密码中的齐夫定律

和简，北京大学
福建师范大学黄心怡
王萍，北京大学

尽管三十多年的密集研究努力，文本密码仍然笼罩在神秘的面纱。在这项工作中，我们在理解密码的基本分布方面向前迈进了一大步。通过对 9720 万个密码(大量混沌数据)的语料库进行线性回归，我们首次表明 Zipf 定律完美地存在于用户生成的密码中，找出了相应的精确分布函数，并研究了我们的观察结果对密码策略和基于密码的加密协议 (如认证、加密和签名)的一些基本影响。

作为这一自然规律的一个具体应用，我们提出了回归中使用的唯一密码的数量和回归线斜率的绝对值一起作为评估密码数据集强度的度量，并以数学上严格的方式证明了其正确性。此外，大量的实验(包括最优攻击、模拟最优攻击和最新的破解会话)证明了该方法的有效性。在四种情况中的两种情况下，我们的度量在简单性方面优于 Bonneau 的 α-猜测，并且就知识而言，它是第一个既容易近似又准确以方便比较的度量，为安全管理员提供了一个有用的工具，以精确掌握他们的密码数据集的强度并更合理地调整密码策略。

类别和主题描述符:C.4.6【操作系统】:安全和保护——认证通用术语:理论、安全、度量

附加关键词和短语:互联网安全、密码、基于密码的协议；齐夫定律

ACM 参考格式:
王博士，简，黄，x，王，2015 年。密码中的 Zipfs 定律。ACM Trans 信息。系统。秘书 1、1、第一条(2015 年 1 月)，33 页。
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. 介绍

用户认证是网络系统保护资源和服务免受未授权访问的最基本的安全机制。尽管已经有很多关于其缺陷的报道，文本密码仍然是用户认证的主要机制，保护着互联网规模的服务提供商的数亿个账户。最近，在提出替代方案(例如，图形密码[Zhu et al. 2014]，单一登录[Sun et al. 2013] 和多因素身份验证[Huang et al. 2014]) 移除密码，但密码比以往任何时候都被更广泛地使用和牢固地确立。由于密码提供了许多其他替代身份验证方案无法比拟的优势[Bonneau et al. 2012；Zhaoet al. 2015；Wang et al. 2014] 此外，替换它们的过渡成本无法有效量化[Herley and Van Oorschot 2012]，在可预见的未来，它们很可能会继续存在并主导认证系统。

作者地址:北京大学电子工程与计算机科学学院，北京 100871；简，北京大学数学科学学院，北京 100871；黄，福建师范大学数学与计算机科学学院，福州 350007。邮箱:wangdingg，demscimath @ pku . edu . cn；xyhuang81@gmail.com；pwang@pku.edu.cn。

DOI:http://dx.doi.org/10.1145/0000000.0000000

尽管它无处不在，但基于密码的认证伴随着生成密码的困境，强大的攻击者很难破解密码，但普通用户很容易记住密码。用户很难记住真正随机的密码，而用户选择的密码可能是高度可预测的[Yan et al. 2004].实际上，用户倾向于使用与其日常生活相关的弱密码(例如，生日、电话号码、爱人、朋友和宠物名)[Brown et al. 2004; Florencioand Herley 2007]，这意味着这些密码来自一个相当小的字典，因此容易受到离线/在线猜测攻击。

为了减轻这种臭名昭著的“安全性-可用性”困境，已经提出了各种密码创建策略，例如随机生成[Yan et al. 2004]，基于规则[Bishopand V Klein 1995; Schechter et al. 2010]，基于熵的[Burr et al. 2006; Burr et al. 2013]和基于裂化的[Houshmand and Aggarwal 2012; Castelluccia et al. 2012].它们强制新创建的密码遵守一些规则并达到可接受的强度。密码强度指示器和规则的多样性带来了不同网站之间的大量不同要求，导致对相同密码的高度冲突的强度输出。例如，密码$1 被 Dropbox 认为“非常弱”，被 Google 认为“一般”，被 Yahoo!

密码强度的上述矛盾结果(有关更具体的示例，请参见[de Carnavalet and Mannan 2014; Wang and Wang 2015]) 是不同网站之间采用的不一致的密码强度度量的直接结果，这可以部分地由每个网站的不同兴趣来进一步解释。人们普遍认为，更严格的政策可能会使密码更难破解，但副作用是用户可能会觉得更难创建和记住密码，从而降低了可用性[Ur et al. 2012].该作品由[Adams and Sasse 1999; Inglesant and Sasse 2010] 还报告称，在特定的使用环境中，不适当的密码政策会增加用户的精神和认知负荷，并对他们的工作效率产生负面影响，最终他们会想尽一切办法来规避这种不友好的政策。

因此，不同类型的应用程序系统通常有非常不同的偏好。对于易贝这样的电子商务网站，雅虎这样的门户网站像 Kasper- sky 这样的订单接受网站，可用性是非常重要的，因为每个登录事件都是一个收入机会。任何破坏用户体验的事情都会损害企业的成功。所以他们倾向于使用限制较少的密码策略[Flore^ncio and Herley 2010]. 另一方面，防止攻击者非法访问安全关键站点上的宝贵资源也非常重要，例如保存敏感文档的云存储站点(例如 Dropbox)和管理课程成绩的大学站点。因此，它们可能要求用户选择的密码受到更复杂的约束(例如，包括混合大小写、数字和特殊字符，并拒绝 pa$$word123 等常用密码)。

由于不同的系统喜欢不同的密码策略，出现了许多关键问题:密码策略设计者如何评估他们的策略？管理员如何为他们的系统选择正确的策略？此外，通常系统的用户(以及它的服务)可能随着时间的推移而动态地改变，这高度导致在一段时间(例如，一年)之后密码数据集中的大的变化，即使密码策略[1]保持不变，尤其是对互联网规模的服务提供商而言。在这种情况下，安全管理员应量化密码的强度，并可能需要调整密码策略。无论是没有注意到密码数据集中的变化，还是采取了不适当的对策，都可能导致严重的(但微妙的)安全性和可用性问题，如上所示。因此，对密码数据集的强度进行适当的评估是必不可少的，没有这种评估，安全管理员就无法确定以下重要问题:应该如何调整密码策略？或者同样地，密码策略应该被增强以提高安全性，保持不变或者甚至放松一点以获得可用性作为回报吗？一言以蔽之，核心症结

1 在本文中，术语“密码策略”和“密码创建策略”将互换使用，而关于锁定和过期的策略[InChiasson and van Oorschot 2015]超出了本文的范围。

设计和选择一个合适的口令策略或对其进行适当调整的关键在于如何准确评估在该策略下创建的口令数据集的强度。注意，在这项工作中，我们假设每个现有的认证系统已经采用了某种密码策略(例如，[Houshmand and Aggarwal 2012; Castelluccia et al. 2012])，其调整主要涉及更改部分规则和密码强度阈值。

1.1. 动机

令人惊讶的是，据我们所知，现有文献并没有对上述如何准确测量给定密码数据集强度的问题提供满意的答案。评估密码数据集强度的两种最常用方法是理论上测量其信息熵(例如[Burr et al. 2006])并根据经验估计其“可猜测性”(例如[Kelley et al. 2012; Mazurek et al. 2013]).然而，前者不是基于经验数据，而且已经证明是不准确的[Weir et al. 2010]，而后者在很大程度上取决于破解算法、参数和输入字典的选择[Ma et al. 2014; Dell'Amico et al. 2010]，有太多的不确定性，无法准确表征给定数据集的强度。后来，Bonneau [2012b]引入了一个巧妙的基于统计的度量标准 G α ()(名为 α-猜测)，其根据攻击者期望的成功率 α 来参数化。这个指标是准确，但本质上是不确定的。例如，G0.50(A)的关系

> G0.50(B)永远无法保证 G0.49(A) > G0.49(B)，这里 A 和 B 是两个密码数据集。这意味着除非计算出完整的 α 猜测曲线(x 轴范围为[0, α])，否则无法得出明确的结论。未能抓住这一微妙之处可能会

引起巨大的误解，就像在[Li et al. 2014].这种不确定性破坏了 α 猜测的简单性。幸运的是，在这项工作中，我们开发了一个简单、准确和确定的(在四分之二的情况下)统计指标。

不可避免地，准确评估密码数据集的强度需要解决一个更基本的问题:如何精确地描述给定密码数据集的特征？或者同样，现实生活中用户生成的密码遵循什么分布？尽管经过了三十多年的深入研究，文本密码仍然笼罩在神秘的面纱中，这个老问题年复一年地被提出，这可以很好地解释为什么今天大多数具有可证明安全性的密码认证密钥交换(PAKE)协议(数以千计，一些著名的包括[Chen et al. 2014; Canetti et al. 2012] 在随机预言模型中[Katz and Vaikuntanathan 2013; Halevi and Krawczyk 1999] 在标准模型中)仍然依赖于一个简单但不可思议的假设:密码遵循均匀分布。

据所知，该工作由 Malone and Maher [2012]可能与我们将在本文中讨论的内容最相关。他们最初试图调查密码的分布，并得出结论，他们的密码数据集“实际上不太可能是 Zipf 分布的”。这样的结论与我们将在当前工作中展示的正好相反。Malone and Maher [2012] 还得出结论，“Zipf 分布与用户选择密码的频率是相对较好的匹配”。有点自相矛盾？关键是，他们使用一种天生有缺陷的方法，试图用 Zipf 对密码分发进行建模(自然，他们失败了)，他们将他们的模型与统一模型进行比较，他们的比较结果显示他们的模型是“相对较好的匹配”。由于几乎任何模型都会优于统一模型，因此他们的模型“相对较好”的结论几乎没有意义。这种令人困惑、不尽如人意的情况激发了这项工作。

1.2. 我们的贡献

在这项工作中，我们通过采用统计技术将对现实生活中的密码的理解和密码数据集的评估带到了一个合理的科学基础上，并试图提供关于上述两个基本问题的令人信服的答案:(1)密码(用户生成的)的基本分布是什么？ 以及(2)如何精确测量给定密码数据集的安全强度？

作为我们的主要贡献，我们采用来自计算统计学的技术来表明 Zipf 定律完美地存在于现实生活的密码中，这是受 Zipf 定律的适用性的启发，以描述令人惊讶的各种自然和社会现象，如互联网拓扑[Faloutsos et al. 1999]和美国公司规模[Axtell 2001].我们删除最不频繁的密码，按降序排列每个剩余唯一密码的频率，并通过使用线性回归调查频率和等级之间的数学关系。在 12 个密码数据集(在服务、大小、泄露方式、本地化和语言方面有很大不同)的大规模语料库上的广泛实验表明，我们的 Zipf 模型能够准确地表征现实生活中密码的分布。换句话说，每个密码数据集都是从底层密码群体中抽取的特定样本，完全遵循 Zipf 定律。这使得 1997 年提出的索赔无效[Malone and Maher 2012] 用户密码“实际上不太可能被 Zipf 分发”。特别地，我们证明了密码的前端自然地遵循 Zipf 定律，而不是他们工作中所报道的“密码选择的长尾”。然后，我们弄清楚为什么会出现这样截然相反的观察结果，为什么我们的方法是至关重要的。此外，我们证明了我们的观察的普遍适用性，强调了密码策略和基于密码的加密协议的基本含义，并对 Wang et al. [2014].这是对第一个问题的有力回答。

我们的第二个贡献是一种新颖的度量，它利用了密码分布函数的具体知识，因此它克服了现有度量中的各种问题(例如，基于破解的方法中的不确定性[Kelley et al. 2012]和 α-猜测中的非确定性 Bonneau 2012b])。我们的指标有助于密码策略设计者和安全管理员以数学上严格的方式简明掌握他们的密码数据集(纯文本或散列形式)的强度，并使他们能够精确地评估正在检查的密码策略的安全属性。这表明了第二个问题的解决。

本文的另一个贡献是通过经验证据展示了我们度量密码数据集强度的方法的有效性。首先，我们在收集的真实密码数据集上模拟最优猜测攻击。然后，我们向前迈出一步，采用最先进的破解算法(即，基于马尔可夫的[Maet al. 2014])来近似最优密码破解攻击。独立感兴趣的可能是我们的观察，在某些情况下，与最佳破解结果相比，基于马尔可夫的破解成功率要低得多，这意味着最先进的破解算法远非最佳算法，并且还有很大的未来改进空间。此外，我们报告了 α 猜测的强度转换中的固有缺陷[Bonneau 2012b]并设法解决它。

路线图。在截面中 2，我们调查相关作品。然后，我们在第二节中证明了密码中存在 Zipf 定律 3。我们观察到的一些基本含义将在第二节中讨论 4。密码数据集强度指标在第节中进行了展示、证明和实证 5、和部分 6 总结论文。

2. 相关著作

在这一节中，我们简要回顾一些关于密码创建策略和密码破解技术的相关工作，为后面的讨论提供一些背景知识。

2.1. 密码创建策略

1990 年，Klein 提出了主动口令检查器的概念，它使用户能够创建口令并检查新口令是否“安全”[Klein 1990].该标准可以分为两种类型。一类是构成可接受密码的确切规则，如最小长度和字符类型要求。另一种类型是基于估计的密码强度使用拒绝功能。这方面的一个例子是不允许的“弱”密码黑名单。尽管作者将这种技术称为“主动密码检查”，但它确实与我们今天所知的密码创建策略相同，因此在本文中，我们可以互换使用这两个术语。

自从Klein的开创性工作以来，已经提出了许多主动式密码检查器，旨在减少将新创建的密码与“弱”密码黑名单进行匹配的时间和空间，例如Opus [Spafford 1992b]和ProCheck[Bergadano et al. 1998].也有尝试在每个站点的基础上设计可调整的规则来塑造密码创建，其中有有影响力的NIST电子认证指南SP-800-63[Burr et al. 2006]. 然而，通过针对在不同规则下创建的真实用户密码来模拟当前密码破解技术的成功率，Weir et al. [2010] 表明仅仅基于规则的策略在确保理想的安全级别方面表现不佳。在Weir等人工作的基础上，Houshmand and Aggarwal [2012]提出了一种新的策略，该策略首先根据基于经验的破解结果分析用户选择的密码是弱密码还是强密码，如果密码是弱密码，则稍微修改密码以创建加强密码。该策略有助于更准确地测量个人密码的强度，此外，由于其调整仅涉及在连续范围内调整阈值，因此可以比以前的策略更灵活地调整该策略。也许最相关的政策与我们评估密码数据集的强度指标有关(参见第5)是由建议的Schechter et al. [2010].他们耐人寻味的想法是使用一个流行的甲骨文来取代传统的密码创建策略，从而拒绝高流行的密码。这种策略在挫败针对具有数百万用户帐户的互联网规模认证系统的基于统计的猜测攻击方面特别有效。如果这一政策到位，我们提出的指标将是不必要的。然而，如何防止攻击者利用他们的神谕来猜测呢密码是一个公开的问题。此外，该策略拒绝出现概率超过阈值(例如，如中所示= 16)的密码[Schechter et al. 2010])，然而它是否会大大降低可用性还没有得到彻底的评估(例如，没有实际的用户案例研究报告)。这项政策的直接后果是
可能会经常禁止用户使用他们想要的通常很流行的密码，从而惹恼用户。例如，大约34.89%的用户www.tianya.cn使用频率大于等于16的密码，这表明超过三分之一的用户有同样的可能对选择和维护新密码感到恼火。
然而，如果这些问题能够得到解决，这种政策将大有希望。

2.2. 密码破解

基于密码的系统容易受到各种攻击，如在线猜测、离线猜测、键盘记录、肩窥和社会工程[Long 2011; Herley 2013].这里我们只考虑在线和离线猜测攻击，其他攻击与密码强度或密码数据集强度无关，因此不在本文的讨论范围之内。虽然非加密技术可以很好地阻止在线猜测，例如在登录失败达到阈值次数后锁定帐户，或者使用更灵活的锁定策略[Van Oorschot and Stubblebine 2006; Alsaleh et al. 2012]，离线猜测攻击是在攻击者控制的本地硬件上执行的，因此，如果有足够的时间和计算能力，她可以进行尽可能多的猜测。Florencio et al. [2014] 讨论了离线猜测构成真正威胁的场景，并确定了密码对这两种猜测的猜测抵抗力之间的巨大“鸿沟”。他们发现，在这个“鸿沟”中，逐步增加密码的强度几乎不会带来什么安全好处，因此他们对促使用户使用更强密码而不是在线猜测的常见做法提出了质疑。然而，不难看出，这样的“鸿沟”将在很大程度上被消除(并且相应的怀疑也是如此)，如果考虑口令(例如，在salted-hash中)已经被泄露，但是这种泄露仅在一段时间(例如，几天)之后才被检测到(并且被处理)的情况，在此期间离线猜测确实造成了现实的威胁。

