On the Implications of Zipf's Law in Passwords

Ding Wang[1] and Ping Wang[1,2(✉)]

[1] School of EECS, Peking University, Beijing 100871, China
{wangdingg,pwang}@pku.edu.cn
[2] School of Software and Microelectronics, Peking University, Beijing 100260, China

Abstract. Textual passwords are perhaps the most prevalent mechanism for access control over the Internet. Despite the fact that humanbeings generally select passwords in a highly skewed way, it has long been assumed in the password research literature that users choose passwords randomly and uniformly. This is partly because it is easy to derive concrete (numerical) security results under the uniform assumption, and partly because we do *not* know what's the exact distribution of passwords if we do not make a uniform assumption. Fortunately, researchers recently reveal that user-chosen passwords generally follow the Zipf's law, a distribution which is vastly different from the uniform one.

In this work, we explore a number of foundational security implications of the Zipf-distribution assumption about passwords. Firstly, we how the attacker's advantages against password-based cryptographic protocols (e.g., authentication, encryption, signature and secret share) can be 2–4 orders of magnitude more accurately captured (formulated) than existing formulation results. As password protocols are the most widely used cryptographic protocols, our new formulation is of practical significance. Secondly, we provide new insights into popularity-based password creation policies and point out that, under the current, widely recommended security parameters, usability will be largely impaired. Thirdly, we show that the well-known password strength metric α-guesswork, which was believed to be parametric, is actually nonparametric in two of four cases under the Zipf assumption. Particularly, nine large-scale, real-world password datasets are employed to establish the practicality of our findings.

Keywords: User authentication · Zipf's law · Password-based protocol · Password creation policy · Password strength metric

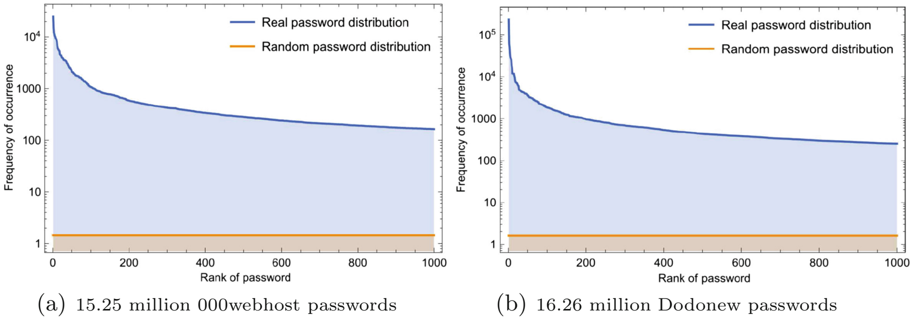
1 Introduction

With so much of our lives digital and online, it is essential that our digital assets are well-protected from unauthorized access. Since passwords are easy to use, low-cost to implement and convenient to change, almost every web service today authenticates its users by passwords, ranging from low value news portals and technical forums, moderate value e-commerce and email to highly sensitive financial transactions and genomic data protection [21]. Although its security

I. Askoxylakis et al. (Eds.): ESORICS 2016, Part I, LNCS 9878, pp. 111–131, 2016.
DOI: 10.1007/978-3-319-45744-4_6

weaknesses (e.g., vulnerable to guessing attacks [29]) and usability issues (e.g., typo and memorability [17,39]) have been constantly articulated, and a variety of alternative authentication methods (e.g., multi-factor authentication and graphical passwords) have also been successively proposed, password-based authentication firmly remains the most prevalent mechanism for access control and reproduces in nearly every new service and system. As no alternative schemes can provide all the benefits that passwords offer [9], and further due to inertia and economic reasons [19], passwords are likely to continue to be the dominant authentication mechanism in the foreseeable future.

Since system-assigned passwords are of poor usability, in most cases users are allowed to select passwords by themselves. It is well-known, however, that users tend to choose weak passwords for convenience (e.g., passwords based on dictionary words, meaningful phrases and personal information [28,39]) and to reuse or slightly modify existing passwords for saving efforts [35]. Thus, it has been widely recognised (see [8,30]) that user-chosen passwords are *unlikely* to be uniformly randomly distributed. For a concrete grasp of the distribution of passwords, we exemplify two large-scale real-world password lists in Fig. 1. Our other datasets also exhibit similar distributions but cannot be shown here only due to space constraints. Clearly, they are all far from a uniform distribution. Now a critical question naturally arises: *if human-chosen passwords do not follow a uniform distribution, then what is their exact distribution?*



(a) 15.25 million 000webhost passwords

(b) 16.26 million Dodonew passwords

Fig. 1. Frequency distribution of two large-scale real-word password datasets.

This question has remained as an open problem for decades, which is partly due to the scarcity of real-world password datasets (because real-life passwords are sensitive and difficult to gather) and partly due to the fact that resolving this problem involves some recent advancements in the inter-discipline knowledge, such as computational statistics and natural language processing (NLP). As a result, among these password-related works that *must* rely on an explicit assumption about the distribution of passwords, most ones (e.g., password-based cryptographic protocols such as authentication [1], encryption [5], signature [18] and secret sharing [40]), *reluctantly*, make the unrealistic assumption that user-chosen passwords are uniformly randomly distributed, while the few remaining ones *wittingly* make various assumptions about the password distribution just for convenience of security analysis (e.g., the binomial distribution in [13] and min-entropy model in [2,27]). As for these works that do *not* have to rely on an explicit assumption about the distribution of passwords, they (e.g., password creation policies [34] and strength metrics for password datasets [8]) generally simply *avoid involving* an assumption about the password distribution.

As we will demonstrate in this work, the above unrealistic assumptions about the distribution of passwords often give rise to serious security and usability issues (e.g., there are, as shown in Sect. 3, *two to four orders of magnitude underestimation* in the attacker's online guessing advantages between a uniform assumption made in [1,5,18,23,25] and the reality. For these works (e.g., [8,34]) that avoid involving an assumption about the password distribution, many important properties or goals that actually rely on an assumption are left undiscussed. For instance, if the password creation policy proposed in [34] is imposed by a web service, what fraction of users will be potentially annoyed? This kind of prediction is important yet virtually impossible if one makes no assumption about the password distribution.

Fortunately, with the help of 14 large-scale datasets and by introducing a number of statistical and NLP techniques, researchers recently reveal that human-chosen passwords generally follow a Zipf distribution [36]. This theory has already been successfully adopted into the "GenoGuard" genome cryptosystem [21] and "CASH" password hash scheme [7]. It implies that the frequency of passwords decreases polynomially with the increase of their rank, and this behavior is distinct from that of a uniform distribution. In this work, we give an improved version (named CDF-Zipf) of the PDF-Zipf model in [36], and show that most of the above-mentioned issues can be well addressed. Specifically, we show how the attacker's advantages in password-based protocols (e.g., [1,5,25,40]) can be 2–4 orders of magnitude more accurately captured, predict what fraction of users will be annoyed under the popularity-based policy [34] when given a specific threshold, and reveal an important property for the well-known α-guesswork in [8].

Our contributions. The key contributions of this work are as follows:

(1) First, we propose to use the formulation $C' \cdot Q(k)^{s'}$ to capture an attacker's advantages in making at most $Q(k)$ on-line guesses against password-based cryptographic protocols, superseding the traditional ones (i.e., $Q(k)/|\mathcal{D}|$ [1,25] and $Q(k)/2^m$ [2,27]), where k is the system security parameter, $\mathcal{D}$ is the password space, C' and s' are the CDF-Zipf regression parameters of $\mathcal{D}$, and m denotes the min-entropy of $\mathcal{D}$. Experiments on 9 large-scale password datasets show the superiority of our new formulation over existing ones. Given a target system, the values of C' and s' can be approximated by leaked datasets from sites with a similar service, language (and policy). For instance, if the protocol is to be deployed in a Chinese e-commerce site, one can set $C' = 0.019429$ and $s' = 0.211921$, which come from the leaked Dodonew site.

(2) Second, based on the Zipf assumption of passwords, we propose a series of prediction models to facilitate the choices of parameters for the promising

popularity-based password creation policy in [34]. Our models provide new insights and highlight that, usability will be largely impaired if the threshold parameter $\mathcal{T}$ is improperly chosen. For instance, when setting $\mathcal{T} = 1/10^6$ (which is widely recommended [17,34]) for Internet-scale sites, our model predicts that an average of 38.73 % of users will be potentially annoyed. Our theory well accords with the extensive experiments.

