Understanding Human-Chosen PINs: Characteristics, Distribution and Security

Ding Wang[†], Qianchen Gu[†], Xinyi Huang[‡], Ping Wang[†*]

[†]School of EECS, Peking University, Beijing 100871, China
[‡]School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China
[*]National Research Center for Software Engineering, Peking University, Beijing 100871, China
{wangdingg,guqianchen}@pku.edu.cn; xyhuang81@gmail.com; pwang@pku.edu.cn

ABSTRACT

Personal Identification Numbers (PINs) are ubiquitously used in embedded computing systems where user input interfaces are constrained. Yet, little attention has been paid to this important kind of authentication credentials, especially for 6-digit PINs which dominate in Asian countries and are gaining popularity worldwide. Unsurprisingly, many fundamental questions (e.g., what's the distribution that human-chosen PINs follow?) remain as intact as about fifty years ago when they first arose. In this work, we conduct a systematic investigation into the characteristics, distribution and security of both 4-digit PINs and 6-digit PINs that are chosen by English users and Chinese users. Particularly, we, for the first time, perform a comprehensive comparison of the PIN characteristics and security between these two distinct user groups.

Our results show that there are great differences in PIN choices between these two groups of users, a small number of popular patterns prevail in both groups, and surprisingly, over 50% of every PIN datasets can be accounted for by just the top 5%∼8% most popular PINs. What's disturbing is the observation that, as *online* guessing is a much more serious threat than *offline* guessing in the current PIN-based systems, *longer* PINs only attain marginally improved security: human-chosen 4-digit PINs can offer about 6.6 bits of security against online guessing and 8.4 bits of security against offline guessing, and this figure for 6-digit PINs is 7.2 bits and 13.2 bits, respectively. We, for the first time, reveal that Zipf's law is likely to exist in PINs. Despite distinct language/cultural backgrounds, both user groups choose PINs with almost the *same* Zipf distribution function, and such Zipf PIN-distribution from one source (about which we may know little information) can be well predicted by real-world attackers by running Markov-Chains with PINs from another known source. Our Zipf theory would have foundational implications for analyzing PIN-based protocols and for designing PIN creation policies, while our security measurements provide guidance for bank agencies and financial authorities that are planning to conduct PIN migration from 4-digits to 6-digits.

Keywords

User authentication; Personal identification number; Zipf's law; Online guessing; Markov-Chain-based cracking.

ASIA CCS '17, April 02-06, 2017, Abu Dhabi, United Arab Emirates

DOI: http://dx.doi.org/10.1145/3052973.3053031

1. INTRODUCTION

As one special kind of passwords, personal identification numbers (PINs) are typically composed of a fixed-length (e.g., 4 or 6) of digits and do not entail any letters or symbols. This makes them especially suitable for resource-constrained environments where users are only offered a numpad but not a common keyboard, such as automated teller machines (ATMs), point-of-sales (POS) terminals and mobile phones. While the past half a century has witnessed the competence of PINs in their banking role, PINs proliferate in a variety of new embedded applications like electronic door locks, SIM cards verification and mobile payment. As long as there are cases where the absence of a full keyboard prevents the use of textual passwords, PINs will remain an important authentication method in the foreseeable future.

There have been a number of standards (e.g., ISO 9564 [13] and the EMV standard [10]) that provide various security guidelines about PIN selection and management. Typical advice such as "select a PIN that cannot be easily guessed (i.e., do not use birth date, partial account numbers or repeated values such as 1111)" [10] might fail to be effective in practice, for they only enumerate *some* kinds of bad practice and never tell common users what constitute good PINs. Users have long been known to have difficulties in selecting and maintaining textual passwords [29, 35], and they are notoriously inclined to favor a small number of popular and predictable choices [5,33]. Expectedly, the immaturity of these primary PIN guidelines (e.g., [10,13]) can only add to concerns about what exactly the security that user-generated PINs can provide.

Despite the ubiquity of PINs, it was not until 2012 that the first academic research on human-chosen PINs was conducted by Bonneau et al. [6]. Particularly, they focused on 4-digit banking PINs that are used in America and Europe. As no real-life dataset of banking PINs has ever been publicly available, Bonneau et al. [6] employed two datasets of 4-digit sequences, which are extracted from 32 million Rockyou passwords and 205K smartphone unlock-codes, to approximate user choices of banking PINs. The underlying rationale is that, the digits and text of a textual password are *generally* semantically independent (despite of some exceptions like `june2001`), and these digit patterns and text patterns reflect distinct user behaviors. And thus the digit sequences existing in passwords highly reveal the user choices of their PINs.

Bonneau et al. [6] reported that Rockyou 4-digit PINs offer about 10.74 bits of entropy and iPhone 4-digit PINs offer about 11.01 bits of entropy. Following this approximating approach, a number of studies [4, 14, 20, 26] have been conducted on 4-digit PINs by using passwords: the work in [4, 20] examine the popular patterns that dominate user choices, while the work in [14, 26] propose new methods (e.g., mapping and morphing) for aiding users to memorize more randomly selected PINs.

1.1 Motivations

Though a handful of studies [4, 6, 14, 20, 26] have reported some results on human-chosen PINs, many important issues remain unresolved, e.g., *what's the distribution of human-chosen PINs? Do longer PINs generally ensure more security? Moreover, to the best of our knowledge, all previous studies deal with 4-digit PINs selected by American and European users, little is known about 6-digit PINs that dominate in Asia and are increasingly gaining popularity worldwide). What is the characteristics of 6-digit PINs and how is their security as compared to that of 4-digit ones?*

Particularly, Chinese users account for the world's largest Internet population (i.e., 710 million [1]) and largest consumer group of bank cards (i.e., 3.5 billion [24]), and they have been shown of great differences in choices when selecting passwords as compared to English users [21, 31], due to language, culture, economy and possibly many other confounding factors. *What about Chinese user PINs*? Are there any characteristics that differentiate Chinese user PINs from English user PINs? What are their strengths and weaknesses as compared to English user PINs? It is expected that, in a broad sense, settling these basic questions would contribute to a much better understanding of human-chosen credentials in terms of language, culture and the informatization process.

It has long been unrealistically assumed that PINs are uniformly distributed (e.g., "we assume a uniform (PIN) probability distribution in our experiments" [17] and "our models are correct under our assumption of uniformly distributed PINs" [27]). Rather than a theoretically desirable uniform distribution, we will show that in reality some PINs occur much more frequently than others. Moreover, both these overly popular and unpopular PINs are statistically significant in every PIN dataset, indicating that such a skewed distribution cannot be described by the common distribution models, such as uniform, normal, log-normal or exponential. Passwords have been found to follow the Zipf's law [31], yet this does not necessarily resolve the question of *what's the exact distribution that Human-chosen PINs follow*, because PINs are of a fixed length and with a much smaller character space. The settlement of this question would have foundational implications, ranging from security formulation of PIN-based cryptographic protocols, PIN creation policies to ecological validity of PIN user studies.

1.2 Contributions

In this work, we conduct the first systematic investigation into the two most widely used types of PINs (i.e., 4-digit and 6-digit ones) used by English and Chinese users, aiming to answer the above fundamental questions. Our key contributions are three-fold:

First, we compare the selection strategies of 4-digit PINs between English users and Chinese users, and initiate the study of human-chosen 6-digit PINs in both user groups. As semantic patterns are difficult to recognize from these massive and chaotic numbers, we resort to visualization techniques (i.e., heatmap and word cloud) and identify the semantic patterns in user choices more easily. As expected, simple patterns like years/dates, numpad-based numbers, digit repetition and sequential up/down are prevalent in both 4-digit and 6-digit PINs of each user group. By building effective models on the basis of semantic patterns observed, we manage to identify a number of unique structural and semantic characteristics that dwell in PINs of each user group, revealing distinct PIN selection behaviors between these two user groups.

Second, we, for the first time, shed light on the underlying distributions of user-chosen PINs by using natural language processing (NLP) techniques. We find that, despite the great differences in characteristics, PINs from these two distinct user groups exhibit quite a similar degree of self-organization: Zipf's law well applies to all our PIN datasets, and the corresponding parameters of their exact distribution functions are nearly the same. This not only outmodes the long-used assumption that PINs are uniformly distributed [12, 17, 27], but also has some foundational implications.

Third, we employ leading metrics (i.e., statistic-based α-guesswork [5] and cracking-based Markov algorithm [21]) to measure PIN strength. Our results show that, when online guessing is currently the primary threat, longer PINs essentially attain marginally (i.e., $<$1 bit) improved security (which is opposed to common belief): 4-digit PINs can offer about 6.6 bits of security against online guessing and 8.4 bits of security against offline guessing, while this figure for 6-digit PINs is 7.2 bits and 13.2 bits, respectively. This provides new insights into the relationship between length and strength of user chosen credentials.

Roadmap. We discuss related work in Sec. 2 and elaborate on PIN characteristics in Sec. 4. Sec. 5 devotes to understanding PIN distributions. Sec. 6 focuses on PIN strength. Sec. 7 concludes.

2. PIN USAGES AND PRIOR ART

Our work builds on a number of previous efforts. In this section, we first provide a panoramic sketch of PIN usages around the world, and then review the known research results on PINs.

2.1 PIN usages around the world

Initially, PINs were used in automated dispensing and control systems at petrol filling stations, and later on they were introduced to "the Chubb system" deployed by the Westminster Bank in the UK in 1967 [3]. Since then, PINs have been popular in the banking industry worldwide. With the rapid development of microelectronic technologies in the 1990s, various embedded devices emerge, and PINs act like passwords to safe guard these devices (e.g., PDAs and smart phones) from unauthorized access. Today, it is unsurprising to see many stores are equipped with a POS terminal to facilitate customers who have a bank card and a PIN.

While most mobile devices employ a 4-digit PIN, the lengths of banking PINs are much diversified, varying from country to country. Most banks in Europe allow for 4-digit PINs only. Most banks in America and Canada allow 4-digit PINs, and some banks (e.g., Bank of America and Royal Bank of Canada) begin allowing customers to use an up to 12-digit PIN number. South America countries like Brazil, Peru and Ecuador mainly employ 4-digit PINs, but some banks also accept up to 6 digits. In Australia and New Zealand, 4-digit PINs predominate, but many of the machines also accept up to 12 digits. Banks in Switzerland use 6 to 8 digit PINs, and banks in Italy typically use 5-digit PINs [6]. ATMs in Egypt and Nigeria only accept 4-digit PINs, while ATMs in South Africa accept both 4-digit and 5-digit ones.

Currently, most East and South Asia countries (e.g., China, Singapore, India, Indonesia and Malaysia) allow for a 6-digit PIN. Bank of Singapore allows 5-digit PINs. ATMs in Japan, South Korea, Thailand and Oman only take 4-digit PINs. But in recent years, there is a trend that many of these countries using 4-digit PINs would migrate to 6-digit PINs. For example, since Jan. 1st 2015 all UAE card-holders are needed to abandon their 4-digit PIN codes and use 6-digit PIN codes to make any purchases with their cards [19]; After the personal data of 20 million South Korean bank customers (i.e., 40% of Koreans) was leaked after a cyber-attack earlier in 2014, the national financial authority plans to introduce a 6-digit PIN system to "make banking more secure" [18].

In a nutshell, while most countries in Europe, America, South America, Africa and Oceania favor 4-digit banking PINs, most countries in Asia are currently using (or are going to migrate to) 6-digit PINs.[1] This implies that 6-digit PINs are now being used by nearly half of the world's population and deserves attention. Thus, in this work we focus on both 4-digit and 6-digit PINs.

[1] We thank many friends for helping identify PIN practices worldwide.

2.2 PIN characteristics and security

If you lose your ATM card on the street, what's the chance that someone correctly guesses your PIN and proceeds to clean out your savings account? The answer is 18.6%, with just three tries, according to Nick Berry, the founder of Data Genetics [4]. He extracted 3.4 million passwords with exactly 4-digit long from leaked password datasets like Rockyou and Yahoo (see Section 4), and used these extracted passwords to approximate 4-digit PINs, under the rationale that "if users select a four-digit password for an online account or other web site, it's not a stretch to use the same number for their four-digit bank PIN codes". He found that there is a "staggering lack of imagination" when users choose their PINs. The most popular PIN (i.e., `1234`) accounts for 10.71% of all the 3.4 million PINs collected, which is larger than that of the lowest 4,200 codes combined. The second most popular one is `1111` (6.01%), followed by `0000` (2%).