因此，对于基于口令的认证系统来说，正确评估其对离线猜测攻击的弹性是至关重要的。在文献中，这通常是通过将搜索空间大小(即猜测的次数)与百分比进行比较来完成的

可以离线恢复的哈希密码。这种措施只取决于攻击技术和用户选择密码的方式，与认证系统的特定性质无关(例如，使用哪种类型的散列函数，PBKDF2 还是 SHA-1？)也不受攻击者能力的影响。相反，系统的性质和攻击者的能力将决定攻击者为每一次猜测所付出的代价[Dell' Amico et al. 2010]. 例如，针对离线攻击的系统对策，如利用加盐来挫败预计算技术(例如，彩虹表[Oechslin 2003]) 或关键强化[Du¨ rmuth 2013]为了让猜测攻击付出更大的代价，在评估密码系统对离线攻击的弹性时，只构成一个关键参数。通过将该成本与搜索空间的度量相结合，可以获得对离线攻击的具体成本效益分析。这种措施也是我们工作遵循的。

密码搜索空间本质上取决于用户如何选择他们的密码。它众所周知，用户倾向于选择密码(例如，来自字典的单词，如著名的“dict-0294 ”)[Outpost9.com's Lab 2014] 或与他们的日常生活相关的东西)Shay et al. 2010; Florencio and Herley 2007].然而，用户很少使用来自这种列表的未修改的元素，例如，因为密码创建策略阻止这种做法，相反，用户以他们仍然可以容易地回忆起它们的方式来修改单词。例如，流行的密码 pa$$word 是通过输入两个容易猜到的字符串密码的字母生成的。

为了模拟这种密码生成实践，研究人员使用了各种启发式方法从像“dict-0294”这样的输入词典中产生单词变体的 gling 规则[Out-post9.com's Lab 2014], 这种技术早在 1979 年莫里斯·汤普森对 3000 个密码的分析中就出现了[Morris and Thompson 1979]. 这一初步工作之后，由独立的作品 Klein [1990]和 Spafford [1992a]. 后来，一些专用的软件工具，如开膛手约翰[Designer 1996] 出现了。后续研究(例如[Kuo et al. 2006; Dell'Amico et al. 2010]) 经常利用这些软件工具来执行字典攻击作为次要目标。

直到最近，密码破解才开始偏离艺术科学。Narayanan and Shmatikov [2005]开发了一种高级破解算法，该算法使用马尔可夫链而不是特定的混乱规则来模拟用户密码创建模式。这种算法生成的密码在发音上与单词相似。在 142 个散列密码的数据集上测试，96 个(67.6%)密码被成功破解。然而，他们的算法不是标准的基于字典的攻击，因为它只能产生语言上可能的密码。此外，测试数据集过于有限，无法令人信服地显示其有效性。

在 2009 年，基于概率上下文无关文法(PCFG)，Weir et al. [2009]提出了一种新的自动推导单词拼写规则的技术，他们进一步使用大型真实数据集来测试其有效性。在这种技术中，密码被认为是字母符号(用 L 表示)、数字(D)和特殊字符(S)的组合。例如，pa$$word123 由 L2S2L4D3 表示。然后，从一组训练的明文密码中获得一组单词篡改规则。为了模拟最佳攻击，该算法以概率递减的顺序产生猜测，并且它能够比开膛手约翰多破解 28%到 129%的密码[Designer 1996].2014 年，Ma et al. [2014]发现，当使用正确的顺序进行调整并采用一些方法来处理数据稀疏性和规范化问题时，基于马尔可夫链的破解算法将比基于 PCFG 的破解算法执行得更好。因此，在这项工作中，我们遵循马等人的基于马尔可夫的算法来破解收集的数据集，并根据提出的度量进行比较。

3. 现实生活中密码的齐夫定律

在本节中，我们首先介绍一些统计技术——线性回归的背景知识，然后描述收集的数据集。此外，我们提供了对密码的基本理解，并表明 Zipf 定律完美地存在于现实生活的密码中。

3.1. 线性回归

在统计学中，线性回归是一种通过将线性方程拟合到观察数据来模拟两个变量之间关系的方法。一个变量被认为是解释变量，另一个被认为是因变量。通常，线性回归指的是一种模型，其中给定 x 的值，条件
y 的均值是 x 的仿射函数:y = a + b x，其中 x 是解释变量
y 是因变量。直线的斜率是 b，a 是截距。

拟合回归线最常用的方法是使用最小二乘法。该方法通过最小化每个数据点到线的垂直偏差的平方和来计算观察数据的最佳拟合线。例如，如果一个点正好位于拟合线上，则它的垂直偏差为 0。更具体地说，从实验中我们收集了一堆数据:(xi，易)，1 ≤ i ≤Σ N .我们期望 y = a + b x + ε，其中

x 是 xi 的算术平均值，对于 y 也是如此

a = y bx，其

误差ΣN(yi ab Xi)2 降至最低。在回归中，决定系数

(用 R2 表示，范围从 0 到 1)是回归线接近真实数据点的程度的统计度量:越接近 1 越好。R2 值为 1 表示所有数据点完全位于回归线上。

3.2. 密码数据集的描述

在我们的工作中，我们收集了 12 个大规模的真实生活密码列表，它们在服务、大小、如何泄露、用户本地化、语言和文化(信仰)背景方面不同，表明我们的模型能够准确地表征真实生活密码的分布。表中总结了所有 12 个数据集 I，被黑客攻破或被匿名内部人士泄露，随后在互联网上公开披露。其中一些还被许多研究密码的科学著作使用(例如[Weir et al. 2010; Komanduri et al. 2014; Ma et al. 2014]).我们意识到，虽然这些数据集是公开的，但其中包含电子邮件、用户名和密码等私人数据。因此，我们将所有用户名视为机密，仅报告有关密码的汇总信息，以便在我们的研究中使用它们不会增加对受害者的伤害。此外，攻击者很可能利用这些帐户作为训练集或破解字典，而我们对它们的研究对安全管理员和普通用户保护他们的帐户具有实际意义。

表一.关于 12 个数据集的基本信息(“PWs”代表密码)

资料组	服务位置		语言	泄露时	如何泄露	总 PWs	独特的 PWs
天涯	社会论坛	中国	中国人	2011 年 12 月 4 日	黑客突破了 30233633		12,614,676
多多新	游戏和电子商务编程游戏社交论坛约会作家论坛黑客论坛	中国 中国中国美国美国美国美国	中国人	2011 年 12 月 3 日	黑客攻破了 16231271		11,236,220
CSDN			中国人	2011 年 12 月 2 日	黑客入侵了	6,428,287	4,037,610
多万			中国人	2011 年 12 月 1 日	知情人透露	4,982,740	3,119,070
聚友网（网站）			英语	2006 年 10 月 1 日	网络钓鱼攻击	41,545	37,144
Single.org			英语	2010 年 10 月 1 日	查询字符串注入 16，250		12,234
信仰作家			英语	2009 年 3 月 1 日	SQL 注入	9,709	8,347
Hack5			英语	2009 年 7 月 1 日	黑客入侵了	2,987	2,351
摇滚你	赌博	美利坚合众国	英语	2009 年 12 月 7 日	黑客入侵了	32,603,388	14,341,564
美国 Yahoo 公司(提供互联网的信息检索服务)	网络门户	美利坚合众国	英语	2012 年 7 月 12 日	SQL 注入	453,492	342,515
Mail.ru	电子邮件	俄罗斯	俄语	2014 年 9 月 9 日	网络钓鱼和恶意软件 4，938，663		2,954,907
Yandex.ru	搜索引擎俄罗斯		俄语	2014 年 9 月 9 日	网络钓鱼和恶意软件 1，261，810		717,203

前四个数据集，即天涯、多多网、CSDN 和多玩，都来自中文网站。我们根据相应网站的域名来命名每个密码数据集(例如，“天涯”数据集来自 www.tianya.cn)。由于 2011 年 12 月在中国发生的几起安全事件，这些文件在互联网上都是公开的[Martin 2012]而我们当时就收藏了。CSDN 是中国程序员最大的社区网站；天涯是有影响力的中文 BBS 多玩是热门游戏论坛；Dodonew 也是一个受欢迎的游戏论坛，它支持

货币交易。除了 Duowan 数据集的一部分，所有的密码都是明文。Duowan 包含哈希(MD5)和纯文本密码，我们将我们的分析限制在 498 万个纯文本密码。

第五个数据集是最初于 2006 年 10 月发布的“Myspace”。Myspace 是美国著名的社交网站，其密码被攻击者破解，攻击者建立了一个虚假的 Myspace 登录页面，然后对用户进行标准的社交工程(即网络钓鱼)攻击。虽然 Myspace 数据集有几个版本，但由于不同的研究人员在不同的时间下载了这个列表，我们从[Bowes 2011]其中包含 41，545 个明文密码。以下两个数据集是“Singles.org”和“Faithwriters”。它们都是由几乎完全信仰基督教的人组成的:www.singles.org 是一个约会网站，表面上是为基督徒和 www.faithwriters.com 是一个面向基督徒的在线写作社区。前者通过查询字符串注入被攻破，16250 个密码被泄露，而后者则因 SQL 注入攻击而泄露了 9709 个密码。

第八个数据集来自 www.hak5.org 它被一个叫 ZFO 的组织攻破了[Constantin 2009].这个数据集只是整个数据集的一小部分 www.hak5.org 数据集。令人惊讶的是，尽管 Hak5 声称是“喜剧、技术、黑客、自制软件、法医和网络安全的鸡尾酒混合物”，但它的数据集却是最弱的(参见第 5.1)的数据集。在这项工作中，我们使用这个数据集作为现实生活中密码分布代表的反例。

除了上述八个数据集之外，我们还采用了四个数据集(即 Rockyou、Yahoo、Yandex.ru 和 Mail.ru)来显示我们在第节中的 Zipf 定律的发现的可推广性 3.4 和 3.5，而且由于篇幅所限，在别处就不分析了。Rockyou 数据集包括 2009 年 12 月游戏论坛 Rockyou 泄露的 3200 万个密码 Allan 2009]; 450K 雅虎密码是由名为 D33Ds 的黑客组织于 2012 年 7 月在网上制作的；最后两个数据集(即 490 万 Mail.ru 和 130 万 Yandex.ru)于 2014 年 9 月被俄罗斯黑客泄露，其中约 90%是活跃的[Mick 2014]，并且据说这些凭证不是通过黑客攻击网站而是通过网络钓鱼和对用户的其他形式的黑客攻击(例如，键盘记录器)来收集的。

3.3. 关于密码数据集的统计信息

在这一小节中，我们报告一些关于收集的数据集的统计信息。首先，在表中总结了字符组成信息 II。有趣的是，中国用户更可能只使用数字来构造他们的密码，而英国用户更喜欢使用字母来构造他们的密码。一个合理的解释可能是中国用户，他们通常使用象形文字，不太熟悉键盘上的英文字母。另一个有趣的观察是，Myspace 用户倾向于通过在一系列小写字母中添加数字“1”来生成他们的密码。

表二。每个密码数据集的字符组成信息

资料组	[a-z]+	[A-Z]+	[A-Za-z]+	[0-9]+	[a-zA-Z0-9]+	[a-z]+[0-9]+	[a-z]+1	[a-zA-Z]+[0-9]+	[0-9]+[a-zA-Z]+	[0-9]+[a-z]+
天涯	9.96%	0.18%	10.29%	**63.77%**	98.05%	14.63%	0.12%	15.64%	4.37%	4.11%
多多新	8.79%	0.27%	9.37%	**20.49%**	82.88%	40.81%	1.39%	42.94%	7.31%	6.95%
CSDN	11.64%	0.47%	12.35%	**45.01%**	96.31%	26.14%	0.24%	28.45%	6.46%	5.88%
多万	10.30%	0.09%	10.52%	**52.84%**	97.59%	23.97%	0.37%	24.84%	6.04%	5.83%
聚友网（网站）	7.18%	0.31%	7.66%	0.71%	89.95%	**65.66%**	**18.24%**	**69.77%**	6.02%	5.66%
Singles.org	60.20%	1.92%	**65.82%**	9.58%	99.78%	17.77%	2.73%	19.68%	1.92%	1.77%
信仰作家	54.40%	1.16%	**59.04%**	6.35%	99.57%	22.82%	4.13%	25.45%	2.73%	2.37%
Hak5	18.61%	0.27%	20.39%	5.56%	92.13%	16.57%	2.01%	31.80%	1.44%	1.21%

桌子 III 显示了每个数据集的长度分布。我们可以看到，最受欢迎的密码长度在 6 到 10 之间，平均占整个数据集的 85.01%。很少有用户选择超过 12 位的密码，Dodonew 是个例外。一个可能的原因是，www.dodonew.com 是一个支持货币交易的网站，其用户认为他们的账户很重要，

因此选择更长的密码。我们的观察特别感兴趣的是，与其他数据集相比，CSDN 数据集的长度为 6 和 7 的密码要少得多。这可能是由于这样的事实 www.csdn.net(以及许多其他网站)从宽松的密码策略开始，后来实施了严格的策略(例如，要求密码长度至少为-8)。我们还注意到，来自 www.christian-singles.org 都不超过 8 个字符，这可能是由于策略禁止用户选择超过 8 个字符的密码。这种政策在许多金融公司中仍然存在 ohnston 2013]，一个合理的原因可能是转向更长的允许密码长度是一个不小的问题。

表三。每个数据集的长度分布信息

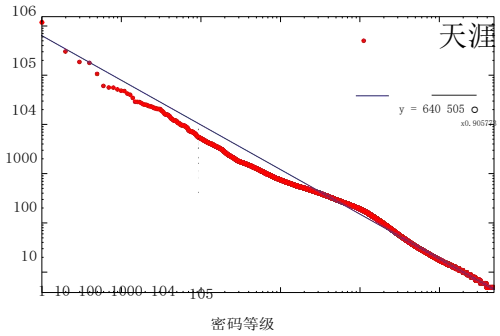
长度	1-3	4	5	6	7	8	9	10	11	12	13-16	17-30	30+全部
天涯	0.61%	0.65%	0.55%	33.77%	13.92%	18.10%	9.59%	10.28%	5.53%	2.88%	4.05%	0.07%	0.00% 100%
多多新	0.36%	0.70%	0.78%	9.71%	13.45%	18.49%	20.29%	14.69%	3.10%	1.34%	10.24%	6.79%	0.04% 100%
CSDN	0.01%	0.10%	0.51%	1.29%	0.26%	36.38%	24.15%	14.48%	9.78%	5.75%	6.96%	0.32%	0.00% 100%
多万	0.02%	0.13%	0.12%	20.62%	17.68%	22.49%	15.12%	11.55%	5.30%	2.72%	4.13%	0.12%	0.00% 100%
聚友网（网站）	0.25%	0.51%	0.79%	15.67%	23.40%	22.78%	17.20%	13.65%	2.83%	1.13%	1.15%	0.48%	0.17% 100%
Singles.org	0.68%	4.74%	7.68%	32.05%	23.20%	31.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00% 100%
信仰作家	0.04%	0.14%	0.99%	31.97%	20.95%	22.71%	10.35%	5.98%	3.24%	1.87%	1.53%	0.20%	0.01% 100%
Hak5	0.10%	0.64%	0.97%	12.96%	8.50%	20.89%	8.94%	30.83%	3.58%	3.08%	6.90%	2.44%	0.17% 100%
平均的	**0.26%**	**0.95%**	**1.55%**	**19.75%**	**15.17%**	**24.19%**	**13.20%**	**12.68%**	**4.17%**	**2.35%**	**4.37%**	**1.30%**	**0.05% 100%**

80 年代，据透露当时最流行的密码是 12345；30 年后，从表中可以看出 IV，123456 带头。一个长期存在的问题是，相当一部分用户喜欢相同的密码，就像事先达成一致一样，这部分是由于人类认知的固有限制。请注意，这种情况不能通过简单地禁止这种流行的密码来从根本上改变。比如 password 被禁，那么 password1 就会流行(见 Myspace 最流行的密码)；如果 password1 被禁，那么 pa$$word1 将会大行其道。希望自适应口令表(例如[Castelluccia et al. 2012])将最终消除这个问题。中文密码的前 10 名大部分是单个数字，而英文密码的前 10 名大部分是单个字母。

