Targeted Online Password Guessing: An Underestimated Threat

Conference Paper · July 2016

DOI: 10.1145/2976749.2978339

5 authors, including:

Ding Wang
Peking University
49 PUBLICATIONS **933** CITATIONS
SEE PROFILE

Zijian Zhang
Peking University
1 PUBLICATION **30** CITATIONS
SEE PROFILE

Targeted Online Password Guessing: An Underestimated Threat

Ding Wang[†], Zijian Zhang[†], Ping Wang[†], Jeff Yan[*], Xinyi Huang[‡]
[†]School of EECS, Peking University, Beijing 100871, China
[*]School of Computing and Communications, Lancaster University, United Kingdom
[‡]School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China
{wangdingg, zhangzj, pwang}@pku.edu.cn; jeff.yan@lancaster.ac.uk; xyhuang81@gmail.com

ABSTRACT

While *trawling* online/offline password guessing has been intensively studied, only a few studies have examined *targeted online guessing*, where an attacker guesses a specific victim's password for a service, by exploiting the victim's personal information such as one sister password leaked from her another account and some personally identifiable information (PII). A key challenge for targeted online guessing is to choose the most effective password candidates, while the number of guess attempts allowed by a server's lockout or throttling mechanisms is typically very small.

We propose *TarGuess*, a framework that systematically characterizes typical targeted guessing scenarios with seven sound mathematical models, each of which is based on varied kinds of data available to an attacker. These models allow us to design novel and efficient guessing algorithms. Extensive experiments on 10 large real-world password datasets show the effectiveness of TarGuess. Particularly, TarGuess I∼IV capture the four most representative scenarios and *within* 100 guesses: (1) TarGuess-I outperforms its foremost counterpart by 142% against security-savvy users and by 46% against normal users; (2) TarGuess-II outperforms its foremost counterpart by 169% on security-savvy users and by 72% against normal users; and (3) Both TarGuess-III and IV gain success rates over 73% against normal users and over 32% against security-savvy users. TarGuess-III and IV, for the first time, address the issue of cross-site online guessing when given the victim's one sister password and some PII.

Keywords

Password authentication; Targeted online guessing; Personal information; Password reuse; Probabilistic model.

1. INTRODUCTION

Passwords firmly remain the most prevalent mechanism for user authentication in various computer systems. To understand password security, a number of probabilistic guessing models, e.g., Markov n-grams [21, 25] and probabilistic context-free grammars (PCFG) [31, 35], have been successively proposed. A common feature of these guessing models is that they characterize a *trawling offline guessing* attacker who mainly works against the leaked password files and aims to crack as many accounts as possible. As highlighted in [16], offline guessing attacks, no matter trawling ones or targeted ones, only pose a real concern in the very limited circumstance: the server's password file is leaked, the leakage goes undetected, and the passwords are also properly hashed and salted. Recent research [7, 16] has realized that it should be the role of websites to protect user passwords from offline guessing by securely storing password files, while normal users only need to choose passwords that can survive online guessing.

Online guessing can be launched against the publicly facing server by *anyone* using a browser at *anytime*, with the primary constraint being the number of guesses allowed. *Trawling online guessing* mainly exploits users' behavior of choosing popular passwords [22, 34], and it can be well addressed by various security mechanisms at the server (e.g., suspicious login detection [14], rate-limiting and lockout [18]). However, *targeted online guessing* (see Fig. 1) can exploit not only weak popular passwords, but also passwords reused across sites and passwords containing personal information. This is a serious security concern, since various Personally Identifiable Information (PII) and leaked passwords become readily available due to unending data breaches [2, 3, 17]. For instance, the most recent large-scale PII data breach in April 2016 [3] involves 50 million Turkish citizens, accounting for 64% of the population. According to the CNNIC 2015 report [1], over 78.2% of the 668 million Chinese netizens have suffered PII data leakage. In a series of recent breaches, over 253 million American netizens become victims of PII and password leakage [27].

This indicates that the existing password creation rules (e.g., [15, 28]) and strength meters (e.g., [24, 32]) grounded on these trawling guessing models [21, 25, 31, 35] can mainly accommodate to the limited *offline guessing* threat, taking no account of the *targeted online guessing* threat which is increasingly more damaging and realistic. This misplaced research focus largely attributes to the failure (see [7, 33]) of the academic world to identify the crux of current practices and to suggest convincingly better password solutions than current practices to lead the industrial world.

The main challenge for targeted online password guessing is to effectively characterize an attacker $\mathcal{A}$'s guessing model, with multiple dimensions of available information (see Fig. 2) well captured, while the number of guesses allowed to $\mathcal{A}$ is small – the NIST Authentication Guideline [18] requires Level 1 and 2 systems to keep login failures less than 100 per user account in any 30-day period. The following explains why it is a challenge.

First, people's password choices vary much among each other. When creating a password, some people reuse an existing password, and some modify an existing password; Some incorporate PII into their passwords, yet others do not; Some favor digits, some favor letters, and so on. Thus, a user population's passwords created for a given web service can differ greatly. Therefore, the

CCS'16, October 24-28, 2016, Vienna, Austria

DOI: http://dx.doi.org/10.1145/2976749.2978339

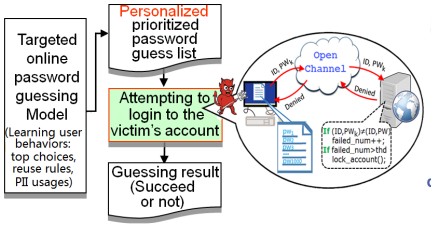


Figure 1: Targeted online guessing.

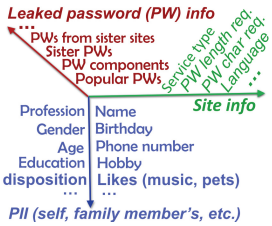


Figure 2: Multiple info for $\mathcal{A}$.

trawling guessing models [21, 25, 31, 35], which aim to produce a *single* guess list for *all* users, are not suitable for characterizing targeted online guessing.

Second, users' PII is highly heterogeneous. Some kinds of PII (e.g., name, and hobby) are composed of letters, some (e.g., birthday and phone number) are composed of digits, and some (e.g., user name) are a mixture of letters, digits and symbols. Some PII (e.g., name, birthday and hobby), as shown in Fig. 2, can be directly used as password components, while others (e.g., gender and education) cannot. As we will show, most of them have an impact on people's password choice. Thus, it is challenging to, at a large-scale, automatically incorporate such heterogeneous PII into guessing models when the guess attempts allowed is limited.

Third, users employ a diversified set of transformation rules to modify passwords for cross-site reuse. As shown in [12, 32], when given a password, there are over a dozen transformation rules, such as insert, delete, capitalization and leet (e.g., `password`→ `passw0rd`) and the synthesized ones (e.g., `password`→ `Passw0rd1`), that a user can utilize to create a new password. How to prioritize these rules for each individual user is not easy.

Moreover, which transformation rules users will apply for password reuse are often context dependent. Suppose attacker $\mathcal{A}$ targets Alice's eBay account which requires passwords of length 8^+, and knows that Alice is in her 30s. With access to a sister password `Alice1978Yahoo` leaked from Alice's Yahoo account, $\mathcal{A}$ will have a higher chance by guessing `Alice1978eBay` than by `Alice1978` due to the inertia of human behaviors. Yet, when Alice's leaked password is `123456`, $\mathcal{A}$ would more likely succeed by guessing `Alice1978` than by `Alice1978eBay`. When site password policies are also considered, the situation may further vary. Such context dependence necessitate an adaptive, semantics-aware cross-site guessing model.

1.1 Related work

Zhang et al. [37] suggested an algorithm for predicting a user's future password with *previous ones* for *the same account*. Das et al. [12] studied the password reuse issue, and proposed a cross-site cracking algorithm. However, their algorithm is *not optimal* for targeted online guessing for four reasons. First, it does not consider common popular passwords (e.g., `iloveyou`, and `pa$$w0rd`) which do not involve reuse behaviors or user PII. Second, it assumes that all users employ the transformation rules in a fixed priority. Yet, as we observe, this priority is actually dynamic and context-dependent. Third, their algorithm does not consider various synthesized rules. Fourth, it is heuristics based.

Li et al. [20] examined how user's PII may impact password security, and found that 60.1% of users incorporate at least one kind of PII into their passwords. They proposed a semantics-rich algorithm, Personal-PCFG, which considers six types of personal information: name, birthdate, phone number, National ID, email address and user name. However, as we will show, its *length-based* PII matching and substitution approach makes it inaccurate to capture user PII usages, greatly hindering the cracking efficiency. Our TarGuess-I manages to overcome this issue by using a *type-based* PII matching approach and gains drastic improvements.

1.2 Our contributions

In this work, we make the following key contributions:

- **A practical framework**. To overcome the challenges discussed above, we propose *TarGuess*, a practical framework to characterize typical targeted online guessing attacks, with sound probabilistic models (rather than ad hoc models or heuristics). TarGuess captures seven typical attacking scenarios, with each based on a different combination of various information available to the attacker.
- **Four probabilistic algorithms**. To model the most representative targeted guessing scenarios, we propose four algorithms by leveraging probabilistic techniques including PCFG, Markov and Bayesian theory. Our algorithms all significantly outperform prior art. We further demonstrate how they can be readily employed to deal with the other three remaining attacking scenarios.
- **An extensive evaluation.** We perform a series of experiments to demonstrate that both the efficacy and general applicability of our algorithms. Our empirical results show that an overwhelming fraction of users' passwords are vulnerable to our targeted online guessing. This suggests that the danger of this threat has been significantly underestimated.
- **New insights.** For example, Type-based PII-tags are more effective than length-based PII-tags in targeted guessing. Simply incorporating many kinds of PII into algorithms will *not* increase success rates, which is counter intuitive. The success rate of a guess decreases with a Zipf's law as the rank of this guess in the guess list increases.

2. PRELIMINARIES

We now explicate what kinds of user personal information are considered in this work and elaborate on the security model.

2.1 Explication of personal information

The most prominent feature that differentiates a targeted guessing attack from a trawling one is that, the former involves user-specific data, or so-called "personal info". This term is sometimes used inter-changeably with the term "personally identifiable info" [10, 20], while sometimes their definitions vary greatly in different situations, laws, regulations [23, 29]. Generally, a user's personal info is "any info relating to" this user [29], and it is broader than PII. For better comprehension, in Table 1 we provide the first classification of personal info in the case of password cracking, making a systematical investigation of targeted guessing possible.

We divide user personal information into three kinds, with each kind having a varied degree of secrecy, different roles in passwords and various types of specific elements. The first kind is user PII (e.g., name and gender), which is natively semipublic: public to friends, colleagues, acquaintances, etc., yet private to strangers. The second kind is user identification credentials, and parts of them (e.g., user name) are public, while parts of them (e.g., password) are exclusively private. The remaining user personal data falls into the third kind and is irrelevant to this work. We further divide user PII into two types: Type-1 and Type-2. Type-1 PII (e.g., name and birthday) can be the building blocks of passwords, while Type-2 PII (e.g., gender and education [22]) may impact user behavior of password creation yet cannot be directly used in passwords. Each type of PII shapes our guessing algorithms quite distinctly.

Here we highlight a special kind of user personal information — a user's passwords at various web services. As shown in [12, 32], users tend to reuse or modify their existing passwords at other sites (called *sister passwords*) for new accounts. However, such sister passwords are becoming more and more easily available due to the unending catastrophic password file leakages (see [2, 4, 27]).

Table 1: Explication of user personal info (NID stands for National identification number, e.g., SSN; PW for password)

Different kinds of personal info		Degree of secrecy	Roles in PWs	**Considered** in this work(✓)	Not Considered in this work(×)
Personally identifiable information (PII)	Type-1	Semipublic	Explicit	Name, Birthday, Phone number, NID	Place of birth, Likes, Hobbies, etc.
	Type-2	Semipublic	**Implicit**	Gender, Age, Language	Faith, Disposition, Education, etc.
User identification credentials		Private	Explicit	Passwords, Personal Identification Numbers	Finger prints, Private keys, etc.
		Public	Explicit	User name, Email address	Debit card number, Health IDs, etc.
Other kinds of personal data		—	—	—	Employment, Financial records, etc.

