人类选择密码中的个人信息及其安全含义研究

⊙、王海宁、孙昆
*威廉玛丽伊利学院计算机科学系，ksun @cs.wm.edu
特拉华大学电气和计算机工程系 hnw@udel.edu

摘要——尽管不推荐，互联网用户经常在他们的密码中包含部分个人信息，以便于记忆。然而，过去还没有系统地研究密码中个人信息的使用及其安全含义。在本文中，我们首先从泄露的数据集中剖析用户密码，以调查用户个人信息如何以及在多大程度上驻留在密码中。特别地，我们提取了由个人信息表达的最流行的密码结构，并展示了个人信息的用法。然后，我们引入一个新的度量标准，称为覆盖率，以量化密码和个人信息之间的相关性。然后，基于我们的分析，我们将概率上下文无关文法(PCFG)方法扩展到语义丰富，并提出个人 PCFG 通过生成个性化猜测来破解密码。通过离线和在线攻击场景，我们证明了个人 PCFG 破解密码比 PCFG 快得多，并使在线攻击更容易成功。

I. 介绍

在可预见的未来，基于文本的密码仍然是一种占主导地位且不可替代的身份验证方法。尽管人们已经提出了不同的认证机制，但是没有一种替代方案可以在不给用户带来任何额外负担的情况下带来密码的所有好处。然而，密码一直被认为是身份认证中最薄弱的环节之一。由于人类记忆的需要，用户密码通常不是真正的随机字符串[2]-[6]。换句话说，人类用户倾向于选择弱密码，仅仅因为它们容易记住。因此，大多数密码只在整个密码空间的一小部分中选择，容易受到暴力和字典攻击。

为了提高密码安全性，在线认证系统开始实施更严格的密码策略。同时，许多网站部署密码强度计来帮助用户选择安全的密码。然而，这些仪表被证明是临时的和不一致的[7]，[8]。为了更好地评估密码的强度，我们需要更深入地了解用户是如何构造他们的密码的。如果攻击者确切地知道用户如何创建他们的密码，猜测他们的密码将变得容易得多。同时，如果用户意识到由常用的密码创建方法引起的潜在脆弱性，则用户可以避免使用相同的方法来创建密码。

为此，研究人员付出了巨大努力来揭示密码的结构。传统词典对密码的攻击表明，用户倾向于使用简单的字典单词来构建他们的密码[9]。语言也很重要，因为用户在构建密码时倾向于使用他们的第一语言[2]。此外，尽管密码不是简单的字典单词，但它们大多在语音上容易记忆。还指出用户可以在他们的密码中使用键盘和日期串[5]、[10]、[11]。然而，大多数研究只发现了表面的密码模式，密码的语义丰富的组成仍然是神秘的，有待完全揭示。幸运的是，一项启发性的工作研究了用户如何通过学习密码中的语义模式来生成他们的密码[12]。

在本文中，我们从一个不同的角度研究密码语义，即个人信息的使用。在这项研究中，我们利用了一个从中国网站上泄露的密码数据集，其中包含个人信息。我们首先测量密码创建中个人信息的使用情况，并给出有趣的观察结果。我们能够获得最流行的嵌入个人信息的密码结构。我们还观察到，男性和女性在使用个人信息创建密码时表现不同。接下来，我们引入一个称为覆盖率的新指标来准确量化个人信息和用户密码之间的相关性。因为它考虑了密码中个人信息的长度和连续性，所以覆盖率是衡量密码强度的有用指标。我们使用覆盖率度量的量化结果证实了我们在数据集上的直接测量结果，显示了覆盖率的有效性。此外，覆盖率很容易与现有工具集成，例如用于创建更安全密码的密码强度计。

为了证明在密码中使用个人信息导致的安全漏洞，我们提出了一种称为个人 PCFG 的语义丰富的概率上下文无关文法(PCFG)方法，该方法通过考虑密码结构中与个人信息相关的符号来扩展 PCFG [13]。个人-PCFG 破解密码的速度比 PCFG 快得多。它还通过大幅提高猜测成功率，使在线攻击更加可行。最后，我们讨论了抵御语义感知攻击(如个人 PCFG)的潜在解决方案。

我们的研究基于从一个中文网站收集的数据集。虽然测量结果可能与其他数据集不同，但我们的观察仍然揭示了密码中如何使用个人信息。只要可记忆性在密码创建中起着重要的作用，那么

无论用户使用哪种语言，个人信息和用户密码之间的关联仍然存在。我们相信，我们在个人信息量化、密码破解和密码保护方面的工作可以适用于来自不同网站的任何其他基于文本的密码数据集。

本文的其余部分组织如下。第二节测量了用户密码中的个人信息，并显示了密码创建中的性别差异。第三节介绍了新的度量标准“覆盖率”,以准确量化个人信息和用户密码之间的相关性。第四节详细介绍了个人 PCFG，并显示了与原 PCFG 相比的开裂结果。第五节讨论了限制和潜在的防御。第六节概述相关工作，最后第七节总结本文。

II. 密码中的个人信息直觉上，人们倾向于根据他们的个人信息，因为人类受记忆能力的限制，随机密码更难记住。通过剖析一个中等规模的泄露密码数据集中的密码，我们表明用户的个人信息在人类选择的密码生成中起着重要作用。了解密码中个人信息的用法及其安全含义可以帮助我们进一步增强密码的安全性。首先，我们介绍整个研究中使用的数据集。

A. *12306 数据集*

近年来，许多密码数据集已经暴露在公众面前，通常包含几千到几百万个真实密码。因此，有一些基于分析这些数据集的密码测量或密码破解研究[2]，[10]。在本文中，一个名为 12306 的数据集用于说明个人信息是如何参与密码创建的。

1) 数据集简介:2014 年底，一个中文数据集被匿名攻击者泄露给公众。据报道，该数据集是通过在网上尝试其他泄露的数据集的用户名和密码收集的。我们称这个数据集为 12306，因为所有密码都来自网站 www.12306.cn，这是中国铁路网上订票系统的官方网站。12306 网站的确切用户数没有数据可查；然而，我们推断该系统至少有数千万注册用户，因为它是整个中国铁路系统的唯一官方网站。