If your ATM card was lost in a wallet along with your identification card, what's the chance for someone to withdraw your money? In a seminal work, Bonneau et al. [6] combined their 2 million approximated 4-digit PINs (extracted from Rockyou and iPhone PIN codes) with the data obtained from a user survey with 1,108 effective US participants, and they estimated that this chance will be from 5.63% to 8.23%, depending on whether your bank has employed a blacklist to disallow weak PINs. They observed that dates dominate user choices of PINs, representing about 23% of users. Other popular patterns include sequential up/down, repetition, etc. They further employed statistical metrics [5] to assess PIN strength, and found that 4-digit banking PINs offer between 12.6 and 12.9 bits of security against offline guessing (which is acceptable), while the success rate of an attacker allowed up to 10 online guesses with the knowledge of birthday info reaches 8.9%.

As far as we know, the above works mostly focus on 4-digit PINs, some also deal with 5-digit PINs, yet little attention has been paid to 6-digit PINs which have been widely used in Asia and are gaining popularity worldwide. What's the characteristics of 6-digit PINs as compared to 4-digit ones? What's the distribution of 6-digit PINs? When measuring PIN security, existing studies only consider an optimal attacker, what's the PIN strength under the real-world attackers? Such basic questions all remain *unsolved*.

2.3 PIN distribution

What's the distribution that user-chosen PINs follow? It seems that this question is unlikely to be answered before a satisfactory solution has been provided to another question: What's the distribution that user-chosen passwords follow? In 2012, Malone and Maher [22] made an attempt to examine whether PINs follow a Zipf distribution, and they concluded that their password datasets are "unlikely to actually be Zipf distributed". In the meantime, Bonneau [5] also investigated the distribution of passwords and reported that a Zipf distribution is problematic for describing their password dataset, because the scale parameter of a Zipf distribution largely depends on the sample size and there is no meaningful way to determine a non-zero minimum password probability.

Different from the studies of [5, 22] that fit *all passwords* in the collected dataset into a Zipf model, the work by Wang et al. [31] first eliminates the least frequent passwords and then fit the remaining passwords to a Zipf model, and remarkably good fitting is achieved. Wang et al.'s underlying rationale is that, these *unfrequent passwords* are noise and do not show their true Zipf behavior due to the law of large numbers, and thus incorporating them into the fitting would only serve to conceal the good Zipf property of *frequent passwords*. They further provided compelling evidence that unfrequent passwords are also highly likely to follow the Zipf's law. More detailed justifications are referred to [31]. This idea has inspired our finding of Zipf's law in PINs.

3. METHODOLOGY FOR PIN CREATION

We now justify our PIN creation methodology, elaborate on the PINs creation process, and describe the resulting PIN datasets.

3.1 Why approximate PINs by passwords

As far as we know, no database of real-world banking PINs has ever been publicly available. Online or on-site user surveys might be conducted to collect some PINs, even a large number of participants can be recruited in through Amazon's Mechanical Turk crowdsourcing service. However, surveying on sensitive topics like web passwords is inherently subject to the ecological validity issue [11], let alone Banking/device-unlocking PINs. Fortunately, there is another source of PIN data. Dozens of high-profile web services (e.g., Dropbox and Yahoo [23]) have recently been hacked and billions of real-life passwords were leaked, and these datasets can be used to approximate PINs mainly due to three reasons.

Firstly, it is reasonable to assume that the digits and texts in a password are *generally* semantically independent. This assumption serves the foundation for the PCFG-based password cracking technique [34], which has been shown a great success for characterizing password selection [21, 33]. Our scrutiny into password datasets also confirms this assumption, despite that we come across a handful of passwords, such as `jamesbond007`, `obama2012` and `woaini1314`, with their digits and letters not independent.

Secondly, the user cognition capacity is rather limited: generally, a user's working memory can only manage a total of 5 to 7 chunks of independent information [16], and they will probably reuse PIN sequences as building blocks for their passwords.

Thirdly, our survey on password behaviors of 442 Chinese user reveals that 14.03% users re-use their banking PINs in web passwords [33], which well accords with English users: "over a third (34%) of users re-purpose their banking PINs in another authentication systems", including *web* services (15%) [6].

Thus, we use digit sequences with fixed length that are extracted from real-life password datasets on proxy of user-generated PINs.

3.2 How to approximate PINs by passwords

The idea of using passwords to approximate human choices of 4-digit PINs first appeared in [6], and this work has inspired a number of further investigations into PINs (e.g., [4, 26]). Initially, we attempted to repeat the experimentations in [6], but we soon were confronted with a difficult question: *how to extract PINs from textual passwords as there are 4+ different ways available*?

The first approach is to choose the passwords (e.g., `5683`) that *only* consist of 4 digits as PINs. In this case, password `12345` or `a1234` will be rejected. The second method is to select passwords that contain digits and the length of consecutive digits is exactly 4. For example, `a1234bc5678` is accepted and we can get two 4-digit PINs (i.e., `1234` and `5678`) from it. Meanwhile, `a12345` is rejected because the length of consecutive digits in it is 5, but not 4. The third way is to choose the passwords containing digits and the length of consecutive digits is no shorter than 4, and only the first 4 digits are used. For example, `a12345b` is accepted and we can get `1234` from it. The fourth way extends the third one, where one can get `1234` and `2345` from `a12345`. Maybe there are some other ways, here we have mainly considered these four cases.

As far as we know, the work in [4] prefers the first approach; the work in [6] favors the second approach; Stanekova and Stanek [26] employ both the second and third approaches. However, the rationale underlying their choices has never been given. In this work, we decide to choose the second one mainly for two reasons. Firstly, even though users may reuse their banking PINs in common sites, it is unlikely that such PINs are reused without any alteration (e.g., no appending, concatenation, insertion, or

Table 1: Basic information about the four real-life password datasets

Dataset	Web service	Location	Language	Original	Miscellany	Length>30	All removed	After cleansing	Unique Passwords
Dodonew	Gaming&E-commerce	China	Chinese	16,283,140	10,774	13,475	0.15%	16,258,891	10,135,260
CSDN	Programmer forum	China	Chinese	6,428,632	355	0	0.01%	6,428,277	4,037,605
Rockyou	Social and gaming forum	USA	English	32,603,387	18,377	3140	0.07%	32,581,870	14,326,970
Yahoo	Portal(e.g., E-commerce)	USA	English	453,491	10,657	0	2.35%	442,834	342,510

Table 2: Basic information about the derived PINs

	Dodonew	CSDN	Rockyou	Yahoo
Total 4-digit PINs	1,223,677	444,204	1,780,587	47,540
Unique 4-digit PINs	10,000	9,951	10,000	8,379
Coverage(4-digit PIN)	100.00%	99.51%	100.00%	83.79%
Total 6-digit PINs	2,876,047	809,899	2,758,491	21,020
Unique 6-digit PINs	465,741	224,250	448,186	14,001
Coverage(6-digit PIN)	46.57%	22.43%	44.82%	1.40%

duplication), and only those with the lowest security-consciousness would directly use a 4-digit PIN as their passwords. Therefore, one may underestimate the security of PINs by using the first approach.

Secondly, though users reuse PIN sequences as building blocks for their passwords, so far there has been no information about how PINs are built into passwords. For instance, it remains an open problem as to whether there are priorities when users applying the alteration strategies? Before this question is answered, the applicability of the PIN extraction approaches like the third one and the fourth one (which may largely overestimate or underestimate PIN usages in passwords) cannot be assessed.

Consequently, we prefer a conservative approach, i.e., the second one: we extract all consecutive sequences of *exactly* 4 and 6 digits from our four password datasets and obtain eight corresponding PIN datasets. Though we get fewer PINs as compared to the third, fourth or other approaches, we avoid destroying the original connotation in digits and introducing uncertainties. For example, in the fourth approach, one can get `1994`, `9940`, `9404`, `4041`, `0411` from a birthday `19940411`. One may get even more useless information from a telephone number in the third approach. Such approaches will surely lead to a skew of the PIN distribution. Meanwhile, we have to admit the disadvantage of our selected approach – it conservatively ignores some PINs dwelling in passwords. As for 6-digit PINs, a similar approach is taken.

3.3 PIN datasets description

Table 1 summarizes some basic info about the four password datasets we collected. They were all hacked by adversaries from prominent sites and somehow made public over the Internet. They have been widely used in research [9, 21, 31, 33]. We first carry out data cleansing for each dataset. Email addresses, user names (and other non-password info) are removed from the original data. We then observe that some strings containing non-ASCII letters (i.e., "miscellany"), and they are unlikely to be passwords and thus are purged. We further remove strings whose lengths are abnormally long (i.e., >30), because they are more likely to be generated by password managers than by human-beings. This process not only increases data quality but also eases later data operations.

Then, we extract all consecutive sequences of *exactly* 4 and 6 digits from these four password datasets, resulting in eight corresponding PIN datasets as summarized in Table 2. While Bonneau et al. [5] observed that users employ 4-digit sequences significantly more often than 3- and 5-digit sequences in passwords, Table 2 shows that users manifest a particular affinity for 6-digit sequences even over 4-digit ones. We leave to future research the interesting question of to exactly what extent our created PIN datasets are comparable to real corpus of PINs used in the banking context.

4. PIN CHARACTERISTICS

We now systematically investigate into the characteristics of 4-digit and 6-digit PINs generated by English and Chinese users.

4.1 Characteristics of 4-digit PINs

Table 3 lists the top ten 4-digit PINs in our datasets. For the two Chinese datasets, `1234`, `1314`, `2008` occupy the top three positions. It is not a surprise to see `1234` being among the most popular PINs, yet the popularity of `1314` and `2008` is a bit puzzling. Then we realize that, `1314` sounds like "forever and ever" in Chinese, and "2008" is just the year where the 29^{th} Summer Olympic Games were held in Beijing, China. For English PINs, the microscopic picture differs: `1234` is indisputably the most popular one, while the rest of the top-10 list are completely occupied by years ranging from `1991` to `2009`, being consistent with [6].

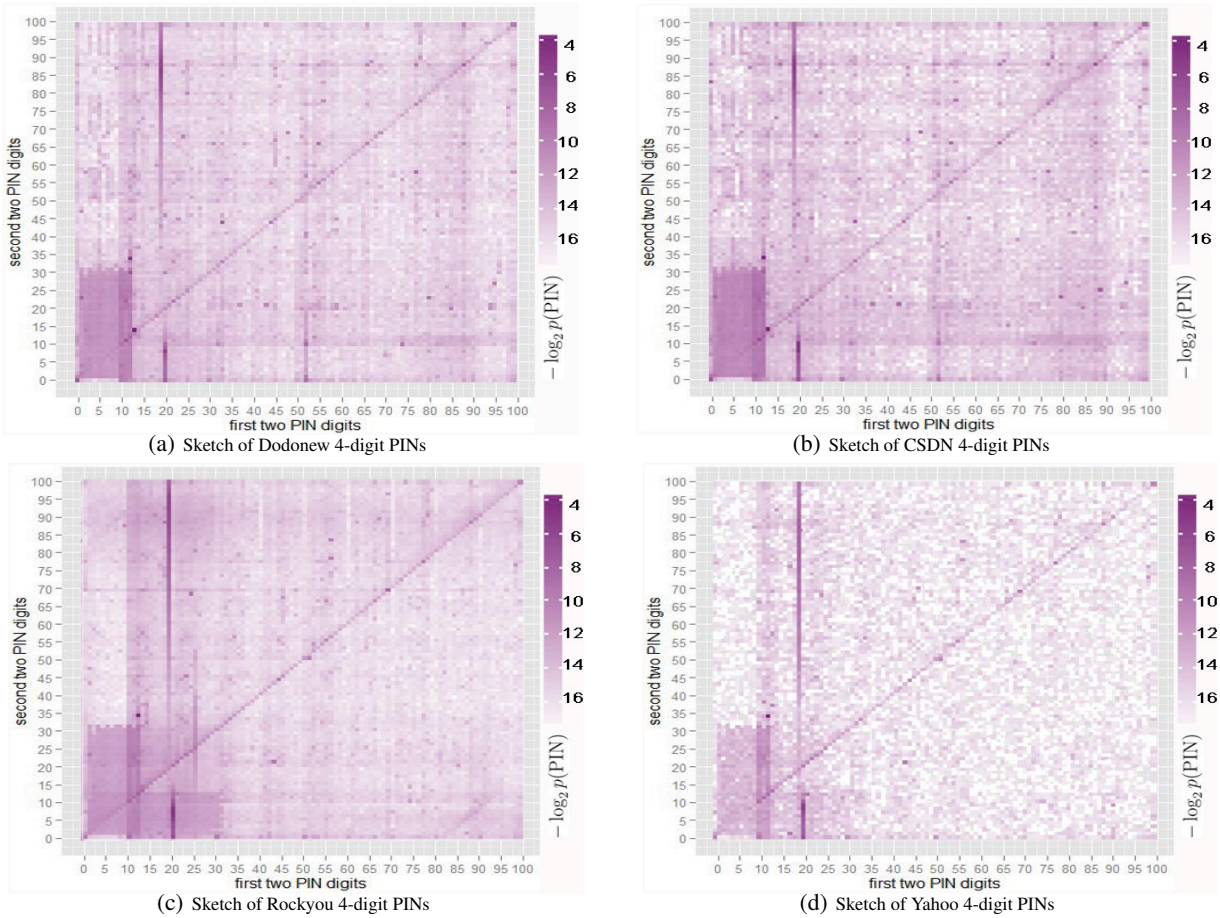
Table 3: Top ten 4-digit PINs in each PIN dataset