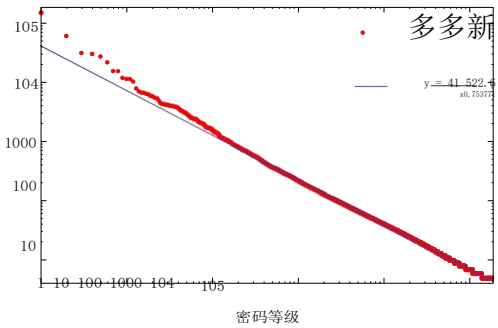
表四。每个数据集的前 10 个最常用密码

车阶	天涯	多多新	CSDN	多万	聚友网（网站）	Singles.org	信仰作家	Hak5
一	**123456**	123456	123456789	**123456**	密码 1	**123456**	**123456**	QsEfTh22
2	111111	a123456	12345678	111111	abc123	**耶稣**	**作 家**	——
3	000000	123456789	11111111	123456789	fuckyou	密码	**耶 稣 1**	蒂莫沙人
前 3 名(%)	5.58%	1.49%	8.15%	5.01%	0.40%	2.10%	**1.03%**	4.62%
四	123456789	111111	亲爱的书	123123	猴子 1	12345678	基督保佑 约 翰	ike02banaA
5	123123	**5201314**	00000000	000000	我爱你 1 我的空间 1	耶稣爱	316	**123456**
6	123321	123123	123123123	**5201314**	操你 1 第一足球 1	公主阳光	耶稣基督密码天堂信仰	zxczxc
七	**5201314**	a321654	1234567890	123321	尼科 1	1234567	作家	123456789
8	12345678	12345	88888888	a123456	0.78%	3.40%	2.17%	西侧
9	666666	000000	111111111	随便				ZVjmHgC355
10	111222 天涯	123456a	147258369	12345678				Kj7Gt65F
前 10 名(%)	7.42%	3.28%	10.44%	6.78%				7.20%

有趣的是，“爱”也是密码的永恒主题:五个数据集都有一个最受欢迎的与“爱”相关的密码。例如，5201314 这个中文发音为“我爱你，永远永远”的密码，在多多网和天涯分别排名第五和第七。信仰在塑造用户密码方面也有作用。例如，密码 jesus1 出现在 Sigle.org 和信仰作家(基督徒网站)的前十名名单中。令人惊讶的是，对于几个数据集，仅前 3 个最流行的密码就占了所有密码的 5%以上。这表明，要闯入这些相应的网站，一个在线(拖网)猜测攻击者每二十次尝试中就有一次会成功。此外，顺便提一下，尽管 Hak5 中流行的密码看起来相当复杂(多样化),实际上大约 66.18%的密码是由小写/大写字母和数字混合组成的，但这个数据集仍然非常集中，我们将在后面的部分中展示 5.1，是其中最弱的。这意味着看似复杂的密码可能不难破解，实际上可能相当脆弱，这进一步表明了对密码的基本理解的必要性和重要性。



图一。天涯的齐夫定律(R2 = 0.994)



图二。多多纽的齐夫定律(R2 = 0.996)

3.4. 密码中的齐夫定律

最初，PCFG 是用于自然语言处理(NLP)的机器学习技术，然而 Weir et al. [2009]设法利用它来自动建立密码篡改规则。最近，NLP 技术在评估语法对长密码和密码短语的脆弱性的影响方面也显示出有用 Rao et al.[2013]以及通过以下方式处理密码中的稀疏性问题 Ma et al. [2014].

受这些早期作品的启发，在这项研究中，我们试图研究齐夫定律，[2]存在于自然语言中，也存在于密码中。Zipf 定律最初被表述为等级-频率关系，以通过以下方式量化自然语言中单词的相对共性 Zipf [1949]. 它指出，给定一些自然语言话语的语料库，其中任何词的频率与其在频率表中的排名成反比。更具体地，对于以频率降序列出的自然语言语料库，单词的排名 r 和其频率 fr 成反比
成比例的，即 fr = C，其中 C 是取决于特定语料库的常数。这意味着最频繁出现的单词将是第二频繁出现的单词的两倍，第三频繁出现的单词的三倍，依此类推。
最近，Zipf 定律被证明可以很好地解释互联网拓扑[Faloutsos et al. 1999]，美国公司规模[Axtell 2001]和 Linux 软件包的分发[Maillart et al. 2008].
有趣的是，通过从数据集中排除最不受欢迎的密码(即，在这项工作中少于三到五个计数的密码)并使用线性回归，我们发现现实生活中密码的分布遵循类似的规律:对于密码数据集，密码的秩 r 及其频率 fr 遵循以下等式

$$fr = \text{卢比}, \quad (1)$$

其中 C 和 s 是取决于所选数据集的常数，而所选数据集又实质上由许多混淆因素决定(例如要保护的 web 服务的类型，
网站采用的基本密码策略，以及用户的人口统计因素，如年龄、性别、教育水平和语言)。Zipf 定律可以更容易地通过在双对数图上绘制数据(在本工作中以 10 为基数)来观察，坐标轴是 log(等级顺序)和 log(频率)。换句话说，log(fr)与 log(r)成线性关系:

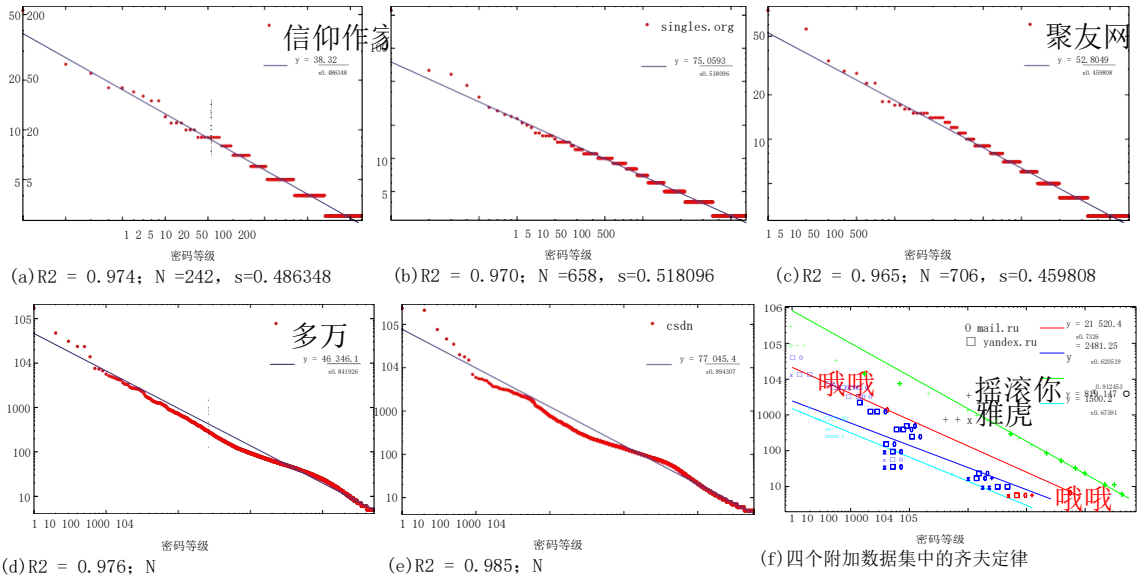
$$\log fr = \log C \; s \; \log r。 \quad (2)$$

从图中可以看出。1，来自网站的 3023 万个密码 www.tianya.cn 符合齐夫定律的程度为决定系数(用 R2 表示)为 0.994204954，约等于 1。这表明回归线
logy = 5.806522 0.905773∫ logx 完美拟合来自天涯的数据。如图 2 所示。2

2Zipf 定律分布也称为帕累托分布或幂律分布，它们是看待同一事物的不同方式，都可以从彼此推导出来[Zipf' sAdamic 2014].

和微型图。3，来自其他十个数据集的密码也总是遵循 Zipf 定律，并且回归线很好地代表了来自相应数据集的数据点。由于空间限制和前述 Hak5 数据集的不完美性质，我们在此不呈现其相关的 Zipf 曲线，尽管实际上其拟合线也具有高决定系数(即，R2 = 0.923)。

更准确地说，正如表中“决定系数”一栏所总结的那样 V，每个线性回归(除了 Hak5)的 R2 都大于 0.965，接近于 1，表明拟合非常好。至于“Hak5”，其 R2 约为 0.923，虽然可以接受，但不如其他数据集。一个看似合理的原因可能是它只包含不到三千个密码，并且可能无法代表整个密码数据集的真实分布 www.hak5.org。还应该注意的是，数据集如何泄露可能会对 R2 产生直接影响。从表中可以看出 V 与网站漏洞泄露的数据集相比，由网络钓鱼攻击泄露的数据集可能具有较低的 R2，因为网络钓鱼攻击不太可能获得网站的整个数据集，而网站漏洞一旦成功，网站的所有(或至少绝大部分)密码将被获取。



(a)R2 = 0.974；N =242，s=0.486348　(b)R2 = 0.970；N =658，s=0.518096　(c)R2 = 0.965；N =706，s=0.459808

(d)R2 = 0.976；N =51797，s=0.841926　(e)R2 = 0.985；N =57715，s=0.894307　(f)四个附加数据集中的齐夫定律

图 3。现实生活中的 Zipf 定律是以对数标度绘制的

我们需要删除最不常用的密码的原因将在第节中阐述 3.5。选择特定的小值(例如 3 或 5)作为最小频率(LF)的阈值基本上是基于统计学的发现(参见图 3)[Clausetet al. 2009])：当样本容量小于样本空间时，首先回归
随着 LF 逐渐增加而大大提高，直到达到最佳点 pˇ，之后回归变得极其缓慢(因为样本量减少)，如下所示
LF 增加。我们进行了一系列实验来确定使回归达到 pˇ 的确切 LF，并且发现，根据经验，对于具有数百万个密码的大型数据集，可以设置 LF = 5，否则设置 LF = 3。更复杂的方法(参见的第 12 页[Clauset et al. 2009]) 可以用来估计这个阈值
更精确地确定分布参数，然而它们不在这项工作的重点之内。尽管如此，表中的回归结果 V 证明我们对最小频率阈值的选择是令人满意的:每个回归都达到接近 1 的 R2。

回归过程中涉及的另外两个关键参数是 N 和 s，它们分别代表回归中使用的唯一密码的数量和回归线斜率的绝对值。虽然没有明显的关系

N 和 s，我们发现：(1)N 与密码总数有密切的联系，N 越大，密码总数越大；(2)参数 s 落在范围[0，1]中，这不同于其他自然/社会现象(例如，太阳耀斑的强度、战争的强度和姓氏的频率[Newman 2005])那些 s > 1 的。

我们强调，排除这些最不常用的密码是因为可用样本的有限大小：尽管密码群体(即，全部人类选择的密码)完全遵循 Zipf 分布，但百万大小的样本(例如，3000 万个天涯和 3200 万个 Rockyou)仍然太小，无法完全展示这一内在特征。这将在第节中得到证明 3.5。我们还推测，只有这些流行的密码会影响(降低)数据集的强度，这将通过严格的证明和第节中的大量实证实验来确定 5。此外，为了有资格作为数据集的适当描述，分布函数 f (x)应保持在至少 2 3 个数量级的范围 xmin f (x) xmax(即 xmax/xmin 102∾3)[Maillart et al. 2008]. 除了 Hack5，我们所有的回归都满足这个条件。

表五. 十二个密码数据集的线性回归(LR)结果(“PWs”代表密码)

	de- PWs 频率的唯一 PWs 绝对值	Zipf 回归系数的	Totoal 最小分数。斜率线(logy)终点(R2)的斜率线(N)的斜率线(LR)中的 PWs
天涯	30，233，633	5 0.50443286	486，118 0.905773 5.806523 0.905773∫logx 0.994204954
多多新	16 231 271	5 0.21640911	187 901 0.753771 4.618284 0.753771∫logx 0.995530686
CSDN	6 428 287	5 0.29841262	57 715 0.894307 4.886747 0.894307∫logx 0.985106832
多玩	4982740	5 0.28653592	51797 0.841926 4.666012 0.841926∫logx 0.976258449
Myspace	41，545	3 0.08094836	706 0.459808 1.722674 0.459808∫logx 0.965861431
Singles.org	16 250	3 0.22135384	658 0.518096 1.875405 0.518096∫logx 0.970277755
信仰作家	9，709	3 0.12472963	242 0.486348 1.583425 0.486348∫logx 0.97417589
hak	5 2987	3 0.15400067	76 0.643896 1.579116 0.643896∫logx 0.922662999
rock you	32 603 388	5 0.49600581	563 074 0.912453 5.913362 0.912453∫logx 0.997298647
雅虎	453 492	3 0.22668537	12 608 0.675910 3.176150 0.675910∫logx 0.983232690
mail . ru	4 938 663	5 0.33034872	83 914 0.732600 4.332851 0.732599∫logx 0.970047769
yandex . ru	1 261 810	5 0.34210777	26 003 0.620519 3.394671 0.620519∫logx 0.972507203

3.5. 我们方法的合理性

我们注意到 Malone and Maher [2012] 也试图调查现实生活中密码的分布。然而，与我们的发现相反，即用户生成的密码是 Zipf 分布的，并且受欢迎的密码(即，整个密码的前端)本身遵循 Zipf 定律，他们得出结论，他们的数据集(包括 32M Rockyou)“实际上不太可能是 Zipf 分布的”，并且“虽然 Zipf 分布不能完全描述我们的数据，但它提供了一个合理的模型，特别是密码选择的长尾。”我们找出了他们不同观察结果的主要原因——他们将数据集的所有密码都拟合到 Zipf 模型中。

更具体地说，不受欢迎的密码(例如 fr < 3)非常常见(见表 V)并构成密码选择的长尾(参见[Malone and Maher 2012] 具体把握)或统计领域中的“嘈杂的尾巴”[Newman 2005]，然而根据大数定律，它们未能反映出它们真正的受欢迎程度。因此，当整个数据集用于回归时，这些占每个数据集很大一部分的不受欢迎的密码将极大地负面影响拟合的良好性。这很好地解释了为什么在[Malone and Maher 2012]并提供了删除不受欢迎的密码的必要性的直接原因。

我们观察到，存在一个更本质的(然而微妙的)原因：即使 pass- word 群体完全遵循 Zipf 分布，百万规模的样本(例如，3000 万天涯和 3200 万 Rockyou)仍然太小，无法完全展示这一内在特征。举个例子，www.csdn.net 采用允许密码由字母和数字组成并且长度为 8 到 16 的策略，这意味着用户的密码(由随机变量 X 表示)将具有大约 X =6216 628 4.8 1029 个可能的(不同的)值在这个政策下。但是我们从泄露中只得到 6.42 106 个 CSDN 密码，_ ≈ *由于 Zipf 分布中概率的多项式递减性质，相对于 X 是非常小的样本(见等式。1)，低概率事件(如 fr < 3)

将压倒小样本中的高概率事件，因此不排除不受欢迎的事件的这种小样本极不可能反映真实的基本分布。由此可见，当拟合相对小的数据集的所有密码时，回归将受到这些不受欢迎的密码的负面影响，并且即使密码的前端展现出良好的 Zipf 属性，也无法观察到显著的规则。

我们强调，虽然这些最不常用的密码本身并不显示 Zipf 行为，但这一事实并不违背我们的断言，即(网站的)密码群体完全遵循 Zipf 分布。桌子 V 显示，通常，数据集越大(或者样本量越大)，常用密码(即回归中使用的密码)的比例就越大。基于这种趋势，可以预计，如果数据集足够大，不受欢迎的密码将会很少，无论是否排除它们，对拟合的良好性都没有什么影响。也就是说，整个数据集将展示一个 Zipf 属性。幸运的是，我们的后续工作之一(参见 http://wangdingg.weebly.com/uploads/2/0/3/6/20366987/pin‧zipf.pdf)一种特殊的密码——人工选择的密码——的分布很好地证实了这一推论。可以看到，大多数检查的 4 位 PIN 数据集可以完全符合 Zipf 模型，即使排除 fr < 10 的 PIN，仍有超过 94%的数据集留在回归，他们完全遵循齐夫定律(R2 > 0.97)。

为了进一步证明我们的断言，即用户选择的密码样本(即数据集)遵循 Zipf 定律，我们研究了从完美的 Zipf 分布中随机抽取的样本的回归行为，并查看这两种类型的样本是否显示出相同的回归行为。我们探讨了可能影响回归的三个参数，即精确分布(3 种)、样本量(8 种)和最小相关频率(5 种)，从而进行了一系列 120(=3 5 8)次回归试验。更具体地说，假设随机变量 X 遵循 Zipf 定律，则有 *N =的 103 个可能值{x1，x2，C/1's x103 }。不失一般性，*秒1 1/2秒 1/纳秒 分布定律被定义为{p(x1)=∑N C=∑N 1，p(x2)=∑N 1，，p(xN) =∑N 1 }，其中样本空间 N 和斜率 s 定义了精确的 Zipf 分布函数。为了稳健起见，每个实验运行 103 次；为了更好的比较，每个实验只有一个参数变化。由于空间限制，表 VI 仅包括 40 个实验，其中 Zipf N 固定为 103，Zipf s 固定为 0.9，样本大小从 102 变化到 2 104，LF 从 1 逐渐增加到 5。读者可以参考[Wang et al. 2015]. 注意，表中的一些整数统计量(如拟合的 N)VI 因为它们是 1000 次重复实验的平均值。