Table 2: A summary of the four most representative scenarios of targeted online guessing

Attacking scenario	Exploiting *public* information (e.g., datasets and policies)	Exploiting user *personal* information †			Existing literature	**Our model**
		One sister password	Type-1 PII	Type-2 PII		
Trawling #1	✓				Ref. [21, 25, 35]	
Targeted #1	✓		✓		Ref. [20]	TarGuess-I
Targeted #2		✓			Ref. [12]	
	✓	✓			None	TarGuess-II*
Targeted #3	✓	✓	✓		None	TarGuess-III
Targeted #4	✓	✓	✓	✓	None	TarGuess-IV

*As public password datasets are readily available, TarGuess-II and [12] is comparable because they exploit the same type of user PII.

†A total of 7(=C_3^1+C_3^2+C_3^3) scenarios result from combining the three types of personal info. With TarGuess-I∼IV, all 7 cases will be tackled in Sec. 4.

2.2 Security model

Without loss of generality, in this work we mainly focus on the client-server architecture, the most common case of user authentication, as shown in the right of Fig. 1. There are three entities involved in a targeted online guessing attack: a user U, an authentication server S and an attacker $\mathcal{A}$.

User U has registered a password account at the server S. This password is only known to S, though U's passwords at other sites may have already been publicly disclosed. S may be remote (e.g., an e-commerce site) or local (e.g., a password-protected mobile device). To be realistic, we assume that S enforces some security mechanisms such as suspicious login detection and lockout [14,18], and thus the number of guesses allowed to $\mathcal{A}$ is limited (e.g., 10^2 [8,18]). $\mathcal{A}$ knows some amount of personal info about U, and may be a curious friend, a jealous wife, a blackmailer, or even an evil hacker group that buys personal info from the underground market.

As there is a messy mixture of multiple dimensions of info (see Fig. 2) potentially available to the attacker $\mathcal{A}$, it is challenging to characterize $\mathcal{A}$. We tackle this issue by assuming that all the public info (e.g., leaked PW lists and site policies) should be available to $\mathcal{A}$, and then by defining a series of attacking scenarios (see Table 2) based on varied types of U's personal info given to $\mathcal{A}$. This is reasonable: (1) $\mathcal{A}$ is smart and likely to exploit the readily available public info to increase her chance; and (2) $\mathcal{A}$ would use different attacking strategies when given different personal info. Once $\mathcal{A}$ has successfully guessed the password, the victim's sensitive info can be disclosed, reputation could be ruined (see [36]), password account may be hijacked and money might be lost (see [26]).

Note that, here we only consider scenarios where $\mathcal{A}$ is with at most one sister password of user U. The underlying reason is that, among the 547.56M of leaked password accounts that we have collected over a period of six years, less than 1.02% (resp. 1.73%) of them have more than one match by email (resp. user name). Similarly, among the 7.96M accounts collected by Das et al. in 2014 [12], only 152 (0.00191%) of them have more than one match by email. Therefore, *it is realistic to assume that most users have leaked one sister password, and $\mathcal{A}$ can exploit U's this sister password for attacking*.

3. HUMAN BEHAVIORS OF PASSWORD CREATION

Here we report a large-scale empirical study of human behaviors in creating passwords, in particular, how often they choose popular passwords, how often to reuse passwords, how often to make use of their own PII.

Table 3: Basic information about our 10 password datasets

Dataset	Web service	Language	When leaked	Total PWs	With PII
Dodonew	E-commerce	Chinese	Dec., 2011	16,258,891	
CSDN	Programmer	Chinese	Dec., 2011	6,428,277	
126	Email	Chinese	Dec., 2011	6,392,568	
12306	Train ticketing	Chinese	Dec., 2014	129,303	✓
Rockyou	Social forum	English	Dec., 2009	32,581,870	
000webhost	Web hosting	English	Oct., 2015	15,251,073	
Yahoo	Web portal	English	July, 2012	442,834	
Rootkit	Hacker forum	English	Feb., 2011	69,418	✓
Xiaomi*	Mobile, cloud	Chinese	May, 2014	8,281,385	
Xato	Synthesised	English	Feb., 2015	9,997,772	

*Xiaomi passwords are in salted-hash and will be used as real targets.

Table 4: Basic information about our personal-info datasets

Dataset	Language	Number of Items	Types of PII useful for this work
Hotel	Chinese	20,051,426	Name, Gender, Birthday, Phone, NID
51job	Chinese	2,327,571	Email, Name, Gender, Birthday, Phone
12306	Chinese	129,303	Email, User name, Name, Gender, Birthday, Phone, NID
Rootkit	English	69,324	Email, User name, Name, Age, Birthday

3.1 Our datasets

Our evaluation builds on ten large real-world password datasets (see Table 3), including five from English sites and five from Chinese sites. They were hacked by attackers or leaked by insiders, and disclosed publicly on the Internet, and some of them have been used in trawling password models [13, 19, 21]. Rootkit initially contains 71,228 passwords hashed in MD5, and we recover 97.46% of them by using our TarGuess-IV and various trawling guessing models [21, 30] in one week. In total, these datasets consist of 95.83 million plain-text passwords and cover various popular web services. The role of each dataset will be specified in Sec. 5.

In particular, two of these ten password datasets contain various types of PII as shown in Table 4. Besides, we further employ two auxiliary PII datasets, aiming to augment the password datasets by matching the email address to facilitate a more comprehensive understanding of the role of PII in user-chosen passwords. While most of the PII attributes in Chinese PII-associated datasets are available, 17.90% of names and 54.04% of birthdays in Rootkit are null. These missing attributes may hinder the effectiveness of targeted attacks against Rootkit users. To the best of knowledge, *our corpus is the largest and most diversified ever collected for evaluating the security threat of targeted online guessing*.

3.2 Popular passwords

Table 5 shows how often users from different services choose popular passwords. It is disturbing that 0.79%∼10.44% of user-chosen passwords can be guessed by just using the top 10 passwords. Generally, top Chinese passwords are more concentrated than English ones [34], which may imply that the former would be

Table 5: Top-10 most popular passwords of each service

Rank	Dodonew	CSDN	126	12306	Rockyou	000webhost	Xato	Yahoo	Rootkit
1	123456	123456789	123456	123456	123456	abc123	123456	123456	123456
2	a123456	12345678	123456789	a123456	12345	123456a	password	password	password
3	123456789	11111111	111111	**5201314**	123456789	12qw23we	12345678	welcome	**rootkit**
4	111111	**dearbook**	password	123456a	password	123abc	qwerty	ninja	111111
5	**5201314**	00000000	000000	111111	**iloveyou**	a123456	123456789	abc123	12345678
6	123123	123123123	123123	**woaini1314**	**princess**	123qwe	12345	123456789	qwerty
7	a321654	1234567890	12345678	123123	1234567	secret666	1234	12345678	123456789
8	12345	**88888888**	**5201314**	000000	**rockyou**	YfDbUfNjH10305070†	111111	sunshine	123123
9	000000	111111111	**18881888**	qq123456	12345678	asd123	1234567	**princess**	qwertyui
10	123456a	147258369	1234567	1qaz2wsx	abc123	qwerty123	dragon	qwerty	12345
% of top-10	3.28%	10.44%	3.52%	1.28%	2.05%	0.79%	1.46%	1.01%	3.94%

†The letter-part (i.e., `YfDbUfNjH`) can be mapped to a Russian word which means "navigator". Why it is so popular is beyond our comprehension.

more prone to online guessing. While most of the top Chinese passwords are only made of simple digits, popular English ones tend to be meaningful letter strings or keyboard patterns. Love plays an important role — `iloveyou` and `princess` are among the top-10 lists of two English sites, while `5201314` and `woaini1314`, both of which sound as "I love you forever and ever" in Chinese, are among the top-10 lists of three Chinese sites. Other factors such as culture (see `18881888`) and site name (see `rockyou` and `rootkit`) also show their impacts on password creation.

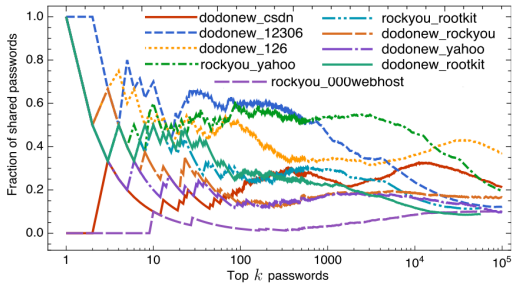


Figure 3: Fraction of PWs shared between two sites.

Fig. 3 illustrates the fraction of top-k passwords shared between two different services with varied thresholds of k. Generally, the fraction of shared passwords from the same language is substantially higher than that of shared passwords from different languages. In addition, the fraction of shared passwords between any two services is less than 60% at any threshold k larger than 10. This implies that both language and service play an important role in shaping users' top popular passwords.

Rockyou and 000webhost share significantly fewer common passwords than other pairs do. We examine these two datasets and find that 99.29% of 000webhost passwords include *both* letters and digits, indicating that this site enforces a password creation policy that requires passwords to include both letters and digits. This can also be corroborated by Table 5 where all top-10 000webhost passwords are composed of both letters and digits. Similarly, we find that CSDN requires passwords to be of length 8^{+}.

3.3 Password reuse

While users have to maintain probably several times as many password accounts as they did 10 years ago, human-memory capacity remains stable. As a result, users tend to cope by reusing passwords across different services [16,32]. Several empirical studies [5, 12] have explored the password reuse behaviors of English and European users, yet as far as we know, no empirical results have been reported about Chinese users, who reached 668 million by Dec., 2015 [11] and account for about 25% (and the largest fraction) of the world's Internet population.

To fill this gap, we intersect 12306 with Dodonew by matching email, and further eliminate the users with identical password pairs. This produces a new list `12306&Dodonew` with two non-identical sister passwords for each user. Similarly, we obtain two more intersected Chinese password lists and three intersected English lists as shown in Fig. 4. During the matching process, we find that 34.02%∼71.11% of Chinese users' sister password pairs are identical (and thus are eliminated), while these figures for English users are 6.25%∼21.96% (see Sec. 5.1). This suggests that our English users reuse less.

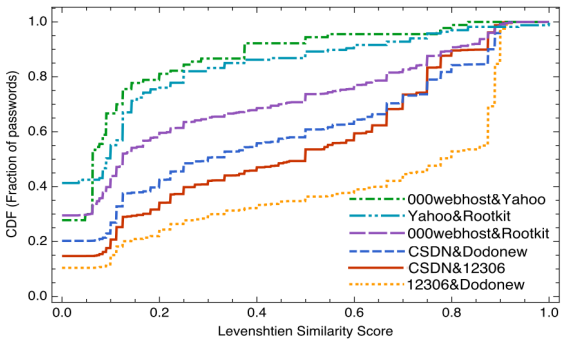


Figure 4: Using the Levenshtein-distance similarity metric to measure the similarity of two passwords chosen by the same user across different services. Results suggest that most users modify passwords in a non-trivial way.

We employ the widely accepted Levenshtein-distance metric to measure the similarity between two different passwords of a given user. Fig. 4 shows that, sister passwords of Chinese users generally have higher similarity than English users, implying that Chinese users modify passwords less complexly. About 30% of the non-identical Chinese password pairs have similarity scores in [0.7, 1.0], while this figure for our English password pairs is less than 20%. We also employ the longest-common-subsequence metric for measurement. Both metrics show similar results. Our results imply that *the majority of users modify passwords in a non-trivial approach, and it would be challenging to model such users's modification behaviors*.

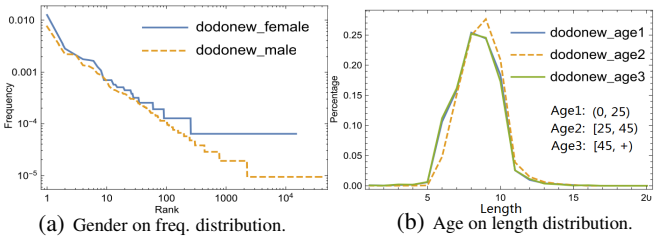
We have observed that our English users reuse less and modify passwords more complexly. A plausible reason for this observation is that the two english sites are not normal: Rootkit is a hacker forum and 000webhost is mainly used by web administrators. Therefore, *the users of both sites are likely to be more security-savvy than normal users*. Thus, the lists `Rootkit&000webhost`, `Rootkit&Yahoo` and `000webhost&Yahoo` will show more secure reuse behaviors than that of normal English/Chinese users.