Rank	Dodonew		CSDN		Rockyou		Yahoo	
1	**1314**	7.25%	**1234**	5.91%	**1234**	3.72%	**1234**	4.51%
2	**1234**	3.45%	**1314**	4.57%	2007	2.23%	2008	2.13%
3	**2008**	2.09%	**2008**	2.70%	2006	2.10%	2009	2.06%
4	1987	2.06%	2010	2.31%	2008	1.73%	2007	1.32%
5	1986	1.82%	2009	2.21%	2005	1.33%	2000	0.99%
6	1988	1.58%	1987	1.86%	1994	1.18%	2006	0.97%
7	1989	1.47%	1988	1.76%	1993	1.13%	2005	0.77%
8	1985	1.43%	1989	1.71%	1992	1.13%	2004	0.66%
9	1984	1.21%	1986	1.36%	1995	1.06%	2002	0.61%
10	1990	1.01%	1985	1.03%	1991	1.02%	2001	0.59%
Total	112,917	25.42%	285,973	23.37%	296,112	16.63%	6,946	14.61%

What's staggering is that, over 23.37% of all 4-digit Chinese PINs could be guessed by just trying these 10 combinations! This figure for English PINs is 14.61%, much lower than that of 4-digit Chinese PINs, yet it is still alarming. Statistically, if 4-digit PINs were uniformly distributed (i.e., with 10^4 possible combinations), we would expect these ten PINs to account for just 0.1% of the total datasets, but not 23.37% or 14.61% as we have actually encountered. While "it's amazing how predictable (US) people are" [4], it is utterly incredible to see how the lack of imagination Chinese users are! In the NIST SP-800-63-2 standard for passwords [7], password security can be largely improved by imposing a policy that disallows overly popular passwords. A similar policy would produce PIN distributions with much better resistance to guessing, yet no standards/guidelines for PINs mention such a policy [6] and as far as we know, `1234` (or `123456`) is allowed on nearly all ATMs and mobile devices. Our Zipf theory in Sec. 5 implies that a threshold-based blacklisting approach is much better than simply blacklisting all popular PINs (see Appendix A), due to the *polynomially* decreasing nature of PIN frequency.

After having gained a concrete grasp of the most popular PINs in each dataset, we now employ a fundamentally different approach to examining semantic patterns and reveal the user aggravated behaviors by plotting each PIN distribution in a 2-dimensional grid (see Fig. 1) using R, a free software programming environment for statistical computing and graphics. Each grid uses the first two PIN digits as the x-axis and the second two PIN digits as the y-axis, and color is employed to represent frequency: the higher the frequency of a PIN is, the darker the color in a cell will be. And thus such grids are also called heatmaps. This idea is inspired by Bonneau et al. [6], and it allows an informative view of the whole PIN dataset. This is in stark contrast to most previous approaches that use statistics to just catch a glimpse of one corner of a dataset.

Fig. 1 illustrates several important features that lie in all four PIN datasets. Firstly, the left bottom corner of four mini-pictures are invariably with a darker color, indicating the popularity of calendar



(a) Sketch of Dodonew 4-digit PINs

(b) Sketch of CSDN 4-digit PINs

(c) Sketch of Rockyou 4-digit PINs

(d) Sketch of Yahoo 4-digit PINs

Figure 1: A visualization of the distribution of 4-digit PINs. The darker the color, the more popular the PIN is.

dates in a MMDD or DDMM format. Being more careful, one can trace the variation in lengths of each month (e.g., there are only 28 days in Feb. and 31 days in Jan., Mar., May, etc.). Secondly, the lines corresponding to `19XX` and `20XX` show that users love to use the year of birth (or possibly the registration year) as their PINs. Thirdly, a diagonal line of PINs with the same first and last two digits (e.g., `3737`) can be clearly identified. What's more, a number of insular cells (e.g., `4869`, `1412`, `5683` and `2580`)[2] stand out like stars in the sky, which reveals some independent events (e.g., homonyms of characters famous novels/cartoons, theme of love and numpad patterns) that may influence PIN popularity. In all, while some PINs stand out as independent events, some other ones exhibit a warp and woof of woven fabric divulging certain subtle links we are unaware of. All this highlights the effectiveness of visualization in the early stage of data analysis, especially helpful for eliminating unnecessary experiments.

There are also some substantial differences in the PIN distributions between both user groups. Perhaps the most obvious one is that, Chinese users prefer the MMDD date format, while English users equally favor the MMDD and DDMM date formats. There are more "stars" standing out in Chinese PINs as compared to English PINs: nearly all the "stars" (e.g., `2580`, `1357`, `4869` and `2468`) appearing in English PINs have also emerged in Chinese PINs, while many "stars" (e.g., `1314`, `3721`, `9527` and `2046`) appearing in Chinese PINs have not emerged in English PINs. Furthermore, English users like to end their PINs with the number `69`, which demonstrates users' affinity and see Fig. 1(c) for the horizontal line of Rockyou PINs. This observation is consistent with [4, 6, 20]. In contrast, Chinese users love to begin their PINs with `52` (which sounds as "I love ...") and to end their PINs with `88` which sounds as "making a fortune". All these highlight some basic linguistic/cultural factors that influence user PIN choices.

The frequent showing up of some numpad-based PINs (e.g., `2580` and `5683`), to some extent, suggests the effectiveness of extracting PINs from textual passwords. We notice that some full-size keyboard for PCs and laptops do also have a numpad, yet the digits on such PC numpads are often inversely arranged (e.g., `7`, instead of `1`, is on the top-left) as compared to digits on numpads for embedded devices. It is, therefore, not as convenient for users to type `2580` on a PC numpad as on an ATM numpad: the former first involves a bottom-up and then a jump to 0, while the latter only involves a vertical top-down swipe. As for `5683`, it can be mapped to "love" on an ATM numpad, but with no evident meaning on a PC numpad. Thus, their frequent showing up can, arguably, confirms the validity of our data. As we will show in Sec. 4.2, numpad-based PINs are even much more popular in the extracted 6-digit PINs.

To quantitatively measure the influencing factors that dominate the user selection of 4-digit PINs, we devise a model (see Table 4) that contains 12 patterns observed above. Note that, "YYYY" stands for a year format like `2008`, and here we only consider the years after 1940 according to our observations from the heatmaps. "Chinese elements" are composed of fourteen PINs, including ten PINs sounding like meaningful phrases in Chinese (i.e., `1314`, `3344`, `5200`, `5210`, `9420`, `8520`, `5257`, `8023`, `7758`, `9958`),

[2] "4869" relates to the famous cartoon character Sherlock Holmes and Conan, "1412" to Magic Kaito; For "5683", on a numpad, "5" can be mapped to "L", "6" to "O", "8" to "V", "3" to "E", which make up "LOVE"; "2580" is the obvious pattern "|" on ATM/phone-style numpads.

Table 4: A simple model for evaluating patterns in Human-chosen 4-digit PINs (with a focus on Chinese user PINs)

Patterns in our model	Random model		Dodonew		CSDN		Rockyou		Yahoo	
	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs
All 4-digit PINs	10000	100%	1223677	100.00%	444204	100.00%	1780587	100.00%	47540	100.00%
YYYY (1940∼2014)	75	**0.75%**	297310	**24.30%**	123635	**27.83%**	560003	**31.45%**	12688	**26.69%**
MMDD (e.g., 0406, 1230)	365	3.65%	278441	22.75%	119304	26.86%	279606	15.70%	6572	13.82%
One digit repeated (e.g.,1111)	10	0.10%	26906	2.20%	11617	2.62%	14975	0.84%	676	1.42%
Numpad pattern (e.g., 2580, 1357)	68	0.68%	19845	1.62%	4298	0.97%	11905	0.67%	298	0.63%
Sequential up/down (e.g., 1234, 7890)	14	0.14%	50879	4.16%	29840	6.72%	75674	4.25%	2421	5.09%
Chinese elements (e.g., 1314, 5210)	14	**0.14%**	110830	**9.06%**	26185	**5.89%**	3722	0.21%	60	0.13%
Total of the above six patterns*	541	**5.41%**	763019	**62.35%**	308233	**69.39%**	938041	**52.68%**	22477	**47.28%**
DDMM (e.g., 0604, 3012)	365	3.65%	191975	15.69%	99260	22.35%	433474	24.34%	9387	19.75%
Couplets repeated (e.g., 1616)	90	0.90%	29413	2.40%	9592	2.16%	59247	3.33%	1708	3.59%
Palindrome (e.g., 1221, 2442)	100	1.00%	48328	3.95%	18985	4.27%	62320	3.50%	1806	3.80%
Beginning with 52 (e.g., 5211)	100	1.00%	19640	1.61%	4155	0.94%	7838	0.44%	245	0.52%
Ending in 88 (e.g., 7688, 5088)	100	1.00%	35091	2.87%	14539	3.27%	28818	1.62%	632	1.33%
Universal elements (e.g., 4869, 5683)	5	0.05%	2926	0.24%	1306	0.29%	5504	0.31%	115	0.24%
Total of the above six patterns*	738	**7.38%**	296539	24.23%	133988	30.16%	559622	31.43%	12751	26.82%
Random(i.e., beyond the above 12 patterns)	8894	**88.94%**	391981	32.03%	116724	26.28%	658093	36.96%	20922	44.01%

*As there are ambiguities when determining to which pattern a PIN belongs to, we manually check: if a PIN shows an obvious pattern tendency, it only belongs to that pattern; if there is almost equal tendency for 2+ patterns, it is counted by these 2+ patterns. To avoid duplicate counting, we compute the two "total" statistics in a top-down order of the patterns as arranged in the table. Once a PIN matches with a pattern, then this PIN is marked as counted.

two PINs related to classic Chinese movies (i.e.,`2046`, `9527`) and two PINs which are names of popular sites in China (i.e., `3721`, `8848`). "Universal elements" consist of five PINs, including two PINs related to world-wide famous Characters in cartoons/novels (i.e., `4869` for Sherlock Holmes and Conan, `1412` for Magic Kaito), one PIN related to love (i.e., `5683` as said earlier) and two PINs related to odd/even sequential numbers (i.e., `2468`, `1357`).

It is worth noting that, in this section (and Sec. 4.2) we take PINs from Chinese user as a case study and mainly focus on devising an effective model for evaluating popular patterns in them. It is not difficult to see that a fine-grained model for English user PINs can be constructed in a similar way. Though our model is generalized mainly from two Chinese datasets (assisted with two English ones), we believe that *it is of universal applicability for Chinese user PINs due to the generality nature of each of its elements*.

We also note that some digit sequences may match several different patterns. For example, `0123` collides with the patterns "sequential numbers" and "MMDD". However, it is more likely that users choose it mainly because it is a memorable sequential number. Meanwhile, there also exist many sequences which don't show an obvious tendency. For example, `1221` matches both "MMDD" and "palindrome", yet we cannot determine its bias towards which pattern. Thus, we prefer to not deal with the ambiguities. For example, `0123` will be deemed as a date *and* as a sequential number. To avoid duplicates when computing the two "total" statistics, we adhere to the notes under Table 4.

Table 4 shows that one could guess about a quarter of Chinese PINs (as well as English PINs) by just trying a set of 75 years ranging from 1940 to 2014. Fourteen Chinese elements account for 9.06% of the Dodonew PIN dataset and 5.89% of the CSDN PIN dataset, respectively. Interesting, the single Chinese element (i.e., `1314`) makes up 7.25% of the Dodonew PINs and 5.91% of the CSDN PINs, respectively. Even this element covers 0.12% of the Rockyou PINs, twelve times higher than a random PIN should do. This suggests that there would be a non-negligible portion of Chinese users who register in the English site Rockyou. While Chinese users prefer the date pattern "MMDD", their English counterparts favour the date pattern "DDMM". Both groups of users equally like to employ simple patterns such as "one digit repeated", "sequential up/down", "couplets repeated" and "palindrome".