(3) Third, we, for the first time, reveal that the widely used password strength metric α-guesswork [8], which was believed to be parametric, is actually non-parametric in two of four cases under the Zipf assumption of passwords. As passwords are generally Zipf-distributed, this result makes α-guesswork much simpler to use — now we only need a single value of the advantage α instead of "all values of α" [8] to inform decisions.

2 Preliminaries

In this section, we first describe the nine datasets used. Then, we briefly review the Zipf model [36], and finally present an improved fitting methodology.

2.1 Descriptions of Real-World Password Datasets

We employ nine large-scale password datasets, a total of 111.94 million real-world passwords, to enable a comprehensive evaluation of the revealed implications. The basic info about these datasets is summarized in Table 1. Some of them have been widely used in password studies [29,35,36]. They were somehow breached by hackers or leaked by insiders, and then publicly disclosed over the Internet. Most of these breaches have been confirmed by the victim sites [31,32]. Our datasets range from low-value gaming and programmer forums, moderate-value social networks, to relatively sensitive email, e-commerce and web hosting service. They have a time span of 10 years, and come from four countries located in three distant continents. They are in four different languages, including two most-spoken ones (i.e., English and Chinese) in the world. To the best of knowledge, our corpus is amongst the largest and most diversified ones ever used for password-related studies. We refer the readers to [29,35,36] for a grasp of user-chosen passwords and learn what might impact users' password choices: language, service type, password policy, culture, faith among others.

2.2 Review of the PDF-Zipf Model

It has long been an open problem as to what is the distribution that user-chosen passwords follow. It is well-known that user-spoken words follow a Zipf's law, whether user-chosen passwords follow the same law? In 2012, Bonnneau [8] and Malone-Maher [30] separately studied this question, and both works plot the probability density function (PDF) of password datasets with the x-axis variable being the rank r of passwords and y-axis variable being the frequency f_r of the

Table 1. Basic info about the nine real-world password datasets

Dataset	Web service	Location	Language	When leaked	Unique PWs	$TotalPWs$
Rockyou	Social forum	USA	English	Dec. 2009	14,326,970	$32,581,870$
000webhost	Web hosting	USA	English	**Oct. 2015**	10,583,709	$15,251,073$
Battlefield	Gaming site	USA	English	June 2011	417,453	$542,386$
Tianya	Social forum	China	Chinese	Dec. 2011	12,898,437	$30,901,241$
Dodonew	Game&Ecommerce	China	Chinese	Dec. 2011	10,135,260	$16,258,891$
CSDN	Programmer forum	China	Chinese	Dec. 2011	4,037,605	$6,428,277$
Mail.ru	Email	Russia	Russian	Sep. 2014	2,954,907	$4,932,688$
Gmail.ru	Email	Russia	Russian	Sep. 2014	3,132,028	$4,929,090$
Flirtlife.de	Dating site	Germany	German	**May 2006**	115,589	$343,064$

password with rank r. They use a Zipf model to approximate the PDF graphs, yet the Kolmogorov-Smirnov (KS) tests suggest a negative answer.

In 2014, Wang et al. [36] made a further attempt. Different from these two studies [8,30] that fit all passwords in a dataset into the Zipf model, the work in [36] first eliminates the least frequent (LF) passwords (e.g., $LF < 4$) from datasets, and then use a Zipf model to approximate the PDF graphs of the remaining passwords. More specifically, Wang et al. found:

$$f_r = \frac{C}{r^s}, \tag{1}$$

where C and s are constants depending on the password dataset and can be calculated using methods like least squares or maximum likelihood estimation (MLE). We denote this Zipf model as the PDF-Zipf model. The KS tests (with samples of size 500K) accept most of the PDF-Zipf fittings. Equation 1 is better illustrated to show the nature of a Zipf's law in a log-log plot (see the green dotted lines in Figs. 2(a) and (b)), where $\log(f_r)$ is linear with $\log(r)$:

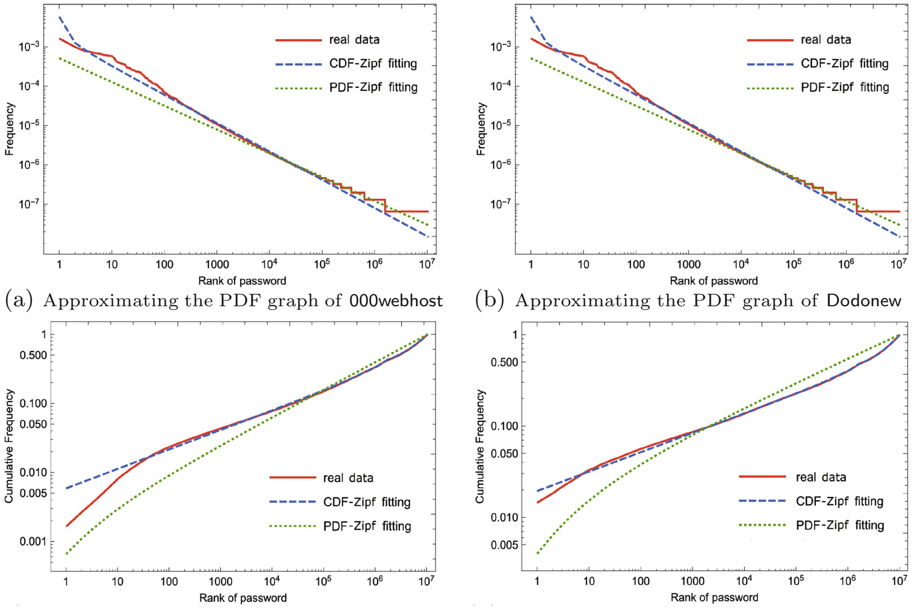
$$\log f_r = \log C - s \cdot \log r. \tag{2}$$

This means that, on a log-log plot the PDF regression line will be a straight line.

2.3 Our CDF-Zipf Model

As shown in Figs. 2(a) and (b) and also in [36], one undesirable feature of the PDF-Zipf model is that, it can not well capture the distribution of the first few most popular passwords (e.g., passwords with rank less than 1000).[1] We have tried various means to adjust the PDF-Zipf parameters to accommodate these most popular passwords, yet we are always caught in a dilemma: if they are well captured, the overall fitting cannot be accepted by KS tests; if they are not considered, the overall fitting will be acceptable.

[1] Note that the least frequent passwords are inherently difficult to be captured by a theoretic model due to the law of large numbers, and see more discussions in [36].



(a) Approximating the PDF graph of 000webhost (b) Approximating the PDF graph of Dodonew

(c) Approximating the CDF graph of 000webhost (d) Approximating the CDF graph of Dodonew

Fig. 2. A grasp of effectiveness of the PDF-Zipf approach and CDF-Zipf approach (Color figure online).

Essentially, the KS test quantifies the distance between the cumulative distribution function (CDF) $F_n(x)$ of an empirical distribution and the CDF $F(x)$ of the theoretic distribution (e.g., obtained by fitting):

$$D = \sup_x |F_n(x) - F(x)|,$$

where n is the sample size and $\sup_x$ is the supremum of the set of distances. This statistic $D \in [0, 1]$ is essentially the max gap between the two CDF curves $F_n(x)$ and $F(x)$, the smaller the better. It is used to conduct KS tests (see [36]).

As the PDF and the CDF of a distribution can be converted to each other, whether we can directly model the CDF of a password distribution? Interestingly, we find the Zipf model well fits to the CDF graphs of *entire* datasets (see the dashed blue lines in Figs. 2(c) and (d)). We call this model the CDF-Zipf model:

$$F_r = C' \cdot r^{s'}, \tag{3}$$

where F_r is the cumulative frequency of passwords up to rank r, C' and s' are constants depending on the password dataset and can be calculated by linear regression. $F_r(\cdot)$ is a step function, because $r = 1, 2, 3, \cdots$. Thus, we have

$$f_r = F_r - F_{r-1} = C' \cdot r^{s'} - C' \cdot (r-1)^{s'}. \tag{4}$$

Table 2. Comparison of our CDF-Zipf model with the PDF-Zipf model [36]. All the timings use second as the unit.