我们对 120 个实验的结果表明，给定 Zipf 分布(即，当 Zipf 参数 N 和 s 固定时)，无论样本大小小于、等于或大于 N，较大的 LF 在开始时都会导致更好的回归(即，拟合的 s 更接近 Zipf s，R2 更接近 1)，但随着 LF 进一步增加，情况会恶化。更具体地说，当样本量小于 N 时，拟合的 s 随着 LF 逐渐增大而先增大后减小；相反，当样本量大于 N 时，随着 LF 的递增，拟合的 s 先减小后增大。因此，我们可以确定最佳拟合(粗体)，从中我们可以看出，样本量越大，回归中使用的流行事件部分就越大。这种行为很好地符合我们对现实生活中的密码数据集的观察。

特别地，当样本量足够大时(例如，104 N = 103)，流行事件(例如，fr 3)总是占每个样本的 95%以上，并且完全遵循齐夫定律(R2 0.99)。这种行为与我们在大头针和根据我们对密码数据集的推断。此外，当样本大小远小于样本空间 N 时，冷门事件占大多数，但我们必须排除它们以获得良好的拟合。这证明了我们在执行回归分析时的数据处理方法(即删除不常用的密码)是正确的，因为现实生活中的密码数据集的大小通常比密码样本小得多

表六。模拟 Zipf 分布时样本大小和最小频率(LF)对线性回归的影响

Zipf	Zipf	样品		唯一的数量密码	使用的密码回归中	使用的密码回归中(%)	合适的*普通*	合适的 *s*	*R2*
1000	0.9	100	一	71.197	100.000	100.00%	71.197	0.429486	0.754566
1000	0.9	100	2	71.262	41.099	41.10%	12.361	0.641264	0.884263
1000	0.9	100	3	70.963	27.201	**27.20%**	5.307	**0.719897**	0.894042
1000	0.9	100	四	71.068	20.585	20.59%	3.173	0.683547	0.916477
1000	0.9	100	5	70.765	17.010	17.01%	2.215	0.622484	0.953243
1000	0.9	200	一	123.933	200.000	100.00%	123.933	0.516278	0.822066
1000	0.9	200	2	124.103	102.971	51.49%	27.074	0.688394	0.923847
1000	0.9	200	3	123.795	73.429	36.71%	12.145	0.761613	0.935451
1000	0.9	200	四	124.121	59.139	**29.57%**	7.392	**0.785336**	**0.930795**
1000	0.9	200	5	123.954	50.151	25.08%	5.242	0.784747	0.921241
				245.459	500.000	100.00%	245.459	0.633549	0.895852
1000	0.9	500	2	246.040	326.859	65.37%	72.899	0.724630	0.951529
1000	0.9	500	3	245.482	250.498	50.10%	34.245	0.796940	0.969880
1000	0.9	500	四	245.697	211.680	42.34%	21.499	0.819386	0.970288
1000	0.9	500	5	245.586	187.536	**37.51%**	15.372	**0.834885**	**0.966581**
1000	0.9	1000	一	389.360	1000.000	100.00%	389.36	0.730031	0.937941
1000	0.9	1000	2	388.014	760.039	76.00%	148.053	0.756649	0.965318
1000	0.9	1000	3	388.733	611.795	61.18%	74.478	0.807381	0.979783
1000	0.9	1000	四	388.774	530.803	53.08%	47.184	0.833071	0.983395
1000	0.9	1000	5	388.839	476.921	**47.69%**	33.829	**0.847137**	**0.983550**
1000	0.9	2000	一	573.821	2000.000	100.00%	573.821	0.835995	0.964407
1000	0.9	2000	2	573.607	1712.451	85.62%	286.058	0.790817	0.977339
1000	0.9	2000	3	574.446	1455.076	72.75%	158.041	0.818059	0.985691
1000	0.9	2000	四	574.011	1287.865	64.39%	102.03	0.840089	0.989460
1000	0.9	2000	5	574.229	1173.160	**58.66%**	73.534	**0.854452**	**0.990812**
				828.243	5000.000	100.00%	828.243	0.963949	0.963691
1000	0.9	5000	2	828.466	4760.094	**95.20%**	588.56	**0.861714**	**0.989008**
1000	0.9	5000	3	827.675	4379.226	87.58%	397.276	0.842637	0.991843
1000	0.9	5000	四	828.601	4014.673	80.29%	276.308	0.849865	0.993588
1000	0.9	5000	5	828.281	3724.258	74.49%	203.349	0.859765	0.994832
1000	0.9	10000	一	953.483	10000.000	100.00%	953.483	1.013698	0.943442
1000	0.9	10000	2	953.545	9884.596	98.85%	838.141	0.929787	0.985080
1000	0.9	10000	3	953.125	9582.080	**95.82%**	686.791	**0.884120**	**0.994655**
1000	0.9	10000	四	953.483	9146.947	91.47%	541.471	0.867965	0.996179
1000	0.9	10000	5	953.365	8683.549	86.84%	425.614	0.866388	0.996641
1000	0.9	20000	一	995.527	20000.000	100.00%	995.527	0.994645	0.943878
1000	0.9	20000	2	995.514	19979.918	99.90%	975.432	0.968837	0.969613
1000	0.9	20000	3	995.521	19886.123	99.43%	928.450	0.938901	0.986291
1000	0.9	20000	四	995.550	19665.232	98.33%	855.099	0.912336	0.994446
1000	0.9	20000	5	995.544	19298.183	**96.49%**	763.027	**0.894282**	**0.997415**

注:关于更详细的结果，读者可参考补充材料[Wang et al. 2015] 这项工作。

空格(如 6.42 106 Xcsdn 4.8 1029)* 简而言之，我们在 12 个密码数据集上的回归中显示的所有行为都与 120 个模拟实验很好地一致。

3.6. 我们观察的普遍适用性

在前面几节的回归中，我们只考虑了在宽松的密码创建策略下生成的数据集。桌子 IIIV 显示在每个数据集中都出现了非常短且只有字母的密码，这表明在任何站点生成密码都没有明显的长度或组成要求。我们认为对这种现象更精确和合理的解释是，这些密码大多数是在未知策略的混合下创建的:最初，没有规则(策略)；后来，应用了一些更严格(或更宽松)的规则；一段时间后，这些网站被黑了。

然而，在某些情况下，这是不正确的，特别是对于安全关键的服务，它们可能在一开始就实施严格的策略。为了进一步建立我们的发现的适用性，考虑了在更受约束(但相当现实)的密码策略下创建的两种特殊类型的数据集:(1)密码长度满足某个最小长度(例如，至少长度-8)的数据集；以及(2)每个密码是字母和数字的混合(例如，至少一个字母和一个数字)的数据集。

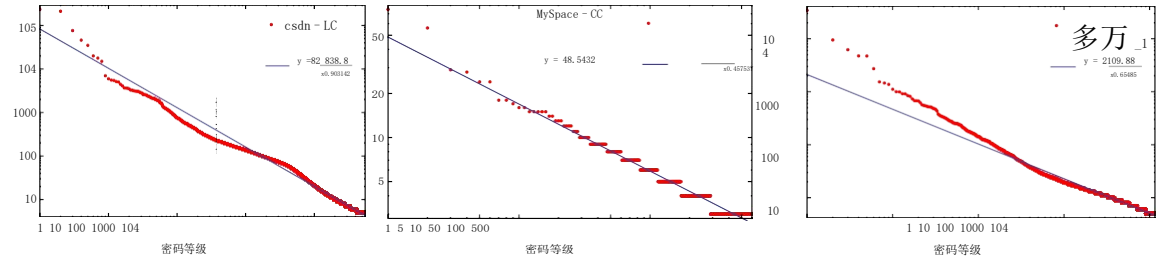
由于我们没有在某些具有长度或组成要求的特定创建策略下准确生成的密码的确切示例(据我们所知，没有公开可用的理想数据)，我们试图通过基于最小长度或组成要求进一步划分这些数据集来对此类策略进行建模。然而，我们被警告说，简单地根据一些人为的策略划分现有的数据集可能是没有意义的，因为在这个过程中，用户行为将在很大程度上被扭曲。这种谨慎的一个附带证据是观察到，在明确策略下创建的密码“不能简单地通过从较大的语料库中选择符合密码的子集来正确地表征”，并且“这样的子集不可能代表在所讨论的策略下创建的密码”[Ur et al. 2012].Mazurek et al.[2013]也报道了类似的观察。幸运的是，在仔细检查了我们的12个数据集(见表II 和桌子III)，我们发现:

(1) 在CSDN，只有2.17%的密码短于八个字符。这些短密码很大程度上是由于最初的宽松政策，其余97.83%的长密码是由于后来的增强密码政策。密码策略的这一转变已经得到证实;
(2) Myspace中高达75.79%(=69.77%+6.02%)的密码同时由字母和数字组成，超过18.24%的用户选择字母序列与数字“1”串接的密码。这高度表明，在黑客攻击发生之前的某个时候，组合需求发生了转变，尽管我们无法证实这种转变。

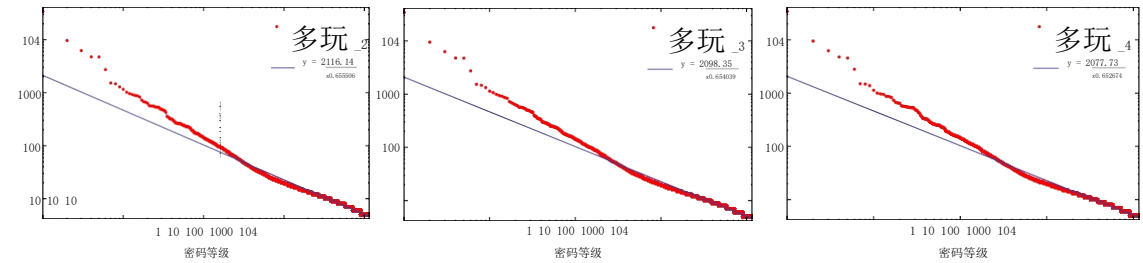
因此，这两个数据集构成了有用的子集，分别代表符合上述两个受约束的密码策略的密码。更具体地说，来自CSDN的97.83%的长密码构成了根据要求密码长度至少为八个字符的策略创建的数据集，来自Myspace的75.79%的密码构成了根据要求密码长度至少为一个字母和一个数字的策略创建的数据集。而我们简称它们为“csdn-lc”和“myspace-cc”，其中“lc”代表“长度受限”，“cc”代表“字符受限”。这两个细化数据集的线性回归结果如图所示。4(a) 和4(b)，分别为。我们可以看到，这两个回归的决定系数(R2)为0.966或更高，表明拟合良好。这表明Zipf定律也适用于在非常严格的策略下创建的密码。

为了研究遵守Zipf定律的数据集的子集是否也遵守该定律，我们进一步对从12个数据集中随机选择的子集进行线性回归。正如预期的那样，任何子集与其父数据集之间的拟合效果没有显著差异(Fisher精确检验，p值0.05)。由于空间限制，图中只描绘了从Duowan中随机选择的四个子集(每个子集的大小为1百万)。4(c)图 。 4(f)。由于这四个回归的R2都是0.977并且非常接近1，这表明Zipf定律在这些子集中非常适合。这意味着，如果我们可以获得一个认证系统的足够大的密码子集，那么整个密码的分布可以在很大程度上通过进行线性回归并使它们符合Zipf定律来确定。然而，数据集的多少部分可以被认为是“足够大”的呢？六分之一，十分之一，或者百分之一怎么样？这暗示了未来研究的自然方向。

在这个阶段，一个自然的问题出现了:我们的观察可以推广到大多数情况下用户生成的密码吗？或者同样，这个问题可以表达为:在这项工作中使用的数据集是否可以代表大多数数据集？答案是高度肯定的。一方面，这项工作中使用的数据集是迄今为止最多样化的(就服务、规模、泄露方式、地点、语言和文化/信仰而言)，也是最大的数据集之一(就密码总数和数据集数量而言)，因此它们具有良好的代表性。在之前的密码研究中，据所知，最多样化的数据集(即，三个来自美国，三个来自中国，每个来自不同的服务)



(a)R2 = 0.985; N =55178, s=0.903142 (b)R2 = 0.966; N =600, s=0.457537 (c)R2 = 0.977; N =11549, s=0.654849



(d)R2 = 0.977; N =11479, s=0.655506 (e)R2 = 0.977; N =11571, s=0.654038 (f)R2 = 0.977; N =11626, s=0.652674

图 4。在受约束的策略下创建的密码和从真实数据集(使用 Duowan 作为典型示例)中随机采样的密码中的 Zipf 定律以双对数标度绘制

据报道[Ma et al. 2014] 最大的数据集(即七个数据集，总共有 1.14 亿个密码)已经在[Li et al. 2014]，而在我们的工作中，我们使用了 12 个不同的数据集，总共有 9720 万个密码。诚然，我们的数据集(以及长度受限和字符受限的数据集)不能表示所有种类的真实数据集，例如，它们都不能表示非常重要的凭证(例如，电子银行)。尽管如此，这些数据集仍然代表了大量多样化的用户生成的密码，并可用于研究一般的密码分布，并且由于决定系数较高，它们所确定的显著特性不太可能是巧合。

另一方面，严格地说，像我们这样的概括是否恰当，没有(或者永远不会有)明确的答案。医生的目标是理解
物理世界是如何工作的永远无法确定他们的理论(如牛顿定律)是否正确，相反，他们只能判断他们的理论是否与最先进的实验一致。同样地，我们的目标是了解现实生活中的密码是如何分布的，但永远无法确定我们的理论是否绝对正确；有了足够的数据和正确的工具，我们只能开发模型来越准确地描述密码分布，这可能是一项永无止境的工作。我们坦率地承认，在这个有趣的课题上需要投入更多的努力。

总的来说，虽然我们的数据并不理想，但我们相信我们的发现确实提供了很多
更好地理解用户生成的密码的分布，可以广泛应用。虽然对这一重要课题知之甚少，但即使是相对有限的探索也是一种进步，更不用说基础研究了。

4. 一些基本含义

在这一节中，我们展示了 Zipf 理论的两个基本含义。我们相信这一理论在其他领域也是令人感兴趣的，并且它为它们的进一步理论发展和实际应用奠定了基础(参见“ geno guard”[Huang et al. 2015])。

4.1. 密码策略的含义

最近，许多关于密码策略的工作(例如，[Schechter et al. 2010; Castelluccia et al.2012]) 已经建议不允许用户选择危险流行的密码(例如，123456 和密码 123),其出现的概率大于预定的概率

阈值(例如= 1/106)。令人惊讶的是，他们的动机主要是基于
根据经验观察，一些用户使用不受欢迎的流行密码等
密码特别容易受到统计攻击，这是一种字典攻击形式(可能是在线或离线)，攻击者根据流行度对字典进行排序，并首先猜测最流行的密码。到目前为止，基本原理还没有给出，许多基本问题仍有待解决。例如，受降低流行阈值影响的那部分用户增长的基本趋势是什么？在给定的阈值下，选择流行密码的用户比例是多少？如果我们限制前 0.0001%最常用的密码，会有多少比例的用户受到影响？限制前 0.01%最流行的密码怎么样？

我们现在准备回答这些问题。在截面中 3，我们已经表明，在大多数情况下，用户生成的密码遵守 Zipf 定律，该定律规定 a 的秩 r
密码及其频率 fr 遵循公式 fr = C，其中 C 是常数，即

通常略小于最流行的密码(由 F1 表示)的频率，即 C = f1F1。为了说明的目的，假设用户密码 X 的频率是连续的实变量，并且在从 X 到 x + dx 的区间中取值的相应概率由 p(X = x)dx 表示。据[Adamic 2014]，现在 p(X = x)服从一个幂律分布。更具体地说，

$$p(X = X) = C' X^{\alpha}, \tag{3}$$

其中 α = 1 + 1/s，s 如等式中所定义。1。至于 C′，它是由归一化要求给出的

$$1 = \int_{xmin}^{\infty} p(X = X)dx = \int_{xmin}^{\infty} C' x^{-\alpha} dx = \frac{C'}{[x^{\alpha+1}]_{x}^{\infty}}, \tag{4}$$

$$\frac{1-\alpha}{}$$

其中，在实际情况下，xmin 不是测得的 x 的最小值，而是满足幂律特性的最小值。当 α = 1 + 1/s > 1 时，我们得到

$$c' = (\alpha - 1) x_{min}^{\alpha - 1}。 \tag{5}$$

因此，特定密码的频率将大于 x (x ⩾ xmin)的概率由下式给出

$$p(X > X) = \int_{x}^{\infty} p(X = X')dx = \frac{c'}{\alpha} x^{\alpha+1} = (x)^{\alpha+1}。 \tag{6}$$

$$\frac{x}{-1}$$

注意，根据定义，P (X > x)也可以看作是累积密码流行度分布函数。基于情商。4 还有情商。5 以及 α = 1 + 1/s > 2 的事实(见表中的 sV)，可以确定阈值 t 下允许的最大频率 x

$$x = \int_{xmin}^{\infty} xp(X = x)\, dx = c' \int_{xmin}^{\infty} x^{\alpha+1} dx = \frac{\alpha - 1}{\alpha - 2} x_{min}^{\alpha+1}。 \tag{7}$$