12306 数据集包含超过 13 万中国人密码。目睹了这么多泄露的大数据集，12306 数据集的大小算是中等。它的特别之处在于，除了明文密码，该数据集还包括几种类型的用户个人信息，如用户名和政府颁发的唯一 ID 号(类似于美国社会安全号)。由于该网站需要真实的身份证号码注册，人们必须提供真实的个人信息才能订票，我们认为该数据集中的信息是可靠的。

表一:最常用的密码。

军阶	密码	数量	百分率
一	123456	389	0.296%
2	a123456	280	0.213%
3	123456a	165	0.125%
四	5201314	160	0.121%
5	111111	156	0.118%
6	沃艾尼	134	0.101%
七	1314 qq123456	98	0.074%
8	123123	97	0.073%
9	000000	96	0.073%
10	1qaz2wsx	92	0.070%

*2) 基本分析:我们首先进行一个简单的分析来揭示 12306 数据集的一些一般特征。为了数据的一致性，我们删除了 ID 号长度不是 18 位的用户。这些用户可能使用其他 id(如护照号码)在系统上注册，占整个数据集的 0.2%。该数据集包含 131，389 个密码，用于清理后的分析。请注意，不同的网站可能有不同的密码创建策略。例如，通过严格的密码策略，用户可以应用篡改规则(例如，abc > @bc 或 abc1)，以满足策略要求[14]。自从 12306 网站更改了它的*密码策略在密码泄露后，我们不知道数据集第一次被破坏时的确切密码策略。然而，从泄露的数据集中，我们推断密码策略相当简单——所有密码都不能短于六个符号。对于可以使用什么类型的符号没有限制。因此，用户不需要对他们的密码应用任何篡改规则。

12306 数据集中密码的平均长度是 8.44。表 1 列出了 12306 数据集中最常见的密码。主要的密码是普通密码(如 123456 、 a123456 等 ） 。 ） 、 键 盘 密 码 （ 如 1qaz2wsx 、 1q2w3e4r 等 。 ） ， 以 及 “ iloveyou” 密码。“5201314”和“我爱你 1314”在中文里都是“我永远爱你”的意思。最常用的中文密码类似于之前的一项研究[10]；但是，12306 数据集要稀疏得多。最流行的密码“123456”在所有密码中所占的比例不到 0.3%，而在[10]中这一数字为 2.17%。我们认为密码稀疏是因为网站的重要性；用户不太倾向于使用像“123456”这样的琐碎密码，因为注册需要真实的身份证号码，所以 sybil 帐户较少。

与[10]类似，我们根据各种指标测量 12306 数据集的抗猜测性，包括最坏情况安全位表示(H)和猜测位表示(g∞)，α-猜测位表示(G˜ 0.25 和 G˜ 0.5)，以及 β 成功率(λ5 和 λ10)。结果如表 II 所示。我们发现，12306 的用户避免使用非常容易被猜到的密码，如“123456”，因为 12306 的最坏情况要高得多 β = 5 和 10 时的安全性和 β 成功率。我们相信用户在创建时会有一定的密码安全顾虑 12306 等关键服务系统的密码。然而，他们的关注似乎仅限于避免极其简单的密码。正如阿尔法猜测值所示，12306 数据集的总体密码稀疏性并没有提高

表二:抵制猜测表四:个人信息。

H∞型	λ5 λ10	G˜	G˜0.5
g		0.25	
8.4 16.85	0.25% 0.44%	16.65	16.8

表三:最常见的密码结构。

军阶	结构	数量	百分率
一	D7	10,893	8.290%
2	D8	9,442	7.186%
3	D6	9,084	6.913%
四	L2D7	5,065	3.854%
5		4,820	3.668%
6	L3D6	4,770	3.630%
七		4,261	3.243%
8	L1D7	3,883	2.955%
9		3,590	2.732%
10		3,362	2.558%

“D”代表数字，“L”代表英文字母。该数字表示线段长度。例如，L2D7 表示密码包含 2 个字母和 7 个数字。

比以前研究的数据集要多。

我们还研究了 12306 密码的基本结构。最流行的密码结构如表 III 所示。与之前的研究[10]类似，我们的结果再次表明，中国用户更喜欢在密码中使用数字，而不是像英语用户那样使用字母。前五种结构都有显著的数字部分，最多在前面附加 2 或 3 个字母。这背后的原因可能是汉字是基于语标的，在创建密码时，数字似乎是最佳选择。

综上所述，12306 数据集是一个具有一般中文密码特征的中文密码数据集。用户选择不那么简单的密码会有一定的安全顾虑。然而，12306 数据集的总体稀疏性并不比以前研究的数据集高。

B. 个人信息

12306 数据集不仅包含用户密码，还包含表 IV 中列出的多种类型的个人信息。

注意，政府发放的身份证号是唯一的 18 位数字，其中包含个人信息本身。数字 1-6 代表主人的出生地，数字 7-14 代表主人的出生日期，数字 17 代表主人的性别——奇数表示男性，偶数表示女性。我们取出 8 位数的出生日期并分别处理，因为出生日期在密码创建中是非常重要的个人信息。所以我们最终有了六类个人信息:姓名、生日、邮箱、手机号、账户名、身份证号(生日除外)。

1) 新的密码表示法:为了更好地说明个人信息与用户密码之间的关系，我们开发了一种新的密码表示法，除了传统的“D”、“L”和“S”符号(代表数字、字母和特殊符号)之外，还添加了更多的语义符号。

类型	描述
姓名	用户的中文姓名
电子邮件地址	用户注册的电子邮件地址
手机	用户的注册手机号码
帐户名	用于登录系统的用户名
身份证	

分别是。我们尝试将用户密码的一部分与六种类型的个人信息进行匹配，并用这些个人信息表示密码。例如，密码“alice1987abc”可以表示为[姓名][生日]L3，而不是传统表示中的 L3D4L3。这
匹配的个人信息由相应的标签表示，在本例中为[姓名]和[出生日期];对于不匹配的段，我们仍然使用“D”、“L”和“S”来描述符号类型。

