PAPER Special Section on Information and Communication System Security

Comparative Analysis of Three Language Spheres: Are Linguistic and Cultural Differences Reflected in Password Selection Habits?*

Keika MORI[†a)], **Takuya WATANABE**[†,††b)], **Yunao ZHOU**[†c)], **Ayako AKIYAMA HASEGAWA**[††d)], ***Nonmembers***, **Mitsuaki AKIYAMA**[††e)], ***and*** **Tatsuya MORI**[†,†††f)], ***Members***

SUMMARY This work aims to determine the propensity of password creation through the lens of *language spheres*. To this end, we consider four different countries, each with a different culture/language: China/Chinese, United Kingdom (UK) and India/English, and Japan/Japanese. We first employ a user study to verify whether language and culture are reflected in password creation. We found that users in India, Japan, and the UK prefer to create their passwords from base words, and the kinds of words they are incorporated into passwords vary between countries. We then test whether the findings obtained through the user study are reflected in a corpus of leaked passwords. We found that users in China and Japan prefer dates, while users in India, Japan, and the UK prefer names. We also found that cultural words (e.g., "sakura" in Japan and "football" in the UK) are frequently used to create passwords. Finally, we demonstrate that the knowledge on the linguistic background of targeted users can be exploited to increase the speed of the password guessing process.

key words: *user authentication, password security, cross-cultural analysis*

1. Introduction

Despite having several security risks, such as cracking or massive breaches, passwords are still the primary authentication mechanism and are used in a diverse range of services because of their simplicity and user-friendliness. The proper use of a password generator/manager is a promising approach towards securing password-based authentication without sacrificing usability too severely. However, the majority of users today still rely on their brains to create and store their passwords, implying that background knowledge about a user could be used to attack their password efficiently.

There have been several prior studies that have analyzed large corpora of leaked passwords [2] with the aim of assessing the risks of password cracking. There have been other studies that aimed at performing in-depth analyses of password creation propensity through a user study approach [3]–[7]. While prior studies on human-generated passwords have focused on the characteristics of passwords created by English speakers, there have been few studies that focus on passwords created by non-native English speakers. While passwords are usually composed of alphanumeric letters**, many languages use other letters, such as Chinese characters, Korean Hangul, or Japanese Hiragana. We believe that such a linguistic difference as well as cultural difference may strongly affect the password creation processes and the resulting password properties. We also believe that such knowledge on the linguistic/cultural background of a target may help an attacker to speed up the password guessing process.

With the above in mind, we aim to understand the propensity of password creation through the lens of *language spheres*. Our research questions are as follows:

RQ1 Are linguistic and cultural differences reflected in users ' password habits?

RQ2 If so, do these differences allow attackers to crack passwords effectively?

To answer **RQ1**, we adopt a two-fold strategy. We first perform an online survey of users from four different cultural spheres—Chinese, Indian, Japanese, and English—and conduct an analysis on leaked passwords that seem to belong to each country. Regarding the online survey, to recruit the participants from each cultural sphere without introducing possible bias factors, we used four crowdsourcing services that operate in each respective country. Because we intend to highlight the characteristics of passwords created by people with different cultural backgrounds/native languages, our questionnaires were created in three languages to ensure the native languages reported are correct. For Indians, we recruited users who speak English and asked them to answer the questions in English. Note that this approach has some limitations, but it does provide a better opportunity to recruit participants in a specific language sphere. We carried out user surveys through crowdsourcing services and compared the propensity of password creation processes, such as the use of particular types of words and their languages, use of random letters, use of password generator, and the management strategies of Chinese, Indian, Japanese, and the UK

Manuscript received August 27, 2019.
Manuscript revised January 7, 2020.
Manuscript publicized April 10, 2020.

[†]The authors are with Waseda University, Tokyo, 169–8555 Japan.
[††]The authors are with NTT Secure Platform Laboratories, Musashino-shi, 180–8585 Japan.
[†††]The author is with NICT, Koganei-shi, 184–8795 Japan.

*Early version of this paper [1] was presented at EuroUSEC.

a) E-mail: keika@nsl.cs.waseda.ac.jp
b) E-mail: takuya.watanabe.yf@hco.ntt.co.jp
c) E-mail: zhouyunao@nsl.cs.waseda.ac.jp
d) E-mail: ayako.hasegawa.vg@hco.ntt.co.jp
e) E-mail: akiyama@ieee.org
f) E-mail: mori@nsl.cs.waseda.ac.jp

DOI: 10.1587/transinf.2019ICP0009

**Although we are aware of some exceptions on this assumption, we omit this issue due to the space limitations.

Table 1 Our findings and main contributions.

No.	Contribution	Corresponding RQ	Section
1	We analysed user-generated passwords from the viewpoint of linguistic and cultural differences. To this end, we adopted a unique approach – combining a user study and leaked password analysis.	–	–
2	Our online user survey revealed that for the three language spheres we studied, participants reported that more than 80% of users do not use a password generator/manager in their daily life, and 35–70% of the users make use of specific words or patterns of digits for their passwords. For each country, there were specific tendencies in word choice for creating passwords.	RQ1	Sect. 3
3	Our large-scale password analysis revealed that some of the characteristics we found in our user study can be observed in the leaked passwords, such as combining several words. On the other hand, we observed that some characteristics we found in the user study were not observed in the leaked password analysis, e.g., use of leet or reordering characters in a word.	RQ1	Sects. 4, 5
4	We demonstrated that knowledge on the linguistic/cultural background of a user can accelerate the password guessing process.	RQ2	Sect. 6

users.

Next, using more than 830 million leaked passwords collected from the three different sources of leaked password datasets, we tested whether the findings obtained through the user study were reflected in the corpus of leaked passwords. The leaked password datasets we obtained for this study contained pairs of email address and plain passwords. By applying domain name heuristics to the user email addresses, we extracted the passwords that are likely associated with users who belong to one of the three language spheres. To analyze the passwords of three language spheres, for each language, we compiled several dictionaries, including ones that contain generic words with lexical categories, specific dictionaries for person names, and patterns of digits such as dates of birth, telephone numbers, etc. As a collective, these dictionaries contain a huge volume of words, so we also developed a simple methodology that leverages multiple Bloom filters to count the frequencies of words in the password dataset in a memory-efficient manner.

To answer **RQ2**, we tested whether knowledge on the linguistic/cultural background of targeted users can be exploited to make the password guessing process faster. To this end, we leveraged the probabilistic context-free grammar (PCFG) as a modern password guessing algorithm. We changed the password corpus data to train the PCFG model and test how linguistic differences in the training data affect the password guessing speed.

Our findings and main contributions are summarized in Table 1.

The remainder of this paper is organized as follows: In Sect. 2, we review related work and compare it with our study. Section 3 describes the details of the user study we performed. Section 4 presents the methodology and results of the large-scale leaked password analysis. In Sect. 5, we examine whether the findings obtained through the user study are observed in the leaked password analysis. In Sect. 6, we show that the knowledge of the linguistic background of targeted users helps an attacker efficiently guess their passwords. Section 7 discusses the limitations of the work, possible extensions of the work, and future research directions. Section 8 concludes the work.

2. Related Work

This section reviews several related works. We first show several studies on the cross-cultural user surveys on security, which is closely related to our approach. Next, we present prior user studies on password habits. We then present several analytical studies on password habits. We compare these prior studies with ours to clearly highlight our contributions.

2.1 Cross-Cultural User Surveys on Security

Several studies have been conducted to analyze how the cultural differences are correlated with user behavior or attitude toward security. Harbach et al. [8] and Sawaya et al. [9] conducted user-based surveys in multiple countries. To this end, these two research groups attempted to translate their survey questions into the participants' native languages. Harbach et al. [8] aimed at investigating user attitude toward smartphone unlocking and they found that the level of protection of smartphone data was significantly different among various countries and Japanese participants tended to consider that their data on their smartphone is sensitive. Sawaya et al. [9] recruited participants from seven countries, i.e., China, France, Japan, Korea, Russia, the United Arab Emirates, and the United States. They investigated security behavior and various other factors such as security knowledge and self-confidence in security, and concluded that Asian participants, especially Japanese, tended to behave less securely. While their study was based solely on the online survey approach, we combined online surveys and leaked password data analysis; such a multiangle approach enabled us to obtain the in-depth insight for studying the research questions. Furthermore, we looked into the users' behavior related to passwords in detail.

