密码中对Zipf定律的新观察

胡振铎和丁王

摘要：密码分布为各种密码研究奠定了基础，准确地描述密码特征受到了广泛的关注。在IEEE TIFS‘17，Wang等人。提出了采用金切片搜索（GSS）拟合方法的CDF-Zipf分布模型来寻找最优参数。他们的模型已被120多个与密码相关的研究所采用。在本文中，我们用一个有原则的方法来解决它们剩余的、基本的密码分布的拟合优度问题。首先，我们证明了最先进的蒙特卡罗方法（MCA，对于优度检验）的置信水平渐近收敛于0。通过对2.28亿个真实世界密码的实验，我们证实了Wang等人。.92的对样本量影响的猜想，即微小的偏差将导致大规模数据集的统计显著性。我们提出了绝对偏差和相对偏差指标，并发现这两个指标中1%的随机偏差足以拒绝CDF-Zipf。其次，我们试图减小经验分布和拟合分布之间不可忽略的差距（累积分布函数（CDF）的最大偏差平均为1.91%）。我们在两个坐标系中探索了8种不同的分布模型，发现有3种模型比CDF-Zipf更准确，但没有一种能通过MCA。特别地，我们揭示了拉伸指数，CDF-Zipf的一种变体，平均可以将最大的CDF偏差从1.91%降低到1.25%。第三，为了取代MCA，我们引入了一种新的基于对数似然的拟合优度测度。我们发现拉伸指数总是比它的对应指数有更大的对数似然。总之，拉伸指数法更适合密码，并进一步支持密码中的Zipf定律。

索引项-密码分布，Zipf定律，优度偏差测量，蒙特卡罗方法，拉伸指数。

Ⅰ.介绍

TEXTUAL密码仍然是基于互联网的认证[15]的主流形式。尽管安全缺陷在40年前就被发现了，并且已经提出了各种替代方案(e。g.，生物识别[31]和多因素认证[45])，密码由于非常简单，部署成本低，仍被广泛应用于认证中。因此，密码可能在可预见的未来[15]，[49]，[53]。

尽管密码无处不在，但它面临着一个困难的挑战[38]，[58]，[59]：真正的随机的密码很难记住，而大多数人类选择的密码是高度可预测的。在实践中，除了流行的密码(e。g.,，普通用户倾向于使用个人身份信息(e。g.，名称和生日[50]，[53]，[54])来构建他们的密码，和58%的～，79%的被调查用户重用（或轻微修改）密码跨网站[21]，[38]，[52]，[56]。这种脆弱的行为使得密码分布不均匀且高度倾斜。然而，有数百项研究(e。g.，[2]，[18]，[29]，[46])假设“密码是均匀分布”。这种不切实际的假设往往会导致对密码相关研究的安全性的低估或高估(e。g.，加密协议[16]、[55]、加密方案[2]、[22]和哈希函数[13])。这些事实强调了准确描述扭曲的密码分布的必要性。

由于人类语言遵循Zipf定律[60]，Malone和Maher [30]首次尝试用Zipf分布来描述人类选择的密码。对4个密码数据集（其中3个数据集的大小小于0）的概率密度函数（pdf）进行了实验。他们得出的结论是，这些数据集“不太可能是Zipf分布的”。2012年，Bonneau [14]也试图将密码数据集的pdf与Zipf定律相匹配，并得出了与Malone和Maher [30]相似的结论。为了弄清楚密码的分布情况，在2017年，Wang等人。[51]研究了密码数据集的累积分布函数（CDF，即pdf的求和）是否遵循Zipf定律，并提出了CDF-Zipf分布模型。他们使用了黄金部分搜索（GSS）拟合方法(一种数值优化方法，见Sec。II-C)找到密码分布的CDF（而不是PDF）的最优参数，并实验了14个3000到3200万大小的数据集。大量的实验表明，他们的模型不仅提供了最小的柯尔莫戈罗夫-斯米尔诺夫（KS）统计量(这是经验分布和拟合分布之间的CDF偏差的最大值，见Sec。，但也可以覆盖整个数据集。到目前为止，他们的模型已经被120多个与密码相关的研究所采用，如密码加密、策略、猜测和加密协议。[1]

经验的和拟合的CDF分布之间的KS统计量归因于两种类型的偏差[17]，[20]，[39]。*第1类偏差来自于统计随机性，这是其固有的，不能通过优化或使用更准确的模型来减少。*第2类偏差来自于使用了不适当的分布模型，可以通过使用更准确的模型[20]，[39]来减少甚至消除。如果KS统计量主要来自第1类偏差，且模型准确，则可以从(i。e.，下面)具有高置信水平[17]、[20]、[39]的拟合分布。否则，就有可能采用一个更合适的模型来减少甚至消除第2类偏差。

为了评估密码是否遵循一个选定的分布，我们需要一个名为拟合优度检验的独立评估来区分这两种类型的偏差[17]，[20]，[39]。如果第2类偏差不可忽略，则所选择的分布模型可能不是最优的，可以进一步改进[17]，

稿件于2021年11月16日收到；2022年4月22日修订；2022年5月10日接受。出版日期：2022年5月18日；当前版本的日期为2022年12月9日。国家自然科学基金资助项目62172240；国家天津自然科学基金资助项目21JCZDJC00190。协调对该手稿的审查并批准其出版的副编辑是Dr。IssaTrare。（*通讯作者：王丁*）

侯振铎，北京大学数学科学学院，北京100871(电子邮件：乔侯13@pku。edu.cn).

丁王曾就职于中国南开大学天津300350网络科学学院，以及中国天津300350天津网络和数据安全技术关键实验室(电子邮件：王丁@南凯。edu.cn).

数字对象标识符10。1109/TIFS.2022.3176185

[1]1124项研究的完整列表可以在https://bit上找到。ly/30hODA0.

[20]， [39].2017年，Wang等人。[51]使用了Clauset等人推荐的蒙特卡罗方法（MCA）。[20]来对他们的CDF-Zipf分布模型做这个测试。MCA的基本思想是首先生成一些(e。g.,104)合成数据集，通过使用相同的分布参数拟合的经验数据集，然后计算比例(i。e.，p值，见Sec。III-A)的第1类偏差大于经验数据集和综合数据集之间的KS统计量。如果这个比例很大(e。g.，>90%)，第1型偏差占优势；否则，第2型偏差不可忽略，这表明所检查模型的不适宜性。

在他们的MCA实验中，Wang等人。[51]发现上面的比例非常小(i。e.，$<10^{-4}$)，但推测这种现象是由于样本量的影响："给定一个足够大的样本，可以发现非常小和不显著的差异具有统计显著意义，统计意义没有说明差异[41]的实际意义。"尽管有这样的说法，王等人。*[51]把未完成的证明作为未来的工作。因此，一个自然的问题出现了：给定CDF-Zipf分布模型，MCA在多大程度上是大尺度数据集的一个不合适的拟合优度度量？*据我们所知，之前没有任何工作解决过这个问题。忽略这个关键问题可能会导致不正确的密码分配声明。

由于CDF-Zipf分布模型不能通过MCA，第二个问题出现了：是否有比CDF-Zipf具有可比性甚至更准确的分布模型？*特别是，这些模型是否能够通过MCA？*如果另一种模型可以实现这一点，那么密码比CDF-Zipf（i。e.，它的KS统计数据将会更小)。否则，如果所有这些更准确的替代模型（其中大多数是常用的）仍然不能通过MCA，那么很自然就会对MCA的内在有效性产生怀疑。在这种怀疑下，需要一个更合适的拟合优度测量。*这就提出了第三个问题：对于大规模密码数据集，是否有比MCA更合适的拟合度度量？*特别是，鉴于各种具有相当准确性的密码分布模型，哪种分布模型更有可能（和合理）遵循密码？总之，本研究首次关注了上述三个关键的研究问题。

A.我们的贡献

在本文中，我们做出了以下贡献。.**MCA定量分析。**我们提供两个钻机

多孔数学证明和大量的实验验证Wang等。他的[51]民间传说对样本量的影响。首次证明了在CDF-Zipf分布模型下，第1类偏差渐近收敛于0。在8个大规模密码数据集的子集上的实验结果证实了数学证明：当次采样数据集的大小为≥为0。2500万美元，MCA拒绝了CDF-Zipf。其次，我们提出了绝对和相对偏差指标来模拟真实世界的密码偏差。证明了KS统计量的最大值随偏差单调增加。有了这一理论保证，我们进行了广泛的实验，并发现在这两个指标中，1%的随机偏差足以使MCA拒绝CDF-Zipf分布。这表明MCA太敏感，不能作为拟合优度度量的有效。

.**替代密码分发。**我们调查四个

秩频坐标系统中的分布模型

tem，在频率-频率坐标系中，总共有8个替代模型来寻找是否有比CDF-Zipf更准确的模型。我们用Wang等人拟合了这些分布模型。的[51] GSS拟合方法。我们发现，两个系统的对数正态和秩频系统的拉伸指数正态与CDF-Zipf比较准确，CDF的最大偏差为1。25%到1。平均48%。

我们还为这些替代模型重新讨论了MCA。MCA

当数据集的大小很大时，无论准确性如何，都会拒绝所有它们(e。g.，≥100万)，但在规模较小时接受多个分布模型，并予以确认

MCA无效。

.**一种新的拟合优度测量方法。**我们介绍

对数似然比检验（LRT），以确定密码更有可能遵循哪个分布。特别是，我们研究了CDF-Zipf和其他三个相对准确的替代模型(i。e.，两个坐标系中的对数正态和秩频坐标系中的拉伸指数)。我们发现拉伸指数模型，CDF-Zipf的一个变体，有一个

显著且不断地增大(1。26×107～6.15×108)

对数似然比比其他三种模型都要高。此外，

其最大CDF偏差为0.46%～2.49%(平均。

1.而CDF-Zipf的比例较大

0.50%～4.54%（avg.1.91%)，特别是在六人出局时

在8个数据集中，拉伸指数更高

速率比CDF-Zipf。因此，拉伸指数可以

更准确地描述了密码分配的特征。我们

还比较了LRT和MCA，并表明LRT在最小化统计误差方面优于MCA。

.**一些见解。**我们得到了很多见解

从我们的理论和实验中，预期和一些令人惊讶的。令我们惊讶的是，我们发现一些分布（除了CDF-Zipf）很难用GSS进行优化，这可能来自于CDF表达式中超越函数中意想不到的奇点。正如预期的那样，数据集越大，密码的统计随机性就越小。

II .初步的

A.数据集和伦理

我们在本研究中使用了8个大规模的数据集，共计2.2892亿个真实世界的密码。如表一所示，从2009年到2020年，所有的数据集都是被黑客攻击或发布的，并且已经公开使用了一段时间了。使用这些数据集的理由有四个方面：(1)最近违反的密码数据集可以代表密码演变的最新趋势；(2)早期违反密码的数据集在其他工作中被广泛使用(e。g.，[12]，[51]，[54])，并可以使这项工作的结果可重复；(3)实际破解研究表明，密码进化缓慢的[14]，因此来自早期和最近被破坏的数据集的属性应该是相似的；(4)这些数据集具有不同的背景(e。g.，语言和服务类型)，并可以反映真实世界的密码。

我们也充分了解这项研究的伦理学。尽管这些数据集已经公开获得，并在以前的研究中广泛使用(e。g.，[12]，[51]，[54])，它们包含私有数据(e。g.，姓名和电子邮件地址)。在本研究中，我们只使用频率等汇总的统计信息，并对他人保密，因此使用这些数据集不会给用户带来额外的风险。最后，我们的研究旨在造福于学术界和工业界。

表i

密码数据集的基本信息

Dataset	Service type	Language	When leaked	How leaked	Total passwords
Yahoo†	Portal	English	Aug., 2013	Released	69,301,337
000webhost	Web Host	English	Oct., 2015	Hacked	15,251,073
Rockyou	Social Forum	English	Dec., 2009	Hacked	32,603,388
Tianya	Social Forum	Chinese	Dec., 2011	Hacked	30,816,592
Chegg	Education	English	Apr., 2018	Hacked	38,997,234
Mathway	Education	English	Jan., 2020	Hacked	16,524,045
Wishbone	Chatting	English	Jan., 2020	Hacked	9,171,560

† Yahoo was first collected by Bonneau [14] and later published by Blocki et al. [11] using differential private techniques. The data is available at https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937.

对于任何给定的具有|DS|密码的数据集，我们将N个唯一密码从最频繁到最小进行排序，并表示第i次唯一传递-的频率

在真实数据集RealSet为f^ *i*：用于密码 *PW1*、*PW2*、...、PWN，它认为$\hat{p}1 \geq \hat{p}2 \geq \geq \ldots \hat{p}N$，在哪里$\hat{p}_i = f\hat{}i/|DS|$为概率密度函数（PDF），RealSet的累积分布函数（CDF）为pdf的总和，i。e.，$\hat{P}_I = \sum_{i=1}^{r} \hat{p}_i$.同样，理论数据集TheoSet的CDF为*波多黎各*=$\sum_{i=1}^{r} pi$.

B.科尔莫戈罗夫-斯米尔诺夫统计学

柯尔莫戈罗夫-斯mirnov（KS）统计量[20]是两个累积分布函数（CDFs）之间的距离的最大值。e.，两个cdf之间的差值超过绝对值的最大值。因此，TheoSet和RealSet之间的KS统计量为

$$DKS = \max_{1 \leq r \leq N} \hat{P} |\mathrm{Pr} - \mathrm{r}|, \quad (1)$$

在哪里*脱氧酮类醇*∈ [0, 1].因为一个小*脱氧酮类醇*意味着一个更准确的描述，目标是最小化*脱氧酮类醇*通过搜索参数(e。g.，C和s为CDF-Zipf)。这种KS统计测量被广泛应用于非参数配件[20]，[37]，所以我们的使用是合理的。

C.金切片搜索拟合方法

王等人。[51]引入了黄金部分搜索（GSS）拟合方法来拟合密码数据集。其关键思想是找到一个以真实数据集RealSet为特征的综合理论数据集TheoSet。例如，在CDF-Zipf分布下，CDF是*波多黎各=贷款*，最小化KS统计量是

分*脱氧酮类醇*. (2)

$$C, s$$

GSS的细节见Alg。附录A中的4。通过这种方法，密码的CDF在数值上依赖于TheoSet中密码的频率。因此，建议运行Alg。4次多次。g.,100)，以最小化随机波动。

D.拟合性试验

由于CDF-Zipf分布模型可以以相对较高的准确性来表征密码分布，因此自然出现了一个问题：密码是否真的遵循它。在统计学中，通过检验分布模型是否成立，采用拟合优度检验来回答这个问题。

通常，拟合优度检验是一种假设检验方法，其方法如下。声称密码遵循X分布(e。g.，CDF-Zipf)被视为零假设*H0*.同样地，声称密码也没有遵循X（或遵循其他一些分布）被视为替代假设*H1*.一个简单的想法

是否*H0*或*H1*持有量是为了区分KS统计量的来源。如果KS统计量主要来自于统计随机性(i。e.，类型1偏差)，所选模型具有较高的置信水平，不太可能得到进一步优化；否则，如果KS统计量主要来自于分布模型的使用(i。e.，第2型偏差)，它很可能会被采用

减少KS统计量的替代模型。该思想被蒙特卡罗方法（MCA）所使用，用来确定是否a

分布模型具有较高的统计可信度。

如果上述简单的MCA不能证明CDF-Zipf的合理性，那么密码可能会遵循其他在对数-对数尺度下大致呈线性的分布。理想情况下，一个良好的分布模型X（不是CDF-Zipf）应该同时(1)提供较小的KS统计量，(2)得到优度检验结果的支持。否则，就需要一种新的方法来对密码数据集进行拟合优度测试。

假设检验存在两类统计误差。第1类统计误差是错误拒绝真的概率*H0*，即拒绝X（e。g.，第2类统计错误是错误接受错误的概率*H0*，当密码不遵循它时，它就会接受X。在本文中，我们将讨论他们在评估拟合优度措施MCA(见Sec。和对数似然比检验(LRT，见Sec。IV-D).

III .MCA的潜在问题分析

在本节中，我们首先使用严格的数学证明和广泛的实验来研究什么数据集的大小可以使MCA拒绝CDF-Zipf。然后，我们引入了两个度量标准来模拟真实世界的偏差，并进行了广泛的实验来找到拒绝CDF-Zipf的阈值。

A.MCA工艺的重新设计

如第二节所述。II-B，我们使用KS统计量来表示测量CDF偏差。首先，我们引入了MCA拟合优度度量的思想。由于理论数据集TheoSet具有分布参数(e。g.，C和s为CDF-Zipf)，拟合TheoSet的KS统计量(记为D'_{KS})可以被视为第1类偏差(i。e.，统计随机性）。CDF偏差(记为*脱氧酮类醇*)，在真实数据集RealSet和理论数据集TheoSet之间同时包含类型1和类型2的偏差。因此，如果D'_{KS}有D'_{KS}> *脱氧酮类醇*，CDF偏差主要来自于第1类偏差，密码应遵循分布模型。

更详细地说，我们做MCA如下：(1)使用黄金搜索（GSS）将RealSet与CDF-Zipf分布模型拟合，获得KS统计量*脱氧酮类醇*；(2)使用相同的分布参数来描述RealSet来生成*J0*理论数据集，我。e.，TheoSet1，• • •，TheoSetJ0；(3)使用GSS分别单独地拟合每个TheoSetj（1≤j≤J0），并计算相应的D'_{KS_j}(4)计算D的比例$'_{KS}$> *脱氧酮类醇*作为p值（分布模型的置信水平）；

我们在分母和分子上同时加上一个来使

p值平滑[23]；(5)我们将p值阈值设置为0.01：如果为p-value>0.01，则为原假设*H0*CDF-Zipf之后的密码应该被接受；否则，应被拒绝。MCA与CDF-Zipf的处理过程为

显示在Alg。1.

