

Evaluating the Guessability of 6-Digit PINs under User Bias and Date-of-Birth Leakage

DOI: <https://doi.org/10.54654/isj.v2i28.6411>

Giang Anh Tuan*, Nguyen Viet Khoa

Abstract—Six-digit PIN security cannot be characterized by the nominal 10^6 -value space or by unconditional entropy alone. This paper presents a fully specified synthetic evaluation of three populations: uniform (M1), survey-calibrated biased (M2), and user-specific date-of-birth linked (M3). The anonymous convenience survey contained 142 responses; one missing strategy response was excluded, one free-text response was recoded, and date-format probabilities were estimated conditionally from 119 applicable responses. Exact population metrics are separated from finite-sample plug-in estimates, and four attacks are evaluated on held-out users across 20 declared seeds. Correct full-support enumeration gives M1 an expected guess number Frequency-Ranked of 500,000.5, Top-10 success whereas reaches $19.833\% \pm 0.252\%$ against M2. M3 has unconditional entropy $H(PIN) = 13.682 \text{ bits}$ but conditional entropy $H(PIN|DOB) = 10.57 \text{ bits}$; with DOB known, Leakage-Assisted Top-10 success reaches $38.736\% \pm 0.409\%$. Thus, population diversity and individual protection are different properties. Blocklist outcomes depend on post-rejection behavior: conditional reselection can reduce global frequency attacks but does not mitigate the modeled DOB-assisted attack, while concentrated workarounds can increase risk. Conclusions are limited to the stated synthetic populations and assumptions; no blocklist size is claimed to be universally optimal.

Tóm tắt— Mức an toàn của mã PIN sáu chữ số không thể được đặc trưng chỉ bằng không gian danh nghĩa 10^6 giá trị hoặc entropy vô điều kiện. Bài báo trình bày đánh giá tổng hợp được đặc tả đầy đủ cho ba quần thể: đồng đều (M1), thiên lệch hiệu chỉnh từ khảo sát (M2), và liên kết với ngày

sinh theo từng người dùng (M3). Khảo sát thuận tiện ẩn danh có 142 phản hồi; một phản hồi thiếu được loại, một phản hồi tự do được mã hóa lại, và xác suất định dạng ngày được ước lượng có điều kiện từ 119 phản hồi phù hợp. Các đại lượng chính xác của quần thể được tách khỏi ước lượng plug-in từ mẫu hữu hạn, và bốn kiểu tấn công được đánh giá trên người dùng kiểm tra qua 20 hạt giống công bố. Việc liệt kê đầy đủ không gian cho M1 số lần đoán kỳ vọng 500.000.5, trong khi tấn công xếp hạng theo tần suất đạt Top-10 $19.833\% \pm 0.252\%$ đối với M2. M3 có entropy vô điều kiện $H(PIN) = 13.682 \text{ bits}$ nhưng entropy có điều kiện $H(PIN|DOB) = 10.57 \text{ bits}$; khi biết ngày sinh, tấn công hỗ trợ rò rỉ đạt Top-10 $38.736\% \pm 0.409\%$. Vì vậy, tính đa dạng của quần thể và mức bảo vệ từng cá nhân là hai thuộc tính khác nhau. Hiệu quả danh sách chặn phụ thuộc vào hành vi chọn lại: chọn lại có điều kiện có thể giảm tấn công theo tần suất toàn cục nhưng không giảm tấn công dựa trên ngày sinh, còn cách né tránh tập trung có thể làm tăng rủi ro. Kết luận chỉ áp dụng cho các quần thể tổng hợp và giả định đã nêu; bài báo không khẳng định một kích thước danh sách chặn tối ưu phổ quát.

Keywords— PIN security; guessability; low-entropy attacks; date-of-birth leakage; weak PIN blocklisting; authentication policy; user-chosen secrets.

Từ khóa— Bảo mật mã PIN; khả năng đoán được; các cuộc tấn công entropy thấp; rò rỉ ngày sinh; danh sách chặn PIN yếu; chính sách xác thực; bí mật do người dùng chọn.

I. INTRODUCTION

Six-digit device-unlock, PINs are used in payment, access-control, and secondary-verification settings because they are memorable and compatible with numeric keypads and attempt limits. Their nominal space contains one million values, but practical security depends on the choice distribution. Dates, repeats, sequences, keypad patterns, and memorable years concentrate probability, allowing a ranked attacker to capture far more mass within 1–10

This manuscript was received on June 5, 2026. It was reviewed on July 15, 2026, revised on August 3, 2026 and accepted August 20, 2026.

*Corresponding author

guesses than uniform chance. DOB knowledge sharpens this bias. A population attacker knows only that dates are popular; a targeted attacker who knows one user's DOB can prioritize that user's encodings. Those values may be spread across the population, raising aggregate diversity while leaving each linked user predictable. This distinction motivates comparing unconditional population entropy with conditional targeted guessability. Names, phone numbers, identity fragments, usernames, addresses, and other PII are outside the model. Empirical PIN corpora reveal real behavior but may be ethically sensitive, selection-biased, or unavailable. Synthetic models are reproducible but can create artifacts; assigning every user one DOB, for example, produces artificial global peaks. We instead give every record an independently assigned DOB and use only that record's date for its individualized attack list. Dates may coincide naturally, but uniqueness is not imposed. This study does not claim frequency ranking, targeted guessing, or blocklisting as new attack or defense classes. Its contribution is their controlled integration in a transparent DOB-conditioned PIN model. First, it provides an auditable mapping from 142 survey responses to a fully specified generator and complete guess orders under an 80/20 protocol across 20 seeds. Second, it computes both $H(\text{PIN})$ and $H(\text{PIN} \mid \text{DOB})$ and evaluates individualized DOB-aware guesses. Third, it retrains attackers after two explicit post-rejection mechanisms and evaluates both population and DOB-aware attacks, without treating any tested blocklist size as a usability optimum. The remainder of the paper is structured as follows. Section II reviews related work on PIN guessability, password metrics, and policy defenses. Section III describes the PIN-generation models, attack strategies, evaluation protocol, security metrics, and blocklist defense. Section IV reports the experimental results. Section V discusses the key findings, practical implications, and limitations. Section VI concludes.