Dataset	PDF-Zipf model [36]				Our CDF-Zipf model			
	Timing	Statistic D[a]	C	s	Timing	Statistic D	C'	s'
Rockyou	23.41	0.193567	0.025464	0.913760	52917.74	**0.045874**	0.037433	0.187227
000webhost	13.23	0.111546	0.000512	0.603784	30236.77	0.006170	0.005858	0.281557
Battlefield	0.35	0.225527	0.003522	0.692898	973.79	0.010557	0.010298	0.294932
Tianya	22.95	0.161718	0.018684	0.895411	44702.48	0.022798	0.062239	0.155478
Dodonew	12.08	0.164640	0.002566	0.740560	29526.20	**0.004926**	0.019429	0.211921
CSDN	3.61	**0.268982**	0.008176	0.853028	10954.37	0.022319	0.058799	0.148573
Mail.ru	3.61	0.168754	0.006142	0.768912	8274.22	0.020773	0.025211	0.218212
Gmail	3.63	0.217463	0.007013	0.793667	8743.09	0.020543	0.020963	0.225653
Flirtlife.de	0.13	**0.062585**	0.016824	0.745634	159.26	0.036448	0.034577	0.291596

[a] The statistic Ds obtained by the CDF-Zipf model are always smaller than the PDF-Zipf model, indicating the former is better. Hereafter we only use parameters fitted from the CDF-Zipf model.

Note that, f_r can be approximated by using the derivative of F_r when seeing F_r as a continuous function: $f_r \approx d(F_r)/dr = C' \cdot s' \cdot r^{s'-1}$, implying a Zipf's law.

We fit the CDF-Zipf model to our nine datasets (see Fig. 3), and always obtain better fittings than the PDF-Zipf model in terms of the KS statistic D (i.e., the max gap between the CDF curves of a fitted model and the real data). Our CDF-Zipf parameters are calculated by linear regression using the well-known golden-section-search method on an Intel i7-4790 3.60 GHz PC. As summarized in Table 2, the *largest* D from fittings under our CDF-Zipf model is *smaller* than the smallest D of the PDF-Zipf model (set `least frequency` $=4$). This means that the max CDF gap under the CDF-Zipf model is always smaller than those of the PDF-Zipf model. This suggests the superiority of our CDF-Zipf model.

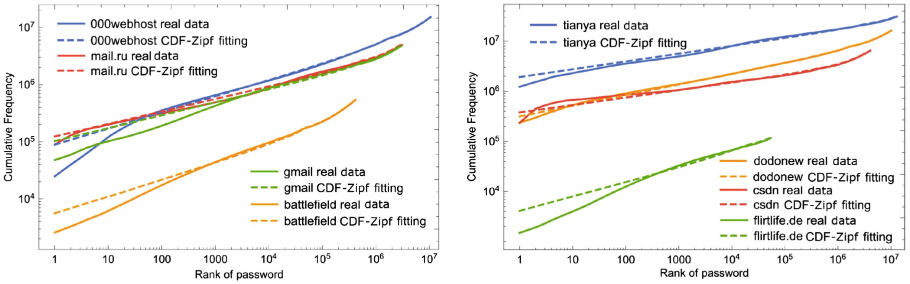
Table 3 shows that our CDF-Zipf model is stable: the parameters fitted from subsets of a dataset remain largely the same with the parameters fitted from the entire dataset. For instance, the parameters (i.e., $C' = 0.019440$ and $s' = 0.211843$) fitted from *1/4* of Dodonew are almost the same with those (i.e., $C' = 0.019429$ and $s' = 0.211921$, see Table 2) fitted from the *entire* Dodonew. All the Max-CDF-gap*s* are <0.035 (avg. 0.015). However, as shown in [36], this feature does not hold in the PDF-Zipf model.

Table 3. The CDF-Zipf model is stable.

Dataset	CDF-Zipf C'	CDF-Zipf s'	Max-CDF-gap[a]
1/4 Rockyou	0.031065	0.205094	0.034697
1/4 000webhost	0.005407	0.287458	0.003948
1/4 Battlefield	0.008033	0.323953	0.007699
1/4 Tianya	0.056322	0.164992	0.019645
1/4 Dodonew	0.019440	0.211843	0.004901
1/4 CSDN	0.059822	0.142107	0.023216
1/4 Mail.ru	0.019689	0.240814	0.011068
1/4 Gmail	0.016879	0.247743	0.013908
1/4 Flirtlife.de	0.026715	0.327783	0.023809

[a] "Max-CDF-gap" measures the largest distance between the CDF curve of each *entire* dataset and that of the CDF-Zipf model fitted with *1/4* dataset.

Summary. Our CDF-Zipf model is superior to the PDF-Zipf model [36] in four-fold: (1) its fitted parameters more accurately approximate the real distribution (data); (2) it does *not* need to eliminate the unpopular passwords (e.g., with $f_r < 4$) when performing a Zipf fitting; (3) the most popular passwords can be



(a) Zipf's law in PWs of English and Russian. (b) Zipf's law in PWs of Chinese and German.

Fig. 3. Zipf's law in nine real-life password datasets from four different populations, using our CDF-Zipf fitting approach. For detailed CDF-Zipf parameters, see Table 2.

well captured; and (4) it is stable. Note that, our CDF-Zipf model achieves these superiority at the cost of about 3000–4000 times higher fitting timings. Still, all our CDF-Zipf fittings can be completed in one day on a common PC.

3 Implication for Password-Based Cryptographic Protocols

In this section, we mainly use the most common password-based cryptographic protocol, i.e. password-based authentication, as a case study to show the implication, and then show its generality to other kinds of password-based protocols.

3.1 Implication for Password-Based Authentication Protocols

It is expected that, the most foundational implication of the discovery of Zipf's law in passwords is for *hundreds* of existing provably secure authentication protocols that involve passwords. According to whether additional authentication factors are involved, password authentication protocols can be classified into password-based single-factor schemes (e.g., two-party [25] and multi-party [15]) and password-based multi-factor schemes (e.g., two-factor [38] and three-factor [20]). Here we first show the implication for password-based single-factor schemes (also called PAKE protocols) and then for multi-factor schemes.

Uniform-based security formulation. In most of the provably secure PAKE protocols (e.g., [1,3,15,24] in the random oracle model and [25,26,40] in the standard model), it is typically assumed that "password pw_U (for each client U) is chosen independently and *uniformly at random* from a dictionary $\mathcal{D}$ of size $|\mathcal{D}|$, where $|\mathcal{D}|$ is a fixed constant independent of the security parameter k" [25], then a security model is described, and finally a "standard" definition of security as the one in [25] is given:

> "$\cdots$ $\cdots$ Protocol $\mathcal{P}$ is a secure protocol for password-only authenticated key-exchange if, for all [password] dictionary sizes $|\mathcal{D}|$ and for all

ppt[probabilistic polynomial time] adversaries $\mathcal{A}$ making at most $Q(k)$ on-line attacks, there exists a negligible function $\epsilon(\cdot)$ such that:

$$\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k) \leq Q(k)/|\mathcal{D}| + \epsilon(k), \tag{5}$$

where $\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k)$ is the advantage of $\mathcal{A}$ in attacking $\mathcal{P}$."

Generally, user-generated passwords offer about 20–21 bits [8] of actual security against an optimal offline dictionary attack, which means the effective password space $\mathcal{D}$ is of size about 2^{20}–2^{21}. This indicates that a system which employs a PAKE protocol achieving the security goal of Eq. 5 can assure that one online guessing attempt will attain a success rate no larger than $1/2^{20}$–$1/2^{21}$. This is not the case in reality, and actually it may convey an overly optimistic sense of security to common users and security engineers.

As shown in Table 4, within 10^3 online guesses, the real attacker's advantages against most of the real-world sites are *three to four orders of magnitude higher* than that of a uniform-modelled attacker. For instance, the actual advantages of the real attacker against the gaming&e-commerce site www.dodonew.com reach 1.45 % when $Q(k) = 3$, 3.28 % when $Q(k) = 10$ and 5.60 % when $Q(k) = 100$. They are far beyond the theoretic results (see the last row in Table 4) given by Eq. 5.

Table 4. The cumulative percentages of top-x most popular passwords of each real-life password dataset ("Uni. dist." stands for uniform distribution).

