务泄露的密码作为我们的基本字典，以及另一个从一个敏感服务泄露的相对较强的密码列表作为我们的训练字典，并确定用户如何使用混乱规则为新服务构建密码。*这个过程自动创建一个模糊概率上下文自由语法（PCFG），并产生我们的基于模糊PCFG的仪表，fumzyPSM。它可以动态响应用户选择密码的变化，并通过与5个有代表性的psm进行比较进行评估。在11个真实密码列表上进行的大量实验表明，一般来说，模糊psm的性能优于所有对应密码，尤其是在区分弱密码方面非常准确，并且适用于在线猜测攻击盛行的服务。

Ⅰ.介绍

由于我们生活中如此多的数字和在线，网络服务对于防止我们的数字资产未经授权的访问是非常重要的。由于文本密码易于理解、易于使用和部署的低成本，自互联网诞生以来，文本密码几乎完全被用于保护对我们的数据的访问，尽管它们有许多已知的弱点[1]-[3]。在过去的几年里，有几十种替代的认证机制(e。g.，图形密码[4]和多因素认证[5])已经被建议，但密码顽固地生存和复制与每一个新的web服务。这种情况很大程度上是因为与密码[6]相比，这些替代方案都有其突出的弱点（es）。更重要的是，由于用户习惯的粘性和模糊的过渡成本[7]，安全性和可用性的增量提高往往不足以达到激活能量替换密码。因此，在可预见的未来，密码很可能会继续主导互联网。

众所周知，普通用户倾向于选择弱用户密码和使他们的帐户处于高风险。为了克服这个问题，几乎所有受人尊敬的web服务现在都执行密码创建策略，要求用户选择的密码满足一些组成规则，通过强制要求一个规则列表，并通过施加密码强度计达到一定的强度阈值(见图中的Gooogle表。1).[1]许多研究已经调查了密码创建策略的影响。1999年，Adams [8]发现，当用户被迫遵守与用户工作实践和组织策略不兼容的密码创建策略时，用户通常会试图规避这些策略。在一项实验室研究中，Proctor等人。[9]测量了各种密码创建策略下密码的创建时间、可记忆性和可裂性，发现更严格的策略使密码更难被破解，但也更难创建和记住。后续研究(e。g.，[10]，[11])还证实，如果政策设计不正确，用户将采取可能损害安全性和生产力的应对策略。

图1.谷歌的密码强度计及其实时反馈。

2012年，Ur等人。[12]透露，只有那些提供关于用户提交的密码强度的准确反馈的密码强度计（psm）才能导致密码强度的显著增加。2015年，他们进一步观察到，很大一部分用户对他们所选择的密码可以提供[13]的真正安全性存在误解。这在很大程度上是由于目前广泛应用在野外的psm的误导性反馈（见[14]，[15]）。除了谷歌的仪表，如图所示。例如，我们可以确认雅虎的PSM衡量密码为“弱”，密码1为“中”，密码123为“强”；万维网联盟（W3C）的PSM宣传自己是“发展开放标准以确保网络的长期增长”，衡量密码为“非常弱（8%）”，密码1为“弱（26%）”，密码1为“足够（46%）”，密码123为“强（67%）”。
这种不准确的反馈会导致普通用户持有a

1https://accounts.谷歌。com/SignUp

9781-4673-8891-7/16 $31.00©2016 IEEE
DOI 10.1109/DSN60.2016.

相信添加一个数字或大写其弱密码的第一个字母会产生安全的密码。

对密码强度的准确测量需要一个适当的度量标准。也许最具影响力的度量标准是在NIST电子认证指南SP-800-63 [16]中提出的基于熵的方法。正如[16]中所承认的那样，NIST PSM本质上是一种基于启发式规则的密码强度的特别估计。g.，"对于一个需要大写和非字母字符的组合规则，分配了6位熵的奖励”，以及“为广泛的字典检查添加了最多6位的熵的奖励”)。大多数在互联网上高调的经前综合症(e。g.，雅虎，Gmail，微软和贝宝)“完美”的[15]抓住了NIST PSM的精神。然而，Weir等。[17]和Kelley等人。[18]说明，这种基于熵的方法只提供了一个非常粗略的密码强度近似，以真实世界的猜测阻力来测量。

最近的研究(e。g.，[12]，[18])表明，衡量密码强度的一个更有效的指标是“可猜测性”，它与现实世界的安全有关，并描述了密码破解算法恢复帐户所需的时间复杂度。这通常是通过测量猜测数(i。e.，搜索空间大小)需要打破该密码[19]。该度量消除了对底层身份验证系统和攻击者的特定性质的依赖性，并且它只与攻击技术和用户密码的复杂性有关。特别需要对给定的数据集进行不同的破解算法，将不同的密码与给定的破解算法进行公平比较。因此，基于猜测性的度量在最新著作中被广泛青睐。g.，[14]，[20]，[21]).

评估密码的“可猜测性”与密码猜测技术的最新进展相吻合。根据是否涉及认证服务器，猜测攻击可以分为：在线猜测和离线猜测（见[22]）；根据是否涉及到用户特定的信息，猜测攻击可以分为：拖网猜测和目标猜测（见[14]）。这项工作主要集中在拖网搜索的猜测，而其他攻击向量(e。g.，目标猜测[14]，键盘记录[23]和肩膀冲浪[24])和他们的对策超出了我们的范围。此后，只要提到“猜测”一词，都是拖网猜测。

拖网式的猜测继续对用户有密码保护的账户构成严重威胁。更具体地说，尽管离线猜测的问题可以通过使用盐和复杂的散列(e。g.，2)，通过使用专用的硬件和软件，攻击者的猜测效率将比预期的高几个数量级（见[25]）。此外，很可能最近泄露了数以万计的纯文本密码(e。g.天涯[27])将进一步提高攻击者的猜测效率，因为它能让攻击者更深入地了解人类选择密码的习惯。例如，这些数据集最近被基于概率上下文无关语法（PCFG）的[28]和基于马尔可夫链的[29]破解算法作为训练集来揭示用户密码行为。类似地，尽管在线猜测的威胁可以通过使用现代的基于机器学习的检测、限速和锁定策略来部分解决。g.，[30]，[31])，攻击者的能力将通过利用密码得到增强泄漏，与以前的蛮力或基于指令的方法相反。当进一步考虑到僵尸网络[32]的流行时，它可以为攻击者提供巨大的计算和带宽资源时，离线和在线的猜测都将更具破坏性。

为了更准确地捕捉攻击者的高级猜测策略，并提供更准确的密码强度估计，卡斯特西亚等人。[33]首次开发了一种基于基于马尔可夫链的概率猜测模型的自适应PSM。他们的基于马尔可夫的PSM可以对用户密码行为的变化做出动态反应，并且被证明比NIST、谷歌和微软的PSM更准确。几乎与此同时，胡什曼德和阿格瓦尔[34]提出了一种自适应的基于PCFG的PSM，如果用户的原始密码的强度低于允许的阈值，它可以建议更好的候选密码。然而，这种基于PCFG的PSM既没有在真实的密码数据集上进行评估以显示其准确性，也没有与其他PSM进行比较以显示其优缺点。

2015年，卡纳瓦莱特和Mannan [15]研究了22个目前部署在流行的网络服务和密码管理器上的psm。他们发现，大多数仪表都是基于特殊设计，不同的仪表给出相同密码高度不一致的结果，许多弱密码被标记为“强”。相比之下，Dropbox（称为Zxcvbn [35]）和Keepass（称为KeePSM [36]）的psm表现最好，这两者也是唯一为它们的设计选择提供一些解释的。

A.动机

令人惊讶的是，据我们所知，学术界没有最先进的pm(i。e.，基于马尔科夫的PSM [33]和基于PCFG的PSM [34])已经被纳入领先的web服务或密码管理器中。这一结论是基于我们对Alexa全球500强网站流量排名前120位的网站进行的调查(http://www。亚历山大。以及在[14]研究约50米和在[15]研究约22米的结果。这在很大程度上可以归因于学术界很少对最先进的psm进行评价或比较(i。e.，基于马尔科夫的PSM [33]和基于PCFG的PSM [34])具有行业中最好的psm(i。e.，Zxcvbn [35]和KeePSM [36])。因此，这些PSM的优点和缺点对安全社区来说仍然是微妙的和未知的，因此PSM不太可能被采用和迁移。

此外，据我们所知，基于马尔科夫的PSM [33]和基于PCFG的PSM [34]从未被公开比较过，关于哪种性能更好仍然是一个悬而未决的问题。卡斯特利西亚等人。[33]推测
马尔可夫模型可以用来创建更好的psm
因为那是“之前的工作(e”。g.，[19])已经证明，基于马尔可夫模型的密码破解技术优于现有的破解技术（如基于PCFG的破解技术）”。然而，正如我们将展示的，事实并非如此，恰恰相反，基于PCFG的PSM通常优于基于马尔可夫的PSM。

这些来自学术界的psm的另一个不令人满意的方面是，它们都隐含地基于一个基本假设，即用户将通过组合单词、数字和/或符号片段为新服务构建新密码(i。e.，基于PCFG的方法[34]，[37])或通过结合马尔可夫n-克(i。e.，基于马尔可夫的方法

[33]).事实上，这在实践中并不是这样的。正如我们对442名参与者的在线调查显示，在注册一项新服务时，大多数用户（77.38%）只是从内存中检索他们现有的密码，然后重用（或稍微修改）它。只有14.48%的用户会从头开始构建一个全新的密码。这两个数字与这次有224名参与者的调查数据非常一致
由Das等人进行。[38].因此，看看psm是否可以建立在这个新的观察结果的基础上是至关重要的。