我们现在解释为什么我们选择p值阈值为0首先，就像阈值0.05 [25]，这个0.01有.01.

表二

参数，偏差的数量级，和使用MONTE CARLO方法（MCA）计算的p值

Dataset	C	$O(\sigma_C)$	s	$O(\sigma_s)$	D_{KS}	$O(\sigma_{D_{KS}})$	C'	$O(\sigma_{C'})$	s'	$O(\sigma_{s'})$	D'_{KS}	p-value
Yahoo	0.033148	10^{-4}	0.180907	10^{-4}	0.040775	10^{-5}	0.033207	10^{-17}	0.180752	10^{-16}	0.000298∼0.000697	$<10^{-4}$
Dodonew	0.019255	10^{-5}	0.211921	10^{-5}	0.004979	10^{-4}	0.019459	10^{-4}	0.211798	10^{-4}	0.000101∼0.000186	$<10^{-4}$
000webhost	0.005738	10^{-5}	0.282561	10^{-4}	0.005022	10^{-4}	0.005665	10^{-5}	0.283099	10^{-4}	0.000506∼0.001544	$<10^{-4}$
Rockyou	0.038208	10^{-5}	0.185939	10^{-16}	0.045357	10^{-4}	0.037543	10^{-6}	0.186960	10^{-15}	0.000310∼0.000693	$<10^{-4}$
Tianya	0.062337	10^{-4}	0.155266	10^{-4}	0.022925	10^{-5}	0.062095	10^{-4}	0.155464	10^{-4}	0.000200∼0.002294	$<10^{-4}$
Chegg	0.008297	10^{-4}	0.234966	10^{-4}	0.008617	10^{-4}	0.008186	10^{-5}	0.236053	10^{-4}	0.000043∼0.001214	$<10^{-4}$
Mathway	0.010541	10^{-5}	0.245255	10^{-4}	0.011059	10^{-4}	0.010548	10^{-5}	0.245325	10^{-4}	0.000413∼0.001594	$<10^{-4}$
Wishbone	0.017144	10^{-4}	0.230503	10^{-4}	0.014775	10^{-4}	0.017125	10^{-4}	0.230626	10^{-4}	0.000192∼0.002275	$<10^{-4}$

† $O(\blacksquare)$ denotes the orders of magnitude of the standard deviations. D_{KS} is the Kolmogorov-Smirnov (KS) statistic resulting from fitting the *real-world* dataset RealSet with C and $\blacksquare$, and D'_{KS} is the KS statistic (denoted in range) resulting from fitting the *theoretical* dataset TheoSet in the Monte Carlo approach (MCA). The results show that (1) $D_{KS} > D'_{KS}$; (2) $\blacksquare$-values $\blacksquare 10^{-4}$), and (3) $C' \blacksquare C$ and $s' \approx s$ hold for all datasets.

基于CDF-的算法1的蒙特卡罗方法 Zipf沙痂

输入：真实世界的数据集RealSet。

输出：置信度水平p值。

1开始

2 (C , s ，*脱氧酮类醇*) = GSS (RealSet);

3为j = 1到*J0*做

4=原（C , s, |DS|）; /*生成

*J0*理论数据集，|DS|是数据集的大小。*/

5 (C, s , D′′′KSj)= GSS（TheoSetj）; /*拟

采用与真实数据集相同的拟合方法。*/

6p值= (#{D′KSj|D′KSj>*脱氧酮类醇*，1≤j≤J0}+ 1)/（J0 + 1）; /*平

滑p值。*/

*7，如果p值>为0。*01然后

8.接受CDF-Zipf分布模型；

9其他

10.拒绝CDF-Zipf分布模型；

11输出：p值。

GSS是一种金剖面搜索拟合方法。4)由Wang等人使用。[51]. 利用[20]中的转换方法。附录A中的3）。

也被广泛应用于各种研究领域(e。g.，流行病学[26]，心理学[48]，和网络安全[19])。其次，上述定义p值的方法表明，阈值越小，MCA就越难拒绝CDF-Zipf。因此，使用0.01可以使我们的分析更加严格，并使用更小的1型统计误差：当p值<为0时。382和*H0*和*H1*以近似相等的先验概率保持不变(i。e.，P（H0）≈P（H1），通常是由Berger等人提出的。[9])，第1类统计误差e1（这是p值的函数）是

基于[43]。因此，通过将p值阈值设置为0.01而不是0.05，第1类统计误差可以从30+90(2)减少值16.67%(3使我们的结果更加可靠。此外，为了确保p值的准确性，我们取*J0*=10000（生成的数量），所以运行-多姆波动ep的p值为<0.005 (ep [20])。这也表明最小的p值为$\frac{1}{10,001}<10^{-4}$

(使用平滑时，参见Alg的第6行。1)，所有类型1的偏差（统计随机性）都小于KS统计量(i。e.，D'_{KS}<*脱氧酮类醇*).第三，我们也并不意味着要排除存在其他阈值的可能性。与之前的研究一样(e。g.，[32]，[36]，[44]，[47])，我们提供了详细信息，包括p值、第1型偏差和使用MCA计算的KS统计数据（见表II和表VIII），因此从业者可以设置他们自己的p值阈值(e。g.，0.005 [7]、[27]、0.05 [25]，和0。1 [20]).我们的理论还将揭示，对于足够大的数据集(e。g.，尺寸为≥，100万，

看到秒。*III-B)，确切的p值阈值与总体结论无关。*

表二显示了用CDFZipf分布拟合8个大规模密码数据集的p值和参数结果。我们可以看到：(1) D'_{KS}至少要小一个数量级吗*脱氧酮类醇*，因此，2型偏差总是主导KS统计量；因此，原假设*H0*遵循CDF-Zipf分布的密码被拒绝；(2)p值始终为<10^{-4}，*而MCA将不接受CDF-Zipf，无论p值阈值如何(e。g.，0.01、0.05和0.1)；(3) C和C*，*以及s*和s值非常接近，差异为C = |C−C|Δ*距离修正从10^{-7}至10^{-4}和s = |s−Δ*|范围从10^{-6}向

在MCA。

10^{-3}；在这种情况下，有理由处理C*≈C和s*≈ s

B.p值与数据集大小之间的关系

我们使用严格的数学证明和广泛的实验来展示p值如何随密码数据集大小的变化。因为p值是D的比例$'_{KS}$>*脱氧酮类醇*，我们关注类型1的最大偏差最大D'_{KS}随着数据集大小|和DS|的变化而变化。

1)理论：在生成理论集的过程中，每一个密码*PWi*可以看作是随机伯努利变量

可与平均p_i^m和标准偏差$\sqrt{p_i^m(1-p_i^m)}$，$\frac{m}{i}$

其中p_i^m的真实概率是什么码*PWi*由

频率f_i的*PWi*在|DS|采样后进行

二项分布与平均$\mu_i = p_i^m|DS|$和

分布模型(e。g.，CDF-Zipf) [14]，[51]。因此，standard deviation $\sigma_i = \sqrt{p_i^m(1-p_i^m)|DS|}$. Since $1-p_i^m < 1$, there is $\sigma_i < \sqrt{p_i^m|DS|} = \sqrt{\mu_i}$. For popular passw或d s with $f_i \geq$ fb (e.g., fb = 10)，使用f_i近似的μ_i是准确的

因为$\sigma_i/\mu_i < \sqrt{1/\mu_i}$.为不受欢迎的密码与$f_i$< fb，探索性实验显示$\sigma_i \to 0$.基于这些观察结果，我们有了以下定理。

*定理1：假设RealSet和C的C和s*和s*满足C≈C*和s*≈；*对于每个密码

$f_i \geq$fb，估计误差$e_i = f_i - \mu_i$遵循正常分布N（0，σ_i^2），以及密码为f的密码$_i$< fb ，$\sigma_i = 0$.在这种情况下，类型1的最大偏差(i。e.，统计随机性）最大D'_{KS}随着|DS|的增加而减少，并且有$\lim_{|DS|\to\infty} D'_{KS} = 0$.因此，p值随着|DS|的增加而降低，并且有$\lim_{J0\to\infty}$p值=0（见附录B中的证明）。

*2)讨论：首先，我们证明了定理1中的条件是现实的。*首先，如表二、C所示*≈C和s*≈适用于我们所有的8个大规模密码数据集，所以我们可以处理C*还有C，还有s*统计上是等价的。这意味着定理1不管C和s的确切值如何都成立。第二，正态性假设，

表三

使用MONTE CARLO方法（MCA）计算的密码数据集的子集的p值

Dataset	Subset size	ΔC	Δs	D_{KS}	D'_{KS}	p-value	Dataset	Subset size	ΔC	Δs	D_{KS}	D'_{KS}	p-value
Yahoo	$0.05M$	0.000033	0.005277	0.004660	0.000320∼0.008640	**0.054**	Tianya	$0.05M$	0.000001	0.000191	0.008620	0.000480∼0.012180	0.005
	$0.1M$	0.000091	0.010370	0.005752	0.000300∼0.007180	0.006		$0.1M$	0.000134	0.000654	0.007910	0.000390∼0.008400	0.003
	$0.25M$	0.000071	0.015701	0.007834	0.000500∼0.003880	$<10^{-4}$		$0.25M$	0.000123	0.000468	0.008195	0.000500∼0.003780	$<10^{-4}$
	$0.5M$	0.000085	0.008823	0.010275	0.000324∼0.003286	$<10^{-4}$		$0.5M$	0.000143	0.000386	0.008951	0.000388∼0.003956	$<10^{-4}$
	$1M$	0.000002	0.004011	0.011862	0.000214∼0.002057	$<10^{-4}$		$1M$	0.000024	0.000004	0.010125	0.000258∼0.003534	$<10^{-4}$
Dodonew	$0.05M$	0.000032	0.000093	0.006177	0.000500∼0.010240	**0.150**	Chegg	$0.05M$	0.000038	0.000652	0.001338	0.000260∼0.004440	**0.032**
	$0.1M$	0.000040	0.000575	0.005494	0.000310∼0.005640	0.006		$0.1M$	0.000001	0.000292	0.000987	0.000110∼0.003240	$<10^{-4}$
	$0.25M$	0.000015	0.000080	0.004864	0.000360∼0.002796	0.004		$0.25M$	0.000005	0.000034	0.000827	0.000148∼0.003420	0.003
	$0.5M$	0.000044	0.000110	0.005001	0.000248∼0.002970	$<10^{-4}$		$0.5M$	0.000005	0.000042	0.000689	0.000146∼0.002154	0.003
	$1M$	0.000004	0.000011	0.004946	0.000197∼0.002882	$<10^{-4}$		$1M$	0.000004	0.000060	0.000545	0.000102∼0.002059	$<10^{-4}$
000webhost	$0.05M$	0.000027	0.000214	0.004180	0.000200∼0.006060	0.007	Mathway	$0.05M$	0.000030	0.000273	0.003438	0.000240∼0.006980	**0.050**
	$0.1M$	0.000011	0.001858	0.004201	0.000220∼0.005480	0.003		$0.1M$	0.000053	0.000836	0.003122	0.000210∼0.006460	**0.020**
	$0.25M$	0.000001	0.000627	0.004242	0.000232∼0.002364	$<10^{-4}$		$0.25M$	0.000004	0.000017	0.003319	0.000196∼0.002448	0.002
	$0.5M$	0.000015	0.000357	0.004186	0.000126∼0.001766	$<10^{-4}$		$0.5M$	0.000002	0.000001	0.003655	0.000202∼0.002408	$<10^{-4}$
	$1M$	0.000003	0.000042	0.004232	0.000095∼0.002406	$<10^{-4}$		$1M$	0.000001	0.000027	0.004427	0.000219∼0.002110	$<10^{-4}$
Rockyou	$0.05M$	0.000004	0.000133	0.002823	0.000460∼0.013540	$<10^{-4}$	Wishbone	$0.05M$	0.008346	0.000024	0.002338	0.000380∼0.009300	**0.140**
	$0.1M$	0.000016	0.000155	0.002294	0.000510∼0.007280	$<10^{-4}$		$0.1M$	0.000020	0.000129	0.001735	0.000320∼0.006180	0.007
	$0.25M$	0.000029	0.000074	0.001621	0.000588∼0.004480	$<10^{-4}$		$0.25M$	0.000013	0.000077	0.001342	0.000276∼0.004892	$<10^{-4}$
	$0.5M$	0.000082	0.000229	0.001328	0.000386∼0.004058	$<10^{-4}$		$0.5M$	0.000015	0.000112	0.001099	0.000188∼0.003582	$<10^{-4}$
	$1M$	0.000008	0.000041	0.001131	0.000210∼0.003073	$<10^{-4}$		$1M$	0.000011	0.000064	0.000896	0.000159∼0.002873	$<10^{-4}$

† $\Delta C = |C - C'|$ and $\Delta s = |s - s'|$ are differences between C and C', as well as s and s'. D_{KS} is the KS statistic, and $1M = 10^6$, i.e., one million. The **bold** p-values are those >0.01. Both ΔC and Δs are very small ($10^{-7} \sim 10^{-3}$, except two cases in Yahoo), so $C \approx C'$ and $s \approx s'$. Besides, only if the subset size is $\leq 0.1M$ can the p-value be >0.01 to support the CDF-Zipf distribution model.

i.e., $\epsilon_i \sim N(0, \sigma^2)$不仅由中央限制保证定理，但也推荐的NIST标准的统计方法[1]。第三，初步结果表明，实际情况如何σ i是否小于理论上的最大值 $\sqrt{\mu_i}$，并变得非常小(e。g., $<10^{-5}$)，在前3000个唯一密码之后。因此，我们有理由假设σ i=0为不受欢迎的密码与fi<fb。

我们还讨论了定理1的含义。它显示，一旦数据集大小|DS|很大(e。g., ≥，第1类偏差(i。e., 统计随机性$'_{KS}$将接近于0。因此，KS统计量主要来自于第2型偏差，i。e., 使用CDF-Zipf分布模型，p值始终为< 10^{-4}当J0=10000。*这意味着，对于一个足够大的数据集，无论确切的p值阈值是什么，MCA都将拒绝CDF-Zipf，而第2类统计错误（当密码真正遵循它时，拒绝CDF-Zipf）是不可能的*。此外，由于林|DS|→∞D'_{KS}=为0(因此是$D'_K \ll$*斯德克斯*)适用于每个生成的TheoSet，一个直接的推论是lim$_{J0\to\infty}$p值=0。这表明了在MCA (i。e., 做一个更大规模的MCA实验)并不会改变D SDKS的本质，而且只要数据集足够大，CDF-Zipf仍然会被拒绝。

3)实证结果：我们在8个大规模数据集的子集上进行实验，以观察p值如何随着数据集大小|DS|的变化而变化，并找到使p值<为0的阈值。01.通过大小为0的随机子采样数据集（不进行替换）。05M, 0.1M, 0.25M, 0.5M，和1M (1M =106, i.e., 100万)，我们计算了这些子集的p值和参数，结果显示在表三中。 表III显示：(1)对于所有子集，C和C之间的差值CΔ\Δ，以及s和s之间的s\，大约是$10^{-7} \sim 10^{-3}$（在雅虎公司中只有两个例外），因此，定理1对于子集也是实用的；(2)D的最大值和最小值$'_{KS}$随着子集大小的增大而减小，p值单调减小：当子集大小为≥为0时。当大小为25M时，p值为<0.01；当大小为≥1M时，p值为<10^{-4}和D'_{KS}<*脱氧酮类醇；此外，即使p值阈值为<0.01 (e。g.,0.005)，且不同数据集的收敛速度不同，当数据集大小超过1M时，MCA总是拒绝CDF-Zipf；这些发现与定理1一致；(3)当数据集大小为<为0时*。1M，p值是>值为0。01（天涯除外），与Wang等人一致。的观察[51]值小数据集(大小约为$10^4 \sim 10^5$e.g., 的Myspace 0.04M密码[54])可以通过MCA。所有这些都证实了定理1的正确性，并揭示了0的阈值。超过25M，MCA拒绝CDF-Zipf。

我们现在从统计误差的角度来解释表三。因为p-values<0.01为子集≥0。25M，p值<10^{-4}对于≥1M，在MCA中接受CDF-Zipf的机会很低。因此，对于大规模数据集，≥为0。25M，我们只需要考虑第1类统计误差，p-value<0.01时为16.67%，p值<10时为1.96%$^{-4}$.类似地，当数据集的大小很小时(e。g., ≤0.1M)，我们只考虑第2类统计误差。我们将在Sec中讨论MCA的这些错误。IV-C详细。

C.对真实世界的频率偏差的刺激

上述发现提出了一个很自然的问题：对于一个大规模的密码数据集，MCA会拒绝CDF-Zipf? 为了回答这个问题，我们首先测量一个密码的频率差(e。g., 在真实数据集RealSet和理论数据集TheoSet中；然后，我们提出了两个指标(i。e., 绝对和相对偏差度量)来模拟真实世界的偏差。

1)点-明智偏差：我们定义点级偏差来测量在真实世界和理论数据集中的第i个密码之间的频率差异。

定义1：对于第i个唯一的密码，点向偏差$\hat{pd}_{vi}$在RealSet中的第i个唯一的密码是

$$\hat{pd}_{vi} = (\hat{f}_i - f_i)/f_i, \tag{4}$$

其中$\hat{f}_i$和f_i是RealSet的第i个唯一密码和吗TheoSet. The sign of $\hat{pd}_{v_i}$ is denoted as $\hat{\delta}_i$ and $\hat{\delta}_i \in \{-1, 1\}$.