Datasets	Top 1	Top 3	Top 10	Top 10^2	Top 10^3	Top 10^4	Top $\frac{1}{10}$	Top $\frac{1}{10^2}$	Top $\frac{1}{10^3}$	Top $\frac{1}{10^4}$
Tianya	3.98 %	5.59 %	7.43 %	11.50 %	16.04 %	25.78 %	58.21 %	41.19 %	27.45 %	16.70 %
Dodonew	1.45 %	2.15 %	3.28 %	5.60 %	8.59 %	13.62 %	39.97 %	22.72 %	13.66 %	8.61 %
CSDN	3.66 %	8.15 %	10.44 %	13.26 %	16.54 %	23.91 %	42.66 %	28.46 %	20.62 %	14.97 %
Rockyou	0.89 %	1.37 %	2.05 %	4.55 %	11.30 %	22.31 %	57.28 %	39.30 %	24.24 %	12.84 %
000webhost	0.16 %	0.34 %	0.79 %	2.32 %	4.30 %	7.71 %	34.09 %	15.17 %	7.83 %	4.36 %
Battlefield	0.48 %	0.71 %	1.14 %	3.21 %	8.13 %	17.91 %	30.57 %	13.49 %	5.78 %	2.16 %
Mail.ru	1.82 %	3.06 %	4.05 %	6.37 %	9.94 %	17.40 %	43.88 %	24.92 %	12.46 %	7.81 %
Gmail.ru	0.97 %	1.43 %	2.08 %	3.88 %	8.66 %	17.77 %	41.65 %	23.79 %	12.63 %	5.76 %
Flirtlife.de	1.30 %	2.00 %	3.47 %	10.83 %	28.51 %	58.01 %	48.73 %	22.52 %	7.92 %	2.55 %
Avg. above	**1.63 %**	**2.76 %**	**3.86 %**	**6.84 %**	**12.45 %**	22.71 %	44.11 %	25.73 %	14.73 %	8.42 %
Uni. dist	0.01‱	0.03‱	0.1‱	1‱	0.10 %	1.00 %	10.00 %	1.00 %	0.10 %	0.01 %

As a prudent side note, some PAKE studies (e.g., [25,26]) complement that the assumption of a uniform distribution of passwords with a constant-size dictionary is made for simplicity only, and their security proofs can be extended to handle more complex cases where passwords do not distribute uniformly, different distributions exist for different clients, or the password dictionary size depends on the security parameter. However, such a complement only serves to obscure their security statements and undermine the readers' understanding of *to exactly what extent they can have confidence in the authentication protocol used to protect systems, because no one knows what the security guarantees would be if "user-chosen passwords do not distribute uniformly"*. This defeats the purpose of constructing provably secure protocols which "explicitly capture the inherently *quantitative* nature of security, via a concrete or exact treatment of security" and "offer *quantitative* security guarantee" [4] in the first place.

Our Zipf-based security formulations. According to the Zipf theory, now it is fundamentally unnecessary to make the uniform assumption of password distribution. Since system-assigned random passwords are hardly usable [39], most services allow users to generate their own passwords. This would generally lead to the passwords complying with the Zipf's law as we have shown in Sect. 2.3. Therefore, it is more desirable to make the Zipf assumption about password distributions. Under the PDF-Zipf model in [36], it is natural to reach that:

$$\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k) = \frac{C/1^s}{\sum_{i=1}^{|\mathcal{D}|}\frac{C}{i^s}} + \frac{C/2^s}{\sum_{i=1}^{|\mathcal{D}|}\frac{C}{i^s}} + \cdots + \frac{C/Q(k)^s}{\sum_{i=1}^{|\mathcal{D}|}\frac{C}{i^s}} = \frac{\sum_{j=1}^{Q(k)}\frac{1}{j^s}}{\sum_{i=1}^{|\mathcal{D}|}\frac{1}{i^s}} + \epsilon(k), \quad (6)$$

Under our CDF-Zipf model (see Eq. 3), it is natural to reach that:

$$\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k) = C' \cdot Q(k)^{s'} + \epsilon(k), \quad (7)$$

where the parameters C, C', s and s' are referred to Eqs. 1 and 3 in Sect. 2.3.

Figure 4 shows that $\mathcal{A}$'s advantage is more accurately captured by Eq. 7 than by Eq. 6. This is expected according to the results in Sect. 2.3.

Our popularity-policy-based formulation. Figure 4 (as well as Figs. 2(c) and (d)) shows that, an attacker who only tries a rather small number (e.g., $Q(k) = 100$) of the most popular passwords can crack a non-negligible proportion of user accounts. In other words, even if the authentication protocol implemented is provably secure, secure user identification still cannot be reached if the passwords of the system obey Zipf's law. Countermeasures like the popularity-based password policy [34] can be taken. In this case, the skewed Zipf distribution seems hardly possible to be mathematically characterized, we are stuck in a conundrum to formulate $\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k)$. Inspired by the essential notion of security that a secure PAKE protocol can provide – only online impersonation attacks are helpful to the adversary in breaking the security of the protocol [2,25], we manage to get out of the problem by giving up the idea of firstly characterizing the exact distribution of passwords and then formulating the definition of security. And instead, whenever a policy like [34] is in place, we provide a tight upper bound for the adversary's advantage. More specifically, Eq. 5 now is amended as follows:

$$\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k) \leq F_1 \cdot Q(k)/|\mathcal{DS}| + \epsilon(k), \quad (8)$$

where F_1, as said earlier, is the frequency of the most popular password in the dataset $\mathcal{DS}$, $|\mathcal{DS}|$ is the (expected) number of user accounts of the target authentication system, and the other notations are the same with those of Eq. 5. Note that, dictionary $\mathcal{D}$ is *the password sample space* and it is a *set*, while dataset $\mathcal{DS}$ is *a (specific) password sample* and it is a *multiset*. Therefore, the value of $F_1/|\mathcal{DS}|$ is exactly the threshold probability $\mathcal{T}$ (e.g., $\mathcal{T} = 1/16384$) that the underlying password policy (see [34]) maintains. For a system to reach a Level 1 certification [10], the success chance of an online guessing attacker should be no larger than 1 in 1024, which indicates $F_1/|\mathcal{DS}| \leq 1/1024$; Similarly, for a Level 2 certification, the system shall ensure $F_1/|\mathcal{DS}| \leq 1/16384$. For example, for the gaming and e-commerce website www.dodonew.com to achieve a Level 2 security, F_1 should have been no larger than $991(\approx 16231271/16384)$. Also note that, Eq. 5 is actually a special case of Eq. 8, where $F_1 = 1$ and $|\mathcal{DS}| = |\mathcal{D}|$.

Min-entropy-based security formulation. In 2015, Abdalla et al. [2] proposed a provably secure PAKE protocol that does not employ the traditional security formulation like Eq. 5, but uses a different one:

$$\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k) \leq Q(k)/2^m + \epsilon(k), \tag{9}$$

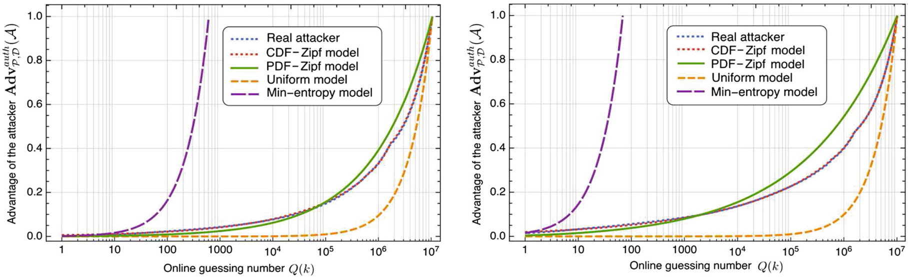
where m is the *min-entropy* [8] of a password dataset.[2] Actually, it is not difficult to see that Abdalla et al.'s this formulation is in essential the same with Eq. 8, because one can derive that $m = -\log_2(F_1/|\mathcal{DS}|)$. However, no rationale or justification for preferring Eq. 9 rather than Eq. 5 has been given in [2]. In comparison, our formulation Eq. 8 is more concrete and easily understood than Eq. 9 from the prospective of password policy.

In addition, as with our Eq. 8, Abdalla et al.'s Eq. 9 (i.e., the min-entropy model) is *only* effective when a popularity-based password policy like [34] is in place, resulting in that the password distribution does not follow the Zipf's law. However, without such a policy in place, passwords are likely to follow the Zipf's law, and thus both Eqs. 8 and 9 will be useless. This has not been pointed out in [2]. Recent studies [12,37] and our exploration of 120 top sites show that, such a policy has not been adopted into leading web services or password managers.