In 2014, Das et al. [12] found that the fraction of identical sister PW pairs of normal English users is 43%, which roughly accords with our Chinese users yet 2∼6 times higher than our English users. They also showed that about 30% of their non-identical English PW pairs have similarity scores in [0.7, 1.0], well in accord with that of our Chinese users. Moreover, the survey results on password reuse behaviors of normal Chinese users [32] are largely consistent with the survey results on normal English users [12]. Both empirical and survey results suggest that *normal Chinese and English users have similar reuse behaviors, while our English users would be good representatives of security-savvy users*.

Table 6: Percentages of users building passwords with (and only with) their *own* heterogeneous personal information[†]

Typical usages of personal information (examples)	PII-Dodonew (161,510)		PII-126 (30,741)		PII-CSDN (77,439)		PII-12306 (129,303)		PII-Rootkit (69,330)		PII-Yahoo (214)		PII-000webhost(2,950)	
Full_name (`lei wang`, `john smith`)	4.68	0.82	3.00	1.32	4.85	1.81	5.02	1.13	1.38	0.75	2.34	1.87	2.44	1.32
Family_name (`wang`, `smith`)	11.15	0.01	6.16	0.00	9.75	0.00	11.23	0.00	2.28	0.78	4.67	1.87	3.73	1.46
Given_name (`lei`, `john`)	6.49	0.07	4.10	0.12	6.26	0.08	6.61	0.07	0.49	0.07	0.93	0.00	0.75	0.20
Abbr. full_name (`wl`, `lwang`, `js`, `jsmith`)	13.64	0.02	6.36	0.00	9.42	0.00	13.13	0.00	0.15	0.01	0.00	0.00	0.20	0.00
Birthday(`19820607`, `06071982`, `07061982`)	3.12	1.00	3.70	2.77	6.29	5.16	4.33	1.77	0.08	0.06	0.47	0.00	0.10	0.07
Year of bithday (`1982`)	8.92	0.00	8.84	0.01	11.37	0.00	10.78	0.00	0.75	0.01	1.40	0.00	1.12	0.00
Date of bithday (`0607`, `0706`)	8.32	0.00	10.48	0.02	11.84	0.00	10.03	0.00	0.44	0.01	0.47	0.00	0.58	0.00
Abbr. bithday(`198267`, `671982`, `761982`, `820607`, `060782`)	2.37	0.59	2.60	1.71	2.89	1.45	3.31	1.12	0.10	0.05	0.00	0.00	0.20	0.14
Family_name+bithday (`wang19820607`, `smith06071982`)	0.08	0.08	0.05	0.05	0.03	0.03	0.14	0.14	0.00	0.00	0.00	0.00	0.00	0.00
Family_name+Abbr. bithday①(`wang198267`, `smith671982`)	0.11	0.11	0.03	0.02	0.05	0.05	0.15	0.14	0.00	0.00	0.00	0.00	0.00	0.00
Family_name+Abbr. bithday②(`wang820607`, `smith060782`)	0.17	0.17	0.07	0.07	0.13	0.11	0.17	0.16	0.00	0.00	0.00	0.00	0.00	0.00
Family_name+year of birth (`wang1982`, `smith1982`)	0.55	0.22	0.20	0.07	0.22	0.07	0.64	0.25	0.01	0.00	0.00	0.00	0.00	0.00
Family_name+date of birth (`wang0607`, `smith0607`)	0.12	0.09	0.05	0.03	0.08	0.04	0.16	0.12	0.01	0.00	0.00	0.00	0.00	0.00
User name (`icemoon12`, `bluebirdz`)	1.54	1.14	0.54	0.38	0.61	0.43	1.96	1.32	1.59	0.92	2.34	1.40	2.20	1.32
Email_prefix (l0veu4ever@example.com)	5.07	3.07	2.52	1.60	4.35	2.48	3.03	1.82	0.77	0.44	4.21	1.87	1.32	0.78
Phone number (11-digit Chinese mobile number `13511336677`)	0.10	0.10	0.48	0.45	0.50	0.45	0.07	0.01	—	—	—	—	—	—
'a'+birthday(`a19820607`, `a06071982`, `a07061982`)	0.16	0.13	0.04	0.02	0.03	0.02	0.16	0.12	0.00	0.00	0.00	0.00	0.00	0.00
Full_name+1 (`wanglei1`, `johnsmith1`)	1.49	0.22	0.51	0.03	0.84	0.03	1.65	0.17	0.06	0.01	0.00	0.00	0.03	0.00

[†]All the decimals in the table use '%' as the unit. For instance, 4.68 in the top left corner means that 4.68% of the 161,510 PII-associated Dodonew users employ their full name to build passwords; 0.82 means that 0.82% of these 161,510 Dodonew users' passwords are *just* their full names.



(a) Gender on freq. distribution. (b) Age on length distribution.

Figure 5: Impact of type-2 PII on user password creation. Both gender and age show tangible impacts.

3.4 Password containing personal info

We show in Table 6 how often users employ their *own* PII to build passwords. Since some password lists have no PII (see Table 3), we correlate them with the PII datasets of the same language in Table 4 by matching email. As a result, seven PII-associated password lists are produced, and they are much more diversified than those in [20]. The sample size of each PII-associated dataset is shown in the first row of Table 4. As expected, highly heterogeneous PII becomes components of passwords, and users like to use names, birthdays and their variations. Particularly, a non-negligible fraction of users employ just their full names (0.75%∼1.87%) as passwords, and 1.00%∼5.16% of Chinese users use just their birthdays as passwords. Surprisingly, email and user name prevail in passwords of both user groups, ranging from 0.77% to 5.07% and from 0.54% to 2.34%, respectively. In comparison, English users exhibit a more secure behavior in PII usages, for our English users represent security-savvy ones.

Fig. 5 illustrates the impact of type-2 PII : (1) passwords of Dodonew female users are more concentrated; (2) passwords of Dodonew users in age≤24 and age≥46 have quite similar length distributions (pairwise χ^2 test, p-value= 0.009), while users in age 25∼45 are significantly different in length distributions (p-values<10^{-6}). Similar results are found in all other datasets.

Type-based PII matching. To achieve accuracy in PII recognition, we propose a *type-based* PII segment matching method: besides the traditional PCFG-based L, D, S tags [35], we employ a few kinds of PII tags (e.g., N for name and B for birthday), and each subscript number of our PII tags stands for a particular *sub-type* of one kind of PII considered. For instance, N_1 denotes the usage of family name (e.g., `li`), B_5 denotes the usage of year in birthday (e.g., `1982`) and so on. More details will be given in Sec. 4.1.

This is inherently different from the *length-based* PII matching method given in an independent study [20]. To avoid mismatching, only PII segments with $len \geq 3$ are considered in [20]. For instance, a match with any length 3^+ substring (e.g., `195`, `952`, `520`) of a birthday `19520123` will be considered as a birthday match. However, this introduces both under-estimations and over-estimations in PII matching. For example, the password `li.520` of a user named "Wei Li" with birthday `19520123`" will be tagged as $L_2S_1Birth_3$, because the family name `li` is of length <3. As 20% of the top-50 Chinese family names are with length <3 (e.g., `li`, `wu` and `he`), a large fraction of users' name usages may be *under-estimated* by [20]. For instance, 30,926 (23.9%) of the 13K 12306 users are with a family name $len \leq 2$, and 4,346 of these 30,926 users indeed use their family name in passwords, yet this fact cannot be captured in [20].

On the other hand, the segments (e.g., `123`, `520` and `201`) in top popular digital passwords (e.g., `123456`, `123456789`, `5201314`) would often coincide with user birthdays and phone numbers, leading to *over-estimations* of their usages in passwords. As we will show in Sec. 4.1, this length-based matching method also introduces a weakness in the guess generation process when performing cracking, while either increasing or decreasing their length threshold will not eliminate the problem.

Summary. Our PII-associated password corpus is so far the largest and most diversified ever collected for evaluating targeted online guessing. Particularly, it, for the first time, covers (security-savvy) English users. While users' three vulnerable behaviors might be potentially exploited to improve cracking, our results show that varied circumstances (e.g., language, service and policy), non-trivial transformation rules and highly heterogeneous PII all would make it a challenging task to *automate* this process, especially when given a limited guessing number (e.g., 100 by NIST [8, 18]).

4. TARGUESS: A FRAMEWORK FOR TARGETED ONLINE GUESSING

We now propose *TarGuess*, a practical framework that effectively addresses the realistic yet challenging problem of modeling various targeted online guessing scenarios.

As shown in Fig. 6, TarGuess consists of three phases (i.e. preparing, training and guessing). The design of the first and third phases is straightforward, and the main task lies in the second one. TarGuess captures four types of the most representative targeted online guessing scenarios, with each type based on varied kinds of personal information available to $\mathcal{A}$ (see Table 2): (i) only

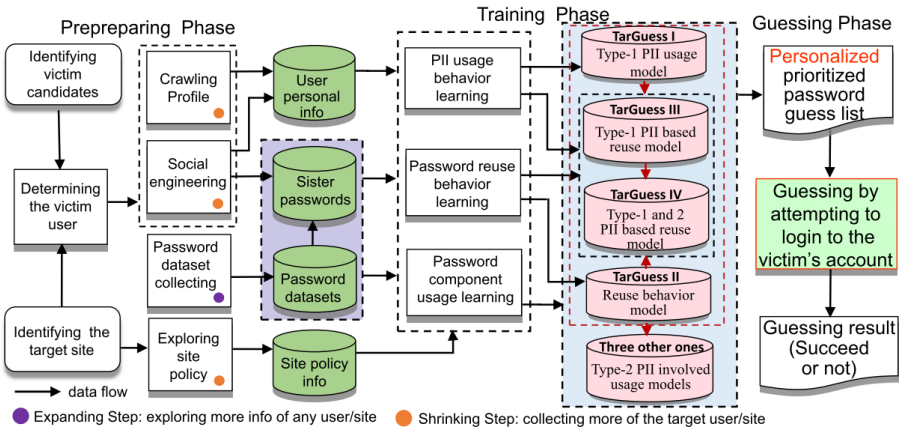


Figure 6: An architectural overview of the TarGuess.

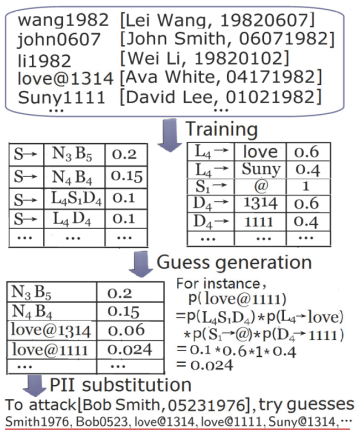


Figure 7: TarGuess-I: an illustration.

type-1 PII; (ii) one sister password; (iii) combination of i and ii; (iv) combination of iii and type-2 PII. To model these four scenarios, we suggest four guessing models (I∼IV) by leveraging a number of probabilistic techniques such as PCFG, Markov and Bayesian theory. We also show that, with TarGuess-I∼IV, the three remaining scenarios can also be well addressed.

4.1 TarGuess-I

TarGuess-I aims to online guess a user U's passwords by exploiting U' some type-1 PII (e.g., name and birthday, not gender). It builds on Weir et al.'s PCFG-based algorithm [35] which has been shown a great success in dealing with *trawling* guessing scenarios. In the training phase of [35], each password is seen as a combination of letter(L)-, digit(D)- and symbol(S)- segments. For example, `loveyou@1314` is parsed into the L-segment "lovey-ou", S-segment "@" and D-segment "1314", and its base structure is $L_7S_1D_4$; `wanglei@1982` is also parsed into $L_7S_1D_4$.