II. BACKGROUND

A. PIN Guessability and Human Bias

The nominal size of a secret space is only an upper bound on the protection obtained when people choose their own secrets. Markert et al. collected user-chosen four- and six-digit smartphone unlock PINs and evaluated them under throttled budgets of 10, 30, and 100 guesses [1, 2]. Their results show that increasing PIN length does not automatically provide the expected combinatorial gain: users continue to select dates, repeats, and other high-frequency values. Their blocklist experiments are particularly relevant here because they show both the concentration of risk in popular PINs and the importance of what users choose after rejection. Our M2 categories represent these broad mechanisms, but our numerical weights come only from the survey reported in Section III rather than being fitted to their corpus. Munyendo et al. studied four-to-six-digit PIN upgrades with 1,010 participants and common and targeted guessing within 10–30 attempts [3]. Their targeted attacker used a participant's prior four-digit PIN and transformation heuristics. In contrast, M3 conditions a synthetic six-digit population on each user's DOB, reports exact full-support and conditional metrics, and evaluates post-blocklist distributions. The different data, enrollment tasks, and side information preclude comparison. direct numerical PIN choice also has a life cycle that is not visible in a frequency table. Khan et al. found reuse, sharing, inheritance, and infrequent updates across devices and contexts [4]; the present simulation models only one PIN per record and does not reproduce cross-account reuse. Uellenbeck et al. found comparable structural bias in Android unlock patterns [5]. Pattern frequencies are not transferred to M2, but the result reinforces why an attacker ranks memorable structures rather than sampling the nominal space uniformly. Target-specific evidence can also come from side channels. Aviv et al. studied touch-screen residue [6], while Cardaioli et al. investigated audio-based PIN inference [7]. These differ from DOB knowledge but share the consequence that guesses are conditioned on one target. M3 isolates that effect without modeling either side channel's observation process.

B. Password Guessability Metrics and Attack Modeling

Shannon's entropy formalizes average uncertainty in a probability distribution [8], but it does not by itself answer how much probability an online attacker captures in the first few guesses. In particular, entropy averages over the entire support, whereas authentication systems often impose a threshold at three, five, or ten failed attempts. Bonneau developed guessability measures that connect a distribution to an attacker's ordered guesses and emphasized the value of success-rate metrics such as λ_β for limited guessing [9]. We consequently report Shannon entropy, min-entropy, expected guess number, and held-out Top-k success rather than treating any single measure as a complete security characterization. Large-scale password studies help explain why these measures diverge. Florencio and Herley documented reuse and practical compromises between security and memorability [10]; Bonneau et al. related those compromises to the persistence of deployable but imperfect authentication [11]. PINs have a smaller alphabet and fixed length, but their observed choice distribution likewise determines resistance to ranked guessing. Research on guessing algorithms further distinguishes a probability model from a guessing procedure. Weir et al. used probabilistic context-free grammars to learn structural password patterns and assign probability across mangling rules [12]. Kelley et al. compared password policies by simulating multiple cracking algorithms and measuring the number of guesses required, demonstrating that conclusions depend on the attack model used for evaluation [13]. Melicher et al. later used neural networks to model password guessability efficiently while preserving an ordered probabilistic view of attacks [14]. These methods target richer alphanumeric spaces than the six-digit domain studied here. Because complete enumeration of 106 values is feasible, our framework uses transparent frequency and rule orders rather than a learned grammar or neural model. Wang et al.'s RFGuess uses random forests and 13 real password datasets [15]. RFGuess-PII

incorporates names, birthdays, phone numbers, and other fields and reports targeted success within 100 guesses. It is broader than our fixed-length DOB rule prefix, which is an interpretable baseline rather than a state-of-the-art PII guesser or upper bound. Without a compatible password corpus and linked PII fields, an RFGuess implementation would not be a like-for-like baseline. We adopt three evaluation principles: separate training and test records, define attack orders over the complete support, and distinguish theoretical metrics from plug-in estimates. The last two are crucial for M1 because 80,000 training records cannot cover one million values even though the source distribution is known.

C. Defenses and Policy

Defenses can intervene during enrollment or authentication. NIST SP 800-63B-4 recommends screening prospective memorized secrets against commonly used, expected, or compromised values and combining verification with rate limiting and other controls [16], [17]. A PIN-specific deployment still requires its own threat and usability analysis: screening changes the distribution of accepted choices, while throttling limits how far an attacker can traverse that distribution. The present work therefore measures low-budget success and treats blocklisting as one component rather than as a replacement for secure storage, attempt limits, or step-up authentication. Usability studies caution that a rule's nominal strictness is not the same as its realized benefit. Komanduri et al. measured how password-composition policies change both strength and creation burden [18]. Their work supports evaluating the choices produced after a policy is applied, not merely deleting forbidden values from an existing histogram. Markert et al. reached a related conclusion for PIN blocklists: design and user reaction affect the resulting security [1]. Our two post-rejection mechanisms respond directly to this issue. Conditional reselection represents ideal compliance with the original model, whereas numeric increment is a deliberately concentrated stress case. Neither is presented as an empirically measured user strategy. Interfaces can influence choice without hard rejection. Ur et al. found that

stringent strength meters can improve created passwords while affecting time and annoyance [19]; Forget et al. proposed persuasive authentication at secret selection [20]. Blocki et al.'s rehearsal model further shows why memory burden matters when proposing stronger secrets [21]. Applied to PINs, these works motivate warnings and usability measurement, but they cannot predict whether a rejected user broadens their choice, abandons enrollment, reuses a secret, or adopts a workaround.