最后但并非最不重要的是，现有的psm只在少数英语密码数据集上显示出其准确性。更具体地说，基于马尔科夫的PSM [33]在两个数据集上进行了实验，一个来自社交论坛Rockyou [26]，另一个来自程序员论坛Phpbb [39]，而基于PCFG的PSM [34]仅仅通过使用Rockyou数据集显示其可行性（并没有显示其准确性）。*一方面，它更希望的密码来自其他更敏感的服务类型(e。g.，电子商务和约会服务)也可以进行测试。另一方面，由于现有的psm主要被设计用于测量英语密码，所以它们是否可以用于测量非英语密码呢？*在撰写本文时，
互联网[40]上有6.68亿中国网民，占全球互联网总人口的四分之一以上，因此探索现有psm对中国网络密码的适用性是很有趣的。

B.我们的贡献

在这项工作中，我们做出了以下关键贡献：

- **用户调查。**要在选择过程中显示用户的行为
  新服务的密码，我们对442名有效参与者进行了用户调查，这是为此目的进行的最大的一次，也是第一个针对中国网民进行的调查。我们的结果一致
  与之前的调查结果相比，[38]上的英语用户。这项调查揭示了普通用户如何使用混乱的规则来修改新在线账户的现有密码，以及我们应该如何衡量密码强度。
- **一个新的仪表。**为了模拟用户的现实行为，
  我们使用一个从一个较不敏感的服务泄露的相对较弱的密码字典作为我们的基本字典，以及另一个从一个敏感服务泄露的密码字典作为我们的训练字典，并确定用户如何使用混乱规则为新服务构建密码。这个过程自动创建一个模糊概率上下文无关语法（PCFG），并产生我们的基于模糊pcfg的仪表，模糊PSM。它可以动态地对用户选择密码的变化做出反应，与5个具有代表性的psm进行广泛的经验比较表明，fuzyyPSM通常表现最好：它在衡量弱密码方面排名第一，在衡量强密码方面排名第二。
- **这是一个系统的评估。**我们表演了一系列的
  实验和使用两种不同的等级相关指标来调查五种最先进的psm的有效性：两种来自学术界(i。e.，基于马尔科夫的[33]和基于PCFG的[34])，两个来自工业界(i。e.，Zxcvbn [35]和KeePSM [36])和一个来自标准机构(i。e.，尼斯特PSM [16])。我们的评估建立在11个大型的真实密码列表上，
  它包括9743万个密码，涵盖各种流行的互联网服务，是有史以来为PSM评估收集的最大语料库。
- **一些见解。**我们得出了一些见解
  从我们广泛的实验来看，可能会有一些令人惊讶的结果。在大多数情况下，基于PCFG的pm优于基于马尔可夫的pm，这与常见的信念（在[33]中成立）相反。正如预期的那样，在所有情况下，学术psm都优于工业世界的psm。结果表明，如果正确选择训练集，最初为英语用户设计的psm可以用于非英语用户。

路线图。第二节提供了一些初步工作。第三节详细介绍了我们的用户调查。第四节详细阐述了我们的仪表模糊psm。第五节着重介绍了实验方法和评价结果。最后，第六节对论文进行了总结。

II .初步的

我们现在解释了作为我们和现有仪表基础的安全模型，为实际理想的仪表提供了一个正式的定义，并介绍了评估仪表的指标。

A.安全模型

如前所述，psm的目标是量化攻击者在猜测正确的密码之前必须投入多少努力。为此，我们必须首先明确攻击者可以发起哪些类型的攻击，因为不同类型的攻击通常涉及完全不同的资源和猜测策略。和大多数相关的工作一样，[19]-[21]，[29]，[33]，[34]，[37]，在这项工作中，我们主要关注拖网搜索猜测攻击，[2]虽然没有考虑针对性的猜测攻击[14]，因为后一种攻击涉及特定用户的数据，攻击者如何利用这些个人数据仍然是一个悬而未决的问题。此外，其他攻击向量，如键盘记录[23]和肩膀冲浪[24]，其中攻击者通过非密码的方式获取密码，与密码强度无关，因此超出了本工作的范围。

在拖网网猜测攻击中，攻击者并不在乎谁是受害者，她唯一的目的是找出与任何一个账户匹配的密码（见[41]）。因此，她最好的策略是首先反复猜测最有可能的密码(i。e.，以概率递减的顺序来尝试猜测)。拖网猜测攻击可以进一步分为在线拖网猜测和离线拖网猜测。在在线拖网式猜测攻击中，攻击者通过尝试与身份验证服务器交互来测试其猜测的正确性。在离线拖网猜测攻击中，攻击者获得了直接访问散列密码，为了破解它，她不需要与服务器交互，可以使用相同的散列算法离线搜索猜测密码(e。g.，scrypt或PBKDF2)[3]然后离线后与目标进行比较。如果这两个散列匹配，则攻击者已找到原始密码。

[2]尽管在大多数psm(e。g.，[33]-[37])和相关文献（[19]-[21]，[29]）假设了一个拖网式的猜测攻击者，但与这项工作相反，这样的假设通常只是含蓄地做出的。这很可能会导致对安全管理员的误解。

[3]尽管长期以来一直建议网站通过使用专用的哈希函数(e。g.，现实是，64%的泄露数据集是明文的或MD5没有盐[42]，甚至11个排名前500的网站[43]以明文存储密码。

表i。不同猜测攻击的比较

猜测攻击类型		使用个人数据	相互作用与服务器	攻击者的主要约束条件	允许的猜测数	考虑在这项工作中
拖网	在线	不	是	检测、锁定	$< 10^4$	是
	离线	不	不	攻击者的力量	$> 10^9$	是
目标	在线	是	是	检测、锁定	$< 10^4$	不
	离线	是	不	攻击者的力量	$> 10^9$	不

表一显示了在线攻击和离线攻击之间最关键的区别是攻击者可以做出的猜测次数。在在线攻击中，攻击者试图攻击的系统是可运行的，而诸如可疑登录检测[30]和锁定策略[31]等安全机制仍然处于活动状态。因此，通常在线攻击者在被锁定之前只能做出一些猜测(e。g.因此，对她来说，最好的策略是尝试这几个最受欢迎的密码[41]，[44]。另一方面，离线攻击者不受防御者的安全机制的限制，并且允许她进行尽可能多的猜测，而唯一的限制是她自己的时间和计算资源。这意味着一个离线的猜测会话可能包含数万亿的猜测[21]，[45]，因此一个离线的攻击者不仅可以尝试流行的密码，而且还可以尝试她猜测的非常罕见的密码。

总而言之，拖网猜测攻击者的最佳猜测策略总是更早尝试更流行的密码，在猜测正确密码之前需要尝试的猜测数量很好地描述了攻击者必须投资的努力。这个数字可以定义为目标密码的强度。

B.实际理想仪表的正式定义

w为了测量给定密码p的强度，psm通常不直接输出p的猜测数。w相反，他们输出每个猜测的熵(e。g.，NIST [16])或与这个猜测的推导相关的概率(e。g.，基于概率模型的一个模型，[33]，[34])。根据熵或概率的值，对所有的猜测都可以进行排序，从而可以确定地确定p的猜测数。w

口令（文本）密码的正式定义首先由Ma等人给出。在IEEE标准普尔2014年[29]指数中。w密码p是一个被限制在字母表Σ中的字符串，其长度介于L之间mi几和Lma北[29].因此，在身份验证系统中允许的密码集是：

$$\Gamma = \bigcup_{l=L_{min}}^{L_{max}} \Sigma^l \tag{1}$$

通常，字母表Σ可以设置为95个可打印的ASCII字符中的任何一个子集。但是，由于这95个字符通常被分为四组：数字、小写字母、大写字母和特殊符号，因此Σ通常被设置为包含这四组的任何组合，如10位数字和36个带有数字的小写字母。我们调查了前50个站点，发现其中11个站点需要Lmi几=6和Lma北=为20，其中7个需要Lmi几=6和Lma北= 16.请注意，在本工作的破解实验中，我们将Σ设置为包含完整的95个字符。

密码强度计。w密码强度计（简称米）是一个函数M（· ），它将字母Σ上的密码p作为输入，并输出[0,1]中的概率（实数）：

wM： p→[0,1]，其中p∈ Γ (2)w

w密码p越强，M (p)的概率就越低。w请注意，只有当M（· ）是一个基于概率模型的PSM时，M（· ）才满足以下条件：

$$\sum_{pw \in \Gamma} M(p) = 1.$$

w例如，[33]，[34]中的psm是两个概率模型-

而Zxcvbn [35]、KeePSM和NIST PSM [16]则没有。ww每个密码p的强度M (p)构成了对对手为了成功所需投资的工作量的估计。在实践中，仪表的值通常分为几个桶，如“弱、中、强”（见苹果的仪表）和“弱、公平、好、强”（见谷歌的仪表）。请注意，仪表可以是强制性的，也可以只是暗示性的。w前一种仪表要求只接受概率M (p)低于预定义阈值的密码p，而后一种仪表要求只向用户提供有关密码强度的信息。w

理想的仪表。$\mathcal{X}$本质上，概率PSM的最终目标是在目标认证系统中重现密码的概率分布。理论上，理想的米（· ）能达到的是，

$$M(p) = p_{pw} w, \quad A\ p \in \Gamma \tag{3}$$

其中p_{pw}*是密码p的真实概率。*$w\mathcal{X}$换句话说，一个理想的仪表可以精确地再现该分布。$\mathcal{X}$在实践中，几乎不可能确定p的值pw.幸运的是，对于流行的密码(e。g.，与$f_{pw} \geq 4$)，*我们可以使用经验概率*$f_{pw}\mathcal{DS}$/| |到近似p_{pw}其相对标准误差约为$1/\sqrt{f_{pw}}$[41]，其中f_{pw}是密码数据集中p的频率。$w\mathcal{DSX}\mathcal{DSX}\mathcal{DS}$这意味着一个人可以从经验概率中抽取一个大样本并按递减顺序排序，那么列表中流行的部分可以用作保证准确性的理想仪表：列表中的顺序只是猜测数字(i。e.，要测量的密码的强度)。