Also note that, if m is defined to be the *entropy* of passwords, then Eq. 9 is virtually equal to Eq. 5 and it provides a *mean* value for the online guessing difficulty, for one can derive that $m = \sum_{r=1}^{|\mathcal{D}|} -p_i \log_2 p_i$, where p_i is the probability of the i^{th} most frequent password in $\mathcal{D}$ (e.g., $p_1 = F_1/|\mathcal{DS}|$). This well explicates why Benhamouda et al. (see Sect. 6.1 of [6]) stated that "equivalently the advantage of any adversary can be bounded" by either Eqs. 5 or 9. However, as we have shown, if m is defined as the *min-entropy* of passwords, Eqs. 5 and 9 (or equally, Eq. 8) will be significantly different from each other.

Comparison and summary. We show in Fig. 4 how the existing two PAKE security formulations (i.e., Eqs. 5 and 9) and the two Zipf-based formulations (i.e., Eqs. 6 and 7) approximate the real attacker (using 000webhost and Dodonew for example). Since online guessing attacks are generally prevented by lockout, rate-limiting techniques or suspicious login detection [16], the attackers cannot make a large number of login attempts, and thus the guess number is often small (generally, $Q(k) \leq 10^4$). One can see that, *our CDF-Zipf*

[2] We note that, in Sects. 5.2–5.4 of [2], m is re-defined to be the *entropy* of passwords. This inconsistence would lead to great differences in security guarantees. We conjecture typos have occurred there.



(a) Online guessing the 000webhost service

(b) Online guessing the Dodonew service

Fig. 4. With $Q(k)$ online guessing attempts, the advantages of the real attacker, the uniform-modeled attacker [1,24,26], min-entropy-modeled attacker [2,27], PDF-Zipf-modeled attacker [36] and our CDF-Zipf-modeled attacker. Our model almost *overlaps* with the real attacker.

model well approximates the real attacker — its advantage curve almost overlaps with that of the real attacker, substantially outperforming the three other models. For instance, the actual advantages of $\mathcal{A}$ against Dodonew reach 5.60 % when $Q(k) = 10^2$ and 8.59 % when $Q(k) = 10^3$. They are far beyond 0.00098 % and 0.0098 % given by Eq. 5 [1,24,26], and far less than 100 % and 100 % given by Eq. 9 [2,27], respectively. Fortunately, our CDF-Zipf model (i.e., Eq. 7) predicts a 5.15 % when $Q(k) = 10^2$ and a 8.40 % when $Q(k) = 10^3$, respectively. In all, CDF-Zipf model performs the best and yields results well accord with the real attacker's guessing advantages.

3.2 Implication for Multi-factor Authentication Protocols

Without loss of generality, here we use the most widely used smart-card-based password authentication protocols as an example. The major goal of designing two-factor schemes is to achieve "truly two-factor security" [38] which means that only an entity who knows *both* the password and the smart card can login to the server. This means an attacker who knows either the password factor or the smart card factor shall be unable to login.

The crux of designing a protocol with "truly two-factor security" lies in how to resist offline password guessing attack, in case the smart card has been stolen and extracted by the attacker [38]. This means now the protocol security only relies on the password. The attacker $\mathcal{A}$ can use the stolen smart card and tries to login with the guessed passwords $pw_1, pw_2, \cdots$, until the server locks out the account. Since such an online guessing is always unavoidable, the best security that a two-factor protocol $\mathcal{P}$ can achieve is to ensure that: such an online guessing attack is the best that $\mathcal{A}$ can do. Accordingly, protocol $\mathcal{P}$ is said secure only if

$$\mathrm{Adv}^{\mathrm{2fa}}_{\mathcal{A},\mathcal{P}}(k) \leq Q(k)/|\mathcal{D}| + \epsilon(k), \tag{10}$$

where $\mathrm{Adv}^{\mathrm{2fa}}_{\mathcal{A},\mathcal{P}}(k)$ is $\mathcal{A}$'s advantage in attacking $\mathcal{P}$ with $Q(k)$ online guesses.

Essentially the same security formulation like Eq. 10 are made in most of the provably secure two-factor schemes (e.g., see Sect. 2.2.1 of [11], Definition 1 of [24]). As discussed in Sect. 3.1, our Zipf theory invalidates such, at best unrealistic and at worst misleading (i.e., convey a false sense of security guarantees), forms of formulation. A formulation like our proposed Eq. 7 is much more accurate and desirable (see Fig. 4):

$$\mathrm{Adv}^{\mathrm{2fa}}_{\mathcal{A},\mathcal{P}}(k) = C' \cdot Q(k)^{s'} + \epsilon(k), \tag{11}$$

where $\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k)$ is the advantage of $\mathcal{A}$ in attacking $\mathcal{P}$ with $Q(k)$ online guesses, and C' and s' are the parameters calculated using the CDF-Zipf model.

3.3 Implication for Other Kinds of Password-Based Protocols

Without loss of generality, here we mainly use typical examples to show the applicability of our formulation Eq. 7 to password-protected secret sharing [23], password-based signatures [18] and password-based encryption [5].

In 2016, Jarecki et al. [23] proposed an efficient and composable password-protected secret sharing (PPSS) protocol. In Definition 2 of [23], it is assumed that "$pw \leftarrow_R \mathcal{D}$", which means password pw are drawn uniformly at random from password space $\mathcal{D}$. Further, they defined the protocol security to be $\mathrm{Adv}^{\mathrm{ppss}}_{\mathcal{A}}(k) = (q_S + q_U)/|\mathcal{D}| + \epsilon$. According to the CDF-Zipf theory, a formulation like our proposed Eq. 7 is much more accurate and desirable:

$$\mathrm{Adv}^{\mathrm{ppss}}_{\mathcal{A}}(k) = C' \cdot Q(k)^{s'} + \epsilon(k), \tag{12}$$

Similarly, for password-based signatures (PBS, e.g. [18]), when defining "Blindness" it generally involves an explicit assumption of the distribution of passwords. Most of related works assume "$pw \leftarrow_R \texttt{PW}$" [18]. For password-based encryption (PBE) schemes, most related works assume "$pw \leftarrow_{\$} A_1(\lambda)$" (see Fig. 7 of [5]) which means passwords are drawn uniformly. All such password- related protocols can be readily designed with the Zipf assumption of password distribution, and give more realistic security formulations like Eq. 7.

Summary. To the best of our knowledge, we, for the first time, pay attention to the joint between passwords and password-based authentication protocols. With the knowledge of the exact distribution of passwords, we manage to develop a more accurate and realistic formulation (i.e., $\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k) = C' \cdot Q(k)^{s'} + \epsilon(k)$) to characterize the formal security results for password-based authentication protocols. Given a target system, the values of C' and s' can be predicted (approximated) by leaked datasets from sites with a similar service, language and policy. For instance, if the protocol is to be deployed in a English gaming site, one can set $C' = 0.010298$ and $s' = 0.294932$, which come from the leaked Battlefiled site. As a rule of thumb, high-value sites can prefer C' and s' from Dodonew/000webhost; medium-value sites prefer those of Gmail/Battlefield; low-value sites prefer those of Tianya/Rockyou. *This enables an accurate, quantitative and practical assessment of the security provisions of a password system about which we have no password data.*

Here we have mainly taken password-based authentication as a case study. As we have sketched in Sect. 3.3, the results revealed herein can also be readily applied to other kinds of password-based cryptographic protocols whose security formulation essentially relies on the *explicit* assumption of the distribution of user-chosen passwords, such as PBE [5], PBS [18] and PPSS [3,23].

4 Implications for Password Creation Policies

Recently, it has been popular (e.g., [17,34]) to advocate a password policy that disallows users from choosing undesirably popular passwords (e.g., `123456` and `letmein`) that are more frequently chosen than a predefined threshold $\mathcal{T}$ (e.g., $\mathcal{T} = 1/10^6$). The motivation underlying such a policy is that some users prefer dangerously popular passwords, and as shown in Eq. 7 of Sect. 3.2, such passwords would make $\mathcal{A}$'s advantage $\text{Adv}_{\mathcal{A},\mathcal{P}}(k) = C' \cdot Q(k)^{s'} + \epsilon(k)$ extremely high with even a small guess number $Q(k)$.

However, under the PDF-Zipf model, Wang et al. [36] suggested a number of prediction models, and pointed out that this popularity-based policy would largely impair usability. For example, given a threshold $\mathcal{T} = 1/10^6$, 60 % of users in most Internet-scale sites will be *potentially* annoyed to abandon their old, popular password and select a new one. As such theoretical predictions is important, and now a natural question arise: how to obtain more accurate prediction models under our CDF-Zipf model?