Here we explain Definition 1 First, if the sign $\hat{\delta}_i$ is positive, then $\hat{f}_i > f_i$, and vice versa Second, the absolute value $|\hat{pd}_{v_i}|$ measures *the degree of point-wise deviation*, namely, the在RealSet和TheoSet中具有相同等级的密码频率之间的相对误差。*我们使用术语点级偏差来强调这种偏差是定义在单个密码上定义的，它可以点对点扩展到一个密码间隔*。因此，通过分别处理符号和度，我们将秩为i的单个密码的点偏差域扩展到[i中的密码区间1，i2]如下。

定义2：如果所有的密码都在排名范围内1，我也有相同的符号。区间上的点向偏差1，i2]是

$$\hat{pd}_{v\ [i_1,i_2]} = \hat{\delta}_{[i_1,i_2]} \min_{i \in [i_1\ i_2]} |\hat{pd}_{vi}|, \tag{5}$$

表iv

TOP - 10个唯一密码的点级偏差†

Rank	Yahoo		Dodonew		000webhost		Rockyou		Tianya		Chegg		Mathway		Wishbone	
	Password	pdv_i	Password	pdv_i	Password	pdv_i	Password	pdv_i	Password	pdv_i	Password	pdv_i	Password	pdv_i	Password	pdv_i
1	123456	-67.16%	123456	-25.52%	abc123	-71.15%	123456	-76.16%	123456	-35.96%	default‡	-68.47%	123456	-74.30%	123456	-62.94%
2	password	-51.37%	a123456	24.58%	123456a	-18.30%	12345	-53.21%	111111	41.25%	123456	3.91%	mathway	-14.07%	password	-12.81%
3	welcome	-69.23%	111111	57.33%	12qw23we	-5.15%	123456789	-29.90%	000000	36.77%	Chegg123	12.78%	123456789	24.10%	123456789	-7.55%
4	ninja	-62.27%	123456789	48.03%	123abc	13.58%	password	-28.28%	123456789	76.50%	password	25.56%	12345	58.26%	wishbone	-29.07%
5	abc123	-54.42%	a321654	50.71%	a123456	23.50%	iloveyou	-25.91%	123123	30.24%	Testing1	46.77%	password	64.92%	12345678910	-21.35%
6	123456789	-46.69%	123123	61.63%	123qwe	39.83%	princess	-41.83%	123321	-12.33%	testing	54.83%	abc123	7.68%	unicorn	-17.06%
7	12345678	-43.09%	5201314	77.92%	secret666	47.13%	1234567	-56.55%	5201314	-4.41%	Chegg1	29.79%	12345678	3.13%	1234567890	-23.65%
8	sunshine	-42.99%	123456a	84.73%	YfDbUfNjH10305070‡	60.20%	rockyou	-52.95%	12345678	4.20%	A1b2c3	13.92%	1234567890	4.90%	12345678	-16.19%
9	princess	-38.09%	0	47.92%	asd123	69.87%	12345678	-48.71%	666666	6.97%	default‡	24.89%	mathsucks	-0.67%	qwertyuiop	-13.68%
10	qwerty	-45.77%	000000	46.77%	qwerty123	82.33%	abc123	-54.55%	111222tianya	9.86%	010203Zaq	32.05%	qwerty	7.71%	1234567	-10.45%
%*	1.89%		3.28%		0.79%		2.05%		7.43%		0.98%		1.20%		1.64%	

† pdv_i is the point-wise deviation of the i-th unique password of the real-world dataset RealSet, and can be extended on interval $[i_1, i_2]$ through Definition 2.

‡ The 8th unique password YfDbUfNjH10305070 of 000webhost is a default value [54]. The top-1 and top-9 defaults in Chegg correspond to MD5 hashes (without salt) that cannot be recovered. Why these passwords are popular may due to system settings (e.g., system-generated passwords) that are accepted by users. The exact reasons are beyond our comprehension and are *unlikely to alter our analysis in any significant way*.

■ % means the proportion of top-10 unique passwords. This shows that Chinese passwords are more concentrated than their English counterparts as concluded in [■■].

在哪里$\hat{\delta}_{i1,i2}$使满意

$$\hat{\delta}_{[i_1,i_2]} = \begin{cases} 1 & \text{If } \hat{pdv}_i \geq 0 \text{ for all } i \in [i_1, i_2] \\ -1 & \text{If } \hat{pdv}_i < 0 \text{ for all } i \in [i_1, i_2]. \end{cases} \quad (6)$$

我们的8个数据集的前10个唯一密码及其点向偏差如表4所示。对于每个数据集，密码汇总在第一列中，它们的点向偏差显示在第二列中。

我们演示了点向偏差的含义
结果见表四。首先，第一个符号的符号(i。e.，top1)
密码是负的，我。$\hat{\delta}$e.,1<0和$\hat{f}_1 < f_1$，表示
RealSet中前1个唯一密码的频率较小
比CDF-Zipf分布给出的TheoSet相比
模型一个可能的原因是密码123456是
在我们的大多数数据集中，最流行的密码(除了
000个网络主机和Chegg由于他们的密码策略[52])，
并被广泛认为较弱，用户自觉回避
选择它。其次，密码与相同的标志往往是
连接的，所以点向偏差的定义
间隔(i。e.，定义2)是实用的。因为两者$\hat{f}_i$和f_i
随着i的增加而逐渐减小，可以看作是一个连续的
函数以i作为变量。基于标志保存

属性，如果pdvi0>0(报告。<0)对于一些我0，存在一个
附近的0，i0+ Len]，pdvi>0(回复。<0)为所有
$i \in [i0, i0 + Len]$。此外，由于CDF-Zipf是准确的，因此
RealSet和TheoSet的斜坡足够近，所以Len是
通常大(e。g.,$\hat{\delta}$[2，136]=1和Len=135)。

图1掌握RealSet、let和TheoSet中前100个唯一密码的频率。在[2,100]中排名的密码符号可以是负的，也可以是正的，这与用户的语言和服务类型有关。例如，在两个中文数据集（Dodonew和天涯）中，第2到第10个唯一密码的标志是都是积极的，我。e.,$\hat{f}$i>fi，这可能是因为那个中国人

密码比英文密码[53]更集中（见表四）。

基于上述观察结果，我们通过调整以下两个指标中的点向偏差（记为pdvs）来生成模拟数据集SmuSet，以模拟真实世界的密码偏差。

*2)绝对偏差度量：我们假设是第一个NO*唯一密码偏离，i
。e.，偏差范围为[1，NO]。首先，我们来确定这个符号。对于RealSet，由于具有相同符号的密码通常是连接的，因此存在每个唯一密码具有相同符号的时间间隔

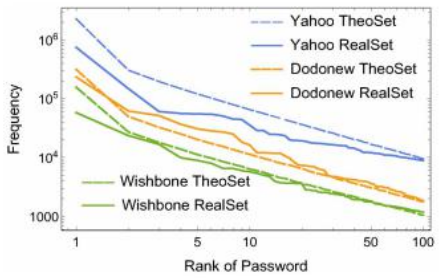

图1.掌握在雅虎、Dodonew和希望骨数据集的前100个唯一密码的频率。可以观察到，在雅虎和愿望骨中低于神集，而RealSet在中高于神集
请排在前两个密码之后。因此，根据定义1，
$\hat{d}$pvi<0为i∈[1,100]在雅虎和希望骨，和pvi>0为i∈$\hat{d}$
[2,100]在Dodonew。这表明具有相同符号的密码经常被连接，因此定义2是定义良好的。

符号(e。g.，[2,10]在黎明时分)。因此，我们将[1，NO]划分为区间i的不相交并集1，i2：

$$[1, N_0] = \bigcup [i_1, i_2] \quad \hat{\delta}_i = \hat{\delta}_{i_1} \text{ for all } i \in [i_1, i_2]$$

$$[i_1, i_2] \bigcap [i_1', i_2'] = \phi \quad \text{for any two different intervals。} \quad (7)$$

在每个间隔[i1，我们设置了点向偏差pdv^{S}模拟数据集
SmuSet基于两个规则：

1)模拟的点向偏差的符号与
the real-world one, i.e., $\delta^S_{[i1,i2]} = \hat{\delta}_{[i1i2]}$

2) The absolute value of the simulated point-wise deviation e.,0 < k$\hat{A}$=
$|pdv^s_{[i_1 i_2]}| \leq |\hat{pdv}_{[i_1 i_2]}|$ on $[i_1, i_2]$.
Hence, the i-th dev iat ed p assword in SmuSet has

$$是= fi\ (1 + pdv\ i^S_{[1,i2]}) = fi\ (1 + \hat{\delta}[i1,i2]k\hat{A}). \quad (8$$

就像我和现实世界里的那个一样。

如表四所示，点向偏差pv$\hat{d}_1$…的
第一个密码通常是不同的符号和一个大得多的
级别因此，我们分别处理第一个密码。

基于这些规则，我们计算了绝对偏差度量中的p值。类似于Alg。1，首先，我们生成JO=10,000个理论集，在绝对偏差度量中偏离它们，得到突变集。其次，我们用Wang等人提出的GSS拟合方法对每个突变集进行拟合。[51]，然后得到过程，s^S，D^S_{KS})作为KS的统计量和表征参数。第三，我们计算每个SmuSet的p值，并取一个平均值（记为p值）来减轻统计波动。该过程见Alg。2.

算法2计算绝对偏差度量（ABSDEV）中的p值

输入：参数C和s，真实数据集RealSet，
点偏差度k_A和范围[1, NO]。
输出：p值^{S}模拟数据集的模拟集合。

1 **开始**
2 **对于j = 1到**JO**做**
3 =（C，s，|DS|）；/*生成JO理论数据集。*/
4 （C，s，D'''_{jjKSj}）= GSS（TheoSetj）；/*用相同的方法拟合真实世界的数据集，从而拟合JO理论数据集。*/
5 **对于j = 1到**JO**做**
6 除以[1, NO] =n[i_1, i2]与我1, i2] U[1'', i] = ϕ和δ [i_1,i2]对每一个都是固定的吗1将偏差范围划分为区间的不相交联合。*/
7 **对于i = 1到**NO**做**
8 f_1^s= f_1(1 + δ 1k_1偏离第一个密码。*/
9 $f_i^s = f_i$(1+ δ^[i_1 , i_2]k$_A$);/ *D eviatp asswords i_1 to i_2 ineac h t he或 'eti ca l达taset*/
排名来自
10 SmuSetj = {f'',''f .''12.}; (C_j^s, s_j^s, D_{KSj}^s)= GSS
11 适合JO采用GSS拟合方法模拟数据集。*/
p值$_j^s$
12 = (#{$D'_{KS\ j}$|$D'_{KS\ j}$ > $D_{KS\ j}^s$, 1 ≤ j ≤ JO} + 1)/(JO
计算每个模拟数据集的p值，并平滑计算p值^S. */
p值^S=（p值$_1^s$+p值$_2^s$..+p值$_{JO}^s$)/JO；/*采取
13 平均值作为p值^S. */
输出：p值^{S}模拟数据集。
14

3)相对偏差度量：我们也探索了相对偏差
偏差度量，其中点向偏差pdv$_i^s$…的
SmuSet中的第i个唯一密码与中的密码成正比

RealSet, i.e., pd_{vis} = p d^v$_i k_R$ with the d
$0 < k_R \leq 1$. In th是c作为e t he frequency f_i^s

eviation parameter f_I^s= fi (1 + p$\hat{d}$vi • k$_R$) = fi (1 −
in SmuSet has
k$_R$) +f^ik_R. (9)

Since each deviated password has δ_i^s = δ^$_i$
to divide [1, N_0] into disjoint unions of int e
, there is no need
rvals. The process
类似于Alg。除了第8行和第9行被替换为
平衡所以我们在这里省略了它的演示。

4)偏差度量理论：我们现在证明了D的最大值D_{ks}^s（拟合SmuSet的KS统计量）随着偏差度k的增加而增加$_A$和k$_R$. 因此，我们只需要尝试有限数量的k$_A$和k$_R$值来找到使MCA拒绝CDF-Zipf的阈值。

在输入证明之前，我们先陈述一下我们的假设。首先，类似于Sec。III-B，我们假设*过程*≈ C\和s^S≈ s\SmuSet。其次，密码的秩在偏差后没有显著变化，即TheoSet中的第i个密码在SmuSet中也排名约为i。假设
两个密码*PWi*和*PWJ*在TheoSet中有fi>fj；(1)如果两者都有*PWi*和*PWJ*有偏差，有fis >f$_j^s$在SmuSet中
根据方程式。8和9，所以它们在偏差范围内的排名没有变化；(2)如果*PWi*是偏离，但*PWJ*不是，偏差程度可以调整以最小化等级的变化。
第三，TheoSet之间的直接最大CDF偏差
用其导出的SmuSet作为最大最大D_{KS}^s,

i.e., max D_{KS}^s= |CDF (SmuSet) − CDF (TheoSet)|.这是
理由是D_{ks}^s由GSS产生的结果不能超过
直接最大CDF偏差。在这些假设的基础上，我们将关于密码偏差的两个定理表述如下。

定理2：在绝对偏差度量中，如果是密码
在SmuSet中，偏离为pdv$_I^s$=$\hat{\delta}$[i1,i2]• k$_A$对于我∈[我1， i2]，
和偏差度为k$_A$满足0≤k$_A$≤ |p$\hat{d}$v [i_1, i2]|，
然后是最大最大D_{KS}^s增加为k$_A$增加量

定理3：在相对偏差度量中，如果是密码
are deviated as $pdv_i^s = \delta\hat{}_i \cdot |pd\hat{}v_{i/kR}$ for $i \in [1, N_0]$ and $0 < k_R \leq 1$, then the ma ximum $\max D_{KS}^s$ increases as k_R 增加（见附录B中定理2和3的证明）。

*5)讨论：我们现在做了一些理由。*首先，我们只探讨只有一个区间[i]的情况1，i2]涉及到绝对偏差的情况。对于每个区间上符号固定的多个区间，我们可以归纳地应用定理2得到全局解。其次，我们澄清为什么不考虑KS统计量D之间的关系$_{ks}^{s}$偏差范围*NO*.在绝对偏差中
情况$\hat{\delta}_i$在多个问题上是复杂的；在相对偏差中
的情况下，也不存在简单的单调性。因此，我们主要
关注如何max D_{kS}^{s}k的变化*A*和k*R*.

定理2和理3表明，在这两个偏差中都满足了-
rics, max D_{KS}^{s}随着偏差度k的增加而增加*A*和k*R*增加，所以p值
^{s}计算1型偏差大于SmuSet的KS统计量(i。e., $D'_{KS}{}_{ks}^{s}$>D)应该减少为k*A*和k*R*增加因此，对于一个给定的密码间隔，它的排名在[i]中1，i2] (resp。)，如果是一个特定的k*A*(报告。*kR*)的值可以使p值<为0。01，所以一个更大的值也可以实现这个目标。有了这个单调性保证，我们只需要尝试一个有限数量的k*A*和k*R*请使用相应的值来查找阈值。

D.数值模拟实验研究

在本节中，我们将描述偏差阈值调查的实验设置和结果，包括绝对偏差和拒绝MCA的相对偏差指标。

*1)实验设置：我们首先考虑绝对偏差度量。*我们设置了偏差范围参数*NO*∈{1,10,100}，将[1，NO]划分为不相交的区间并集，并分别处理第一个唯一的密码，如Alg所示。2.我们还设置了|pdv$_1$,i*2]/>/pdv[i*$_3$,i4]|为其他连续的间隔来平滑偏差。不需要考虑的原因
其他密码(e。g., 第一个100个～10,000个唯一密码)
有两方面： (1)这些密码通常比前100个密码的频率要小得多，所以它们的偏差影响不那么显著；(2)这些密码的点向偏差往往有很大的不同，这使模拟变得复杂。例如，在雅虎

$pd\hat{}v_{115} = 0.68\%$ but $pd\hat{}v_{1\ 000}$
willu nnecessarily comp李cate t
$= 35\%$, so consideringthem
he simulation.