Our new algorithm. To capture PII semantics, besides the L, D, S tags as with PCFG [35], we introduce a number of *type-based* PII tags (e.g., N_1∼N_7 and B_1∼B_{10}). For a *type-based* PII tag, its subscript number stands for *a particular sub-type* of one kind of PII usages but not the *length* matched, as opposed to the L, D, S tags. For instance, N stands for name usages, while N_1 for the usage of full name, N_2 for the abbr. of full name (e.g., `lw` from "`lei wang`"), ···; B stands for birthday usages, B_1 for full birthday in YMD format (e.g., 19820607), B_2 for full birthday in MDY, ···. This gives rise to a PII-enriched context-free grammar $\mathcal{G}_I = (\mathcal{V}, \Sigma, \mathcal{S}, \mathcal{R})$, where:

1) $\mathcal{S} \in \mathcal{V}$ is the start symbol.
2) $\mathcal{V} = \{\mathcal{S}; \mathsf{L}, \mathsf{D}, \mathsf{S}; \mathsf{N}_1, \cdots, \mathsf{N}_7; \mathsf{B}_1, \cdots, \mathsf{B}_{10}; \mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3; \mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3; \mathsf{P}_1, \mathsf{P}_2; \mathsf{I}_1, \mathsf{I}_2, \mathsf{I}_3\}$ is a finite set of variables,[1] where: (1) N_1 and N_2 have been specified earlier, N_3 for family name (e.g., `wang`), N_4 for given name, N_5 for the 1st letter of the given name + family name (e.g., `1wang`), N_6 for last name+ the 1st letter of the given name (e.g., `wang1`), N_7 for family name with its 1st letter capitalized (e.g., `Wang`); (2) B_1 and B_2 have been specified earlier, B_3 stands for full birthday in DMY (e.g., `07061982`), B_4 for the date in birthday, B_5 for the year in birthday, B_6 for Year+Month (e.g., (e.g., `198206`), B_7 for Month+Year (e.g., `061982`), B_8 for the last two digits of year + date in MD format (e.g., `820607`), B_9 for date in MD format + the last two digits of year (e.g., `060782`), B_{10} for date in DM format + the last two digits of year (e.g., `070682`); (3) A stands for account name usages, A_1 for full account name (e.g., `icemoon12`), A_2 for the (first) letter-segment of account name (e.g., `icemoon`), A_3 for the (first) digital-segment of account name (e.g., `12`); (4) E stands for email prefix usages, E_1 for the full email prefix (e.g., `loveu1314` from `loveu1314@example.com`), E_2 for the first letter-segment of email prefix (e.g., `loveu`), E_3 for the first digital-segment of account name (e.g., `1314`); (5) P stands for mobile phone number usages, P_1 for the full number, P_2 for the first three digits, P_3 for last four digits; (6) I stands for the Chinese Notional Identification number, I_1 for the last 4 digits, I_2 for the first 3 digits, I_3 for the first 6 digits.[2]
3) Σ={95 printable ASCII codes, $Null$} is a finite set disjoint from $\mathcal{V}$ and contains all the terminals of $\mathcal{G}_I$.
4) $\mathcal{R}$ is a finite set of rules of the form $A \to \alpha$, with $A \in \mathcal{V}$ and $\alpha \in \mathcal{V} \cup \Sigma$ (see Fig. 7).

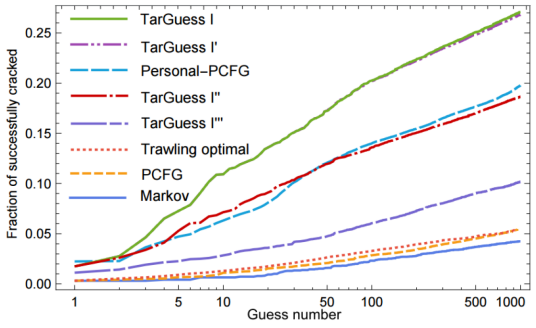


Figure 8: A comparison of TarGuess-I (and its variants) with Personal-PCFG [20], trained on the 50% of 12306 dataset and tested on the remaining 50%. Both TarGuess-I and Personal-PCFG [20] employ six kinds of the 12306 type-1 PII, while TarGuess-I′ eliminates phone # and NID, TarGuess-I″ further eliminates email and user name, and TarGuess-I‴ further eliminates birthday. TarGuess-I and I′ greatly outperform [20].

A probabilistic context-free grammar (PCFG) is a CFG that, for a specific left-hand side (LHS) variable (e.g., L_4), all the probabilities associated with its rules (e.g., $\mathsf{L}_4 \to$ `love` and $\mathsf{L}_4 \to$ `Suny`) can add up to 1 [35]. This condition is satisfied by $\mathcal{G}_I$.

[1]The number of PII-based variables and their specific definitions depend on the nature of the PII to be trained (e.g., phone number in US is 10 digits while 11 digits in China) and on the granularity the attacker $\mathcal{A}$ prefers (e.g., $\mathcal{A}$ may prefer 4 types of name usages but not 7 as we do). Here we give a typical definition for attacking Chinese users, and it is easily tailored to other user groups. Besides, it's feasible to generalize $\mathcal{G}_I$ by pre-defining a number of running-modes: Chinese-mode, US-mode, German-mode, etc. Then, $\mathcal{G}_I$ needs no customization, and the inputs to the algorithm TarGuess-I are the victim's PII-attributes plus a running-mode.

[2]Our definitions are gained by recursively adjusting and training on the PII-associated 126 dataset. To follow the basic machine learning principles, this dataset hereafter will never be used as the test set.

It is not difficult to see that the training phase of TarGuess-I can indeed automatically derive a PCFG. For more background, see [35]. Using this grammar, in the guess generation process (see Fig. 7) we can further derive: (1) all the terminals (e.g., `love@1314`) which are instantiated from base structures that only consist of L, D and S tags; and (2) all the pre-terminals (e.g., N_3B_5 and N_31234) which are intermediate guesses consisting of PII-based tags. The final guess candidates come from these terminals as well as from instantiating all the pre-terminals with the victim's PII.

Note that, to improve accuracy, we match using the longest-prefix rule and also only consider PII-segments with $len \geq 2$. For example, if `john06071982` matches John Smith's account name "john0607", it will be parsed into A_1B_5 using the longest-prefix rule, but not N_3B_2. In addition, we have only considered full MMDD dates in the definition of $B_1 \sim B_{10}$, yet many users tend to use an abbr. of date when possible (e.g., "198267" instead of "19820607"). Thus, when matching a birthday-based segment in the training phase, if an abbreviation happens, the tag related to the corresponding full segment will be counted by one; In the password generation process, both full and abbreviated date segments will be produced. For instance, both "john06071982" and "john671982" will be produced if the structure N_3B_2 is used for guess generation.

Our type-based PII tags are widely applicable. In the above, we have shown how to employ type-based PII tags to build a *semantic-aware* grammar using PCFG. Actually, they can also be employed by various other guessing algorithms (e.g., Markov-based [21] and TarGuess-II in Sec. 4.2) to build PII-enriched cracking algorithms. For instance, to build a PII-enriched Markov-based algorithm, we only need to incorporate the type-based PII tags $\{N_1, \cdots, N_7; B_1, \cdots, B_{10}; A_1, A_2, A_3; E_1, E_2, E_3; P_1, P_2; I_1, I_2, I_3\}$ into the alphabet Σ (e.g., $\Sigma = \{95$ printable ASCII characters$\}$ in [21]) of the Markov n-gram model, and then all operations for these type-based PII tags are the same with the original characters in Σ.

$\mathcal{G}_I$ is highly adaptive. On the one hand, whenever we want to consider new semantic usages (e.g., website name) or new type-1 PII usages (e.g., hobby), we can simply define new corresponding *type-based* tags (e.g., W for website name and H for hobby), the same as we define the N and B tags. In the training and guess generation phases, all the operations related to H and W are similar to that of N and B tags. It is "Plug-and-Play". On the other hand, even if TarGuess-I defines the B tag yet the training set has no birthday information, $\mathcal{G}_I$ still works properly—it will not parse passwords using B tags and simply parse birthday information in passwords using the D tag. That is, $\mathcal{G}_I$ is "self-dumping" and we do not need to specially eliminate the B-related tags in such cases.

An independent study. In a recent paper (published in April 2016), Li et al. [20] presented a *length-based* PII matching method. Our work is independent from theirs.

Besides the L, D, S tags in PCFG [35], Li et al. introduced six kinds of PII tags: N for name, B for birthdate, E for email, A for account (user) name, P for phone number, and I for NID. In contrast to our type-based approach, each PII-based tag in [20] uses a subscript number to denote the length len of the matched PII segment (only $len \geq 3$ are considered in [20]), following the same approach of the L, D and S tags as in PCFG [35]. As a result, in [20], `wanglei@1982` now is parsed into the N segment "wanglei", S segment "@" and B segment "1982", and its base structure will be $N_7S_1B_4$; "`loveyou@1314`" is parsed into $L_7S_1D_4$. Within 100 guesses, their "Personal-PCFG" algorithm cracked about 17% of their test dataset by using "perfect dictionaries".

However, we discovered a weakness in their Personal-PCFG. Their algorithm uses *length-based* tags, the same as Weir et al.'s algorithm [35]: it differentiates a segment's length, but is insensitive to a segment's subtype. For example, both `john@1982` and `wang@1982` will be parsed into $N_4S_1B_4$, because both "`wang`" and "`john`" are of length 4, despite the fact that "`wang`" is a family name while "`john`" is a given name. As shown in Fig. 9, this not only introduces both under-estimations and over-estimations in the training phase, but also leads to illogical situations in the guess generation phase. In the training phase, since "`wang`", "`smith`" and "`li`" are all family names and "`1982`" is a user's year of birth, the probability of N_4B_4 shall be 0.6 but not 0.4, the probability of L_2B_4 shall be 0 but not 0.2. In the guess generation phase, "`lee`" is of length 3, but there is no base structure N_3B_4 available, and thus the guess `lee1977` will be given a probability 0 and thus not be generated by Personal-PCFG.

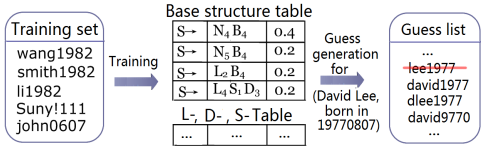


Figure 9: A weakness of Personal-PCFG [20].

According to our grammar $\mathcal{G}_I$, both `wang1982` and `li1982` will be parsed into the same base structure N_3B_5, `john1982` is parsed into N_4B_5, and thus the guess `lee1977` can be generated using the base structure N_3B_5 for "David Lee" born in 1977. This addresses the weakness in [20].

Evaluation. For fair comparison, we leverage the 12306 dataset as with [20] and follow their experimental setups (see Fig. 8) as closely as possible. The only exception is that, we do not use "the perfect (L-) dictionary" which is collected directly from the *test* set, because this not only introduces overfitting issues [21, 35] but also is unrealistic in practice. Instead, all our experiments directly learn the L- dictionary from the 12306 *training* set, a recommended practice in password cracking [13, 21]. Fig. 8 shows that, within $10\sim10^3$ guesses, our TarGuess-I *outperforms* Persona-PCFG [20] by 37.11%∼73.33%, and *outperforms* the three trawling online guessing algorithms [6, 21, 35] by at least 412%∼740%.

We have further examined to what extent each individual PII would impact TarGuess-I. As shown in Fig. 8, within 100 guesses, our TarGuess-I can successfully crack a 12306 user's password with an average chance of 20.26% when given this user's email, account name, name, birthday, phone number and NID. This figure is 20.18% when given email, account name, name and birthday. This figure is 13.61% when given name and birthday; this figure is 6.04% when given only name. Our results suggest that *email, account name, name and birthday would be very valuable for an online attacker*, while phone number and NID provide marginally improved success rates. Interestingly, Personal-PCFG [20] exploits two more PII attributes than TarGuess-I′ yet is much less effective, suggesting that simply incorporating more PII information into algorithms will not always yield more effectiveness.

Summary. We are the first to propose type-based PII tags for building a semantics-aware PCFG. Such PII tags can also be employed by other trawling algorithms (e.g., Markov n-grams [21]) to build targeted ones. Within 10^2 guesses, TarGuess-I has a success rate of 20.18% when given a 12306 user's email, user name, name and birthday. Within $10\sim10^3$ guesses, TarGuess-I outperforms Personal-PCFG [20] by cracking 37.11%∼73.33% more passwords. Particularly, TarGuess-I is highly adaptive.