D. Position of This Work

This work complements rather than supersedes those larger empirical studies. Its data are synthetic six-digit populations calibrated by 142 survey responses; its side information is an independently assigned DOB; and its budget is 1–10 guesses. The distinctive focus is exact support metrics, $H(\text{PIN} \mid \text{DOB})$, and attacker retraining after policy. Within-category supports remain declared assumptions, so the simulation lacks the external validity of a representative deployment corpus. M3 makes the user record, rather than the population, the unit of leakage: an independently sampled DOB remains associated with its record, and LA observes that DOB while FR observes only training frequencies. This supports direct comparison of $H(\text{PIN})$ with $H(\text{PIN} \mid \text{DOB})$. Post-policy behavior is likewise varied explicitly through idealized reselection and a concentrated workaround. All reported rates are therefore consequences of the stated populations and attacker knowledge, not universal PIN-system estimates.

III. MODELS AND ATTACK METHODS

A. Research Questions and Threat Model

The evaluation addresses four questions. RQ1 compares exact and finite-sample metrics across M1–M3. RQ2 compares Random, Frequency-Ranked, and independent Rule-Based attacks at one, three, five, and ten guesses. RQ3 measures how DOB knowledge and leakage fraction affect conditional uncertainty and LA. RQ4 varies blocklist size, population, attacker knowledge, and response to rejection. This separates distribution, attacker capability, and policy behavior.

TABLE I. M2 TOP-LEVEL WEIGHTS AND WITHIN-CATEGORY SUPPORT

Category	Weight	Support / rule
Personal date	46/141	valid DOB-range day; Q2 format mix
Anniv./historical	21/141	valid 1950-2025 day; Q2 format mix
Repeated	21/141	000000, ..., 999999 (10 values)
Sequential	16/141	10 declared ascending/descending strings
Significant	17/141	10 declared values (e.g., 520520, 131452)
Uniform random	20/141	all 10^6 values

The threat is throttled online guessing with $k \in \{1,3,5,10\}$. All attackers know the PIN format and declared setting. FR learns from training records; RB receives public rules but no counts; LA additionally receives the target's DOB and uses FR as fallback. LA does not observe the source label or whether the PIN came from M3's leakage branch.

Exact DOB knowledge isolates the effect of correct personal information; partial or erroneous leakage and auxiliary attributes are not simulated. Device protections such as biometric gates and hardware attempt counters would change the feasible budget but not the orders defined here. Offline compromise is outside scope because all one million PINs can be enumerated; storage protection then dominates. The results characterize comparative online guessability, not complete account-compromise probability.

B. PIN-Generation Models

The framework generates $n = 100,000$ user records per model. Every record contains the PIN, an independently assigned DOB, the source category, any source date and date format, and an indicator for the M3 leakage branch. M1 samples each value in $[000000, 999999]$ with probability 10^{-6} . M2 is the mixture in Table I. Categories are sampled first and values are then sampled uniformly within the repeated, sequential, significant-number, or uniform-random category. Dates are sampled uniformly over valid calendar days, not by independently sampling day/month fields. Personal dates and assigned

DOBs use the inclusive interval 16 May 1956–16 May 2011; anniversary/historical dates use 1 January 1950–31 December 2025. The fixed reference date is 16 May 2026. For both date categories, the conditional format probabilities are DDMMYY 48/119, MMDDYY 30/119, YYMMDD 18/119, and padded year 23/119; the last is split equally between YYYY00 and 00YYYY. This reuse of Q2 for anniversaries is an explicit modeling assumption. The exact sequential list is 123456, 654321, 012345, 987654, 234567, 345678, 456789, 876543, 765432, and 543210. The significant-number list is 000000, 666666, 888888, 999999, 123123, 112233, 121212, 101010, 520520, and 131452. Each list member has equal within-category probability. Output-level overlap is allowed: for example, 000000 receives probability mass from repeated, significant, and uniform-random branches. Such mass is added rather than resolved by category priority; the stored source label is used only for auditing. M3 draws from a user-specific DOB branch with probability $p_l = 0.30$ and otherwise draws the full M2 mixture with probability 0.70. Thus date-based categories remain present in the M2 branch but are independent of the assigned user's DOB; an incidental match is possible. The M3 leakage branch is the only branch explicitly linked to that user's DOB. DOBs are independently assigned and need not be unique.

C. Survey and Mapping Procedure

The anonymous Google Form received 142 submissions from 27 March to 6 April 2026. It was a convenience sample circulated through the authors' university network. The export contains no names, email addresses, demographics, or actual PINs; therefore no demographic participant profile or subgroup analysis is claimed. Participation was voluntary. The form had no separate consent checkbox; submission was treated as consent for anonymous aggregate analysis. The project record contains no formal ethics-board approval, and none is claimed. The four questions asked for: (Q1) the most likely strategy for a new six-digit PIN in a non-critical app; (Q2) the most common DOB format; (Q3) PIN reuse frequency; and (Q4) perceived ten-guess risk when an acquaintance knows personal

information. Q1 counts were date 45, anniversary/historical 21, repeated 21, sequential 16, significant 17, random 20, one free-text "bias birthday," and one missing response. The free text was recoded as date and the missing response excluded, giving the exact M2 denominator 141. Q2 counts were DDMMYY 48, MMDDYY 30, YYMMDD 18, padded year 23, and never-use-dates 23. Excluding the last category yields the conditional denominator 119. Q3 and Q4 are descriptive and do not set model weights. The full questionnaire, raw aggregate export, recoding script, and resulting aggregate table accompany the code.