我们认为像[Name][Birthdate]L3 这样的表示法比 L5D4L3 更好，因为它们用更详细的语义更准确地描述了用户密码的组成
信息。使用这种表示，我们将以下匹配方法应用于整个 12306 数据集，以查看这些个人信息标签如何出现在密码结构中。

2) 匹配方法:我们提出一种匹配方法来定位用户密码中的个人信息。基本思想是，我们首先生成密码的所有子串，并按长度降序排序。然后我们将这些子字符串从最长到最短匹配到所有类型的个人信息。如果找到匹配项，匹配函数将递归应用于剩余的密码段，直到不再找到匹配项。我们要求要匹配的段至少有 2 个符号长。与任何个人信息都不匹配的段将使用传统的“LDS”标签进行标记。

我们描述用于匹配每种类型的个人信息的方法如下。对于中文名字，我们将其转换成拼音形式，这是汉字的字母表示。然后，我们将密码段与一个名字的 10 种可能排列进行比较，例如姓氏+名字和姓氏首字母+名字。如果片段与排列中的一个完全相同，我们认为它是匹配的。对于生日，我们列出了 17 种可能的排列，并将密码段与这些排列进行比较。如果线段与任何排列相同，我们认为它是匹配的。对于帐户名、电子邮件地址、手机号码和 ID 号码，我们进一步将一个段的长度限制为至少 3，以避免由于巧合而导致的不匹配。此外，由于人们倾向于通过分成 3 位数的组来记忆数字序列，我们认为至少 3 的匹配可能是真正的匹配。

请注意，对于密码段，它可能匹配多种类型的个人信息。在这种情况下，所有可能的匹配都会计入结果中。

3) 匹配结果:在将匹配方法应用于 12306 个数据集之后，我们发现 131389 个密码中的 78975 个(60.1%)包含六种个人密码类型中的至少一种

表五:最常见的密码结构。

军阶	结构	数量	百分率
一	[ACCT]	6,820	5.190%
2	D7	6,224	4.737%
3	[名称][BD]	5,410	4.117%
四	[BD]	4,470	3.402%
5	D6	4,326	3.292%
6	[电子邮件]	3,807	2.897%
七	D8	3,745	2.850%
8	L1D7	2,829	2.153%
9	【名称】D7	2,504	1.905%
10	[ACCT][BD]	2,191	1.667%

表六:个人信息使用情况。

军阶	信息类型	数量	百分率
一	出生年月日	31,674	24.10%
2	帐户名	31,017	23.60%
3	名字	29,377	22.35%
四	电子邮件	16,642	12.66%
5	识别号	3,937	2.996%
6	手机	3,582	2.726%

信息。显然，个人信息经常用于密码创建。我们相信，如果我们了解更多的用户个人信息，这一比例还会更高。我们在表 V 中列出了十大密码结构，在表 VI 中列出了密码中个人信息的使用情况。如上所述，一个密码段可能匹配多种类型的个人信息，我们对所有这些匹配进行统计。因此，百分比之和大于 60.1%。在 131，389 个密码中，我们获得了 153，895 个密码结构。根据表 V 和表 VI，我们可以看到人们在创建密码时很大程度上依赖于个人信息。在 6 种类型的个人信息中，出生日期、帐户名和姓名最受欢迎，出现率超过 20%。12.66%的用户在密码中包含邮箱。然而，只有很少比例的人在密码中包含他们的手机和身份证号码(不到 3%)。

4) 性别密码偏好:由于我们的数据集中的用户 ID 号实际上包含性别信息(即 ID 号中的倒数第二位数字代表性别)，我们比较了男性和女性的密码结构，以查看密码偏好是否有任何差异。由于数据集在性别上有偏差，有 9，856 名女性和 121，533 名男性，我们随机选择 9，856 名男性并与女性进行比较。

男性和女性的平均密码长度是
分别为 8.41 和 8.51 个字符，这表明性别对密码长度的影响并不大。然后，我们将匹配方法应用于每个性别。我们观察到 61.0%的男性密码包含个人信息，而只有 54.1%的女性密码包含个人信息。我们在表七中列出了每个性别的前 10 个结构，在表八中列出了个人信息的使用。这些结果表明，男性用户比女性用户更有可能在密码中包含个人信息。此外，我们还有另外两个有趣的观察结果。第一，总数

表七:不同性别中最常见的结构。

军阶	男性的		女性的	
	结构	百分率	结构	百分率
一	[ACCT]	4.647%	D6	3.909%
2	D7	4.325%	[ACCT]	3.729%
3	[名称][BD]	3.594%	D7	3.172%
四	[BD]	3.080%	D8	2.453%
5	D6	2.645%	[电子邮件]	2.372%
6	[电子邮件]	2.541%	[名称][BD]	2.309%
七	D8	2.158%	[BD]	1.968%
8	L1D7	2.088%	L2D6	1.518%
9	【名称】D7	1.749%	L1D7	1.267%
10	[ACCT][BD]	1.557%	L2D7	1.240%
钠	总数	28.384%	总数	23.937%

表八:不同性别最常出现的个人信息。

军阶	男性的		女性的	
	信息类型	百分率	信息类型	百分率
一	[BD]	24.56%	[ACCT]	22.59%
2	[ACCT]	23.70%	[BD]	20.56%
3	[姓名]	23.31%	[姓名]	12.94%
四	[电子邮件]	12.10%	[电子邮件]	13.62%
5	[ID]	2.698%	[细胞]	2.982%
6	[细胞]	2.506%	[ID]	2.739%

女性的密码结构数量为 1 756 个，比男性多 10.3%。此外，28.38%的男性密码属于前 10 位结构，而只有 23.94%的女性密码属于前 10 位结构。因此，男性创造的密码更密集，更容易预测。第二，男性和女性在姓名信息的使用上存在显著差异。23.32%的男性密码中包含自己的名字。相比之下，只有 12.94%的女性密码包含她们的名字。我们注意到名字是男性和女性在个人信息使用上的主要区别。