It is staggering to see that a small set of 541 4-digit PINs (constructed from the top-6 patterns) can account for 62.35% and 69.39% of the entire Dodonew PIN dataset and CSDN PIN dataset, respectively. The results for English 4-digit PINs are also impressive: about 50% are covered by 5.41% of all 10^4 possible combinations. Further with six other patterns, over 67.97% and 73.72% of these two Chinese PIN datasets can be covered, respectively. If we had combined more complex patterns (e.g., YYMD and vertical swipe) with this model, even more PINs could be covered. However, we intentionally donot focus on using such complex patterns, because we aim to show that even using a simple model like ours, which is merely comprised of a few simple patterns, can successfully cover a significant fraction of the PINs. We also find that PINs from each dataset offer significantly different semantic distributions (pairwise χ^2 test, $p < 0.01$).

Actually, it is interesting to see that our model is also quite suitable for characterizing English PINs (e.g., 52.68% Rockyou 4-digit PINs can be covered by the first six patterns), and this figure would be higher if we had taken into account the English elements (e.g., popular dates that go beyond birthdays include historical years `1492` and `1776`, and the number `007` for James Bond).

Summary. Whereas there are some notable differences in 4-digit PIN choices between English users and Chinese users, both groups of users tend to choose PINs in a predictable way. The effectiveness of our relatively simple model suggests that, the identified twelve general patterns well reveal the behaviors when users selecting their 4-digit PINs. In Section 7, we will show the security vulnerabilities associated with such (weak) patterns in user-generated PINs.

4.2 Characteristics of 6-digit PINs

To the best of our knowledge, so far there has been no published investigation into the domain of 6-digit PINs, though they have long been widely used by billions of card holders in the world, especially in Asia. Still, some conjectures about 6-digit PINs have been made, such as "the greater the number of digits required, the more predictable PIN selections become" [4, 20]. Will this conjecture be true? Are there any prominent features of human-chosen 6-digit PINs as compared to 4-digit PINs? In the following, besides suggesting compelling answers to these two questions, we also make the first attempt to identify the dominant factors that influence user behaviors of 6-digit PIN choices.

Table 6 shows the top-10 most popular PINs in each 6-digit PIN dataset. As expected, `123456` tops the list, followed by `111111` and `123123`. Surprisingly, these top-3 PINs can occupy from 12.60% to 21.21% of our PIN datasets. Such figures are much higher than those of the top three 4-digit PINs as shown in Table 3. Based on similar observations, the works in [4, 20] conjectured that "the greater the number of digits required, the more predictable PIN selections become". We confirm that this is largely true as we will demonstrate in the following explorations.

Table 5: A simple model for evaluating patterns in Human-chosen 6-digit PINs (with a focus on Chinese user PINs)

Patterns in our model	Random model		Dodonew		CSDN		Rockyou		Yahoo	
	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs	# of matched PINs	% of all PINs
All 6-digit PINs	1000000	100%	2876047	100.00%	809899	100.00%	2758491	100.00%	21020	100.00%
YYYYMM (194001∼201412)	900	0.090%	35708	1.24%	11396	1.41%	17048	0.62%	95	0.45%
YYMMDD (400101∼141231)	27375	2.738%	648974	**22.56%**	236572	**29.21%**	277568	10.06%	1824	8.68%
YYYYMD (194011∼201499)	6075	0.608%	149789	5.21%	43237	5.34%	54005	1.96%	267	1.27%
One digit repeated (e.g., 111111)	**10**	**0.001%**	123109	**4.28%**	14126	**1.74%**	63176	**2.29%**	610	**2.90%**
Numpad patterns (e.g., 147258)	262	**0.026%**	706938	**24.58%**	122002	**15.06%**	425381	**15.42%**	3025	**14.39%**
Sequential numbers (e.g., 123456)	**11**	**0.001%**	503972	**17.52%**	90692	**11.20%**	351008	**12.72%**	2506	**11.92%**
Chinese elements (e.g., 585520)	**21**	**0.002%**	63963	**2.22%**	10846	**1.34%**	927	**0.03%**	47	**0.22%**
Total of the above 7 patterns*	34491	**3.449%**	1623168	**56.44%**	423948	**52.35%**	823927	**29.87%**	5644	**26.85%**
YYDDMM (400101∼143112)	27375	2.738%	296178	10.30%	103707	12.80%	280387	10.16%	2139	10.18%
MMYYYY (011940∼122014)	900	0.090%	3032	0.11%	1036	0.13%	37488	1.36%	358	1.70%
MMDDYY (010140∼123114)	27375	2.738%	180278	6.27%	42942	5.30%	787165	**28.54%**	6347	**30.20%**
MDYYYY (111940∼992014)	6075	0.608%	12861	0.45%	3772	0.47%	81640	2.96%	773	3.68%
Couplets repeated (e.g., 121212)	90	0.009%	18000	0.63%	3553	0.44%	62112	**2.25%**	621	**2.95%**
Double sequential (e.g., 112233)	17	0.002%	13939	0.48%	3289	0.41%	9671	0.35%	62	0.29%
Triple repeated(e.g., 136136)	990	0.099%	135250	4.70%	25597	3.16%	41187	1.49%	498	2.37%
Triple sequential (e.g., 111222)	19	0.002%	6775	0.24%	1806	0.22%	4934	0.18%	59	0.28%
Palindrome (e.g., 123321, 179971)	1000	0.100%	155827	5.42%	23723	2.93%	80606	2.92%	787	3.74%
Universal elements (e.g., 314159)	9	0.001%	3890	0.14%	2907	0.36%	4250	0.15%	42	0.20%
Total of the above 10 patterns*	57606	**5.761%**	632258	**21.98%**	177879	**21.96%**	1095346	**39.71%**	9142	**43.49%**
Random(i.e., beyond the above 17 patterns)	919094	**91.909%**	1068148	37.14%	328390	40.55%	1094711	39.69%	8202	39.02%

*To avoid duplicate counting, we employ the same method as in Table 4.

Table 6: Top ten 6-digit PINs in each PIN dataset

Rank	Dodonew		CSDN		Rockyou		Yahoo	
1	123456	17.05%	123456	10.73%	123456	11.69%	123456	11.05%
2	111111	2.15%	123123	0.97%	654321	0.56%	111111	0.92%
3	123123	2.01%	111111	0.68%	111111	0.51%	123123	0.63%
4	000000	1.29%	123321	0.53%	000000	0.49%	654321	0.54%
5	**321654**	1.00%	000000	0.47%	123123	0.40%	000000	0.41%
6	123321	0.55%	654321	0.27%	666666	0.29%	**030379**	0.39%
7	**520131**	0.53%	112233	0.24%	121212	0.21%	666666	0.31%
8	**520520**	0.44%	**123654**	0.21%	112233	0.21%	123321	0.30%
9	112233	0.30%	520520	0.21%	**789456**	0.21%	121212	0.29%
10	**147258**	0.30%	666666	0.20%	**159753**	0.20%	**101471**	0.28%
Total	736,843	25.62%	117,516	14.51%	407,429	14.77%	3,178	15.12%

Every PIN in Table 6 (except for these in bold) conforms to one of three basic patterns: digit repetition, sequential up/down and palindrome. As for these ten bolded PINs, six PINs (i.e., `147258`, `123654`, `321654`, `789456` and `159753`) obviously comply with a numpad pattern (e.g., `159753` is a "×" mark over the numeric keypad), two PINs (i.e., `520520`, `520131`) intriguingly sound like "I love you ⋯" in Chinese and two PINs from Yahoo dataset (i.e., `030379`, `101471`) seem to be of no obvious meaning or simple patterns. It is alarming to see that, for every 6-digit PIN dataset, its top-10 PINs can account for more than a seventh of the entire dataset. Still, this figure is of no significant difference as compared to the top-10 4-digit PINs (see Table 3).

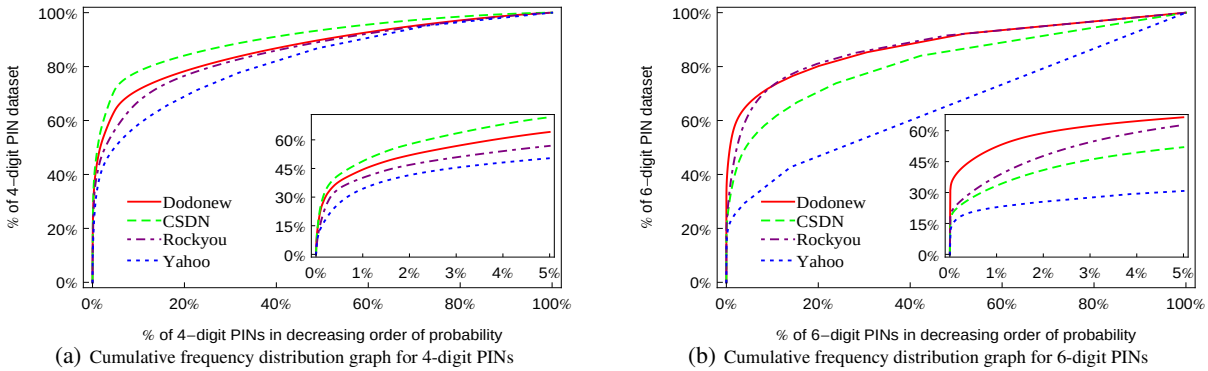
To further identify the dominant factors that influence user choices of 6-digit PINs, we once again resort to a visualization technique (i.e., word cloud) due to its intuitiveness and informativeness, and make use of the wordle diagram http://www.jasondavies.com/wordcloud/. We provide the raw PINs from each dataset to the wordle tool which gives greater prominence to PINs that are more frequent. As a result, the PINs shown in the cloud picture are sized according to the number of their occurrences.

Due to space constraints, the word clouds for our four 6-digit PIN datasets can be found in Appendix B. One can see that, patterns like dates and single-digit/couplets/triple repetition are as prevalent as that of 4-digit PINs. Interestingly, users prefer using pairs of numbers that have smaller space gaps between them. For example, combinations like 12 and 78 are used much more frequently than 17 and 28. One plausible reason is that a smaller space gap is easier to type. As with 4-digit PINs, meaningful 6-digit PINs are favored by either Chinese users (e.g., `131452` and `110120`) or English users (e.g., `420420` and `696969`), or both groups of users (e.g., `007007`). Note that, `110` and `120` are the alarm/emergency call in China; `420` is a code for marijuana.

What's quite different from 4-digit PINs is that *numpad patterns* seem to be much more popular in 6-digit PINs: about 26% to 36% of the top-150 PINs in each dataset are with some kind of numpad patterns like "×" (e.g., `159753`), "+" (e.g., `258456`), "=" (e.g., `123654`), "≡" (e.g., `134679`), "⊤" (e.g., `123580`) and "‖" (e.g., `147369`). Interestingly, once again we see corroborative evidence that extracting exactly 6-digit sequences from passwords is a good proxy for 6-digit real-world PINs. One can see that many of these numpad-pattern PINs (e.g., `142536` and `258456`) are utterly awkward to type on a laptop/PC keyboard because the order of these digits is interlaced on such keyboards. As said earlier, even though some PC keyboards also have a numpad area, yet the digits on such PC numpads are often inversely arranged as compared to a phone-style numpad. Thus, these PINs (e.g., `123580` and `123654`) are still inconvenient to type on a PC-based numpad, yet a phone-style numpad just facilitates such digit sequences. This suggests that, besides their preference of maintaining easy to type (and remember) PINs for their credit cards/mobile devices, many users also seem to tend to re-use their PINs in online passwords, even though the mnemonics related to numpads (e.g., `123580` relates to a "⊤") no longer apply to the PC keyboards/numpads.

To quantitatively measure user behaviors, we devise a simple model (see Table 5) that contains 17 patterns observed above. Here we only consider the years after 1940 due to the distribution of years revealed in Fig. 1. The "numpad pattern" incorporates $262=(16^2+6)$ PINs, for there are sixteen 3-digit numpad sequential numbers (i.e., "`123`, `456`, `789`, `147`, `258`, `369`, `159`, `357`; `321`, ⋯, `753`") and we also consider six hybrid numpad sequences such as `159874`. "Chinese elements" are composed of 21 PINs, including 18 PINs (see Appendix C) which sound as meaningful phrases in Chinese and three PINs which are combinations of well-known calls in China (i.e., `110120`, `110110`, `110119`). "Universal elements" consist of 9 individual PINs, including 5 PINs (i.e., `112358`, `314159`, `141592`, `271828`, `142857`) which are important constant numbers and popular in our datasets, two PINs relate to odd/even sequential numbers (i.e., `135790`, `246810`), one PIN stands for James Bond (i.e., `007007`) and one PIN (i.e., `911911`) relates to the U.S. emergency number.