4.2 TarGuess-II

TarGuess-I aims to online guess a user U's password PW_x at one service (e.g., CSDN) when given U's one sister password PW_s leaked from another service (e.g., Dodonew). This is a challenging task for two reasons. Firstly, the online guessing number allowed is small. Secondly, there are over a dozen transformation rules, such

as insert, delete, capitalize, leet (e.g., password→ passw0rd) and the synthesized ones (e.g., password→ Passw0rd), at users' choices to create PW_x by modifying PW_s. This process depends on the password creation policy, the value of service and each user's creativity. Moreover, even if $\mathcal{A}$ knows that U is likely to insert three digits, which digital sequence will be *exactly* used by U? It is worth noting that, users also love to use top popular passwords instead of modifying PW_s (see Sec. 3.2). The sole work by Das et al. [12] considers cross-site guessing by using U's one sister password, yet it is ineffective due to four reasons as we have shown in Sec. 1.

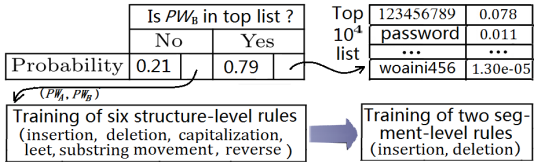


Figure 10: The training process of TarGuess-II

In this work, we prefer a data-driven approach. We use two lists of passwords as training sets, one leaked from a similar policy/service with the target site, the other is similar with that of PW_s, and look for the same users in these two lists by matching email. Further, the identical PW pairs are eliminated, and this creates a new list of non-identical sister PW pairs $\{PW_A, PW_B\}$. Then, we measure how PW_B is modified from PW_A, or whether PW_B is simply a popular password. To determine whether PW_B is a popular one, we build a top-10^4 list $\mathcal{L}$ for the target service (e.g., CSDN) from various leaked lists with consideration of policy and language, e.g., $\mathcal{L}$=$\{pw$ |len(pw)$\geq$ 8 and the value of $P_{csdn}(pw) * P_{126}(pw) * P_{dodonew}(pw)$ ranks top-$10^4\}$.

As shown in Fig. 10, in the training phase of TarGuess-II, first of all it determines whether $PW_B \in \mathcal{L}$ or not. If yes, the occurrence of $PW_B \in \mathcal{L}$ increases by one; If not, the pair (PW_A, PW_B) first goes through the structure-level training and then the segment-level training. In the structure-level process, TarGuess-II first parses passwords with L, D, S tags as with TarGuess-I. For instance, abc123 is parsed into $\mathsf{L}_3\mathsf{D}_3$. According to [12, 32], we consider six main types of structure-level transformation rules: insertion (e.g., $\mathsf{L}_3\mathsf{D}_3\rightarrow \mathsf{L}_3\mathsf{D}_3\mathsf{S}_2$), deletion (e.g., $\mathsf{L}_3\mathsf{D}_3\mathsf{S}_2$) $\rightarrow \mathsf{L}_3\mathsf{D}_3$), capitalization C, leet L, substring movement SM (e.g., abc123→123abc) and reverse R (e.g., abc123→321cba). There are two segment-level rules: insertion (e.g., $\underline{\mathsf{L}_3 : \text{abc}} \rightarrow \underline{\mathsf{L}_4 : \text{abcd}}$) and deletion (e.g., $\underline{\mathsf{L}_3 : \text{abc}} \rightarrow \underline{\mathsf{L}_2 : \text{bc}}$). For each of these eight types of rules, there exist a number of sub-types and we consider the most common ones. More specifically, for both levels of insertion (see Fig. 11), there are tail insertion ti (resp. ti') and head insertion hi (resp. hi'); For both levels of deletion, there are tail deletion td(resp. td') and head deletion hd(resp. hd'); For capitalization, there are four types as in Table 7; For leet, we consider 5 sub-types as in Table 8; For reverse, we consider 2 sub-types as in Table 10. Note that, a combination of our insertion and deletion operations can transform abc123 to abc!123, achieving the middle insertion.

The first step of the structure-level training is to employ the Levenshtein-distance (LD) algorithm (with only insertion and deletion enabled) to measure the similarity score d_1=LD(PW_A, PW_B) between the pair PW_A and PW_B. Then, we use each structure-level rule (except for insertion and deletion) in the C, L, R, SM order to obtain PW'_A based on PW_A, when considering that their popularity order is C>L>SM>R [12, 32] and that the rule R would result in a more drastic change than SM. Upon each rule, we compute d_2=LD(PW'_A, PW_B). If d_2>d_1, such a rule is called a "live" one, then PW_A is updated to PW'_A, and the occurrence of the corresponding rule (see Tables 7 to 10) increases by one. Then, we execute the next rule on PW'_A to produce PW''_A, and compute d_3=LD(PW''_A, PW_B). If d_3>d_2, this rule is "live" and counted.

Upon all these live rules, assume PW'''_A will be created from the original PW_A. To avoid futile transformations to dilute the effective ones, we require that if LD(PW'''_A, PW_B) is smaller than a predefined threshold (e.g., 0.5 as suggested in this work), then all these "live" rules are *un*-counted, and the training process switches to the *next* password pair in the training set. Otherwise, both PW'''_A and PW_B are parsed with L, D and S tags to be, e.g., $\mathsf{L}_4\mathsf{D}_3\mathsf{S}_2$ and $\mathsf{L}_6\mathsf{S}_1$. Since we do not consider the length of a PW segment in the structure-level, $\mathsf{L}_4\mathsf{D}_3\mathsf{S}_2$ and $\mathsf{L}_6\mathsf{S}_1$ will be seen as LDS and LS. Now we use the LD metric to compute d_3=LD("LDS", "LS") and meanwhile, the LD algorithm returns a LD edit route which records how to arrive "LS" from "LDS": first use the rule td on the S_2-segment, then use the rule td on the D_3-segment, and finally use the rule ta on the S_1-segment, producing PW''''_A with a base structure $\mathsf{L}_4\mathsf{S}_1$. Accordingly, the occurrence of all the corresponding items in Fig. 11(a) is updated.

Now we come to the segment-level training phase, and the focus is in the inner of the L-, D- or S- segment of a password. For PW''''_A (whose structure is $\mathsf{L}_4\mathsf{S}_1$) and PW_B ($\mathsf{L}_6\mathsf{S}_1$), we use the LD metric to measure the similarity of their L-segments. As with the structure-level training, the LD metric is used to update the occurrence of all the corresponding rules in Fig. 11(b). In our experiments, we find that the probabilities in the right-most row of Fig. 11(b) are better by computing using Markov n-grams [21] which are trained on a million-sized large password list, than by using the training as stated above. This is mainly because the size of the non-identical PW pairs in our training sets is only moderate and may lead to the sparsity issue. Fortunately, Markov n-gram model trained on million-sized PW lists can overcome this issue.

Our above two training phases give rise to a password-reuse based context-free grammar $\mathcal{G}_{II} = (\mathcal{V}, \Sigma, \mathcal{S}, \mathcal{R})$, where:

1) $\mathcal{V} = \{\mathcal{S}; \mathsf{L}, \mathsf{D}, \mathsf{S}; L, R, C, SM; ti, td, hi, hd; ti', td', hi', hd'\}$ is a finite set of variables.
2) $\mathcal{S} \in \mathcal{V}$ is the start symbol.
3) Σ={95 printable ASCII codes; C_1,$\cdots$,C_4; R_1,R_2; L_1, $\cdots$, L_5; Yes, No; No′} is a finite set disjoint from $\mathcal{V}$.
4) $\mathcal{R}$ is a finite set of rules of the form $A \rightarrow \alpha$, with $A \in \mathcal{V}$ and $\alpha \in \mathcal{V} \cup \Sigma$ (see Fig. 11 and Tables 7 to 10).

Note that, $\mathcal{G}_{II}$ is a probabilistic context-free grammar due to the fact that, for a specific left-hand side (LHS) variable (e.g., $R \rightarrow$) of $\mathcal{G}_{II}$, all the probabilities associated with its rules (e.g., $R \rightarrow$No, $R \rightarrow R_1$ and $R \rightarrow R_2$) can add up to 1. Using $\mathcal{G}_{II}$, in the guess generation phase we can create a list of guesses with possibilities. For instance, when given password, Pr("Pa$$word123")= Pr($\mathcal{S} \rightarrow \mathsf{L}_8$)* Pr($\mathsf{L}_8 \rightarrow$ ti)* Pr($ti \rightarrow \mathsf{D}_3$)* Pr($\mathsf{D}_3 \rightarrow$123) *$P(C \rightarrow C_1)$ * Pr($L \rightarrow L_2$) * Pr($L \rightarrow L_2$) * Pr($R \rightarrow$No)* Pr($SM \rightarrow$No) =1 * 0.1 * 0.15 * 0.08 * 0.03 * 0.01 * 0.01 * 0.97 * 0.97 = 3.39 * 10^{-9}, where the related probabilities are referred to Tables 7 to 10. Then, all the probabilities of guesses generated by $\mathcal{G}_{II}$ should be multiplied by the factor α. This α represents the fraction of users who do *not* choose top passwords (e.g., 0.21 in Fig. 10). Then, the probability of each password in the top-10^4 list are multiplied by $1 - \alpha$. Finally, these two probability-associated lists are merged and sorted in decreasing order, and then we select the top k (e.g., k=10^3) as the final guess candidates.

In Fig. 12, we provide a comparison of TarGuess-II with Das et al.'s algorithm [12]. These two algorithms are comparable because they employ the same personal information of the victim. When given a user U's Dodonew password, within 100 guesses, Das et al.'s algorithm [12] gained a success rate of 8.98% against U's CSDN account, while the figure for TarGuess-II is 20.19%, reaching

Table 7: Training of capitalization C (C_1: Cap. all; C_2: Cap. the 1st letter; C_3: Lower all; C_4: Lower 1st)

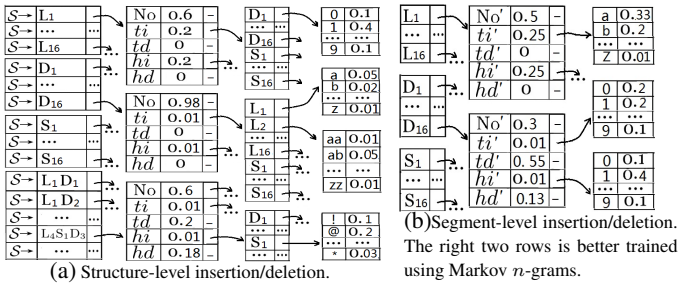
	No	C_1	C_2	C_3	C_4
Probability	0.95	0.01	0.03	0.003	0.007

Table 9: Training of substring movement

	substring moved SM	
	Yes	No
Prob.	0.03	0.97

Table 10: Training of reverse operation R (R_1: Reverse all; R_2: Reverse each segment)

	No	R_1	R_2
Probability	0.97	0.02	0.01



(a) Structure-level insertion/deletion.

(b)Segment-level insertion/deletion. The right two rows is better trained using Markov n-grams.

Figure 11: Training of two levels of insertion and deletion. As over 99% of passwords are with $len \leq$16 [21], only segments with $len \leq$16 are considered by us. The right-most two rows in Fig. 11(a) is better trained by using PCFG [35] on a million-sized password list.

a 124.83% improvement. In a series of 10 experiments in Sec. 5, under the same personal information and within 100 guesses, TarGuess-II outperforms their algorithm [12] by 8.12%∼300% (avg. 111.06%). One may conjecture that the two variations of TarGuess-II employ more personal information than one sister PW, and thus they are more powerful. In what follows, we, for the first time, provide more than anecdotal evidence to back this conjecture.

4.3 TarGuess-III

TarGuess-III aims to online guess a user U's passwords by exploiting U's one sister password *as well as* some PII. This is realistic: if the attacker $\mathcal{A}$ wants to target U and knows U's one sister password, it is likely that $\mathcal{A}$ can also obtain some PII (e.g., email, name) about U. As far as we know, no public literature has ever paid attention to this kind of attacking scenario. Here we mainly consider type-1 PII (e.g., name and birthday), while type-2 PII (e.g., gender and age) will be dealt with in Sec. 4.4.