我们将受阈值 t 潜在和实际影响的用户账户(密码频率超过 x)的确切比例表示为 Wp(X > x)
西澳(X > x)，分别在哪里

$$W(X > x) = \int_{x'}^{\infty} \int_{x}^{\infty} p(X = X')(x' - x)dx' = (x) \tag{8}$$

$$\int_{x}^{\infty} (x' - x)p(X = x')dx' \quad \text{电话/音(cross-talk的缩写)}$$

$$Wa(X > x) = \int_{xmin}^{\infty} X' p(X = X')dx' = \alpha - 1(x^{-} \quad \text{部}) . \tag{9}$$

3 请注意，Wp(X > x)和 Wa(X > x)确实是两个独立且有用的指标，用来衡量可用性受影响的程度。例如，现在如果 Notewww.dodonew.com 使用 = 1/1024 实施基于流行度的策略，那么将有 Wp(X > x)=3.33%的帐户的密码比
= 1/1024，这意味着这 3.33%的账户中的每一个都具有被要求更换新账户的同等潜力

密码。但是，实际上只有 2.51%的 Wa(X > x)=账户需要选择不同的密码，因为在 Wa(X > x)=2.51%的账户已经被更改后，剩余的 WP(x >x)= 0.82%的 Wa(X > x)=0.82%的账户的密码没有 = 1/1024 那么受欢迎。

使用等式。6~9，我们可以得到用户帐户的分数，其每个密码位于最流行的部分 P（X > x ）：

$$WP(x > x\ ) = (p(x > x\ ))(-\alpha + 2)/(-\alpha + 1)。 \quad (10)$$

由于 α = 1 + 1/s，等式。10 可以重写为

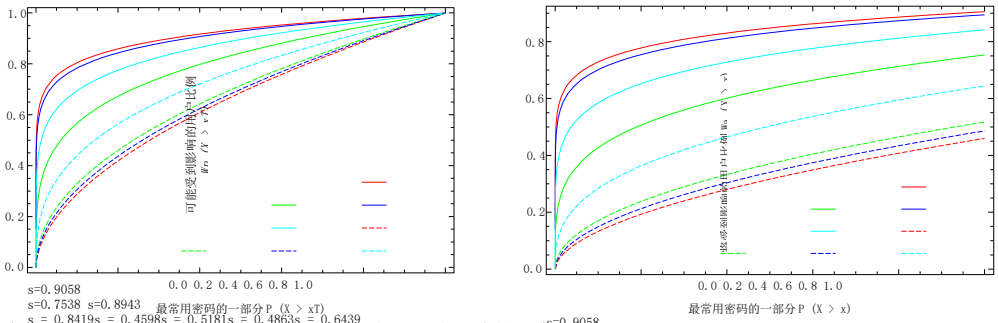
$$w(X > X) = (P(X > X))(1)/(= (P(X > X\ ))1s。 \quad (11)$$

同理，情商。9 可以重写为

$$W_x(X > x) = (P(X > x))(11)/(1) = s(P(X > X\ ))1s。 \quad (12)$$

$$\alpha - 1$$

这表明，关于累积密码流行度分布函数 p(X > x)，两个可用性降低指标 Wp(X > x)和 Wa(X > x)遵循具有正指数 1 s 的帕累托定律。为了更好地理解，在图。5 我们描绘了 Wp(X > x)和 Wa(X > x)相对于 p(X > x)的曲线形式，对于表中列出的各种 s 值 V。



(a)潜在受 P（X > xT)影响的用户(b)实际受 P（X > xT)影响的用户

图 5。如果密码遵循指数为 s 的 Zipf 法则分布，则可能/实际受到基于流行度的策略影响的用户比例，如表中所列 V。

Wp 和 Wa 在曲线的最开始急剧增加(见图。5)明确揭示了流行密码过于流行，即使只检查了流行密码的边缘部分，也有不可忽略的一部分用户会感到不方便。例如，根据等式。12，Wa=2.51%用户在 s = 0.7538，= 1/1024，P = 0.0001%时会很烦。看看我们的理论是否符合
实际上，我们还总结了中八个真实密码数据集的统计结果
桌子 VII。可以确定的是，理论 Wa 超过经验 Wa 13 倍。从我们的理论模型得到的结果大于实验统计结果的主要原因是，有很大比例的密码
不频繁的(即，它们的频率低于 xmin)，通常称为“噪声尾部”[Newman 2005] 在统计领域。此外，为了简单起见，我们将用户密码的频率(它是一个离散的整数)建模为一个连续的实变量，这将不可避免地引入一些偏差。

尽管上述理论模型并不完全准确，但据我们所知，它对
第一次确实揭示了受流行阈值影响的那部分用户的基本趋势，并提供了有见地的、简明的和实用的指标，便于策略设计者和安全管理员在可用性和安全性之间提供更可接受的折衷。例如，根据我们的理论不难
请注意，对于互联网规模的网站，设置= 1/106 可能不合理，因为超过 60%的用户可能会感到恼火。然而，Schechter et al. [2010]和 Florenˆcioet al. [2014]只是明确地(或隐含地)建议这样一个值。另一边
手，齐普夫定律在第一节中揭示 3.4 表明大多数的频率

表七。密码流行度阈值对具有不期望的流行度的密码部分(即 P)和将实际受到影响的用户帐户部分(即 Wa)的影响

密码	= 1/1024		= 1/10000		= 1/16384		= 1/1000000	
资料组	*P*	*伍人*	*P*	*伍人*	*P*	*伍人*	*P*	*伍人*
天涯	0.0001%	6.6023%	0.0015%	10.7586%	0.0023%	11.6473%	0.4416%	30.9110%
多多新	0.0001%	1.3926%	0.0009%	3.1556%	0.0014%	3.6298%	0.2958%	11.2351%
CSDN	0.0002%	9.4648%	0.0029%	12.2806%	0.0049%	12.8732%	0.8441%	24.6874%
多万	0.0004%	5.8130%	0.0048%	8.8648%	0.0079%	9.6064%	1.6607%	24.4955%
聚友网（网站）	0.0054%	0.1228%	0.5358%	2.1952%	1.9007%	4.6961%	-	-
Singles.org	0.1553%	2.6154%	14.1818%	24.7138%	-	-	-	-
信仰作家	0.1917%	1.3390%	-	-	-	-	-	-
Hak5	3.2327%	10.3113%	-	-	-	-	-	-

注:破折号“-”代表“不适用”,因为1/ 大于相应数据集的大小。

流行的密码以大约对数的速度下降，因此只有有限比例的密码过度流行。因此，我们只需要防止这些过度的密码，并设置一个适当的流行阈值。例如，当设置为适中值 1/16384 时，大多数系统中只有不到 13%的用户会感到烦恼
符合 2 级认证[Burr et al. 2013]，这表明= 1/16384
会更容易被大多数互联网规模的电子商务网站所接受。这是第一次，
为基于流行度的密码策略的必要性和可行性(以及预防措施)提供了合理的理论基础。我们还要强调的是，我们在这里绘制的图片是对政策可用性的初步的、可信的(而不是结论性的)评估，彻底的实地研究本质上仍然是必要的。

4.2. 基于密码的身份验证的含义

我们的观察的另一个基本含义是对于数以千计的涉及口令的可证明安全的认证协议，像只有口令的单因素方案(例如，两方[Katz et al. 2009]和多方[Chen et al. 2014]) 和基于密码的多因素方案(例如，双因素[Wang et al. 2014] 和三因素[Huang et al.2014])。在这里，我们首先展示了仅密码方案的含义，也称为 PAKE 协议。在大多数可证明安全的 PAKE 协议中(例如[Bellare et al. 2000; Canetti et al.2012; Pointcheval 2012; Abdalla et al. 2015a; Chen et al. 2014] 在随机预言模型和[Halevi and Krawczyk 1999; Katz et al. 2009; Katz and Vaikuntanathan 2013; Yi et al.2014] 在标准模型中)，通常假设“密码 pwC(对于每个客户端 C)是从大小的字典中独立且均匀地随机选择的，其中是独立于安全参数 k 的固定常数”，然后描述安全模型，最后是安全的“标准”定义，如[Katz et al. 2009] 是鉴于: D |D|

协议 P 是仅用于口令认证密钥交换的安全协议，如果对于所有[口令]字典大小|D|并且对于所有 PPT[概率多项式时间]对手 A 进行最多 Q(k)次在线攻击，存在
可忽略的函数 ϵ()使得:

$$\mathrm{Adv}_{P,D}(k) \leq Q(k)/|D| + \epsilon(k), \qquad (13)$$

Adv , (k)在进攻中的优势在哪里。"[4]

美联社

据[Bonneau 2012b]，用户生成的密码通常提供大约 20 ^ 21 位的实际安全性来抵御最佳离线字典攻击，这意味着有效的密码空间的大小大约为 220 ^ 221。这表明采用 PAKE 协议的系统实现了等式 1 的安全目标。13 可以保证 个在线猜测

我们注意到一些 PAKE 协议(例如，We[Chen et al. 2014;Bellare et al. 2000])将安全性的定义放宽到 Adv, (k)c Q(k)/ + ϵ(k)，其中 c 是一个恒定的正整数，表示现在允许每次在线尝试猜测 c 个密码。然而，这并不一定意味着相应的
协议实际上面临着 每次在线尝试都能猜出多个密码的威胁，这种放松可能是由于还原论证明中的一些技术原因，而不是 .的固有缺陷

尝试将获得不大于 1/220 1/221 的成功率，这显然不是事实，实际上在实践中可能有些误导。例如，相对于游戏和电子商务网站的实际优势 www.dodonew.com 当 Q(k)=3 时达到 1.49%，当 Q(k)=10 时达到 3.28%，这远远超出了理论结果
按 Eq。13。可以预见的是，与大多数真实网站相比，的优势将是
很大程度上被低估了，并且可能会向普通用户和安全管理员传达一种过于乐观的安全感。

作为谨慎的旁注，这些作品中的一些(例如，[Katz and Vaikuntanathan 2013; Katzet al. 2009; Halevi and Krawczyk 1999])补充说明，具有恒定大小字典的密码均匀分布的假设只是为了简单起见，并且它们的安全性证明可以扩展到处理更复杂的情况，其中密码不均匀分布、不同客户端的不同分布或者密码字典大小取决于安全参数。然而，这样的补充只会模糊他们的安全声明，并破坏读者(例如，工业、政府和学术界的人)对他们在多大程度上可以信任用于保护系统的认证协议的理解，因为没有人知道如果“用户选择的密码不均匀分布”，分布会是什么。这违背了构造可证明安全的协议的目的，该协议“通过对安全的具体或精确处理，明确地捕获安全的固有的量化性质”并且“提供量化的安全保证”[Bellare 1999] 首先。

根据我们的理论，现在根本没有必要(不切实际地)制造
密码均匀分布的假设，相反，可以直接对密码分布进行 Zipf 假设。因为系统分配的随机密码几乎不可用[Shay et al. 2012],大多数系统允许用户生成他们自己的密码，这将高度导致密码符合 Zipf 发行版，如我们在第节中所示 3.4。然而，根据 Zipf 假设，这是非常可能的
那 $s\ s\ s$ $k\ 1$ 将大得惊人(由于

副词 $(k) = C/1 + C/2 + + C/k = j=1\ js$ Σ Σ Σ Σ

调和数的逼近对数函数性质 Paule and Schneider2003])因此，即使 k 很小，系统也会处于严重的危险之中。例如，当仅对网站进行 3 次在线攻击时，该值达到 1.49%www.dodonew.com 这使得货币交易成为可能。从表中可以看出 IV，八个网站中的七个有超过 3.28%的帐户被攻击者攻破的机会，攻击者仅仅进行十次在线模仿尝试。

来自 Eq。12 (或 Fig。5(a))我们还可以看到，相当小一部分最流行的密码(用 P 表示)可以占用户帐户(用 Wp 表示)的不可忽略的比例，这表明在线猜测攻击者只需尝试 P 个不同的密码就有机会 Wp 成功。也就是说，即使所采用的认证协议是可证明安全的，如果系统的密码遵守 Zipf 定律，也仍然不能达到安全的用户识别。这再次强调了加密方法应该与系统解决方案相结合，以确保系统安全。为此，密码不应遵循 Zipf 分布。这表明应该采取一些必要的对策(例如，利用限制过于流行的密码的策略)，这可能导致密码具有倾斜的 Zipf 分布。虽然关于密码的均匀分布假设是不现实的，Zipf 分布是不安全的，并且倾斜的 Zipf 分布似乎很难被严格地表征，但是我们陷入了一个难题，无法制定安全结果的定义，如等式 1。13。受安全 PAKE 协议所能提供的基本安全概念的启发，只有在线假冒攻击有助于对手破坏协议的安全性[Halevi and Krawczyk 1999]，我们试图通过放弃先描述口令的分布，然后制定安全的定义的想法来摆脱这个问题。相反，我们提供了
一个严格的上限
为了对手的利益。更确切地说，情商。13 现修改如下:

$$\mathrm{Adv}_{\mathcal{A}}(k) \leqslant \mathrm{F1}\cdot Q(k)/|\mathcal{DS}| + \epsilon(k),$$

(14)如前所述，其中 F1 是数据集中最流行的密码的频率
是目标身份验证系统的(预期)用户帐户数量，
其他符号与 Eq 的符号相同。13。注意，dictionary 是密码样本空间，它是一个集合，而 dataset 是一个(特定的)密码样本，它是一个多重集合。因此，F1 的值正好是阈值概率(例如，$T_{1/16384}=|\mathcal{DS}|/16384$)，由底层密码策略维护。对于达到 1 级认证的系统[Burr et al. 2013]，在线猜测攻击者的成功几率应该不大于 1/1024，表示 F1/1/1024；同样，对于 2 级认证，系统应确保 F1/ 1/16384。例如，对于游戏和电子商务网站 www.dodonew.com 要实现 2 级安全性，F1 应该不大于 991(16231271/16384)。
还要注意，Eq。13 其实是情商的特例。14，其中 F1 = 1 且=
。而且，情商。14 也可用于粗略地描述中的正式安全结果
用户密码完全遵循 Zipf 定律的情况。虽然不精确，情商。14 (例如，让 F1/ = 1/1024)仍然比像 Eq 这样的更合理(和现实)。13 目前在加密协议社区中广泛使用的(例如[Katzet al. 2009; Pointcheval 2012; Chen et al. 2014; Yi et al. 2014; Abdalla et al. 2015a])。
我们碰巧发现一系列由 Abdalla et al. [2015b]使用不同于传统的安全模式:

$$\mathrm{Adv}_{\mathcal{A}}(k) \leqslant Q(k)/2^{m} + \epsilon(k), \tag{15}$$

其中 m 是密码的最小熵。[5] 其实不难看出，这种提法(即 Eq。15)和情商在本质上是一样的。14，因为可以推导出 $m = \log_2(\mathrm{F1}/)$[Bonneau 2012b].相比之下，我们的表述更加具体和易懂。更何况，没有任何理由或理由偏爱情商。15
但情商不行。13 是在[Abdalla et al. 2015b].幸运的是，在部分 3.4 我们终于解决了这个基本问题:密码是否可能是均匀分布的？如果不是，那么什么才是正确的分布呢？
人们还可以看到，如果 m 被定义为密码的熵，那么等式。15 实际上等于 Eq。13 并且它提供了在线猜测难度的平均值，因为可以推导出 $m = \sum i\log_2 p_i$，其中 pi 是第 I 个最大的概率
频繁输入密码(例如，p1 = F1/)。这很好地解释了为什么 Benhamouda 等人。
(参见的第 6.1 节[Benhamouda et al. 2013]) 陈述“相当于的优势”
任何对手都可以被“任一等式”所限制。13 还是情商。15。然而，正如我们已经展示的，如果 m 被定义为密码的最小熵(这在大多数情况下是正确的定义)，等式。13 还有情商。15 (或者同样，Eq。14)将彼此显著不同。
与 PAKE 协议不同，在 PAKE 协议中，用户必须与服务器交互来注册他们的密码，大多数多因素方案提供了一个属性，称为“DA2-Local-Secure”[Wang et al. 2014]，方便用户在本地自由更改密码(即无需与服务器交互)。由于没有与服务器的交互，基于流行度的密码策略无法实施，用户密码几乎肯定会遵循 Zipf 分布。然而，当评估是否可以提供“真正的多因素安全性”时，这些方案通常执行简化的安全性证明，并获得类似等式 1 的安全性结果。13 (参见[Y]的定义 lang et al. 2008])，假设除了密码因素之外的其他因素已经被破坏。如上所述，我们的理论不鼓励这种简单但不现实的、实际上误导性的(即虚假的安全感)形式的表述。像我们提出的等式这样的公式。14 对于这种情况更准确和合适。这进一步表明了必要性

[5] 我们注意到，在[We note that, in Sections 5.2Abdalla et al. 2015b], *m 被重新定义为密码的熵。这种不一致将导致安全保障方面的巨大差异。我们推测那里出现了错别字。*

放弃属性“DA2-Local-Secure”并要求用户通过与服务器交互来改变他们的密码(即，更喜欢属性“DA2-Interactive”[Wang et al. 2014])，回答了[Wang et al. 2014]:当一个理想方案达到所有标准(包括 10 个可取的属性和 9 个安全目标)是无法实现的，那么应该放弃哪个标准？

据我们所知，我们第一次注意到了口令和基于口令的认证协议之间的结合。有了密码的确切分布的知识，我们设法开发一个更准确、更现实和更通用的公式来描述基于密码的认证协议的形式安全结果。这里，我们主要将基于口令的认证作为案例研究，并且可以容易地发现，我们在这里揭示的结果也可以容易地应用于其他类型的基于口令的密码协议，其安全性公式本质上依赖于口令分布的显式假设，例如基于口令的加密(例如[Juels and Ristenpart 2014])、基于密码的签名(例如[Gjosteenand Thuen 2012]) 和密码保护的秘密共享(例如，[Bagherzandi et al. 2011]).

