On the Economics of Offline Password Cracking

Jeremiah Blocki
Purdue University

Ben Harsha
Purdue University

Samson Zhou
Purdue University

Abstract—We develop an economic model of an offline password cracker which allows us to make quantitative predictions about the fraction of accounts that a rational password attacker would crack in the event of an authentication server breach. We apply our economic model to analyze recent massive password breaches at Yahoo!, Dropbox, LastPass and AshleyMadison. All four organizations were using key-stretching to protect user passwords. In fact, LastPass’ use of PBKDF2-SHA256 with \(10^5\) hash iterations exceeds 2017 NIST minimum recommendation by an order of magnitude. Nevertheless, our analysis paints a bleak picture: the adopted key-stretching levels provide insufficient protection for user passwords. In particular, we present strong evidence that most user passwords follow a Zipf’s law distribution, and characterize the behavior of a rational attacker when user passwords are selected from a Zipf’s law distribution. We show that there is a finite threshold which depends on the Zipf’s law parameters that characterizes the behavior of a rational attacker — if the value of a cracked password (normalized by the cost of computing the password hash function) exceeds this threshold then the adversary’s optimal strategy is always to continue attacking until each user password has been cracked. In all cases (Yahoo!, Dropbox, LastPass and AshleyMadison) we find that the value of a cracked password almost certainly exceeds this threshold meaning that a rational attacker would crack all passwords that are selected from the Zipf’s law distribution (i.e., most user passwords). This prediction holds even if we incorporate an aggressive model of diminishing returns for the attacker (e.g., the total value of 500 million cracked passwords is less than 100 times the total value of 5 million passwords). On a positive note our analysis demonstrates that memory hard functions (MHFs) such as SCRYPT or Argon2i can significantly reduce the damage of an offline attack. In particular, we find that because MHFs substantially increase guessing costs a rational attacker will give up well before he cracks most user passwords and this prediction holds even if the attacker does not encounter diminishing returns for additional cracked passwords. Based on our analysis we advocate that password hashing standards should be updated to require the use of memory hard functions for password hashing and disallow the use of non-memory hard functions such as BCRYPT or PBKDF2.

1. Introduction

In the last few years breaches at organizations like Yahoo!, Dropbox, LastPass, AshleyMadison, LinkedIn, eBay and Adult FriendFinder have exposed over a billion user passwords to offline attacks. Password hashing algorithms are a critical last line of defense against an offline attacker who has stolen password hash values from an authentication server. An attacker who has stolen a user’s password hash value can attempt to crack each user’s password offline by comparing the hashes of likely password guesses with the stolen hash value. Because the attacker can check each guess offline it is no longer possible to lockout the adversary after several incorrect guesses.

An offline attacker is limited only by the cost of computing the hash function. Ideally, the password hashing algorithm should be moderately expensive to compute so that it is prohibitively expensive for an offline attacker to crack most user passwords e.g., by checking millions, billions or even trillions of password guesses for each user. It is perhaps encouraging that AshleyMadison, Dropbox, LastPass and Yahoo! had adopted slow password hashing algorithms like BCRYPT and PBKDF2-SHA256 to discourage an offline attacker from cracking passwords. In the aftermath of these breaches, the claim that slow password hashing algorithms like BCRYPT or PBKDF2 are sufficient to protect most user passwords from offline attackers has been repeated frequently. For example, LastPass claimed that “Cracking our algorithms [PBKDF2-SHA256] is extremely difficult, even for the strongest of computers.” Security experts have made similar claims about BCRYPT e.g., after the Dropbox breach a prominent security expert confidently stated that “all but the worst possible password choices are going to remain secure” because Dropbox had used the BCRYPT hashing algorithm.

Are these strong claims about the security of BCRYPT and PBKDF2 true? Despite all of their problems passwords remain prevalent and are likely to remain entrenched as the dominant form of authentication on the internet for years to come because they are easy to use and deploy, and users are already familiar with them. It is therefore imperative to develop tools to quantify the damages of password breaches, and provide guidance to organizations on how to store passwords. In this work we seek to address the following question:

Can we quantitatively predict how many user passwords a rational attacker will crack after a breach?

We introduce a game-theoretic model to answer this question and analyze recent data-breaches. Our analysis strongly challenges the claim that BCRYPT and PBKDF2-SHA256 provide adequate protection for user passwords. On the positive side our analysis indicates that more modern
password hashing algorithms [8] (e.g., memory hard functions [9]) can provide meaningful protection against offline attackers.

1.1. Contributions

We first develop a new decision-theoretic framework to quantify the damage of an offline attack. Our model generalizes the stackelberg game-theoretic model of Blocki and Datta [10]. A rational password attacker is economically motivated and will quit guessing once his marginal guessing costs exceed his marginal reward. The attacker’s marginal reward is given by the probability \( p_i \) that the next (ith) password guess is correct times the value of an additional cracked password to the adversary e.g., the additional revenue of selling that password on the black market or the expected amount of additional money that could be extorted from this user. Given the average value \( \nu \) of each cracked password for the adversary, the cost \( k \) of computing the password hash function and the probability distribution \( p_1 > p_2 > \ldots \) over user selected passwords, our model allows us to predict exactly how many passwords a rational adversary will crack. Unlike the model of Blocki and Datta [10] we can use our framework to model a setting in which the attacker encounters diminishing returns as we would expect in most (black)markets i.e., the total value of 500 million cracked passwords may be significantly less than 100 times the total value of 5 million passwords.

Second, we present the strongest evidence to date that Zipf’s law models the distribution of user selected passwords (with the possible exception of the tail of the distribution). These findings strongly support previous conclusions of Wang and Wang [11]. In particular, we show that Zipf’s law closely fits the Yahoo! password frequency corpus. This dataset was collected by Bonneau [12] and later published by Blocki et al. [13]. In contrast to datasets from password breaches the Yahoo! dataset was collected by trusted parties, and is representative of active Yahoo! users (researchers have observed that hacked datasets contain many passwords that appear to be fake [14]). Our sample size, 70 million users, is also more than twice as large as the datasets Wang and Wang [11] used to support their argument that Zipf’s law closely models password datasets.

Third, we show that there is a finite threshold \( T(\cdot) \) which characterizes the behavior of a rational value v-adversary whenever the distribution over passwords follows Zipf’s law. In particular, if the first cracked password has value \( v \geq T(\cdot) \times k \) then the adversary’s optimal strategy is always to continue guessing until he cracks the user’s password. The threshold \( T(y, \tau, \alpha) \) is parameterized Zipf’s law parameters \( y \) and \( \tau \) and a parameter \( \alpha \) representing the rate of password value decay. We remark that, even if Zipf’s law fails to model the tail of the password distribution, the threshold \( T(y, \tau, \alpha) \) still provides a useful characterization of the attacker’s behavior. In particular, if \( (1-x)\% \) of passwords in a distribution follow Zip’s law and the other \( x\% \) follow some unknown (possibly uncrackable) distribution then our bounds imply that an attacker will compromise at least \( (1-x)\% \) of user passwords whenever \( v \geq T(y, \tau, \alpha) \times k \).

Fourth, we also derive model independent upper and lower bounds on the fraction of passwords that a rational adversary would crack. While these bounds are slightly weaker than the bounds we can derive using Zipf’s law these bounds do not require any modeling assumptions e.g., it is impossible to determine for sure whether or not Zipf’s law fits the tail of the password distribution. Interestingly, the lower bounds we derive suggest that state of the art password crackers [15] could still be improved substantially.

Finally, we derive model independent upper and lower bounds on the % of passwords cracked by a rational adversary. These bounds do not rely on the assumption that Zipf’s law models the tail of the password distribution. Nevertheless, our predictions are still quite dire e.g., a rational adversary will crack 51% of Yahoo! passwords at minimum. Our analysis indicates that, to achieve sufficient levels of protection with BCRYPT or PBKDF2, it would be necessary to run these algorithms for well over a second on modern CPU which would constitute an unacceptable authentication delay in many contexts [17]. On a more positive note our analysis suggests that the use of more modern password hashing techniques like memory hard functions can provide strong protection against a rational password attacker without introducing inordinate delays for users during authentication. In particular, our analysis suggests that it could be possible to reduce the % of cracked passwords below 22.2% without increasing authentication delays to a full second.