Given a limited number of guesses, more information available to TarGuess-III generally means more messy things to be considered and thus more challenges to be addressed. Suppose $\mathcal{A}$ wants to target Alice Smith's account at eBay which requires passwords to be of length 8^+, and knows Alice was born in 1978 and one of Alice's passwords `Alice1978Yahoo` was leaked from Yahoo. Given guesses `Alice1978eBay`, `Alice1978` and `12345678`, which one shall $\mathcal{A}$ try first? If Alice's leaked password is `123456`, will the choice vary? Answering this question necessitates an adaptive, PII-aware cross-site guessing model.

Fortunately, we find that TarGuess-III can fulfill this goal by introducing the PII-based tags (which we have proposed in $\mathcal{G}_I$ of TarGuess-I) into the grammar $\mathcal{G}_{II}$ of TarGuess-II. In this way, we can build a PII-enriched, password reuse-based grammar $\mathcal{G}_{III}$. More specifically, besides the L, D, S tags in $\mathcal{G}_{II}$, our grammar $\mathcal{G}_{III}$ further includes six types of PII usages as with $\mathcal{G}_I$, and adds a number of *type-based* PII tags (e.g., N_1∼N_7 and B_1∼B_{10} as shown in Sec. 4.1) into $\mathcal{V}$ of $\mathcal{G}_{II}$.

In the training phase, all the PII-based password segments (each of which is parsed with one kind of PII tag) only involve the six structure-level transformation rules as defined in $\mathcal{G}_{II}$, and all the other things in $\mathcal{G}_{III}$ remain the same with that of $\mathcal{G}_{II}$. In the guess generation phase, from $\mathcal{G}_{III}$ we derive: (1) all the terminals (e.g., `love@1314`) which are instantiated from base structures that only consist of L, D and S tags; and (2) all the pre-terminals (e.g., $\mathsf{N}_4\mathsf{B}_5$ and $\mathsf{N}_3$1314) which are intermediate guesses that consist of PII-based tags. For these intermediate guesses, we further instantiate them with the target user's PII.

Table 8: Training of the leet transformation rule L

	No	$L_1: a \leftrightarrow @$	$L_2: s \leftrightarrow \$$	$L_3: o \leftrightarrow 0$	$L_4: i \leftrightarrow 1$	$L_5: e \leftrightarrow 3$
Prob.	0.95	0.02	0.01	0.01	0.005	0.005

As with $\mathcal{G}_I$, our $\mathcal{G}_{III}$ is also highly adaptive. The reasons are similar with that of $\mathcal{G}_I$ (see Sec. 4.1). This means that a new semantic tag, namely W_1 for website name, can be easily incorporated into $\mathcal{G}_{III}$ as with these PII tags. Now, $\mathcal{G}_{III}$ can parse `Alice1978Yahoo` into the structure $\mathsf{N}_4\mathsf{B}_5\mathsf{W}_1$, and the guess `Alice1978eBay` can be generated with the highest probability, because *no* transformation rules will be involved in the process from $\mathsf{N}_4\mathsf{B}_5\mathsf{W}_1$ to `Alice1978eBay`.

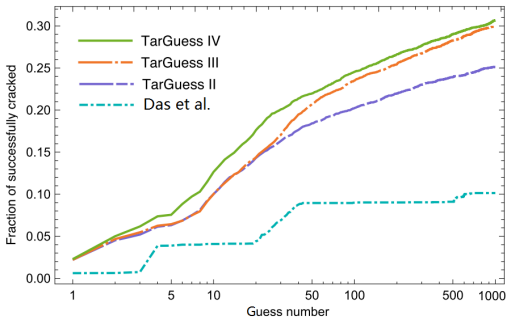


Figure 12: A comparison of TarGuess II∼IV and Das et al.'s algorithm [12], trained on the 66,573 non-identical PW pairs of 126→CSDN and tested on the 30,8045 non-identical PW pairs of Dodonew→CSDN. Besides a sister password, TarGuess-III uses four types of 51job type-1 PII and TarGuess-IV further uses the gender information.

As shown in Fig. 12, even if U does not exactly reuse her Dodonew password for her CSDN account, TarGuess-III can still achieve a success rate of 23.48% when allowed to try only 100 guesses, being 16.3% more effective than TarGuess-II. Among these un-cracked PW pairs by TarGuess-III, over 80% are significantly different (with LD similarity scores<0.5).

4.4 TarGuess-IV

TarGuess-IV aims to online guess a user U's passwords by exploiting U's one sister password as well as both type-1 and type-2 PII. A major challenge is that, type-2 PII (e.g., gender) can not be directly measured using any PII tag-based PCFG grammars or Markov n-grams. We tackle this issue by proving a theorem and leveraging the Bayesian theory.

THEOREM 1. *Let pw denote the event that the password* `pw` *is selected by U for a service, pw' denote the event that* `pw'` *is selected by U for another service and was leaked, A_i ($i = 1, 2, \cdots, n$) denote one kind of user PII attributes, including both type-1 and type-2 ones. We have*

$$\Pr(pw|pw', A_1, A_2, \cdots, A_n) = \frac{\prod_{i=1}^{n} \Pr(pw|pw', A_i)}{\Pr(pw|pw')^{n-1}},$$

under the assumptions that $A_1, A_2, \cdots, A_n$ are mutually independent, and that they are also mutually independent under the events pw' and (pw, pw').[3]

Proof. Since $A_1, \cdots, A_n$ are assumed to be mutually independent, we have $\Pr(A_1, \cdots, A_n) = \prod_{i=1}^{n} \Pr(A_i)$. Since

[3]Note that, the assumptions in Theorem 1 do not contradict with the fact that $A_1, \cdots, A_n$ are *dependent* on the events pw' and (pw, pw').

they are also assumed to be mutually independent under the events pw' and (pw, pw'), thus $\Pr(A_1, \cdots, A_n|pw)$=$\prod_{i=1}^{n}\Pr(A_i|pw)$, $\Pr(A_1, \cdots, A_n|pw') = \prod_{i=1}^{n}\Pr(A_i|pw')$ and $\Pr(A_1, \cdots, A_n|pw, pw') = \prod_{i=1}^{n}\Pr(A_i|pw, pw')$. Now, we can derive:

$$\begin{aligned}
&\Pr(pw|pw', A_1, A_2, \cdots, A_n)\\
=&\frac{\Pr(pw, A_1, A_2, \cdots, A_n|pw')}{\Pr(A_1, A_2, \cdots, A_n|pw')}\\
=&\frac{\Pr(A_1, A_2, \cdots, A_n|pw, pw') \cdot \Pr(pw|pw')}{\prod_{i=1}^{n}\Pr(A_i|pw')}\\
=&\frac{\left(\prod_{i=1}^{n}\Pr(A_i|pw, pw')\right) \cdot \Pr(pw|pw')}{\prod_{i=1}^{n}\Pr(A_i|pw')}\\
=&\frac{\prod_{i=1}^{n}\left(\Pr(A_i|pw, pw') \cdot \Pr(pw|pw')\right)}{\left(\prod_{i=1}^{n}\Pr(A_i|pw')\right) \cdot \Pr(pw|pw')^{n-1}}\\
=&\frac{\prod_{i=1}^{n}\frac{\Pr(pw, A_i|pw')}{\Pr(A_i|pw')}}{\Pr(pw|pw')^{n-1}} = \frac{\prod_{i=1}^{n}\Pr(pw|pw', A_i)}{\Pr(pw|pw')^{n-1}}. \qquad \blacksquare
\end{aligned}$$

This theorem indicates that the problem of predicting a user U's $\mathtt{pw}'$ at one service, when given U's PII information A_1, A_2, $\cdots$, A_n and the sister password pw at another service, can be addressed by a "divide-and-conquer" approach. More specifically, we can first compute $\Pr(pw|pw', A_i)$ for each i and $\Pr(pw|pw')$, and then compute the final goal $\Pr(pw|pw', A_1, A_2, \cdots, A_n)$. Fortunately, $\Pr(pw|pw')$ can be computed using TarGuess-II, and $\Pr(pw|pw', A_i)$ can be obtained by using TarGuess-III when A_i is a type-1 PII. When there are 2^+ type-1 PII attributes to be considered (suppose A_l, A_m, A_n), they together first can be deemed as *one* virtual attribute (e.g., to be A'_l) in Theorem 1 and then be addressed simultaneously by running TarGuess-III. The only issue left is how to compute $\Pr(pw|pw', A_i)$ when A_i is a type-2 PII.

To address it, we introduce the Bayesian theory. Without loss of generality, assume A_k is a type-2 PII. First,

$$\begin{aligned}
\Pr(pw, pw', A_k) =& \Pr(pw, A_k|pw') \cdot \Pr(pw')\\
=& \Pr[(A_k|pw)|pw'] \cdot \Pr(pw|pw') \cdot \Pr(pw').
\end{aligned}$$

It is reasonable to approximate $\Pr[(A_k|pw)|pw']$ by $\Pr(A_k|pw)$. Consequently, we have:

$$\begin{aligned}
\Pr(pw|pw', A_k) =& \frac{\Pr(pw, A_k, pw')}{\Pr(pw', A_k)}\\
=& \frac{\Pr[(A_k|pw)|pw'] \cdot \Pr(pw|pw') \cdot \Pr(pw')}{\Pr(pw', A_k)} \qquad (1)\\
\approx& \Pr(pw|pw') \cdot \frac{\Pr(A_k|pw) \cdot \Pr(pw')}{\Pr(pw', A_k)},
\end{aligned}$$

where $\Pr(pw|pw')$ is called the "prior" in Bayesian theory, the factor $\Pr(A_k|pw) \cdot \Pr(pw')/\Pr(pw', A_k)$ represents the impact of (pw', A_k) on the probability of event (pw', pw).

As a result, $\Pr(pw|pw', A_k)$ can be fully computed: (1) $\Pr(pw|pw')$ can be computed using TarGuess-II; (2) $\Pr(pw', A_k)$ = c_1 is a constant, because the event (pw', A_k) is a known and fixed fact when we attack U, and the ordering of guesses do not need the exact value of c_1; (3) $\Pr(pw')$=c_2 is, similarly, a constant since the event pw' is a known and fixed fact when we attack U; and (4) $\Pr(A_k|pw)$ can be computed by counting the training set—the password pw is selected by what fraction of users that are with an attribute A_k. When the training data is sufficiently large (e.g., $>10^6$), $\Pr(A_k|pw)$ can be obtained by direct counting. Otherwise, smoothing techniques (e.g., Laplace and Good-Turing) shall be used to overcome the sparsity issue to assure accuracy.

We note that some PII attributes are inherently *dependent* between each other (e.g., birthday vs. age, and first name vs. gender). Fortunately, since the majority of PII attributes (see Table 1) are mutually *independent*, the practicality of Theorem 1 will not be affected much. This is especially true when many attributes are simultaneously exploited. We observe that, even if TarGuess-III only employs birthday and TarGuess-IV employs one more PII (i.e., age), TarGuess-IV still performs better than TarGuess-III by now only adjusting the *non*-birthday-involved guesses using Eq. 1.

As shown in Fig. 12, by exploiting an additional PII (i.e., gender), TarGuess-IV can achieve improvements over TarGuess-III by 4.38%∼18.19% within 10∼10^3 guesses, reaching a success rate of 24.51% with 10^2 guesses and 30.66% with 10^3 guesses, respectively. This indicates that *type-2 PII, which, as far as we know, has never been considered in the literature of password cracking, is indeed valuable for* $\mathcal{A}$.

4.5 Dealing with other attacking scenarios

As mentioned in Table 2, seven scenarios can be resulted from the various combinations of the 3 types of personal info that we focus in this work. This means that, beyond the four most representative scenarios #1∼#4 that we have considered above, three other ones remain: #5 (type-2 PII), #6 (type-1 PII + type-2 PII) and #7 (1 sister PW + type-2 PII). Besides, there are scenarios involving 2^+ sister PWs: #8 (2^+sister PWs) and #9 (2^+sister PWs+some PII).

When A_k is a type-2 PII, it is natural to derive from Eq. 1 that:

$$\Pr(pw|A_k) \approx \Pr(pw) \cdot \frac{\Pr(A_k|pw)}{\Pr(A_k)}, \qquad (2)$$

where $\Pr(A_k) = c_3$ is a constant, and both $\Pr(pw)$ and $\Pr(A_k|pw)$ can be obtained by counting the training set, as discussed in Sec. 4.4. Eq. 2 well addresses Scenarios #5.