因此，任何现实世界的PSM (e。g.，[33]，[34])可以看作是这个理想仪表的近似值。然后，我们可以通过测量PSM结果的差异来评估真实世界PSM的好坏： 1)每个密码的概率差异；2)每个密码的猜测数的差异。我们注意到，如果PSM的焦点是其对数估计的猜测数估计的精度，那么第一种差异不仅难以测量，而且可以很好地描述后一种差异（见第II-a节）。作为一个例子，假设两米，两米$_1$（· ）和M$_2$（· ），允许M$_1$（）是一个理想的仪表和M$_2$（）输出如下：

$$M_2(pw_i) = \begin{cases} M_1(pw_1) + (M_1(pw_2) - M_1(pw_3))/2, \\ M_1(pw_2) - (M_1(pw_2) - M_1(pw_3))/2, \\ M_1(pw_i), \quad \text{for all } \ i \geq 3 \text{ and } pw_i \in \Gamma \end{cases} \tag{4}$$

我们可以确认$\sum_{pw_i \in \Gamma} M_2(pw_i) = 1$和那个，如果M$_1$(p$w_1$) ≥ M$_1$(p$w_2$) ≥ M$_1$(p$w_3$)≥· · · ，然后是M$_2$(p$w_1$) ≥ M$_2$(p$w_2$) ≥ M$_2$(p$w_3$) ≥ · · · .这意味着M$_2$（· ）是一个基于概率模型的计量器$_2$（· ）可以为一个给定的密码产生与理想的计量表M相同的猜测数$_1$（· ）。因此，在我们的安全模型下，它们的行为是无法区分的。这就产生了以下放松（但更实用）的理想仪表。

一个实际上理想的仪表。我们说一个电表M（· ）是一个几乎理想的电表，如果它可以输出相同的猜测数为一个给定的密码作为一个理想的电表所做的。在形式上，M（· ）满足这一点，只要ppwi≥ ppwj，

$$M(pw_i) \geq M(pw_j)，适用于所有pw_i, pw_j \in \Gamma \quad (5)$$

其中ppw$w$是从底层分布中抽取的密码p的真实概率。$\mathcal{X}$在猜测数方面，M（· ）与理想的仪表是有区别的。这个定义的显著优点是，它使我们能够专注于猜测数，并使用经过充分研究的指标（见第II-C节）来准确地测量任何真实的仪表的缺陷。

我们注意到岩藻等人。[33]还定义了一个“理想表”，但我们可以确认，实际上，他们的理想表并不是一个理想表，因为根据他们的定义，它不能在目标认证系统中复制密码的概率分布。他们的理想仪表本质上是我们上述实际理想仪表的一些特别的例子。

C.两个秩相关度量

根据上述定义，要测量PSM M（· ）的优度，可以测量它与实际理想的仪表的距离。这可以通过测量每个米产生的排名猜测数（或相等的，概率）之间的相关性来实现。由于psm产生的概率通常是高度偏态的(见图。[29]的2)，应使用非参数等级相关度量。斯皮尔曼系数p [46]和肯德尔T [47]是这两个应用最广泛的指标。p和T都是[-1,1]中的实数，1表示两个排名之间的一致性是完美的，-1是最差的，0是独立的。更多的详细信息请参考完整的版本(请参见http://t。cn/RG8Ewf3)的本工作。

III .用户调查

为了更准确地评估用户选择的密码的强度，我们首先需要更好地了解用户的行为。e.，用户在注册新帐户时如何创建密码。为此，我们使用Sojump。这是中国最受欢迎的在线调查平台。我们的问题的副本可以在http://www上找到。sojump.com/jq/6443561.第三部分与这项工作有关。为了更好地比较中国用户与英语用户的行为（见[38]中的美国定位调查），我们的一些问题是专门设计为尽可能接近[38]的。

我们在发布前优化了调查问题，通过迭代地从我们实验室其他团队的学生那里获得反馈，以尽量减少用户的拒绝和退出（见[11]）。最后，还有一些潜在的冒犯性问题(e。g.，“你有多少个密码？”以及“如果你为这个新的电子邮件帐户重用现有的密码，它可能来自哪个类别？”“)被放弃了，我们的调查大约需要22到35分钟。我们的团队成员每人都需要邀请10到20个可靠的朋友、亲戚或同事来完成调查。我们进一步询问了一些对我们的邀请反应非常积极的参与者(e。g.，他们认为我们的调查对社会有重要价值)来分发调查，他们也被要求通知我们他们发送的邀请的数量。我们总共发送了983份邀请函，我们得到了442份有效的回复。我们首先详细说明调查结果，然后报告受访者的人口统计数据。

我们通过指定一个场景开始调查，要求参与者创建一个将经常使用的电子邮件帐户，并发布了一系列关于他们的密码选择的问题。我们的第一个问题是弄清楚他们新账户的密码和现有账户的密码有多相似。图2显示，77.38%的用户会重用或简单地修改现有的密码。这个值与英语用户的值令人难以想象地一致（77% [38]）。虽然重用现有密码的中国用户比英语用户少6.2%，但创建一个全新密码的英语用户比中国用户多14.86%。1999年，Adams和Sasse [8]发现，在他们的139名问卷受访者中，有50%使用了他们自己的方法，通过转换现有密码来创建令人难忘的多个密码。这可能意味着，随着时间的推移（以及人们必须维护的网络账户数量不断增加），重用密码的程度会变得越来越严重。

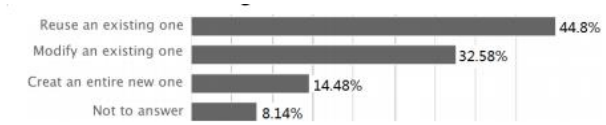


图2.当您被要求为一个将经常使用的新电子邮件帐户创建密码时，您将会怎么做？

如果用户修改了现有的密码，那么修改后的密码与原始密码的相似程度仍然是一个问题。因此，我们提出了一个如图所示的相关问题。3.我们可以看到，“非常相似”和“相同”的答案至少占61.77%，而“相似”至少占20%。这表明，超过80%的用户将提交与现有密码相似的新电子邮件帐户的密码。这种观察是在巨大的对比看似直观但不现实的假设（通常隐式）在大多数现有的psm，当用户注册时，一个全新的密码是由混合段的字母、数字和/或符号（见PCFG-based PSM [34]）或结合n克（见马尔科夫PSM [33]）。这对于衡量密码强度有很大的影响：因为大多数用户的新账户密码将与他们的密码相似

现有的密码，而且攻击者很有可能已经了解了后者(e。*g.，通过不断的密码泄露[26]，[27]，[48]），因此新账户的密码强度主要与修改过程中引入了多少不确定性有关*。这激发了

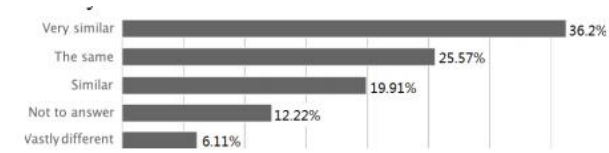


图3.您如何描述新帐户的密码与现有帐户的密码的相似性？

图4帮助我们理解为什么用户会修改（但不重用）一个现有的密码。正如预期的那样，51%的用户致力于提高安全性，这与用户通常认为简单的修改会大大提高他们密码的强度的观察结果相一致——“我补充道‘！’最后使它安全“[13]。相比之下，英语用户的这个数字只有24%的[38]。一个合理的原因可能是，在过去的几年里，高调的中国网络服务比英语服务泄露的用户账户，[27]，[29]，

中国用户更关心当前基于密码的认证的安全性。实现密码策略和提高记忆能力的数字分别为42.76%和32.58%，与[38]中报道的英语用户相当。

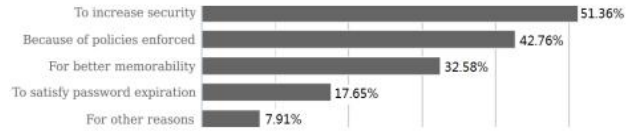


图4. 如果修改新帐户的现有密码，为什么要修改（多项选择）？似乎大多数用户的目标都是提高安全性。

用户使用转换规则对现有密码的修改响应如图所示。5. 显然，连接(e。g.，″在开头/结尾加一个数字/符号”)带头，然后是“大写”和“leet”。g.，a↔@和4↔)。此外，“子弦移动”、“反向”和“添加特定于站点的信息”也有它们的位置。虽然这些数字不能直接与[38]进行比较，因为我们的问题不是单一选择的，但它们在很大程度上与关于英语用户的研究结果一致。

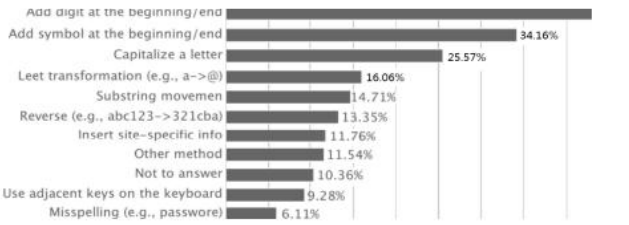


图5. 修改现有密码（多项选择）时使用哪些转换规则？大多数用户更喜欢连接。

在下面的几个问题中，我们将研究上述转换规则在密码中发生的位置。图。6和7表明，用户喜欢将数字/符号放在最后、中间和开始。这大致符合英语用户[38]的情况。至于资本化，图。8显示，47.96%的中国用户倾向于在密码的开头进行大写，而英语用户的这一数字相当相似（44%）。然而，
22.62%
中的中国用户选择不使用转换规则
而英语用户的这个数字只有6%。这些结果与我们在泄露的真实密码数据集中的观察结果相吻合（见表八）。