2.2 User Studies on Password Creation

There have been several studies that have analyzed users'

password habits and choices through online studies or monitoring the behavior on their end devices [4]–[7]. These studies attempt to understand users' password habits/strategies through both surveys and experimental studies. Wash et al. [5] revealed that people often reuse passwords across different websites. Pearman et al. identified several intrinsic strategies people use when creating and reusing passwords [6]. Riley et al. [3] conducted a user survey to understand users' practices of password creation and storage. They asked participants about their habits on the Internet, real strategies to create a password, and practices they think are safe. They revealed that users did not employ the best practice they knew. Ur et al. [4] interviewed 49 participants about their password creation strategies. In their study, they asked participants to create passwords for three websites (banking, email, and news website). Not only did they identify users' misconceptions about strong passwords, but they also found that their thoughts on the value of each account were different from that those assumed in the security research community.

In these prior studies, authors recruited English speakers as the participants of their studies, primarily using US-based crowdsourcing services such as Amazon MTurk or recruiting university students. Although the participants may include non-native English speakers, the studies do not consider the linguistic/cultural differences of participants, assuming English is the primary language used by all participants. In our study, we shed light on users rooted in different cultures or countries and conducted our study across three different language spheres. Our comparative analysis unveiled that the differences of password habits and password creation strategies in those countries are statistically significant.

2.3 Analysis of Leaked Passwords

While large-scale data breaches, especially password leakages, have caused serious risks in terms of identity theft, ironically, leaked passwords have been used as an irreplaceable data source for password research and have contributed to password-security policies. In fact, analyzing leaked passwords is another promising channel to understand user's password habit.

In this regard, the closest work to ours is the work done by Li et al. [2]. They analyzed a large corpus of Chinese web passwords and reported that Chinese speakers prefer digits and include Pinyin which is a system to representing Chinese pronunciation with alphabets in their passwords. Zeng et al. [10] also investigated Chinese passwords. They studied the lexical sentiment in passwords and found that users tend to use positive words, especially words representing joy. While they also looked into a non-English password corpus, what distinguishes their study and ours is that our work is a *comparative* analysis among three language spheres, rather than being focused on the property of passwords for services used in a single language sphere. AlSabah et al. [11] analyzed passwords of users from different cultural/linguistic backgrounds (Arabs, Indian and Pakistani, Filipinos, and English speakers). They used datasets with rich meta-data (i.e., names, phone numbers, emails, addresses, recovery questions and answers) and found certain differences in passwords of users with different backgrounds. However, they did not conduct the user survey. User study is another way to understand users' behavior. We conducted both the leaked data analysis and the user survey to know it deeply.

Leaked password datasets are used in other lines of research. Thomas et al. [12] analyzed breached password datasets collected from various publicly available information sources such as paste sites, search indexes, public forms, and private forums. They found that the majority of breached passwords originated from private forums and 7–25% of stolen passwords matched a victim's valid Google account. Das et al. [13] analyzed 6,077 unique user passwords and found that 43–51% of users reused the same password across multiple sites.

2.4 Password Guessing

State-of-the-art password guessing approaches go beyond naive traditional techniques such as brute force guessing or dictionary attacks. Modern password guessing approaches leverage statistical methodologies such as Markov models or probabilistic context-free grammars (PCFG). Recently, several researchers have proposed using neural network models. Narayanan et al. was the first to propose Markov model-based password guessing [14], and Ma later studied it more comprehensively [15]. The advantage of a Markov model is that it works well for modeling language, i.e., it can predict the probability of the next character in a password based on the previously generated characters. Weir et al. produced PCFG to model the structures of passwords based on their probability distributions [16]. This represents passwords as word-mangling templates and terminals and generates guesses in highest probability order. It achieves an improvement over John the Ripper password cracker, ranging from 28–129% more passwords being cracked. The password guess generator of PCFG is open-source. Melicher was the first to use a long short-term memory (LSTM) neural network to extract password features from hidden semantics within passwords and make predictions [17].

As a reasonable choice of a modern password guessing approach, we adopt PCFG, primarily due to its performance and availability. We note that the aim of our study is not to propose a novel password guessing technique but to test whether the linguistic/cultural background of users can accelerate the process of password guessing.

3. Survey of Password Habits

In this section, we study how users create passwords through online surveys. We first present the survey design. Next, we show the descriptive statistics of participants. We then present the analysis of the password habits of participants.

3.1 Survey Design

We designed the length of the survey to take 10–15 minutes for each participant and the survey was conducted in July 2019. Before starting the survey, we clarified our purpose and the usage of the answers. We obtained informed consent from the participants. For those who agreed to participate in our experiment, we asked the following questions: demographic information, knowledge about password security, their ways of managing passwords, and habits of password creation.

Since our survey involves participants from four different countries and three languages, we designed our survey so that the difference in language will not affect the survey results. To this end, we used four different online survey systems widely used in each country. Our expectation is that users who primarily speak in their own language may prefer to read and answer the questions in that language. We show the English version of the questionnaire in Appendix.

We recruited participants from China, Japan, and the UK whose first language was Chinese, Japanese, or English, respectively. Regarding Indians, we recruited participants who speak English. Further, we recruited residents of China, India, Japan, and the UK. The participants were asked about their resident location in the questionnaires. Each Indian participant was offered 2.1 USD; the Japanese, 300 JPY; and the UK participants, 2.5 GBP. The payment was adjusted to be well above the minimum wage of each country. We could not adjust the payment for Chinese participants because the crowdsourcing platform automatically determines the price of work. To conform with the ethical considerations, we obtained informed consent from all the participants before the survey.

3.2 Descriptive Statistics of Participants

We received 315 responses from China through Sojump, 300 from India through Amazon Mechanical Turk, 300 from Japan through Lancers, and 301 from the UK through Prolific. The number of participants was adjusted based on the previous studies [18], [19]. We omitted invalid answers; e.g., inconsistent answers or answers from participants who do not live in appropriate countries. Finally, We analyzed 287, 254, 284, and 282 responses from China, India, Japan, and the UK, respectively. The demographics of the participants are listed in Table 2, and the description of devices they use is summarized in Table 3. We observe that the majority of the participants use a PC across the four countries. Table 4 summarizes the breakdown of participants who have a degree in computer science and/or information security. A majority of the participants (80–90%) in China, Japan, and the UK do not have these degrees, while approximately 90% of our participants in India do. Table 5 shows the breakdown of participants who have had opportunities to hear about the information about the risk of poorly created/managed passwords. For Japan and the UK, majority of the participants reported that they have not received such information, while majority of the participants from China and India reported that they have received such information.

Table 2 Demographics of the participants.

	Gender # participants	Age (Years) 18–19 / 20–29 / 30–39 / 40–49 / 50–59 / 60–
CN	F: 135 M: 151 O: 1	4 / 43 / 43 / 8 / 2 / 0 (%)
IN	F: 103 M: 151 O: 0	0 / 82 / 16 / 2 / 0 / 0 (%)
JP	F: 121 M: 154 O: 9	0 / 12 / 34 / 34 / 17 / 3 (%)
UK	F: 200 M: 82 O: 0	3 / 27 / 28 / 17 / 16 / 9 (%)

Table 3 What kind of computing devices do you use? (Multiple choices allowed.)

Devices	CN (%)	IN (%)	JP (%)	UK (%)
PC	90	81	94	91
Smartphone	99	69	68	93
Tablet	52	16	13	51

Table 4 Do you have a degree in computer science or information security? Are you taking a degree in them currently?