4.1 Our Prediction Models

In Sect. 2.3, our CDF-Zipf model reveals that in reality, users choose passwords far from uniform and the CDF of passwords follows a Zipfian distribution. More specifically, this means that *the rank r of a password* and *the cumulative frequency F_r of passwords up to r* obey the equation $F_r = \frac{C'}{r^{s'}}$, where C' and s' are constants calculated from the CDF-Zipf regression. In other words,

$$P(rank \leq r) = C' \cdot r^{s'}, \tag{13}$$

Generally, there is finite number (say N) of distinct passwords. Thus, we have:

$$P(rank \leq N) = C' \cdot N^{s'} = 1 \quad \Rightarrow \quad N = (\frac{1}{C'})^{1/s'}. \tag{14}$$

Consequently, the number of these top η (e.g., $\eta = 1\,\%$) of passwords is $\eta \cdot N$. Then, the cumulative frequency of these top $\eta \cdot N$ passwords will be:

$$P(rank \leq \eta \cdot N) = C' \cdot (\eta \cdot N)^{s'} = C' \cdot (\eta \cdot (1/C')^{1/s'})^{s'} = C' \cdot \eta^{s'} \cdot (1/C') = \eta^{s'}. \tag{15}$$

This indicates that $W_p(\eta) = P(rank \leq \eta \cdot N) = \eta^{s'}$ of passwords will be potentially affected. For better illustration, assume the frequency f_r of a password with rank r is a continuous real variable. Thus, the frequency of a password with rank r is

$$f_r = \frac{d(F_r)}{dr} = \frac{d(C' \cdot r^{s'})}{dr} = C' \cdot s' \cdot r^{s'-1}. \tag{16}$$

Now we can obtain the relationship between the top fraction η and the popularity threshold $\mathcal{T}$. Assume the rank of the password with frequency equals $\mathcal{T}$ to be $r_{\mathcal{T}}$. On the one hand, Eq. 16 suggests that: $\mathcal{T} = C' \cdot s' \cdot r_{\mathcal{T}}^{s'-1}$; On the other hand, $\eta = r_{\mathcal{T}}/N$. Further according to Eq. 14, we get

$$\eta = \frac{r_{\mathcal{T}}}{N} = (\frac{\mathcal{T}}{C' \cdot s'})^{\frac{1}{s'-1}} \cdot \frac{1}{N} = (\frac{\mathcal{T}}{C' \cdot s'})^{\frac{1}{s'-1}} \cdot (C')^{\frac{1}{s'}}. \tag{17}$$

Equation 17 suggests that the frequency of a password with rank $\eta \cdot N$ will be $C' \cdot s' \cdot (\eta \cdot N)^{s'-1}$. Therefore, among these $W_p(\eta)$ of passwords that are *potentially* affected, the fraction of users that *actually* will *not* be affected is $C' \cdot s' \cdot (\eta \cdot N)^{s'-1} \cdot \eta \cdot N = s' \cdot \eta^{s'}$. Thus, the fraction will be actually affected is:

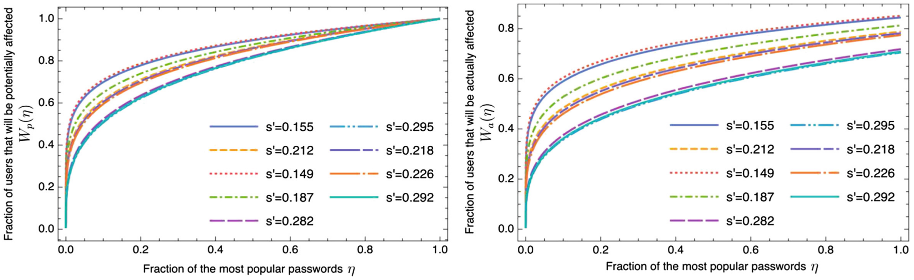
$$W_a(\eta) = W_p(\eta) - s' \cdot \eta^{s'} = (1 - s') \cdot \eta^{s'}. \tag{18}$$

There is a subtlety to be noted. $W_p(\eta)$ and $W_a(\eta)$ are indeed two independent and useful indicators to measure the extent to which usability will be affected. For instance, suppose an Internet-scale English social-network site www.example.com wants to enforce a popularity-based policy with $\mathcal{T} = 1/10^6$, then we can predict (by using CDF-Zipf parameters of Rockyou) that there will be $W_p(\eta) =$ 36.08 % accounts with passwords more popular than $\mathcal{T} = 1/10^6$. This means each of these 36.08 % accounts has *an equal potential* to be required to change a new password. However, there will only be $W_a(\eta) =$ 29.32 % accounts that are *actually* required to choose a different password for the reason that, after $W_a(\eta) =$ 29.32 % accounts have already been changed, the remaining $W_p(\eta) - W_a(\eta) =$ 6.76 % accounts will be with passwords less popular than $\mathcal{T} = 1/10^6$ and comply with the policy.

4.2 Our Empirical Results

In Fig. 5 we depict the form of the curves of $W_p(\eta)$ and $W_a(\eta)$ against η for various values of s' as listed in Table 2. The rapid increase of W_p and W_a at the top-10 % of their curves (see Fig. 5) clearly reveals that, a significant fraction of users will be annoyed despite that only a marginal fraction of popular passwords are prohibited. Since $W_p(\eta) = \eta^{s'}$, an average of $W_a =$ 38.73 % (Max = 51.72 %, Min = 19.46 %) of users in the 7 million-size sites (see Table 1) will be potentially inconvenienced when $\mathcal{T} = 1/10^6$. To see whether our theory accords with the reality, we also summarize the statistical results from 9 real-life password datasets in Table 5. Generally, when $\mathcal{T} < 1/16384$ the theoretical W_a is *lower* than the empirical W_a by a factor < 1, much smaller than that of [36]. The means that our predictions would serve as conservative indicators of usability degradations.

Summary. Our above prediction models (i.e., η, $W_p(\eta)$ and $W_a(\eta)$) indicate that $\mathcal{T} = 1/10^6$ might be too restrictive for Internet-scale sites in terms of usability. In contrast, less than 17 % of users in most systems will be potentially annoyed



(a) Users *potentially* affected with top fraction η (b) Users *actually* affected with top fraction η

Fig. 5. The fraction of users that will be potentially/actually affected by a popularity-based policy, when passwords follow a Zipf law with s' as listed in Table 2.

Table 5. Effects of policy threshold $\mathcal{T}$ on the proportion (i.e., η) of undesirable popular passwords and on the proportion (i.e., W_a) of users that will be actually annoyed.

Password dataset	$\mathcal{T} = 1/1024$		$\mathcal{T} = 1/10000$		$\mathcal{T} = 1/16384$		$\mathcal{T} = 1/1000000$	
	η	W_a	η	W_a	η	W_a	η	W_a
Rockyou	0.0000 %	1.22 %	0.0020 %	4.13 %	0.0040 %	5.70 %	0.4863 %	27.30 %
000webhost	0.0000 %	0.07 %	0.0007 %	1.36 %	0.0106 %	3.31 %	0.2668 %	7.45 %
Battlefield	0.0007 %	0.42 %	0.0393 %	2.34 %	1.0698 %	9.69 %	100.0000 %	100.00 %
Tianya	0.0001 %	6.61 %	0.0014 %	10.76 %	0.0022 %	11.64 %	0.4394 %	30.92 %
Dodonew	0.0001 %	2.30 %	0.0011 %	4.61 %	0.0020 %	5.19 %	0.3962 %	14.71 %
CSDN	0.0002 %	9.46 %	0.0029 %	12.28 %	0.0049 %	12.87 %	0.8441 %	24.69 %
Mail.ru	0.0003 %	3.08 %	0.0043 %	5.40 %	0.0879 %	9.54 %	2.8439 %	26.26 %
Gmail.ru	0.0002 %	1.24 %	0.0049 %	2.94 %	0.1339 %	9.66 %	2.4732 %	23.34 %
Flirtlife.de	0.0594 %	3.02 %	1.9939 %	19.06 %	3.3126 %	24.19 %	100.0000 %	100.00 %

when we set $\mathcal{T} = 1/16384$ which complies with a Level 2 certification [10]. We suggest that $\mathcal{T} = 1/16384$ would be more acceptable for most Internet-scale sites, especially those care for user experience, such as e-commerce and gaming.