综上所述，男性的密码一般由更多的个人信息组成，尤其是用户名。此外，男性的密码多样性较低。我们的分析表明，男性的密码比女性的更容易被破解。至少从与个人信息相关的攻击的角度来看，我们的观察与[15]中得出的结论不同，即男性的密码比女性略强。

III. 相关量化

虽然上面的统计数字显示了每种类型的个人信息和密码之间的相关性，但它们无法准确衡量个人密码中个人信息的参与程度。因此，我们引入了一个新的度量标准——覆盖率——以准确和系统的方式量化个人密码创建中涉及的个人信息。

A. 新闻报道

覆盖率的值范围从 0 到 1。覆盖范围越大意味着相关性越强，覆盖范围“0”表示密码中不包含个人信息，覆盖范围“1”表示整个密码完全匹配

只有一种个人信息。虽然覆盖率主要用于衡量单个密码，但平均覆盖率也反映了一组密码的相关程度。在下文中，我们描述了计算覆盖率的算法，并阐述了覆盖率的关键特征。

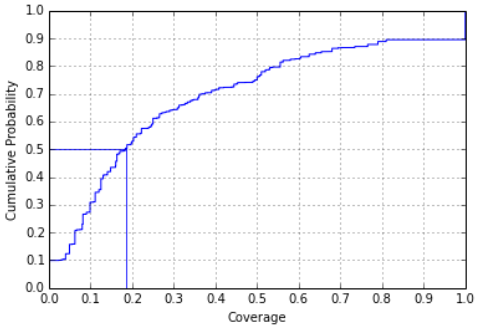
为了计算覆盖率，我们将字符串形式的密码和个人信息作为输入，并使用滑动窗口方法来进行计算。我们维护一个动态大小的窗口，从密码的开头滑动到结尾。窗口的初始大小是 2。如果窗口覆盖的段与某一类型的个人信息匹配，我们将窗口大小放大 1。然后，我们再次尝试将较大窗口中的段与个人信息进行匹配。如果找到匹配，我们进一步扩大窗口大小，直到出现不匹配。此时，我们将窗口大小重置为初始值 2，并将窗口滑动到导致前一个窗口不匹配的密码符号。同时，我们维护一个与密码长度相同的数组 tag array 来记录每个匹配密码段的长度。在我们滑动窗口通过整个密码字符串后，标记数组用于计算覆盖值——匹配密码段长度的平方和除以密码长度的平方。数学上我们有

$$CV\ G = \sum_{i=1}^{n} \left(\frac{li}{L}\right)^2 \quad (1)$$

其中 n 表示匹配密码段的数量，li 表示对应的匹配密码段的长度，L 是密码的长度。请注意，如果至少有一个 2 符号长的密码段与特定个人信息的子字符串匹配，则找到匹配项。然后，我们展示一个计算用户密码覆盖率的例子。爱丽丝，1988 年 8 月 16 日出生，密码“alice816！！”。我们在 Alice 上应用了覆盖计算算法。彻底滑动窗口后，标签数组为[5，5，5，5，3，3，0，0]。数组中的前五个元素，即 5，5，5，5，5，表示前五个密码符号匹配特定类型的个人信息(本例中为姓名)。数组中接下来的三个元素，即 3，3，3，表示这三个符号匹配特定类型的个人信息(本例中为出生日期)。数组中的最后两个元素，即 0，0，表示最后两个符号不匹配。根据等式 1 覆盖率计算如下:$CV\ G = \sigma 2\{\} \frac{1}{i=1} \frac{52+32}{L2} = \frac{0.34}{102}$。

覆盖范围独立于密码数据集。只要我们能建立一个完整的个人信息字符串列表，覆盖率就能准确量化一个用户的密码与其个人信息之间的相关性。对于相同长度的个人信息片段，覆盖强调匹配的连续性。连续的比赛比零散的比赛更强。也就是说，对于长度为 L 的给定密码，长度为 L(L ^ 1)的匹配段比两个匹配段具有更强的与个人信息的相关性
长度为 l1 和 l2 的线段，l = l1 + l2。例如，长度为 6 的匹配段预期具有更强的
长度为 3 的两个匹配段的相关性。覆盖的这一特征是所希望的，因为多个较短的片段(即，源自不同类型的个人信息)通常更难猜测，并且可能由于巧合而涉及错误的匹配。因为很难区分真实的

图 1:覆盖分布- 12306。



从一个巧合的匹配，我们想尽量减少错误匹配的影响，采取平方的匹配段计算覆盖有利于连续匹配。

B. *12306 的覆盖结果*

我们计算了 12306 数据集中每个用户的覆盖值，并将结果显示为图 1 中的累积分布函数。为了容易理解覆盖率的价值，我们讨论以
大约 0.2%的覆盖率。假设我们有一个 16 符号长的
个例子来说明
密码。一个长度为 5 的匹配段将产生 0.25 的覆盖率。长度为 3 的两个匹配段(即，总共 6 个符号与个人信息匹配)产生 0.18 的覆盖率。此外，长度为 2 的 5 个匹配段(即，所有符号都匹配，但是以分段的方式)产生 0.2 的覆盖范围。显然，覆盖率为 0.2 表明个人信息和密码之间有相当高的相关性。

用户覆盖范围的中值是 0.186，这意味着很大一部分用户密码与个人信息具有相对较高的相关性。此外，大约 10.5%的用户的覆盖率为 1，这意味着 10.5%的密码与一种类型的个人信息完全匹配。另一方面，大约 9.9%的用户没有覆盖，这意味着他们的密码中没有使用个人信息。

整个 12306 数据集的平均覆盖率为 0.309.我们还计算了男性和女性的平均覆盖率
女性群体，因为我们观察到男性用户更多
可能会在他们的密码中包含个人信息
第二部分-B4。男性群体的平均覆盖率为 0.314，女性群体的平均覆盖率为 0.269。这与我们之前的观察一致，并表明男性用户的相关性高于女性用户。反过来，这也表明覆盖率在量化密码和个人信息之间的相关性方面非常有效。