其次，我们考虑了偏差的程度。我们调查
真实世界的偏差度p$\tilde{d}$v [i$_1$, i2]在每个不相交处
间隔时间1，i2]，并将其设置为模拟值的最大值
偏差度，i。e., k*A*= /*pdv i*$_{[1,i2]}^{s}$ |= 1%， 5%， 10%，25%, ...，|p$\tilde{d}$v [i$_1$,i2]|，(其中|p$\tilde{d}$v [i$_1$,i2]|=$_i$
2]$\min_{i\in[i_1}\tilde{d}$|pvi|，见
定义2）。我们还考虑了两个层次的p值阈值，如在第二秒中。III-B：(1)p值<0。01（标准阈值）和(2)p值^{s}<10^{-4},
i.e., D_{KS}^{s}>max D'_{KS}，其中所有生成的模拟数据集的第2型偏差>0。

在相对偏差度量中，除了绝对度量中的偏差范围外，我们还设置了*NO*= N， i.e., 所有的密码都有偏差。如第二节所述。III-C，我们直接基于等式偏离密码9，并设置偏差参数k*R*=值分别为1%、5%、10%、25%、50%和100%。

2)实验结果：表V为KS统计数据和p值^{S}雅虎、Dodonew和许什骨的组合（其他类似的）。左边的列记录绝对偏差的结果，右边的列记录相对偏差的结果。

表V显示了：(1)在这两个偏差指标中，
KS统计量D_{KS}^{s}随偏差而单调增加

表v

模拟数据集的KS统计量和p值†

Dataset	Absolute point-wise deviation‡									Relative point-wise deviation								
	Point-wise deviation	D^s_{KS}	p-value^s	Point-wise deviation	D^s_{KS}	p-value^s	Point-wise deviation	D^s_{KS}	p-value^s	Point-wise deviation	D^s_{KS}	p-value^s	D^s_{KS}	p-value^s	D^s_{KS}	p-value^s	D^s_{KS}	p-value^s
Yahoo	Dev-range $N_0=1$			Dev-range $N_0=10$ $[1,10]=\{1\}\bigcup[2,10]$			Dev-range $N_0=100$ $[1,100]=\{1\}\bigcup[2,10]\bigcup[11,100]$			Dev-range $N_0=1$			Dev-range $N_0=10$		Dev-range $N_0=100$		All unique PWs	
	-1%	0.000928	$<10^{-4}$	-1%	0.000842	$<10^{-4}$	-1%	0.001066	$<10^{-4}$	1%	0.000986	$<10^{-4}$	0.000937	$<10^{-4}$	0.000898	$<10^{-4}$	0.017486	$<10^{-4}$
	-5%	0.001409	$<10^{-4}$	-5%,-1%	0.001206	$<10^{-4}$	-5%,-1%	0.000952	$<10^{-4}$	5%	0.001076	$<10^{-4}$	0.001355	$<10^{-4}$	0.000939	$<10^{-4}$	0.017633	$<10^{-4}$
				-5%	0.001002	$<10^{-4}$	-5%	0.001317	$<10^{-4}$	10%	0.001520	$<10^{-4}$	0.001473	$<10^{-4}$	0.001213	$<10^{-4}$	0.018583	$<10^{-4}$
	-10%	0.001542	$<10^{-4}$	-11%	0.001900	$<10^{-4}$	-10%,-1%	0.002226	$<10^{-4}$	25%	0.002595	$<10^{-4}$	0.002756	$<10^{-4}$	0.002749	$<10^{-4}$	0.021969	$<10^{-4}$
	-25%	0.004221	$<10^{-4}$	-20%	0.002415	$<10^{-4}$	-10%,-5%	0.001968	$<10^{-4}$	50%	0.005425	$<10^{-4}$	0.005384	$<10^{-4}$	0.006131	$<10^{-4}$	0.028538	$<10^{-4}$
				-25%	0.004221	$<10^{-4}$	-25%,-1%	0.004710	$<10^{-4}$	100%	0.010961	$<10^{-4}$	0.012264	$<10^{-4}$	0.012122	$<10^{-4}$	0.043985	$<10^{-4}$
Dodo-new	Dev-range $N_0=1$			Dev-range $N_0=10$ $[1,10]=\{1\}\bigcup[2,10]$			Dev-range $N_0=100$ $[1,100]=\{1\}\bigcup[2,10]\bigcup[11,100]$			Dev-range $N_0=1$			Dev-range $N_0=10$		Dev-range $N_0=100$		All unique PWs	
	-1%	0.001203	**0.185**	-1%,1%	0.001352	**0.029**	-1%,1%	0.000851	**0.361**	1%	0.000218	0.003	0.000980	**0.265**	0.000837	**0.381**	0.002467	$<10^{-4}$
	-5%	0.000990	**0.240**	-5%,5%	0.001221	**0.055**	-5%,1%	0.000770	**0.408**	5%	0.000004	**0.376**	0.000980	**0.256**	0.000764	**0.423**	0.002586	$<10^{-4}$
	-10%	0.001270	**0.079**	-10%,1%	0.001187	**0.074**	-10%,1%	0.001323	**0.029**	10%	0.000005	**0.015**	0.001426	**0.032**	0.001054	**0.169**	0.002628	$<10^{-4}$
				-10%,5%	0.001775	**0.012**	-10%,5%,1%	0.001625	0.008	25%	0.000182	**0.026**	0.001230	**0.040**	0.001283	**0.050**	0.002513	$<10^{-4}$
	-25%	0.002953	$<10^{-4}$	-25%,1%	0.002915	$<10^{-4}$	-25%,1%	0.002849	$<10^{-4}$	50%	0.000394	0.005	0.002232	0.004	0.002391	$<10^{-4}$	0.002693	$<10^{-4}$
				-25%,5%	0.002962	$<10^{-4}$	-25%,5%,1%	0.002918	$<10^{-4}$	100%	0.000400	$<10^{-4}$	0.003855	$<10^{-4}$	0.004687	$<10^{-4}$	0.004941	$<10^{-4}$
Wish-bone	Dev-range $N_0=1$			Dev-range $N_0=10$ $[1,10]=\{1\}\bigcup[2,10]$			Dev-range $N_0=100$ $[1,100]=\{1\}\bigcup[2,62]\bigcup[63,100]$			Dev-range $N_0=1$			Dev-range $N_0=10$		Dev-range $N_0=100$		All unique PWs	
	-1%	0.000823	**0.689**	-1%	0.000800	**0.710**	-1%,1%	0.001813	**0.010**	1%	0.000806	**0.709**	0.000818	**0.698**	0.000831	**0.681**	0.004274	$<10^{-4}$
	-5%	0.001078	**0.357**	-5%,-1%	0.001049	**0.355**	-5%,-1%,1%	0.001189	**0.172**	5%	0.001094	**0.290**	0.001125	**0.247**	0.001173	**0.189**	0.004558	$<10^{-4}$
				-5%	0.001219	**0.143**	-5%,1%	0.001155	**0.210**	10%	0.000979	**0.466**	0.001102	**0.278**	0.001158	**0.206**	0.004874	$<10^{-4}$
	-10%	0.001991	0.004	-10%,-1%	0.001916	0.007	-10%,-1%,1%	0.001763	**0.011**	25%	0.002060	0.005	0.001974	0.005	0.001862	0.008	0.006160	$<10^{-4}$
	-25%	0.002952	$<10^{-4}$	-25%,-1%	0.002709	$<10^{-4}$	-25%,-1%,1%	0.002816	$<10^{-4}$	50%	0.003594	$<10^{-4}$	0.003337	$<10^{-4}$	0.003425	$<10^{-4}$	0.008999	$<10^{-4}$
	-50%	0.005170	$<10^{-4}$	-50%,-1%	0.005182	$<10^{-4}$	-50%,-1%,1%	0.005180	$<10^{-4}$	100%	0.006563	$<10^{-4}$	0.006391	$<10^{-4}$	0.006394	$<10^{-4}$	0.014763	$<10^{-4}$

D_{KS} is the KS statistic of fitting the simulated dataset SmuSet, and the ■-value^s is the corresponding ■-value. The **bold** ■-value^s results are the ones ≥ 0.01 threshold, meaning that CDF-Zipf can pass Monte Carlo approach (MCA), so passwords are supposed to follow it. We can see that with small (i.e., 1■25%) deviations in both absolute and relative deviation metrics, the ■-value <0.01 and even ■10^{-4}, so MCA is not an effective method in telling whether a large-scale password dataset follows CDF-Zipf.

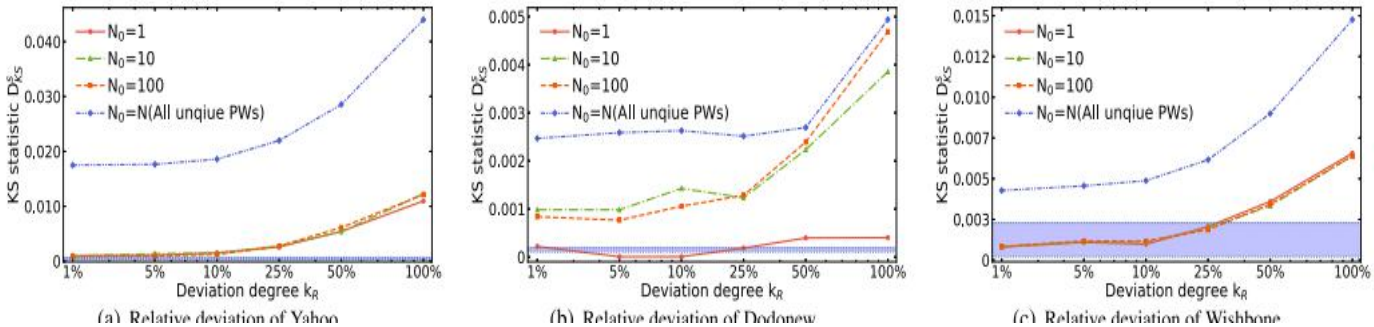

图2．雅虎、Dodonew和许骨数据集的相对偏差。每条曲线代表一个不同的偏差范围。阴影区域对应于KS统计量D的最大值和最小值r_{KS}拟合TheoSet（用于测量统计随机性）。只有在阴影区域内，蒙特卡洛

方法（MCA）p值^{s}可能是>0。01．然而，所有的D^s_{KS}与k相比，突变体超过了阴影区域$R\geq$ 50%．

pd0度^{s}只有少数例外(e。g.， pd0$^s_1$= −5%
与定理2和3一致；

(2) In both casees, pdv^s_t only ned to be 1%∼25% to make p-value^s<0.01, and 1%∼50% to make p-valu e$^s<10^{-4}$; Particularly, when *all* p ass swords aredeviated (i e, $N_0=N$ i n the

相对情况下)，偏差小到1%就足以使p值$^s<10^{-4}$.这意味着即使p值阈值是< 0 (e。.01g.,0.005)，MCA将总是拒绝对于大规模数据集的具有真实世界偏差的CDF-Zipf。*所有这些都证实了Wang等人的观点*。对CDF-Zipf样本量影响的[51]猜想，如此小和不显著的偏差确实会导致大规模数据集的统计显著性。我们还展示了如何计算KS统计量D^s_{KS}随相对偏差度k的变化R在无花果。2.

3)总结：我们使用理论和实验来研究MCA拒绝CDF-Zipf的程度。我们首次证明了在CDF-Zipf分布模型和GSS拟合方法下，第1类偏差(i。e.， 统计随机性)随着数据集大小的增加而渐近地收敛于0。因此，一个大数据集的密码数据集(e。g.， ≥100万)主要来自第2型偏差（使用CDF-Zipf分布模型），p值为$<10^{-4}$.对8个大规模数据集的实验结果表明，数据集的大小只需要为≥0。25M，使p值为<0。01拒绝CDF-Zipf，≥1M使p值为$<10^{-4}$.接下来，我们提出绝对和相对偏差指标来模拟真实世界的偏差，并揭示

这两个指标中1%的随机偏差就足够了
MCA拒绝CDF-Zipf。这些严格的数学证明和广泛的实验证实了王等人。关于样本量对CDF-Zipf影响的[51]猜想。

增值新型号的密码分发

由于MCA总是拒绝大规模数据集的CDF-Zipf，一个问题自然出现了：是否有更好的分布模型可以提供更高的精度并通过MCA? 我们比较了CDF-Zipf与其他可能的分布的拟合精度，并使用MCA来计算它们的p值来回答这个问题。

A. 考虑替代分布的必要性

我们展示了探索替代分布的必要性。如果密码遵循CDF-Zipf (i。e.，*波多黎各*= Crs)，CDF曲线在对数-对数尺度下应是一条直线。然而，由于在Sec中显示的点级偏差。(特别是表IV和图。1)，一条扭曲但不是直线在实践中更现实。因此，在对数对数尺度下，可能有更精确的大致线性曲线的分布模型。然而，据我们所知，之前没有研究探索过这些替代分布模型。

我们给出了图中五种分布的一个例子。3．所有的cdf（详细信息见表VI）在对数-对数尺度下大致呈线性关系，但实际上只有蓝色的曲线是从CDF-Zipf中绘制出来的。此外，当使用线性回归来拟合这些数据时

表VI

CDF -ZIPF的替代分发模型†

Distribution‡	Distribution model in rank-frequency (RF) coordinate system		Distribution model in frequency-frequency (FF) coordinate system, but converted to RF system	
	PDF kernel	CDF	PDF kernel	CDF
CDF-Zipf [51]	r^{s-1}	Cr^s	r^{s-1}	Cr^s
Exponential [20]	$\exp(-\lambda r)$	$1-\exp(-\lambda r)$	$\ln\frac{r}{\lambda\lvert DS\rvert}$	$-\frac{r}{\lambda\lvert DS\rvert}\ln\frac{r}{\lambda\lvert DS\rvert}$
Lognormal [20]	$\frac{1}{r}\exp(-\frac{(\ln r-\mu)^2}{2\sigma^2})$	$\Phi(\frac{\ln r-\mu}{\sqrt{2}\sigma})$	$\exp(\sigma\Phi^{-1}(\frac{\exp(\mu+\sigma^2/2)}{\lvert DS\rvert}r+\frac{1}{2}))$	$\Phi\left(2(\sigma-\Phi^{-1}(1-\frac{\exp(\mu+\sigma^2/2)}{\lvert DS\rvert}r))\right)-\frac{1}{2}$
Zipf-cutoff [20]	$r^{-\alpha}\exp(-\lambda r)$	$Q(1-\alpha,\lambda r)$	$Q^{-1}(1-\alpha,\frac{(1-\alpha)\cdot r}{\lambda\lvert DS\rvert})$	$Q\left(2-\alpha,Q^{-1}(1-\alpha,\frac{(1-\alpha)r}{\lambda\lvert DS\rvert})\right)$
Stretched-exponential [20]	$r^{\alpha-1}\exp(-\lambda r^{\blacksquare})$	$1-\exp(-\lambda r^{\blacksquare})$	$\left(\ln\frac{\Gamma(1+1/\alpha)r}{\lambda\lvert DS\rvert}\right)^{\frac{1}{\alpha}}$	$Q\left(1+\frac{1}{\alpha},-\ln\frac{\Gamma(1+\frac{1}{\alpha})r}{\lambda\lvert DS\rvert}\right)$

† The kernel is the term in the PDF expression which only includes variables but no coefficients. All distributions are considered in both rank-frequency (RF) and frequency-frequency (FF) coordinate systems. For distributions in the FF systems, they have been converted to the RF system as shown in the right half of the table.
‡ Since the power-law and Zipf are equivalent, their PDF kernel has the same form in two coordinate systems. For lognormal, $\Phi(\blacksquare)$ is the CDF of the standard normal distribution $N(0,1)$, and Φ^{-1} is the inverse function of Φ. For Zipf-cutoff, Q is the regularized gamma function defined as $Q(s,x)=\frac{\Gamma(s,x)}{\Gamma(s)}$, where $\Gamma(s,x)=\int_x^\infty t^{s-1}\exp(-t)dt$ and $\Gamma(\blacksquare)=\int_{\blacksquare}^\infty t^{s-1}\exp(-t)dt$ [redacted] $s,x>\blacksquare$ to make $Q(s,x)$ to be a real number, so $\alpha<1$.