The artifact reports two-sided 95% Wilson intervals for each proportion. Personal date is 32.62% (25.44–40.73%, $n = 141$), and DDMMYY is 40.34% among applicable responses (31.96–49.32%, $n = 119$). These descriptive intervals do not establish convenience-sample representativeness; M2 uses the disclosed point estimates.

The example displayed beside the MMDDYY option was 980305, which is actually YYMMDD for 5 March 1998. The generator follows the category label and produces 030598. This questionnaire defect is disclosed as a limitation.

D. Train/Test Evaluation Protocol

Each sample is split 80/20 at the user-record level. Training records estimate FR rankings; test records evaluate success. Main experiments use seeds 0–19. Leakage and blocklist sensitivity use seeds 0–9. Generation, splitting, random guessing, and defense reselection use deterministic, separated seed streams recorded in the run manifest. Reported deviations are sample standard deviations across seeds.

E. Attack Strategies

Random (RD) samples a uniformly random permutation of all 106 PINs. Frequency-Ranked (FR) sorts observed training PINs by descending count, breaks count ties by numeric PIN, and then appends every unseen PIN in numeric order. Hence FR is a complete permutation and does not silently omit unseen values. Rule-Based (RB) is implemented independently of FR and never

reads training counts. Its fixed order is: the ten repeated PINs in numeric digit order; the declared sequential list; the declared significant list; valid date encodings ordered by conditional format probability and calendar date; then all remaining values numerically. First occurrence removes overlaps. For seed 0 in M2, FR begins 888888, 666666, 000000, 999999, 222222, whereas RB begins 000000, 111111, 222222, 333333, 444444. They are different algorithms and differ at Top-1, Top-3, and Top-5. Their Top-10 sets happen to contain the same ten repeated PINs, explaining equality only at Top-10.

Leakage-Assisted (LA) constructs a separate prefix for each test user. For DOB 1998-03-05 it is 050398 (DDMMYY), 030598 (MMDDYY), 980305 (YYMMDD), 199800 (YYYY00), and 001998 (00YYYY), ordered by the Q2 conditional probabilities. If formats collide (e.g., day equals month), their probabilities are added and the duplicate is kept only at its earliest position. LA then appends FR guesses not already present. Success is the fraction of users whose PIN lies in the first k entries of their individualized order.

F. Security Metrics

For the fully specified generative distribution P , the framework computes Shannon entropy $H(P)$ [8], min-entropy $H_\infty(P)$, and optimal expected guess number $G(P)$. It also reports finite-sample plug-in estimates $\hat{H}$ and $\hat{H}_\infty$ from training counts. These are not theoretical entropy values. The earlier value near 36,979 for M1 was observed-support plug-in guesswork: unseen values had empirical mass zero. It is retained only as a diagnostic in the repository and is not reported as population guesswork.

$$H(P) = - \sum_x p_x \log_2 p_x$$

$$H_\infty(P) = -\log_2 \max_x p_x$$

$$G(P) = \sum_{i=1}^{10^6} i p_{(i)}, \lambda_k = Pr[PIN \in g_{1:k}]$$

where $p_{(i)}$ follows the optimal probability order. M1 therefore has $G = (10^6 + 1)/2 = 500,000.5$. For M3, conditional entropy is:

$$H(PIN|DOB) = E_d[H(P(PIN|d))]$$

which measures residual average uncertainty after DOB is known. Top- k is evaluated on held-out users.

G. Weak PIN Blocklisting Defense

The N most frequent training PINs form B_N , with $N \in \{10, 50, 100, 200, 500\}$. We evaluate two explicit post-rejection behaviors. Under conditional reselection, a blocked user redraws from the same model (and, in M3, retains the same DOB) until an allowed PIN appears; this is the idealized renormalization assumption. Under the incremental-workaround stress case, a blocked value is moved to the next numerically allowed PIN, concentrating rejected mass. The latter is not claimed as a measured human behavior; it demonstrates sensitivity to workaround concentration. For every size and policy, the attacker relearns a post-policy training ranking and is evaluated on a separately transformed test set. FR and per-user LA are evaluated for M1–M3. No tested N is called optimal because the study contains no usability or abandonment measurements.

H. Reproducibility and Verification

The source code, configuration, survey aggregates, tests, and per-seed outputs are available in the public reproducibility repository [22]. All generator and experiment parameters are stored in one JSON file, including survey counts, weights, supports, date intervals and encodings, leakage fractions, sample size, split ratio, blocklist sizes, and seeds. Main tables use seeds 0–19; sensitivity analyses use seeds 0–9. Separate deterministic streams drive generation, splitting, random attack order, and defense reselection. A run manifest records the effective configuration and outputs. Exact metrics are not sample estimates: the implementation enumerates all 106 values and accumulates every branch contribution, including M3's DOB distribution and conditional entropy. Probability vectors are checked to sum to one. Hatted empirical rows are computed separately from training counts. The artifact retains per-seed results, representative FR/RB prefixes, frequency

tables, survey aggregation, and both policies. Eleven tests cover M1 theory, normalization, survey mapping, DOB linkage and collisions, unseen-PIN inclusion, RB independence, LA fallback, and blocked-value exclusion. These checks make the computational claims auditable but do not establish external validity.