C. *覆盖率使用*

覆盖率对于构建密码强度表非常有用，据报道，密码强度表大多是临时的[7]。大多数计量器根据密码结构和长度或常用密码(例如，臭名昭著的“密码”)进行评分。还有一些血糖仪可以进行简单的社交资料分析，比如拒绝输入密码

当它包含用户名或帐户名称时。然而，这些简单的分析机制很容易被破坏，而密码仍然很脆弱。使用覆盖范围的度量标准，可以改进密码强度指示器，以更准确地测量密码的强度。此外，将覆盖率作为强度度量的一部分来实现是很简单的(只需要几行 Javascript 就可以了)。更重要的是，由于用户无法通过简单的篡改方法轻易破解覆盖测量，他们被迫选择更安全的密码。

覆盖率也可以集成到现有的工具中，以增强它们的功能。有几种基于马尔可夫模型的工具可以在用户创建密码时预测下一个符号[14]、[16]。这些工具根据从 dictionaries 或泄露的数据集学习的马尔可夫模型对下一个符号的概率进行排名，然后显示最可能的预测。因为大多数用户会惊讶地发现他们脑海中的下一个符号与工具的输出完全匹配，
他们可能会选择一个更不可预测的符号。

覆盖范围有助于确定个人信息预测在概率上是否排名足够高，以提醒用户避免在密码创建中使用个人信息。

IV. 个人-PCFG

在通过测量和量化研究了个人信息和用户密码之间的相关性之后，我们进一步研究了它们在从攻击者的角度破解密码方面的潜在用途。基于 PCFG 方法[13]，我们开发了个人 PCFG 作为一个面向个人的密码破解程序，它可以通过利用已知的个人信息对目标用户产生个性化的猜测。

A. 攻击场景

我们假设攻击者知道一定数量的关于目标的个人信息。攻击者可能是一个邪恶的邻居、一个好奇的朋友、一个嫉妒的丈夫、一个黑邮件发送者，甚至是一个从其他公司购买个人信息的公司。在这种情况下，通过亲自了解受害者或在网上搜索，特别是在社交网站(SNS) [17]，[18]，很容易获得有针对性的个人信息。个人 PCFG 可以用于离线和在线攻击。

在传统的离线密码攻击中，攻击者通常从受害系统中窃取哈希密码，然后试图找出这些密码的未哈希值。由于安全散列函数不能简单地被逆转，最流行的攻击策略是通过暴力猜测和验证密码。通过从密码字典中散列一个密码(需要添加 salt),并将结果与泄露的密码数据库中的散列值进行比较，来验证每个猜测。高概率密码猜测通常可以匹配密码数据库中的许多散列值，因此应该首先尝试。对于离线攻击，个人 PCFG 在猜测正确密码方面比传统方法快得多，因为它可以生成高概率的个性化密码并首先验证它们。

对于在线攻击，由于攻击者甚至没有哈希密码数据库，他或她会尝试登录通过猜测密码直接进入真实系统。在线攻击比离线攻击更难成功，因为在线服务系统通常在给定的时间段内对登录尝试有限制。如果在没有输入正确密码的情况下达到了尝试配额，则帐户可能会被锁定一段时间，甚至永久锁定，除非采取某些措施(例如，呼叫服务提供商)。所以网络攻击需要精准的猜测，通过整合个人信息就可以实现。个人- PCFG 能够在仅仅 5 次猜测内破解 20 个密码中的大约 1 个。

B. 重访 PCFG

个人-PCFG 是基于 PCFG [13]的基本思想，并提供了进一步提高其效率的扩展。在介绍个人 PCFG 之前，我们先简单回顾一下 PCFG 的原则。PCFG 预处理密码并生成基本密码结构，如“L D S”。

密码。从高概率结构开始，PCFG 方法使用从训练集中学习到的相同长度的段来替换“D”和“S”段。根据从训练集中学习到的出现概率对这些替代段进行排序。所以大概率段会先试。一个基础结构可以有多个
替换，比如“L5D3S1”可以有“L5123！”以及“L5691！”作为它的替代品。这些新的表现被称为前终末结构。目前没有“L”段
替换，因为 alpha 字符串的空间太大，无法从训练集中学习。接下来，从高概率到低概率对这些前置终端进行排序。最后，使用字典替换“L”段以产生实际的猜测。由于 PCFG 可以首先生成统计上的高概率密码，因此可以显著减少传统字典攻击的猜测次数。

C. 个人-PCFG

个人-PCFG 利用了 PCFG 的基本思想。除了 PCFG 的“L”、“D”和“S”符号，我们还增加了更多的语义符号，包括“B”代表生日，“N”代表姓名，“E”代表电子邮件地址，“A”代表帐户名，“C”代表手机号码，“I”代表身份证号码。更丰富的语义使得个人 PCFG 在猜测密码时更加准确。为了实现个人 PCFG，在原有的 PCFG 方法中增加了一个额外的个人信息匹配阶段和一个自适应替换阶段。因此，个人 PCFG 总共有 4 个阶段，每个阶段的输出将作为输入提供给下一个阶段。最后一个阶段的输出是尝试的实际猜测。我们现在用简单的例子详细描述每个阶段。

1) 个人信息匹配:给定一个密码字符串，我们首先将整个密码或密码的子字符串与其个人信息进行匹配。匹配算法类似于第二部分-B2 中的算法。然而，这次我们也记录匹配段的长度。我们用相应的符号替换密码中匹配的段，并用长度标记符号。不匹配的段保持不变。例如，我们假设爱丽丝出生于 1988 年 8 月 16 日，她的密码是“helloalice816！”。匹配阶段将“alice”替换为“N5 ”,将“816”替换为

《B3》。剩下的“你好”保持不变。因此，这一阶段的结果是“helloN5B3！”。

2) 密码预处理:这个阶段类似于原始 PCFG 的预处理例程；然而，基于个人信息匹配阶段的输出，已经匹配到个人信息的片段将不被
已处理。例如，示例结构“helloN5B3！”将在此阶段更新为“L5N5B3S1”。现在
口令由个人 PCFG 的语义符号充分描述，这一阶段的输出为个人 PCFG 提供了基础结构。