1.2. Discussion

In light of our analysis we contend that there is a clear need to update standards for password storage to provide developers with clear guidance about the importance of using memory hard functions such as SCRYPT [9] or Argon2id [18]. In a recent recent user study Naikshetra et al. [19] asked developers to select a password hash function for a new social networking platform. None of the

2. Wang and Wang [11] observed that the tails of empirical password datasets are not inconsistent with a Zipf’s law distribution. However, we cannot be entirely confident that Zipf’s law models the tail of the distribution since, by definition, we do not have many samples for passwords in the tail of the distribution.
developers in this study selected a memory hard function and the strongest password hashing algorithms selected were PBKDF2 with 20,000 hash iterations and BCRYPT with 1,024 iterations. The selection of PBKDF2 with 20,000 hash iterations would be deemed acceptable under 2017 NIST standards — PBKDF2 with at least 10,000 iterations is presented an acceptable selection for password hashing. In this sense, LastPass’ use of PBKDF2-SHA256 with 100,000 iterations greatly exceeds current NIST standards. Nevertheless, our analysis suggests that even PBKDF2-SHA256 with 100,000 hash iterations is insufficient to protect a majority of user passwords while memory hard functions such as SCRYPT or Argon2id would provide meaningful protection. In addition to memory hard functions we also advocate for the use of secure distributed password hashing protocols whenever feasible so that an attacker cannot mount an offline attack without breaching multiple authentication servers.

2. Economic Model

2.1. Preliminaries

Given a dataset D of N user passwords we use fi to denote the frequency of the i-th most common password in the dataset and we use pwdi to denote the i-th most common password in the dataset. We use p1, p2, ... to denote the actual distribution over passwords pwd1, pwd2, ... That is p1 is the probability that a random user selects password pwd1. We use p̂i = fi/N to denote an empirical estimate of pi given a dataset D which was sampled from the real password distribution. We also use λi = ∑i=1tpj to denote the cumulative probability of the i most likely passwords. Equivalently, λi denotes the probability that an adversary cracks the user’s password within the first i guesses.

We say that the probability distribution pi ≥ p2 ≥ ... follows Zipf’s law if pi = i−α for some constants s and z. We say that a probability distribution follows a CDF-Zipf distribution if λi = yiτ for some constants r and y.

Offline Attack. To authenticate users password authentication servers traditionally store salted password hashes. In more detail to authenticate user u the authentication stores a record like the following: (u, si, H(pwdi | si)). Here, u is the the username and pwdi is the user’s password, si is a random string called the salt value used to protect against rainbow table attacks and H is a cryptographic hash function. An adversary who breaches the authentication server will be able to obtain the hash value along with the secret salt value. This adversary can now attempt as many guesses as he desires offline by computing the hashes of likely passwords guesses H(g1, si), H(g2, si), ... and comparing these values with

the stolen password hash. The attacker is only limited by the resources that he is willing to invest trying to crack the user’s password.

Key Questions and Parameters. We aim to address the following questions: How many guesses will our rational adversary attempt? What fraction of the user passwords will an adversary manage to break? The answer to these questions will depend on several factors. How valuable is a cracked password to the adversary? How much does it cost to compute H each time we validate a new password guess? What does the distribution over user passwords look like?

We use ν to denote the value of a cracked password to the adversary measured in units of $C_H$, where H is an underlying cryptographic hash function like SHA256. We can estimate ν by looking at black market prices for cracked passwords. For example, Fossi et al. [10] found that the market price for hacked passwords tends to lie in the range $[4, 30]$. A more recent analysis of Yahoo! passwords found that they sell for between $0.7$ and $1.2$ [25] — the drop in price may be due to an increased supply of Yahoo! passwords. Herley and Florencio found that dishonest behavior can significantly inhibit trade on black markets [26]. Thus, these prices may underestimate the true value of a cracked password.

Password hash functions are often constructed from an underlying cryptographic hash function H. For example, PBKDF2-SHA256 simply iterates the SHA256 hash function multiple times. We use k to denote the cost of a computing the final password hash function — once again measured in units of $C_H$. We use ν = ν × C_H (resp. k3 = k × C_H) to denote the value (resp. cost) in USD given an estimate of $C_H$.

2.2. Rational Adversary

We model a rational adversary who has obtained the salted password hash of a user’s password. Our model generalizes the stackelberg game-theoretic framework of Blocki and Datta [10] by introducing a parameter 0 ≤ α ≤ 1 which models diminishing returns. We assume that adversary knows the password distribution p1, p2, ... as well as the corresponding passwords pwd1, pwd2, ... However, the adversary does not know which password the user selected.

Attacker Game. We model password cracking using a single-shot game. In the game we sample a random password pwd from the password distribution Pr[pwd] = p1. The adversary picks a threshold t ≥ 0. The threshold t specifies an ordered list L(t) = pwd1, ..., pwd of the t most likely passwords. If the real password is contained in the list of adversary guesses, pwd ∈ L(t), then the adversary receives a payment of ν and we charge the adversary j · k, where j is the index of the correct password guess pwd = pwdj. If the real password is not contained in the list pwd1, ..., pwd of adversary guesses then the adversary receives no payment (ν = 0) and the adversary is charged t · k. Notice that t = 0 corresponds to the strategy in which the adversary gives up without guessing, and t = ∞ is a strategy in which the adversary tries all possible passwords.
corresponds to the strategy in which the adversary never quits. Observe that \( \lambda_t = \sum_{j=1}^{t} p_j \) denotes the fraction of user passwords that are cracked by a threshold \( t \) adversary.

About the Attacker. In our analysis we consider an attacker that is

1. **Informed:** The attacker knows the password distribution \( p_1, p_2, \ldots \) and the associated passwords \( \text{pwd}_1, \text{pwd}_2, \ldots \). However, the attacker does not know which password a particular user \( u \) selected.

2. **Uninformed:** The attacker does not have personal knowledge about the user that can be exploited to improve the guessing attack.

3. **Rational:** The attacker is economically motivated, and will stop attacking the user once marginal guessing costs exceed the marginal guessing rewards.

Discussion. Our attacker model captures the most common types of password attacks. It is generally reasonable to assume that the attacker knows the password distribution — possibly excluding of the tail of the distribution. In particular, previous password breaches provide plenty of training data for the attacker and it is reasonable to assume that password cracking models will continue to improve as attackers obtain more and more training data from future password breaches. We focus on an untargeted attacker in our analysis. However, we stress that our model may also be useful when considering a targeted attacker with background knowledge of the user (e.g., name, birthdate, hobbies etc...). In particular, let \( p_i \) denote the probability that a targeted adversary’s \( i \)’th guess is correct. Wang et al. observed that a targeted distribution over user passwords \( p_1, p_2, \ldots \) still seems to follow Zipf’s law.

Rational Attacker Behavior. If the adversary chooses a threshold \( t \) then his expected guessing costs are

\[
C(t) = t \left( 1 - \sum_{j=1}^{t} p_j \right) k + k \sum_{j=1}^{t} j \cdot p_j.
\]

Similarly, his expected reward is

\[
R(t) = v \left( \sum_{j=1}^{t} p_j \right)^{\alpha}
\]

where the parameter \( 0 \leq \alpha \leq 1 \) allows us to model diminishing returns for the attacker as he obtains additional cracked passwords. For example, let \( t_{1\%} \) (resp. \( t_{2\%} \)) be given such that \( p_1 + \ldots + p_{t_{1\%}} = 0.01 \) (\( p_1 + \ldots + p_{t_{2\%}} = 0.02 \)) then for \( \alpha < 1 \) we have \( R(t_{2\%}) = 2^\alpha R(t_{1\%}) \times R(t_{1\%}) \) even though an adversary cracks twice as many passwords by increasing his threshold from \( t_{1\%} \) to \( t_{2\%} \).

**Diminishing Returns:** We note that the original model of Blocki and Datta [10] is a special case of our model when \( \alpha = 1 \) (no diminishing returns). There are a number of reasons why an attacker may encounter diminishing returns \( (\alpha < 1) \) for additional cracked passwords. First, if the attacker plans to sell the passwords on the black market then basic economics suggests that increasing the supply of cracked passwords is likely to drive down prices. In the case of a large breach like Yahoo! (500 million passwords) it is conceivable the number of available passwords on the black market might quickly increase by two orders of magnitude. Second, the more user accounts that are hacked/actively exploited the more likely it is that the original breach will be detected. If the breach is detected then an organization can ask (or require) users to change their passwords or require two-factor authentication, which will reduce the value of each cracked password.

**Interpreting model parameter \( v \):** We note that we have \( v = R(\infty) \), where \( R(\infty) \times N \) denotes the total value of a completely cracked password dataset of size \( N \). Thus, the parameter \( v \) denotes the average value of a cracked password given that all password have been cracked. We can estimate this parameter based on black market sales data. For example, suppose that we know that \( R(t_{1\%}) = 4 \times 1\% \) e.g., from equilibrium black market prices when only 1% of cracked passwords are on the market. In this case we can extrapolate

\[
v = R(\infty) = R(t_{100\%}) = 100^a R(t_{1\%}) = \frac{4}{100^a} \cdot \frac{1}{100^a} .
\]

**Rational Attacker Behavior:** Formally, the rational adversary will select the threshold \( t^* \) maximizing his overall utility

\[
t^* = \arg \max_t \left( R(t) - C(t) \right).
\]

Intuitively, a rational adversary should stop guessing if the marginal cost of one more password guess exceeds the marginal benefit of that guess. Thus, we will have \( MC(t^*) = C(t^*) - C(t^* - 1) \approx MR(t^*) = R(t^*) - R(t^* - 1) \)

The marginal cost of increasing the threshold from \( t - 1 \) to \( t \)

\[
MC(t) = C(t) - C(t - 1) = k \left( 1 - \sum_{j=1}^{t-1} p_j \right) .
\]

Intuitively, the attacker pays an extra cost \( k \) to hash \( \text{pwd}_i \) if and only if the first \( t - 1 \) guesses are incorrect. Similarly, the attacker’s marginal revenue is \( MR(t) = R(t) - R(t - 1) \) when \( \alpha = 1 \) we have \( MR(t) = v \times p_t \) otherwise

\[
MR(t) = v \left( \sum_{j=1}^{t} p_j^{\alpha} - \sum_{j=1}^{t-1} p_j^{\alpha} \right) \times p_t .
\]

Note that \( \lambda_t \) denotes the expected fraction of passwords compromised by a rational attacker. Given a specific assumption about the password distribution (e.g., Zipf’s law) we can derive bounds on \( \lambda_t \).