IV. EXPERIMENTAL RESULTS

A. Security Metrics

Table II generative-distribution separates metrics exact from training-sample plug-in estimates. M1 has the required values $H = H_\infty = \log_2(10^6) = 19.932 \text{ bits}$ and $G = 500,000.5$. Its lower plug-in entropy, 16.209 bits, is a finite-sample effect rather than a property of the uniform population. Exact M2 guesswork is 78,830.9, not the observed-support diagnostic near 7,385. Exact M3 unconditional entropy exceeds M2, but conditioning on DOB reduces it from 13.682 to 10.570 bits.

TABLE II. EXACT POPULATION METRICS AND EMPIRICAL PLUG-IN ESTIMATES (MEAN $\pm$ STD, 20 SEEDS)

Metric	M1 Uniform	M2 Biased	M3 Leakage
$H(P)$ exact (bits)	19.932	12.965	13.682
$H_\infty(P)$ exact (bits)	19.932	5.214	5.728
$G(P)$ exact	500,000.5	78,830.9	59,562.5
$H(\text{PIN} \mid \text{DOB})$ exact	19.932	12.965	10.570
$\hat{H}_{\text{train}}$ (bits)	16.209 $\pm$ 0.001	11.584 $\pm$ 0.016	12.493 $\pm$ 0.016
$\hat{H}_{\infty, \text{train}}$ (bits)	14.396 $\pm$ 0.218	5.183 $\pm$ 0.014	5.693 $\pm$ 0.030

M1 has the largest exact–plug-in gap because 80,000 training records cover only a fraction of one million values. Assigning unobserved values zero empirical mass lowers H_{train} to 16.209 bits despite the uniform source. This is sampling sparsity, not generator weakness. M2 and M3 samples cover their concentrated regions better, but their plug-in values remain experimental descriptions rather than exact metrics. M2’s $H_\infty = 5.214$ bits reflects strong peak

concentration, consistent with FR’s 2.685% Top-1 success. M3’s DOB-linked mass is more dispersed unconditionally, but its lower conditional entropy predicts that a DOB-aware attacker can recover that mass with a short individualized list, as Table III confirms.

B. Attack Success Rates

TABLE III. TOP-K HELD-OUT SUCCESS (% , MEAN $\pm$ STD, 20 SEEDS)

Model	Atk	Top-1	Top-3	Top-5	Top-10
M1	FR	<0.01	<0.01	<0.01	<0.01
	RB	<0.01	<0.01	<0.01	<0.01
	LA	<0.01	<0.01	<0.01	<0.01
	RD	<0.01	<0.01	<0.01	<0.01
M2	FR	2.685 $\pm$ 0.115	8.076 $\pm$ 0.166	12.299 $\pm$ 0.202	19.833 $\pm$ 0.252
	RB	2.696 $\pm$ 0.106	5.716 $\pm$ 0.117	8.727 $\pm$ 0.195	19.833 $\pm$ 0.252
	LA	0.002 $\pm$ 0.003	0.007 $\pm$ 0.006	0.252 $\pm$ 0.033	12.509 $\pm$ 0.207
	RD	<0.01	<0.01	<0.01	<0.01
M3	FR	1.872 $\pm$ 0.072	5.678 $\pm$ 0.119	8.601 $\pm$ 0.153	13.772 $\pm$ 0.235
	RB	1.923 $\pm$ 0.099	4.007 $\pm$ 0.103	6.042 $\pm$ 0.138	13.772 $\pm$ 0.235
	LA	12.339 $\pm$ 0.250	24.243 $\pm$ 0.362	30.173 $\pm$ 0.470	38.736 $\pm$ 0.409
	RD	<0.01	<0.01	<0.01	<0.01

Table III reports held-out success. Against M2, FR reaches 19.833% $\pm$ 0.252% at Top-10. RB is not identical to FR at Top-1/3/5; equality at Top-10 follows from the shared set of ten repeated PINs. Against M3, LA reaches 38.736% $\pm$ 0.409%, compared with FR at 13.772% $\pm$ 0.235%. Because the five surveyed DOB encodings exhaust the leakage branch, LA already reaches 30.173% $\pm$ 0.470% at Top-5. FR and RB have nearly equal M2 Top-1 success, but FR is higher by 2.360 and 3.573 percentagepoints at Top-3 and Top-5. FR learns the order of repeated values; RB follows digit order. At ten guesses both have exhausted the same repeated-digit set, so equality is a boundary effect, not shared implementation. LA’s assigned DOB is independent of M2 PINs, so only its later FR fallback recovers population mass. In M3, LA captures 12.339% with one candidate and 30.173% with five; the five candidates span all modeled DOB formats and approximately recover the 30% linked branch. FR fallback raises Top-10 to 38.736%, confirming that the gain comes from target-specific linkage.

C. Sensitivity Analysis

Table IV shows that LA rises from 21.118% at $p_l = 0.10$ to 56.235% at $p_l = 0.50$, while FR falls because target-specific probability mass is dispersed across many population values. The sensitivity trend separates two perspectives on the same population. Increasing p_l transfers mass from the shared M2 component to user-specific DOB encodings. A single global ranking therefore loses concentration, and FR Top-10 falls from 17.650% to 9.747%. Once DOB is observed, however, LA can place the relevant encodings first, and its Top-10 rate rises from 21.118% to 56.235%. The trends should not be extrapolated beyond the tested values or interpreted as estimates of real-world DOB use; p_l is a controlled model parameter. The standard deviations are small relative to these between-setting changes because every run contains 20,000 held-out records. They quantify variation over the declared simulation and split seeds, not uncertainty about whether the 142-response convenience sample represents a wider population. External uncertainty in the mixture weights is not captured by the seed-wise deviations and remains a limitation.