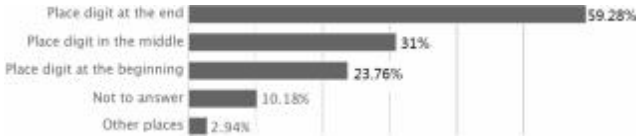


图6. 如果网站要求密码中包含一个数字，你通常会将这个数字放在哪里？有多项选择的

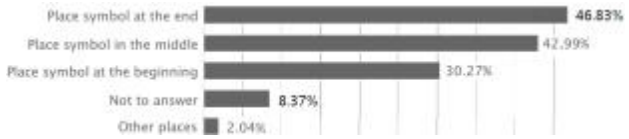


图7. 如果网站要求密码中包含一个符号，您通常会将该符号放在哪里？有多项选择的

我们还获得了一些关于参与者的基本人口统计信息：三分之二是男性；80.55%的人年龄在18岁的～之间，15.67%的人年龄在35岁以上；80.55%是

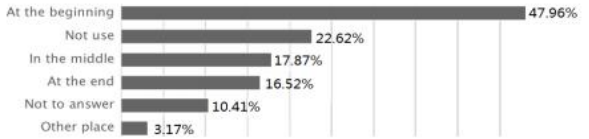


图8. 如果你的新密码涉及到大写，它会发生在哪里（多项选择）？大约有一半的用户将第一个字母大写。

攻读或至少获得学士学位，43.44%正在攻读或至少拥有硕士学位。这是意料之中的，因为我们的团队成员是年轻的博士候选人，因此他们的个人关系主要是年轻人和受过教育的人。更详细的信息请参考[45]。相比之下，我们的参与者比[38]更多样化，其中93%的参与者是18岁，92%的参与者至少拥有学士学位。然而，我们的参与者比普通人群更年轻，受教育程度更高，而且他们可能更有安全意识和技术意识。因此，我们很可能低估了这个问题。

道德和限制。据我们所知，这项调查是第一个针对全球互联网人口最多的中国用户的调查。我们得到了本中心伦理委员会的批准。参与者被匿名调查，结果都被汇总，这样就不能推断出用户可识别的信息。虽然我们的参与者在数量和多样性方面优于早期的相关研究（英语用户[38]，[49]，[50]），但我们的调查受到任何密码相关调查的固有局限性：用户可能提供错误的数据，生态效度难以保证。尽管如此，在许多方面，我们的研究结果都与真实的数据集相一致。和英语用户。g.，密码重用的速度和添加数字/符号的位置)。

增值我们提出的方法模糊psm

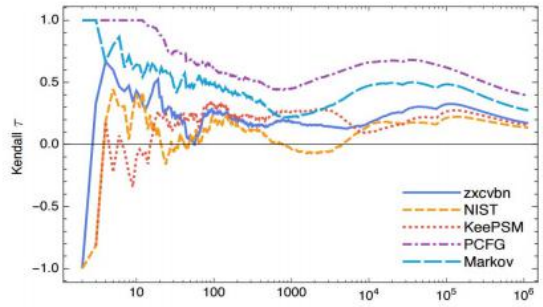
我们现在描述了模糊PSM，一种新的基于模糊PCFG的PSM。为了解释为什么我们的仪表更喜欢类PCFG模型而不是类马尔可夫模型，我们首先对现有的psm进行了简要的比较，并展示了它们的主要局限性。

A. 简要比较一下现有的psm

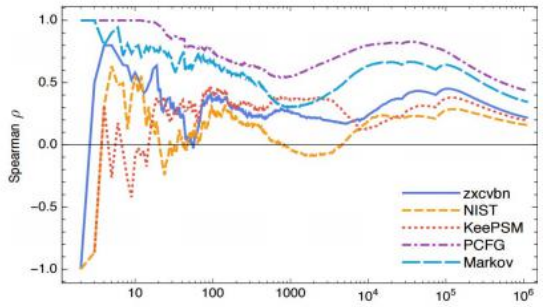
据我们所知，这两种最先进的PSMs（i。e.，基于马尔科夫的PSM [33]和基于PCFG的PSM [34])从未被公开比较过，关于哪一个表现更好仍然是一个悬而未决的问题。卡斯特利西亚等人。[33]推测，马尔可夫模型可以用来创建更好的psm，因为“以前的工作(e。g.，[19])已经证明，基于马尔可夫模型的密码破解技术优于现有的破解技术（如基于PCFG的破解技术）”。事实上，最近的一些研究(e。g.，[21]，[29]，[45])也证实了基于马尔可夫的模型在密码破解方面的优越性。然而，正如我们将在这里展示的(和在秒。V)，这种趋势在衡量密码强度的情况下并不成立，而恰恰相反，基于PCFG的PSM通常优于基于马尔可夫的PSM。

作为一个具体的例子，我们使用这两个基于机器学习的psm来测量从CSDN披露的643万个CSDN密码。这是中国软件开发者的热门社区。一般来说，基于机器学习的仪表的有效性取决于两个因素：算法本身和所使用的训练集。为了排除训练集的影响，如[17]，[33]，我们随机的

将CSDN数据集平均分成四个部分，第1部分用于训练，第4部分用于测试。由于这两个PSM都没有开源代码，我们根据[29]的最新改进，实现了基于PCFG的PSM [34]和基于markov的PSM [33]。更具体地说，对于基于PCFG的PSM，与字母段相关的概率是直接从训练过程中学习到的，而对于基于马尔可夫的PSM，我们使用了回退方法。为了确保我们的实现的正确性，我们使用了这些猜测[4]这两个psm输出在[29]、[45]中重复开裂实验，开裂结果完全一致。我们计划开源这项工作中的所有代码，以促进研究社区。



(a)真实世界的PSM与理想PSM的肯德尔T的比较



(b

图9.在肯德尔和斯皮尔曼系数方面的比较（1/4 CSDN密码用于训练，另外1/4用于测试）。

然后我们采用了两个等级相关度量(i。e.，肯德尔T和斯皮尔曼p)测量由真实世界的PSM输出的猜测数字的距离，由理想的仪表输出(见秒。微光这就产生了无花果。分别为9(a)和9(b)。请注意，北轴上的值k表示图中的数据点是为排名前1、2、· · · 、k的密码集计算的。这两个指标提供了几乎相同的结果，它们的相似性在我们以后的所有实验中都成立。相比之下，肯德尔T度量对微小的差异更为敏感，因此以后我们只提供基于肯德尔T的结果来节省空间。此外，这里我们还提供了行业领先的比较(i。e.，Zxcvbn [35]，Keepass PSM [36]和NIST PSM [16])，它们都是基于规则的，而不是基于机器学习的PSM。 这两个指标都表明，基于PCFG的仪表在现有的psm中表现最好，而来自行业的三个基于规则的psm不如来自学术界的两种基于机器学习的psm。虽然后来的发现在意料之中，但前者有点令人惊讶。*这推翻了人们普遍认为的猜想，即基于马尔可夫的模型“可以用来创建更好的主动密码强度计（比PCFG）”[33]。这是相当非预期的和反直觉的，因为基于马尔可夫的模型在密码破解方面比基于PCFG的模型更好（见[20]，[29]，[45]）。*

B.解释了PCFG PSM的意外优势

为了解释这一看似矛盾的观察结果，我们提供了一个基于PCFG的PSM [34]和基于马尔科夫的PSM [33]之间的差异的微观掌握。更具体地说，我们检查了每个PSM为测试密码输出的猜测数的差异。在无花果。10，每个红色的数据点代表一个来自训练集的密码。其协调(北$_i$,y$_i$)意味着该密码的强度可以承受北$_i$猜测尝试由理想的仪表测量，并承受y$_i$猜测由基于PCFG的仪表测量的尝试尝试。类似地，每个蓝色的数据点代表一个来自同一训练集的密码及其坐标(北$_i$,y$_i$)表示理想仪表和马尔可夫计给出的猜测数。请注意，北轴上的值k表示图中的数据点是为排名前1、2、· · · 、k的密码集计算的。不难看出，一个仪表越好，数据点越接近，就会形成绿线。这表明，基于PCFG的电表的性能优于基于马尔可夫的电表，这在测量弱（顶部）密码时尤其明显。

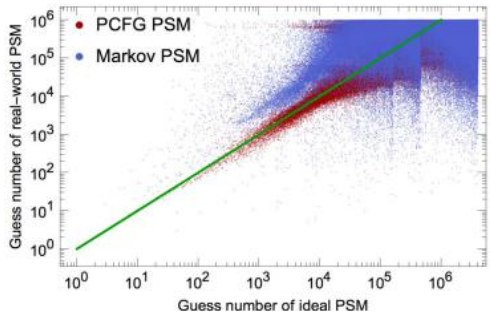


图10.基于PCFG的PSM与基于马尔科夫的PSM在猜测数方面的比较（1/4 CSDN用于训练，另外1/4 CSDN用于测试）。

为了更具体地掌握Fig。我们在表二中总结了每个PSM给出的一些典型的弱密码的猜测数字。这六个密码的弱点可以通过训练集中的小排名（见第2列）和理想仪表给出的小猜测数（见第3列）来证明。在六种情况中的四种，基于PCFG的仪表比基于马科的仪表输出更好的猜测数估计。总之，我们的fuzzyPSM，如后面进一步说明，输出最准确的密码强度估计。