CS Degree	CN (%)	IN (%)	JP (%)	UK (%)
Yes.	16	88	6	5
No.	82	11	94	94
Other.	2	1	1	0

Table 5 Have you received any information about the risks of not managing password properly? (Multiple choice allowed.)

	CN (%)	IN (%)	JP (%)	UK (%)
Yes, at school.	28	26	6	8
Yes, at work.	37	77	15	24
Yes, at other places.	11	20	7	20
No I haven't.	38	11	76	56

3.3 Password Habits

We now present the results of the structured questionnaire, which aims at studying how a user creates passwords. Specifically, we studied, password creation approaches, the password composition process, words used for passwords and languages. We expect that these factors are correlated with the linguistic/cultural differences of language spheres as well as the weakness of the passwords.

Password Creation Approaches We asked the participants how they created their passwords; they were given four choices: "think by themselves", "use password generator", "use initial passwords", and "others." These choices are based on previous works focusing on users password-creation habits [4]. We asked the participants who answered "other" to describe the strategy. Figure 1 presents the results. We see that "Think by themselves," which is prone to be cracked in most cases, was the most common password creation approach in all the four countries. The use of password generator has not been a primary method for password creation in the four countries; among them, the UK had the highest adoption rate of 9.6%. We also notice that Chinese and Indian methods of creating passwords were

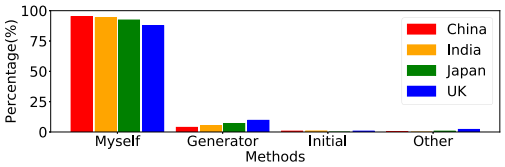

Fig. 1 How do you create passwords?

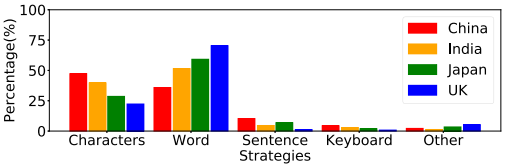

Fig. 2 How do you come up with a password?

Table 6 What words or numbers do you use? (Multiple choices allowed.)

	CN (%)	IN (%)	JP (%)	UK (%)
Personal words				
First name	44	59	18	6
Last name	43	47	9	4
Nickname	29	50	23	7
Birthday	44	61	16	14
Phone number	24	39	3	2
Credit card	2	9	1	1
Person you love	44	30	21	12
Important date	31	24	14	30
Family word	8	15	14	24
Generic words				
Famous person's name	10	28	10	10
Place name	10	21	8	21
Love word	7	13	4	1
Music word	10	9	12	13
Sport word	4	6	5	6
Animal word	4	8	10	23
Religion word	0	3	0	1
Membership ID word	7	14	5	7
Motto	12	2	8	6

similar with each other, while English methods were different, which was proved to be statistically significant through a Chi-square test (significance level of 0.01).

Password Composition Process Next, to those who answered that they create passwords by themselves, we asked their thinking processes. The result is shown in Fig. 2. Here, we see clear differences among the language spheres. While "choosing characters randomly" was not a common strategy among the Japanese and the UK participants, it was more common than "choosing words" among the Chinese participants.

Words Used for Passwords For those who answered that they create passwords from base words or numbers, we asked what word they use. Table 6 summarizes the results. Again, we see intrinsic differences among the cultural spheres. While generic words such as place names or animal words are preferred by the UK participants, personal

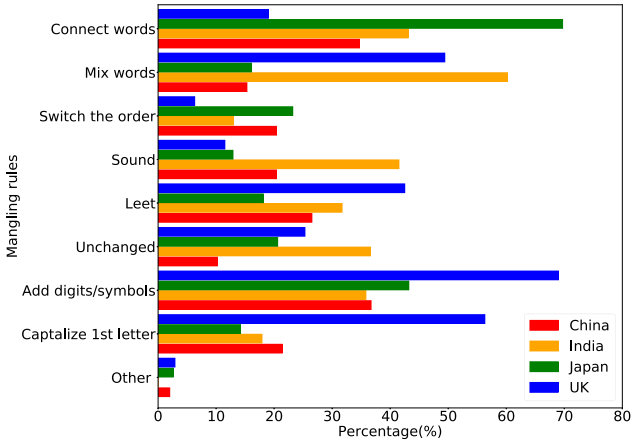

Fig. 3 How do you randomize the words or the numbers? (Multiple choices allowed.)

Table 7 If you create passwords from base words or sentences, what language do you use? (Multiple choices allowed.)

	CN (%)	IN (%)	JP (%)	UK (%)
Prefer CN	62	0	0	0
Prefer JP	0	0	75	0
Prefer EN	38	98	26	100
Prefer others	0	2	1	1

words such as names or birthdays were preferred by Chinese and Indian participants. We note that these differences were statistically significant with the Chi-square test (the significance level of 0.01).

Mangling Rules We also studied the differences in the use of "mangling rule," which is a technique to transform a dictionary word into an obfuscated word; e.g., "Donald Trump" may be transformed into "d0n4ld 7rump" by using the "Leet" technique (which is one of the mangling rules). Adding '!' or '123' at the end of passwords is another example of mangling rule. Figure 3 shows the results. We see the differences of mangling rules among the cultural spheres. Japanese participants preferred to connect words. Indian and the UK participants were fond of mixing words or replacing certain characters with other characters such as Leet. Among the Chinese participants, adding digits/symbols was the most common method. However, the adoption rate of each method in China was lower than 40% and there was no popular method.

Languages Finally, we asked the language they use when creating passwords by themselves. The result is shown in Table 7. As expected, the UK participants mostly use English and some other languages such as Spanish and Czech. In contrast, Chinese and Japanese participants use both English and their first languages. However, a majority of the Indian participants use English instead of their native language such as Tamil and Hindi. Such differences may impact the strategy of selecting effective dictionary when cracking passwords.

4. Analysis of User-Generated Passwords: Leaked Passwords Approach

We analyzed leaked passwords to test if participants' answers in the online survey corresponded to their actual behavior. In this work, we focus our attention on the participants who reported that they create passwords from base words. We study whether dictionary words, meaningful digits, or personal words are included in the leaked passwords. We also study which language is commonly used and what kind of mangling rule was frequently used in the real world. As our password corpus and dictionaries were huge, we leveraged Bloom filters to process the enormous number of words.

4.1 Dataset

Our dataset included sets of email addresses and passwords that were leaked from multiple websites. As these lists contain email address– password pairs, they can be used for an attack called "credential stuffing." In late 2016, a large corpus called "Exploit.in" including email addresses and passwords from various websites appeared in public [20]. The "Exploit.in" dataset contains nearly 600 million unique email address–password pairs. In addition to "Exploit.in", we use other lists that were leaked from two Chinese websites called "7k7k" [21] and "人人网" [22], which are a gaming site in China and a social networking site in China, respectively. Table 8 shows the volumes of the datasets we used.

4.2 Associating Passwords with Language Spheres

Using the email address–password pairs, we attempted to extract passwords that are likely generated by people from each language sphere. To this end, we leveraged the domain names contained in the email addresses. From an email address, we extracted its domain name and checked the audience geography of the website with that domain name. We used the service provided by Alexa Website Traffic, Statistics, and Analytics [23]. From the audience geography of a domain name, we can estimate the primary language of the visitors who access the site. That is, if more than 90% of the visitors are located in either China, India, Japan, or the UK, we labeled the email address and password as Chinese, Indian, Japanese, or English data, respectively. As the number of domain names included in the dataset was large, we limited our search to a set of top-level domain names (TLDs). Namely, we adopted ".com", ".org", and ".net" as the TLDs used in four countries and adopted ".cn", ".in", ".jp", or ".uk" as the respective TLD country codes (ccTLDs).