For better comprehension, we also provide the combined dictionaries (with duplicates removed) in Table 5. Our results show to what extent human-beings are lacking of imagination: with just seven popular patterns observed, we can cover over 50% of Chinese user PINs and 25% of English user PINs by using a small dictionary that consists of only 3.449% of all the 10^6 possible



(a) Cumulative frequency distribution graph for 4-digit PINs

(b) Cumulative frequency distribution graph for 6-digit PINs

Figure 2: The percentage of PIN dataset being covered by the top x% PINs. The 20/80 rule is evident.

PINs; Further with ten more patterns, we can cover about 60% of PINs of both user groups by using a dictionary consisted of 8.09% of the 10^6 possible PINs. Though popular PINs in 6-digit datasets are more concentrated than that of 4-digit datasets, yet there are larger fractions of random 6-digit PINs, implying that 6-digit user-chosen PINs are more secure against offline guessing. This partially confirms the conjecture made in [4, 20].

In 6-digit PINs, years and dates are not as popular as that of 4-digit PINs, yet three other patterns (i.e., "one digit repeated", "numpad patterns" and "sequential numbers") are much more prevalent. For example, a mere 11 sequential numbers can cover over 11.20% of every PIN dataset. Besides, meaningful Chinese elements (e.g., `520134`) are abundant and they also occupy a non-negligible fraction of English PINs (up to 0.22%), implying that many users in English sites are from China. As for "universal elements" (e.g., `007007`), they are so dangerously popular that an attacker's cost/success ratio can be as low as $1/360 \sim 1/140$. We also find that 6-digit PINs from each dataset offer significantly different semantic distributions (pairwise χ^2 test, $p < 0.01$).

Particularly, as high as 14.39%∼24.58% of PINs of every dataset can be guessed by the 262 numpad-pattern PINs (which are a mere 0.026% of all possible 10^6 6-digit PINs), indicating an over 553(=14.39/0.026) times increase in success rates than 262 random PINs should do. In 2014, Das et al. [8] reported that about 43%∼51% of users re-use their textual passwords across various sites, while our results arguably imply that a non-negligible fraction of "persistent" users (1.31%∼4.35%) who reuse their 6-digit PINs in textual passwords even though the mnemonics related to numpads no longer apply to laptop/PC keyboards/numpads. An important caveat is that some of these 262 PINs can also be based on other patterns such as "sequential up/down" (e.g., `123456`) and "universal element" (e.g., `135790`), meaning some overestimation of these "persistent" users. Still, most of these 262 PINs mainly facilitate typing on a numpad (especially unfriendly on a laptop), and *this result provides compelling evidence for PIN reuse in passwords and highlights this highly vulnerable human behaviors*.

We note that the coverage of our evaluation model may possibly be expanded by leveraging some complex patterns (e.g., a mixture of odd and even sequences like `135246`), but its effectiveness heavily depends on the target data. We intentionally do not incorporate such complex patterns, because *we aim to show that our simple model can well capture the dominant factors that influence PIN choices of Chinese users* (which provides substantial insight into users' PIN selection process). That being said, an evaluation model for 6-digit PINs of English user can be built in a similar way.

Summary. As compared to 4-digit PINs, 6-digit ones are more likely to be of numpad-based patterns, language-based specific elements and sequential numbers. While popular 6-digit PINs are more concentrated than 4-digit ones, a larger fraction of 6-digit PINs do not follow any obvious pattern. This has critical *real-world implications*: 6-digit PINs are more prone to small number of guessing attempts (i.e., online guessing [33]) yet more secure against larger numbers of guessing (i.e., offline guessing). Since the former threat is a much more realistic and serious one, this calls in question the necessity of migration to longer PINs (e.g., [18]).

5. PIN DISTRIBUTION

After having gained a comprehensive grasp of the PIN characteristics and seen that some PINs occur significantly more frequently than others, one may naturally wonder a more fundamental question: *what is the exact distribution that PINs follow*? Since the PIN frequency distribution indicates the degree of PIN concentration, the settlement of this question would have foundational implications for PIN-based cryptographic protocols, PIN strength meters, creation policies and ecological validity of PIN user studies. In this section, we make the first attempt to address this issue.

5.1 Cumulative frequency distribution

In Section 4, we have seen the frequency distributions of top-10 PINs from both user groups. How about top-100 PINs, top-10^3 PINs, and so on? We answer this question by presenting a cumulative frequency distribution graph (see Fig. 2), where the x-axis is the top x% of PINs and the y-axis is the percentage of total datasets covered by these top x% of PINs. Statistically, the top-100 PINs (i.e., top 1% for 4-digit PINs and top 0.01% for 6-digit PINs) of each PIN dataset can cover at least 30% and 40% of the 4-digit datasets and 6-digit datasets, respectively.

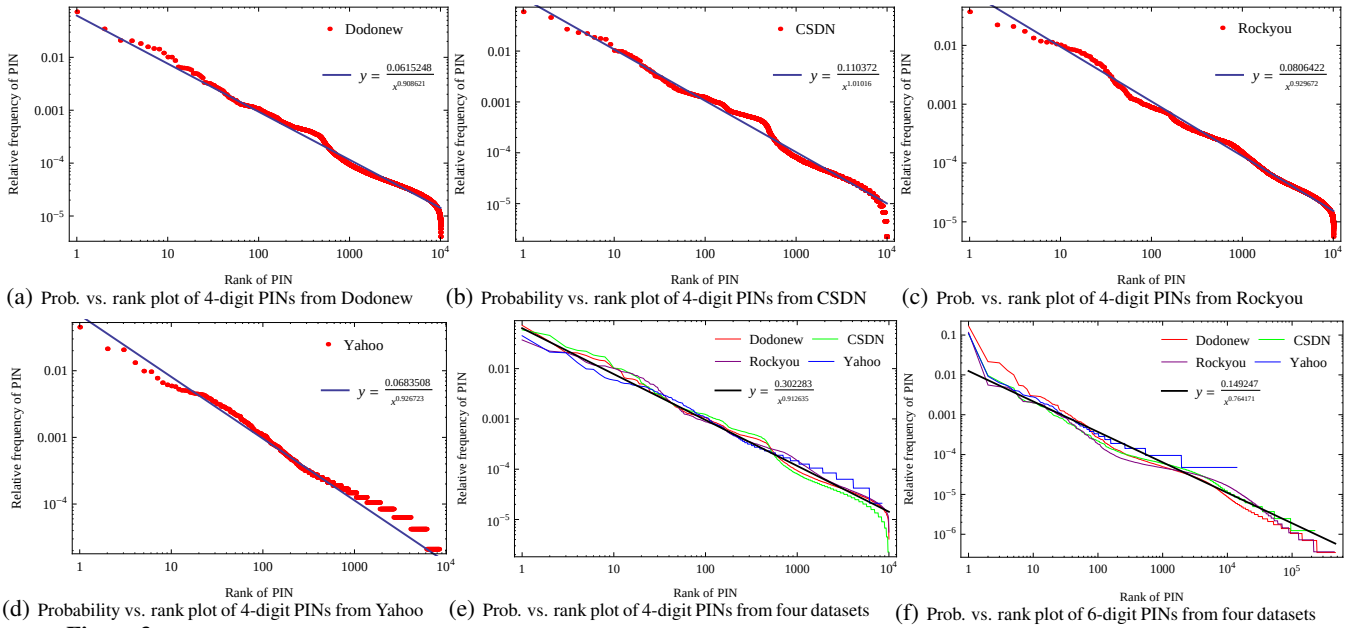
Alarmingly, the 50% cumulative chance threshold of 4- and 6-digit PINs (except for Yahoo) is passed at just the top 2.75% and top 2.35%, respectively. This indicates a 18.18(=50/2.75) and a 21.28 (=50/2.35)) times increase in an attacker's success rates, respectively, if she somehow knows the underlying PIN distributions.

5.2 Frequency distribution

The above CDF graphs (see Fig. 2) show that *both these overly popular PINs and unpopular PINs are statistically significant in every PIN dataset*. This essentially indicates that such a skewed distribution cannot be described by the common distribution models, such as normal, log-normal, exponential or Poisson.

Fortunately, we observe that Fig. 2 is much similar to the Fig. 5(a) in [31], which is reminiscent of the Zipf's law that occurs in an extraordinarily diverse range of phenomena such as the species per genus and US firm sizes [2]. Initially, this law was used to describe that the frequency of any word in a natural language corpus is inversely proportional to its rank in the frequency table arranged in decreasing order. Formally, it is formulated as

$$f_r = \frac{C_0}{r}, \tag{1}$$



(a) Prob. vs. rank plot of 4-digit PINs from Dodonew (b) Probability vs. rank plot of 4-digit PINs from CSDN (c) Prob. vs. rank plot of 4-digit PINs from Rockyou

(d) Probability vs. rank plot of 4-digit PINs from Yahoo (e) Prob. vs. rank plot of 4-digit PINs from four datasets (f) Prob. vs. rank plot of 6-digit PINs from four datasets

Figure 3: Zipf's law in user-generated PINs plotted on a log-log scale. The fittings are remarkably good for PINs of both user groups.

Table 7: Least square linear regression (LR) results of 4-digit PIN datasets. Large R^2s indicate the soundness of our Zipf models.

4-digit PIN Dataset	Total PINs	Unique PINs	Least freq. used	PINs used in LR	% of PINs used in LR	Unique PINs used in LR	Zipf law regression line	Absolute value of the slope (s)	Coefficient of determination (R^2)	KS test p-value
Dodonew	1,223,677	10,000	10	1,223,243	99.965%	9,945	-1.210950-0.908621*x	0.908621	0.978354	0.300168
CSDN	444,204	9,951	10	417,773	94.050%	5,185	-0.957140-1.010156*x	1.010156	0.975388	1.851E-08
Rockyou	1,780,587	10,000	10	1,780,587	100.000%	10,000	-1.093437-0.929672*x	0.929672	0.987545	0.024229
Yahoo	47,540	8,379	10	25,655	53.965%	579	-1.165256-0.926723*x	0.926723	0.988836	0.626134
Overall*	N/A	N/A	N/A	N/A	N/A	25,709	-1.196391-0.912634*x	0.912635	0.972065	N/A

*"N/A" because the regression of "Overall" is not on PINs but on the total 25,709 unique data points of the four red curves (see Figs. 3(a) to 3(d)).

where f_r is the frequency of the word ranked r in the corpus, and C_0 is a constant determined by the corpus. However, in most cases (e.g., US firm sizes [2]) other than natural languages, a more general form of Zipf's law applies:

$$f_r = \frac{C_0}{r^s}, \tag{2}$$

where the exponent s is a real number and close to 1. Note that, Eq. 2 can be equally expressed as

$$p_r = \frac{f_r}{|\mathcal{DS}|} = \frac{C_0/|\mathcal{DS}|}{r^s} = \frac{C}{r^s}, \tag{3}$$

where $|\mathcal{DS}|$ is the size of dataset, p_r is the relative frequency (or so-called probability of occurrence) of the rth ranked item (i.e., $p_r = f_r/|\mathcal{DS}|$), and $C(= C_0/|\mathcal{DS}|)$ is a constant determined by the dataset. We observe that, interestingly, while Eq. 2 is more intuitive than Eq. 3, the latter facilitates better comparison between different fitting instantiations.

Generally, for better comprehension, we can plot the data on a log-log graph (base 10 in this work), with the x-axis being log(rank order r) and y-axis being log(probability p_r). In other words, $\log(p_r)$ is linear with $\log(r)$:

$$\log p_r = \log C - s \cdot \log r. \tag{4}$$

We plot the probability vs. rank of 4-digit PINs on a log-log scale. Due to space constraints, Fig. 3 illustrates the graphs for each 4-digit PIN dataset as well as the aggregated graphs for both 4-digit and 6-digit PINs (see the *red* curves). Due to space constraints, here we only give the aggregated graphs for the 6-digit PINs.

It is worth noting that the relative frequency of each PIN in all our datasets drops *polynomially* as its rank becomes lower. An exception is that the probability of very few PINs at the tail of the rank lists drop much more sharply than polynomially (see Figs. 3(a) to 3(c)) or much more slowly than polynomially (see Fig. 3(d)). In other words, most of the data points fall approximately on a straight log-log line. This strongly indicates that an overwhelming majority of PINs well follow a Zipf distribution, with the parameter s given by the absolute value of slope of the straight line.