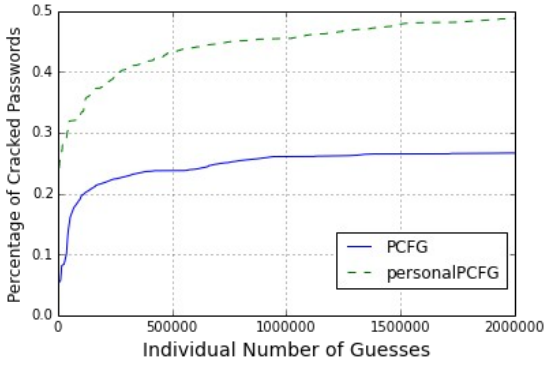
3) 猜测生成:与最初的 PCFG 类似，我们以递减的概率顺序，用从训练集中学习到的实际字符串重新放置“D”和“S”符号。“L”符号被字典中的单词替换。与 PCFG [13]类似，我们即时输出结果，因此我们不需要等待所有可能的猜测被计算和排序。请注意，我们没有替换任何代表个人信息的符号，因此猜测仍然不是真正的猜测。在此步骤中，我们不处理个人信息，因为每个用户的个人信息是不同的，并且个人信息符号只能在目标明确之前进行替换。因此，在这个阶段，我们的基本结构只生成前置词，前置词是部分猜测，包含部分实际猜测和部分个人 PCFG 语义符号。举个例子，
示例“L5N5B3S1”实例化为“helloN5B3！”如果“hello”是输入词典中第一个 5 个符号长的字符串，并且“！”在 1 个事件中发生的概率最高
训练集中的符号特殊字符。注意，对于“L”段，相同长度的每个单词具有相同的概率。“hello”的概率简单来说就是 1，其中 N 是输入字典中长度为 5 的单词总数。

4) 自适应替换:在原始 PCFG 中，guess 生成的输出可以应用于任何目标用户。然而，在个人 PCFG 中，猜测将进一步用个人信息来实例化，这些信息仅特定于一个目标用户。每个个人信息符号由相同长度的相应个人信息代替。如果有多个长度相同的候选项，则所有候选项都将
包括在试验中。在我们的例子“helloN5B3！”，“N5”会直接换成“爱丽丝”。然而，由于“B3”有许多
“19880816”的候选段和任何长度为 3 的子串都可能是候选项，猜测包括所有子串，如“helloalice198！”，“helloalice988！”，。。。，“helloalice816！”。然后，我们逐一尝试这些候选人的猜测，直到我们发现有一个候选人与 Alice 的密码完全匹配。注意，不是有多个候选人，而是不是所有的个人信息段都可以被替换，因为相同长度的段可能不总是可用的。例如，一个前
终端结构“helloN6B3！”不适用于 Alice，因为她的名字最多有 5 个符号长。在这种情况下，不应该为 Alice 生成来自该结构的猜测。

D. 裂化结果

我们使用具有 131，389 个用户的 12306 数据集来比较个人 PCFG 和原始 PCFG 的性能。我们使用数据集的一半作为训练集，另一半作为测试集。对于“L”段，两种方法都需要使用字典，这一点很关键

图 2：PCFG 对个人-PCFG(离线)。



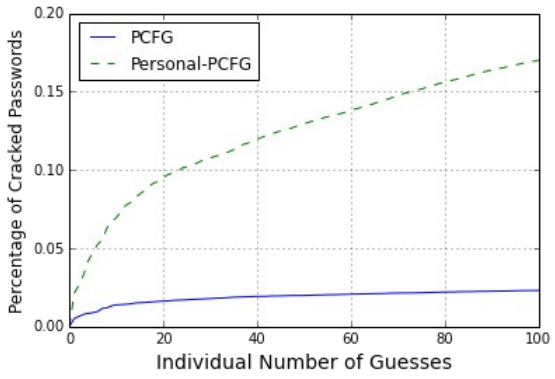
为了破解密码。为了消除不公平的词典选择的影响，我们在两种方法中都使用“完美的”词典。完美字典是我们直接从测试集中收集的字典，因此字典中的任何字符串都是有用的，密码中的任何字母段都必须出现在字典中。因此，一个完美的字典是有效找到正确的字母字符串的保证。在我们的研究中，PCFG 完美词典和个人 PCFG 完美词典都包含 15，000 到 17，000 个词条。

我们用个人猜测的次数来衡量个人 PCFG 的有效性，并与 PCFG 进行比较。个人猜测次数被定义为为破解每个个人账户而生成的密码猜测次数，例如，每个个人账户 10 次猜测尝试，这与密码数据集的大小无关。在个人 PCFG 中，个人猜测的总次数(即猜测的总次数)与密码数据集的大小成线性关系。相比之下，在像 PCFG 这样的传统破解策略中，每个猜测都应用于整个用户群，因此猜测的个体数量等于猜测的总数。尽管个人 PCFG 和传统破解方法之间存在这种差异，但密码破解的性能瓶颈在于大量的哈希运算。由于 salt 机制，对于个人 PCFG 和其他密码破解者，哈希的总数受 G N 的限制，其中 G 是个人猜测的次数，N 是数据集的大小。

给定不同的猜测次数，我们计算这些破解的密码在整个 pass- word 试验集中所占的百分比。图 2 显示了原始 PCFG 和个人 PCFG 在离线攻击中的比较结果。这两种方法都有一个快速的开始，因为它们总是先尝试高概率的猜测。图 2 清楚地表明个人 PCFG 破解密码的速度比 PCFG 快得多。例如，在 500，000 次猜测的中等规模下，个人 PCFG 实现了与原始 PCFG 的超过 2 亿次猜测所能达到的相似的成功率。此外，个人 PCFG 能够覆盖比 PCFG 更大的密码空间，因为个人信息提供了丰富的个性化字符串，这些字符串可能不会出现在字典或训练集中。