Competition. We do not attempt to directly model the behavior of an adversary who faces competition from other password crackers. Many breaches (e.g., Yahoo!, LinkedIn, 5. However, the cracked passwords arguably still have significant value after the breach is detected for two reasons. First, many users will not update their passwords unless they are required to do so. Second, many of the users that do update their passwords may do so in a predictable way. Third, many users will have the same password for other accounts.
we can reduce \( \lambda \) with PBKDF2-SHA256 to authenticate e.g., it would at least a second to compute authentication times. In particular, the number of sequential attempts by the defender is constrained by server workload and by the cost of computing the underlying hash function \( H \) and requires \( \tau \) blocks of memory. Thus, the Area x Time (AT) complexity of computing the Memory Hard Function scales with \( \tau^2 \) because the adversary must allocate \( \tau \) blocks of memory for \( \tau \) units of time. In particular, we use \( k = \tau \times C_H + \tau^2 \times C_{mem} \) to model the approximate cost of computing a memory hard function which iteratively makes \( \tau \) calls to the underlying hash function \( H \) and requires \( \tau \) blocks of memory. By contrast, the AT complexity of BCRYPT and PBKDF2 is just \( k = \tau \) since these functions can be computed with a single block of memory. Here, \( C_{mem} \) is a constant representing the core memory-area ratio. That is the area of one block of memory on chip divided by the area of a core evaluating \( H \) on chip. In this paper we use the estimate \( C_{mem} \approx 1/3000 \) as in [30, 31] though we stress that our analysis could be easily repeated with different parameter choices.

Model Limitations. To keep exposition simple we do not attempt to incorporate any model of equilibrium prices for cracked passwords on the black market and instead assume that the value of a cracked password \( \psi \) is static for all users. A targeted adversary may have higher valuations for specific user passwords e.g., celebrities, politicians. Similarly, an attacker who floods a black market with cracked passwords may drive equilibrium prices down. Our primary findings would not be altered in any significant way by including such a model unless equilibrium prices drop by 1–2 orders of magnitude [16]. We also remark that our intention is to model an untargeted economically motivated attacker and not a nation state focused on cracking the passwords of a particular person of interest. However, it may still be reasonable to believe that a nation state attacker will be largely be constrained by economic considerations (e.g., expected value of additional intelligence gained by cracking the password versus expected cost to crack password).

3. Yahoo! Passwords Follow Zipf’s Law

Zipf’s law states that the frequency of an element in a distribution is related to its rank in the distribution. There are two variants of Zipf’s law for passwords: PDF-Zipf and CDF-Zipf. In the CDF-Zipf model we have \( \lambda_i = \sum_{i=1}^N \psi_i = y \cdot t^r \), where the constants \( y \) and \( r \) are the CDF-Zipf parameters. In the PDF-Zipf model we have \( \lambda_i = \frac{C}{t^r} \), where \( s \) and \( C \) are the PDF-Zipf parameters. Normalizing by \( N \) the number of users we have \( \psi_i = \frac{C}{t^r} \), where \( z = \frac{C}{N} \).
Wang et al. [12] previously found that password frequencies tend to follow PDF-Zipf's law if the tail of the password distribution (e.g., passwords with frequency \( f_i < 5 \)) is dropped. Wang and Wang [11] subsequently found that CDF-Zipf's model is superior in that the CDF-Zipf fits were more stable than PDF-Zipf fits and that the CDF-Zipf fit performed better under Kolmogorov-Smirnov (KS) tests. Furthermore, the CDF-Zipf model can fit the entire password distribution (e.g., without excluding passwords with frequency \( f_i < 5 \)). These claims were based on analysis of several smaller password datasets (N ≤ 32.6 million users) which were released by hackers.

In 2016 Yahoo! allowed the release of a differentially private list of password frequencies for users of their services [13]. We refer an interested reader to [12, 13] for additional details about how the Yahoo! data was collected and how it was perturbed to preserve differential privacy. The Yahoo! dataset is superior to other datasets in that it offers the largest sample size \( N \approx 70 \) million and the dataset was collected and released by trusted parties. We show that the Yahoo! dataset is also well modeled by CDF-Zipf’s law. Our analysis comprises the strongest evidence to date of Wang and Wang’s premise [11] that password distributions follow CDF-Zipf’s law due to the advantages of the Yahoo! dataset. We focus on the CDF-Zipf’s law model in this section since it can fit the entire password distribution [11]. We also verified that the Yahoo! dataset is also well modeled by PDF-Zipf’s law if we drop passwords with frequency \( f_i < 5 \) like Wang et al. [12], but we omit this analysis from the submission due to lack of space.

The rest of this section is structured as follows: First, in section 3.1 we discuss the advantages of using the Yahoo! dataset over leaked datasets like RockYou. In 3.2 we show that the noise that was added to preserve differential privacy will have a negligibly small impact on CDF-Zipf fittings. In section 3.3 we use subsampling to show that the CDF-Zipf fittings for Yahoo! converge to a stable solution. Finally, in section 3.4 we present the CDF-Zipf fitting for the entire Yahoo! dataset.

### 3.1. On Ecological Validity

The Yahoo! frequency corpus offers many advantages over breached password datasets such as RockYou or Tianya.

- The Yahoo! password frequency corpus is based on 70 million Yahoo! passwords — more than twice as large as any of the breached datasets analyzed by Wang and Wang [11].
- The records were collected in a trusted fashion. No infiltration, hacking, tricks, or general foul play was used to obtain any of this data. There was no ulterior motive behind collecting these passwords other than to provide valuable data in a way that can be used for scientific research. By contrast, it is possible that hackers strategically omit (or inject) password data before they release a breached dataset like RockYou or Tianya! Why should we trust rogue hackers to provide researchers with representative password data?
- Breached password datasets often contain many passwords/accounts that look suspiciously fake. In 2016 Yang et al. [14] suggested that such passwords can be removed with DBSCAN [33]. Cleansing operations ended up removing a reasonable portion of the dataset (e.g., 5 million passwords were removed from RockYou’s data). With the Yahoo! data such cleansing is not needed, as it was collected in a manner that ensured collected passwords were in use. Previous work that has been done on Zipf distributions in breached password datasets [11] did not perform any sort of sanitizing step on the data. It is unclear how such operations would affect the Zipf law fit.
- The information is released in a responsible way that preserves users’ privacy. The differential privacy mechanism means that even with the released data it is not possible to determine any new information about Yahoo’s users that an adversary would not be able to obtain anyways.
- Data from the Yahoo! password frequency corpus ultimately is derived from the passwords of active Yahoo! users who were logging in during the course of the study as opposed to passwords from throwaway accounts that have been long forgotten.

### 3.2. On the Impact of Differential Privacy on CDF-Zipf Fits

The published Yahoo! password frequency lists were perturbed to ensure differential privacy. Before attempting to fit this dataset using Zipf’s law we seek to answer the following question: Does this noise, however small, affect our CDF-Zipf fitting process in any significant way? We claim that the answer is no, and we offer strong empirical evidence in support of this claim. In particular, we took the RockYou dataset (N ≈ 32.6 million users) and generated 30 different perturbed versions of the frequency list by running the \((\epsilon, \delta)\)-differentially private algorithm of Blocki et al. [13]. We set \( \epsilon = 0.25 \), the same value that was used to collect the Yahoo! dataset that we analyze. For each of these perturbed frequency lists we compute a CDF-Zipf law fit using linear least squares regression. To apply Linear Least Squares regression we apply logarithms to the

<table>
<thead>
<tr>
<th>List Version</th>
<th>( y )</th>
<th>( \sigma_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>RockYou Standard</td>
<td>0.0285</td>
<td></td>
</tr>
<tr>
<td>RockYou Diff. Private</td>
<td>0.0302</td>
<td>1.348 * 10^{-6}</td>
</tr>
<tr>
<td>RockYou Standard</td>
<td>0.2108</td>
<td></td>
</tr>
<tr>
<td>RockYou Diff. Private</td>
<td>0.2077</td>
<td>2.94 * 10^{-6}</td>
</tr>
<tr>
<td>RockYou Standard</td>
<td>0.7687</td>
<td></td>
</tr>
<tr>
<td>RockYou Diff. Private</td>
<td>0.9681</td>
<td>6.50 * 10^{-7}</td>
</tr>
</tbody>
</table>

TABLE 1: Impact of Differential Privacy on CDF Fit
CDF-Zipf equation \( \lambda_t = y \cdot t^r \) to obtain a linear equation
\( \log \lambda_t = \log y + r \log t \).