We extracted domain names with the following criteria: for the domain names under the four ccTLDs, we picked up the ones that were associated with more than 10 distinct email addresses. Similarly, for the three TLDs, ".net", ".edu", and ".org", we picked up the ones associated with more than 100 distinct email addresses, and for domain names under ".com", we picked up the ones with more than 1,000 distinct email addresses. Finally, we eliminated the email address–password pairs, which contained non-ASCII characters. As a result, we eliminated 7,761 pairs from the Chinese, 560 pairs from the Indian, 39 pairs from the Japanese, and 74 pairs from the UK data. In this way, we obtained sets of email addresses and passwords classified by users' countries. The volumes of data are presented in Table 9. Previous works have shown that major services require users to create passwords of six characters or longer [24]–[26]. In this study, we decided to use data that includes passwords longer than six characters.

4.3 Extracting Word-Based Passwords

To detect word-based passwords, we first converted upper case letters in passwords into lower case letters, and then took the following two steps. The first step was to check mangling rules and extract words observed in leaked passwords. The second step was to analyze the words (languages and categories of the words). Our procedure is illustrated in Fig. 4. According to several password properties, we classified the passwords into eight groups. To this end, we formulated a rule that compiles the heuristics in a mutually exclusive, collectively exhaustive (MECE) manner.

In Fig. 4, G1 is the group of passwords composed of one word including names, dates, phone numbers, credit card numbers, and words in dictionaries. Passwords that are words (names or dictionary words) switched the order of the characters belonging to G2. Passwords that are words (names or dictionary words) converted with leet are G3. G4 includes the passwords of multiple words, and when at least one word of the multiple words is converted with leet, the password belongs to G5. When the password is created by adding numbers or digits to a word, it belongs to G6, and when the password is a mixture of a word and digits/symbols, it is in G7. G8 is the group of passwords including words. Each password belongs to one group and the groups do not overlap anywhere.

4.3.1 Dictionaries and Regular Expressions

In Sect. 3, we found that users prefer to create pass-

Table 8 Volumes of the leaked password datasets used in our study.

Dataset	Amount	Web service	Breach date
Exploit.in	805,499,579	–	October 2016
7k7k	19,138,452	Game	January 2011
人人网	4,768,600	Social media	December 2011

Table 9 Number of passwords analyzed.

Country	# total	# Longer 6
China	5,881,906	5,720,606
India	890,079	830,733
Japan	462,048	437,147
the UK	388,276	370,596

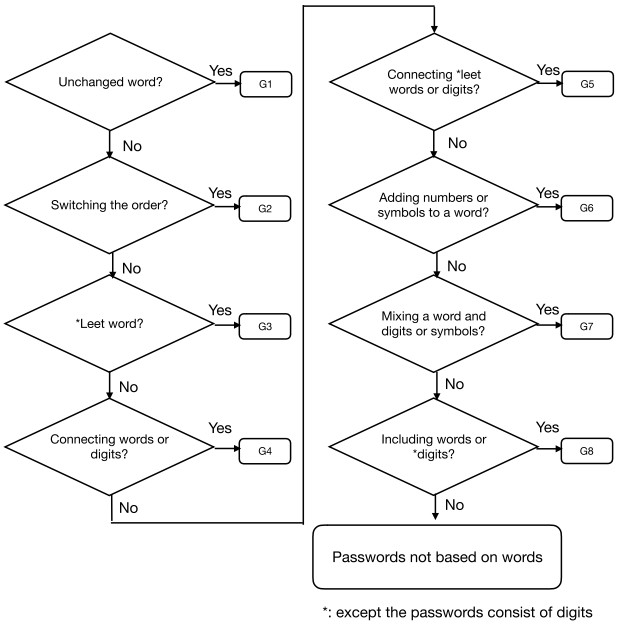

Fig. 4 Procedure of the word analysis.

words using words. "Name", "Birth Date", "Phone number", and "Animal words" were frequently used. We confirmed whether the results corresponded with leaked passwords. We used name lists of Facebook users [27] to check if passwords included names. In particular, for the Japanese dataset, we also used name lists in mecab-ipadic-neologd [28]. Some names have the same spelling as dictionary nouns. When a password consists of such a word, we labeled it as a password with a generic word. To check if a potential name is truly a name or a generic word, we used words tagged "Temporal Noun (NT)" or "Other Noun (NN)" in Chinese Treebank 8.0, "generic nouns (名詞, 一般)" in mecab-ipadic-neologd, and "noun, singular (NN)" or "noun plural (NNS)" in MASC Sentence Corpus. Regarding "Birth Date" and "Phone number", we created regular expressions to check if such information was included in passwords. We do not have users' personal information, so we just check if passwords include a date in the range of 1900 to 2099. We checked the following patterns: YYYYMMDD, MMDDYYYY, DDMMYYYY, YYYY, MMDD, or DDMM. In addition, we check "Credit card number" by using regular expressions. We prepared regular expressions for the following cards: Amex, BC-Global, Carte Blanche, Diners Club, Discover, Insta Payment, JCB, KoreanLocal, Laser, Maestro, Mastercard, Solo, Switch, Union Pay, Visa, and Visa Master. As for generic words, we prepared dictionaries of Japanese, Chinese, and English words. We only used words that consist of more than three characters. Details about the dictionaries are presented in Table 10.

4.3.2 Bloom Filter

As shown in Table 10, the size of each dictionary is relatively large. Therefore, storing all the dictionaries in a memory space requires a large amount of memory capacity. To address this issue, we leveraged a Bloom filter, using which we can make the lookup process scalable. A Bloom filter [29] is a data structure consisting of an M-bits array. All bits are set to 0 at first. To insert a word in the data structure, k hash functions are computed for each word. The outputs of the hash functions should be smaller than M. Say we have a word "e" to store in the bloom filter. First, we compute k hash functions for the word. Then the bits of index $h_1(e)$, $h_2(e)$, ..., $h_k(e)$ are set to 1. To check if a word is in the Bloom filter, the same process is done for the word. When all bits are set 1, we consider the word to be in it. We do not skip words, but we may mistakenly detect a word that is not in the filter. By changing the length of the bits array or the number of hash functions, we can control the false positive rate. We decided that the false positive rate should be $p = 0.000001$. In order to make the filter more robust, we adopted a variant of the Bloom filter introduced in [30], which prepares a bit array for each hash function. This Bloom filter ensures that k bits are set for one word.

Table 10 List of dictionaries used in our study. The "Used" column refers to the set of words that eliminates digit-only words and one/two-letter words.

	Dictionaries	Total #	Used #
Name	facebook-firstnames-withcount.txt	4,347,667	4,346,965
	facebook-lastnames-withcount.txt	5,369,437	5,368,735
	Mecab (tagged as persons' name)	599,934	598,106
Chinese	Chinese Treebank 8.0	114,174	113,378
Japanese	Mecab	2,394,665	2,394,026
English	MASC Tagged Corpus	40,286	32,200
	Wordnet	354,117	353,804

4.3.3 Mangling Rules

We analyzed the usage rates of "Connecting words", "Mixing words", "Switching the orders", "Leet", "Unchanged", and "Adding digits or symbols to a word" in leaked passwords. "Connecting words" is a strategy of connecting words, names, dates, phone numbers, or credit numbers, and "Unchanged" is a way of using them as is. "Adding digits or symbols" means adding digits or symbols at the beginning or end of a word. "Switching the orders" and "Leet" are common mangling rules and Hashcat [31] supports these rules. We define "Switching the orders" as reversing the word, putting the first letter at the end, or putting the last letter at the beginning. Regarding "Leet", we checked for the following replacements: a:4, a:@, b:6, c:<, c:{, e:3, g:9, i:1, i:!, o:0, q:9, s:5, s:$, t:7, t:+, x:%. For "Mixing words", We decided to check passwords that consist of a word whose characters are surrounded with digits or symbols (e.g., p1a2s3s4, 12pa!!ss). Finally, we checked the passwords that include words. We labeled the passwords as "Include" if we could not determine the mangling rules.