5 Implications for the α-guesswork Metric

Now we show that the password strength metric α-guesswork [8], which previously was deemed as fully parametric and has been widely used (e.g., [14,22]), is actually non-parametric in two of four cases when measuring passwords. This inherent property would make α-guesswork much simpler to use.

5.1 Review of the α-guesswork Metric

For ease of understanding, here the notations follow [8]. $\mathcal{X}$ stands for the password distribution, and each password x_i is randomly drawn from $\mathcal{X}$ with a probability p_i, where $\sum p_i = 1$. Without loss of generality, suppose $p_1 \geq p_2 \geq \cdots \geq p_{\mathcal{N}}$, where $\mathcal{N}$ is the total number of possible individual passwords in $\mathcal{X}$.

Before we present α-guesswork $G_\alpha(\mathcal{X})$, two other statistic-based metrics (i.e., β-success-rate $\lambda_\beta(\mathcal{X})$ and α-work-factor $\mu_\alpha(\mathcal{X})$) are reviewed:

$$\lambda_\beta(\mathcal{X}) = \sum_{i=1}^{\beta} p_i, \tag{19}$$

which measures the expected advantages of $\mathcal{A}$ restricted to β guesses per account.

$$\mu_\alpha(\mathcal{X}) = min\{j| \sum_{i=1}^{j} p_i \geq \alpha\}, \tag{20}$$

where $0<\alpha\leq 1$. $\mu_\alpha(\mathcal{X})$ denotes the least number of fixed guesses per account when $\mathcal{A}$ aims to guess no less than a fraction α of total accounts. Therefore, λ_{μ_α} stands for $\mathcal{A}$'s actual advantages when given μ_α guesses per account and $\lambda_{\mu_\alpha} \geq \alpha$. With the above two definitions, α-guesswork is specified as:

$$G_\alpha(\mathcal{X}) = (1 - \lambda_{\mu_\alpha}) \cdot \mu_\alpha + \sum_{i=1}^{\mu_\alpha} p_i \cdot i, \tag{21}$$

The rationales behind Eq. 21 are referred to [8].

5.2 Defect in the α-guesswork Metric

The α-guesswork metric has been widely employed in recent studies (e.g., [14,22]), and the related paper also won the "NSA 2013 annual Best Scientific Cybersecurity Paper Award" [33]. However, it is subject to a defect — it is non-deterministic. More specifically, it is always parameterized on the success rate α (e.g., a relationship of $G_{0.51}(\mathcal{X}_A) > G_{0.51}(\mathcal{X}_B)$ can never guarantee that $G_{0.52}(\mathcal{X}_A) \geq G_{0.52}(\mathcal{X}_B)$), as admitted in [8] that "we can't rely on any single value of α, each value provides information about a fundamentally different attack scenario." Thus, for a fair comparison, entire curves (i.e., with α ranging from 0 to 1) of $G_\alpha(\mathcal{X}_A)$ and $G_\alpha(\mathcal{X}_B)$ have to be drawn. This makes it quite cumbersome to use. This defect is inherently due to the fact that α-guesswork does not employ the knowledge of the explicit distribution of passwords.

5.3 Our New Observations

Interestingly, we observe that, based on the Zipf assumption of passwords (which is generally the case in reality), G_α can be shown to be no longer parameterized in two of four cases. Note that, λ_β stands for the success rate by β guesses under the *optimal* attack. This means the curve of λ_β is essentially the same the CDF curve of distribution $\mathcal{X}$. The latter, as shown in Sect. 2.3 can be well approximated by our CDF-Zipf model. Therefore, we have

$$\lambda_\beta = \sum_{i=1}^{\beta} p_i \approx F_\beta = C' \cdot \beta^{s'}. \tag{22}$$

where β is the number of online guesses in an optimal order.

Theorem 1. *For two password distributions $\mathcal{X}_A$ and $\mathcal{X}_B$, suppose $C'_A \leq C'_B$, $s'_A \leq s'_B$. Then*

$$G_\alpha(\mathcal{X}_A) \geq G_\alpha(\mathcal{X}_B),$$

where $0 \leq \alpha \leq 1$. If either inequalities of the above two conditions is strict, then $G_\alpha(\mathcal{X}_A) > G_\alpha(\mathcal{X}_B)$, where $0 < \alpha \leq 1$.

Proof. Firstly, since $C'_A \leq C'_B$ and $s'_A \leq s'_B$, we can get that $\lambda_\beta(\mathcal{X}_A) \leq \lambda_\beta(\mathcal{X}_B)$:

$$\lambda_\beta(\mathcal{X}_A) = \sum_{i=1}^{\beta} p_i^A \approx F_\beta(\mathcal{X}_A) = C'_A \cdot \beta^{s'_A} \leq C'_B \cdot \beta^{s'_B} = F_\beta(\mathcal{X}_B) \approx \sum_{i=1}^{\beta} p_i^B = \lambda_\beta(\mathcal{X}_B), \quad (23)$$

where $1 \leq \beta \leq \max\{|\mathcal{X}_A|, |\mathcal{X}_B|\}$. Secondly, with Eq. 20 and $\lambda_\beta(\mathcal{X}_A) \leq \lambda_\beta(\mathcal{X}_B)$ (i.e., Eq. 23), it is natural to get $\mu_\alpha(\mathcal{X}_A) \geq \mu_\alpha(\mathcal{X}_A)$. Thirdly, we can derive

$$\begin{aligned} G_\alpha =& (1-\lambda_{\mu_\alpha}) \cdot \mu_\alpha + \sum_{i=1}^{\mu_\alpha} p_i \cdot i = \sum_{i=1}^{\mu_\alpha} \sum_{j=1}^{i} p_i + (1-\lambda_{\mu_\alpha}) \cdot \mu_\alpha \\ =& \sum_{j=1}^{\mu_\alpha} \sum_{i=j}^{\mu_\alpha} p_i + \sum_{j=1}^{\mu_\alpha} (1-\lambda_{\mu_\alpha}) = \sum_{j=1}^{\mu_\alpha} (1-\lambda_{\mu_\alpha} + \sum_{i=j}^{\mu_\alpha} p_i) \\ =& \sum_{j=1}^{\mu_\alpha} (1-\lambda_{j-1}). \end{aligned}$$

Since $\mu_\alpha(\mathcal{X}_A) \geq \mu_\alpha(\mathcal{X}_B)$ and $\lambda_j(\mathcal{X}_A) \leq \lambda_j(\mathcal{X}_B)$, we get

$$G_\alpha(\mathcal{X}_A) \geq G_\alpha(\mathcal{X}_B).$$

If either of the two conditions in Theorem 1 is strict, then it holds that $G_\alpha(\mathcal{X}_A) > G_\alpha(\mathcal{X}_B)$, where $0 < \alpha \leq 1$.

Corollary 1. *Suppose $C'_A \geq C'_B, s'_A \geq s'_B$. Then*

$$G_\alpha(\mathcal{X}_A) \leq G_\alpha(\mathcal{X}_B),$$

This corollary holds due to the evident fact that it is exactly the converse-negative proposition of Theorem 1.

The above theorem and corollary indicate that, given two password datasets A and B, we can first use liner regression to obtain their fitting lines (i.e., C'_A, s'_A, C'_B and s'_B), and then compare C'_A with C'_B and s'_A with s'_A, respectively. This gives rise to four cases, among which are the two cases (i.e., $\{C'_A \geq C'_B, s'_A \geq s'_B\}$ and $\{C'_A \leq C'_B, s'_A \leq s'_B\}$) where we can show α-guesswork [8] is deterministic. This makes it much simpler to use in these two cases.

6 Conclusion

In this paper, we have revealed three important implications of the Zipf's law in passwords. While most password-related cryptographic protocols, security policies and metrics either are based on the uniform assumption of password distribution or simply avoid making an explicit assumption of password distribution, it is of great importance to study the implications when user-chosen passwords actually follow the Zipf's law, a distribution far from uniform. We have provided more accurate security formulations for provably secure password protocols, suggested policy parameters with better security and usability tradeoff, and proved a new, inherent property for the metric α-guesswork. Particularly, extensive experiments on 9 large-scale password datasets, which consist of 112 million real-world passwords and cover various popular Internet services and diversified user bases, demonstrate the validity of our proposed implications. Besides, Zipf's law can be useful for other situations, e.g., to evaluate the goodness/validity of algorithms/studies (such as honeywords generation, password hash and user studies) in which a human password distribution needs to be reproduced.