表三总结了PCFG和基于马尔可夫的破解模型产生的不可用猜测的数量。如果猜测是由猜测模型产生的，但它不出现在测试集中，那么它被称为不可用。这种不可用的猜测部分地表明了一个破解模型的优点：较少的不可用的猜测将意味着一个更好的模型。我们可以看到，当猜测数很小时，基于PCFG的模型产生的不可用的猜测较少，但当猜测数大于10时，这种情况就会逆转[6].这就解释了为什么（大量猜测）基于马尔可夫的模型通常比基于PCFG的模型破解更多的密码（见[20]，[29]，[45]）。目前，通过使用643万个CSDN密码

4.基于概率模型的psm本质上是密码破解工具[28]，[33]，它们可以按概率的递减顺序输出猜测。

表二。每个PSM给出的猜测数
密码（使用1/4个CSDN密码作为训练集）。

典型密码	训练集	理想的PSM	PCFG PSM	马尔可夫PSM	fuzzyPSM
123qwe	10182	8527	771719	324	4764
123qwe123qwe	2778	3105	194671	236965	4505
password123	2315	2522	4219	6124	2397
密码123	27097	10170	29341	31009	23438
密码	18	21	20	35	46
p@ssw0rd	76	259	337982	290	274

*深灰色的猜测数表示它最接近理想PSM给出的猜测数，浅灰色的猜测数表示第二好的猜测数。

表III。基于pcfg和马尔可夫的破解模型产生的不可用的猜测数

	前10名[2]	前10名[4]	前10名[6]	前10名[7]		
	PCFG	马尔科夫	PCFG	马尔科夫	PCFG	马尔科夫	PCFG	马尔科夫		
无法使用的猜测	1		12		160		4509	372883	359401	

作为一个例子，我们设法调和了这两个看似矛盾的观察结果：基于PCFG的模型更擅长测量密码，而基于马尔科夫的模型更擅长破解密码。一个看似合理的基本原理是，平滑技术(e。g.，在基于马尔可夫的模型中，使它们更好地破解密码。e.，预测更多看不见的密码)，但这反过来使基于马尔可夫的pms受到稀疏性问题的影响，在测量弱密码方面更差。

C. 一个基于模糊PCFG的仪表

虽然基于PCFG的PSM在现有的PSM中表现最好，但它仍然受到固有的局限性。在基于PCFG的模型中，密码被认为是由三种段组成的。e.L、D和S，每个都分别代表字母、数字和符号的字符串序列。例如，p@ssw0rd由段p、@、ssw、0和rd段组成，基结构为L1S1L3D1L2；密码123由段密码和123组成，具有基本结构L8D3；123qwe123qwe由段123、qwe、123和qwe组成，基结构为D3L3D3L3.“p@ssw0”的概率（强度）计算为P（“p@ssw0”）=P(L1S1L3D1L2)· P (L1→ p)· P (S1→ @)· P (L3→ ssw)· P (S1→ 0) · P (L2→ rd).在2009年的原始方案[28]中，基结构、数字段和符号段的概率是通过计数从训练集中学习出来的，而字母段的概率是通过使用外部输入字典计算出来的。马等人。[29]认为，通过从训练集中学习来计算字母段的概率会更好。这个建议已经被广泛接受了，[20]，[29]，我们也采用了它。

从上面的例子PCFG-based的PSM措施密码，不难发现这种PSM本质上假设密码创建的新服务是由L、D和S段从零开始，它主要考虑连接规则。然而，它忽略了两个基本事实：(1)绝大比例的用户从现有的密码构建密码(见图。2)；和(2)其他的转换规则，如大写和leet也很流行(见图。5).虽然还有其他几个PSMs (e。g.，
[16]，[33]，[35]-[37])已经考虑了第二个现实，但到目前为止，大多数psm都没有考虑第一个现实。

为了模拟用户的密码重用行为，我们采用：(1)从一个不那么敏感的服务泄露的密码作为我们的基本字典$\mathcal{B}$来构建一个基本的密码解析特里树；(2)另一个相对强的密码列表一个敏感的服务，作为我们的培训字典。$\mathcal{T}$请注意，现在我们的三元树只包括低参数和长度不小于3的基本密码。然后，我们解析训练字典中的每个密码，并了解用户如何使用哪些混乱规则来为新服务构造密码。这个过程自动创建一个模糊概率上下文无关语法（PCFG），并产生我们的基于模糊PCFG的指标，模糊PSM。我们的仪表包括三个阶段：培训、测量和更新。

在训练阶段，我们表示-
确定了一定的频率表iv。
与培训相关的模式碱基概率
口令我们假设所有的传球-
单词是用片段来构建的
从基本的密码集$\mathcal{B}$=开始$\mathcal{S}$
{b1,b2，· · · ,bN}，和$\mathcal{B}\subseteq\mathcal{B}$。$\mathcal{S}$我们
表示长度的基本密码
n为基结构B_n和
考虑一下最受欢迎的三个话题吧
转换规则(i。e.，接合
国家，资本化和leet)。
培训中的每个密码
$\mathcal{T}$集合由三角树解析
使用最长的前缀匹配。其他的
更复杂的解析方法(e。g，贝叶斯解析[51])可以用来提高精度，但我们在这项工作中保持我们的仪表尽可能简单，并表明，即使是这个简单的仪表也能够提供准确的估计。

STRUCTURES AND SEGMENTS

左侧	RHS	概率
$\mathcal{S}$→	B6	0.4
$\mathcal{S}$→	B8	0.4
$\mathcal{S}$→	B8B3	0.1
$\mathcal{S}$→	B6B6	0.1
B1→	1	0.8
B1→	a	0.2
B3→	123	1
B6→	123456	0.7
B6→	123qwe	0.2
B6→	龙	0.1
B8→	密码	0.85
B8→	[illegible]	0.15

在形式上，我们的上下文无关语法（CFG）被定义为G =（V，Σ，，h），其中：$\mathcal{S}$

1) V = {,B$\mathcal{S}$1,B2，· · · ,Bk，资本化，L1,L2，· · · ,L6}是一组有限的变量，其中k不大于系统中允许的密码的最大长度。
2) Σ =$\mathcal{B}$ ={b$\mathcal{S}$1,b2，· · · ,bN是，是与V不相交的有限集。请注意，所有95个可打印的ASCII代码都在Σ中。

$\mathcal{S}$3)∈V是开始符号。
4) h是以a→a形式的有限规则集，具有a∈V和一个∈V∪Σ（见表IV到表VI）。

$\mathcal{T}$例如，密码123∈将定义基本结构B11并且不涉及转换操作，因为密码123∈$\mathcal{B}$，因此密码123可以被解析为一个段；密码123∈将定义基结构B$\mathcal{T}$11并涉及一个大写操作，因为密码123∈$\mathcal{B}$，但密码123$\mathcal{B}$；p@ssw0rd∈将定义基本结构B$\notin\mathcal{T}$8并涉及到一个精简的操作(i。$\notin\mathcal{T}$e.，o→0)，因为ssw∈$\mathcal{B}$但是ssw0$\mathcal{B}$；123qwe123qwe∈定义了基结构B6B6(i.e.，涉及一个连接操作)，因为最长的前缀是123qwe∈$\mathcal{B}$。当遇到一个密码时。g.，Tyxdqd123∈)，不能被解析(e。$\mathcal{T}$g.，由于在三角树中没有带有前缀tyx的项)，所以我们将此密码定义为基本结构B8B3，它同样可以通过使用传统的PCFG方法[34]进行解析。

训练阶段可以自动推导出训练集中所有密码的所有基本结构、基本段（包括基本密码和其他段）和转换操作及其相关的发生概率。$\mathcal{T}$要了解这方面的一个例子，请参阅表四到表六。这构建了一个概率上下文无关的语法，然后它可以用来衡量密码强度(i。e.，估计出现给定密码的概率)

表v。 A基本密码段的首字母大写的概率

	首字母大写	
	是	不
概率	0.03	0.97

测量阶段。例如，从训练阶段开始，我们了解到基础结构B8发生概率为0。4.在训练集中，B8→p@s剑发生的概率为0。15和L4→是的：o↔0发生的概率为0.004。因此，P（“p@ssw0rd1”）=P(→B$8B1) · P (B8→ p@ssword)· P (B1→ 1) · P (Capitalize→No) · P (L1→编号）· P（L2→编号）· P（L2→编号）· P（L3→是的)P（大写→No）· P(L4→No)= 0.1 * 0.15 * 0.8 * 0.97 * 0.994 * 0.995 * 0.995 * 0.004 * 0.97 * 0.997 = 0.00004431.所有相关概率均见表四至表六。这个过程如图所示。11.