TABLE IV. TOP-10 SUCCESS RATE (%) VS. LEAKAGE FRACTION P_l (10 SEEDS)

p_l	FR Top-10 (%)	LA Top-10 (%)
10%	17.650 $\pm$ 0.261	21.118 $\pm$ 0.239
20%	15.763 $\pm$ 0.147	30.009 $\pm$ 0.348
30%	13.789 $\pm$ 0.293	38.767 $\pm$ 0.495
40%	11.869 $\pm$ 0.269	47.563 $\pm$ 0.418
50%	9.747 $\pm$ 0.189	56.235 $\pm$ 0.336

D. Blocklist Results

The blocklist results in Table V are not monotone and depend on reselection. Under conditional reselection, M2 FR falls from a 19.884% baseline to 1.257% at $N=50$, but rises to 1.351% at $N=100$ as the post-policy top ten

changes and finite-sample rankings vary. More importantly, the same policy increases M3 LA from 38.767% to 41.218% at $N = 50$: rejecting common population PINs does not remove a user's DOB encodings, and conditional redrawing can concentrate mass among allowed candidates. Under the workaround stress case, displaced mass can make FR worse than baseline. These findings invalidate a general "N=50 is optimal" claim.

Conditional reselection spreads blocked M2 mass across allowed support. Even then, Top-10 need not fall at every adjacent N : FR is relearned and different values can cross the ten-guess boundary. The small increases from $N = 50$ to $N = 100$ are consistent with re-ranking and finite-sample variation. For M3, a global blocklist cannot place every user's date encodings near the top. Removing common non-personal values can increase the conditional share of allowed DOB candidates, so LA can worsen. A personalized DOB check would target this mechanism but requires separate privacy and usability assessment. The workaround case maps rejected users to near by allowed values, creating peaks learned by a retrained attacker. Although numeric increment is a stress construction, it shows that policy effects depend on post-rejection choices. Deployment studies should measure accepted choices, failures, and abandonment before selecting N .

TABLE V. POST-BLOCKLIST TOP-10 SUCCESS (%), MEAN, 10 SEEDS). C=CONDITIONAL RESELECTION; W=INCREMENTAL WORKAROUND. BASELINES: M2 FR 19.884, M3 FR 13.789, M3 LA 38.767

N	M2 C-FR	M2 W-FR	M3 C-FR	M3 C-LA	M3 W-LA
10	14.618	21.054	9.578	39.897	39.810
50	1.257	21.054	1.447	41.218	38.532
100	1.351	21.086	1.596	40.419	35.945
200	0.032	24.432	0.082	39.651	38.816
500	0.035	24.432	0.036	39.538	38.582

V. DISCUSSION

A. Unconditional Entropy and Conditional Guessability

The earlier phrase “entropy–guess ability inversion” is more precisely a comparison between two information states. M3 has greater unconditional population entropy than M2 (13.682 vs. 12.965 bits), because user-specific dates spread probability over more PIN values. The targeted attacker, however, observes DOB. Under that conditioning, M3 entropy is only 10.570 bits and LA Top-10 is 38.736%. There is no contradiction: $H(\text{PIN})$ and $H(\text{PIN}|\text{DOB})$ answer different questions. The result supports reporting both unconditional population diversity and conditional targeted guess ability rather than interpreting either in isolation.

B. Practical Implications

WebAuthn creates relying-party-scoped public-key credentials, not six-digit shared secrets [23]. NIST SP 800-63B-4 covers syncable authenticators but still permits a numeric local activation PIN, requires at least four characters, recommends at least six and blocklist screening, and caps consecutive failed activation attempts at ten [16]. Thus, our results do not assess WebAuthn/passkey cryptography; they apply to directly verified six-digit PINs and, cautiously, the human-choice layer of PIN-activated authenticators under comparable retry budgets.

The first implication concerns attack budgets. In M2, much of the Top-10 risk is already present by the fifth guess: FR rises from 2.685% at Top-1 to 12.299% at Top-5 and 19.833% at Top-10. A nominal lockout policy therefore cannot be evaluated only by confirming that some limit exists; the chosen threshold and reset process determine how much high-probability mass is exposed. Rate limiting should be combined with device binding, anomaly detection, escalating delay, or step-up authentication so an attacker cannot obtain fresh budgets through resets or parallel channels.

The second implication is that personal data should be treated as authentication-relevant side information. M3 does not make DOB itself an authenticator; rather, it models users who

transform a known attribute into a PIN. The 38.736% LA Top-10 result applies to the stated 30% linkage assumption, not to all users. Nevertheless, it shows why applications should avoid suggesting dates as memorable examples and should warn users that profile information can make a numeric secret predictable. Minimizing unnecessary exposure of DOB and similar identifiers can also reduce an attacker’s inputs.

Third, a population blocklist and a personalized check solve different problems. A population list targets choices that are frequent across users, such as repeated digits. A personalized check targets values derived from one account holder’s DOB or other stored attributes. Combining them might cover both M2 and M3 mechanisms, but a personalized check also creates privacy and governance questions: which attributes may be used, whether derived candidates are retained, how false positives are explained, and whether the check reveals account data to an observer. Those issues require design review beyond a guessability simulation.

Finally, security should be measured on the accepted distribution after rejection, with the attacker retrained. Before deployment, an operator should measure rejection frequency, enrollment time, abandonment, reuse, systematic modifications, and the post-policy Top-k curve. The simulation supplies testable hypotheses under two mechanisms; it does not replace behavioral measurement.