Our results, shown in Table 1, strongly suggest that the differential privacy mechanism does not impact the parameters \( y \) and \( r \) in a CDF-Zipf fitting in any significant way. In particular, the parameters \( y \) and \( r \) we obtain from fitting the original data with a CDF-Zipf model are virtually indistinguishable from the parameters we obtain by fitting on one of the perturbed datasets. Similarly, differential privacy does not affect the \( R^2 \) value of the CDF-Zipf fit. Here, \( R^2 \) measures how well the linear regression models the data (\( R^2 \) values closer to 1 indicate better fittings). Thus, one can compute CDF-Zipf’s law parameters for the Yahoo! data collected by [13] and [12] without worrying about the impact of the \((\epsilon, \delta)\)-differentially private algorithm used to perturb this dataset. We also verified that the noise added to the Yahoo! dataset will also have a negligible affect on the parameters \( s \) and \( z \) in a PDF-Zipf fitting.

3.3. Testing Stability of CDF-Zipf Fit via Subsampling

There are two primary ways to find a CDF-Zipf fit: Golden Section Search (GSS) and Linear Least Squares (LLS). Wang et al. [11] previously found that CDF-Zipf fits stabilize more quickly with GSS than with LLS. This was particularly important because the largest dataset they tested had size \( \approx 3 \times 10^7 \). In this section we test the stability of LLS by subsampling from the much larger Yahoo! dataset. In particular, we subsample (without replacement) datasets of size 15 million, 30 million, 45 million and 60 million and use LLS to compute the CDF-Zipf parameters \( y \) and \( r \) for each subsampled dataset. Our results are shown in Table 2 graphically in Figure 1. While the CDF-Zipf fit returned by LLS does take longer to stabilize our results indicate that it does eventually stabilize at larger (sub)sample sizes (e.g., the Yahoo! dataset).

We also found that the PDF-Zipf parameters \( s \) and \( z \) stabilize before \( N = 7 \times 10^7 \) samples.

3.4. Fitting the Yahoo! data set with CDF-Zipf

We used both LLS regression and GSS to obtain separate CDF-Zipf fittings for the Yahoo! dataset. The results, shown in Table 3 and graphically in Figure 1, showed that both methods produce high quality fittings. In addition to the parameters \( y \) and \( r \) we report \( R^2 \) values and Kolmogorov-Smirnov (KS) distance. The KS test can be thought of as the largest distance between the observed discrete distribution \( F_n(x) \) and the proposed theoretical distribution \( F(x) \). Formally,

\[
D_{KS} = \sup |F_n(x) - F(x)|
\]

Intuitively, smaller \( D_{KS} \) values (resp. larger \( R^2 \) values) indicates better fits.

3.4.1. Discussion. Both LLS and GSS produce high quality CDF-Zipf fittings (e.g., \( R^2 = 0.9544 \)) for the Yahoo! dataset. LLS regression outperforms the golden section search under both \( R^2 \) and Kolmogorov-Smirnov (KS) tests. Wang and Wang [11] had previously adopted golden section search because the results stabilized quickly. While this was most likely the right choice for smaller password datasets like RockYou, our analysis in the previous section suggest that LLS eventually produces stable solutions when the sample size is large (e.g., \( N \geq 60 \) million samples) as it is in the Yahoo! dataset. Thus, in the remainder of the paper we use the CDF-Zipf parameters \( y = 0.0211 \) and \( r = 0.2166 \) from LLS regression. We stress that the decision to use the CDF-Zipf parameters from LLS instead of the parameters returned by GSS does not affect our findings in any significant way.

We remark that LLS is also more efficient computationally. While we were able to run GSS to find a CDF-Zipf fit for the Yahoo! dataset (\( N \approx 7 \times 10^7 \)), running GSS on a dataset of \( N = 1 \) billion passwords (e.g., the size of the most recent Yahoo! breach [34]) would be difficult if not intractable. By contrast, LLS could still be used to find a CDF-Zipf fitting and our analysis suggests that the fit would be superior.
4. Analysis of Rational Adversary Model for Zipf’s Law

In this section, we show that there is a finite threshold $T(y, r, a)$ which characterizes the behavior of a rational offline adversary when user passwords follow CDF-Zipf’s law with parameters $y$ and $r$ i.e., $\lambda_n = y_n^r$. In particular, Theorem 1 gives a precise formula for computing this threshold $T(y, r, a)$. If $v/k \geq T(y, r, a)$ then a rational value $v$ adversary will proceed to crack all user passwords as marginal guessing rewards will always exceed marginal guessing costs for a rational attacker. In Table 4, we use this formula to explicitly compute $T(y, r, a)$ for the Yahoo! dataset as well as for nine other password datasets analyzed by Wang and Wang [11].

We note that we choose to focus on CDF-Zipf’s law in this section as it is believed to be better than PDF-Zipf models. However, we stress that similar bounds can be derived using PDF-Zipf’s law though we omit these results from the submission for lack of space.

**Theorem 1.** Let $k$ denote the cost of attempting a password guess. If

$$v/k \geq T(y, r, a) = \max_{0 \leq t \leq 2(ya^t r^{-1})} \left(1-y(t-1)^r\right)$$

where

$$Z = \left[\left(\frac{1}{y}\right)^{1/r}\right] + 1$$

then a value $v$ rational attacker will crack 100% of passwords chosen from a Zipf’s law distribution with parameters $y$ and $s$.

**Proof:** Suppose a password frequency distribution follows Zipf’s Law, for some parameters $0 < r < 1$ and $y$, so that $\lambda_n = y_n^r$. Since the marginal revenue is $MR(n) = v(\lambda_n^a - \lambda_{n+1}^a)$ and the marginal cost is $MC(n) = k\left(1 - \sum_{n=1}^{t} p_n\right)$, a rational adversary can be assumed to continue attacking as long as $MR(n) \geq MC(n)$. Therefore, the attacker will not quit as long as

$$v(\lambda_n^a - \lambda_{n+1}^a) \geq k\left(1 - \sum_{n=1}^{t} p_n\right)$$

$$v(ya^a r^a - y(t-1)^r) \geq k(1-y(t-1)^r)$$

In particular, the attacker will not quit as long as

$$v \leq \frac{1-y(t-1)^r}{ya^a r^a - y(t-1)^r}.$$ 

Notably, if $\frac{v}{k} \geq \max_{0 \leq t \leq 2(ya^t r^{-1})} \left(1-y(t-1)^r\right)$ for all $t$, then a rational adversary will eventually crack all passwords. Since $ya^t r^a - y(t-1)^r = \int_{t-1}^{t} ya^t r^a dx$ we have $ya^t r^a - y(t-1)^r \leq ya^t r^a - y(t-1)^r \leq y(t-1)^r$ and

$$\max_{0 \leq t \leq 2(ya^t r^{-1})} \left(1-y(t-1)^r\right) \geq \max_{0 \leq t \leq 2(ya^t r^{-1})} \left(1-y(t-1)^r\right).$$

Let $f(t) = \left(1-y(t-1)^r\right)$. We note that if $v/k \geq \max_{0 \leq t \leq 2(ya^t r^{-1})} \left(1-y(t-1)^r\right)$ then for any value of $t$ exceeding $Z$ it will always be true that $\frac{v}{k} \geq 0$, and thus an adversary will be expected to crack all passwords. Then it follows that

$$f'(t) = \frac{(t-1)^r t^{-1}(1-ar^1 y^{1-a})}{(1-ar t)(t-1)^r} \leq \frac{y^{1-ar t}(t-1)^{r-1}}{\text{at}^{r-1}}$$

then $f'(t) \leq 0$ if and only

$$(1-ar)y^{(1-t)}y^{(1-t)} \leq y^{1-ar t}(t-1)^{r-1}$$

Since $(t-1)(1-ar) \leq (t-1)(1-ar) + tr$, then the last expression certainly holds true if $(1-ar) \leq y(t-1)^{r-1}((t-1)(1-ar) + tr)$.

Since these bounds are tight, we can derive by setting $Z = \left[1 + \left(\frac{1}{y}\right)^{1/r}\right]$, it follows that $f'(t) \leq 0$ for all $t > Z$.

5. Analysis of Previous Password Breaches

In this section, we apply our economic model to analyze the consequences of recent password breaches and the impact of defenses that could have been adopted.