4.4 Analysis of Word-Based Passwords

Basic statistics

Password length, structure, and common passwords are presented in Table 11. We used "Password Analysis and Cracking Toolkit" (PACK) [32] to check the length and structure. Similar to the findings of previous studies, frequently used passwords are "123456" and "password". More than half of Chinese passwords consist of only digits, and the site name "tianya" and "5201314", which sounds like "我愛你一生一世 (I love you forever)", are popular. Users in India and the UK prefer letters to digits. In the top-10 passwords of UK users, both generic words ("password","liverpool") and personal words ("charlie","thomas") appear. "liverpool" and "chelsea", which are the names of English football clubs, were common. In the top-10 passwords of Indian users, we found names of deities in Hinduism such as "Krishna" and "Ganesh." Regarding Japanese, they use both letters and digits. "Sakura" means cherry blossoms in Japanese, and "yokohama" is a city in Japan. "11922960" sounds like the phrase "いい国作ろう (Let's make our country great)", which is related to Japanese history.

Word-based Passwords

The fractions of passwords that are likely created from base words were 41.3%, 82.9%, 80.1%, and 90.2% for Chinese, Indian, Japanese, and the UK, respectively. As Sect. 3 showed, the percentage of word-based passwords is high in Japan and the UK. The difference in the percentages was statistically significant in the Chi-squared test (significance level of 0.01). Regarding the languages of the words, native languages are frequently used. In China and Japan, English words are also popular (Table 12).

Of the passwords created from base words, we calculated the percentages of the passwords that use each mangling rule (Fig. 5). The mangling rule most frequently used in China and India was "Unchanged", the most popular one in Japan was "Connecting words", and the most common in the UK was "Adding digits/symbols". Their adoption percentages were different among countries, and the differences were statistically significant. Of the users who create passwords by connecting words, 56% in China, 79% in India, 48% in Japan, and 77% in the UK use two words, and 7% in China, 4% in India, 5% in Japan, and 2% in the UK of them repeat the same word.

In all four countries, "Date", "Name", and "Word in dictionaries" are popular. In spite of user studies showing that Chinese people tend to use personal information, words in the dictionary that mainly include generic words are frequently used in Chinese leaked passwords. Also, we observed "Names" in English passwords, while most of the UK participants in our user study did not report that they used people ' s name. The use rates of "Names", "Dates", and "Phone numbers" were statistically different among countries. We looked into the words used to create passwords. We only checked 100 frequently used words in passwords that consist of one word. In Chinese and Indian passwords, technical words like "computer" and "internet" are frequently used. We found some nicknames (e.g., "xiaoxiao", "yangyang") in Chinese passwords. We did not have nickname lists, so these nicknames were classified as dictionary words. Comic book character names (e.g., "doraemon", "naruto") are common in Japan, and animal names (e.g., "monkey", "elephant") frequently appear in English passwords. Also, we found foods like "muffin" and

Table 11 Basic statistics of the leaked passwords.

	CN (%)	IN (%)	JP (%)	UK (%)
Password Length				
1	6 (26)	6 (25)	8 (48)	8 (26)
2	8 (22)	8 (22)	6 (13)	6 (22)
3	7 (17)	7 (14)	9 (9)	7 (16)
4	10 (12)	9 (10)	7 (9)	9 (14)
5	9 (11)	10 (3)	10 (7)	10 (9)
Password Structure				
	CN (%)	IN (%)	JP (%)	UK (%)
1	DDDDDD (20)	LLLLLL (14)	LLLLLLLL (22)	LLLLLL (11)
2	DDDDDDDD (12)	LLLLLLLL (10)	DDDDDDDD (7)	LLLLLLLL (9)
3	DDDDDDD (12)	LLLLLLL (9)	LLLLDDDD (5)	LLLLLLL (7)
4	DDDDDDDDD (4)	DDDDDD (6)	DDDDDD (4)	LLLLLLDD (4)
5	DDDDDDDDDD (3)	LLLLLLLLL (4)	LLLLLL (4)	LLLLLLLLL (3)
Password Ranking				
	CN (%)	IN (%)	JP (%)	UK (%)
1	123456 (3.30)	123456 (1.96)	123456 (0.15)	password (0.21)
2	111111 (0.83)	password (0.35)	password (0.07)	123456 (0.21)
3	123456789 (0.52)	ashishbiyani (0.33)	123456789 (0.06)	charlie (0.14)
4	123123 (0.38)	123456789 (0.21)	12345678 (0.06)	liverpool (0.12)
5	111222tianya (0.27)	12345678 (0.16)	1qaz2wsx (0.05)	chelsea (0.09)
6	12345678 (0.27)	krishna (0.15)	sakura (0.04)	thomas (0.08)
7	5201314 (0.23)	sairam (0.12)	Exigent (0.04)	george (0.07)
8	super123 (0.22)	indian (0.10)	1234567890 (0.03)	charlie1 (0.07)
9	D1lakiss (0.19)	ganesh (0.10)	11922960 (0.03)	tigger (0.07)
10	123321 (0.15)	sachin (0.09)	yokohama0 (0.03)	password1 (0.07)

Table 12 Languages used in leaked passwords.

	CN (%)	IN (%)	JP (%)	UK (%)
Chinese words	13	4	5	3
Japanese words	2	6	18	4
English words	9	22	17	35

Table 13 Categories of words in word-based passwords.

	CN (%)	IN (%)	JP (%)	UK (%)
Date	26	5	17	4
Phone	5	3	0.3	0.1
Credit card	0.02	0.04	0.03	0.03
Name	19	40	44	33
Words in dictionary	54	37	48	45

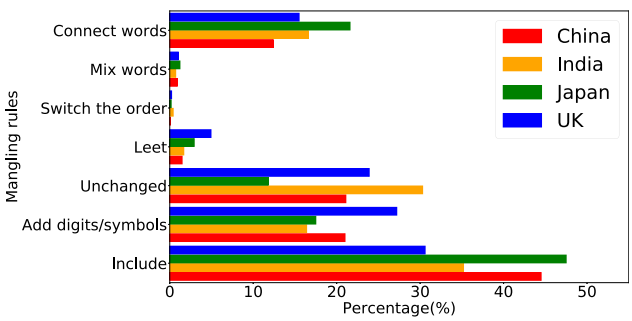

Fig. 5 Mangling rules of word-based passwords.

"cookie" in English passwords. Regarding sports, while we found various kinds of sports (e.g., "tennis", "soccer", "baseball") in Japanese passwords, only "football", "cricket", and "golfer" appeared in the top 100 frequently used English words. We also determined that the names of deities in Hinduism (e.g. "Krishna", "Ganesh", and "Lakshmi") were popular in India. As we did not use a Hindi dictionary that covers generic words in Hindi, these words were classified under "Name."

5. Comparing User Study and Leaked Passwords

In the user study shown in Sect. 3, 34.1%, 48.4%, 54.6%, and 61.7% of China, India, Japan, and the UK participants created passwords by themselves using base words/digits, respectively. However, 41.3%, 82.9%, 80.1%, and 90.2% of passwords seemed to be derived from words or meaningful digits, respectively. The percentages of leaked passwords were much higher than those of the user study. Possible reasons for the results are the false positives of word-detection, users' unconscious use of words, the differences between services of the supposed sites in the user study, and the services of actual leaked sites.

Languages and Based Words The user study and leaked passwords correspond to the language they use to create passwords. Users are most likely to use their native language in China, Japan, and the UK, and English is also as common as in China and Japan.

Regarding the categories of the words, as users answered in our user study, we observed several dates in the Chinese and names in the Indian and Japanese passwords. Further, we found phone numbers in Chinese and Indian leaked passwords. Names are used frequently in the UK, but UK users in our user study said they did not use their personal information. They seem to choose easy words that can be remembered unconsciously.