There are several approaches to determine the parameters of a Zipf distribution when given empirical data, among which is the widely used least-squares linear regression [2] and we adopt it for its simplicity. This method calculates the best-fitted line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the fitted line. The Coefficient of Determination (denoted by R^2) is used as an indicator of the quality of the fitting — the fraction of variance of the observed data that can be explained by the fitted line: the closer to 1 the better. For instance, an R^2=0.978 means that 97.8% of the total variation in the observed data can be explained by the fitted line, while 2.2% of the variation remains unexplained. Besides R^2, we also employ the Kolmogorov-Smirnov (KS) test to evaluate the goodness-of-fit.

The regression results on each 4-digit PIN dataset and 6-digit PIN dataset are listed in Table 7 and Table 8, respectively. The corresponding regression line for each dataset is depicted in Fig. 3. Besides linear regressions on individual datasets, we also perform a similar linear regression (with each dataset in a group contributing equally) on two groups of PIN datasets, see Figs. 3(e) and 3(f).

It is worth noting that, as shown in Table 7 and 8, we have excluded the least frequent PINs (i.e., with f <10) in our regression, because these PINs at the very tail of the rank lists apparently do not show a Zipf's law behavior (see Fig. 3): their frequencies drop either much more quickly or slowly than polynomially. Including these PINs in the regression would only deteriorate the good property of the popular PINs, which are the overwhelming

Table 8: Least square linear regression (LR) results of 6-digit PIN datasets. Large R^2s indicate the soundness of our Zipf models.

6-digit PIN Dataset	Total PINs	Unique PINs	Least freq. used	PINs used in LR	% of PINs used in LR	Unique PINs used in LR	Zipf law regression line	Absolute value of the slope (s)	Coefficient of determination (R^2)	KS test p-value
Dodonew	2,876,047	465,741	10	1,941,541	67.507%	26,120	-1.585365-0.874400*x	0.874400	0.972208	0.013623
CSDN	809,899	224,250	10	410,126	50.639%	9,978	-2.278761-0.641415*x	0.641415	0.956394	1.071E-08
Rockyou	2,758,491	448,186	10	1,964,132	71.203%	39,511	-1.898256-0.749407*x	0.749407	0.930145	0.093668
Yahoo	21,020	14,001	10	4,324	20.571%	67	-1.550087-0.992080*x	0.992080	0.952208	3.996E-15
Overall	N/A	N/A	N/A	N/A	N/A	3,447,258	-1.902151-0.764170*x	0.764171	0.899897	N/A

majority of the dataset. A more essential reason is that these low frequency PINs are unlikely to exhibit their true probability distribution according to the law of large numbers (see [31]).

Also note that, the selection of 10 as the threshold of least frequency used in regression is based on a series of exploratory experiments. Fortunately, the regression results in Table 7 and Table 8 reveal that using 10 as the threshold is satisfactory: every linear regression on 4-digit PIN datasets is with its R^2 larger than 0.975, which closely approaches to 1 and thus suggests a sound fitting; Equally good regression results are obtained for all 6-digit PIN datasets. Besides satisfactory R^2, all regressions on 4- and 6-digit PINs have incorporated at least 94% and 50% of the corresponding datasets, respectively, while the only exception for Yahoo PINs can be largely attributed to the insufficient data volume. In addition, the majority of the KS tests are with a p-value > 0.01, indicating that we should accept the null hypothesis that PINs follow the corresponding Zipf distribution functions. All these suggest that user-generated PINs, no matter 4-digit ones or 6-digit ones, English ones or Chinese ones, follow a Zipf distribution.

Now a natural question arises: *whether digit sequences of other length (e.g., 3, 5, 7, 8, 9, 10) extracted from passwords also follow the Zipf's law?* We have performed similar experiments as Fig. 3, and found that *only* digit sequences of length 3, 4, and 6 follow this law. This is somewhat unexpected. A plausible reason is that users love to use digit chunks of length 3, 4, and 6 as their secrets. This partly justifies our PIN creation methodology in Sec. 3.

Summary. Our results show that despite great language, culture differences, PINs from both user groups share almost the same Zipf distribution, thereby *obsoleting the long-used uniform assumption* [12, 17, 27] about user-chosen PINs in cryptography research. Two critical implications of our Zipf theory are shown in Appendix A.

6. PIN STRENGTH

We now address: *How much security can user-generated PINs provide*? *Between these two user groups, whose PINs are generally more secure*? Generally, there are two kinds of security threats against PINs: online guessing (whereby the allowed guess number is limited [33]) and offline guessing. Other attacks like malware/shoulder surfing, are largely unrelated to PIN strength. To measure PIN strength, there are two broad approaches available, i.e., statistic-based (see [6]) and cracking-based (see [21]). The former measures resistance against the optimal attacker, while latter is against real attackers. To be robust, we will employ both.

6.1 Statistical results

Here we mainly adopt the five metrics (i.e., min-entropy, β-success-rate, shannon entropy, guesswork and α-guesswork) that have been widely utilized: the first two metrics are used for measuring online guessing, the next two are for offline guessing and the last one is multi-purpose. Since we are interested in online guessing in different throttling options, we also examine offline guessing which can be seen as intensive online guessing. Here we use the same notation and terminology (see Table 9) from [5] and take 4-digit PINs for example: the probability distribution of PINs is denoted by $\mathcal{X}$ which is over the set $\{0000, \cdots, 9999\}$; the user PIN is a stochastic variable X which is randomly drawn from $\mathcal{X}$, and $\mathcal{X}$ may take a value $x_i \in \{0000, \cdots, 9999\}$ with the probability p_i, where $p_1 \geq p_2 \geq \cdots \geq p_N$ and $N = 10^4$.

The statistical results are shown in Table 10. "Average_4" stands for the average of metric results of the above four 4-digit PIN datasets, similarly for "Average_6"; "Random_4" stands for the dataset consisting of 10^4 distinct 4-digit PINs, similarly for "Random_6". 4-digit Rockyou PINs and "Random_4" have also been gauged in [6], and our corresponding results well agree with [6].

Table 10 shows that, on average, 4-digit PINs can provide an average of 6.62 bits ($=\widetilde{\lambda}_{30}$) of resistance against online guessing and 8.41 bits ($=\widetilde{G}_{0.5}$) of resistance against offline guessing, while this figure for 6-digit PINs is 7.24 bits and 13.21 bits, respectively. This means both types of PINs offer less than 50% of security as compared to random PINs of the same length. We also find that 4-digit PINs within the different user group offer significantly different security distributions (pairwise Wilcoxon test, $p < 0.01$), while PINs within the same user group are not (p=0.015 for CSDN&Dodonew; p=0.046 for Rockyou&Yahoo). For 6-digit PINs, Rockyou&CSDN are not significantly different (p=0.39358).

Table 10 also shows that, 4-digit PINs of Chinese users are less secure as compared to their English counterparts against *both online guessing and offline guessing*. As for 6-digit Chinese PINs, CSDN PINs are comparable to English PINs, while Dodonew PINs are always weaker than English PINs. When compared to graphic passwords (i.e., Android unlock patterns [25, 28]), only 6-digit PINs can provide comparable security against offline guessing, and both types of PINs are less secure against online guessing.

While being 150% of length with 4-digit PINs, 6-digit PINs generally can offer expected increase (i.e., from 133.18% to 164.77%) in security against offline guessing, yet the increase (i.e., 0.62 bit) in security against online guessing is not significant (less than 10% relative increase), which is opposed to common belief. The key issue of PINs lies in a few excessively popular ones, which is the very nature of a Zipf law distribution (see Sec. 5.2) *but not their length*. This would have implications for bank agencies and authorities that have conducted (or plan to) PIN migration from 4-digits to 6-digits or even longer ones (see the migration in Korean [18] and UAE [19]): *as online guessing is the primary threat to PIN-based systems, the additional security gained by enforcing a longer PIN requirement would* not *outweigh the increased costs in deployment and usability (e.g., memorization and typing).*

6.2 Cracking-based experimental results

Currently, the state-of-the-art password cracking algorithms include PCFG-based [34] and Markov-Chain-based [21]. As the former mainly focuses on exploiting various structures/patterns (e.g., the structure of "pa$$word123" is $\mathsf{L}_2\mathsf{S}_2\mathsf{L}_4\mathsf{D}_3$) that dominate textual password choices, it is not suitable for cracking PINs which are fixed-length digit-only sequences. In contrast, there is no conception of "structure" in Markov-based approach [21], and thus it is effective in cracking PINs and we adopt it here.

In our experiments, since PINs are of fixed length, there is no normalization problem. Hence, as recommended in [21], we mainly consider two smoothing techniques (i.e., Laplace and Good-Turing) to deal with the data sparsity problem and use varying orders to avoid overfitting. In each attack, we use PINs from one source as training sets and generate PIN guesses in decreasing order of probability. Then, we try these guesses sequentially to attack

Table 9: Statistical metrics [5] for measuring PIN resistance against online and offline guessing

Metric	Formula	Term	Description
$H_1(\mathcal{X})$	$\sum_{i=1}^{N} -p_i \cdot \log p_i$	Shannon entropy	A measure of the uncertainty of $\mathcal{X}$ to an attacker
$H_\infty(\mathcal{X})$	$-\log_2(p_1)$	Min-entropy	An asymptotic limit on the number of random bits extracted from $\mathcal{X}$
$G(\mathcal{X})$	$\sum_{i=1}^{N} p_i \cdot i$	Guesswork	Expected number of guesses in optimal order to find the password X
$\widetilde{G}(\mathcal{X})$	$\log_2(2 \cdot G(\mathcal{X}) - 1)$	Guesswork in bits	Bit representation of $G(\mathcal{X})$
$\lambda_\beta(\mathcal{X})$	$\sum_{i=1}^{\beta} p_i$	β-*success rate*	Expected success rates for an attacker given β guesses
$\widetilde{\lambda}_\beta(\mathcal{X})$	$\log_2(\beta/\lambda_{\beta(\mathcal{X})})$	β-*success rate* in bits	Bit representation of $\lambda_\beta(\mathcal{X})$
$G_\alpha(\mathcal{X})$	$(1-\lambda_{\mu_\alpha}) \cdot \mu_\alpha + \sum_{i=1}^{\mu_\alpha} p_i \cdot i$	α-*guesswork*	Expected number of guesses per account to achieve a success rate α
$\widetilde{G}_\alpha(\mathcal{X})$	$\log_2(\frac{2 \cdot G_\alpha(\mathcal{X})}{\lambda_{\mu_\alpha}} - 1) + \log_2(\frac{1}{2-\lambda_{\mu_\alpha}})$	α-*guesswork* in bits	Bit representation of $G_\alpha(\mathcal{X})$

Table 10: Statistical results (in bits) on strength of 4-digit PINs and 6-digit PINs

PIN Dataset		Online guessing (i.e., in small guess number) resistance							Offline/intensive guessing resistance			
		H_∞	λ_3	λ_6	λ_{30}	λ_{60}	$\widetilde{G}_{0.1}$	$\widetilde{G}_{0.2}$	$\widetilde{G}_{0.3}$	$\widetilde{G}_{0.5}$	H_1	$\widetilde{G}$
4-digit PINs	Dodonew	3.79	4.55	5.04	6.48	7.24	4.21	5.19	5.99	8.06	10.26	11.36
	CSDN	4.08	4.51	4.94	6.33	7.11	4.25	4.99	5.55	7.46	9.77	10.90
	Rockyou [6]	4.75	5.22	5.61	6.66	7.38	5.48	6.08	6.61	8.78	10.74	11.50
	Yahoo	4.47	5.11	5.65	7.00	7.61	5.30	6.69	7.43	9.33	11.01	11.54
	Average_4	4.27	4.81	5.27	**6.62**	7.33	4.81	5.74	6.40	**8.41**	10.44	11.33
	Rankdom_4	13.29	13.29	13.29	13.29	13.29	13.29	13.29	13.29	13.29	13.29	13.29
6-digit PINs	Dodonew	2.55	3.82	4.64	6.67	7.55	2.55	3.74	6.66	12.43	13.52	16.77
	CSDN	3.22	4.60	5.46	7.48	8.36	3.22	9.23	12.13	13.90	14.81	16.31
	Rockyou	3.10	4.56	5.43	7.44	8.31	3.10	8.88	12.54	14.05	15.01	16.75
	Yahoo	3.18	4.57	5.43	7.36	8.21	3.18	8.01	10.83	12.48	12.30	13.22
	Average_6	3.01	4.35	5.19	**7.24**	8.11	3.01	7.46	10.54	**13.21**	13.91	15.76
	Rankdom_6	19.93	19.93	19.93	19.93	19.93	19.93	19.93	19.93	19.93	19.93	19.93
Average_6/Average_4		**70.50%**	**90.34%**	**98.50%**	**109.36%**	**110.54%**	**62.63%**	130.11%	164.77%	157.14%	133.18%	139.18%
Android unlock patterns	Defensive [28]	–	–	–	–	–	8.72	9.10	–	12.69	–	–
	offensive [28]	–	–	–	–	–	7.56	**7.74**	–	8.19	–	–
	With meter [25]	–	–	–	–	–	8.96	10.33	11.32	12.92	–	–
	Without meter [25]	–	–	–	–	–	**7.38**	9.56	10.83	12.61	–	–

PINs from another source. This is just what the real-world guessing attackers do. Our results show that there is not much difference between Laplace Smoothing and Good-Turing (GT) Smoothing. The detailed PIN generation procedure is given in Appendix D.