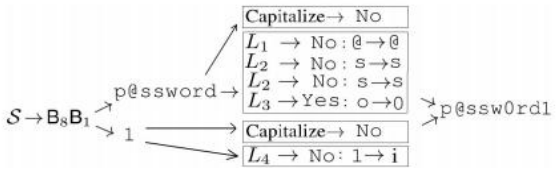


图11.密码“p@ssw0rd1”来自我们的模糊PCFG

更新阶段用于捕获随着时间的推移在系统中密码分布的动态变化。例如，在p@ssw0rd123qwe被系统接受后，所有与基本结构B相关的概率8和B6，终端p@ssword和123qwe，以及规则L3→：是的，应进行更新。这将使我们的PCFG模型更接近真实的密码分布，从而产生一个自适应的仪表。[33]，[34]中的两个psm也提供了这一特性。

PCFG是一种语法，它具有与每个规则相关联的概率，这样对于一个特定的左侧（LHS）变量，所有相关联的结果加起来为1个[28]。我们可以看出，我们的语法G确实是一个PCFG。G的语言是由从开始符号派生出的所有终端组成的字符串集。$\mathcal{S}$当使用真实的密码来实例化G时，基本结构表中超过80%的项是→B的形式$\mathcal{S}_m$(但不是→B$\mathcal{S}_m$B$_n$或者更复杂的)。这与[34]中基于PCFG的PSM形成了鲜明的对比，在[34]中，其基本结构表中通常有超过50%的项目是L的形式$_m$D$_n$或者更复杂的。换句话说，我们的PCFG对密码的传统结构不敏感。e.，因此，fuzzyPSM中的“结构”概念不如基于PCFG的PSM更清晰。模糊逻辑的目的是模仿人类的思维方式，而我们的模糊psm的目的是模仿人类为新服务创建密码的方式。因此，我们称我们的PCFG为一个模糊的PCFG。

限制。模糊psm很简单，但有几个限制。首先，它主要捕获了前3个流行的转换规则(i。e.，连接，大写和倾斜)，而其他规则，如子弦的移动和反向被留作未来的研究。其次，对于大写，它只考虑一个基本密码段的第一个字母的大写。第三，对于leet，它只考虑了前6名。*尽管如此，fuzzyPSM显示了构建一个PCFG的潜力，它可以更真实地描述用户在创建密码时的行为，正如我们将在下一节中演示的那样，它在区分弱密码方面通常优于所有现有的psm。*

表vi。李变换的概率

	某些字母的出现											
	L1：a ↔ @		L2：s ↔ $		L3：o ↔ 0		L4：i ↔ 1		L5：e ↔ 3		L6：t ↔ 7	
	是	不	是	不	是	不	是	不	是	不	是	不
概率	0.006	0.994	0.005	0.995	0.004	0.996	0.003	0.997	0.002	0.998	0.001	0.999

V.实验方法和结果

在本节中，我们首先描述我们的实验方法，包括所使用的11个数据集和各种训练的选择。测试场景。然后，我们使用在Sec中引入的两个等级相关度量。II-C提供了fuzzyPSM与五种领先的psm的比较评价。

表VII。关于11个密码数据集的基本信息

数据集	Web服务位置语言唯一的pw总pw				
天涯黎明	社交论坛，游戏和电子商务	中国	中文汉语	12,898,437	30,901,241
CSDN真微博石友战场雅虎Phpbb单身。org信仰作家	程序员论坛约会网站社交论坛社交论坛游戏网站门户网站程序员论坛基督教约会基督教写作	中国	汉语英语	10,135,260	16,258,891
		中国	英语英语	4,037,605	6,428,277
		美国	英语英语	3,521,765	5,260,229
		美国		2,828,618	4,730,662
		美国		14,326,970	32,581,870
		美国		417,453	542,386
		美国		342,510	442,834
		美国		184,341	255,373
		美国		12,233	16,248
		美国		8,346	9,708

A.数据集描述

我们使用了11个大规模的真实密码数据集，总计9740万套，以便对我们的仪表进行全面评估。关于这些数据集的基本信息见表七。其中大部分都是由这些黑客组织在互联网上公开披露的。这些服务范围从社交论坛、游戏、约会、门户网站到电子商务。这些密码有两种世界上最常用的语言，而且来自两个遥远的大陆。据我们所知，我们的语料库是迄今为止用于评估psm的最大和最多样化的一个(e。g.，在[33]，[34])中使用了三个英文数据集。

B.密码特征

我们现在已经具体掌握了用户的真实密码的情况。为此，我们研究了前10个密码、字符组成和不同密码数据集之间的重叠部分。许多其他特征，包括前10个密码，长度分布和频率分布都参考了完整的版本(见http://t.cn/RG8Ewf3)的这项工作。

表八为密码的字符组成信息。最显著的是，较大比例的英文密码只由小写字母组成，而类似比例的中文密码只由数字组成。总之，很少有密码（包括英文和中文）会包含大写字母（参见列[A-Z])或符号（参见列^[a-zA-Z0-9]+$）。

图12显示了两个不同服务之间共享的密码的比例。一般来说，考虑到不同的阈值，以及来自不同语言的共享密码，共享密码的比例小于60%。比使用同一语言的共享密码要低得多。

总结。我们的结果显示，来自每个数据集的密码可能具有完全不同的性质和分布，这归因于一些混杂因素，如语言、服务、文化、信仰和密码策略。这使得PSM很难在实际分布(i。e.，“最好的”训练集)提前。我们只能尝试使训练集尽可能接近目标站点的位置。

表VIII。关于每个密码数据集的字符组成信息

数据集	^[a-z]+$	[a-z]	^[A-Z]+$	[A-Z]	^[A-Za-z]+$	[a-zA-Z]	^[0-9]+$	[0-9]	仅限符号	^[a-zAZ0-9]+$	^[0-9]+[a-z]+$	^[a-zA-Z]+ [0-9]$+	^[0-9]+[a-zA-Z]+$	^[a-z]+1$
天涯	9.91%	34.63%	0.18%	2.96%	10.24%	35.66%	**63.77%**	89.49%	0.03%	98.08%	4.12%	15.73%	4.39%	0.12%
多诺	10 30%	66 32%	0 30%	3 67%	10 92%	69 05%	**30 76%**	88 52%	0 02%	98 33%	7 55%	45 74%	7 93%	1 40%
中国程序员大本营“网站简称	11 64%	51 39%	0 47%	4 57%	12 35%	54 33%	**45 01%**	87 10%	0 03%	96 31%	5 88%	28 45%	6 46%	0 24%
真艾	6 41%	37 33%	0 24%	3 40%	6 74%	39 54%	**59 52%**	92 87%	0 02%	95 79%	5 24%	21 09%	5 69%	0 08%
微博	19 07%	44 77%	0 64%	3 66%	20 55%	46 71%	**53.04%**	78 78%	0 06%	97 79%	2 80%	18 74%	2 91%	1 24%
摇滚你	**41.71%**	80.58%	1.50%	5.94%	**44.07%**	83.89%	15.94%	54.04%	0.02%	96.25%	2.54%	30.18%	2.75%	4.55%
战场	**32.11%**	89.71%	0.29%	9.60%	**34.01%**	90.69%	9.23%	65.49%	0.01%	98.06%	3.05%	39.58%	3.39%	5.08%
雅虎	**33.09%**	92.83%	0.40%	8.51%	**34.64%**	94.06%	5.89%	64.74%	0.00%	97.15%	5.31%	41.85%	5.64%	4.80%
Phpbb单身。org	**50 18%**	86 18%	0 74%	7 70%	**53 07%**	87 83%	12 06%	46 14%	0 03%	98 34%	2 03%	20 94%	2 35%	2 33%
信仰作家	**60 21%**	87 84%	1 92%	8 14%	**65 82%**	90 42%	9 58%	34 06%	0 00%	99 79%	1 77%	19 68%	1 92%	2 73%
	54 37%	91 74%	1 16%	8 84%	**58 98%**	93 64%	6 36%	40 88%	0 00%	99 52%	2 37%	25 45%	2 73%	4 13%

请注意，第一行是用正则表达式编写的。例如，^[a-z]+$表示仅由小写字母组成的密码；^[a-Za-z]+$表示仅由字母组成的密码；[a-z]表示包含小写字母作为子字符串的密码；^[0-9]+[a-z]+$表示由数字和小写字母组成的密码。

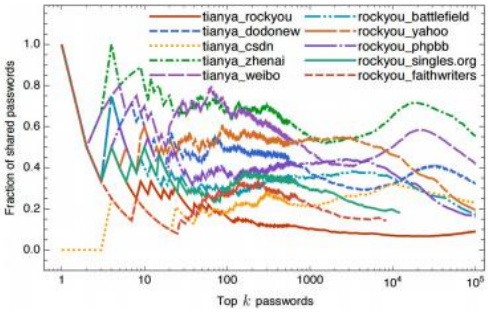


图12. 两个服务之间共享的密码的分数。

幸运的是，找到一个“好”的训练集相对容易：因为语言和服务是两个最重要的因素[37]，[45]，可以选择一个数据集相同的语言和服务测试集训练集，有足够的（数百）公开泄露数据集（见[27][48]）这样的选择。我们离开未来的工作去开发一个量化的、成熟的度量来衡量训练集(s)的优度。此外，当新用户注册时，密码分配将发生更改。这意味着没有一个静态规则集可以表征密码分配中的动态变化，这需要自适应仪表。

C. 实验装置

如上图所示，密码的分布受到多种因素的影响，其中语言占主导地位。因此，我们将训练集和测试集分为两组。为了模拟密码重用的用户行为，我们使用Rockyou和天涯作为fuzzyPSM的基本字典。为了更好地评估，我们考虑了两种情况：理想情况和真实情况。在理想情况下，每个PSM的训练集由从测试集中随机抽取的1/4个密码组成，从测试集中随机抽取的1/4个密码被实际测量。正如在IV-A中所述，这消除了不适当的训练集的影响，并且评估结果仅依赖于仪表本身。

然而，理想的情况在实践中是不现实的，因为人们不能获得从与测试集相同的分布中抽取的训练集。现实的情况是，在网站建立后，网站使用来自相同语言和策略和类似服务的密码来训练PSM，同时，将用户提交的密码插入到训练集中，动态更新PSM。在不丧失一般性的情况下，这个场景可以被建模为：PSM对从类似服务泄露的密码和1/4测试集的密码进行训练，然后用于测量测试集中的所有密码。此外，我们还研究了psm在训练来自一种语言的密码但用于测量来自另一种语言的密码时的表现。这一切都是

表ix。针对每个PSM的培训和测试场景

场景	基础*词典	培训集(s)		测试集，（每一组均进行单独测试）
		理想情况	现实世界	
英语	摇滚你	四分之一测试集	Phpbb，1/4测试集	雅虎，战场，单身人士，忠实作家
中国人	天涯	四分之一测试集	微博，四分之一测试集	Dodonew，CSDN，Zhenai
跨域网。#1	摇滚你		Phpbb，1/4 Dodonew	多诺
跨域网。#2	天涯		微博，1/4雅虎	雅虎