C. Limitations

Survey validity. The 142-response convenience sample is substantially smaller than the 1,010-participant study of Munyendo et al. [3]. No demographic fields were collected, so the available participant description is exhaustive and representativeness by age, occupation, culture, country, or application context cannot be assessed. The question concerns a hypothetical PIN for a non-critical application rather than an observed enrollment decision, and stated strategy may differ from behavior. One Q1 response was recoded and one excluded under disclosed rules. Q2’s date-format distribution is reused for anniversary dates as a modeling assumption. The incorrect MMDDYY example could also have

influenced responses even though the corrected generator follows the label. Wilson intervals quantify point-estimate sampling variation under a binomial approximation, but they do not remove convenience-sampling bias. Larger and more diverse behavioral studies are required before transferring the weights to a deployed population.

Construct validity. The generator is synthetic and assumes uniform choices within categories. Actual users may prefer specific days, years, keypad paths, local holidays, numerological values, or transformations not in the declared lists. Category overlap is computed exactly, but the supports themselves are not claimed to be exhaustive. M2 therefore represents one transparent survey-calibrated population rather than a general empirical model of six-digit PIN choice.

External validation. No actual PINs were collected, and output frequencies were not validated against an empirical corpus. No ethically usable corpus linking six-digit choices to users' DOBs was incorporated. A larger consented study with a held-out population and multiple PII fields is needed before estimating deployment prevalence or comparing directly with RFGuess-PII.

Leakage validity. M3 assumes that the attacker knows DOB exactly and that the linked branch uses one of five surveyed encodings. Real attackers may possess partial, incorrect, or multiple attributes, and real users may transform dates in other ways. DOB frequencies are sampled uniformly over valid days rather than from demographic birth distributions. Consequently, the results neither upper-bound nor lower-bound attacks using phone numbers, identity fragments, anniversaries, addresses, or correlated personal attributes.

Policy validity. The two reselection mechanisms cover only two possible responses. Conditional redraw is idealized, while numeric increment is a stress case and not observed human behavior. Users might instead alter one digit, choose another memorable pattern, reuse a PIN from a different context, or abandon enrollment. The blocklists are learned from each

synthetic training population and may not transfer across countries, applications, or time. No conclusion about an optimal size is possible without behavioral and usability data.

Statistical and operational scope. Seed-wise standard deviations measure Monte Carlo and train/test-split variation conditional on fixed parameters; they are not confidence intervals for survey representativeness or model misspecification. The evaluation assumes a stationary distribution and does not model temporal adaptation by users or attackers. Finally, the threat model is low-attempt online guessing. An offline attacker can enumerate the complete six-digit space, making secure storage, salting, key stretching, hardware protection, and breach response essential regardless of the distributional results.

VI. CONCLUSION

This study shows that six-digit PIN security cannot be characterized by the nominal 106-value space or by unconditional entropy alone. Using a fully specified, survey-calibrated simulation, complete guess rankings, and independently assigned user-specific DOBs, we separated population-level choice DOB-conditioned targeted bias risk from Correct enumeration restores the uniform model's expected guess number to 500,000.5, whereas human-choice bias allows Frequency-Ranked guessing to compromise $19.833\% \pm 0.252\%$ of the modeled M2 population within ten attempts. The central finding is that population diversity and individual protection are different properties. Although M3 has unconditional entropy of 13.682 bits, conditioning on DOB reduces entropy to 10.570 bits, and the Leakage-Assisted attack reaches $38.736\% \pm 0.409\%$ Top-10 success. Thus, a distribution can appear more diverse globally while remaining highly predictable to an attacker with user-specific information. The defense results further show that blocklisting cannot be evaluated independently of user response. Conditional reselection can reduce global frequency-based guessing, but it does not mitigate the modeled DOB-assisted attack; concentrated workarounds can even increase risk. These findings do not

establish a universally effective blocklist size. They instead support evaluating PIN systems through complete attack orders, conditional metrics, realistic post-rejection behavior, and held-out user-level testing. Where feasible, relying-party-scoped public-key authentication avoids making a short PIN the remote credential; where a numeric PIN remains necessary, its accepted-choice distribution and retry controls still require evaluation. Future work should validate the survey weights and reselection assumptions in diverse behavioral studies and investigate consented multi-PIN baselines privacy-preserving personalized screening.