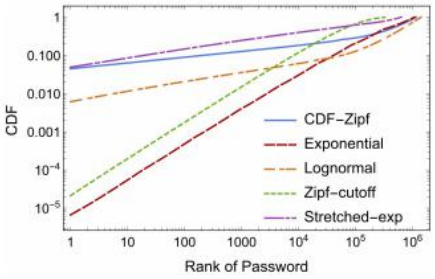

图3. 一个另类分布模型的展览。除CDF-Zipf外，其他四种分布模型（拉伸-exp代表拉伸指数，详见表六）在对数-对数尺度下也大致呈线性，因此密码也可能遵循其中一种。

CDFs，我们发现确定系数为0。850<R2<0.999(其中 *R2*∈[0,1]和越大的是*R2*，数据越接近一条线)。在这四种（可能还有其他）替代模型中，有可能有些人(s)比CDF-Zipf更准确。我们将在下面的内容中确认这一点。

1)两个坐标系：我们提出了两种不同的坐标系统用于分析密码频率，并研究了这两个系统之间的关系。

首先，我们讨论了在Sec中使用的秩-频率坐标系（简称为射频系统）。罗马数字 3在这个系统中，x轴记录了一个密码的秩r，和y轴
记录频率f_r.虽然在实践中，y坐标
isconvedrte 向 rec或d th)e CDF 波多黎各= $\sum_{i=1}^{r} fi/|$ |D𝒮|是 thedataset司ze [51], Cdf 是equiva勒nt D S| (w她 e
向frequency
本质上，当CDF在y轴上时，秩-频率的本质仍然保持不变。

除了秩频率之外，另一种查看数据的方法是计算数字(i。e.，唯一密码的频率)*nk*每个用户都恰好由k个用户使用。在该系统中，x轴记录了密码的频率fr，这与射频系统的y轴相同。因此，y轴记录了频率nf$_I$*唯一的密码发生的*f_I*乘以*我们将这个坐标系表示为频率系（简称为FF系），x作为变量。在实践中，FF系统被广泛应用于物理[20]、复杂网络[5]、市场营销[24]等研究领域，因此我们的考虑是合理的。

我们现在展示了射频系统和FF系统之间的关系。如果密码PW的等级为r和频率为f_I在一个数据集中，它在RF系统中的坐标是（r，fr）。另外，如果有n个f_I唯一的密码，频率为f_I在数据集中，FF系统中PW的坐标为(fr，*n fr*)。

因此，FF系统中的互补CDF具有

$$P(X \geq x) = r/N, \tag{10}$$

其中，N为唯一密码的个数。基于这一性质，我们可以使用与亚当的工作[3]相同的技术，将FF中的任何分布转换为射频系统。

B. 替代分布模型

2009年，克劳塞特等人。[20]比较了FF系统中四种具有幂律（相当于CDF-Zipf [3]）的替代分布模型。受此想法的启发，我们考虑了射频和FF系统中的这四种模型，总共有8种备选模型。射频系统中的PDF和CDF表达式为
详见表六的前两列。

我们证明了在表六中研究两个坐标系中的模型。首先，无论坐标系如何，这些分布在对数-对数尺度下大致呈线性分布，如图所示。3，所以有必要在这两个系统中都考虑它们。其次，这些分布模型也很有意义。例如，对数正态分布通常表征变量的随机乘法过程[33]。另一个例子是拉伸指数，它的生成类似于CDF-Zipf，但有更严格的约束[28]。第三，尽管可能有其他的分发模型可以遵循密码，但它们要么不太常见，要么与我们考虑的模型(e。g.，逆伽马分布）。因此，我们对替代分布模型的考虑要尽可能全面。

我们使用等式10作为将FF系统中的分布转换为射频系统的关键（更多细节见附录C）。转换后的PDF内核和CDF表达式显示在表六的最后两列中。这是理由
转换在于，RF系统中的GSS拟合方法对CDF-Zipf [51]表现最好： GSS不仅导致最小的KS统计量，而且可以覆盖整个数据集。因此，GSS也可以优化射频系统中的其他模型。因此，即使在RF系统中转换的pdf和cdf是复杂的，并且不常用，我们仍然适合使用密码，因为它们的等价物在FF系统中是常见的。

我们根据它们来命名替代的分布模型
起源：对于射频系统中原始的四个模型，我们将它们命名为指数模型射频，lognormal射频，Zipf截止射频和拉伸指数射频.*因此，我们将FF系统中的FF替换为RF，然后转换为RF系统*。拟合参数

表VII

备选分布模型的GSS拟合结果†

Dataset	Distribution	θ_1	θ_2	$D^{\ddagger}_{KS}$	p-value‡	Dataset	Distribution	θ_1	θ_2	$D^{\ddagger}_{KS}$	p-value‡
Yahoo	CDF-Zipf(C,s)	0.033148	0.180907	**0.040775**	$<10^{-4}$	Tianya	CDF-Zipf(C,s)	0.062022	0.155290	**0.022940**	$<10^{-4}$
	Exponential(λ)$_{RF}$	5.24 ■ 10^{-7}	—	**0.394789**	$<10^{-4}$		Exponential(λ)$_{RF}$	1.11 ■ 10^{-6}	—	**0.300070**	$<10^{-4}$
	Exponential(λ)$_{FF}$	0.288010	—	**0.277861**	$<10^{-4}$		Exponential(λ)$_{FF}$	0.144383	—	**0.292062**	$<10^{-4}$
	Lognormal(μ,σ)$_{RF}$	15.01919	6.359214	**0.018063**	$<10^{-4}$		Lognormal(μ,σ)$_{RF}$	13.29193	6.780193	**0.022918**	$<10^{-4}$
	Lognormal(μ,σ)$_{FF}$	-25.90537	5.160470	**0.014101**	$<10^{-4}$		Lognormal(μ,σ)■	-32.91796	5.856597	**0.019428**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{RF}$	0.000027	-6159.767	**0.300296**	$<10^{-4}$		Zipf-cutoff(λ,α)$_{RF}$	0.000123	-5802.884	**0.313710**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{FF}$	0.106206	0.996573	**0.154641**	$<10^{-4}$		Zipf-cutoff(λ,α)■	0.054899	0.983886	**0.263220**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{RF}$	0.020419	0.238575	**0.023873**	$<10^{-4}$		Stretched-exponential(λ,α)$_{RF}$	0.041705	0.214004	**0.012353**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{FF}$	19.34152	0.050895	**0.035185**	$<10^{-4}$		Stretched-exponential(λ,α)$_{FF}$	36.71749	0.026701	**0.021080**	$<10^{-4}$
Dodonew	CDF-Zipf(C,s)	0.019255	0.211921	**0.004979**	$<10^{-4}$	Mathway	CDF-Zipf(C,s)	0.010541	0.245255	**0.011059**	$<10^{-4}$
	Exponential(λ)$_{RF}$	3.07×10^{-7}	—	**0.178653**	$<10^{-4}$		Exponential(λ)$_{RF}$	2.23×10^{-7}	—	**0.145354**	$<10^{-4}$
	Exponential(λ)$_{FF}$	0.856726	—	**0.170507**	$<10^{-4}$		Exponential(λ)$_{FF}$	1.152190	—	**0.147374**	$<10^{-4}$
	Lognormal(μ,σ)$_{RF}$	15.98301	6.364878	**0.019549**	$<10^{-4}$		Lognormal(μ,σ)$_{RF}$	16.22701	5.446235	**0.007611**	$<10^{-4}$
	Lognormal(μ,σ)$_{FF}$	-26.81404	5.035345	**0.021462**	$<10^{-4}$		Lognormal(μ,σ)■	-20.13930	4.289310	**0.005521**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{RF}$	0.000036	-6003.371	**0.184108**	$<10^{-4}$		Zipf-cutoff(λ,α)$_{RF}$	0.000094	-6062.899	**0.320328**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{FF}$	0.349199	0.993975	**0.154641**	$<10^{-4}$		Zipf-cutoff(λ,α)■	0.199062	0.997929	**0.152297**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{RF}$	0.013778	0.254693	**0.013313**	$<10^{-4}$		Stretched-exponential(λ,α)$_{RF}$	0.006868	0.292796	**0.004638**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{FF}$	20.66430	0.052236	**0.036053**	$<10^{-4}$		Stretched-exponential(λ,α)$_{FF}$	14.24821	0.072425	**0.016458**	$<10^{-4}$
000webhost	CDF-Zipf(C,s)	0.005738	0.282561	**0.005084**	$<10^{-4}$	Chegg	CDF-Zipf(C,s)	0.008178	0.235996	**0.008171**	$<10^{-4}$
	Exponential(λ)$_{RF}$	4.98×10^{-8}	—	**0.181766**	$<10^{-4}$		Exponential(λ)$_{RF}$	3.60×10^{-8}	—	**0.139127**	$<10^{-4}$
	Exponential(λ)$_{FF}$	1.440604	—	**0.109905**	$<10^{-4}$		Exponential(λ)$_{FF}$	1.993823	—	**0.130947**	$<10^{-4}$
	Lognormal(μ,σ)$_{RF}$	16.29745	4.835015	**0.017584**	$<10^{-4}$		Lognormal(μ,σ)$_{RF}$	18.86203	6.624927	**0.015775**	$<10^{-4}$
	Lognormal(μ,σ)$_{FF}$	-15.49782	3.703538	**0.013719**	$<10^{-4}$		Lognormal(μ,σ)■	-29.71207	5.114532	**0.003523**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{RF}$	0.000020	-4394.015	**0.122391**	$<10^{-4}$		Zipf-cutoff(λ,α)$_{RF}$	0.000010	-9932.162	**0.138310**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{FF}$	0.630114	0.994591	**0.096630**	$<10^{-4}$		Zipf-cutoff(λ,α)■	0.733076	0.997743	**0.118005**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{RF}$	0.003668	0.330296	**0.008472**	$<10^{-4}$		Stretched-exponential(λ,α)$_{RF}$	0.006048	0.268884	**0.004963**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{FF}$	18.60976	0.062591	**0.015775**	$<10^{-4}$		Stretched-exponential(λ,α)$_{FF}$	46.07984	0.024394	**0.013773**	$<10^{-4}$
Rockyou	CDF-Zipf(C,s)	0.038208	0.185939	**0.045357**	$<10^{-4}$	Wishbone	CDF-Zipf(C,s)	0.017144	0.230503	**0.014775**	$<10^{-4}$
	Exponential(λ)$_{RF}$	5.63×10^{-7}	—	**0.288448**	$<10^{-4}$		Exponential(λ)$_{RF}$	5.52×10^{-7}	—	**0.183852**	$<10^{-4}$
	Exponential(λ)$_{FF}$	0.159546	—	**0.279785**	$<10^{-4}$		Exponential(λ)$_{FF}$	0.175439	—	**0.082278**	$<10^{-4}$
	Lognormal(μ,σ)$_{RF}$	13.44272	5.420420	**0.006333**	$<10^{-4}$		Lognormal(μ,σ)$_{RF}$	14.86753	5.343869	**0.010363**	$<10^{-4}$
	Lognormal(μ,σ)$_{FF}$	-20.92089	4.713954	**0.015878**	$<10^{-4}$		Lognormal(μ,σ)■	-30.92769	5.413191	**0.019383**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{RF}$	0.000078	-5106.710	**0.303566**	$<10^{-4}$		Zipf-cutoff(λ,α)$_{RF}$	0.000063	-5712.203	**0.190858**	$<10^{-4}$
	Zipf-cutoff(λ,α)$_{FF}$	0.998474	0.070754	**0.248137**	$<10^{-4}$		Zipf-cutoff(λ,α)■	0.345017	0.995556	**0.157640**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{RF}$	0.022520	0.253404	**0.024906**	$<10^{-4}$		Stretched-exponential(λ,α)$_{RF}$	0.012051	0.277269	**0.007151**	$<10^{-4}$
	Stretched-exponential(λ,α)$_{FF}$	19.06201	0.050322	**0.029889**	$<10^{-4}$		Stretched-exponential(λ,α)$_{FF}$	55.37731	0.019271	**0.042505**	$<10^{-4}$

† The RF in subscript means the distribution model is original in the rank-frequency (RF) system, and the FF subscript means the distribution model is original in the frequency-frequency (FF) system, but converted to the RF system. θ_1 and θ_2 denote the parameters characterizing each distribution, e.g., $\theta_1 = C$ and $\theta_2 = s$ for the CDF-Zipf distribution model. The two exponential distribution models, exponential$_{RF}$ and exponential$_{FF}$ only have one parameter ■, and thus θ_1 is needed.

‡ ■ D_{KS} among CDF-Zipf and alternative models are colored, and the ■-value denotes Monte Carlo approach (MCA) results of alternative models.

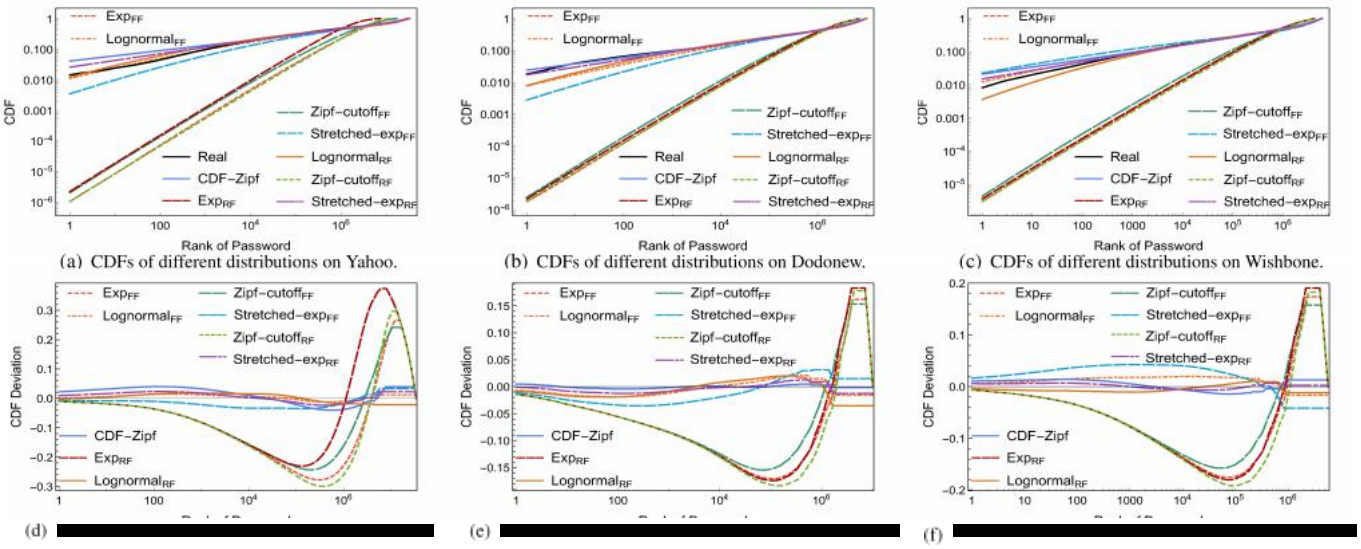

图4. 真实世界、CDF-Zipf和8种备选分布模型的CDFs和CDF偏差（其最大绝对值为KS统计量）。下标中的RF表示分布模型在秩频（RF）系统中，FF表示分布模型在频频（FF）系统中并转换为RF系统。

(表示为θ1和θ2)，这些模型与8个大规模数据集的KS统计数据如表7所示。我们还在图中展示了Yahoo、Dodonew和Wishbone数据集的CDF和KS统计图。4.