5.1. Breaches

We focus on the following breaches in our analysis:

5.1.1. Yahoo! Attackers stole password hashes for 500 million Yahoo! users in 2014, though the breach was unknown to the general public until 2016 [35]. While Yahoo!
used BCRYPT to hash passwords, they have not publicly specified the number of hash iterations $\tau$ that they used. However, we do have empirical password frequency data from 70 million Yahoo! users which allowed us to derive CDF-Zipf parameters $y = 0.0211$ and $\tau = 0.2166$ for Yahoo! passwords. Thus, we can predict the $\%$ of cracked passwords for different values of $\tau$ that Yahoo! might have chosen.

5.1.2. Dropbox. Attackers stole password hashes for $\approx 68.7$ million Dropbox users though the breach was unknown to the general public until 2016. Dropbox used BCRYPT at level 8 (i.e., $\tau = 2^8 = 256$ hash iterations) to hash passwords. We don’t have empirical password data from Dropbox users from which we can derive Zipf’s law parameters $y$ and $\tau$. However, we have Zipf’s law parameters for many other datasets such as RockYou, Tianya, CSDN and Yahoo! allowing us to predict how many passwords a value $v$ adversary would crack if, say, Dropbox passwords and RockYou passwords have similar strength. Arguably, Dropbox passwords could be quite valuable as they are often used to protect sensitive data.

5.1.3. AshleyMadison. Attackers stole nearly 40 million AshleyMadison password hashes in 2015 and released the stolen data publicly a month later. AshleyMadison primarily used BCRYPT at level 12 (i.e., $\tau = 2^{12} = 4,096$ hash iterations) to hash passwords. However, CynoSure Prime noticed that some passwords were effectively protected with MD5 instead of BCRYPT due to an implementation error. CynoSure Prime managed to crack approximately 11 million of these MD5 hashes in just 10 days, though it has been claimed that most of the passwords protected by BCRYPT are uncrackable. Similar to Dropbox, we do not have Zipf’s law parameters for AshleyMadison users. However, it is plausible to believe that these parameters are comparable to the parameters derived from other datasets such as Yahoo! or RockYou!

5.1.4. LastPass. LastPass was using PBKDF2-SHA256 with $\tau = 10^5$ rounds of iteration when they were breached in 2015. Similar to AshleyMadison and Dropbox breaches we don’t currently have Zipf’s law parameters for LastPass passwords though we can still predict how many passwords would be breached under the assumption that these passwords have similar strength to passwords in other datasets like RockYou or Yahoo! Arguably master passwords will be more valuable to an attacker than regular passwords as a master password will unlock multiple user accounts. On the other hand previous research has not found a clear correlation between password strength and account value.

5.2. Results

In section we showed that, if passwords follow CDF-Zipf’s law with parameters $y$ and $\tau$, and $v/k \geq T(y, \tau, a)$ then a rational adversary will crack 100% of user passwords. Figure (a) plots $v = k \times T(y, \tau, 0.8)$ for various thresholds from Table including Yahoo! and RockYou. Thus, for a point $(v, \tau)$ lying on the blue line, a value $v$ rational adversary will crack 100% of Yahoo! passwords when he can compute the hash function at cost $k = \tau$. Note that $\tau = k$ for hash functions like BCRYPT and PBKDF2 — the ones implicitly assumed that the attacker is willing to fabricate an ASIC to evaluate PBKDF2-SHA256 or BCRYPT. We contend that this is a plausible scenario for a rational attacker, since fabrication costs would amortize over the number of user accounts being attacked (e.g., 500+ million). Furthermore, we note that an attacker who is not willing to pay to fabricate an ASIC could obtain similar performance gains using a field programmable gate array (FPGA).
The plots predict that, unless we set $\tau \gg 10^7$, the adversary will crack 100% of passwords in almost every instance. In particular, the levels of key-stretching performed by Dropbox, AshleyMadison and even LastPass are all well below the thresholds necessary to protect Yahoo!, RockYou or CSDN passwords.

While we do not have CDF-Zipf parameters for other breaches such as AshleyMadison, Dropbox, or LastPass, we do have the value $\tau = k$ for each of these breaches. Figure 2(c) plots $v = k \times T(y, r, 0.8)$ only this time we hold $k$ constant and allow $T(y, r, 0.8)$ to vary. For example, in the black line we fix $k = \tau = 10^5$ since LastPass used PBKDF2-SHA256 with $\tau = 10^5$ hash iterations and allow $T(y, r, 0.8)$ to vary. The vertical lines represent the thresholds $T(y, r, 0.8)$ we derive from CDF-Zipf’s law fits for RockYou, Tianya and Yahoo! Table 5 shows the value of $T(y, r, 0.8)$ obtained from 10 different password datasets.

Observe that in all of cases we had $T(y, r, 0.8) \leq 7.64 \times 10^7$. As in Figure 2(b) the $y$-axis in Figure 2(c) is scaled to show the value $v^5$ in USD (estimated). Thus, if Dropbox (resp. AshleyMadison/LastPass) passwords have comparable strength to Yahoo! passwords (resp. Tianya, RockYou) then a rational adversary would crack 100% of these passwords. Indeed, Figure 2(c) shows that unless the thresholds $T(y, r, a)$ for Dropbox/LastPass/AshleyMadison are significantly larger than the previously observed thresholds, a rational adversary would be compelled to crack all passwords, given the range of password values. For example, even if the threshold $T(y, r, a)$ for Dropbox exceeds the threshold for Yahoo! by four orders of magnitude then the adversary will still crack 100% of these passwords.

<table>
<thead>
<tr>
<th>$R(t_{10})$ (USD)</th>
<th>$a = 0.8$</th>
<th>$a = 0.9$</th>
<th>$a = 1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>0.28</td>
<td>0.44</td>
<td>0.70</td>
</tr>
<tr>
<td>1.20</td>
<td>0.48</td>
<td>0.76</td>
<td>1.20</td>
</tr>
<tr>
<td>4.00</td>
<td>1.59</td>
<td>2.52</td>
<td>4.00</td>
</tr>
<tr>
<td>30.00</td>
<td>11.94</td>
<td>18.93</td>
<td>30.00</td>
</tr>
</tbody>
</table>

TABLE 5: $v$ conversion chart

5.3. Discussion

Figures 2(a), 2(b) and 2(c) paint a grim picture. PBKDF2 and BCRYPT most likely provide dramatically insufficient protection for most AshleyMadison, Dropbox, Yahoo! and LastPass users — even if we used the lowest estimation of the value parameter $v$ from Table 5 ($v^5 = 0.28$ USD) and we assume that the attacker faces substantial diminishing returns ($a = 0.8$) for additional cracked passwords. Furthermore, it would not have been possible to provide sufficient protection for users using PBKDF2 or BCRYPT without introducing intolerable authentication delays ($\geq 1$ second).

Our analysis assumes that the password distribution truly follows CDF-Zipf’s law. While previous research (e.g., [32], [11] and our own results in Section 3) strongly supports the hypothesis that most of the password distribution follows Zipf’s law, it is not possible to definitively state that the tail of the password distribution does not follow Zipf’s law since each of the passwords in the tail were (by definition) observed with low frequency. We stress that even if CDF-Zipf’s law does not fit the tail of the password distribution that $T(y, r, a)$ still characterizes adversary behavior. For example, suppose that the $(100 - x)\%$ of passwords follow a Zipf’s law distribution with parameters $y, r$ while $x\%$ of passwords in the tail of the password distribution do not. In this case, whenever $v/k \geq T(y, r, a)$ we a rational adversary will crack at least $(100 - x)\%$ of the user’s passwords which follow Zipf’s Law.

5.4. Memory Hard Functions

Memory hard functions potentially provide a way of increasing computation cost without drastically increasing computation time. As the name suggests memory hard functions require a large amount of memory to evaluate. Thus, the cost of purchasing/renting hardware for password cracking, approximated by a functions Area x Time (AT) complexity, can be substantial for an attacker. Specifically, AT complexity of SCRYPT [9], scales quadratically with the number of time steps $T$. Thus, as discussed in Section 2 we estimate $k^5 = \tau C_H + \tau^2 C_{mem}$, where $C_H \approx 77 \times 10^{-15}$ [29] and $C_{mem} \approx \frac{77}{\tau} \approx 10^7$ as in [30], [31].

In the last section we assumed that the attacker faced aggressive diminishing marginal returns for additional cracked passwords and we used the lowest possible estimations of adversary value finding that an attacker still cracks 100% of passwords from a Zipf’s law distribution. By contrast, in this section we operate under the conservative assumptions that the attacker does not face diminishing returns and we use the larger estimations of adversary value in our analysis. Nevertheless, we find that the use of MHFs can substantially reduce the % of cracked passwords.