Acknowledgment. We are grateful to the anonymous reviewers for their invaluable comments. Ping Wang is the corresponding author. This research was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61472016.

References

1. Abdalla, M., Benhamouda, F., MacKenzie, P.: Security of the J-PAKE password-authenticated key exchange protocol. In: Proceedings of IEEE S&P 2015, pp. 571–587 (2015)
2. Abdalla, M., Benhamouda, F., Pointcheval, D.: Public-key encryption indistinguishable under plaintext-checkable attacks. In: Katz, J. (ed.) PKC 2015. LNCS, vol. 9020, pp. 332–352. Springer, Heidelberg (2015)
3. Bagherzandi, A., Jarecki, S., Saxena, N., Lu, Y.: Password-protected secret sharing. In: Proceedings of ACM CCS 2011, pp. 433–444 (2011)
4. Bellare, M.: Practice-oriented provable-security. In: Proceedings of ISC 1997, pp. 221–231 (1997)
5. Bellare, M., Hoang, V.T.: Adaptive witness encryption and asymmetric password-based cryptography. In: Katz, J. (ed.) PKC 2015. LNCS, vol. 9020, pp. 308–331. Springer, Heidelberg (2015)
6. Benhamouda, F., Blazy, O., Chevalier, C., Pointcheval, D., Vergnaud, D.: New techniques for SPHFs and efficient one-round PAKE protocols. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013, Part I. LNCS, vol. 8042, pp. 449–475. Springer, Heidelberg (2013)
7. Blocki, J., Datta, A.: CASH: a cost asymmetric secure hash algorithm for optimal password protection. In: IEEE CSF 2016 (2016). arxiv.org/pdf/1509.00239v1.pdf
8. Bonneau, J.: The science of guessing: analyzing an anonymized corpus of 70 million passwords. In: Proceedings of IEEE S&P 2012, pp. 538–552 (2012)
9. Bonneau, J., Herley, C., Oorschot, P., Stajano, F.: The quest to replace passwords: a framework for comparative evaluation of web authentication schemes. In: Proceedings of IEEE S&P 2012, pp. 553–567 (2012)

10. Burr, W., Dodson, D., Perlner, R., Gupta, S., Nabbus, E.: NIST SP800-63-2: electronic authentication guideline. Technical report, National Institute of Standards and Technology, Reston, VA, August 2013
11. Byun, J.W.: Privacy preserving smartcard-based authentication system with provable security. Secur. Commun. Netw. **8**(17), 3028–3044 (2015)
12. Carnavalet, X., Mannan, M.: A large-scale evaluation of high-impact password strength meters. ACM Trans. Inform. Syst. Secur. **18**(1), 1–32 (2015)
13. Castelluccia, C., Dürmuth, M., Perito, D.: Adaptive password-strength meters from markov models. In: Proceedings of NDSS 2012, pp. 1–15 (2012)
14. Chatterjee, R., Bonneau, J., Juels, A., Ristenpart, T.: Cracking-resistant password vaults using natural language encoders. In: Proceedings of IEEE S&P 2015, pp. 481–498 (2015)
15. Chen, L., Lim, H.W., Yang, G.: Cross-domain password-based authenticated key exchange revisited. ACM Trans. Inform. Syst. Secur. **16**(4), 1–37 (2014)
16. Dürmuth, M., Freeman, D., Biggio, B.: Who are you? A statistical approach to measuring user authenticity. In: NDSS 2016, pp. 1–15 (2016)
17. Florêncio, D., Herley, C., van Oorschot, P.: An administrators guide to internet password research. In: Proceedings of USENIX LISA 2014, pp. 44–61 (2014)
18. Gjøsteen, K., Thuen, Ø.: Password-based signatures. In: Petkova-Nikova, S., Pashalidis, A., Pernul, G. (eds.) EuroPKI 2011. LNCS, vol. 7163, pp. 17–33. Springer, Heidelberg (2012)
19. Herley, C., Van Oorschot, P.: A research agenda acknowledging the persistence of passwords. IEEE Secur. Priv. **10**(1), 28–36 (2012)
20. Huang, X., Xiang, Y., Bertino, E., Zhou, J., Xu, L.: Robust multi-factor authentication for fragile communications. IEEE Trans. Depend. Secur. Comput. **11**(6), 568–581 (2014)
21. Huang, Z., Ayday, E., Hubaux, J., Juels, A.: Genoguard: protecting genomic data against brute-force attacks. In: Proceedings of IEEE S&P 2015, pp. 447–462 (2015)
22. Huh, J.H., Oh, S., Kim, H., et al.: Surpass: system-initiated user-replaceable passwords. In: Proceedings of CCS 2015, pp. 170–181 (2015)
23. Jarecki, S., Kiayias, A., Krawczyk, H., Xu, J.: Highly-efficient and composable password-protected secret sharing. In: Proceedings of IEEE EuroS&P 2016, pp. 276–291 (2016)
24. Jarecki, S., Krawczyk, H., Shirvanian, M., Saxena, N.: Device-enhanced password protocols with optimal online-offline protection. In: ASIACCS 2016, pp. 177–188 (2016)
25. Katz, J., Ostrovsky, R., Yung, M.: Efficient and secure authenticated key exchange using weak passwords. J. ACM **57**(1), 1–41 (2009)
26. Katz, J., Vaikuntanathan, V.: Round-optimal password-based authenticated key exchange. J. Crypt. **26**(4), 714–743 (2013)
27. Kiefer, F., Manulis, M.: Zero-knowledge password policy checks and verifier-based PAKE. In: Kutyłowski, M., Vaidya, J. (eds.) ESORICS 2014, Part II. LNCS, vol. 8713, pp. 295–312. Springer, Heidelberg (2014)
28. Li, Y., Wang, H., Sun, K.: A study of personal information in human-chosen passwords and its security implications. In: Proceedings of INFOCOM 2016, pp. 1–9 (2016)
29. Ma, J., Yang, W., Luo, M., Li, N.: A study of probabilistic password models. In: Proceedings of IEEE S&P 2014, pp. 689–704 (2014)
30. Malone, D., Maher, K.: Investigating the distribution of password choices. In: Proceedings of WWW 2012, pp. 301–310 (2012)

31. Martin, R.: Amid Widespread Data Breaches in China, December 2011. http://www.techinasia.com/alipay-hack/
32. Mick, J.: Russian Hackers Compile List of 10M+ Stolen Gmail. Yandex, Mailru, September 2014. http://t.cn/R4tmJE3
33. 1st NSA Annual Best Scientific Cybersecurity Paper Competition, July 2013. http://cps-vo.org/group/sos/papercompetition2012
34. Schechter, S., Herley, C., Mitzenmacher, M.: Popularity is everything: a new approach to protecting passwords from statistical-guessing attacks. In: Proceedings of HotSec 2010, pp. 1–8 (2010)
35. Wang, D., He, D., Cheng, H., Wang, P.: fuzzyPSM: a new password strength meter using fuzzy probabilistic context-free grammars. In: Proceedings of DSN 2016, pp. 595–606 (2016)
36. Wang, D., Jian, G., Huang, X., Wang, P.: Zipf's law in passwords. IEEE Trans. Inform. Foren. Secur. (2016, in press). http://t.cn/RqT51U8
37. Wang, D., Wang, P.: The emperor's new password creation policies: an evaluationof leading web services and the effect of role in resisting against online guessing. In: Proceedings of ESORICS 2015, pp. 456–477 (2015)
38. Wang, Y.: Password protected smart card and memory stick authentication against off-line dictionary attacks. In: Gritzalis, D., Furnell, S., Theoharidou, M. (eds.) SEC 2012. IFIP AICT, vol. 376, pp. 489–500. Springer, Heidelberg (2012)
39. Yan, J., Blackwell, A.F., Anderson, R.J., Grant, A.: Password memorability and security: empirical results. IEEE Secur. Priv. **2**(5), 25–31 (2004)
40. Yi, X., Hao, F., Chen, L., Liu, J.K.: Practical threshold password-authenticated secret sharing protocol. In: Pernul, G., Ryan, P.Y.A., Weippl, E. (eds.) ESORICS. LNCS, vol. 9326, pp. 347–365. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24174-6_18