表七显示了还有其他三个分布，i。e.，lognormal射频，lognormalff和拉伸指数射频，提供了与CDF-Zipf分布相当的准确性。特别是对数正态射频或对数正常ff是最准确的在三个数据集(i。e.，雅虎，Rockyou，Chegg)，所以是拉伸指数射频(i.e.，天涯，数学之路，愿望骨)。现在，这些相对准确的分布模型是否能够通过MCA拟合优度测试是至关重要的：如果有一个模型可以通过MCA，密码比最先进的CDF-Zipf分布模型更有可能遵循它。因此，我们需要对备选分布进行拟合优度检验。

表VIII

射频系统中拉伸指数分布模型的密码子集和整个数据集的p值†

Dataset	Dataset size	D_{KS}	D'_{KS}	p-value	$\Delta\lambda$	$\Delta\alpha$	Dataset	Dataset size	D_{KS}	D'_{KS}	p-value	$\Delta\lambda$	$\Delta\alpha$
Yahoo	$0.05M$	0.004348	$0.000420 \sim 0.008380$	**0.058941**	10^{-4}	10^{-3}	Tianya	$0.05M$	0.007597	$0.000600 \sim 0.011220$	**0.010989**	10^{-4}	10^{-4}
	$0.1M$	0.004147	$0.000400 \sim 0.006650$	**0.035964**	10^{-5}	10^{-4}		$0.1M$	0.007633	$0.000440 \sim 0.009490$	0.002997	10^{-4}	10^{-4}
	$0.25M$	0.005693	$0.000372 \sim 0.004884$	$<10^{-4}$	10^{-5}	10^{-4}		$0.25M$	0.009860	$0.000396 \sim 0.004860$	$<10^{-4}$	10^{-5}	10^{-5}
	$0.5M$	0.006759	$0.000294 \sim 0.002962$	$<10^{-4}$	10^{-5}	10^{-4}		$0.5M$	0.011142	$0.000396 \sim 0.003090$	$<10^{-4}$	10^{-5}	10^{-5}
	Entire set	0.023873	$0.000733 \sim 0.000990$	$<10^{-4}$	10^{-6}	10^{-4}		Entire set	0.012353	$0.000274 \sim 0.001570$	$<10^{-4}$	10^{-4}	10^{-4}
Dodonew	$0.05M$	0.007377	$0.000480 \sim 0.009960$	0.001000	10^{-5}	10^{-4}	Chegg	$0.05M$	0.005935	$0.000200 \sim 0.005820$	$<10^{-4}$	10^{-3}	10^{-2}
	$0.1M$	0.006994	$0.000320 \sim 0.007350$	0.001190	10^{-5}	10^{-4}		$0.1M$	0.006080	$0.000150 \sim 0.003950$	$<10^{-4}$	10^{-5}	10^{-5}
	$0.25M$	0.005081	$0.000272 \sim 0.004516$	$<10^{-4}$	10^{-7}	10^{-5}		$0.25M$	0.005475	$0.000248 \sim 0.002492$	$<10^{-4}$	10^{-6}	10^{-5}
	$0.5M$	0.004433	$0.000192 \sim 0.003234$	$<10^{-4}$	10^{-6}	10^{-6}		$0.5M$	0.004772	$0.000174 \sim 0.002098$	$<10^{-4}$	10^{-5}	10^{-4}
	Entire set	0.013313	$0.000133 \sim 0.001081$	$<10^{-4}$	10^{-6}	10^{-4}		Entire set	0.012353	$0.000365 \sim 0.000816$	$<10^{-4}$	10^{-6}	10^{-5}
000webhost	$0.05M$	0.006582	$0.000160 \sim 0.006100$	$<10^{-4}$	10^{-5}	10^{-3}	Mathway	$0.05M$	0.005060	$0.000200 \sim 0.006440$	0.007992	10^{-5}	10^{-5}
	$0.1M$	0.006723	$0.000180 \sim 0.005120$	$<10^{-4}$	10^{-5}	10^{-3}		$0.1M$	0.004073	$0.000210 \sim 0.004590$	0.001998	10^{-3}	10^{-1}
	$0.25M$	0.006016	$0.000332 \sim 0.001988$	$<10^{-4}$	10^{-3}	10^{-4}		$0.25M$	0.003451	$0.000252 \sim 0.003304$	$<10^{-4}$	10^{-5}	10^{-4}
	$0.5M$	0.004818	$0.000234 \sim 0.001858$	$<10^{-4}$	10^{-3}	10^{-4}		$0.5M$	0.002505	$0.000244 \sim 0.002192$	$<10^{-4}$	10^{-3}	10^{-4}
	Entire set	0.008472	$0.000429 \sim 0.001195$	$<10^{-4}$	10^{-6}	10^{-5}		Entire set	0.004638	$0.000212 \sim 0.001349$	$<10^{-4}$	10^{-5}	10^{-5}
Rockyou	$0.05M$	0.006677	$0.000560 \sim 0.010600$	**0.018000**	10^{-5}	10^{-4}	Wishbone	$0.05M$	0.005762	$0.000440 \sim 0.009520$	**0.016983**	10^{-5}	10^{-4}
	$0.1M$	0.008781	$0.000490 \sim 0.006980$	$<10^{-4}$	10^{-5}	10^{-5}		$0.1M$	0.004427	$0.000340 \sim 0.006200$	**0.010989**	10^{-5}	10^{-4}
	$0.25M$	0.011188	$0.000456 \sim 0.003676$	$<10^{-4}$	10^{-5}	10^{-4}		$0.25M$	0.004552	$0.000364 \sim 0.004608$	**0.019802**	10^{-5}	10^{-4}
	$0.5M$	0.013306	$0.000416 \sim 0.002798$	$\blacksquare 10^{-4}$	10^{-6}	10^{-5}		$0.5M$	0.005141	$0.000242 \sim 0.003024$	$\blacksquare 10^{-4}$	10^{-5}	10^{-5}
	Entire set	0.024906	$0.000653 \sim 0.001072$	$<10^{-4}$	10^{-6}	10^{-4}		Entire set	0.007151	$0.000346 \sim 0.001634$	$<10^{-4}$	10^{-5}	10^{-5}

† $\Delta\lambda = |\lambda - \lambda'|$ and $\Delta_\alpha = |\alpha - \alpha'|$ record differences between λ and λ' as well as ■ and α'. The **bold** ■-values are those >0.01. Both Δ■ and Δ■ are very small, and only when the size of a subset is no greater than $0.5M$ ($\blacksquare M = 10^{\blacksquare}$, i.e., one million) can the ■-value be >0.01 to make MCA supports the model.

C. 在替代分发版中重新访问MCA

我们对所有8个备选模型进行了MCA处理，结果显示MCA拒绝了所有它们（见表VII中的p值列）。这就提出了MCA是否适合用于大型数据集的问题。

不失一般性，我们使用射频系统中拉伸指数分布的MCA结果(i。e.，拉伸指数的射频)为例，与其他7个模型的情况相似。拟合结果(包括参数λ和α，KS统计量*脱氧酮类醇*)的子集和整个数据集如表八所示。

结果表明：(1)λ(λ和λ之间的差异$\Delta\backslash\Delta$)和α(在α和α之间\)相对较小(介于$10^{-6}\sim 10^{-3}$只有一个例外在Chegg)，所以在那里在拉伸指数下，可以得出与定理1（在CDF-Zipf下成立）类似的结论射频分布；(2)对于大小大于0的子集和整个数据集。5M (i.e.,0.500万)，而一些KS的统计数据*脱氧酮类醇*值更小。g.，与CDF-Zipf相比，它们仍然不断地大于相应的第1类偏差(i。e.，统计随机性$'_{KS}$，所以p值是<值10^{-4}.似乎，无论分布模型如何，只要数据集的大小很大，MCA总是会拒绝(e。g.， ≥0.5M)。

在评估了8个备选模型之后，可以在MCA中识别出三个问题。首先，比如Sec中的CDF-Zipf的情况。III-B，MCA的第1类统计误差（当密码真正遵循时拒绝假设的X）理论上，当p值为<10时，理论上低至1.96%$^{-4}$，(见等式3)，但它实际上很大，因为MCA拒绝了所有的候选人。一个可能的原因是等式的P（H0）≈P（H1）的条件3为MCA(见Sec。在实践中并不成立。这可能是当*H1*是否定的*H0*如此*H1*可以包含多个发行版(e。g.，lognormal射频和拉伸指数射频)和P（H1）>P（H0）。第二，MCA产生不可忽略的p值(e。g.，$>10^{-4}$只有当数据集大小很小时。g.， ≤0.25M)。也就是说，在数据集较小时，MCA将接受一个分布，但当数据集较大时则拒绝该分布，这表明该方法是不完整的。第三，对于小数据集，MCA接受多种分布。例如，当雅虎子集为0时。05M ，CDF-Zipf的p值为>0.01，对数正态值射频，lognormalff，拉伸指数射频，也就是说，MCA将接受所有这三种分布。因为密码不太可能同时遵循多个分布，所以对于一个小的数据集，第2类统计错误（当密码不真正遵循它时接受一个分布）是不可忽略的。

因此，我们可以分析备选分布模型的偏差度量，并得到类似于定理2和理3的结果(对于CDF-Zipf，见Sec。III-C).由于这些缺陷，有必要采用一种新的拟合优度测量方法来评估这些分布模型。

D. 对数似然比检验

我们引入了一种新的基于可能性的拟合优度度量来替代无效的MCA。在统计学中，对数似然比检验（LRT）捕捉到了具有最大联合概率的事件(i。e.，可能性)最有可能从经验数据[17]，[20]，[39]中观察到。在实践中，可能性的日志形式(i。e.，对数似然法)经常用于方便计算。当分布模型X和Y都可能表征密码分布时，具有较大对数似然的模型更有可能发生，从而被观察到。LRT不仅广泛应用于各个领域。g.，物理和生物学[4]，[20]，[40])，也推荐NIST标准统计方法[1]。

通常，LRT处理相同的分布模型。例如，在CDF-Zipf中，是零假设*H0*是： C =*C0*和=0，以及备择假设*H1*是： C = C1和s=1≠(或C C0和ss≠0).因此，对数似然比为

$$LR = \sum_{i=1}^{N} (\ln(p_{H1_i}) - \ln(p_{H0_i})), \tag{11}$$

其中，N为唯一密码的个数，和$p_{H0\ i}$和p_{H1_i}模型的概率密度函数（PDFs）是由*H0*和*H1*假设。如果LR > 0，*H1*更有可能*H0*反之亦然

受这个想法的启发，我们开始了*H0*作为CDF-Zipf之后的密码，以及*H1*作为密码遵循X（一个替代的分发模型）。因为所有的模型都具有

表ix

针对小密码子集的替代模型与CDF -ZIPF的对数似然比†

Dataset	Dataset size	Lognormal_{RF}	Lognormal_{FF}	Stretched -exponential$_{RF}$
Yahoo	$0.05M$	-5.11×10^{3}	-2.31×10^{3}	4.85×10^{5}
	$0.1M$	-1.00×10^{4}	-3.44×10^{3}	1.01×10^{6}
	$0.25M$	-3.98×10^{4}	-2.02×10^{4}	2.58×10^{6}
	$0.5M$	-1.21×10^{5}	-4.64×10^{4}	5.18×10^{6}
	[redacted]	[redacted] $\times10^{7}$	-3.76×10^{7}	5.54×10^{8}
Dodonew	$0.05M$	8.62×10^{2}	-8.91×10^{2}	4.95×10^{5}
	$0.1M$	-3.28×10^{3}	-2.74×10^{3}	1.04×10^{6}
	$0.25M$	7.36×10^{3}	-1.48×10^{4}	2.71×10^{6}
	$0.5M$	-3.49×10^{4}	-4.79×10^{4}	5.54×10^{6}
	[redacted]	[redacted] $\times10^{6}$	-4.19×10^{6}	1.67×10^{8}
000webhost	$0.05M$	-8.53×10^{2}	-1.13×10^{4}	5.20×10^{5}
	$0.1M$	-1.18×10^{3}	-3.02×10^{3}	1.09×10^{6}
	$0.25M$	1.17×10^{3}	-9.46×10^{4}	2.90×10^{6}
	$0.5M$	-1.85×10^{4}	-2.55×10^{4}	5.99×10^{6}
	[redacted]	[redacted] $\times10^{6}$	-4.22×10^{6}	1.73×10^{8}
Rockyou	$0.05M$	-9.05×10^{3}	-2.42×10^{3}	4.63×10^{5}
	$0.1M$	-2.48×10^{4}	-1.09×10^{4}	9.35×10^{5}
	$0.25M$	-6.80×10^{4}	-3.45×10^{4}	2.35×10^{6}
	$0.5M$	-1.14×10^{5}	-7.71×10^{4}	4.66×10^{6}
	[redacted]	[redacted] $\times10^{7}$	-2.07×10^{7}	2.19×10^{8}

Dataset	Dataset size	Lognormal_{RF}	Lognormal_{FF}	Stretched -exponential$_{RF}$
Tianya	$0.05M$	-2.33×10^{3}	-3.82×10^{3}	4.43×10^{5}
	$0.1M$	-8.28×10^{3}	-1.17×10^{4}	9.09×10^{5}
	$0.25M$	-3.28×10^{4}	-3.33×10^{4}	2.29×10^{6}
	$0.5M$	-9.33×10^{4}	-1.01×10^{5}	4.56×10^{6}
	[redacted]	[redacted] $\times10^{7}$	-2.22×10^{7}	2.08×10^{8}
Chegg	$0.05M$	-2.10×10^{3}	-1.30×10^{3}	5.24×10^{5}
	$0.1M$	1.56×10^{3}	-1.14×10^{3}	1.11×10^{6}
	$0.25M$	-5.79×10^{2}	-1.27×10^{4}	2.94×10^{6}
	$0.5M$	-2.94×10^{4}	-1.93×10^{5}	6.09×10^{6}
	[redacted]	[redacted] $\times10^{6}$	-4.17×10^{6}	5.01×10^{8}
Mathway	$0.05M$	-1.66×10^{3}	-2.54×10^{3}	5.14×10^{5}
	$0.1M$	-1.09×10^{4}	-3.17×10^{3}	1.08×10^{6}
	$0.25M$	-1.26×10^{4}	-1.49×10^{4}	2.83×10^{6}
	$0.5M$	-3.41×10^{4}	-3.13×10^{4}	5.80×10^{6}
	[redacted]	[redacted] $\times10^{7}$	-1.71×10^{7}	1.82×10^{8}
Wishbone	$0.05M$	-3.18×10^{3}	-9.50×10^{2}	4.94×10^{5}
	$0.1M$	-4.69×10^{3}	-3.10×10^{3}	4.93×10^{5}
	$0.25M$	-2.64×10^{4}	-3.75×10^{4}	2.63×10^{6}
	$0.5M$	-7.71×10^{4}	-1.06×10^{5}	5.33×10^{6}
	[redacted]	[redacted] $\times10^{6}$	-4.63×10^{6}	8.90×10^{7}

† All [redacted]-values=0, and are calculated by treating LR[redacted] $= \ln(p_{H_{1i}}) - \ln(p_{H_{0i}})$ as a random variable. Details can be seen in Clauset et al.'s work [20].

由GSS拟合方法得到的参数，我们可以找出最可能最拟合的模型。需要注意的是，只有当一个模型是准确的时，才需要LRT(e。g.， DKS<0。1)，因为不准确的模型(e。g.， 两个zipf截止模型，见表七)将被排除。因此，我们专注于对数法态射频，lognormalff，和拉伸指数射频模型

anterna-的子集和整个数据集的LRT结果
使用CDF-Zipf作为的有效模型*H0*是
如表IX所示。整个数据集的结果表明
(1)只有拉伸指数射频在对数似然方面可以持续显著优于CDF-Zipf；虽然其他两种对数正态分布模型很有希望，但它们的对数似然明显低于CDF-Zipf，更不用说拉伸指数了射频，所以它们不太可能出现；
(2)拉伸指数的KS统计量射频小到

0.004638～0.024906 (avg.0.012459)，而CDF-Zipf更大，是0.004979～0.045357(平均。0.019142)；(3)开

8个数据集中的6个，总之，密码更有可能出现更多遵循拉伸指数射频分布

*1)在LRT和MCA之间的比较：我们比较了LRT和MCA，并揭示了LRT如何解决在MCA中确定的三个问题(见Sec。*IV-C).首先，对于问题1，在LRT中，对于每个备选模型，*H1*是由一个特定的分布模型组成的，而不是否定的吗*H0*，因此，[9]中的条件P（H0）≈P（H1）很可能成立，而等式3适用。此外，由于所有计算出的p值都接近于0，而不管数据集的大小如何，类型1的统计误差也不断地接近于0。第二，对于问题-2，LRT的结果显示了拉伸指数射频无论数据集的大小如何，其性能都优于其他数据集。这解决了不完全性问题，即分布模型对于小数据集被接受，但对于大数据集被拒绝。第三，甚至是问题3

如果数据集的大小小至0。05M～0.5M，拉伸指数的射频在所有数据集中具有最大的对数似然性。这意味着LRT不会接受小数据集的多个分布，而且第2类统计误差相对较低。

2)讨论：我们讨论CDF-Zipf、Zipf截止和拉伸指数模型。如表六所示，在每个系统中，这些分布的PDF内核是相似的。我们以拉伸指数（表示为变量为x）为例，并展开了指数的泰勒级数

PDF内核中的术语exp（－λxα）如下所示

$$\exp(-\lambda x)^{\alpha} = 1 - \lambda x^{\alpha} + \frac{\lambda^2}{2} x^{2\alpha} + O(x)^3. \quad (12)$$

如果我们在展开式中只取常数项1，拉伸指数被简化为CDF-Zipf。如果我们取α = 1，则为

拉伸指数的指数exp（－λx）等于它
因此CDF-Zipf和Zipf截止可以被看作是拉伸指数的变体。

这也部分解释了为什么拉伸指数在两个坐标上都比zipf截止要精确得多

系统有了一个额外的α，exp（－λxα）(回复。经验（－λr当x出现时，不要太小。r)很大，所以它可以调整Zipf项 $x^{-\alpha}$ (报告。rs^{-1})更灵活。数值证据也证实了这一点：
(1)对于zipf的截止ff，有α≈1，
这是一个非常小的exp（－λx）的直接结果；(2)

Zipf截止射频α0的表现更差。≪因此，在这两种系统中的拉链截止模型都是不准确的。

此外，拉伸指数射频模型也是有效的，因为它只比CDF-Zipf多花费大约25%的时间。例如，当在IntelE5-2680v4 2.4 GHz CPU上安装3200万个Rockyou密码时，拉伸指数的运行时间为2.33小时射频CDF-Zipf的工作时间为1.86小时。

最后讨论了LRT方法。虽然它可以有效地区分分布，但总是有潜在的改进空间。一方面，其他方法，如曼-惠特尼U检验也可能有效，但它们往往复杂且计算昂贵，因此在实践中应用较少。另一方面，作为创新的统计检验(e。g.，[7]，[8]，[27]，[35])的开发，未来可能会出现更合适的方法来解决大规模密码数据集上的拟合优度问题。