Figure 6 plots $v^5$ (estimate) versus the minimum value of $\tau$ necessary to prevent a rational attacker from cracking 100% of passwords. For example, the blue line predicts that if Yahoo! had adopted memory hard functions with only $\tau = 2^{20}$ iterations (0.1 seconds) then a value $S$30 adversary will not crack all passwords selected from a CDF-Zipf’s law distribution with the parameters $y = 0.0211$ and $r = 0.2166$, the parameters for our CDF-Zipf’s fit for Yahoo! passwords. By contrast, Yahoo! would need to set $\tau = 2^{26} \approx 7$ seconds) when using a function like PBKDF2 or BCRYPT just to ensure that the adversary does not crack 100% of passwords when $a = 1$.
Suppose that we are given \( N \) independent samples of users’ passwords. For example, the plots predict that we can significantly reduce the % of cracked passwords (easily below 1%) with out introducing unacceptably long authentication delays when passwords follow a Zipf’s law distribution. By contrast, the plots predict that we would need to set \( \tau \approx 2^{32} \) (400+ seconds) to achieve the same result using PBKDF2 or BCRYPT when \( \alpha = 1 \).

Figure 2 predicts that MHFs prevent a rational adversary from cracking all passwords from a Zipf’s law distribution. Of course, if the adversary still cracks 99.9% of passwords then this result would not be particularly exciting. Figure 4 plots % cracked passwords vs. \( \tau \) against a value \( v^3 = \$4 \) adversary. These plots provide an optimistic outlook for the following attacks.

### 6. Model Independent Analysis

In this section we derive model-independent upper and lower bounds on the % of users whose passwords would be cracked by a rational adversary. The advantage of a model independent analysis is that the bounds we derive apply even if we do not make any assumptions about the shape of the password distribution. As we observed previously it is not possible to definitively claim that the tail of the password distribution follows Zipf’s law — even if the tail of the distribution is not known to be inconsistent with Zipf’s law [32], [11]. The disadvantage of a model independent analysis is that the bounds we are able to derive may not always be tight as the bounds we may be able to derive using specific modeling assumptions e.g., Zipf’s law. In this section we assume for the sake of simplicity that \( \alpha = 1 \) i.e., the marginal value of each additional cracked password remains constant.

Suppose that we are given \( N \) independent samples
will certainly be cracked by a value $V$ who chose password $pw^d$. As before, we will let $f_i$ denote the number of users who chose password $pw^d_i$ in a dataset and without loss of generality assume that these frequencies are sorted so that $f_i \geq f_{i+1}$. We can use $f_i$ to obtain an estimate $\hat{p}_i = \frac{f_i}{\sum f_i}$ for $p_i$, the true probability that that a random user selects the password $pw^d_i$. While we do have $\hat{p}_i \geq \hat{p}_{i+1}$ we stress that we may no longer assume that $p_i \geq p_{i+1}$ since our empirical value $\hat{p}_i$ (resp. $\hat{p}_{i+1}$) may over/under estimate the true probability $p_i$.

### 6.1. Lower Bound

Theorem 2 lower bounds the number of passwords that will be cracked by a rational adversary in expectation. The expectation is taken over $N$ passwords sampled from $X^N$. Theorem 2. If $\frac{N}{X} \geq NL$ and $a = 1$ then a rational adversary will crack at least

$$\sum_{i; f_i \geq j} f_i - \frac{N}{(j-1)!L^{j-1}}$$

user passwords, in expectation.

The proof of Theorem 2 is in appendix A. The proof begins with the observation that a password $pwd^i = pw^d_i$ will certainly be cracked by a value $V$ adversary if $p_i \geq \frac{1}{NL}$. We then introduce the notion of a $(j, L)$-bad overestimate. In particular, a $(j, L)$-bad overestimate for $pwd^1$ occurs when $p_i < \frac{1}{NL}$ but $f_i \geq j$. If we have $f_i \geq j$ then either $p_i \geq \frac{1}{NL}$ and the password will be cracked, or we have a $(j, L)$-bad overestimate for the password $pwd^1$. We can then show that $\frac{N}{(j-1)!L^{j-1}}$ upper bounds the expected number of passwords $pwd^1$ with $(j, L)$-bad overestimates.

### 6.2. Upper Bound

In contrast to Theorem 2, Theorem 3 upper bounds the % of passwords that we expect an attacker to compromise.

Theorem 3. If $\frac{V}{X} \leq NL \left(1 - \frac{1+\epsilon}{N} \sum_{i=1}^t f_i \right)$ then, except with probability $\exp \left(-\frac{e^cN - \sum_{i=1}^t p_i}{2} \right)$, a rational adversary will crack at most $\sum_{i; f_i > j} f_i + \mu(N, L, j) \mu(N, L, j)$ user passwords where $\mu(N, L, j) = \sum_{i; f_i > j} f_i \sum_{\ell = 0}^{j-1} \binom{N-1}{\ell} \left(\frac{1}{NL} \right) ^ \ell \binom{N - 1 \ell}{NL} ^ {N - \ell - 1}$.

The proof of Theorem 3 is in appendix A. Briefly, we apply Chernoff bounds to show that, if $\frac{V}{X} \leq NL \left(1 - \frac{1+\epsilon}{N} \sum_{i=1}^t f_i \right)$, then with high probability the number of user passwords in our dataset that a rational adversary cracks is at most $\sum_{i; f_i \geq j} f_i + \sum_i f_i \times C_i$.

Here, $C_i$ denotes the event that we have a $(j, L)$-bad underestimate for the password $pwd^d_i$. We then separately upper bound the sum $\sum_i f_i \times C_i$ to obtain the bound in Theorem 3.

### 6.3. Applications

Theorems 2 and 3 allows us to derive specific upper and lower bounds by plugging in different values of $j$ and $L$. For example, by increasing $j$ we decrease the term $\frac{N}{(j-1)!L^{j-1}}$ in Theorem 2 but we also decrease the sum $\sum_{i; f_i \geq j} f_i$. Increasing (resp. decreasing) $L$ is equivalent to assuming the adversary has a higher (resp. lower) value for cracked passwords, which intuitively allows us to establish higher lower bounds (resp. smaller upper bounds) on the percentage of passwords cracked.

### 6.3.1. Lower Bounds

Applying Theorem 2 we can derive specific lower bounds for each of the datasets studied as well as for the Yahoo! frequency corpus. For most datasets we obtain our lower bound by setting $j = 2$ and $L = 10$. For the Yahoo! and RockYou datasets we obtained better lower bounds by setting $j = 3$ and $L = 10$. The result appears below:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Unique PWs</th>
<th>Total PWs</th>
<th>$\frac{V}{X}$</th>
<th>$%$ cracked</th>
</tr>
</thead>
<tbody>
<tr>
<td>RockYou</td>
<td>14,326,970</td>
<td>32,581,870</td>
<td>3.2582 × 10^8</td>
<td>46.03</td>
</tr>
<tr>
<td>00webhost</td>
<td>10,583,709</td>
<td>15,251,073</td>
<td>1.3253 × 10^8</td>
<td>20.60</td>
</tr>
<tr>
<td>Battlefield</td>
<td>417,453</td>
<td>542,366</td>
<td>5.4259 × 10^3</td>
<td>13.03</td>
</tr>
<tr>
<td>Tianya</td>
<td>12,898,437</td>
<td>30,981,241</td>
<td>3.0973 × 10^5</td>
<td>48.26</td>
</tr>
<tr>
<td>Dodonew</td>
<td>10,135,260</td>
<td>16,258,891</td>
<td>1.6259 × 10^8</td>
<td>27.66</td>
</tr>
<tr>
<td>CSDN</td>
<td>4,037,605</td>
<td>6,428,277</td>
<td>6.4283 × 10^7</td>
<td>27.19</td>
</tr>
<tr>
<td>Mail.ru</td>
<td>2,954,907</td>
<td>4,932,688</td>
<td>4.9329 × 10^7</td>
<td>30.01</td>
</tr>
<tr>
<td>Gmail</td>
<td>3,112,028</td>
<td>4,929,090</td>
<td>4.9274 × 10^7</td>
<td>26.46</td>
</tr>
<tr>
<td>Furlife.de</td>
<td>115,589</td>
<td>343,064</td>
<td>3.2406 × 10^7</td>
<td>56.04</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>2.94 × 10^8</td>
<td>7.7 × 10^8</td>
<td>7.7 × 10^8</td>
<td>51.86</td>
</tr>
</tbody>
</table>

Remark: When $j = 1$ we have $\sum_{i; f_i \geq j} f_i - \frac{N}{(j-1)!L^{j-1}} = N - N = 0$ meaning that Theorem 2 provides no lower bound on the % of cracked passwords. At first glance this may appear to be a shortcoming of the theorem. However, we observe that it is impossible to obtain better lower bounds without making assumptions about the password distribution. In particular, let $X_1$ (resp. $X_2$) be the uniform distribution over a set of $2^{3n}$ (resp. $2^{8n}$) passwords. Observe that $X_1$ and $X_2$ can induce dramatically different rational attacker behavior (e.g., if the value of a password is $2^{3n}k$, the adversary will crack 100% of passwords if the true password distribution is $X_1$ and 0% of passwords if the true distribution is $X_2$). However, if we draw $N = 2n$ samples from $X_1$ and $X_2$, then the frequency lists for the two password distributions will be indistinguishable ($f_1 = f_2 = \ldots = f_N = 1$) by birthday bounds ($N \ll 2.3n$).