To tackle Scenario #6, we need to develop a new formulation. From Theorem 1, we can derive that

$$\Pr(pw|A_1, A_2, \cdots, A_n) = \frac{\prod_{i=1}^{n}\Pr(pw|A_i)}{\Pr(pw)^{n-1}}, \qquad (3)$$

where $\Pr(pw|A_i)$ can be obtained by using TarGuess-I when A_i is a type-1 PII, or be obtained by using Eq. 2 when A_i is a type-2 PII. This addresses Scenario #6.

As our Theorem 1 is suitable for both type-1 and type-2 PII, Scenario #7 can be readily tackled by first using TarGuess-II to generate a list of guesses and then adjusting the probabilities of the guesses according to Eq. 1.

Scenarios #8 and #9 cannot be readily addressed using the models proposed above. A simple approach to tackle them is to employ our TarGuess in a repeated manner, yet this is not optimal and we leave these two scenarios for future work. Still, as we have shown in Sec. 2, only a marginal fraction of users have leaked two or more passwords, and thus these two scenarios are far less common than the seven targeted guessing scenarios we have addressed.

Summary. We have designed a series of sound probabilistic models for targeted online guessing, with each characterizing one of the seven types of attacking scenarios. Our TarGuess-I and II significantly outperform the related algorithms [12, 20], while TarGuess-III and IV, for the first time, tackle the realistic issues of *combining* users' leaked passwords and PII to facilitate online guessing. Based on TarGuess-I∼IV, we further show how to address the three remaining scenarios. Extensive experiments in the following section further demonstrate the effectiveness of our TarGuess-I∼IV.

5. EXPERIMENTS

We now describe our experimental setups and comparatively evaluate TarGuess-I∼IV with five leading algorithms.

5.1 Experiment setup

Among the nine algorithms to be evaluated, three (i.e., Markov [21], PCFG [35] and Trawling optimal [6]) only need some training passwords, four (i.e., Das et al.'s algorithm [12], TarGuess-II∼IV) work on password pairs of the same user, and four (i.e., Personal-PCFG [20], TarGuess-I, III and IV) involve various types of user

Table 11: Training and test settings for each attacking scenario under 9 algorithms†

Experimental scenario*	Training set(s), with policy and language consistent with the test set								Test set
	PCFG	Markov	Tra. opt.	Personal-PCFG	TarGuess-I	Das et al.	TarGuess-II‡	TarGuess-III, IV‡	(size; service)
#1: 12306→Dodo	126	126	Dodo	PII-12306	PII-12306	126	12306, 126	12306, PII-126	49,775; Dodo
#2: 12306→CSDN	8^+Dodo	8^+Dodo	CSDN	8^+PII-Dodo	8^+PII-Dodo	8^+Dodo	12306, 8^+Dodo	12306, 8^+PII-Dodo	12,635; CSDN
#3: Dodo→CSDN	8^+Dodo	8^+Dodo	CSDN	8^+PII-Dodo	8^+PII-Dodo	8^+Dodo	Dodo, 8^+126	Dodo, 8^+PII-12306	5,997; CSDN
#4: Dodo→12306	Dodo	Dodo	CSDN	PII-Dodo	PII-Dodo	Dodo	Dodo, 126	PII-Dodo, 126	49,775; 12306
#5: CSDN→12306	Dodo	Dodo	12306	PII-Dodo	PII-Dodo	Dodo	CSDN, Dodo	CSDN, PII-Dodo	12,635; 12306
#6: CSDN→Dodo	126	126	Dodo	PII-12306	PII-12306	126	CSDN, 126	CSDN, PII-126	5,997; Dodo
#7: Rootkit→Yahoo	Rockyou	Rockyou	Yahoo	PII-Rootkit	PII-Rootkit	Rockyou	Rootkit, Xato	Rootkit, PII-Xato	214; Yahoo
#8: Rootkit→000web	L+D Rockyou	L+D Rockyou	000web	L+D PII-Rootkit	L+D PII-Rootkit	L+D Rockyou	Rootkit, L+D Xato	Rootkit, L+D PII-Xato	2,949; 000web
#9: 000web→Rootkit	Rockyou	Rockyou	Rootkit	PII-Xato	PII-Xato	Rockyou	000web, Xato	000web, PII-Xato	2,949; Rootkit
#10: Yahoo→Rootkit	Rockyou	Rockyou	Rootkit	PII-Xato	PII-Xato	Rockyou	Yahoo, Xato	Yahoo, PII-Xato	214; Rootkit

†Tra. opt.=Trawling optimal; Dodo=Dodonew; 000web=000webhost; 8^+=$len \geq 8$; PII-X=the PII-associated list X; L+D=Passwords with both letters and digits.
*$A \rightarrow B$ means that: (1) for the four password-reuse-based algorithms (i.e., Das et al.'s algorithm [12], TarGuess-II∼IV), a user U's password at service A can be used by $\mathcal{A}$ to help attack U's account at service B; and (2) for the other five algorithms, U's password at A is not involved, and only U's password at B is used as the target. Note that, every user's passwords in both A and B now have been associated with PII (see Tables 12 and 13) to facilitate the four PII-based algorithms.
‡When training TarGuess-II∼IV, U's one sister password comes from the 1st dataset, and $\mathcal{A}$ uses it to guess U's password from the 2nd dataset.

personal info. However, only two of our original datasets (i.e., 12306 and Rootkit) are associated with PII (see Table 3). Thus, as mentioned in Sec. 3.4, we build PII-Dodonew with size 161,510 by matching Dodonew with 51job and 12306 using email address, and PII-000webhost with size 2,950 by matching 000webhost with Rootkit. Matching by email ensures that *all our PII-associated English passwords are created by Rootkit hackers, who well represent security-savvy users*. Since Rockyou does not contain email or user name, we further match Xato with Rootkit to obtain 15,304 PII-associated Xato passwords to supplement Rockyou.

As shown in Table 12, we further build three lists of *password pairs* for Chinese users by matching Dodonew and CSDN with the two PII-associated Chinese password lists using email address. For instance, the list Dodonew↔12306 has a total of 49,775 password pairs, of which 14,380 pairs are with non-identical passwords. Similarly, we build three lists of password pairs for English users (see Table 13), but eliminate one of them (i.e., Yahoo↔000webhost) because the limited size of test set (i.e., 96) would make it impossible to reflect the true nature of an algorithm. These five pair-wised lists lead to ten experimental scenarios in Table 11.

Table 12: Basic information of the matched Chinese datasets

Original dataset	PII-12306(129,303)		PII-Dodonew(161,510)	
	Total	Non-identical(%)	Total	Non-identical(%)
Dodonew	49,775	14,380 (28.89%)	—	—
CSDN	12,635	5,538 (43.83%)	5,997	3,957(65.98%)

Table 13: Basic information of the matched English datasets

Original dataset	PII-Rootkit(69,330)		PII-000webhost(2,950)	
	Total	Non-identical(%)	Total	Non-identical(%)
000webhost	2,949	2,510 (85.11%)	—	—
Yahoo	214	167 (79.04%)	96	90 (93.75%)

To make our experiments as realistic as possible, our choices of the training set(s) for a given test set (attacking scenario) adhere to three rules: (1) they never come from the same service; (2) they are of the same language and password policy; and (3) the training set(s) shall be as large as possible. Rule 1 prevents our experiments from the overfitting issue, while rules 2 and 3 ensure the effectiveness of each algorithm. For fair comparison, we further make sure that all the 9 algorithms work on the same test set, and that for the same type of algorithms (e.g., TarGuess-I and [20]), their training sets exploit the same personal information.

5.2 Evaluation results

The guess number allowed is the most scarce resource when performing an online attack, while computational power and bandwith are not essential. For instance, in every of these ten experiments, the training phase can be completed on a common PC in less than 65.3s, while the generation of 1000 guesses for each user takes less than 2.1s. Thus, we use the guess-number-graph to evaluate the effectiveness of our four probabilistic algorithms with five leading algorithms (i.e., PCFG [35], Markov [21], Trawling optimal [6], Personal-PCFG [20] and Das et al.'s cross-site algorithm [12]).

Figs. 13(a)∼13(j) show that, under the same personal information available to $\mathcal{A}$ and within 100 guesses: (1) TarGuess-I drastically outperforms its foremost counterpart, Personal-PCFG [20], by 11.17%∼509% (avg. 84%); (2) TarGuess-II drastically outperforms Das et al.'s algorithm [12] by 8.12%∼300% (avg. 111.06%) when cracking non-identical password pairs; and (3) TarGuess-III and IV can gain a success rate of 73.09% when attacking Chinese normal users and of 31.61% when attacking English security-savvy users. As the number of guesses increases, in most cases the superiorities of TarGuess-I∼IV over their counterparts are further enhanced. Here we focus on cracking efficiency of cross-cite algorithms for non-identical PW pairs, because cracking non-identical pairs is their primary goal. As mentioned in Sec. 3.1, many PII attributes in English test sets (i.e., Rootkit and its matched lists) are missing, otherwise our cracking results would have been higher.

In particular, TarGuess-IV, for the first time, characterizes a very powerful yet realistic attacker who can launch cross-site guessing by exploiting a victim's one sister password as well as both type-1 and type-2 PII. As shown in Figs. 13(a)∼13(f), within 10 guesses, TarGuess-IV can gain success rates 45.49%∼85.33% (avg. 65.70%) against accounts of normal users at various web services; Within 10^2 guesses, the figures are 56.96%∼88.02% (avg. 73.08%); Within 10^3 guesses, the figures are 62.95%∼89.87% (avg. 77.32%). To achieve such high success rates, the state-of-the-art trawling algorithms [21, 30] need 10^{13} guesses per user and take several days by using high performance computers.

We discover that password strength against both targeted guessing and trawling guessing follows *a Zipf distribution*. The first few guesses are extremely effective, e.g., at 100 guesses, TarGuess-I can gain a success rate of 20% against every Chinese service. Yet, as the number of guesses increases, the success rate of each attempt decreases rapidly. Interestingly, we find that for each of the eight real-world algorithms (i.e., excluding the trawling optimal one [6]), the ratio f_n of the number of successfully cracked accounts *to* the number of guesses per account n can be well approximated by the Zipf's law [34]: $f_n = C * n^s$, where generally $s \in [0.15, 0.30]$ and $C \in [0.001, 0.01]$ are constants dependent on the test set. Such a relationship is a straight line when plotted on a log-log scale (see Fig. 13(k) which depicts the scenario of 12306→Dodonew). The three parallel layers in Fig. 13(k) just correspond to three kinds of guessing algorithms: trawling, targeted using PII, targeted using a sister PW. This diminishing principle has an important implication: $\mathcal{A}$ would stop at some point as her gains do not outweigh her costs, and there would be three such points corresponding to three different attacking strategies, yet existing guidelines [16, 18] only consider the trawling point.