个人 PCFG 不仅提高了离线攻击中的破解效率，还提高了在线攻击中的猜测成功率。由于系统的原因，在线攻击只能在一定时间内尝试少量的猜测

图 3：PCFG 对个人-PCFG(在线)。



对登录尝试的限制。因此，我们将每个目标帐户的猜测次数限制为最多 100 次。我们在图 3 中展示了结果，说明个人 PCFG 能够破解比原始 PCFG 多 309%到 634%的密码。然后，我们在图 4 中展示了几个有代表性的猜测数字。对于一个允许 5 次尝试输入正确密码的典型系统，个人 PCFG 仅在 5 次猜测内就能破解 4.8%的密码。与此同时，最初的 PCFG 的成功率只有 0.9%，而 PCFG 要猜 2000 次才能达到 4.8%。因此，个人 PCFG 在少量猜测内破解密码更有效。

因此，由于将个人信息集成到密码猜测中，个人 PCFG 在在线和离线攻击方面都大大优于 PCFG。个人 PCFG 对个人信息的额外要求可以通过亲自了解受害者或在社交网站(SNS)上搜索来满足。

V. 讨论

A. 限制

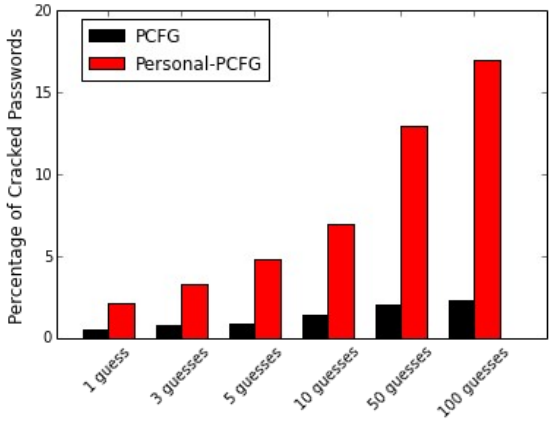
在这项研究中只使用了一个数据集。12306 网站的大部分用户是中国人，男女数量不平衡。因此，分析结果可能存在文化、语言和性别偏见。此外，覆盖率指标和个人 PCFG 的有效性仅在单个网站上得到验证。然而，与个人信息一起泄露的公开可用的密码数据集非常罕见。为了扩展这项工作，我们计划在未来从多个泄露的密码数据集中获取个人信息。

B. 潜在的防御

使用密码管理器可以缓解这个问题，因为用户不需要记住每个站点的密码。在这些网站密码的半自动创建中，引入了更多的随机性，涉及的个人信息更少。然而，用户的主密码仍然容易受到个人 PCFG 的攻击。

在创建密码时减少个人信息相关性的一个简单方法是由用户自己在心理上篡改密码。对现有密码进行简单的篡改就可以轻易地破坏个人信息的完整性和连续性。用户可以简单地选择这样的失真

图 4:代表点(在线)。



在每对现有密码字母/数字之间添加一个额外的符号(例如，字母、数字或特殊字符)。我们已经观察到，这种简单的变形能够显著降低覆盖范围的价值(即密码中的个人信息相关性)，并且个人 PCFG 变得无效。即使攻击者知道用户篡改了他们的密码，由于篡改的方式多种多样，并且学习密码中的个人信息模式的难度越来越大，所以仍然很难成功破解密码。还有其他解决方案来减少用户密码中的个人信息，如第 III-C 节中提到的个人信息感知密码计量器。但是，这些防御方法的有效性需要通过严格的安全分析和用户研究来充分验证，这是我们未来的工作。

VI. 相关著作

研究人员在测量现实生活中的密码方面做了出色的工作。在最早的作品之一[9]中，Morris 和 Thompson 发现密码非常简单，因此容易受到字典攻击。Malone 等人[3]研究了几个大型泄露数据集上的密码分布，发现用户密码很好地符合 Zipf 分布。Gaw 和 Felton [19]展示了用户如何管理他们的密码。Mazurek 等人[15]测量了一所大学的 25，000 个密码，并揭示了人口统计或其他因素之间的相关性，如性别和研究领域。Bonneau [2]从超过 7000 万个密码中研究了语言对用户密码的影响。通过测量 1100 多名银行客户的 4 位 pin 码的可猜测性[20]，Bonneau 等人发现出生日期广泛出现在 4 位 pin 码中。李等人[10]对中文密码进行了大规模的测量研究，研究了超过 1 亿个真实生活中的密码，并给出了中文密码与其他语言密码的差异。

有几部作品在研究密码的特定方面。Yan 等人[6]和 Kuo 等人[21]研究了基于助记符的密码。Veras 等人[5]展示了密码中日期的重要性。Das 等人[22]研究了用户如何为不同的网站篡改一个密码。Schweitzer 等人[11]研究了密码的键盘模式。除了密码本身，人们对密码安全的习惯和心理也进行了研究[23]。

已经证明 NIST 熵不能准确描述密码的安全性[24]。Bonneau 等人[2]，[20]使用的 α 猜测和 β 成功率被认为是衡量密码数据库强度的更准确的指标。其他研究人员也使用这些指标[10]。

密码破解已经被研究了三十多年。攻击者通常试图从哈希密码数据库中恢复密码。虽然反转哈希函数是不可行的，但早期的工作发现密码容易受到字典攻击[9]。然而，近年来随着密码策略变得更加严格，简单的字典密码变得不那么常见了。Narayanan 和 Shmatikov [4]基于密码需要在语音上与用户的母语相似的事实，使用马尔可夫模型来产生猜测。2009 年，Weir 等人[13]利用概率上下文无关文法(PCFG)破解密码。Veras 等人[12]试图在密码中使用语义模式。OMEN+ [25]改进马尔可夫模型[4]破解密码。它甚至包括证明个人信息在密码破解中有用的实验。然而，他们的实验是在基于马尔可夫模型的小得多的范围内进行的，并且改进是有限的。