### 6.3.2. Upper Bounds

Similarly, we may use Theorem 2 to derive model independent upper bounds on the percentage of Yahoo! passwords cracked by a rational adversary as shown in Figure 5. As Figure 5 shows we could potentially use memory hard functions to reduce the % of cracked passwords to \approx 20% without increasing authentication time...
past 1 second. This is particularly, impressive when one considers that an attacker only needs a single guess to achieve success rate 1%!

![Graph](image)

Fig. 5: Memory Hard Functions: % cracked by value \(v\in\{54,30\}\) adversary against an ideal MHF with running time parameter \(\tau\).

### 7. Related Work

#### 7.0.1. Password Cracking

The issue of offline password cracking has been known for decades [40]. Password cracking tools have improved steadily as researchers have explored probabilistic password models [41]. Probabilistic Context Free Grammars for passwords [42], [43], [44]. Markov chain models [45], [46], [47], [48] and even neural networks [49]. Attackers may also use public resources (e.g., quotes from the Internet Movie Database or project Gutenberg to crack sentence based passwords [50], [51]) as well as ‘training data’ from previous breaches at companies like RockYou or Tianya to improve cracking algorithms. Improved password cracking tools make it all the more crucial to develop secure tools for key-stretching (e.g., data-independent MHFs) to minimize the number of guesses an attacker can try. Allodi has studied the economics of the black market for certain attacks and malware, which may be useful in understanding how password cracking markets may work [52].

#### 7.0.2. Improving Password Strength

Efforts to encourage (or force) users to select stronger passwords have shown limited success [53], [54], [55], [56], [57], [58] and often induce high usability costs [59]. Users can be encouraged to select stronger passwords by providing feedback during the password creation policy (e.g., [60], [61], [62]) or by providing clear instructions for the user to follow when creating passwords [60], [63]. Another extensive line of research explored the use of password composition policies in which a user is required to select a password satisfying certain requirements e.g., contains numbers and/or capital letters [53], [54], [55], [56], [58], [64]. Password composition policies also introduce a high usability cost [57], [65], [66], [59], and they typically do not increase password strength significantly. In fact, sometimes these policies result in weaker user passwords [67], [54]. Similarly, password strength meters often provide inconsistent feedback [61], [62] and they often fail to persuade users to select strong passwords.

Another line of research has focused on helping users to generate and remember passwords. One prominent suggestion is to turn a phrase or a sentence into a password. It has been claimed that these passwords are as strong as random ones [50], [51], and this has been promoted by NIST and by security experts such as Bruce Schneier [68]. However, subsequent research indicates that these suggestions are less secure than previously believed [69], [70]. Another line of research seeks to develop and promote secure and usable strategies for password management when the user needs to create and remember multiple passwords [71], [72], [73], [74]. However, all of these schemes require a motivated user. Bonneau and Schechter [29] and Blocki et al. [65] showed that users are capable of memorizing higher entropy secrets (e.g., 56 bits) by following spaced repetition schedules.

#### 7.0.3. Other Defenses Against Offline Attacks

If an organization has multiple authentication servers then they could distribute storage and/or computation of the password hashes across multiple servers [75], [22], [23], [24]. Juels and Rivest [76] proposed storing the hashes of fake passwords (honeywords) and using a second auxiliary server to detect authentication attempts with honeywords (alerting the organization that an breach has occurred). The expensive requirement to purchase and maintain extra servers may prevent widespread adoption of these proposals. Even if these defenses were adopted there is still a clear need to use secure key-stretching mechanisms — an adversary who breaches both servers can still mount an offline attack. Another line of research has sought to include the solution(s) to hard artificial intelligence problems in the password hash so that an offline attacker needs human assistance to verify each password guess [77], [78], [79]. These solutions increase user workload during authentication e.g., by requiring the user to solve a CAPTCHA puzzle [72], [79].

#### 7.0.4. Modeling the Distribution of User Selected Passwords

Malone and Kevin initially explored the feasibility of modeling the distribution of user password choices using Zipf’s law [80]. Wang et al. [32] and Wang and Wang [11] continued this line of work by providing improved tech-
niques to fit Zipf’s law parameters to a dataset. Bonneau [12] took a different approach: collect and analyze a massive password frequency corpus with permission from Yahoo! The Yahoo! dataset was recently released using a differentially private algorithm [13]. We elaborate on Zipf’s law and the Yahoo! frequency corpus at length in the body of the paper.

7.1. Key-Stretching

Key-stretching was proposed as early as 1979 [40] with the goal of protecting lower-entropy secrets like passwords against offline attacks by making it economically infeasible for an offline attacker to try millions or billions of guesses. Traditionally key stretching has been performed using hash iteration e.g., PBKDF2 [2] and BCRYPT [1]. However, password hash functions like PBKDF2 and BCRYPT require minimal memory to evaluate and thus passwords protected by these hash functions are highly vulnerable to attackers with customized hardware [81]. Memory hard functions (MHFs), first explicitly introduced by Percival [9], are a promising tool for constructing an ideal key-stretching function. MHFs are motivated by the observation that the cost of storing/retrieving items from memory is relatively constant across different computer architectures. At a high level a memory hard function is moderately expensive to compute and most of the costs associated with computing the function are memory related (e.g., storing/retrieving items from memory). Ideally we want the Area x Time complexity of computing a MHF to scale with $\tau^2$, where $\tau$ denotes the running time on a standard PC. Intuitively, to compute the MHF once the attacker must dedicate $\tau$ blocks of memory for $\tau$ time steps, which ensures that the cost of computing the function is equitable across different computer architectures (memory on an ASIC is still expensive). By contrast, Area x Time complexity to compute BCRYPT or PBKDF2 is simply $\tau$. Recall that we want to increase costs quickly to minimize delay during authentication. If costs scale with $\tau^2$ then we can rapidly drive up costs, and if computation requires memory then an adversary will not be able to significantly reduce guessing costs by constructing an ASIC. Almost all of the entrants to the recent Password Hashing Competition (PHC) [8] claimed some form of memory-hardness.

7.1.1. Data (In)dependent Memory Hard Functions.

There is a type of MHF called a data-independent MHF (iMHF) which is designed to be resistant to side-channel attacks such as cache timing [82], [83]. These functions have a data access pattern independent of the input. Multiple attacks have been shown in several iMHFs [30], [84], [31], [85], [86], [87], [88]. Data dependent MHFs such as SCRYPT [9] have the previously mentioned side-channel vulnerabilities. Even so, SCRYPT has been found to be optimally memory hard with respect to AT complexity [39], [89]. The authors of Argon2 [18], winner of the password hashing competition [8], now recommend running in hybrid mode Argon2id to balance side-channel resistance and resistance to iMHF attacks.

8. Discussion

Our economic analysis decisively shows that traditional key-stretching tools like PBKDF2 and BCRYPT fail to provide adequate protection for user passwords, while memory hard functions do provide meaningful protection against offline attackers. It is time for organizations to upgrade their password hashing algorithms and adopt modern key-stretching such as memory hard functions [9], [8]. Alternatively, could a creative organization adapt customized Bitcoin mining rigs for use in password authentication? For example, the Antminer S9 [81], currently available on Amazon for approximately $3,000, is capable of computing SHA256 14 trillion times per second. If the organization stored salted and peppered $u_i, s_{ui}, SHA256(pwd_{ui}|s_{ui}|p_{ui})$ then it could potentially use the Antminer S9, or a similar Bitcoin mining rig, to validate a password by quickly enumerating over a (very) large space of secret pepper values $p$ (briefly, a secret salt value that is not stored which even an honest party must brute force).

While our analysis demonstrates that the use of memory hard functions can significantly reduce the fraction of cracked passwords, the damage of an offline attack may still be significant. Thus, we recommend that organizations adopt distributed password hashing [75], [22], [23], [24] whenever feasible so that an attacker who only breaches one authentication server will not be able to mount an offline attack. Furthermore, we recommend that organizations take additional measures to mitigate the affect of an authentication server breach. Solutions might include mechanisms detect password breaches through the use of honey accounts or honey passwords [76], multi-factor authentication and fraud detection/correction algorithms to prevent suspicious/harmful behavior [91].

While solid options for password hashing and key-derivation exist [9], [8], [18], [87] the reality is that many organizations and developers select suboptimal password hashing functions [92], [19]. Thus, there is a clear need to provide developers with clear guidance about selecting secure password hash functions. On a positive note recent 2017 NIST guidelines do suggest the use of memory hard functions. However, NIST guidelines still allows for the user of PBKDF2 with just 10,000 hash iterations. Based on our analysis we advocate that password hashing standards should be updated to require the use of memory hard functions for password hashing and disallow the use of non-memory hard functions such as BCRYPT or PBKDF2. It may be expedient for policy makers to audit and/or penalize organizations that fail to follow appropriate standards for password hashing.