When sister passwords are available, TarGuess-IV can reach a success rate of 77% against normal users with 100 guesses; Even when sister passwords are *un*available, TarGuess-I can still achieve

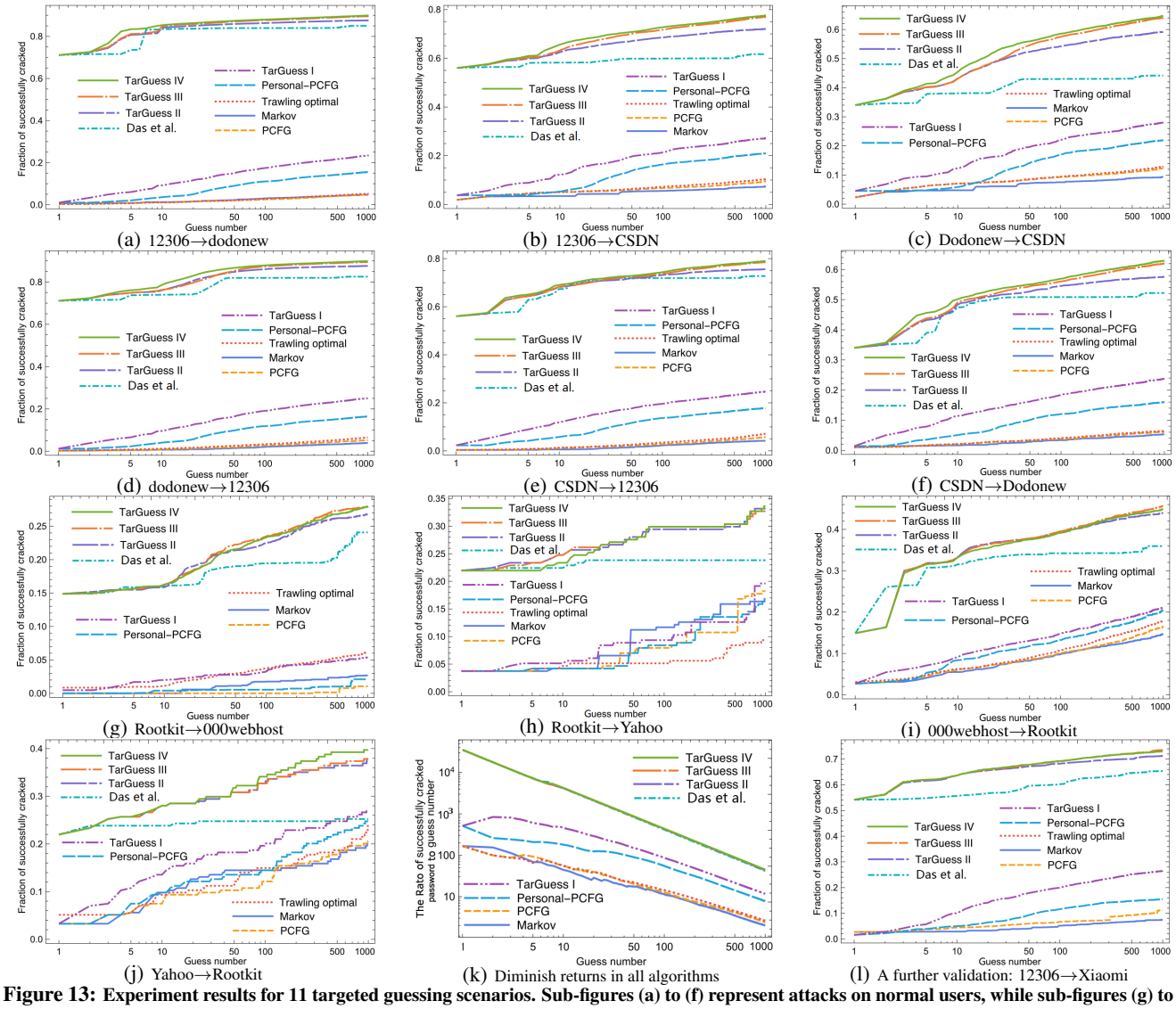


Figure 13: Experiment results for 11 targeted guessing scenarios. Sub-figures (a) to (f) represent attacks on normal users, while sub-figures (g) to (j) represent attacks on security-savvy users. Our TarGuess-I∼IV are highly effective. Sub-figure (k) depicts the diminish returns in 12306→Dodonew.

about 20% success rates against normal users with just 100 guesses, 25% with 10^3 guesses, and 50% with 10^6 guesses. This suggests that the majority of normal users' passwords are prone to a *small* number of targeted online guesses (e.g., 100 as allowed by NIST [8, 18]), invalidating the 2016 NIST claim that online guessing "can be *readily* addressed by throttling the rate of login attempts permitted" [18]. As normal users' passwords are even not strong enough to resist online guessing and still far away from the [10^6, 10^{14}] "online-offline chasm" [16], efforts directed towards resisting offline attacks (e.g., 10^{10} guesses or beyond [24, 28]) could have been better placed. Sites shall primarily focus on defending against online attacks and protecting the password hash file, while users shall mainly avoid the three bad behaviors examined in Sec. 3 to survive online guessing. In all, *targeted online guessing is a much more damaging threat to password security than trawling guessing and than the community (see [8, 18]) might have expected*. Our models will facilitate better evaluation of existing and future password policies (e.g., [9,24,28]), and they will also be helpful for forensic investigators to recover passwords in an offline manner.

5.3 A further validation

In the above experiments, our test sets are in plain-text. Would our algorithms be still effective when cracking "real accounts" about which little is known? We confirm this with a further experiment to crack Xiaomi cloud passwords, which are MD5 hashed with salt, leaked from the world's 3rd largest phone maker.

We attack Xiaomi as we crack Dodonew in Scenario #2 of Table 11. The test set contains 5,284 Xiaomi hashes obtained after matching the 8M Xiaomi dataset with the 130K 12306 dataset using email. As shown in Fig. 13(l), within $10\sim10^3$ guesses, TarGuess-I outperforms Personal-PCFG [20] by 70.58%∼119%; TarGuess-II outperforms Das et al.'s algorithm [12] by 73.66% ∼ 405%; TarGuess-III and IV can gain success rates of 63.61%∼73.56%. These results well accord with our above 10 experiments, especially the Chinese ones, suggesting the generality of our models.

6. CONCLUSION

We have presented the first systematic evaluation of the extent to which an online guessing attacker can gain advantages by exploiting various types of user personal info including leaked passwords and common PII. Our study is grounded on a framework that consists of 7 sound probabilistic models, with each addressing one typical attacking scenario. Particularly, TarGuess-I∼IV characterize the four most representative scenarios, and for the first time, the problem of how to model context-aware, semantic-enriched cross-site password guessing attacks has been well addressed.

Extensive experimental results show that TarGuess-I and II drastically outperform their foremost counterparts, and TarGuess-III and IV can gain success rates as high as 73% with just 100 guesses against normal users and 32% against security-savvy users. Our results suggest that the currently used security mechanisms would be largely ineffective against the targeted online guessing threat, and this threat has already become much more damaging than expected. We believe that the new algorithms and knowledge of effectiveness of targeted guessing models can shed light on both existing password practice and future password research.

Acknowledgment

The authors are grateful to the anonymous reviewers for their constructive comments. We also give our special thanks to Dinei Florêncio, Cormac Herley, Hugo Krawczyk, Haining Wang, Yue Li, Joseph Gardiner, Haibo Cheng and Qianchen Gu for their insightful suggestions and invaluable help. Ping Wang is the corresponding author. This research was in part supported by the National Natural Science Foundation of China (NSFC) under Grants Nos. 61472016 and 61472083, and by the National Key Research and Development Plan under Grant No. 2016YFB0800600.

7. REFERENCES

[1] *Nearly 80 percent of Internet users suffer identity leaks*, July 2015. http://bit.ly/2b9TEdn.

[2] *All Data Breach Sources*, May 2016. https://breachalarm.com/all-sources.

[3] *Turkey: personal data of 50 million citizens leaked online*, April 2016. http://bit.ly/1TPA4j4.

[4] *Amid Widespread Data Breaches in China*, Dec. 2011. http://www.techinasia.com/alipay-hack/.

[5] D. V. Bailey, M. Dürmuth, and C. Paar. Statistics on password re-use and adaptive strength for financial accounts. In *Proc. SCN 2014*, pages 218–235.

[6] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *Proc. IEEE S&P 2012*, pages 538–552.

[7] J. Bonneau, C. Herley, P. van Oorschot, and F. Stajano. Passwords and the evolution of imperfect authentication. *Commun. ACM*, 58(7):78–87, 2015.

[8] W. Burr, D. Dodson, R. Perlner, and et al. NIST SP800-63-2: Electronic authentication guideline. Technical report, NIST, Reston, VA, Aug. 2013.

[9] X. Carnavalet and M. Mannan. A large-scale evaluation of high-impact password strength meters. *ACM Trans. Inform. Syst. Secur.*, 18(1):1–32, 2015.

[10] A. Chaabane, G. Acs, M. A. Kaafar, et al. You are what you like! information leakage through users' interests. In *Proc. NDSS 2012*, pages 1–15.

[11] C. Custer. *China's Internet users zoom to 668 million*, Jan. 2016. http://www.apira.org/news.php?id=1736.

[12] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The tangled web of password reuse. In *Proc. NDSS 2014*.

[13] M. Dell'Amico and M. Filippone. Monte carlo strength evaluation: Fast and reliable password checking. In *Proc. ACM CCS 2015*, pages 158–169.

[14] M. Dürmuth, D. Freeman, and B. Biggio. Who are you? A statistical approach to measuring user authenticity. In *Proc. NDSS 2016*, pages 1–15.

[15] S. Egelman, A. Sotirakopoulos, K. Beznosov, and C. Herley. Does my password go up to eleven?: the impact of password meters on password selection. In *Proc. ACM CHI 2013*, pages 2379–2388.

[16] D. Florêncio, C. Herley, and P. van Oorschot. An administrator's guide to internet password research. In *Proc. USENIX LISA 2014*, pages 44–61.

[17] *Now it's easy to see if leaked passwords work on other sites*, July 2016. http://bit.ly/29AJANh.

[18] P. A. Grassi and J. L. Fenton. NIST SP800-63B: Digital authentication guideline. Technical report, NIST, Reston, VA, 2016. https://pages.nist.gov/800-63-3/sp800-63b.html.

[19] S. Ji, S. Yang, X. Hu, W. Han, Z. Li, and R. Beyah. Zero-sum password cracking game: A large-scale empirical study on the crackability, correlation, and security of passwords. *IEEE Trans. Depend. Secur. Comput.*, 2015. Doi: 10.1109/TDSC.2015.2481884.

[20] Y. Li, H. Wang, and K. Sun. A study of personal information in human-chosen passwords and its security implications. In *Proc. IEEE INFOCOM 2016*, pages 1–9.

[21] J. Ma, W. Yang, M. Luo, and N. Li. A study of probabilistic password models. In *Proc. IEEE S&P 2014*, pages 689–704.

[22] M. L. Mazurek, S. Komanduri, T. Vidas, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur. Measuring password guessability for an entire university. In *Proc. CCS 2013*, pages 173–186.

[23] E. McCallister, T. Grance, and K. Scarfone. NIST SP800-122: Guide to protecting the confidentiality of personally identifiable information (PII). Technical report, NIST, Reston, VA, April, 2010.

[24] W. Melicher, B. Ur, S. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. Cranor. Fast, lean and accurate: Modeling password guessability using neural networks. In *Proc. USENIX SEC 2016*, pages 1–17.

[25] A. Narayanan and V. Shmatikov. Fast dictionary attacks on passwords using time-space tradeoff. In *Proc. ACM CCS 2005*, pages 364–372.

[26] J. Onaolapo, E. Mariconti, and G. Stringhini. What happens after you are pwnd: Understanding the use of leaked account credentials in the wild. In *Proc. SIGCOMM IMC 2016*.

[27] *Four Years Later, Anthem Breached Again: Hackers Stole Credentials*, Feb. 2015. http://t.cn/RqWrMKC.

[28] R. Shay, S. Komanduri, A. Durity, and et al. Designing password policies for strength and usability. *ACM Trans. Inf. Syst. Secur.*, 18(4):1–34, 2016.

[29] *Senate Bill No. 1386: Personal information*, Sep. 2002. http://bit.ly/1WJIIpK.

[30] B. Ur, S. M. Segreti, L. Bauer, and et al. Measuring real-world accuracies and biases in modeling password guessability. In *USENIX SEC 2015*, pages 463–481.

[31] R. Veras, C. Collins, and J. Thorpe. On the semantic patterns of passwords and their security impact. In *Proc. NDSS 2014*.

[32] D. Wang, D. He, H. Cheng, and P. Wang. fuzzyPSM: A new password strength meter using fuzzy probabilistic context-free grammars. In *Proc. IEEE/IFIP DSN 2016*, pages 595–606. http://bit.ly/2ahJ8CO.

[33] D. Wang and P. Wang. The emperor's new password creation policies. In *Proc. ESORICS 2015*, pages 456–477.

[34] D. Wang and P. Wang. On the implications of Zipf's law in passwords. In *Proc. ESORICS 2016*, pages 111–131.

[35] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. In *Proc. IEEE S&P 2009*, pages 391–405.

[36] *This could be the iCloud flaw that led to celebrity photos being leaked*, Sep. 2014. http://bit.ly/Y5vnNc.

[37] Y. Zhang, F. Monrose, and M. Reiter. The security of modern password expiration:an algorithmic framework and empirical analysis. In *Proc. ACM CCS 2010*, pages 176–186.