Due to space constraints, Fig. 4 only illustrates the cracking results about 4-digit PINs and the aggregated results about 6-digit PINs based on Laplace-Smoothing, and the cracking results based on GT smoothing are omitted. Fig. 4 shows that, the larger the order is, the better the Markov-based attack performs. The "optimal" curves in Fig. 4 represent the theoretically optimal attacks related to the statistic metrics in Sec. 6.1. Remarkably, our best attacking curves (i.e., order-3 ones) in Figs. 4(a)∼4(d) nearly overlap with the optimal attacking curves. This suggests that our training sets and algorithm parameters are rightly chosen and that Markov-based attacks, when appropriately tuned, can indeed be effective. More importantly, *this highly indicates the potential that the distribution of PINs from one source (about which we may only know little info) can be well predicted by using PINs from another known source*.

Fig. 4(e) consists of all the best attacks against each 4-digit PIN dataset. No matter in terms of resistance against online or offline guessing, PINs from Yahoo offer the highest strength, PINs from Rockyou are the next most secure, followed PINs from Dodonew, while PINs from CSDN offer the least security. This well accords with the statistics-based results (see the upper half of Table 10).

In contrast, 6-digit PINs do not show an indisputable hierarchy of security (see Fig. 4(f)). Table 10 shows that most of the eleven statistical results of Yahoo 6-digit PINs are higher than Dodonew and lower than CSDN, yet Fig. 4(f) illustrates that Yahoo 6-digit PINs are more secure than both Dodonew and CSDN. However, if we took no account of Yahoo PINs in Fig. 4(f), then all the cracking results would be consistent with the statistics-based results. One plausible reason is that, the 6-digit Rockyou PINs are unsuitable to be used as the training set for cracking Yahoo, and violating this may convey a misleading sense of security. This once again highlights the importance of the selection of appropriate training sets, revealing the inherent limitations of cracking-based evaluation metrics. Still, cracking-based approach can be improved with better knowledge of PIN distribution and characteristics.

7. CONCLUSION

We have conducted a systematic investigation into the characteristics, distribution and security of PINs chosen by English and Chinese users. By exploiting visulization techniques and building semantic models, we have identified various differences in structural and semantic patterns between PINs of these two user groups; By employing NLP techniques, we have revealed that PINs follow a Zipf distribution; By adopting the leading statistic metric α-guesswork and cracking algorithms, we have highlighted that 6-digit PINs essentially offer marginally improved security over 4-digit PINs. To our knowledge, this is the first work that examines 6-digit PINs which dominate in Asia and are gaining popularity worldwide. It is expected that this work would help users and security engineers gain a deeper understanding of the vulnerability of human-chosen PINs, and shed light on future PIN research.

Acknowledgment

The authors are grateful to the shepherd, Prof. Di Ma, of our paper. We also thank the reviewers for their constructive comments. Ping Wang is the corresponding author. This research was supported by the National Key Research and Development Plan under Grant No. 2016YFB0800603, and by the National Natural Science Foundation of China (NSFC) under Grants Nos. 61472016 and 61472083.

8. REFERENCES

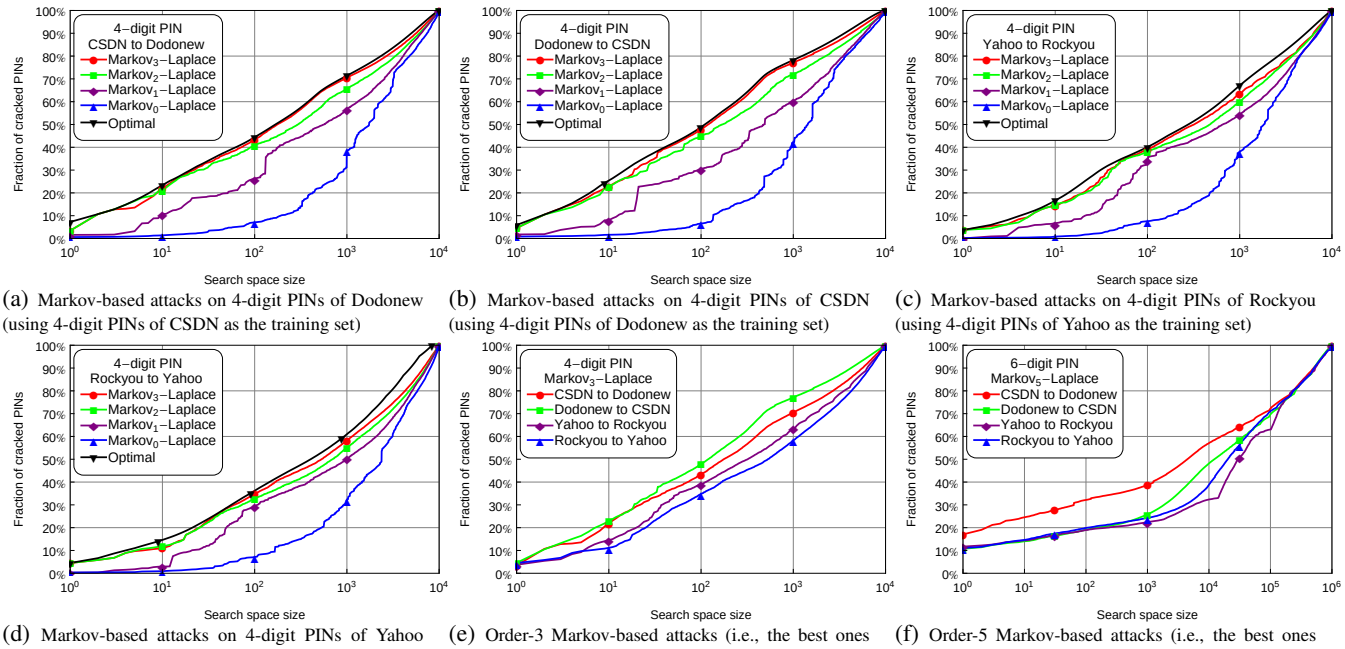
[1] *China now has 656m mobile web users, and 710m total Internet users*, Aug. 2016. http://bit.ly/2avZdlK.

[2] R. L. Axtell. Zipf distribution of US firm sizes. *Science*, 293(5536):1818–1820, 2001.

[3] B. Bátiz-Lazo and R. Reid. The development of cash-dispensing tech- nology in UK. *IEEE Ann. Hist. Comput.*, 33(3):32–45, 2011.

[4] N. Berry. PIN analysis, Sep. 2012. http://www.datagenetics.com/blog/september32012/index.html.

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



(a) Markov-based attacks on 4-digit PINs of Dodonew (using 4-digit PINs of CSDN as the training set)

(b) Markov-based attacks on 4-digit PINs of CSDN (using 4-digit PINs of Dodonew as the training set)

(c) Markov-based attacks on 4-digit PINs of Rockyou (using 4-digit PINs of Yahoo as the training set)

(d) Markov-based attacks on 4-digit PINs of Yahoo (using 4-digit PINs of Rockyou as the training set)

(e) Order-3 Markov-based attacks (i.e., the best ones against 4-digit PINs) on 4-digit PINs of each dataset

(f) Order-5 Markov-based attacks (i.e., the best ones against 6-digit PINs) on 6-digit PINs of each dataset

Figure 4: Markov-based attacks on user-generated PINs using *Laplace Smoothing* to deal with data sparsity and using varing orders to deal with overfitting. Attacks (a)∼(d) are against 4-digit PINs of one dataset, while (e) and (f) are against *all* 4-digit and 6-digit PIN datasets, respectively.

[6] J. Bonneau, S. Preibusch, and R. Anderson. A birthday present every eleven wallets? The security of customer-chosen banking PINs. In *FC 2012*, pages 25–40.

[7] W. Burr, D. Dodson, R. Perlner, W. Polk, S. Gupta, and E. Nabbus. NIST SP800-63-2 – electronic authentication guideline. Technical report, NIST, Reston, VA, Aug. 2013.

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

[9] X. de Carnavalet and M. Mannan. From very weak to very strong: Analyzing password-strength meters. In *NDSS 2014*.

[10] EMVCo Ltd. *Issuer PIN Security Guidelines*, 2010. http://vi.sa/2lxbIFJ.

[11] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *SOUPS 2013*.

[12] D. Florêncio, C. Herley, and B. Coskun. Do strong web passwords accomplish anything? In *Proc. HotSec 2007*.

[13] *ISO 9564: Financial services – Personal Identification Number (PIN) management and security*, 2011. https://www.iso.org/obp/ui/#iso:std:54083:en.

[14] M. Jakobsson and D. Liu. Your password is your new PIN. In *Mobile Authentication*, pages 25–36. Springer, 2013.

[15] J. Katz, R. Ostrovsky, and M. Yung. Efficient and secure authenticated key exchange using weak passwords. *J. ACM*, 57:1–41, 2009.

[16] M. Keith, B. Shao, and P. J. Steinbart. The usability of passphrases for authentication: An empirical field study. *Int. J. of human-computer studies*, 65(1):17–28, 2007.

[17] B. Köpf and D. Basin. Automatically deriving information-theoretic bounds for adaptive side-channel attacks. *J. Comput. Secur.*, 19(1):1–31, 2011.

[18] *Korean Banks to Possibly Adopt a 6-Digit PIN System*, Oct. 2015. http://bit.ly/2lx3MnB.

[19] *Chip and PIN: Final deadline in UAE*, 2014. https://www.souqalmal.com/financial-education/ae-en/chip-and-pin/.

[20] L. Lundin. *PINs and Passwords, Part 1*, Aug. 2013. http://bit.ly/2kVF9hh.

[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] D. Malone and K. Maher. Investigating the distribution of password choices. In *Proc. WWW 2012*, pages 301–310.

[23] M. Nicholas. *The 10 Biggest Data Breaches of Summer 2016*, Sep. 2016. https://blog.dashlane.com/biggest-data-breaches-summer-16/.

[24] Reuters. *Russia launches China UnionPay credit card*, 2014. http://rt.com/business/180696-china-russia-union-pay/.

[25] Y. Song, G. Cho, S. Oh, H. Kim, and J. H. Huh. On the effectiveness of pattern lock strength meters: Measuring the strength of real world pattern locks. In *Proc. CHI 2015*.

[26] L. Stanekovaa and M. Stanek. Analysis of dictionary methods for pin selection. *Comput. Secur.*, 39:289–298, 2013.

[27] G. Steel. Formal analysis of PIN block attacks. *Theor. Comput. Sci.*, 367(1):257–270, 2006.

[28] S. Uellenbeck, M. Dürmuth, C. Wolf, and T. Holz. Quantifying the security of graphical passwords: The case of android unlock patterns. In *Proc. ACM CCS 2013*.

[29] B. Ur, J. Bees, S. Segreti, and et al. Do users' perceptions of password security match reality? In *Proc. ACM CHI 2016*.

[30] R. Veras, J. Thorpe, and C. Collins. Visualizing semantics in passwords: The role of dates. In *Proc. ACM VizSec 2012*.

[31] D. Wang, G. Jian, X. Huang, and P. Wang. Zipf's law in passwords. *IEEE Trans. Inform. Foren. Secur.*, 2016. In press, http://bit.ly/2kfSaVP.

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

[33] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang. Targeted online password guessing: An underestimated threat. In *Proc. ACM CCS 2016*, pages 1242–1254.

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

[35] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Secur. Priv.*, 2(5):25–31, 2004.