3)总结：我们首先研究了两个坐标系中的8种备选分布模型，发现有3种模型可与最先进的CDF-Zipf [51]相媲美。其次，我们重新讨论了替代模型上的MCA，并发现了这一点MCA拒绝了所有的要求。第三，我们引入了一种新的基于对数似然的良好拟合度量，并发现秩频坐标系中的拉伸指数始终具有最大的对数似然。我们还考虑了上述拉伸指数的拟合效率，发现它只花费了约25% (e。g.，3000万大小的数据集半小时)比CDF-Zipf多的时间，这在实践中是可以接受的。第四，我们揭示了LRT优于MCA

最小化统计误差的项，因此它更适合于密码分布的拟合优度使用。

V. 结论

本文以一种有原则的方法研究了密码分布的拟合优度问题。*特别是，我们使用理论和实验来验证样本量影响的民间传说，并定量地揭示了当真实世界的密码数据集较大时，蒙特卡罗方法（MCA）是不可取的(e。g.，≥025万)*。我们还研究了真实世界的密码偏差，并使用了模拟，发现1%的随机偏差就足以使MCA拒绝CDF-Zipf。这表明MCA对于测试大规模密码数据集是否遵循CDF-Zipf分布是无效的。

我们进一步研究了密码可能遵循的8种替代分发模型。我们在两个不同的坐标系中探索了这些模型，发现有三种模型都比较准确。特别是，拉伸指数模型的最大CDF偏差为0.004638～0.024906 (avg.0.012459)，而CDF-Zipf是0。004979～0.045357 (avg.0.019142).此外，在六

CDF-Zipf。*我们还重新考察了所有替代分布模型上的MCA，并进一步证明了它的无效。*作为替代，我们引入了似然比检验（LRT）作为一种更好的拟合优度度量，这表明密码更有可能遵循拉伸指数，我们相信，这项工作提供了更好地理解密码分发，并有助于对涉及密码分发的应用程序的评估。

在8个数据集中，拉伸指数比

附录

A. 随机数与金剖面搜索（GSS）拟合方法

我们介绍了转换方法[20]，[39]，用于生成遵循给定分布的随机数。假设*pr*为给定分布的概率密度函数（PDF），而u∈(0,1]是一个均匀分布的随机数(e。g.，一个标准的伪随机数[6])。

$$pr = p(u) \overset{杜都}{=} 博士 \overset{}{=} 博 \tag{13}$$

整合了双方关于r的问题，我们有

$$波多黎各 = \int_{r}^{\infty} prdr = \int_{u}^{1} du = 1 - u \tag{14}$$

所以r=“P^{-1}」(1-u)，其中P^{-1}的逆函数是什么吗累积分布函数（CDF）。该过程见Alg。3.因此，我们可以根据表六中的CDF表达式，根据每个分布生成随机数。此外，该算法还使我们能够进行Wang等人使用的黄金搜索（GSS）拟合方法。[51]，该过程如Alg所示。4.

B. 定理1、2、3的证明

在进入定理1之前，我们首先陈述一致收敛的概念(i。e.，均匀收敛)和用于检验数学函数序列是否均匀收敛的狄利克雷检验。

算法3根据给定分布生成随机数据（原）

输入：参数θ_1和θ_2给定分布X，数据集大小|DS|。
输出：理论数据集TheoSet。

```
1开始
2为i = 1到|DS| do
3  u = U (0,1]; /* U (0,1]表示均匀分布的随机数u∈(0,1]。*/
4          r = 「P^-1」(u) ; /* P^-1 (u)是的逆函数波多黎各…的
           分布区。*/
5温度[r]=温度[r]+1；6r的计数频率为fr;
7为r, f∈温度做
8  TheoSet.附加（f）;
9.将TheoSet按降序排列;
10输出：TheoSet。
```

算法4黄金剖面搜索（GSS）

输入：真实世界的数据集RealSet，数据集大小为|DS|，以及分布X。
输出：KS统计量*脱氧酮类醇*，对应的参数θ_1，θ_2在给定的分布X下。

```
开始
1   N为唯一密码的数量;
2   对于i = 1到最大限度的迭代来做
3     对于j = 1到最大限度的迭代来做
4       TheoSet = THEOGENX(θ1, θ2, |DS|); /*生成一个
        数据集与|DS|密码后的X(特征为θ1, θ2). */
5       DKS = max_{1≤r≤N} |CDF (TheoSet) - CDF (RealSet)|; /* CDF
6       是一个数据集的累积分布。*/
        (θ1, θ2) = (GSS 1d(θ1), GSS 1d(θ2)); /*GSS 1d是
7     普通的一维黄金切片搜索。*/
    输出：（θ1, θ2, 脱氧酮类醇）;
8
```

定义3：一个函数序列gi (x)（i∈N）一致收敛于域D上，如果对于每个e > 0有一个N (e)∈N，这样对于所有i≥N (e)和所有x∈D，一个有

|gi (x)-g (x)| < e，并记为gi (x)二g (x)。

引理1：（狄利克雷检验[42]）。对于函数序列$\sum_{i=0}^{\infty}$ai (x)bi (x)，其中x∈D，如果满足以下两个条件，则$\sum_{i=0}^{\infty}$ai (x)bi (x)在D上均匀收敛。

1)对于每个给定的x∈D，{ai (x)}$_{i=0}^{\infty}$对于i和ai (x)二0是单调的。

2)部分和$|\sum_{i=0}^{n}$bi (x)|≤M为一些M为任何x∈D和n∈N。

引理2：误差函数erf (y) =$2\pi^{-1/2}\int_0^y e^{-t^2}dt$

有erf (x) = 2(Φ($\sqrt{2}$x)-1，及其逆erf^{-1} (x)函数

当x1时，可以展开如下

$$\operatorname{erf}^{-2}(x) \sim \sum_{i=0}^{\infty} \eta^{-i} Q_i(\ln \eta)$$
$$= \eta - \frac{1}{2}\ln\eta + \eta^{-1}\left(\frac{1}{4}\ln\eta - \frac{1}{2}\right) + \eta^{-2}\left(\frac{1}{16}\ln^2\eta - \frac{3}{8}\ln\eta + \frac{7}{8}\right) + \eta^{-3}\left(\frac{1}{48}\ln^3\eta - \frac{7}{32}\ln^2\eta + \frac{17}{16}\ln\eta - \frac{107}{48}\right) + \ldots,$$

其中，η=-ln(π$^{1/2}$（1-x）)，Qi（ln η）是的一个多项以ln(η)为变量，和erf^{-2}(x)是正方形式

erf-的^{1}(x)．这些细节可以在[10]中看到。

现在，我们证明了定理1、2和3。

*定理1：假设Real Set和C的C和s*和s*满足C≈C*和s*≈；对于每个密码*

$f_i \geq fb$，估计误差$\epsilon_i = f_i - \mu_i$遵循正常分布$N(0, \sigma_i^2)$，以及密码为f的密码$i < fb$，$\sigma_i = 0$。在这种情况下，类型1的最大偏差（i。e.，统计随机性）最大D'_{KS}随着|DS|的增加而减少，并且有$\lim_{|DS|\to\infty} D'_{KS} = 0$。因此，p值随着|DS|的增加而降低，并且有$\lim_{JO\to\infty}$p值=0。

证明：首先，我们确定唯一传递的数量字*附注*谁的$f_i \geq fb$。由于fb是边界，我们有|DS|p*附注*=*fb*带p*附注*= C • s • *附注*$^{s-1}$如此

$$\text{附注} = \left(\frac{f_b}{|\mathcal{DS}|C \cdot s}\right)^{-\frac{1}{1-s}} \tag{16}$$

第二，因为$i \sim N(0, \sigma\epsilon_i^2)$，还有$\xi r =$*波多黎各*-$P_r^m = \sum_{i=1}^{r} \epsilon i / |DS|$遵循$N\left(0, \frac{\sum_{i=1}^{r} \sigma i^2}{|DS|^2}\right)$基于正态分布的性质。此外，考虑到$\sigma i < \sqrt{\mu_i}$当$f_i \geq fb$和$\sigma i f$时$=0 i < fb$，我们有

$$\sqrt{\frac{\sum_{i=1}^{r}\sigma_i^2}{|\mathcal{DS}|^2}} < \sqrt{\frac{\sum_{i=1}^{r}\mu_i}{|\mathcal{DS}|^2}} = \sqrt{\frac{\sum_{i=1}^{r}\mu_i}{|\mathcal{DS}|}\cdot\frac{1}{|\mathcal{DS}|}} = \sqrt{\frac{P_r^m}{|\mathcal{DS}|}}, \tag{17}$$

其中P_r^m*是真正的CDF*。因此，对应的概率为，

$$P\left(D_{KS} \leq 2a_0\sqrt{\text{附注}\, P_r^m/|DS|} \,||\, > 2(1 - (a\Phi_0))\right) \tag{18}$$

保存*因此，对于最大的阶统计量x |ξr|，我们有*马$_r$P(马$_r$x

$$\begin{aligned} &= 1 - P(\hat{\exists}_r \,|\xi r| > \overline{P_r^m/|\mathcal{DS}|}) \\ &\text{a0 mP(|DS|)} \\ &\geq 1 - \text{附注}\left(1 - P(|\xi r| \leq a_0\sqrt{P_r^m/|\mathcal{DS}|})\right) \\ &\geq 2Nb((a\Phi_0) - 1) + 1. \end{aligned} \tag{19}$$

基于此结果，我们研究了第1类偏差D'_{KS}（i。e.，统计随机性），由拟合TheoSet产生。作为C*≈ C，s*≈ s，D'_{KS}是两个随机生成的TheoSets之间的最大CDF偏差吗$_r$(1)和P_r(2).因此，$D'_{KS} = x\ |P$马$_r$(1)- P_r(2)$_r$2ma|≤x |ξr|和

$$\begin{aligned} &P\left(D_{ks} \leq 2a_0\sqrt{P_r^m/|\mathcal{DS}|}\right) \\ &> P\left(\text{马}x\ |\xi r| \leq a_0\sqrt{P_r^m/|\mathcal{DS}|}\right) \\ &> 2Nb((a\Phi_0) - 1) + 1. \end{aligned} \tag{20}$$

设置α=2Nb(aΦ0) - 1) + 1，max D'_{KS}有

$$\begin{aligned} \max D'_{KS} &= 2\sqrt{P_{N_b}^m/|\mathcal{DS}|}\cdot\Phi^{-1}\left(1 - \frac{1-\alpha}{2N_b}\right) \\ &= 2\sqrt{2P_{N_b}^m/|\mathcal{DS}|}\cdot\text{erf}^{-1}\left(1-(1-\alpha)\left(\frac{f_b}{|\mathcal{DS}|Cs}\right)^{\frac{1}{1-s}}\right), \end{aligned} \tag{21}$$

在哪里$\Phi^{-1}(x)$ =erf^{-1}（2x-1）基于他们的定义。据此，我们将证明林|DS|→∞max D'_{KS}= 0.

首先，如引理2中所示，我们取x = (1-(1-α)(|SD, $\frac{f_b}{\mathcal{S}|C}^{\frac{1}{1-s}}$) and $\eta = -\ln(\pi^{1/2}(1-$∞ , there,h)ave $x \to 1$ and $\eta \to \infty$. S x)), so when $|\mathcal{DS}| \to$ econd, we let $a_i(\eta) =$ $\eta^{-(i-1)}$和bi(η) =η^{-1}Qi（ln η）为i≥1。我们现在证明那个人$\eta \to \infty \frac{\ln^i \eta}{\eta}$=0来显示lim η→∞bi(η) = 0用于任何i≥1。当k = 1，lim$_{\eta\to\infty}\frac{\ln \eta}{\eta}$= lim$_{\eta\to\infty}\frac{1}{\eta}$= 0（L‘Hopital规则）。通过数学归纳法，我们假设lim$_\eta \frac{\ln^i \eta}{\eta}$0，当k i 1有时

$$\lim_{\eta\to\infty}\frac{\ln^{k+1}\eta}{\eta} = \lim_{\eta\to\infty}\frac{(\ln^{i+1}\eta)'}{\eta'} = (i+1)\lim_{\eta\to\infty}\frac{\ln^i\eta}{\eta} = 0. \tag{}$$

因此，lim$_{\eta\to\infty}\frac{\ln^i \eta}{\eta}$=0为任何i≥1。自从bi(η)=$\sum_{k=0}^{i} ck^{\ln k}(\eta)$—η（c0，c1，...*ci*系数），lim η→∞bi(η)=0和lim η→∞$|\sum_{i=0}^{\eta} bi(\eta)|$ = 0对于任何给定的n≥1。这意味着存在η0和M，对于任何η > η0和n ≥ 1，$|\sum_{i=0}^{\eta} bi(\eta)| < M$。此外，自从ai(η) = η以来$^{-(i-1)}$二0当η→∞和{ai(η)}时$_{i=0}^{\infty}$单调减小，$\sum_{i=1}^{\infty}\eta^{-i}Qi$（ln η）=$\sum_{i=1}^{\infty} ai(\eta)bi(\eta)$均匀收敛（参见引理1）。因此，极限计算和无限求和运算是可交换的，i。e.，

$$\lim_{\to\infty}\sum_{i=1}^{\infty}\eta^{-i}Q_i(\ln\eta) = \sum_{i=1}^{\infty}\lim_{\eta\to\infty}\eta^{-i}Q_i(\ln\eta) = \sum_{i=1}^{\infty}0 = 0, \tag{2}$$

3）所以最大D'_{KS}这取决于等式的前两项15 i.e.，

$$\text{李}\lim_{|DS|\to\infty}\frac{1}{(\eta - \frac{1}{2}\ln\eta)/|DS|}. \tag{24}$$

自从η=-ln(π1/2((1 - α))())，Eq. $\frac{f_b}{Cs|\mathcal{DS}|}^{\frac{1}{1-s}}$24是根据L‘Hopital规则证明收敛于0。因此，有林姆|DS|→∞max D'_{KS}=为0，因此也因此为lim|DS|→∞D'_{KS}=0基于挤压定理。因此，p值=(#{$D'_{KS}j$ /$D'_{KS}j$>脱氧酮类醇，1 ≤ j ≤*JO*}+1)/（JO+1）（参见Alg中的第6行。1)随着|DS|的增加而减少，并变为1/（JO + 1）*JO*→ ∞.

定理2：在绝对偏差度量中，如果是密码在SmuSet中，偏离为$pdv_i^s = \hat{\delta}[i_1, i_2]$• k*A*对于我∈[我1，i2]和偏差度为k*A*满足0≤k*A*≤ |p$\hat{d}$v [i1，i2]|，然后是最大最大D_{KS}^s增加为k*A*增加量

证据：如第二节所述。三、c级，D级的超端$'_{KS}$具有最大D_{kS}^sCDF（SmuSet）CDF（TheoSet），即，

$$\max D_{KS}^s = \begin{cases} \max\limits_{1\leq r<r_1}\dfrac{k_A P_r P_{[i_1,i_2]}}{1+\hat{\delta}\cdot k_A P_{[i_1,i_2]}} & 1 \leq r < r_1 \\ \max\limits_{r_1\leq r<r_2}\dfrac{k_A P_r(1-P_{[i_1,i_2]})}{1+\hat{\delta}\cdot k_A P_{[i_1,i_2]}} & r_1 \leq r < r_2 \\ \max\limits_{r_2\leq r\leq N}\dfrac{k_A P_{[i_1,i_2]}(1-P_r)}{1+\hat{\delta}\cdot k_A P_{[i_1,i_2]}} & r_2 \leq r \leq N. \end{cases} \tag{25}$$

where $P_r = \sum_{i=1}^{r} f_i$ $P_{[i_1 i_2]} = \sum_{i1}^{i_2} p_i$ and $\delta\hat{} = \delta\hat{}_{[i,i_1 z]}$ see that: (1) D_{KS}^{s} i ncrea ses as r increasesw 他n $1 \leq$. We can

$r < r_2$,

(2)当r的增加而减小$2 \leqslant r \leqslant N$, so D_{KS}^{s}在rr处得到最大值2并有

$$\max D_{KS}^{s} = \frac{k_A P_{[i_1,i_2]}(1 - P_{[i_1,i_2]})}{1 + \hat{\delta} k_A P_{[i_1,i_2]}}. \tag{26}$$

假设我1是固定的，我们研究了最大D的单调性$_{KS}^{s}$对我2和k$_{A}$，我们的偏导数如下。

$$\begin{cases} \dfrac{\partial \max D_{KS}^{s}}{\partial i_2} = \dfrac{k_A(-\hat{\delta}k_A \cdot P_{[i_1,i_2]}^2 - 2k_A \cdot P_{[i_1,i_2]} + 1)}{(1+\hat{\delta}k_A P_{[i_1,i_2]})^2} \\ \dfrac{\partial \max D_{KS}^{s}}{\partial k_A} = \dfrac{P_{[i_1,i_2]}(1-P_{[i_1,i_2]})}{(1+\hat{\delta}k_A P_{[i_1,i_2]})^2}. \end{cases} \tag{27}$$