REFERENCES

- [1] P. Markert, D. V. Bailey, M. Golla, M. D'urmuth, and A. J. Aviv, "This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs", *Proc. 41st IEEE Symposium on Security and Privacy (S&P)*, pp. 286–303, 2020. DOI: 10.1109/SP40000.2020.00100.
- [2] H. T. Phuong and T. S. Nam, "Đánh giá độ mạnh mật khẩu sử dụng ngôn ngữ tiếng Việt dựa trên ước lượng entropy," *Journal of Science and Technology on Information security*, vol. 2, no. 8, pp. 58–68, 2020. DOI: 10.54654/isj.v8i2.34.
- [3] C. W. Munyendo et al., "The Same PIN, Just Longer: On the (In)Security of Upgrading PINs from 4 to 6 Digits," *Proc. 31st USENIX Security Symposium*, Boston, MA, USA, pp. 4023–4040, 2022.
- [4] H. Khan, J. Ceci, J. Stegman, A. J. Aviv, R. Dara, and R. Kuber, "Widely Reused and Shared, Infrequently Updated, and Sometimes Inherited: A Holistic View of PIN Authentication in Digital Lives and Beyond," in *Proc. Annual Computer Security Applications Conference (ACSAC)*, Virtual Event, USA, pp. 249–262, 2020. DOI: 10.1145/3427228.3427240.
- [5] S. Uellenbeck, M. D'urmuth, C. Wolf, and T. Holz, "Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Berlin, Germany, pp. 161–172, 2013. DOI: 10.1145/2508859.2516700.
- [6] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith, "Smudge Attacks on Smartphone Touch Screens," in *Proc. 4th USENIX Workshop on Offensive Technologies (WOOT)*, Washington, DC, USA, pp. 1–7, 2010.
- [7] M. Cardaioli, M. Conti, K. Balagani, and P. Gasti, "Your PIN Sounds Good! On the Feasibility of PIN Inference Through Audio Leakage", *Cryptography and Security*, 2019. DOI:10.48550/arXiv.1905.08742.
- [8] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
- [9] J. Bonneau, "The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, San Francisco, CA, USA, pp. 538–552, 2012. DOI: 10.1109/SP.2012.49.
- [10] D. Florencio and C. Herley, "A Large-Scale Study of Web Password Habits," in *Proc. 16th International World Wide Web Conference (WWW)*, Banff, AB, Canada, pp. 657–666, 2007. DOI: 10.1145/1242572.1242661.
- [11] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano, "Passwords and the Evolution of Imperfect Authentication," *Communications of the ACM*, vol. 58, no. 7, pp. 78–87, 2015. DOI: 10.1145/2699390.
- [12] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, "Password Cracking Using Probabilistic Context-Free Grammars," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, Oakland, CA, USA, pp. 391–405, 2009. DOI: 10.1109/SP.2009.8.
- [13] P. G. Kelley et al., "Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, San Francisco, CA, USA, pp. 523–537, 2012. DOI: 10.1109/SP.2012.38.
- [14] W. Melicher et al., "Fast, Lean, and Accurate: Modeling Password Guessability Using Neural Networks," in *Proc. 25th USENIX Security Symposium*, Austin, TX, USA, pp. 175–191, 2016.
- [15] D. Wang, Y. Zou, Z. Zhang, and K. Xiu, "Password Guessing Using Random Forest," in *Proc. 32nd USENIX Security Symposium*, Anaheim, CA, USA, pp. 965–982, 2023.
- [16] D. Temoshok et al., "Digital Identity Guidelines: Authentication and Authenticator Management", *NIST Special Publication 800-63B-4*, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2025. DOI: 10.6028/NIST.SP.800-63b-4.
- [17] G. A. Tuan, N. H. Ha, D. N. Quang, L. H. Ton, and T. T. Hieu, "Weak Links in Smart Surveillance: An Empirical Security Evaluation of Evil Twin Attacks on IoT Cameras," *Journal*

of Science and Technology on Information security, vol. 3, no. 26, pp. 62–69, 2025. DOI: 10.54654/isj.v3i26.1163.

- [18] S. Komanduri et al., “Of Passwords and People: Measuring the Effect of Password-Composition Policies,” in *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, Vancouver, BC, Canada, pp. 2595–2604, 2011. DOI: 10.1145/1978942.1979321.
- [19] B. Ur et al., “How Does Your Password Measure Up? The Effect of Strength Meters on Password Creation,” in *Proc. 21st USENIX Security Symposium*, Bellevue, WA, USA, pp. 65–80, 2012.
- [20] A. Forget, S. Chiasson, and R. Biddle, “Persuasion as Education for Computer Security,” in *Proc. World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education (E-Learn)*, Quebec City, QC, Canada, pp. 822–829, 2007.
- [21] J. Blocki, M. Blum, and A. Datta, “Naturally Rehearsing Passwords,” in *Advances in Cryptology ASIACRYPT 2013*, Bangalore, India, LNCS 8270, pp. 361–380, 2013. DOI: 10.1007/978-3-642-42045-0_19.
- [22] N. V. Khoa and G. A. Tuan, “Low-Entropy PIN Guessability: Source Code and Reproducibility Artifact”, (2026). Access time: 13/8/2026, https://github.com/gianganh Tuan/low_entropy_attt.
- [23] W3C Web Authentication Working Group, “Web Authentication: An API for Accessing Public Key Credentials - Level 3” (2026). Access time: 13/8/2026, <https://www.w3.org/TR/webauthn-3/>.

ABOUT THE AUTHOR

Giang Anh Tuan



Workplace: University of Science and Technology of Hanoi, Vietnam

Email: giang-anh.tuan@usth.edu.vn

Education: PhD (Paris Sud XI, 2014)

Recent research interests: cybersecurity, network security, and authentication systems performance

Tên tác giả: **Giang Anh Tuấn**

Cơ quan công tác: Đại học Khoa học và Công nghệ Hà Nội, Việt Nam

Email: giang-anh.tuan@usth.edu.vn

Quá trình đào tạo: Tiến sỹ (Paris Sud XI, 2014)

Hướng nghiên cứu hiện nay: an toàn, an ninh mạng, đánh giá hiệu năng các hệ thống xác thực.

Nguyen Viet Khoa



Workplace: University of Science and Technology of Hanoi, Vietnam

Email:

khoanv.23bi14223@usth.edu.vn

Education: B.Sc. in Information and Communication Technology, University of Science and Technology of Hanoi, Vietnam, 2026.

Recent research interests: authentication security, PIN guessability, simulation-based security evaluation.

Tên tác giả: **Nguyễn Việt Khoa**

Cơ quan công tác: Đại học Khoa học và Công nghệ Hà Nội, Việt Nam

Email: khoanv.23bi14223@usth.edu.vn

Quá trình đào tạo: Cử nhân Công nghệ thông tin Đại học Khoa học và Công nghệ Hà Nội năm 2026.

Hướng nghiên cứu hiện nay: an toàn xác thực, dự đoán PIN, đánh giá an toàn trên mô phỏng.