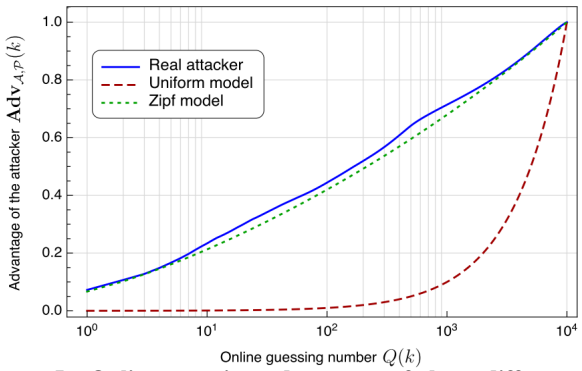


Figure 5: Online guessing advantages of three different attackers against the Dodonew 4-digit PIN based authentication service. The Zipf attacker well approximates the real attacker.

APPENDIX

A. SOME IMPLICATIONS

In this section, we briefly sketch two implications that our Zipf theory is highly likely to carry.

A.1 Implication for PIN creation policies

Perhaps the most immediate implication of the discovery of Zipf's law in user-generated PINs is for PIN creation policies. The *polynomially decreasing nature* of the popular PINs suggests that enforcing a simple blacklist (e.g., blacklisting the top 100 PINs as suggested in [6]) is inherently insufficient. On the one hand, the CDF graphs (see Fig. 2) evidently show that a small fraction of top frequent PINs can account for a tremendous percentage of the total PIN accounts, and simply blacklisting them would annoy a large fraction of customers (see the analysis of W_p in [31]).

On the other hand, if some most frequent PINs are banned, there is no way to prevent other PINs to become as frequent as these banned PINs. Fig. 3 illustrates that the frequency distribution curves of PINs are smooth, indicating there is a steady supply of popular PINs. What's more, as the total PIN space is small (e.g., 10^4), blacklisting some PINs might help an attacker to reduce her search space. This suggests that a threshold approach would be more desirable: set a popularity threshold $\mathcal{T}$ (e.g., $\mathcal{T} = 10^3$ for a 4-digit PIN bank authentication system with 10^7 customers) for the system, only these PINs whose frequency fall below the threshold are allowed. If a user-chosen PIN unfortunately falls above $\mathcal{T}$, the system suggests the user several alternative ones with a low popularity and with the least *edit distance* from the user's originally input PIN. In this way, a better trade-off between security and usability can be achieved.

A.2 Implication for PIN-based protocols

As far as we know, the security formulation results of most existing password- and PIN-based cryptographic protocols (e.g., [12, 15, 17, 27]) are based on the unrealistic assumption that *user-chosen PINs/passwords are uniformly distributed*. Though these protocol designers often cast doubt on such an assumption, *they are stuck in the question*: if PINs/passwords do not obey a uniform distribution, then which distribution will they follow? It was not until very recently that Wang et al. [31] revealed that user-generated textual passwords comply with a Zipf distribution. How about digit PINs? Fortunately, we have provided a promising answer in Sec. 5.2, and it is more practical to design PIN-based cryptographic protocols (e.g., authentication, signature and secret sharing) with an assumption of the Zipf distribution. Here we use the PIN-based *authentication* and key exchange (PAKE) protocols as an example.

Based on our analysis (see [32]) of the implications for password-based protocols, here we show the implication of Zipf theory for PIN-based protocols. In most provably secure PAKE protocols, generally it is formalized in a way that "password/PIN pw_C (for each client C) is chosen independently and *uniformly at random* from a dictionary $\mathcal{D}$ of size $|\mathcal{D}|$, where $|\mathcal{D}|$ is a fixed constant independent of the security parameter k" [15], then a security model is described, and finally a "standard" definition of security as the one in [15] is given:

> "$\cdots$ $\cdots$ Protocol $\mathcal{P}$ is a secure protocol for password/PIN-only authenticated key-exchange if, for all [password/PIN] dictionary sizes $|\mathcal{D}|$ and for all ppt[probabilistic polynomial time] adversaries $\mathcal{A}$ making at most $Q(k)$ on-line attacks, there exists a negligible function $\epsilon(\cdot)$ such that:
>
> $$\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k) \leq Q(k)/|\mathcal{D}| + \epsilon(k), \qquad (5)$$
>
> where $\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k)$ is $\mathcal{A}$'s advantage in attacking $\mathcal{P}$."

The best strategy for a real-world attacker $\mathcal{A}$ is to try the most likely guess first, then the second most likely one and so on [5]. Under the Zipf assumption, it is natural to see that $\mathcal{A}$'s advantage $\mathrm{Adv}_{\mathcal{A},\mathcal{P}}(k)$ can be formulated as:

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

where s is the absolute value of the slope of Zipf line of $\mathcal{D}$.

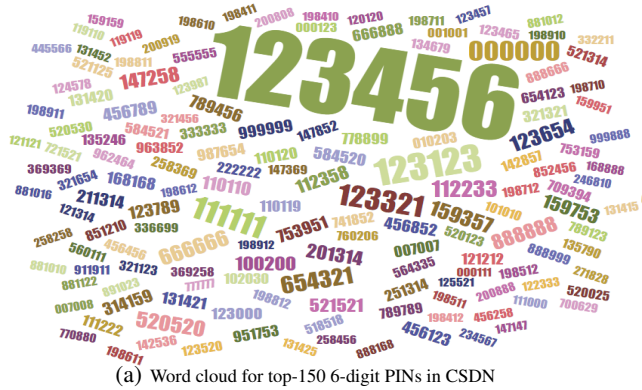
Fig. 5 shows that $\mathcal{A}$'s advantage is more accurately captured by our Zipf model than the uniform model. The latter tends to greatly underestimates $\mathcal{A}$'s online guessing advantage. For instance, at 100 guesses (i.e., $Q(k)$=10^2), the uniform model estimates $\mathcal{A}$'s advantage against the Dodonew service to be 1.0%, yet the real value is 44.7%. Our Zipf attacker achieves a success rate 42.4%, well predicting the real value. This is particularly useful when bank agencies evaluate the security risks that user-chosen PINs (i.e., probably the weakest link in the security chain) bring about.

B. WORD CLOUDS FOR 6-DIGIT PINS

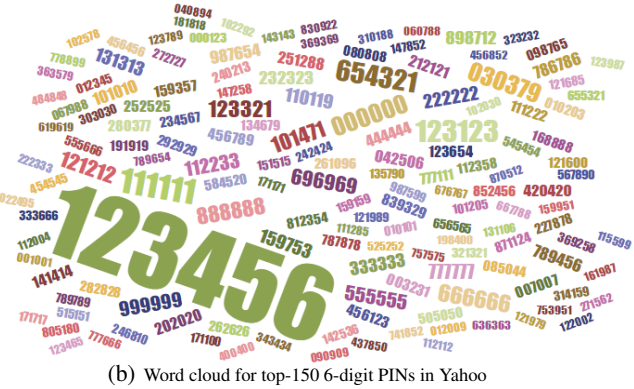
To further identify the dominant factors that influence user choices of 6-digit PINs, we once again resort to a visualization technique (i.e., word cloud) due to its intuitiveness and informativeness [30], and make use of the wordle diagram http://www.jasondavies.com/wordcloud/. We provide the raw PINs from each dataset to the wordle tool which gives greater prominence to PINs that are more frequent. As a result, the PINs shown in the cloud picture are sized according to the number of their occurrences.

The word clouds for the top-150 PINs in each PIN dataset can be found in Fig. 6. One can see that patterns like dates and single-digit/couplets/triple repetition are as prevalent as that of 4-digit PINs. An interesting point is that, users prefer using pairs of numbers that have smaller numerical gaps between them. For example, combinations like 12 and 78 are used much more frequently than 17 and 28. One plausible reason may be that a smaller numerical gap is easier to type on a numpad (and keyboard). As with 4-digit PINs, a number of meaningful 6-digit PINs are chosen by either Chinese users (e.g., `131452` and `110120`) or English users (e.g., `420420` and `696969`), or both groups of users (e.g., `007007`). More specifically, `131452` sounds like "Forever and ever I love" in Chinese pronunciation; `110` and `120` are the alarm call and emergency call in China, respectively; `420` has become a popular code for marijuana; `69` seems to be a favorite number of English users; Fans of James Bond should be proud to see `007`.

What's quite different from 4-digit PINs is that *numpad patterns* seem to be much more popular in 6-digit PINs: about 26% to 36% of the top-150 PINs in each dataset are with some kind of numpad patterns like "×" (e.g., `159753`), "+" (e.g., `258456`), "=" (e.g., `123654`), "≡" (e.g., `134679`), "⊤" (e.g., `123580`)



(a) Word cloud for top-150 6-digit PINs in CSDN



(b) Word cloud for top-150 6-digit PINs in Yahoo

(c) Word cloud for top-150 6-digit PINs in Dodonew

(d) Word cloud for top-150 6-digit PINs in Rockyou

Figure 6: Word clouds for four 6-digit PIN datasets (The top-150 most popular ones of each dataset are depicted)

Algorithm 1: Generating PIN guesses using Markov-Chains

Input: A training set $\mathcal{TS}$; Max PIN length `maxLen`; Markov order `mkOrder`
Output: A PIN guess list L in probability-decreasing order

Training:
  for $pin \in \mathcal{TS}$ **do**
    for $i \leftarrow 1$ **to** $length(pin)$ **do**
      $preStr \leftarrow subStr(pin, max(0, i - \texttt{mkOrder}), i - 1)$
      $nextChar \leftarrow getChar(pin, i)$
      $trainingResult.insert(preStr, nextChar)$
Laplace smoothing ($\delta = 0.1$)**:**
  function $traResult.getProb(preStr, nextChar)$
    $count = traResult.getCount(preStr, nextChar) + \delta$
    $countSum = traResult.getCount(preStr, charSet) + \delta * traResult.charSet.size()$
    return $count/countSum$
function $ProduceGuess(pin, prob)$
  if $length(pin) = \texttt{maxLen}$ **then**
    $guessSet.insert(pin, prob)$
  else
    $preStr \leftarrow pin.tailStr(\texttt{mkOrder})$
    for $char \in [0 - 9]$ **do**
      $newPin \leftarrow pin + char$
      $newProb \leftarrow prob * traResult.getProb(preStr, nextChar)$
      $ProduceGuess(newPin, newProb)$
Produce guesses:
  $ProduceGuess(null, 1)$; $L \leftarrow guessSet.sort()$

and "||" (e.g., `147369`). Interestingly, once again we see very compelling evidence that extracting exactly 6-digit sequences from passwords is a great proxy for 6-digit real-world PINs. One can confirm that many of these numpad-pattern-based digit sequences (e.g., `142536` and `258456`) are utterly awkward to type on a laptop/PC keyboard because the order of these digits is interlaced on such keyboards, but it is quite handy on a phone numpad (i.e., following a "⊤" trace). ATMs and other terminals (e.g., electronic doors) that employ a phone-style numpad just facilitate such digit sequences. This suggests that, besides their preference of easy to type (and remember) PINs for credit cards or mobile devices, users also seem to persist in re-using the same digits of their PINs for their online passwords, even though the mnemonics related to numpads (e.g., "⊤–first horizontally the top and then straight down the middle") no longer apply to laptop/PC keyboards.

A few other *interesting tidbits* can also be revealed from Fig. 6. In the lower-left corner of Fig. 6(a), one can find the PIN `314159`, a random six-digit number, is it? It is deceptively random-looking, but one would at once become enlightened if she calls to remembrance the number π. Then, one will not be surprised to find the PIN `141592` occurring high up in the PIN lists. Some other important numbers like the base of the natural logarithm e and the Fibonacci sequence can also find their silhouettes (e.g., `271828` and `112358`) in Fig. 6. A scrutiny into the original datasets can further identify many other numbers that are frequently used by both groups of users, e.g, the miraculous number `142857` which was said to be first found in the Egyptian pyramids.

C. MEANINGFUL CHINESE PINS

The following are eighteen popular 6-digit PINs of Chinese users which have been identified in both two Chinese 6-digit PIN datasets. They all sound like meaningful and interesting phrases in Chinese pronunciation: `584520`, `520520`, `201314`, `521521`, `211314`, `131421`, `131420`, `251314`, `709394`, `521314`, `584521`, `721521`, `564335`, `123520`, `518518`, `520110`, `520184`, `520134`, if one knows that, in Chinese, "5" sounds like "I"; '18" sounds as "get rich"; "20" and "21" sound like "love you"; "25" sounds like "love me"; "84" sounds like "pledge"; "335" sounds like "miss me"; "1314" sounds like "forever and ever". Accordingly, it is intuitive to understand any combination of them, with one exception that "520110" sounds like "I love you one for a hundred million years".

D. A PIN-CRACKING ALGORITHM

Here we provide a Markov-based PIN-guesses generation procedure using Laplace smoothing as shown in Algorithm 1.