We recommend that users primarily focus on selecting passwords that are strong enough to resist targeted online attacks [27] as there is a often a vast gap between the required entropy to resist online and offline attacks [7]. Extra user effort to memorize a high entropy password might be completely wasted if an organization adopts poor password hashing algorithms like SHA1, MD5 [56] or the identity
function [92]. This effort would likely be more productively spent on trying to reduce password reuse [72].

9. Acknowledgments

We would like the thank the reviewers for their insightful comments. We would also like to thank Ding Wang for sharing code for computing Zipf fittings. The work was supported by the National Science Foundation under NSF Awards #1649515 and #1704587. Ben Harsha was partially supported by a Intel Graduate Research Assistantship through CERIAS at Purdue. The opinions expressed in this paper are those of the authors and do not necessarily reflect those of the National Science Foundation or Intel.

References

[22] J. Camenisch, A. Lysyanskaya, and G. Neven, “Practical yet univer-


Appendix

Reminder of Theorem 2. If \( Y \geq NL \) and \( \alpha = 1 \) then a rational adversary will crack at least

\[
\sum_{i:f_i \geq j} f_i \geq \frac{N}{(j-1)!L^{j-1}}
\]

user passwords, in expectation.

Proof of Theorem 2

We observe that a user password \( pwd^i = pwd_i \) will be certainly cracked if \( V \times Pr[pwd_i] = Vp_i \geq k \) since the marginal cost of including an extra guess pwd in the dictionary is at most k. Thus, the adversary will compromise at least \( \sum_{i:p_i \geq k/V} f_i \) accounts. The problem with this lower bound is that we need to know \( p_i = Pr[pwd_i] \) for each password \( pwd_i \) to compute it. However, the values \( p_i \) are unknown if we do not make assumptions about the shape of the password distribution. However, we can lower bound this quantity. In particular, we say that the estimate \( \hat{p}_i = f_i/N \) is a \((j,L)\)-bad overestimate if \( p_i < \frac{1}{
\sum_{i:f_i \geq j} f_i} B_i \) computes the total number of users whose password got a \((j,L)\)-bad overestimate. The proof now follows from Claims 4 and 5. Claim 4 lower bounds the fraction of cracked passwords in terms of the events \( B_i \).

Claim 4. If \( Y \geq NL \) then the number of user passwords in our dataset that a rational adversary cracks is at least

\[
\sum_{i:f_i \geq j} f_i \geq \sum_{i} f_i \times B_i
\]
Proof: Suppose that a user selects a password pwdᵢ with \( pᵢ \geq \frac{1}{NL} \). Since \( Vpᵢ > k \geq \maxₘ \mathcal{MC}(t) \) the marginal reward of guessing pwdᵢ always exceeds the marginal cost. Thus, a rational attacker must eventually guess pwdᵢ. However, if \( pᵢ < \frac{1}{NL} \) then either \( fᵢ < j \) or \( pᵢ \) is a \((j,L)\)-bad overestimate of \( pᵢ \). Let \( S \) denote the set of users who picked a password \( i \) such that \( fᵢ \geq j \) and let \( T \subseteq S \) denote the set of users whose password got a \((j,L)\)-bad overestimate. Any user in the set \( S \setminus T \) will be compromised eventually. Thus, at least \( |S \setminus T| = \sum_{i:fᵢ \geq j} fᵢ - \sum_i fᵢ \times B₁ \) since \( |T| = \sum_i fᵢ \times Bᵢ \) and \( |S| = \sum_i fᵢ \). Now we have

\[
\sum_{j=1}^{N} X_j = \sum_{i=1}^{t} fᵢ \leq \sum_{j=1}^{t} f_j .
\]

□

Claim 8. We have

\[
\sum_{i=1}^{t} pᵢ \leq \left( 1 + \epsilon \right) \sum_{i=1}^{N} X_i
\]

except with probability

\[
\Pr\left[ \sum_{i=1}^{N} X_i \leq \frac{N}{1 + \epsilon} \sum_{i=1}^{t} pᵢ \right] \leq \exp\left( -\frac{2N}{\epsilon^2} \sum_{i=1}^{t} pᵢ \right) .
\]

Proof: Since \( \Pr\left[ Xᵢ = 1 \right] = \sum_{i=1}^{t} pᵢ \), then

\[
\mathbb{E}\left[ \sum_{i=1}^{N} Xᵢ \right] = N \sum_{i=1}^{t} pᵢ .
\]

Then applying Chernoff bounds,

\[
\Pr\left[ \sum_{i=1}^{N} Xᵢ \leq \frac{N}{1 + \epsilon} \sum_{i=1}^{t} pᵢ \right] \leq \exp\left( -\frac{2N}{\epsilon^2} \sum_{i=1}^{t} pᵢ \right) .
\]

□

Claim 9. With high probability,

\[
\mathcal{MC}(t) \geq \left( 1 - \frac{1 + \epsilon}{N} \sum_{i=1}^{t} fᵢ \right) k.
\]

Proof: By Claims 7 and 8

\[
\sum_{i=1}^{t} pᵢ \leq \frac{1 + \epsilon}{N} \sum_{i=1}^{t} fᵢ .
\]

The proof follows from the observation that \( \mathcal{MC}(t) = \left( 1 - \sum_{i=1}^{t-1} \right) k \) and \( pᵢ \geq 0 \).

Now, we define \( \hat{pᵢ} = fᵢ/N \) as a \((j,L)\)-bad underestimate if \( pᵢ > \frac{1}{NL} \), but \( fᵢ \leq j \). Then define \( Cᵢ \) as the indicator variable for the event that \( \hat{pᵢ} \) is a \((j,L)\)-bad underestimate and \( fᵢ \geq 1 \).

Claim 10. If \( \frac{1}{k} \leq \mathbb{N} \left( 1 - \frac{1 + \epsilon}{N} \sum_{i=1}^{t} fᵢ \right) \) then the number of user passwords in our dataset that a rational adversary cracks is at most

\[
\sum_{i:fᵢ \geq j} fᵢ + \sum_{i} fᵢ \times Cᵢ .
\]

Proof: Suppose that a user selects a password pwdᵢ with \( pᵢ \leq \frac{1}{NL} \). Since \( Vpᵢ \leq NL \left( 1 - \frac{1 + \epsilon}{N} \sum_{i=1}^{t} fᵢ \right) k \leq \mathcal{MC}(t) \) the marginal reward of guessing pwdᵢ never exceeds the marginal cost. Thus, a rational attacker never
chooses to guess pwd. If \( p_i > \frac{1}{NL} \) then either \( f_i > j \)
or \( p_i \) is a \((j, L)\)-bad underestimate of \( p_i \). Let \( S \) denote the
set of users who picked a password \( i \) such that \( f_i > j \) and
let \( T \subseteq S \) denote the set of users whose password got a
\((j, L)\)-bad underestimate. Only the users in the set \( S \cup T \)
may be compromised eventually. Thus, at most \( |S \cup T| \leq \sum_{i: f_i > j} f_i + \sum_{i: f_i \leq j} f_i \times C_i \) since \( |T| = \sum_{i: f_i \leq j} f_i \times C_i \) and
\( |S| = \sum_{i: f_i > j} f_i \).

Then the following immediately holds, noting that there can
be at most NL passwords which are \((j, L)\)-bad underesti-
mates:

**Corollary 11.** If \( \frac{V}{\ell} \leq NL \left(1 - \frac{1 + e}{N} \sum_{i=1}^{t} f_i \right) \) then the
number of user passwords in our dataset that a rational
adversary cracks is at most

\[
\sum_{i: f_i \geq j} f_i + \sum_{i} f_i .
\]

**Claim 12.**

\[
\Pr[C_i \mid f_i \geq 1] \leq \sum_{k=0}^{j-1} \binom{N-1}{\ell} \left( \frac{1}{NL} \right)^\ell \left( \frac{NL-1}{NL} \right)^{N-\ell-1}
\]

**Proof:** Recall that for \( C_i = 1 \), we require \( p_i > \frac{1}{NL} \geq \frac{j}{N} \)
but \( f_i \leq j \). Then for \( j < N/2 \),

\[
\Pr[C_i \mid f_i \geq 1] = \sum_{k=0}^{j-1} \binom{N-1}{\ell} p_i^k (1 - p_i)^{N-\ell-1}
\]

\[
= \sum_{k=0}^{j-1} \binom{N-1}{\ell} \left( \frac{1}{NL} \right)^\ell \left( \frac{NL-1}{NL} \right)^{N-\ell-1}
\]

\[
\sum_{i:0<i\leq j} f_i \times \mathbb{E}[C_i \mid f_i \geq 1].
\]

**Proof:** Follows immediately from Claim 12 by substituting
into \( \mathbb{E}[C_i \mid f_i \geq 1] \) in the above sum.
Fig. 7: No Diminishing Returns (α = 1)