已经有通过强制用户选择更安全的密码来保护密码的研究，其中密码强度计似乎是一种有效的方法。Castelluccia 等人[26]建议使用[4]中的马尔可夫模型来衡量用户密码的安全性。与此同时，流行网站所采用的商业密码计量器被证明是不一致的[7]。有一些工作专注于使用经过训练的泄露密码或字典向用户提供反馈[14]、[16]。

VII. 结论

在本文中，我们对用户个人信息如何驻留在人类选择的密码中进行了全面的定量研究。据我们所知，我们是第一个系统分析密码中个人信息的人。我们有一些有趣的定量发现，比如 12306 数据集中有 3.42%的用户在密码中使用生日，男性用户比女性用户更有可能在密码中包含自己的名字。然后，我们引入一个新的度量标准，覆盖率，来精确地量化个人信息和密码之间的相关性。我们基于覆盖率的量化结果进一步证实了我们关于个人信息在密码创建中的严重参与的披露，这使得用户密码更容易受到有针对性的密码破解。我们在 PCFG 的基础上开发了个人 PCFG，但考虑了更多破解密码的语义符号。个人-PCFG 通过将用户个人信息整合到猜测中来生成个性化的密码猜测。我们的实验结果表明，个人 PCFG 在密码破解方面明显快于 PCFG，并降低了发动在线攻击的可行性。最后，我们讨论了这项工作的局限性和解决方案，以防止包括个人信息的弱密码。

感谢

这项工作得到了美国 ARO 拨款 W911NF-15-1-0287 和 ONR 拨款 N00014-15-1-2396 和 N00014-15-1-2012 的部分支持。

参考

[1] J.Bonneau 、C. Herley 、P. C. Van Oorschot 和 F. Stajano，“寻求替换密码:网络认证方案的比较评估框架”，IEEE 安全与隐私，2012 年。

[2] J.Bonneau，“猜测的科学:分析 7000 万密码的匿名语料库”，IEEE 安全与隐私，2012 年。

[3] D.Malone 和 K. Maher，“调查密码选择的分布”，载于 ACM WWW，2012 年。

[4] A.Narayanan 和 V. Shmatikov，“使用时空折衷对密码的快速字典攻击”，美国计算机学会 CCS，2005 年。

[5] R.Veras，J. Thorpe 和 C. Collins，“密码中的可视化语义:日期的作用”，IEEE VizSec，2012 年。

[6] J.严，a .布莱克威尔，r .安德森，a .格兰特，“密码记忆性和安全性:实证结果”，IEEE 安全与隐私杂志，2004 年。

[7] X.de Carne de Carnavalet 和 M. Mannan，“从非常弱到非常强:分析密码强度表”，NDSS，2014 年。

[8] 南埃格尔曼、a .索蒂拉科普洛斯、I .穆斯卢霍夫、k .别兹诺索夫和
C. 赫利，“我的密码会升到 11 吗？:密码计量器对密码选择的影响”，发表在《计算系统中人的因素的 SIGCHI 会议论文集》中。美国计算机学会，2013 年。

[9] R.莫里斯和 k .汤普森，“密码安全:案例史”
美国计算机学会通讯，1979 年。

[10] Z.李，韩文伟，徐文伟，“中国网络密码的大规模实证分析”，在 Proc。USENIX 安全局，2014 年。

[11] D.Schweitzer，J. Boleng，C. Hughes 和 L. Murphy，“可视化键盘模式密码”，IEEE VizSec，2009 年。

[12] R.Veras、C. Collins 和 J. Thorpe，“密码的语义模式及其安全影响”，NDSS，2014 年。

[13] 米（meter 的缩写））Weir，S. Aggarwal，B. De Medeiros 和 B. Glodek，“使用概率上下文无关语法的密码破解”，IEEE 安全与隐私，2009 年。

[14] 米（meter 的缩写））Weir，S. Aggarwal，M. Collins 和 H. Stern，“通过攻击大量暴露的密码来测试密码创建策略的指标”，载于 ACM CCS，2010 年。

[15] 米（meter 的缩写））L. Mazurek、S. Komanduri、T. Vidas、L. Bauer、N. Christin、L. F. Cranor、P. G. Kelley、R. Shay 和 B. Ur，“衡量整个大学的密码可猜测性”，载于 ACM CCS，2013 年。

[16] 南 Komanduri 、R. Shay 、L. F. Cranor 、C. Herley 和 S. Schechter，“心灵感应词:通过读取用户的思想来防止弱密码”，载于《USENIX 安全》，2014 年。

[17] R.Gross 和 A. Acquisti，“在线社交网络中的信息披露和隐私”，WPES ACM，2005 年。

[18] B.Krishnamurthy 和 C. E. Wills，“关于通过在线社交网络泄露个人身份信息”，载于 ACM COSN，2009 年。

[19] 南 Gaw 和 E. W. Felten，“在线账户的密码管理策略”，载于 ACM SOUPS，2006 年。

[20] J.Bonneau，S. Preibusch，R. Anderson，“每 11 个钱包一份生日礼物？客户选择的银行 pin 的安全性”，金融加密和数据安全。斯普林格，2012。

[21] C.Kuo，S. Romanosky 和 L. F. Cranor，“基于短语的密码助记法的人类选择”，载于 ACM SOUPS，2006 年。

[22] A.Das、J. Bonneau、M. Caesar、n .和 X. Wang，“密码重复使用的纠结网络”，2014 年。

[23] D.弗洛伦西奥和 c .赫利，“网络密码习惯的大规模研究”，载于美国计算机学会网站，2007 年。

[24] 页（page 的缩写）G. Kelley，S. Komanduri，M. L. Mazurek，R. Shay，T. Vidas，L. Bauer，
名词（noun 的缩写）Christin、L. F. Cranor 和 J. Lopez，“再次猜测(一次又一次):通过模拟密码破解算法测量密码强度”，IEEE 安全与隐私，2012 年。

[25] C.Castelluccia，A. Chaabane，M. Du rmuth 和 D. Perito，“当隐私与安全相遇:利用个人信息破解密码”，arXiv 预印本 arXiv:1304.6584，2013 年。

[26] C.来自马尔可夫模型的自适应密码强度计。2012 年在 NDSS。