Mangling Rules Participants in the user study said that they mixed words to create passwords and changed a certain character to another using "leet". However, we did not observe many characters replaced using leet or mixed words in leaked passwords. We found that the most frequently used strategy in Chinese and Indian passwords was "unchanged". Users believe that they are adopting a secure password creation strategy; however, in fact, they are not. Further, we observed a few consistencies in the choice of mangling rules. Both the user study and the leaked passwords indicated that "connecting words" was common in Japan and "adding digits/symbols" was popular in the UK.

6. Password Guessing

In this section, we aim to examine how well attackers can guess passwords by utilizing the linguistic background of the targeted users. To this end, we tested the following two scenarios.

- **Scenario 1**: An attacker knows the linguistic background of their targeted users and can use the password data leaked from the websites for users in the same language sphere of the targeted users.
- **Scenario 2**: An attacker does not know the linguistic background of targeted users and uses the password data leaked from websites for users in the various language spheres.

For comparison's sake, we further set the following baselines that represent an immature attacker and an idealized attacker, respectively.

- **Baseline 1**: An immature attacker who utilizes leaked passwords that are easily found by anyone on the Internet.
- **Baseline 2**: An idealized attacker who can perfectly order guesses of passwords. We compute the metrics named "guesswork," which was proposed in Ref. [33]; i.e., we compute $G_\alpha = \sum_{i=1}^{N} p_i \cdot i$, where p_i is the probability that the i-th most common password is sampled out of the entire password set.

6.1 PCFG

PCFG [16] guesses passwords by using rules, which are generated in the training step or prepared manually. The rules contain the probabilities of structures appearing in the training dataset (e.g., "A_6:0.07", "A_8:0.06") and the frequencies of the character strings (e.g., "password:0.02", "sunshine:0.006"). For scenario 1, we prepared training and test sets for each country. We randomly sampled 150,000 passwords each for training sets and test sets to ensure that the numbers of training/test data from the four countries are the same. For scenario 2, we randomly sampled 37,500 passwords from four countries and prepared mixed training and test sets composed of 150,000 passwords. As the baseline 1, we randomly sampled 150,000 passwords from the "RockYou" dataset and used them as a training set. Finally, we compiled six training sets: CN, IN, JP, UK, mixed, and RockYou, and four test sets: CN, IN, JP, and UK. We created rules and generated guesses from the corresponding training set. Each test set is tested three times with guesses from the corresponding country's training set, mixed trining set, and RockYou training set.

6.2 Result

We generated 10^9 of guesses and calculated the percentages of passwords found in the test sets. We show the results in Fig. 6. We used passwords from each country to train a PCFG and generated guesses, and then we calculated the percentage of cracked passwords of each country. In the figures, the solid lines represent the first scenario, the dashed lines represent the second scenario, the dotted lines represent the baseline 1, and the baseline 2 is represented by the dash-dotted lines.

Comparing the first scenario and the second scenario, we found that when the PCFG was trained with each country's passwords, the speed of cracking passwods was higher

Table 14 Training and test sets for PCFG.

	CN		IN		JP		UK		MIX	RockYou
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Train
Chinese passwords	150,000	150,000	0	0	0	0	0	0	37,500	0
Indian passwords	0	0	150,000	150,000	0	0	0	0	37,500	0
Japanese passwords	0	0	0	0	150,000	150,000	0	0	37,500	0
UK passwords	0	0	0	0	0	0	150,000	150,000	37,500	0
Passwords from RockYou	0	0	0	0	0	0	0	0	0	150,000
total	150,000	150,000	150,000	150,000	150,000	150,000	150,000	150,000	150,000	150,000

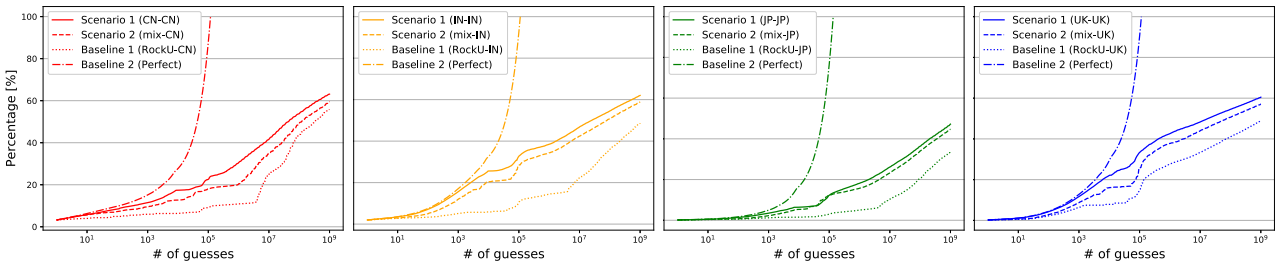

Fig. 6 The percentage of passwords guessed after a given number of guesses: Chinese, Indian, Japanese, and English.

than those with mixed passwords.

To crack 10% of the test passwords, the attacker had to generate 1,408 guesses (for Chinese), 618 guesses (for Indian), 58,860 guesses (for Japanese), and 1,520 guesses (for UK) considering the rules of mixed passwords. However, only 319 guesses, 254 guesses, 71,420 guesses, and 669 guesses for the respective countries were required when attackers leverage each country ' s passwords; i.e., the password guessing process against Chinese, Indian, and the UK passwords has become faster by leveraging the knowledge on the language of targeting passwords. For Japanese passwords, the PCFG trained with mixed passwords eventually outperformed the targeted passwords. This is because PCFG learned propensity of password creation like a structure or sequences of characters.

Regarding the first scenario, comparing the efficiency between countries, we found that 5.81% of Chinese passwords, 3.46% of Indian passwords, 1.13% of UK ones and 0.48% of Japanese ones were cracked within 10 guesses. We note that 3.24% of Chinese passwords and 1.98% of Indian passwords are recovered with just one guess. The guessed password was "123456," which was the most popular password in the Chinese and Indian password dataset. As such, Chinese/Indian common passwords are too popular, which is what lead to this result. The results show that Japanese passwords are relatively difficult to guess. Japanese prefer connecting letters and digits, and guessing the correct combination is laborious. Also, they tend to use various kinds of words like "Names", "English words", and "Japanese words". These observations may be the reasons for the result.

7. Discussion

7.1 Limitations

One limitation of our work is that our dataset was a "combo list", which is a compilation of credential data leaked from various websites, and we do not know the composition requirements and the scenario based on which the passwords were created. In general, users' password creation habits depend on the password policy or the kinds of service. Therefore, we cannot conclude that the propensity we found is always consistent. We also note that two Chinese leaked password sets also had different password composition policies, implying that it is not possible to measure/analyze passwords in a cross-cultural manner given a constraint that all the password sets should have the same password composition policies.

In addition, there was an intrinsic time lag between the time the leaked passwords became available and the time our survey was carried out. Some of the inconsistencies in the results of the user study and leaked passwords analysis might be attributed to the time lag. As users may have become more knowledgeable on the security of passwords, given the several large-scale password leakage incidents, their ways of creating passwords might also have changed. Understanding how such a gap affects the analysis is left for the future study.

7.2 Ethics

Our survey included potentially sensitive questions, such as password creation strategies, password management, etc. Therefore, we obtained informed consent from all participants before questioning them. We clarified that they were

able to quit anytime, that their responses would be used solely for this research, and that their privacy will be protected when the results of this survey are published. The leaked datasets we analyzed consisted of mail address—password pairs. We stored all data securely and did not expose them or test the validity of the data with real services. We only used them for our research.

7.3 Future Work

In this work, we found intrinsic differences among passwords generated by users from multiple-language spheres, and this tendency will help an attacker to guess passwords. However, we did not investigate why these differences exist in nature. A future could include asking users open-ended questions and conducting deep analysis. As a step toward securing human-generated passwords, we require a mechanism to improve the strength of human-generated password without sacrificing usability. In the future, studies could determine better ways to urge users with different cultural backgrounds to create secure passwords.