5. 密码数据集的强度度量

在本节中，我们将解决如何准确测量给定密码数据集强度的问题。作为我们的 Zipf 理论的一个实际应用，提出了一个优雅而精确的基于统计的密码数据集强度度量。

5.1. 我们的指标

通常，一个聪明的离线猜测攻击者，[6]总是首先尝试最可能的密码，然后尝试第二个最可能的密码，依此类推，直到匹配目标密码。在极端情况下，如果攻击者也获得了明文形式的整个密码数据集，因此她可以获得正确的密码顺序，这种攻击称为最优攻击[Dell' Amico et al. 2010;Bonneau 2012b].[7]因此，我们可以使用破解结果

λ ∫ (n)是给定密码数据集的强度度量：

$$\lambda_{\int}(n) = \frac{1}{|DS|}\sum_{r=1}^{n} r$$

其中是密码数据集的大小，n 是猜测的次数。

在截面中 3，我们已经证明了密码的分布服从 Zipf 定律，即，

因此， λ ∫ (N)本质上由N和s决定(注意N是

唯一密码的数量，s 是拟合线斜率的绝对值)：

$$\lambda_{\int}(n) \approx \lambda(n) = \frac{\sum_{r=1}^{n} r^{-s}}{\sum_{r=1}^{N} r^{-s}}. \qquad (17)$$

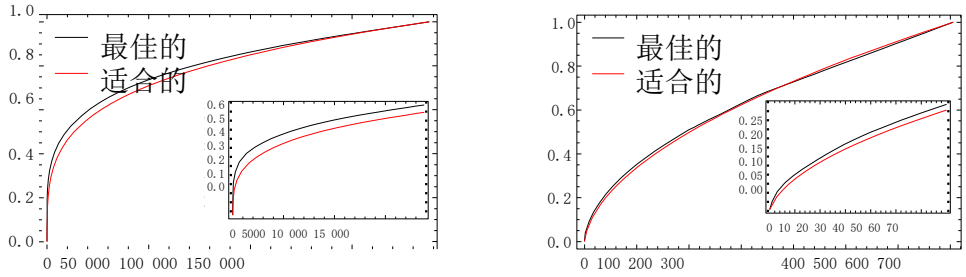
应该注意的是，在等式中。17， λ ∫(n)不完全等于最右边的值，尽管我们的回归线与实际数据非常吻合。我们根据等式绘制 λ∗(n)作为 n 的函数。16 并且根据等式 λ(n)是 n 的函数。17，并把这两条曲线放在一起，看看它们是如何相互吻合的。在图。6(a)，我们描绘了来自天涯数据集的 3023 万个密码的 λ ∫(n)和 λ(n),并且获得了这两条曲线的 1.32%的平均偏差(即，声音拟合)。由于空间原因

限制条件，这里我们不能举例说明其他数据集(如天涯和 Myspace)的相关图片，但我们总结了表中每个数据集的两条曲线 λ ∫(n)和 λ(n) (1 ≤ n ≤ |DS|)之间的平均偏差 VIII。

---

6 本节中提到的攻击都是离线攻击，因为我们的目的是测量整个数据集的强度，通常以加盐哈希 (或未加盐哈希)中的密码可以成功恢复的百分比来表征(参见第 The2.2).

7 请注意，最佳攻击具有理论价值(即提供上限)来描述攻击者可以采用的最佳攻击策略。在实践中，如果攻击者已经获得了所有的明文密码，她就没有必要命令这些密码来自己破解。



(a) Dodonew 的平均偏差:1.76%　　(b) Myspace 的平均偏差:0.88%

图 6。最佳攻击与我们在两个示例数据集(Dodonew 和 Mysapce)上的度量的一致性

从表中可以看出 VIII 对于每个数据集， λ ∫(n)曲线与 λ(n)曲线很好地重叠。具体来说，除了 Hak5，平均偏差都低于 2%(即从 0.54%到 1.93%)，表明 λ(n)与最优攻击结果 λ ∫(n)具有良好的一致性。如同图。6，当 n 较小时，每个数据集的两条曲线首先稍微偏离，然后随着 n 的增加逐渐融合。这主要是由
前几个高频密码与 Zipf 拟合线的偏差(见图。3).

表八。每个数据集的 λ ∫(n)和 λ(n) (1 ≤ n ≤)之间的平均偏差

天涯 Dodonew	CSDN	多玩 Myspace	Singles.org	Faithwriters	Hak5
平均离差	1.32%	1.76%	1.93%		
0.86%	0.88%	1.43%	0.54%	3.05%	

既然最佳攻击可以用 λ(n)很好地近似，那么很自然地建议将对(NA，sA)作为度量密码数据集 A 强度的指标，其中 NA 是回归中使用的唯一密码的数量， sA 是拟合线斜率的绝对值。请注意，本质上，测量密码数据集
相当于衡量创建该数据集的策略。在下文中，我们提出了一个定理和一个推论，并表明我们的度量不仅能够确定一个网站的密码数据集的强度在一段时间后是否变弱，而且还可以用于比较来自不同网站的数据集的强度。这一功能相当吸引人，因为只有在比较之后才会对安全性有信心——通过与其他类似网站进行比较，安全管理员现在对他们的数据集能够提供的强度水平有了更清晰的了解。最近一连串灾难性的网络账户泄露(见[Katalov 2013]为不完整的列表)提供了精彩的材料，以促进这种比较。

定理 5.1。假设 NA ≥ NB，sA ≤ sB。然后

$$\lambda A(n) \leqslant \lambda B(n),$$

其中 0 n NA(如果 n > NB，定义 λB(n) = 1)。如果上述两个条件中的任何一个不等式是严格的，那么 λA(n) < λB(n)，其中 0 < n < NA。

这个定理将在第一节中得到证明 5.2，在第节中 5.3 它与破解结果的一致性将通过模拟的最优攻击和最先进的破解算法(即基于马尔可夫的[Ma et al. 2014])。

推论 5.2。假设 NA ≤ NB，sA ≥ sB。然后

$$\lambda A(n) \geqslant \lambda B(n),$$

其中 0 n NB(如果 n > NA，定义 λA(n) = 1)。如果上述两个条件中的任何一个不等式是严格的，那么 λA(n) > λB(n)，0 < n < NB。

这个推论成立是因为它正是定理的逆否定命题 5.1。

上述定理和推论表明，给定两个密码数据集 A 和 B，我们可以首先使用线性回归获得它们的拟合线(即 NA、s A、NB 和 sB)，然后分别比较 NA 与 NB、sA 与 sA。这就产生了四种情况:(1)如果 NA NBand sAsB，数据集 A 比数据集 B 强；(2)如果 NA NBand sA sB，A 弱于 B；(3)对于 NA NB、sA sB 或 NA NB、sA>sB 的其余两种情况，λA(n)和 λB(n)之间的关系在变量 n，因此它是非确定性的(即无法得出直接结论)。在…里
在这种情况下，我们可能不得不画出曲线(搜索空间 N 对成功率),其中 N 的范围从 1 到 N，类似于其他方法，例如基于破解的方法(例如，基于 PCFG 和基于马尔可夫的方法[Ma et al. 2014]).注意，在所有四种情况下，基于统计的 α 猜测[Bonneau 2012b]是不确定的，即，它固有地参数化成功率 α（例如，关系 G0.49(A) > G0.49(B)永远不能确保 *G0.50(A) G0.50(B))*。Bonneau [2012b] 告诫说“我们不能依赖任何单一的价值观对于 α，每个值都提供了关于一个完全不同的攻击场景的信息。”
从这个角度来看，我们的衡量标准更加简单。我们还发现并修复了 α-猜测的强度转换中的一个固有缺陷，详情可在附录中找到 A.1。

一些言论。注意，与 NIST SP800- 63-2 文件中推荐的熵度量一样[Burr et al. 2013] 以及[Bonneau 2012b]，我们的度量主要对明文或非加盐哈希的密码数据集有效，而不适用于加盐哈希的密码。这是所有基于统计的度量的固有限制(例如，[Burr et al. 2013; Bonneau 2012b]和我们的)。对于加盐散列密码，需要求助于基于攻击的方法(例如[Kelley et al. 2012;Ma et al. 2014])，尽管这是以降低准确性为代价的(我们将在第节中展示)5.3 当前形式的基于攻击的方法具有太多的不确定性，并且远非理想)。同样值得注意的是，可能存在导致良好度量的弱策略，比如要求用户键入他们的用户名作为密码的开头。显然，这将使所有密码更加独特，并导致更好的度量，但如果攻击者知道底层策略，这根本不会增加密码的抵抗力。这构成了基于统计的度量的另一个限制。在这种情况下，还需要求助于基于攻击的方法。

然而，由于几个原因，这些和其他限制并不影响我们的度量的适用性。首先，我们的度量可以依赖于整个数据集的子集，并且仅涉及要在相对长的时间段(例如，一年)之后执行的离线操作，因此网站可以实现在线的加盐密码来认证用户，并且在物理上离线并且受到良好保护的非加盐哈希中维护密码子集，以方便我们的测量。第二，最近的入侵事件清楚地表明，拥有未加盐密码存储的互联网规模的网站远非罕见的例外。最有说服力的证据在于，大多数先前泄露的数据集来自著名的 IT 公司或领先的组织(如脸书、LinkedIn、Adobe、Dropbox、IEEE，仅举几例[Katalov 2013]) 仍然是未加盐的形式。现在，是这些遗留网站采取行动的时候了，其中一个重要的部分是访问其密码政策的安全条款。我们的指标是正确的选择。第三，众所周知，许多国家的当局(例如美国的 NSA)一直要求互联网提供商和网站向他们提供用户密码数据集(明文形式)[McCullagh 2013]. 在这种情况下，这些网站还应保留一份未加盐密码的副本，以确保符合规定。最后但同样重要的是，即使没有来自真实网站的明文(或未加盐散列)密码可用，现场实验(例如，[Egelman et al. 2013])可用于收集用户生成的密码。有了这些字段密码，我们的指标可用于帮助密码策略设计者和安全管理员在给定密码策略投入实际使用之前评估其安全性。

简而言之，尽管有其局限性，我们的度量在许多现实场景中是实用的。此外，如前所述，在四分之二的情况下，我们的度量具有无与伦比的优势

由于其确定性特征，与现有技术的度量标准(例如，基于攻击的[Ma et al. 2014]和基于统计的 α 猜测[Bonneau2012b])。然而，它在剩下的两种情况下是不确定的，在这两种情况下，我们必须画出 λA(n)的整个曲线和 λB(n),其中n的范围从1到最大NA，NB，这非常类似于“猜测”基于攻击的方法中的“曲线”[Ma et al. 2014]和 α －猜 测 [Bonneau 2012b]. 我们强调，只有当底层分布遵循Zipf定律时，我们的指标才是可行的，并且优于这些现有的指标，而在其他情况下(密码分布偏离Zipf)，这些现有的指标只是派上了用场。

5.2. 定理的证明

显然，当NA = NB，sA = sB时，该定理成立。首先我们在下证明这个定理条件 sA = sB = s，NA > NB。回想一下 fr = C，我们用概率来表示秩为r的密码是p (= fr = C). 那么 ΣNA CA = 1， ΣNB CB = 1，并且

λA(n) λB(n) = = CB。所以当 1 ≤ n ≤ NB时.

当NB + 1 ≤ n ≤ NA —1、我们可以得到

λA(n) λB(n)< 1 1 = 0。

接下来我们在 NA = NB = N，sA < sB，

0 < CA = 1 ... = CB。

当1≤N≤n1时，

λ(n) λ(n) = ΣCA ΣCB = C C (Σ 1 Σ 1 − Σ 1 Σ 1)

=C C (Σ 1 Σ 1 + Σ 1 Σ 1 − Σ 1 Σ 1

$$\sum_{r1=1}^{} \quad sA \quad sB \quad 1\ r2=1\ 2$$

$$= C\ C(\sum 1((R1)sA\ sB\ 1))。$$

对于 r1 > r2，sA < sB，所以(rA)sA sB < 1。此外，我们还有

$$\lambda A(n)\ \lambda B(n) < 0。$$

现在唯一剩下的情况就是 NA > NB，sA < sB。我们选择一个密码数据集 C 满足条件 NC = NA，sC = sB，则

$$\lambda A(n) < \lambda C(n)\ 1 \leqslant n \leqslant NA\ \ 1$$

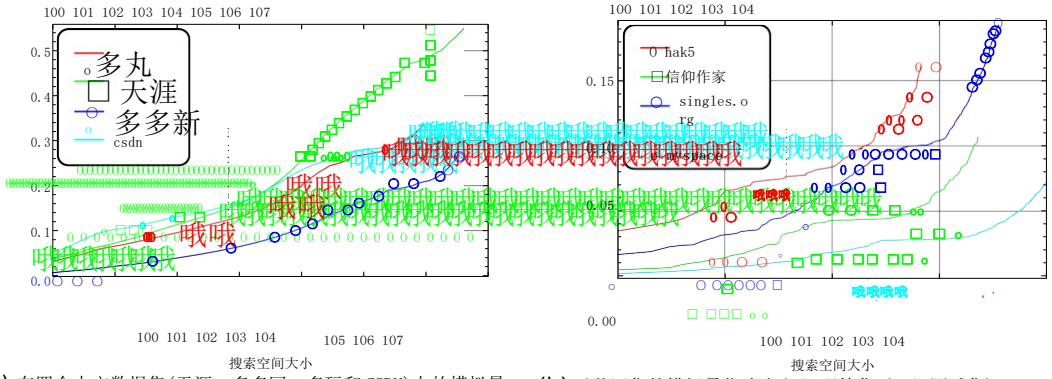
$$\lambda C(n) < \lambda B(n)\ 1 \leqslant n \leqslant NA\ \ 1$$

因此 λA(n) < λB(n)。这就完成了证明。

5.3. 实验结果

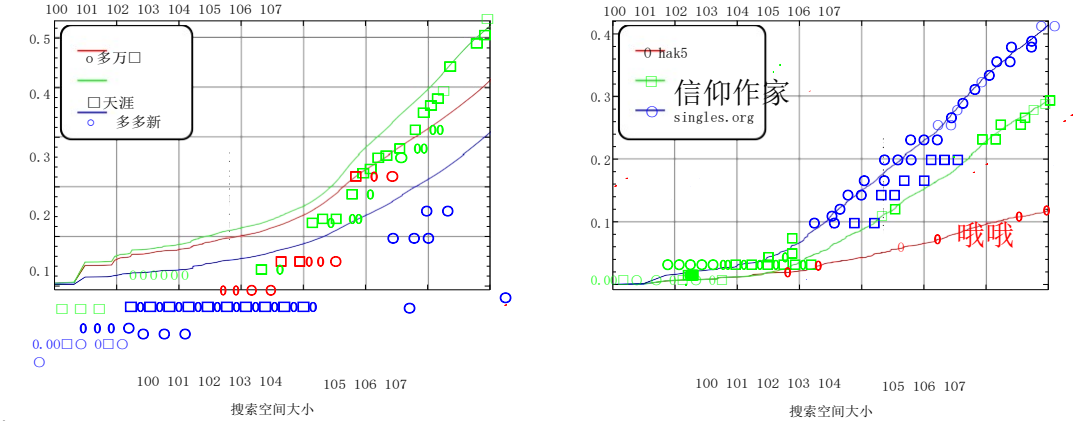
在这一小节中，我们进一步使用模拟的最优攻击和最先进的密码攻击算法来显示我们在第节中的度量 5.1 实际上是有效的。因为基于马尔可夫的破解算法通常比基于 PCFG 的算法执行得更好[Ma et al. 2014]，这里我们更喜欢基于马尔可夫的算法。