$\frac{\partial \max D_{KS}^{s}}{\partial k_A}$一方面，>0始终成立。另一方面手，使> 0，i。$\frac{\partial \max D_{KS}^{s}}{\partial i_2}$e.，有需要

$$-k\hat{\delta}_A \cdot P_{[i1,i2]}^2 - 2P[i_1,i2] + 1 \tag{28}$$

> 0.$\hat{\delta}$如果是1，对于任意的k$_A$那儿有

$$\frac{-1-\sqrt{1+k_A}}{k_A} < P_{[i_1,i_2]} < \frac{-1+\sqrt{1+k_A}}{k_A}. \tag{29}$$

制作等式28持有，$P[i_1,i2]$<inf $\approx$ 0.$\frac{-1+\sqrt{1+k_A}}{k_A}$41，也就是对我们所考虑过的密码感到满意(e。g.，前100名独特的口令在这种情况下，最大DKS随i的增加而增加2增加量否则，如果该符号是负数，则为i。e.，$\hat{\delta}$=−1，有需要$P[i_1,i2]$<inf$^{1-\sqrt{1-k_A}}_{k_A}$<$\frac{1}{2}$或$P[i_1,i2]$>sup$^{1+\sqrt{1-k_A}}_{k_A}\to\infty$制作平衡28.持有，这也很满意。总而言之，最大D$_{KS}^{s}$随着i的增加2当=1或−1出现时会增加。$\hat{\delta}$

定理3：在相对偏差度量中，如果是密码 are deviated as $pdvi_s = \hat{\delta}_i \cdot |pdv_{i/kR}$ $0 < k_R \leq 1$, then那 ma ximum max for $i \in [1, N_0]$ and增加量

证明：类似于绝对点问偏差的证明，最大最大D$_{KS}^{s}$可以表示为，

$$\max D_{KS}^{s} = \begin{cases} \max\limits_{1\leq r<N_0} |\dfrac{(1-P_r)W_r + P_r W_{[r+1,N_0]}}{1+W_{N_0}}|k_R & 1 \leq r < N_0 \\ \max\limits_{N_0\leq r\leq N} \dfrac{|W_r|(1-P_r)}{1+W_{N_0}}k_R & N_0 \leq r \leq N, \end{cases} \tag{30}$$

在哪里$波多黎各_r = \sum_{i=1}^{r} pi$是TheoSet中的CDF，$与_r = \sum_{i=1}^{r} pi \cdot$ pdvi和$W[r+1,N0] = \sum_{i=r+1}^{N_0} pi \cdot pdvi$.因此，D$_{KS}^{s}$增加量如k$_R$增加量然而，自从p的符号$\hat{d}$vi不是固定的，我们不知道最大值D在哪里$_{KS}^{s}$得到其最大值。

C.分配的转换

在本节中，我们使用幂律的转换(i。e.，展示了如何将分布模型从频率-频率（表示为FF）转换为秩-频率（表示为RF）坐标系。

命题1：使用PDF格式的幂律分布 $p(x) = (\alpha-1)x^{-\alpha}$，在FF系统中，互补的CDF记录了一个随机变量X的给定频率r的概率。可以表示为$r = x^{1-\alpha}$.

证明：在FF系统中，在RF系统中可以被转换为 $P(X \geq x) = \int_{x}^{\infty} p(x)dx = x^{-\alpha+1}$. (31)

另一方面，在RF系统中，如果密码PW发生x次的秩为r，则事件X≥x可以解释为唯一密码的比例等级不等于等级r。因此，存在P（X≥x）= r/N（N是唯一密码的数量）。由于x = |DS|pr，（|DS|是数据集的大小，pr是第r个密码的PDF），所以我们有

$$P(X \geq |DS|pr) = x^{-\alpha+1} = r/N. \tag{32}$$

这样，$pr^{\frac{1}{1-\alpha}} \propto r$，和CDF$波多黎各 \propto r^{\frac{2-\alpha}{1-\alpha}}$.因此FF系统中的幂律可以转换为射频系统中的CDF-Zipf，它们是等价的。

参考文献

[1](10月。2021).*统计方法的电子手册*。在线可用： http://www。itl.nist.gov/div898/handbook/

[2] M.阿卜杜拉，F。本哈穆达和P。麦肯齐，“JPAKE密码认证的密钥交换协议的安全协议”。*IEEE模拟。安全。隐私，2015年5月，页*。571－587.

[3] L.A.*亚当克，Zipf，幂律，和帕累托-一个排名教程*。印度钦奈：信息动力学实验室，惠普实验室，2000年。在线可用： https://www。hpl.hp.com/research/idl/papers/ranking/ranking.html

[4] M.阿尼西莫娃，J。P.比拉夫斯基和Z。“似然比检验检测适应性分子进化的准确性和威力”，国立台湾大学研究所硕士论文。*比奥尔。Evol*。，卷。18日，没有。8, pp.1585-1592年8月。2001.

[5] A.L.Barabasi, H.JeongZ。内达，E。拉瓦斯，A。舒伯特和T。“科学合作的社会网络的进化”，物理学。一、*统计。机械。应用程序*。，卷。311, pp.590-614年8月。2001.

[6] E.巴克和J。*Kelsey，NIST特别出版物800-90A修订版1：建议使用确定性随机比特生成器生成随机数*。在线可用： https:// nvlpubs。nist.gov/nistpubs/SpecialPublications/NIST.弹簧1。.800-90pdf

[7] D.J.本杰明等人。，《重新定义统计意义》，《自然哼》。*贝哈夫*。，卷。2、没有。1, pp.6－ 10, 2018.

[8] Y.本杰米尼和Y。“控制错误发现率：一种实用而强大的多重测试方法”，J。*罗伊斯达社会B, Methodol.*，卷。57岁，没有。1, pp.289-300年8月。1995.

[9] J.O.伯杰，L。D.棕色和R。L.“固定和顺序简单假设检验的统一条件频率和贝叶斯检验，”安。*停滞不前*。，卷。22日，没有。4, pp.1787-1807年12月。1994.

[10] J.M.布莱尔，C。A.爱德华兹和J。H.约翰逊，“误差函数的逆的有理切比雪夫近似”，数学。*压缩*。，卷。30岁，没有。136, pp.827－830, 1976.

[11] J.Blocki。达塔和J。邦诺，“不同的私人密码频率列表”，在Proc。*网络。分界线。西斯特。安全。模拟*。, 2016, pp.1－43.

[12] J.Blocki, B.Harsha和S。周，“论离线密码破解的经济学”，在《程序》中发表。*IEEE模拟。安全。隐私》（SP），2018年5月，页*。853－871 .

[13] J.布洛基和A。Sridhar，“客户端-现金：保护主密码免受离线攻击”，在Proc。*第11届ACM亚洲会议。压缩。通勤。安全*。2016年5月，页。165－ 176 .

[14] J.“猜测的科学：分析一个包含7000万个密码的匿名语料库”。*IEEE模拟。安全。隐私，2012年5月，页*。538－552.

[15] 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，卷*。58岁，没有。7, pp.78－87, 2015.

[16] T.布拉德利，J。卡梅尼施，S。Jarecki。莱曼，G。Neven和J。徐，“密码认证公钥加密”，在程序。ACNS, 2019, pp.442－462 .

[17] O.*《科学与工程中的统计理论与方法论》，《技术计量学》，第1卷*。7、没有。3, pp.451-453年8月。1965.

[18] S.A.乔杜里，A。Irshad, K.Yahya，N。库马尔，M。阿拉扎布和Y。B.Zikria，“旋转支持隐私：针对基于云的物联网环境的一种改进的轻量级认证方案”，ACM Trans。*互联网技术*。，卷。21日。3, pp.6月1日至19日。2021.

[19] Y.程，C。徐，Z。海和Y。李彦，“通过深度注意的编解码器模型生成密码助记符”，IEEE翻译。*取决于。安全的计算*。，卷。19日，没有。1, pp.77-90年1月。2022.

[20] A.克劳塞特，C。R.沙利兹和M。E.J.纽曼，“经验数据中的幂律分布”，暹罗第一版。，卷。51岁，没有。4, pp.661－703, 2009.

[21] A.达斯，J。博诺，M。凯撒，N。博里索夫和X。王，“密码重用的复杂网”，在程序。*网络。分界线。西斯特。安全。模拟*。, 2014, pp.23－26.

[22] P.达斯，J。黑塞和A。莱曼，“DPaSE：分布式密码认证的对称加密”。*ASIACCS，2022，pp.*1 - 15.在线可用：https://asiaccs2022。conferenceservice.jp/acceptedpapers/all/

[23] A.C.戴维森和D。V.*欣克利，引导程序方法及其应用程序*。剑桥，美国K.：剑桥大学。出版社，1997年。

[24] T.N.DinhH。张，D。T.阮和M。T.泰国，“大规模社交网络中时间关键活动的高效病毒式营销”，IEEE/ACM Trans。*网络*。，卷。22日，没有。6，pp.2001 - 2011，十二月2014.

[25] R.A.费雪，《研究人员的统计方法》，《统计学的突破》。中国，瑞士：施普林格，1992年，第二页。66 - 70.

[26] S.N.古德曼，《p值和贝叶斯：一个适度的建议》，流行病学，第1卷。12日，没有。3，pp.295-297,2001年5月。

[27] J.P.“建议将p值阈值降低到。005年，“詹姆斯，卷。319，没有。14，pp.1429 - 1430，2018.

[28] H.JeongZ。Neda和AL。.-巴拉巴西，“衡量进化网络中的优先依恋”，欧元。*物理学。拉脱维亚的*，卷。61岁，没有。4，p.567，2003.

[29] F.基弗和M。“零知识密码策略检查和基于验证者的PAKE”，在Proc。*ESORICS，2014，pp.*295 - 312.

[30] D.马龙和K。“调查密码选择的分布”，在专业课程。*21st Int.会议万维网，4月*。2012，pp.301 - 310 .

[31] F.马西斯，H。I.Fawaz和M。“虚拟现实中知识驱动的生物识别认证”。*扩展Abstr。CHI Conf.哼唱因素计算。西斯特*。，Apr.2020，pp.1 - 10.

[32] P.梅耶尔，Y。Zou，F.肖布和A。J.“现在我有点生气了：‘个人’的意识、感知和对影响他们的数据泄露的反应。”*USENIX SEC，2021，pp*。393 - 410.

[33] M.“幂律和对数正态分布的生成模型的简史”，互联网数学。，卷。1、没有。2，pp.226 - 251，2004.

[34] R.莫里斯和K。汤普森，“密码安全：一个案例的历史”，被评论道。*ACM，卷*。22日，没有。11，pp.594 - 597，1979.

[35] S.P.阮，美国。H.帕姆，T。D.阮和H。T.勒，“一种基于推理模型的假设检验的新方法”。IUKM，2016，pp.532 - 541.

[36] S.Oesch和S。Ruoti说，“那是那时，现在是：对密码生成、存储和自动填充基于浏览器的密码管理器的安全性评估。”*USENIX SEC，2020年，页*。2165 - 2182 .

[37] M.L.《对洛特卡定律的实证检验》，J。*美国人社会通知。科学*。，卷。37日，没有。1，pp.26 - 33，1986.

[38] S.皮尔曼等人。，“让我们来仔细看看：观察它们的自然栖息地的密码。”*ACM SIGSAC Conf。压缩。通勤。安全*。十月2017，pp.293 - 310 .

[39] W.H.按B。P.弗兰纳里，S。A.特库科尔斯基和W。T.*兽医学家，帕斯卡尔的数值食谱：科学计算的艺术*。剑桥，美国K.：剑桥大学。出版社，1989年。

[40] R.Protassov，D.A.范戴克。康纳斯，V。L.Kashyap和。“统计学，小心处理：用似然比检验检测多个模型成分，”天文台。*J.*，卷。571，没有。1，p.545，2002.

[41] R.M.“样本量对显著性检验意义的影响”，Amer。*停滞不前*。，卷。40岁，没有。4，pp.313 - 315，1986.

[42] W.*《数学分析原理》，第1卷*。美国，纽约：麦格劳-希尔，1964年。

[43] T.Sellke，M.J.巴亚里和J。O.伯格，“检验精确零假设的ρ值的校准”，Amer。*统计学家，卷*。55岁，没有。1，pp.62-71年，2月。2001.

[44] R.Shay等。，“正确的马电池短语：探索系统分配的密码短语的可用性”。*8th Symp.可用的隐私安全。（汤），2012，页*。1 - 20.

[45] M.雪瓦尼安和S。论文，“2D-2FA：双因素认证中的一个新维度”。*安努。压缩。安全。应用程序。会议*十二月2021，pp.482 - 496 .

[46] J.Srinivas，。K.Das，M。Wazid和N。库马尔，“匿名轻量级混沌基于地图的工业物联网认证关键协议协议，”IEEE跨版。*取决于。安全的计算*。，卷。17日，没有。6，pp.1133-1146年，11月。2020.

[47] J.谭，L。鲍尔，N。克里斯汀和L。F.鹤，“实用的建议，更强，更有用的密码结合最小强度，最小长度，和区块列表的要求，”在Proc。*ACM SIGSAC Conf。压缩。通勤。安全*。十月2020，pp.1402 - 1426 .

[48] J.蔡，R。El-Gabalawy，W.H.雪橇，S。M.南威克和R。H.“美国退伍军人创伤后成长：退伍军人全国健康和复原力研究的结果”，心理学。*医学*，卷。45岁，没有。1，pp.165-179年1月。2015.

[49] L.瓦斯(8月。2016).*比起使用密码，人们更喜欢使用生物识别技术*。在线可用的：https://裸体安全。索福斯。com/2016/08/16/people-like-using-passwords-way-more-than-biometrics/

[50] R.Veras，C。柯林斯和J。“关于密码的语义模式及其对安全的影响”。*网络。分界线。西斯特。安全。模拟*。，2014，pp.1 - 16.

[51] D.王，H。程，P。王，X。黄和G。吉安，“Zipf密码法”，IEEE Trans。*影响取证安全，卷*。12日，没有。11，pp.2776-2791年11月。2017.

[52] D.王，D。他，H。程和P。王，“模糊psm：使用模糊概率上下文无关语法的一种新的密码强度计”，在专业部分。*第46圈。IEEE/IFIP Int.会议可靠的心理。网络。（DSN），6月*。2016，pp.595 - 606.

[53] D.王，P。王，D。他和Y。田，《生日、姓名和双重安全：了解中国网民的密码》。*USENIX SEC，2019，页*。1537 - 1555 .

[54] D.王，Z。张，P。王，J。Yan和X。黄，“有针对性的在线密码猜测：一个被低估的威胁”。*ACM SIGSAC Conf。压缩。通勤。安全*。十月2016，pp.1242 - 1254 .

[55] K.C.王和M。K.Reiter，“检测用户在自己账户上的凭证填充”。*USENIX SEC，2020年，页*。2201 - 2218 .

[56] K.C.王和M。K.“如何结束密码重用在网上，”，在Proc。*网络。分界线。西斯特。安全。模拟*。，2019，pp.1 - 15.

[57] S.沃丁斯基。(11月。2021).*2021年最糟糕的200个密码就在这里，哦，天哪*。在线可用：https://gizmodo。com/the-200-worstpasswords-of-2021-are-here-and-oh-my-god1848073946

[58] J.燕，A。布莱克威尔，R。安德森和A。“密码记忆与安全：经验结果”，IEEE安全。*隐私，卷*。2、没有。5，pp.25 - 31，Sep./Oct.2004.

[59] W.杨，N。李，哦。乔杜里，A。熊和R。W.“基于句子的密码生成策略的实证研究”。*ACM SIGSAC Conf。压缩。通勤。安全*。十月2016，pp.1216 - 1229 .

[60] G.K.*Zipf，人类Behav。原则上最少的努力：对人类生态学的导论*。剑桥，美国K.：拉维尼奥图书公司，2016年。

侯振铎获得了B元奖。S.2015年6月，中国成都四川大学数学学位。他目前正在攻读博士学位。D.北京大学数学科学学院学士学位，北京，中国。他的研究兴趣包括应用密码学和基于密码的认证。曾获国家奖学金和北京大学教学资助奖学金。

丁王获得了博士学位。D.信息学学位2017年北京大学安保。目前为南开大学的正教授。他在IEEE安全与隐私、IEEE安全与隐私、ACM中国化学会、NDSS、USENIX安全、IEEE TRANSAC - TIONS关于可靠和安全计算、IEEE关于信息取证和安全的交易等领域发表了80多篇论文。他的研究已经被200多家媒体报道，如《每日邮报》、《福布斯》、IEEE频谱公司和美国通信公司2017年“庆祝中国计算机科学研究的文章选择”上，SP800-63-2的修订。他的研究兴趣集中在密码、认证和可证明的安全性上。他作为PC主席/TPC成员参与了超过60个国际会议，如ACM中国化学会2022、NDSS 2022、PETS 2023/2022、ACSAC 2020-2022、ACM AsiaCCS 2022/2021、ICICS 2018-2022和SPNCE 2020-2022。曾获“ACM中国杰出博士论文奖”、2018年中国文学最佳论文奖、中国密码研究学会杰出青年奖、教育部自然科学一等奖。