*基础字典仅由fuzzyPSM使用；“跨-lan。”意思是“跨语言”。

汇总见表IX。Rockyou和天涯是每个语言组[29]、[45]中最弱的，因此它们被选为基础字典；Phpbb和微博在每个组中都具有中等强度，因此在现实世界中它们被选为我们的模糊psm的训练集。

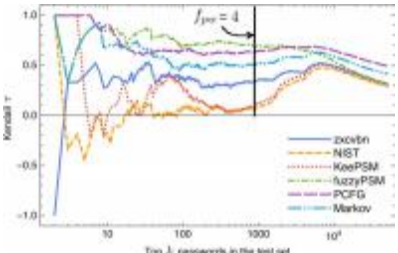
D. 评价结果

在本节中，我们比较了所提出的模糊psm与其他五种领先的psm的性能，包括两个来自学术界(i。e.，基于马尔可夫的[33]和基于PCFG的[34])，两个来自行业(i。e.，Zxcvbn [35]和KeePSM [36])和一个来自标准体(i。e.，净[16])。表九中的理想情况和真实情况将导致9(见图。13(a)到13(i))和7（见图。分别进行13(j)至13(p))次单独实验。我们发现，在我们所有的实验中，基于斯pearmanp的结果与基于肯德尔T的结果没有明显的差异(见图。9(a)和无花果。例如，9(b))，因此我们只说明后者以节省空间。

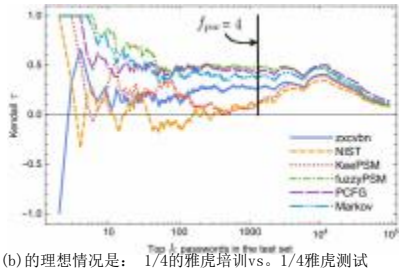
在大多数情况下，我们的fuzyyyPSM(用图中的绿线表示。当要测量的密码的频率不小于4时，13)表现最好。这一点在现实世界中尤为突出。详见Sec。对于流行的密码(e。g.，与$f_{pw} \geq 4$)*我们可以使用经验概率*f_{pw}*DS*/||近似其实概率p_{pw}其相对标准误差约为$1/\sqrt{f_{pw}}$[41]，其中f_{pw}是密码数据集中p的频率吗*w DS*.换句话说，只有这些密码与$f_{pw} \geq 4$*显示了它真正的强度，理想的仪表可以有意义地测量它们*。对于f的密码$_{pw}$<4，由于理想的仪表无效，基准不可靠，因此所有的比较结果都没有什么价值。因此，fuzyPSM在所有检测的psm中表现最好。

VI. 结论

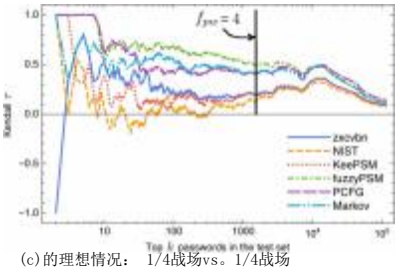
在本文中，我们提出了一种简单而有效的密码强度计，模糊psm，以便在用户选择密码时提供可靠的反馈。受用户密码实践调查的启发，fuzzyPSM为描述密码创建中的真实用户行为迈出了第一步。大量的实验表明，fuzzyPSM优于学术界、工业界和标准机构的领先psm。特别是11个大规模的密码数据集，其中包括9740万个真实生活中的密码，涵盖了各种流行的服务和多样化的用户基础(e。g.，语言和文化)，被用来建立模糊psm的实用性。



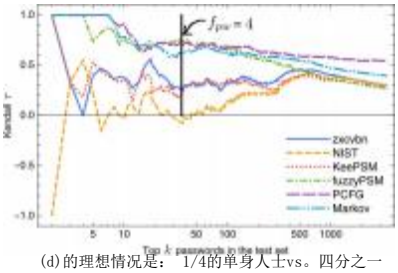
(a)的理想情况： 1/4 Phpbb训练vs。1/4 Phpbb测试



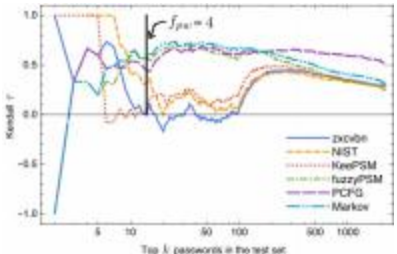
(b)的理想情况是： 1/4的雅虎培训vs。1/4雅虎测试



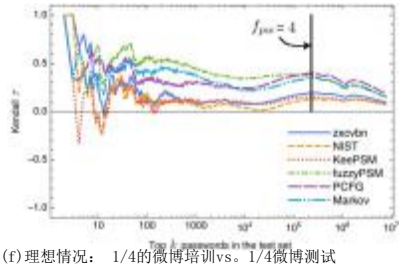
(c)的理想情况： 1/4战场vs。1/4战场



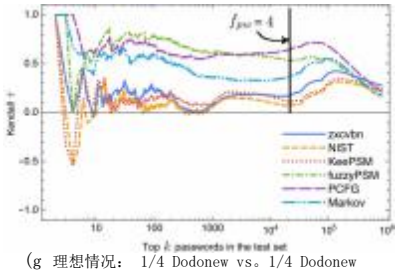
(d)的理想情况是： 1/4的单身人士vs。四分之一的单身



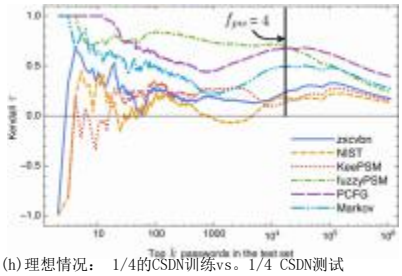
(e 理想情况： 1/4作者vs。1/4 Faithwriters



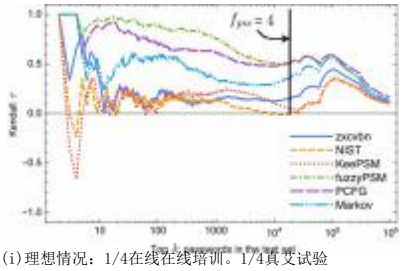
(f)理想情况： 1/4的微博培训vs。1/4微博测试



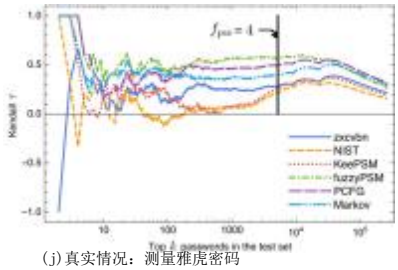
(g 理想情况： 1/4 Dodonew vs。1/4 Dodonew



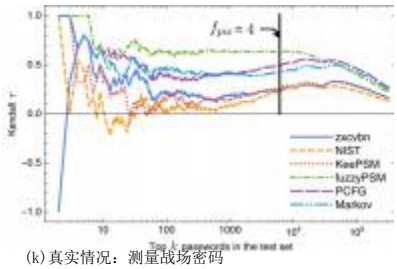
(h)理想情况： 1/4的CSDN训练vs。1/4 CSDN测试



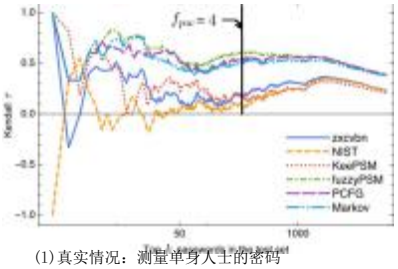
(i)理想情况：1/4在线在线培训。1/4真艾试验



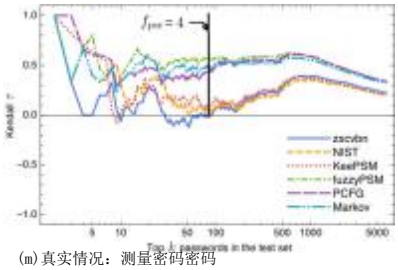
(j)真实情况：测量雅虎密码



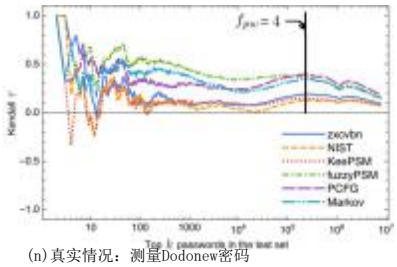
(k)真实情况：测量战场密码



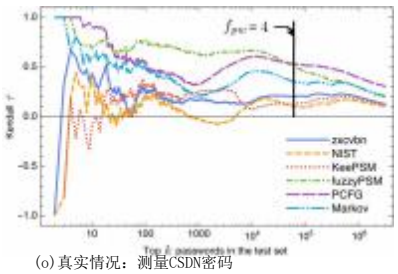
(l)真实情况：测量单身人士的密码



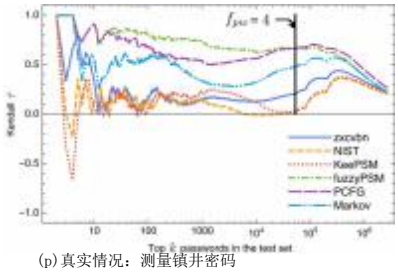
(m)真实情况：测量密码密码



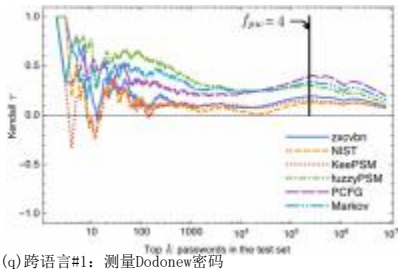
(n)真实情况：测量Dodonew密码



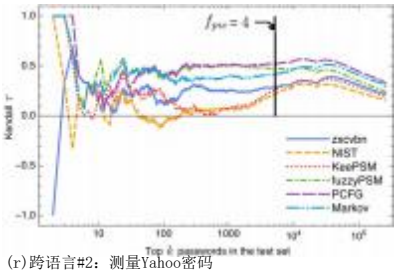
(o)真实情况：测量CSDN密码



(p)真实情况：测量镇井密码



(q)跨语言#1：测量Dodonew密码



(r)跨语言#2：测量Yahoo密码

图13. 理想情况（见子图(a)～(i)）和真实情况（见子图(j)～(p)）的实验结果。子图(q)和(r)显示了跨语言测量的无效。绿色的线表示模糊的，psm通常表现最好。

确认

我们感谢匿名的审稿人提供的宝贵评论。王平是通讯作者。本研究由国家自然科学基金资助项目（NSFC）资助。61472016和61501333。

参考文献

[1] J.波诺，C。赫利，P。范·奥尔肖特和F。“密码与不完善认证的演变”，《评论。*ACM，vol*。58岁，没有。7，pp.78–87，2015.