由于最优攻击作为任何真实攻击的最终目标在理论上是重要的，所以它决不能被视为现实攻击，因为它假设



(a) 在四个中文数据集(天涯、多多网、多玩和CSDN)上的模拟最优攻击

(b) 对英语集的模拟最优攻击(1/4 训练集对 1/4 测试集)



(c) 中文数据集上基于马尔可夫的攻击(以 CSDN 为训练集，退避平滑和结束符号归一化)

(d) 基于马尔可夫的对英语数据集的攻击(Myspace 作为训练集，退避平滑和结束符号归一化)

图 7。在不同数据集组上模拟最优攻击和基于马尔可夫的攻击

攻击者拥有目标身份验证系统的所有明文密码。为了查看我们的指标是否与实际攻击一致，我们稍微放宽了这一假设，并假设攻击者已经获得了目标系统四分之一的纯文本帐户(密码)，并使用它们以任何形式(加盐哈希或未加盐哈希)猜测其余三分之一的用户帐户(这是帐户总数的另一个四分之一)。请注意，这一新的假设要现实得多，因为本文中提到的大多数被入侵的网站已经以纯文本的形式泄露了很大一部分帐户。因此，这种新的攻击场景相当实用，我们称之为“模拟最优攻击”。为了更好地展示，我们将八个主要数据集分为两组:[8] 数据集大小都大于一百万的第一组和小于一百万的第二组。第一组的模拟最优攻击结果如图所示。7(a)，第二组的结果如图所示。7(b)。不难看出，对于同一组中的任意两个数据集，攻击结果符合我们在表中列出的度量结果 V。例如，从图。7(a) 我们知道，对于任何搜索空间大小(即每 n)，数据集 Duowan 都弱于数据集 Dodonew，这意味着 Ndodonew > Nduowan，sdodonew <sduowan。这一含义与表中的统计数据相符 V。

此外，我们执行更真实的猜测攻击(即基于马尔可夫的攻击)评估我们指标的有效性。在模拟的最优攻击中，我们根据大小和语言将八个主要数据集分成两组。对于中国组，我们使用 CSDN 作为马尔可夫训练集；对于英语组，我们使用

8 如前所述，由于空间限制，四个辅助数据集(见表 AsI)仅被证明是 Zipf 分布的，实际上，在本工作中揭示的所有其他一般性质也由它们保持。

Myspace 作为马尔可夫训练集。如[Ma et al. 2014], 主要有三种平滑技术(即拉普拉斯、Good-Turing 和 backoff)来解决数据稀疏问题，以及两种归一化技术(即基于分布和基于结束符号)来解决不平衡的密码长度分布问题。马等人发现，将退避平滑与基于结束符号的归一化相结合的攻击方案执行得最好，因此我们采用这种方案。这两组密码的破解结果如图所示。7(c) 和图。7(d)，分别为。

测试表明，基于马尔可夫的攻击在大多数数据集上的结果是
与我们的指标一致，唯一违反我们的指标的例外是数据集 Hak5。根据表格 V，NHak5 小于任何其他数据集，而 sHak5 大于同一组中的任何其他数据集，这意味着 Hak5 是最弱的一个。然而，图。7(b)结果表明，在基于马尔可夫的猜测攻击下，Hak5 是三个英语测试集中最强的。这种不一致可能是因为真实密码数据集的非代表性，或者是因为我们为基于马尔可夫的猜测攻击选择的训练集不合适。

特别感兴趣的可能是我们的观察，基于马尔可夫的攻击似乎比模拟的最优攻击效果差得多。例如，在 105 次猜测时，基于马尔可夫的攻击在中国数据集上实现的成功率为 14.5% 28.1%，低于模拟最优攻击的成功率。这种差距在英语数据集上更加明显。不应该
令人惊讶的是，成功率的差距是由于破解算法的固有弱点——它们的性能在很大程度上取决于训练集、平滑/归一化技术以及外部输入字典的选择，而这些选择受到太多不确定性的影响。这解释了为什么为了达到更高的成功率，我们根据人群将数据集分为两组，使用不同的训练集，并在基于马尔可夫的实验中特别选择平滑/归一化技术。这也强调了使用经验攻击结果的内在局限性(例如[Weir et al. 2010; Kelley et al. 2012; Castelluccia et al.2012]) 作为密码数据集的强度度量，暗示了我们度量的必要性。简而言之，仍然有开发更实用的攻击算法的空间，这些算法具有更少的不确定性，但更有效。

6. 结论

在这项工作中，我们提供了一个新的用户生成密码的分布前景。通过采用计算统计学的技术，我们第一次证明了 Zipf 定律简明地描述了密码的偏斜分布。我们进一步研究了我们观察结果的普遍适用性，并讨论了理解密码分布的多种好处。特别是，大多数现有的 PAKE 协议(成千上万，一些值得注意的包括[Chen et al. 2014;Canetti et al. 2012; Katz et al. 2009; Bellare et al. 2000])在口令均匀分布的假设下已经被证明是安全的，然而我们已经表明，它们的安全结果公式未能捕捉到现实生活中攻击者的实际优势，并且可能具有一些意想不到的后果。因此，我们提出了一个修正案，以更准确地描述 PAKE 协议的形式安全结果。

除了理论上的兴趣，我们还展示了 Zipf 理论的实际应用
提出了一种新的基于统计的密码数据集强度度量标准。我们的指标在准确性方面优于大多数现有的基于统计的指标(例如[Burr et al.2013]) 并且在四分之二的情况下，甚至在简单的情况下(例如[Bonneau 2012b]).最令人感兴趣的是其数据集强度测量的确定性，这有助于不同数据集之间更简单和精确的强度比较。我们已经以数学上严格的方式正式证明了我们的度量，并且还修正了 α-猜测的强度转换中的一个固有缺陷 Bonneau 2012b].通过在我们的大规模语料库上进行广泛的破解实验，我们进一步评估了我们的度量的有效性，并展示了它的实用性。我们相信这项法律的公布也是其他领域的兴趣所在，这项工作为他们的研究奠定了基础

进一步的理论发展和实际应用(例如，最近的“基因卫士”[Huang et al. 2015]依赖于我们的理论定律和数值结果)。

在这个有趣、重要而又具有挑战性的课题上，还有更多工作要做。比如导致 Zipf 定律出现在像用户生成认证凭证这样的混沌过程中的底层机制是什么？随着时间的推移，一个系统的密码分布会如何变化(演变)？极高价值的账户(如电子银行密码)是否遵守 Zipf 定律？这是一件喜忧参半的事情，随着更多的高价值网站被入侵，更多的数据集被公开，未来进行此类调查的机会只会增加。

这项工作的发现还提出了许多其他问题，它们可能需要全面的实地研究。例如，当我们为采用某种基于流行度的密码策略的必要性提供了合理的理由时，我们应该如何设置流行度阈值呢？并且，对于一个特定的阈值，可用性在实践中会受到多大程度的影响？多因素认证协议有必要放弃支持用户在不与服务器交互的情况下修改密码的特性吗？通过应用机器学习的机制(例如，线性回归)，Zipf 理论有望赋予系统安全中的密码使用以数学上的严密性(参见第 10．2．1 节)。4.1).与此同时，Zipf 理论将传统上属于系统安全范围的基于流行度的密码策略的创造性防御策略引入到了密码学中(参见。4.2).因此，这将引发关于密码研究进展(如安全性、可用性和管理)对基于密码的密码学领域(如认证、加密和签名)的重要影响的讨论。

附录

A.1 .发现并修复 α-猜测的强度转换中的固有缺陷

为了克服现有密码强度度量中的各种问题(例如，不可比性、不准确性和不可重复性)，Bonneau [2012b] 提出了 α 猜测法，它依赖于密码的统计分布，并以攻击者期望的成功率 α 为参数。它很好地抓住了这样一个事实，即一个实际的攻击者通常满足于破解帐户的薄弱部分。这一指标已在学术界广泛使用[Chatterjee et al. 2015; Li et al. 2014; Bailey et al. 2014] 还获得了NSA 2013年度“安全竞赛科学奖”[NSA Press Release 2013].在这里，我们报告了一个内在的缺陷，在它的力量转换，并进一步设法修复它。

为了更好地理解，这里我们遵循[Bonneau 2012b]越近越好。概率分布用X表示，每个密码xi是随机的
以概率pi从X画出Σ，例如

那 *pi=1。不失一般性，因为-*
pi≥pi2≥n . *其中n为*
的{ ssibΣle } X .中的事件总数

$0<\alpha\leq 1$, $\alpha(X)=$
确定A至少需要破解总密码的一部分 α 的每个账户的最小固定猜测次数，以及 λ β (X)= 1

$\sum_{i=1}^{\beta}$ *= pi 表示A的预期成功*
每个帐户只能猜测 β 次。因此，当给定每个帐户的 α 次猜测和 λ α α 时，λ α 测量的实际成功。使用这些术语，α-猜测被定义为:

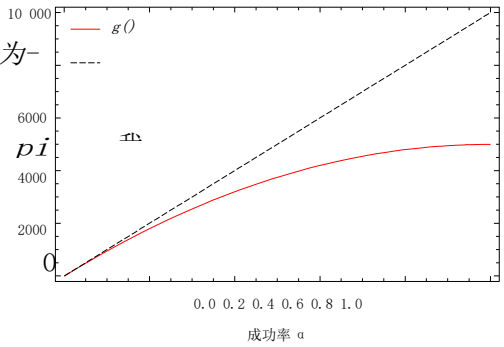


图8:gα(N和α(n如何随 α 变化

pi i(x)=(h8) α) α +

*Gα ()描述了每个帐户达到成功率 α 的预期猜测次数。情商的直觉。*18 是:(1)对每一个账户不在她的字典，她会作出 α 猜测，导致第一项；以及(2)针对字典中的所有账户，她以最佳顺序进行，并且所需的预期猜测次数构成了第二项。Gα ()很好地模拟了真实世界攻击者的现实，他们关心成本效益，以阻止对最强帐户的攻击。

为了更容易与其他现有指标进行比较，也为了程序员和密码学家更好地理解，Bonneau [2012b] 通过计算“具有相同猜测度量值的离散均匀分布(pi = 1/对于所有1 i)的对数大小”，进一步将Gα (X)转换为比特单位(即Gα ())。因为想要破解账户比例 α 的攻击者将“每Gα /α 次猜测获得一次成功猜测”，将“每(+ 1)/2次猜测破解一个账户”。这给出了公式(参见的第19页[Bonneau 2012a]获得更详细的解释):

$$G\alpha(X) = N + \quad g(X) = \mathrm{LG}\ N = \mathrm{LG}[2g\alpha(X)1] \tag{19}$$

*对于任何均匀分布的U ， g\α (x)应该是常数，但是*Bonneau [2012b]
发现事实并非如此。所以，他加上了“修正系数”lg 1 到G \uα (X)，给出:

$$g(X)=LG[2g\alpha(X)1]+LG \tag{20}$$

然而，我们将证明等式左边的等式。19 天生就有缺陷。从图2(a)中可以看出[Bonneau 2012b]，有人认为是Gα (U) = α (U)。恰恰相反，我们的图7很好地充当了一个具体的反例，即Gα (U104)= α (U104)。本质上，根据等式。18，一个人可以得到

$$G(U) = \sum_{i=1} + (1 - \lambda) \frac{(1+\alpha)\alpha}{\lambda} + (1 - \lambda) \quad (21)$$

$$=$$

另一方面，根据[Bonneau 2012b]，我们得到了

$$\alpha(U) = N \lambda \alpha(U) \quad (22)$$

基于情商。22，情商。21 可以改写为

$$G(U) = (1+N\lambda\alpha)N\lambda\alpha + (1\lambda)N \quad (23)$$

$$2N \qquad \alpha \qquad \alpha + (2 \quad 2$$

来自 Eq。22 还有情商。23，很明显，Gα(U) $\alpha(U)$。基于情商。23，对于 U 和为了具有同等的安全性，我们得到

$$G(U) = G \quad (24)$$

(十) 等式 23

$$= 2g\alpha(x) - \lambda\alpha(u)$$

$$\alpha \qquad = N$$

注意，对于 $0<\alpha\leqslant 1$，$0\overset{\alpha}{\leqslant}\lambda\alpha(u) \text{——} \alpha <$ $\frac{(2\lambda\alpha(u))\lambda\alpha(u)}{-}$ 且 $0\leqslant\lambda\alpha(X)\alpha < pn$，其中 $pn\leqslant pn1\leqslant\leqslant P1$ 和 $\Sigma n1\ pi\ <\alpha\leqslant\Sigma n$ $pi = \lambda$。这表明 $1 \leqslant \lambda(X)\lambda\alpha(U) \leqslant qn$，给 $|\lambda(X)\lambda\alpha(U)| \leqslant \max\{1, qn\}$。注意，只有当 α 足够大(0.5 作为[Bonneau 2012b])，Gα()会显示优于 α(X)；qn 随着 α 的增加而减小。当 $\alpha \geq 0.2$ 时，$qn < 1$ 成立 X

我们的 12 个数据集。此外，对于人工生成的密码，通常 2012b;Li et 1000 215[Bonneau N

al. 2014].所有这些给出了这样的关系，当 α 足够大，$\lambda() \alpha \lambda()$. 因此，$\lambda()$ 和 $X\lambda()$ 都可以统一为 λ。这首次解释了为什么等式(10)和(11)中的 $\lambda \alpha$ [Bonneau 2012b]省去分配或。基于这个观察和情商。24，对于 $0 < \alpha < 1$，等式左边的等式。19 可以证明是不正确的:

$$g\alpha(x) = 1+n(2\lambda\alpha) = n+1 \quad (25)$$

$$\lambda\alpha 2 \qquad 2$$

只有当 $\alpha = 1$ 时，因为 $1 = \alpha \leqslant \lambda \leqslant 1$，λ 才会等于 1，Gα( ) = +1。

此外，使用等式。24，Gα()的“有效密钥长度”(即比特强度)可以自然地公式化为

$$g(X) = LG\,N = LG\,2g\alpha(X) - \lambda\alpha = LG[2g\alpha(X) - 1] + LG \quad (26)$$

由此可见，在 α-猜测的强度转换中不需要添加 X 为的“修正系数”，从而证明并修正了 [Bonneau2012a; Bonneau 2012b]. 而有效密钥长度度量 Gα(X)是压倒性的优于 Gα(X)(例如，[Chatterjee et al. 2015; Bailey et al. 2014; Li et al. 2014]) 而且它普遍认为 Gα(U) = α(U)，我们的上述贡献不仅在于发现并修正了 g \uα(x)推导中的一个固有缺陷，而且同样重要的是，还在于

揭示了一种反直觉的关系·Gα(U) $\alpha(U$

参考

米歇尔·阿卜杜拉、法布里斯·本哈穆达和菲利普·麦肯齐。2015 年 a。J-PAKE 口令认证密钥交换协议的安全性。进行中。IEEE S&P 2015。IEEE，1–17。

Michel Abdalla、Fabrice Benhamouda 和 David Pointcheval。2015b。明文可校验攻击下的不可区分公钥加密。在 2015 年 PKC 奥运会上，j .卡茨(编辑。).LNCS，第 9020 卷。斯普林格，332–352。

拉达·阿·亚当。2014.Zipf，幂律和帕累托-排名教程。http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html。

安妮·亚当斯和玛蒂娜·安吉拉·萨斯。1999.用户不是敌人。Commun。美国计算机学会第 42 届会议，第 12 期(1999 年)，第 40-46

页。
C.艾伦。2009.3200万Rockyou密码被盗。http://www.hardwareheaven.com/news.php?newsid=526。
Mansour Alsaleh、Mohammad Mannan和P Van Oorschot。2012.重新审视对大规模在线密码猜测攻击的防御。IEEE Trans。可靠和安全的计算9，1 (2012)，128–141。

罗伯特·阿克斯泰尔。2001.美国公司规模的 Zipf 分布。科学 293，5536 (2001)，1818-1820。

阿里·巴格赞迪、斯坦尼斯劳·贾雷基、尼泰什·萨克塞纳和严斌·卢。2011.密码保护的秘密共享。在…里 *继续。CCS 2011。ACM，433 – 444。*

丹尼尔·V·贝利、马库斯·杜·姆思和克里斯托弗·帕尔。2014.金融账户密码复用和适配强度统计。进行中。SCN 2014。LNCS，第 8642 卷。斯普林格，218 – 235。

米希尔·贝拉雷。1999.面向实践的可证明安全性。1999 年在 ISC。).LNCS 火山。 1561.柏林施普林格/海德堡，1-15。

米希尔·贝拉雷，大卫·波因特切瓦尔和菲利普·罗加威。2000.抗字典攻击的认证密钥交换。进行中。欧洲加密 2000，巴特普雷尼尔(编辑。).LNCS，第 1807 卷。斯普林格，139-155。

法布里斯·本哈穆达、奥利维尔·布拉齐、克莱恩·谢瓦利埃、大卫·波因特谢瓦尔和达米安·弗格诺德。2013.SPHFs 和高效单轮 PAKE 协议的新技术。在《密码 2013》中，冉·卡内蒂和胡安娜。加里(编辑。).LNCS，第 8042 卷。施普林格柏林/海德堡，449 – 475。

F.贝尔加达诺、b .克里斯波和 g .鲁福。1998.主动密码检查的高字典压缩。ACM Trans 告知。系统。安全。1，1 (1998)，3 – 25.