While this work focused on the creation of passwords, extending the study to other topics, such as the management of passwords, is the next step toward establishing better password practices on the basis of language sphere. Conducting research on password management tools and addressing limitations will be necessary in future work.

Users in China, India, Japan, and the UK predominantly use personal information to create their passwords. It is common that the email addresses (especially for work) include names. Attackers can obtain both users' nationalities and their personal information from their email addresses, which helps attackers to guess passwords effectively. Users should be urged not to use previously leaked passwords, but it is also important to educate users in creating passwords that cannot be guessed easily using open data (e.g., email address or personal information on social media).

Looking into countries where cultures and environments are completely different from these four countries is left for our future study. We expect that people living in the "Next Eleven" countries might have different password creation habits from what we found in this study. Focusing on them may help us to understand the relationship between security and industrial development.

8. Conclusions

Users tend to create and memorize their passwords. This way of creating passwords has been studied in English-speaking users. In this study, we focused on Chinese, Indian, Japanese, and English users from two points of view: a user survey and leaked passwords analysis. Both the user study and leaked password data showed that the majority of users create passwords by themselves based on some words. The word categories they choose from and the way they mangle the words differed among countries. Finally, we demonstrated that knowledge of the linguistic background of targeted users contributes to increase the speed of password guessing process.

References

[1] K. Mori, T. Watanabe, Y. Zhou, A.A. Hasegawa, M. Akiyama, and T. Mori, "Comparative analysis of three language spheres: Are linguistic and cultural differences reflected in password selection habits?," 2019 IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2019, Stockholm, Sweden, June 17–19, 2019, pp.159–171, 2019.
[2] Z. Li, W. Han, and W. Xu, "A large-scale empirical analysis of chinese web passwords," Proc. 23rd USENIX Security Symposium, pp.559–574, 2014.
[3] S. Riley, "Password security: What users know and what they actually do," Usability News, vol.8, no.1, 2006.
[4] B. Ur, F. Noma, J. Bees, S.M. Segreti, R. Shay, L. Bauer, N. Christin, and L.F. Cranor, ""I added '!' at the end to make it secure": Observing password creation in the lab," Eleventh Symposium On Usable Privacy and Security, SOUPS 2015, pp.123–140, 2015.
[5] R. Wash, E. Rader, R. Berman, and Z. Wellmer, "Understanding password choices: How frequently entered passwords are re-used across websites," Twelfth Symposium on Usable Privacy and Security, SOUPS 2016, pp.175–188, USENIX Association, 2016.
[6] S. Pearman, J. Thomas, P.E. Naeini, H. Habib, L. Bauer, N. Christin, L.F. Cranor, S. Egelman, and A. Forget, "Let's go in for a closer look: Observing passwords in their natural habitat," Proc. 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17, pp.295–310, ACM, 2017.
[7] A. Hanamsagar, S.S. Woo, C. Kanich, and J. Mirkovic, "Leveraging semantic transformation to investigate password habits and their causes," Proc. 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pp.570:1–570:12, ACM, 2018.
[8] M. Harbach, A.D. Luca, N. Malkin, and S. Egelman, "Keep on lockin' in the free world: A multi-national comparison of smartphone locking," Proc. 2016 CHI Conference on Human Factors in Computing Systems, 2016.
[9] Y. Sawaya, M. Sharif, N. Christin, A. Kubota, A. Nakarai, and A. Yamada, "Self-confidence trumps knowledge: A cross-cultural study of security behavior," Proc. 2017 CHI Conference on Human Factors in Computing Systems, 2017.
[10] J. Zeng, J. Duan, and C. Wu, "Empirical study on lexical sentiment in passwords from chinese websites," Computers & Security, vol.80, pp.200–210, 2019.
[11] M. AlSabah, G. Oligeri, and R. Riley, "Your culture is in your password: An analysis of a demographically-diverse password dataset," Computers & Security, vol.77, pp.427–441, 2018.
[12] K. Thomas, F. Li, A. Zand, J. Barrett, J. Ranieri, L. Invernizzi, Y. Markov, O. Comanescu, V. Eranti, A. Moscicki, D. Margolis, V. Paxson, and E. Bursztein, "Data breaches, phishing, or malware?: Understanding the risks of stolen credentials," Proc. 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, pp.1421–1434, 2017.
[13] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang, "The tangled web of password reuse," Proc. 21st Annual Network and Distributed System Security Symposium, NDSS, 2014.
[14] A. Narayanan and V. Shmatikov, "Fast dictionary attacks on passwords using time-space tradeoff," Proc. 12th ACM Conference on Computer and Communications Security, CCS '05, pp.364–372, ACM, 2005.
[15] J. Ma, W. Yang, M. Luo, and N. Li, "A study of probabilistic password models," IEEE Symposium on Security and Privacy, pp.689–704, IEEE Computer Society, 2014.
[16] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, "Password cracking using probabilistic context-free grammars," Proc. 30th IEEE Symposium on Security and Privacy, S&P 2009, pp.391–405,

2009.
[17] W. Melicher, B. Ur, S.M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L.F. Cranor, "Fast, lean, and accurate: Modeling password guessability using neural networks," Proc. 25th USENIX Conference on Security Symposium, SEC '16, pp.175–191, USENIX Association, 2016.
[18] H. Krasnova and N.F. Veltri, "Privacy calculus on social networking sites: Explorative evidence from Germany and USA," 43rd Hawaii International Conference on Systems Science (HICSS-43 2010), 5–8 Jan. 2010, Koloa, Kauai, HI, USA, pp.1–10, 2010.
[19] Y. Wang, G. Norcie, and L.F. Cranor, "Who is concerned about what? A study of American, Chinese and Indian users' privacy concerns on social network sites," 4th International Conference on Trust and Trustworthy Computing, TRUST 2011, Pittsburgh, PA, USA, June 22-24, 2011, Lecture Notes in Computer Science, vol.6740, pp.146–153, Springer, Berlin, Heidelberg,2011.
[20] T. Hunt, "Password reuse, credential stuffing and another billion records in have i been pwned," https://www.troyhunt.com/password-reuse-credential-stuffing-and-another-1-billion-records-in-have-i-been-pwned/
[21] 7k7k, "7k7k.com," http://www.7k7k.com/
[22] R. Inc., "Renrenwang," http://browse.renren.com/
[23] Alexa, "Website traffic, statistics, and analytics," https://www.alexa.com/siteinfo
[24] J. Bonneau and S. Preibusch, "The password thicket: Technical and market failures in human authentication on the web," WEIS, 2010.
[25] J.R. Saini, "Analysis of minimum and maximum character bounds of password lengths of globally ranked websites," International Journal of Advanced Networking Applications, 2014.
[26] Y. Li, H. Wang, and K. Sun, "Email as a master key: Analyzing account recovery in the wild," INFOCOM, pp.1646–1654, IEEE, 2018.
[27] skullsecurity, "Facebook lists," 2010, https://wiki.skullsecurity.org/index.php?title=Passwords
[28] T. Sato, "mecab-ipadic-neologd: Neologism dictionary for mecab," https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md
[29] B.H. Bloom, "Space/time trade-offs in hash coding with allowable errors," Commun. ACM, vol.13, no.7, pp.422–426, 1970.
[30] F. Chang, W.-C. Feng, and K. Li, "Approximate caches for packet classification," INFOCOM, pp.2196–2207, IEEE, 2004.
[31] hashcat, "hashcat advanced password recovery," https://hashcat.net/hashcat/
[32] P. Kacherginsky, "Pack (password analysis and cracking kit)," https://github.com/iphelix/pack
[33] J. Bonneau, "The science of guessing: Analyzing an anonymized corpus of 70 million passwords," IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA, pp.538–552, 2012.

Appendix: Questionnaire for User Survey

A.1 Consent for Participation in the Study

The researcher requests your consent for participation in a study about password management. This consent form asks you to allow the researcher to use your comments to enhance understanding of the topic.