[2] J.燕，A。F.布莱克威尔，R。J.安德森和A。“密码记忆和安全性：经验结果。”IEEE安全系统。*& priv.*，卷。2、没有。5，pp.25–31，2004.

[3] W.《重新思考密码》杂志。*ACM，vol*。56岁，没有。2，pp.40–44，2013.

[4] Z.赵，GJ。.-Ahn和H。胡，“图片手势认证：实证分析、自动攻击和方案评估”，ACM跨式。*通知。西斯特。安全。*，卷。17日，没有。4，pp.1–37，2015.

[5] D.王和P。王，“一石两鸟：超越传统界限的双因素认证”，IEEE Trans。取决于。安全。Comput，2016，出版，可在http://t获得。cn/RGCaI8f.

[6] J.波诺，C。赫利，P。奥尔肖特和F。“取代密码的追求：web身份验证方案的比较评估框架”。*IEEE标准普尔2012，页*。553–567.

[7] C.赫利和P。范·奥肖特，“一个承认密码持久性的研究议程”，IEEE Secur。*& Priv.*，卷。10，没有。1，pp.28– 36，2012.

[8] A.亚当斯和M。A.*“用户不是敌人。”ACM，vol*。42岁，没有。12，pp.40–46，1999.

[9] R.W.过程器，MC。.-留置权，KP。.-L.Vu，E.E.舒尔茨和G。“提高用户认证的计算机安全性：主动密码限制的影响”，比哈夫。*物品甲安菲他明安装。& Comput.*，卷。34岁，没有。2，pp.163–169，2002.

[10] S.艾格尔曼。Sotirakopoulos，K.贝兹诺索夫和C。“我的密码会到11吗？”：密码计对密码选择的影响。”*ACM CHI 2013，页*。2379–2388.

[11] R.谢伊，S。Komanduri，P.“遇到更强的密码要求：用户的态度和行为”。*汤2010，pp*。1–20.

[12] B.Ur，P.G.凯利，S。Komanduri，J.李，M。马斯，M。马祖里克，T。帕萨罗，R。谢伊，T。维达斯，L。鲍尔等人。，“你的密码如何匹配的？”强度计对密码创建的影响。*USENIX SEC 2012，页*。65–80.

[13] B.Ur，F.NomaJ。蜜蜂，S。M.塞格雷蒂，R。谢伊，L。鲍尔，N。克里斯汀和L。F.克兰纳，“我补充了一句！”最后确保它安全：在实验室观察密码创建。*汤2015，pp*。123– 140.

[14] D.王和P。王，“皇帝的新密码创建政策”。*ESORICS 2015，pp*.456–477.

[15] X.卡纳瓦莱特和M。曼南，“高影响密码强度计的大规模评估”，ACM Trans。*通知。西斯特。安全。*，卷。18日，没有。1，pp.1–32，2015.

[16] W.毛刺，D。多德森，R。Perlner，S.Gupta和E。Nabbus，“NISTSP800-63-2：电子认证指南”，美国国家标准与技术研究所，莱斯顿，弗吉尼亚州，技术部。众议员，8月。2013.

[17] M.Weir，S。阿格瓦尔，M。柯林斯和H。“通过攻击大量被泄露的密码来测试密码创建策略的指标”。*ACM中国化学会2010，页*。162– 175.

[18] P.G.凯利，S。Komanduri，M.L.马祖里克，R。谢伊，T。维达斯，L。鲍尔，N。Christin，L.F.克兰诺和J。洛佩兹，“反复猜测（一遍又一遍）：通过模拟密码破解算法来测量密码强度”，在Proc中说。*IEEE标准普尔2012，页*。523–537.

[19] M.Dell'Amico，P.米奇亚迪和Y。《密码强度：实证分析》，在序言中提到。*INFOCOM 2010页*。1–9.

[20] M.戴尔的Amico和M。“蒙特卡罗强度评估：快速可靠的密码检查”，在项目。*ACM中国化学会2015*。

[21] B.Ur，S.M.SegretiL。鲍尔和等人，“测量真实世界的准确性和偏差”，在Proc。*USENIX SEC 2015，页*。463–481.

[22] D.Florncio，C.Herley和P。C.“密码投资组合和有限努力的用户：可持续地管理大量的账户”。*USENIX SEC 2014，页*。575–590.

[23] F.张，K。LeachH。王和A。“信任登录：安全”商品操作系统密码登录，在亚洲商会2015。

[24] B.Strahs，C。Yue和H。王，“通过增强哈希保护密码”。*丽莎2009年，页*。93– 106.

[25] M.Dürmuth和T。Kranz，“用gpu和fpga猜测密码”，在Proc中。*密码2014，pp*。19–38.

[26] C.*艾伦，3200万罗克尤密码被盗，12月*。2009，http://www.hardwareheaven.com/news.php?newsid=526.

[27] R.*马丁，在中国广泛的数据泄露中。*2011,http://www.techinasia.com/alipay-hack/.

[28] M.Weir，S。AggarwalB。de Medeiros和B。“使用与概率上下文无关的语法进行密码破解”。*IEEE标准普尔2009，页*。391–405.

[29] J.妈妈，W。杨，M。罗和N。李，“概率密码模型的研究”，在序言中。*IEEE标准普尔2014，页*。689–704.

[30] M.Dürmuth，D.弗里曼和B。比乔，“你是谁？”一种衡量用户真实性的统计方法，”在NDSS 2016，页。1– 15.

[31] M.阿尔萨利赫，M。曼南和P。范·奥肖特，“重新审视防御大规模的在线密码猜测攻击”，IEEE Trans。*取决于。安全。压缩。*，卷。9，没有。1，pp.128– 141，2012.

[32] S.S.席尔瓦，R。M.席尔瓦，R。C.平托和R。M.《僵尸网络：调查》，康出版社。*网络。*，卷。57岁，没有。2，pp.378–403，2013.

[33] C.Castelluccia，M.Dürmuth和D。佩里托，“自适应密码强度米从马尔科夫模型，”，在Proc。*NDSS 2012，pp*.1– 15.

[34] S.霍什曼德和S。“使用概率技术建立更好的密码”。*ACSAC 2012，pp*.109– 118.

[35] D.*惠勒，zxcvbn：现实的密码强度估计，2012年4月，*https://t.co/kLO4dLzqFS.

[36] D.*Reichl，关于KeePass的质量/强度评估的细节，2015年7月，*http://keepass.info/help/kb/pw_质量_est.html.

[37] M.雅各布森和M。《理解密码的好处》，在Proc中提到。*HotSec 2012，pp*.1–6.

[38] A.达斯，J。博诺，M。凯撒，N。博里索夫和X。王，“密码重用的复杂之网”，在程序中。*NDSS 2014，pp*.1– 15.

[39] R.*格雷厄姆，PHPbb密码分析，2009年6月，http://www。*darkreading.com/risk/phpbb-password-analysis/d/d-id/1130335?

[40] C.*库斯特，2015年7月，中国网民数量达到6.68亿，https:*//www.techinasia.com/chinas-internet-thirdlargest-country-earth/.

[41] J.“猜测的科学：分析一个包含7000万个密码的匿名语料库”。*IEEE标准普尔2012，页*。1– 15.

[42] D.杰格，H。格劳普纳。Sapegin，F.程和C。梅内尔，“收集和分析身份泄露的安全意识，”在Proc。*密码2014，pp*。102– 115.

[43] E.鲍曼，Y。Lu和Z。林，“半个世纪的实践：谁还在存储明文密码？”在项目。*ISPEC 2015，pp*.253–267.

[44] D.Florncio，C.Herley和P。范·奥肖特，《互联网密码研究》，管理员指南。*USENIX LISA 2014，页*。44–61.

[45] D.王，H。程，P。王，X。黄和C。朱棣，“了解中国用户的密码：调查和实证分析”，CACR报告，2015年中国密码报告，http://t.cn/RG8RacH.

[46] E.E.*库瑞顿，“平均先锋等级标准相关性”，心理测量术，第1卷*。23日，没有。3，pp.271–272，1958.

[47] L.M.阿德勒，“肯德尔对任意关系的修改，”J。*美国人停滞不前。使发生联系*，卷。52岁，没有。277，pp.33–35，1957.

[48] M.卡梅伦，所有数据泄露来源，2月2日。2016，https://breachalarm.com/allsources.

[49] S.T.哈克，M。赖特和S。“用户的层次结构？”网络密码：感知、做法和敏感性”。*J.HumComput。种马*，卷。72岁，没有。12，pp.860–874，2014.

[50] E.斯托伯特和R。《密码生命周期：用户在管理密码中的行为》。*汤2014，pp*。243–255.

[51] K.“动态贝叶斯网络：表示、推理与学习”，博士论文。D.毕业论文，加州大学伯克利分校，2002年。