马特·毕晓普和丹尼尔·克莱因。1995.通过主动密码检查提高系统安全性。计算机与安全 14，3 (1995)，233 – 249。

J.博诺。2012 年 a。猜测人类选择的秘密。博士论文。剑桥大学。

J.博诺。2012 年 b。猜测的科学:分析 7000 万密码的匿名语料库。进行中。第 33 届 IEEE 研讨会。关于安全和隐私。IEEE，538 – 552。

J.博诺、c .赫利、p .奥尔肖特和 f .斯塔亚诺。2012.寻求取代密码:网络认证方案的比较评估框架。进行中。IEEE S&P 2012。IEEE，553 – 567。

罗恩·鲍斯。2011.密码字典。https://wiki.skullsecurity.org/Passwords。

艾伦·S·布朗、伊丽莎白·布莱肯、桑迪·佐科利和道格拉斯国王。2004.生成并记住密码。应用认知心理学 18，6 (2004)，641 – 651。

W.Burr、D. Dodson、R. Perlner、W. Polk、S. Gupta 和 E. Nabbus。2006 年 4 月。NIST sp 800-63-电子认证指南。技术报告。弗吉尼亚州 NIST 莱斯顿。

W.Burr、D. Dodson、R. Perlner、W. Polk、S. Gupta 和 E. Nabbus。2013 年 8 月。NIST sp 800-63-2-电子认证指南。技术报告。弗吉尼亚州 NIST 莱斯顿。

Ran Canetti、Dana Dachman-Soled、Vinod Vaikuntanathan 和 Hoeteck Wee。2012.基于不经意传输的高效口令认证密钥交换。进行中。PKC 2012，马克·费希林，约翰尼斯·布赫曼和马克·马努里斯(编辑。).LNCS，第 7293 卷。施普林格柏林/海德堡，449 – 466。

Claude Castelluccia、Markus Du rmuth 和 Daniele Perito。2012.来自马尔可夫模型的自适应密码强度计。进行中。NDSS 2012。1 – 15.

拉胡尔·查特吉、约瑟夫·博诺、阿里·朱尔斯和托马斯·里斯坦帕尔。2015.使用自然语言编码器的防破解密码库。进行中。IEEE S&P 2015。IEEE，1 – 18。

陈丽群，胡伟林和杨国民。2014.再论跨域的基于口令的认证密钥交换。ACM Trans 告知。系统。安全。16，4 (2014)，15.

索尼娅·基亚松和保罗·范·奥尔肖特。2015.量化密码过期策略的安全优势。设计、代码和密码学 (2015)。Doi:10.1007/s10623-015-0071-9。

亚伦·克劳塞特，科斯马·罗希拉·沙立兹和马克·EJ·纽曼。2009.经验数据中的幂律分布。 *暹罗评论 51，4 (2009)，661 – 703。*

卢西恩·康斯坦丁。2009.黑帽子拥有的安全专家。http://news.softpedia.com/news/Security-Gurus-0 被黑帽拥有-117934.shtml

卡内·德·卡纳维莱和穆罕默德·曼南。2014.从非常弱到非常强:分析密码强度计。进行中。NDSS 青奥会。互联网协会，1-16 岁。http://dx.doi.org/10.14722/ndss.2014.23268。

米（meter 的缩写））戴尔阿米科、p .米歇尔迪和 y .鲁迪尔。2010.密码强度:实证分析。进行中。INFOCOM 2010。IEEE 通信学会，983 – 991。

南设计师。1996.开膛手约翰密码破解。http://www.openwall.com/john/。

米（meter 的缩写））杜鲁斯。2013.有用的密码哈希:如何浪费计算周期的风格？进行中。第 22 届新安全模式研讨会 (NSPW，2013)。ACM，31-40 岁。

南 Egelman、A. Sotirakopoulos、I. Muslukhov、K. Beznosov 和 C. Herley。2013.我的密码会升到 11 吗？:密码计量器对密码选择的影响。进行中。2013 年的迟。ACM，2379 – 2388。

米查利斯·法鲁索斯、彼得罗斯·法鲁索斯和克里斯特斯·法鲁索斯。1999.互联网拓扑中的幂律关系。进行中。SIGCOMM 1999。美国纽约州纽约市 ACM，251 – 262。

迪内·弗洛伦西奥和科马克·赫利。2007.对网络密码习惯的大规模研究。在 WWW 2007 中。美国纽约州纽约市 ACM，657 – 666。

迪内·弗洛伦西奥和科马克·赫利。2010.安全策略从何而来？。进行中。第六届可用隐私和安全研讨会(SOUPS 2010)。美国华盛顿州雷蒙德市 ACM，1 – 14。

迪内·弗洛伦西奥、科马克·赫利和P·范·奥尔肖特。2014.互联网密码研究管理员指南。进行中。USENIX丽莎2014。35－52.

K.Gjosteen和O. Thuen。2012.基于密码的签名。在EuroPKI 2011中。LNCS，第7163卷。17－33.

南哈列维和科劳兹克。1999.公钥加密和密码协议。ACM Trans告知。系统。安全。2，3（1999），230－268.

科马克·赫利。2013.对于攻击者来说，什么时候定位是有意义的？IEEE安全。& Priv。11，2（2013），89－92.科马克·赫利和保罗·范·奥尔肖特。2012.承认密码持久性的研究议程。*IEEE安全与隐私10，1（2012），28－36。*

Shiva Houshmand和Sudhir Aggarwal。2012.使用概率技术建立更好的密码。在…里*继续。ACSAC 2012。美国计算机学会，109-118。*

黄心怡、杨翔、伊莉莎·伯蒂诺、周剑英和徐莉。2014.针对脆弱通信的鲁棒多因素认证。IEEE Trans。靠。安全。计算机。11，6（2014），568－581.

黄志聪、Erman Ayday、Jacques Fellay、Jean-Pierre Hubaux和Ari Juels。2015.GenoGuard:保护基因组数据免受暴力攻击。进行中。IEEE S&P 2015。IEEE，1－16。

菲利普·G·英格尔桑特和M·安吉拉·塞斯。2010.不可用密码策略的真实成本:在野外使用密码。进行中。第28届ACM计算系统中人的因素会议(CHI 2010)。ACM，383－392。

凯西·约翰斯顿。2013.为什么你的密码不能有符号。http://arstechnica.com/security/2013/04/why-your-password-cant-have-symbols-or-be-longer-than-16-characters/。

阿里·朱尔斯和托马斯·里斯坦帕。2014.蜂蜜加密:超越暴力界限的安全性。进行中。*欧洲密码2014。斯普林格，293－310。*

弗拉基米尔·卡塔洛夫。2013.雅虎！、Dropbox和Battle.net被黑:停止连锁反应。http://blog . crack password . com/2013/02/Yahoo-Dropbox-and-battle-net-hacked-stopping-the-chain-reaction/。

乔纳森·凯兹、拉斐尔·奥斯特罗夫斯基和默蒂·容。2009.使用弱口令的高效安全的认证密钥交换。美国化学文摘第57卷，第1页(2009年)，第1-41页。

乔纳森·凯兹和维诺德·维昆塔纳森。2013.基于轮最优口令的认证密钥交换。*密码学杂志26，4（2013），714－743。*

帕特里克·盖奇·凯利、萨兰加·科曼杜里、米歇尔·L·马祖雷克、理查德·谢伊、蒂莫西·维达斯、卢乔·鲍尔、尼古拉斯·克里斯汀、洛里·费思·克兰诺和胡里奥·洛佩兹。2012.再次猜测(一次又一次):通过模拟密码破解算法来测量密码强度。进行中。IEEE S&P 2012。IEEE，523－537。

丹尼尔·克莱因。1990.挫败黑客:对密码安全性的调查和改进。进行中。第二届USENIX安全研讨会。5－14.

萨兰加·科曼杜里、理查德·谢伊、洛里·费思·克兰诺、科马克·赫利和斯图尔特·谢克特。2014.心灵感应词:通过读取用户的思想来防止弱密码。进行中。USENIX证券交易委员会2014。591－606.

辛西娅·郭，萨沙·罗曼诺斯基和罗莉·费思·克兰诺。2006.基于助记短语的密码的人工选择。进行中。汤2006。美国计算机学会，67-78。

李志功、韩伟力和徐。2014.中文网络密码的大规模实证分析。进行中。USENIX安全2014。559－574.

J.很长。2011.没有技术黑客:社会工程指南，翻垃圾箱，和肩冲浪。Syngress。马致远，杨伟宁，，李宁辉。2014.概率密码模型的研究。进行中。*IEEE S&P 2014。IEEE，689－704。*

T.Maillart、D. Sornette、S. Spaeth和G. Von Krogh。2008.开源Linux发行中Zipf法律机制的实证检验。物理评论快报101，21（2008），218701。

D.马龙和k .马希尔。2012.调查密码选择的分布。进行中。WWW 2012。301－310.里克·马丁。2012.在中国大范围数据泄露的情况下。https://sg.finance.yahoo.com/news/Amid-Widespread-Data-Breaches-pennolson-706259476.html。

米（meter的缩写））Mazurek、S. Komanduri、T. Vidas、L. Bauer、N. Christin、L. Cranor、P. Kelley、R. Shay和B. Ur。2013.衡量整个大学的密码猜测能力。进行中。CCS 2013。美国计算机学会，173-186。

德克兰·麦库拉。2013.联邦政府要求网络公司交出用户账户密码。http://www.cnet.com/news/feds-tell-web-firms-to-turn-over-user-account-passwords。

J.米克。2014.俄罗斯黑客汇编了1000多万被盗的Gmail，Yandex，Mailru。http://www.dailytech.com/俄语+黑客+编译+列表+of+10+百万+被盗+Gmail+Yandex+Mailru/article 36537 . htm

R.莫里斯和k .汤普森。1979.密码安全:案例史。Commun。美国计算机学会22，11（1979），594-597。阿尔温德·纳拉亚南和维塔利·什马蒂科夫。2005.利用时空权衡对密码的快速字典攻击。进行中。CCS 2005。ACM，364－372。

米（meter的缩写））纽曼。2005.幂定律、帕累托分布和齐夫定律。当代物理学46，5（2005），323－351。

国家安全局2013年新闻稿。第一届年度最佳科学网络安全论文竞赛。http://cps-vo.org/group/sos/papercompetition2012。

菲利普·奥克斯林。2003.进行更快的密码分析时间-内存权衡。进行中。加密2003年，丹博纳(编辑。).LNCS，第2729卷。施普林格/柏林海德堡，617－630。

Outpost9.com 的实验室 2014。单词表。Outpost9.com 的实验室。http://www.outpost9.com/files/WordLists.html.

彼得·保罗和卡斯滕·施奈德。2003.一族新调和数恒等式的计算机证明。
应用数学进展 31，2 (2003)，359－378。

大卫·波因特切瓦尔。2012.基于口令的认证密钥交换。进行中。PKC 2012，马克·费希林，约翰尼斯·布赫曼和马克·马努里斯(编辑。).LNCS，第 7293 卷。施普林格/柏林海德堡，390－397 英镑。

Ashwini Rao、Birendra Jha 和 Gananand Kini。2013.语法对长密码安全性的影响。进行中。第三届 ACM 会议。关于数据和应用安全和隐私(CODASPY 2013)。ACM，317－324。

斯图尔特·谢克特、科马克·赫利和迈克尔·米森马赫。2010.流行就是一切:保护密码免受统计猜测攻击的新方法。进行中。HotSec 2010。1－8.

R.Shay、P. Kelley、S. Komanduri、M. Mazurek、B. Ur、T. Vidas、L. Bauer 和 L. Cranor。2012.正确的马电池钉书钉:探索系统分配密码的可用性。进行中。汤 2012。ACM，1－20。

R.Shay、S. Komanduri、P. Kelley、P. Leon、M. Mazurek、L. Bauer、N. Christin 和 L. Cranor。2010.遭遇更强的密码要求:用户态度和行为。进行中。汤 2010。ACM，1－20。

E.斯帕福德。1992 年 a。关于可重用密码选择的观察。进行中。USENIX 安全研讨会。1－14.

E.斯帕福德。1992 年 b。Opus:防止弱密码选择。计算机与安全 11，3 (1992)，273－278。

San-Tsai Sun、Eric Pospisil、Ildar Muslukhov、Nuray Dindar、Kirstie Hawkey 和 Konstantin Beznosov。2013.调查用户对网络单点登录的看法:概念差距和接受模型。ACM Trans 互联网技术。13，1 (2013)，1－35.

Blase Ur、Patrick Gage Kelley、萨兰加·科曼杜里、乔尔·李、迈克尔·马斯、米歇尔·马祖雷克、蒂莫西·帕萨罗、理查德·谢伊、蒂莫西·维达斯、卢乔·鲍尔和其他人。2012.你的密码怎么样？强度计对密码创建的影响。进行中。USENIX 安全 2012。65－80.

保罗·范·奥尔肖特和斯图尔特·斯塔布尔宾。2006.利用登录历史和人在回路中对抗在线字典攻击。ACM Trans 告知。系统。安全。9，3 (2006)，235－258.

王鼎、何德彪、王萍和朱兆贤。2014.分布式系统中的匿名双因素认证:某些目标无法实现。IEEE Trans。关于可靠和安全的计算(2014)。http://dx.doi.org/10.1109/TDSC.2014.2355850.

、建、、和。2015.补充数据:模拟给定(完美)Zipf 分布时，样本大小和最小频率阈值对线性回归的影响。(2015 年 3 月)。http://wangdingg . wee bly . com/uploads/2/0/3/6/20366987/模拟 zipf.pdf。

王鼎和王萍。2015.皇帝的新密码创建政策:对领先的 Web 服务的评估和抵抗在线猜测的作用。进行中。ESORICS 2015。斯普林格。

马特·韦尔、苏迪尔·阿格沃尔、迈克尔·柯林斯和亨利·斯特恩。2010.通过攻击大量泄露的密码来测试密码创建策略的指标。进行中。CCS 2010。ACM，162－175。

马特·韦尔、苏迪尔·阿格沃尔、布雷诺·德·梅德罗斯和比尔·格洛德克。2009.使用概率上下文无关文法的密码破解。进行中。第 30 届 IEEE 研讨会。关于安全和隐私。IEEE，391－405。

严建新、布莱克威尔、安德森和格兰特。2004.密码的可记忆性和安全性:实证结果。IEEE 安全与隐私 2，5 (2004)，25－31。

、黄丹青、王、邓小铁。2008.基于智能卡和密码的双因素双向认证。J. Comput。系统。Sci。74，7 (2008)，1160－1172.

荀毅、冯昊和伊莉莎·伯蒂诺。2014.基于身份的双服务器口令认证密钥交换。在…里
继续。ESORICS 2014，M. Kutylowski 和 J. Vaidya(编辑。).LNCS，第 8713 卷。斯普林格，257－276。

赵子明，安盖尔俊和洪欣胡。2015.图片手势认证:实证分析、自动攻击和方案评估。ACM Trans 告知。系统。安全。17，4 (2015)，1－37.

朱，严杰夫，，鲍，毛，和。2014.验证码作为图形密码——一个基于人工智能难题的新的安全原语。IEEE Trans。告知。取证安全 9，6 (2014)，891－904。

乔治·金斯利·齐夫。1949.人类行为和省力原则。艾迪森-韦斯利出版社。