Participation in this study is not forced by anyone. If you decide not to participate, you can abandon the task at anytime (and will not have rewards for it). Please be aware that if you decide to participate, you may stop participating at any time and you may decide not to answer any specific question.

The researcher will maintain the confidentiality of the data. Any information that is obtained in connection with this study and that can be identified with you will remain confidential.

By submitting this form you are indicating that you have read the description of the study, and that you agree to the terms as described.

Thank you in advance for your participation!

1. I agree to participate in the research study. I understand the purpose and nature of this study. I understand that I can withdraw from the study at any time.

- Yes
- No

2. I grant permission for the data generated from this study to be used in the researcher's publications on this topic.

- Yes
- No

3. Please check the following box to indicate agreement to participate in this study.

- I agree

A.2 Demographics

4. How old are you?
5. What is your nationality?
6. What country do you live in?
7. What is your first language?
8. What other languages can you speak?
9. What is your gender?

- Female
- Male
- Prefer not to say
- other:[user's input]

10. What is your occupation?

- Accounting
- Finance
- Freelance
- Engineering
- Health Care
- Government
- Sales
- Transportation
- Student
- Prefer not to say
- other:[user's input]

11. What kind of internet services which require passwords do you use?

- Social media
- Online Shopping

- Banking
- Email
- Video Service
- Payment service
- other:[user's input]

12. What kind of computing devices do you use?

- PC
- smartphone
- tablet
- other:[user's input]

A.3 Knowledge

13. Did you take a degree in computer science or information security? / Are you taking a degree in them?

- Yes
- No
- Prefer not to say
- other:[user's input]

14. Have you received training on how to manage your passwords at work or school?

- Yes. I received training at work.
- Yes. I received training at school.
- Yes. I received training at other places.
- No

15. What do you remember from the training?
16. Have you received any information about the risks of not doing password management?

- Yes. I received information at work.
- Yes. I received information at school.
- Yes. I received information at other places.
- No

17. What information do you remember regarding the risks of not doing password management?
18. What steps do you take to create a strong password?

A.4 Passwords for important accounts

Please answer the questions about your passwords for important accounts. (e.g. primary e-mail account)

19. How do you create passwords? Please select the one which best describes the method for your important accounts.

a create it myself >Q20
b use a password generator >Q29
c use initial passwords >Q31
d other:[user's input]

20. How do you create your passwords? [Please answer, if you answered (a) in Q19.]

a use characters individually
b use words or numbers
c use sentences
d use keyboard layout
e other:[user's input]

21. If you create passwords from base words or sentences, what language do you use? (e.g. English, German, French, Spanish)
22. How do you choose characters? Please select all that apply. (Multiple choices allowed) [Please answer, if you answered (a) in Q20.]

a decide a base sentence and choose one character from each word from the sentence
b decide some words and choose one character from each word
c decide a character which you came up with suddenly
d type on your keyboard randomly
e other:[user's input]

23. What words or numbers do you use? (Multiple choices allowed) [Please answer, if you answered (b) in Q20.]

- your first name
- your last name
- nickname
- your birthday
- phone number
- credit card number
- famous person's name
- famous person's birthday
- the person you love
- important date (anniversary, and so on)
- website name
- the date you register the website
- place name
- words related to love
- words related to music
- words related to sport
- words related to your family
- words related to animals
- words related to your religion
- pets name
- your favorite words
- ID numbers for another membership
- motto
- other

24. How do you randomize the words or the numbers? (Multiple choices allowed) [Please answer, if you answered (b) in Q20.]

a use the word itself
b replace certain characters with other characters (e.g. password ->p@ssw0rd, e->3, i->1)
c replace certain words/numbers with other numbers/ words which have similar sound. (e.g. ate ->8)
d switch the order of each letter/word. (e.g. password ->drowssap/wordpass)
e mix words or numbers

f connect words or numbers
g shorten a word (e.g. password ->pswd)
h add some digits or symbols at the end
i add some digits or symbols at the beginning
j capitalize the first letter of a word
k other:[user's input]

25. How do you choose sentences? (Multiple choices allowed) [Please answer, if you answered (c) in Q20.]

a personal sentence (e.g. I went to New York on April 11th.)
b famous quotes (e.g. Genius is one percent inspiration and ninety-nine percent perspiration.)
c general sentence (e.g. It is fine today.)
d other:[user's input]

26. Please tell us the length of the password created by the password generator. [Please answer, if you answered (b) in Q19.]
27. Please tell us the generation rules. (Multiple choices allowed) [Please answer, if you answered (b) in Q19.]

- Include Symbols (!, @, #, $...)
- Include Numbers (0123...)
- Include Lowercase (abc...)
- Include Uppercase (ABC...)

28. Why do you create your passwords in that way?

A.5 Related information

29. Have you ever had your passwords leaked?

- Yes
- No
- Prefer not to say

30. How did you notice it?
31. What kind of password creation strategies did you used to use?
32. Do you change password creation strategies depending on accounts? How? (Email/Banking/Game/Shopping...etc)
33. Do you reuse your password?

- I use the same password for all websites.
- I use the same password for websites which provide the same service.
- I use the same password for websites which provide the different services.
- I use the different passwords for each website.
- other:[user's input]

34. How many accounts do you have?

- 1
- 2-5
- 6-10
- 11-20
- 21-50
- 51-100
- 101+

35. How many distinct passwords do you have?

- 1
- 2-5
- 6-10
- 11-20
- 21-50
- 51-100
- 101+

36. How do you manage your passwords?

- I remember all passwords.
- I remember some of my passwords.
- I write down the passwords in my notebook or diary.
- I recorded my passwords in my PC or smartphone.
- I save my passwords in my browser.
- I use a password management software.
- other:[user's input]

37. Please tell us your opinion about password management software and give your reasons.

- I'm using one.
- I used to use one.
- I want to try.
- I won't use it.
- I don't know about it.
- other:[user's input]

38. Please give your reasons for the question above.

Keika Mori recieved B.E. degree in computer science and engineering from Waseda University in 2018. She is a graduate student at Department of Computer Science and Communication Engineering, Waseda University. She has been conducting projects in usable security.

Takuya Watanabe recieved M.E. degree in computer science and engineering from Waseda University, Japan in 2016. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2016, he has been engaged in research of consumer security and privacy. He is now with the Cyber Security Project of NTT Secure Platform Laboratories.

Yunao Zhou received B.E degree in the Information Security from Xidian University, Xi'an, China. Currently, he is a graduate student at Department of Computer Science and Communication Engineering, Waseda University.

Ayako Akiyama Hasegawa received her B.S. and M.S. degrees in information science from Ochanomizu University in 2013 and 2015, respectively. She also received her B.S. degree in human science from Musashino University in 2019. She is currently a researcher at NTT Secure Platform Laboratories, Tokyo, Japan. Her current research interests are mainly on usable security and privacy.

Mitsuaki Akiyama received his M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology, Japan in 2007 and 2013. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2007, he has been engaged in research and development on cybersecurity. He is currently a Senior Distinguished Researcher with the Cyber Security Project of NTT Secure Platform Laboratories. His research interests include cybersecurity measurement, offensive security, and usable security and privacy.

Tatsuya Mori is currently a professor at Waseda University, Tokyo, Japan. He received B.E. and M.E. degrees in applied physics, and Ph.D. degree in information science from the Waseda University, in 1997, 1999 and 2005, respectively. He joined NTT lab in 1999. Since then, he has been engaged in the research of measurement and analysis of networks and cyber security. From Mar. 2007 to Mar. 2008, he was a visiting researcher at the University of Wisconsin-Madison. He received Telecom System Technology Award from TAF in 2010 and Best Paper Awards from IEICE and IEEE/ACM COMSNETS in 2009 and 2010, respectively. Dr. Mori is a member of ACM, IEEE, IEICE, and IPSJ.