Five Word Password Composition Policy

Sirvan Almasi
Imperial College London
sirvan.almasi17@imperial.ac.uk

William J. Knottenbelt
Imperial College London
w.knottenbelt@imperial.ac.uk

***Abstract*—Password composition policies (PCPs) are critical security rules that govern how users create passwords for online authentication. Despite passwords remaining the primary authentication method online, there is significant disagreement among experts, regulatory bodies, and researchers about what constitutes effective password policies. This lack of consensus has led to high variance in PCP implementations across websites, leaving both developers and users uncertain. Current approaches lack a theoretical foundation for evaluating and comparing different password composition policies. We show that a structure-based policy, such as the three-random words recommended by UK's National Cyber Security Centre (NCSC), can improve password security. We demonstrate this using an empirical evaluation of labelled password datasets and a new theoretical framework. Using these methods we demonstrate the feasibility and security of multi-word password policy and extend the NCSC's recommendation to five words to account for non-uniform word selection. These findings provide an evidence-based framework for password policy development and suggest that current web authentication systems should adjust their minimum word requirements upward while maintaining usability.**

I. Introduction

The lack of consensus on password composition policy (PCP) and the lack of adoption of best practices on the web [1]–[8] suggests that our knowledge on password security is not yet complete. Although passwords remain the main method of authentication, experts and regulatory bodies disagree fundamentally about what makes a good password policy. State agencies recommend one set of rules, while professional organisations suggest another, and academic researchers advocate for yet different approaches. This lack of consensus reveals a troubling gap in our understanding of password security. What password composition rules should web developers implement to best protect their users?

National organisations such as NIST (US) [9] and NCSC (UK) [10] provide similar guidance on PCP, UK's NCSC additionally recommends minimum of three word password structure. This paper is most interested in structure related policies to enhance user secret security–should a system or web admin recommend such three word password structure policies? In this paper, we develop a model that allows us to reason analytically about PCPs. Using this, we tackle the PCP recommended by the UK's NCSC (minimum of three random words). We present a novel theoretical framework for analysing password composition policies, bridging the gap between security requirements and user behaviour.

The gap between research and practice compounds the inconsistency in the PCP recommendation. Consider Dropbox, which developed `zxcvbn` [11] Password Strength Meter (PSM). Despite creating this tool, which has 14.7K Github stars and has been validated by independent research [12], Dropbox's own web authentication policy contradicts current best practices. Their website requires passwords with a minimum of eight characters, one letter, one digit, and one symbol. Even more concerning, they do not appear to use their own PSM. `zxcvbn` [11] has some weaknesses; it awards its maximum score of 4/4 to `girlsruledaworld` and a strong 3/4 to `asdf1fdsa1`—passwords that we would consider vulnerable as they are common phrases or sequences.

Threat Scenario. We assume that some users will choose memorable passwords, eschewing password managers or WebAuthn alternatives. Although password managers offer stronger security, our model addresses the reality that many users, from a broad population, will rely on human-generated passwords. Passwords are also useful in offline and air-gapped systems too where out-of-band systems are limited. Therefore, system and web admins have to consider such scenarios. We therefore analyse password security against the most challenging threat model: an offline attack where adversaries can make unlimited guessing attempts without rate limiting.

Our Contributions are as Follows:

- Empirical validation of a structure-based password policy in addition to existing policies. This provides quantitative support for existing password composition recommendations, particularly the NCSC's three-word policy
- A formal mathematical model describing the recommended number of words to mitigate user behaviours such as choosing words from a non-uniform distribution.

This work addresses core areas of web security, and our results provide actionable insights for improving password policies across web platforms while maintaining usability. Our findings and insights provide immediate applications for system and web admins. Our work also provides a theoretical and reasoning basis for future password composition policy development and research.

Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb) 2025
28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-9-0
https://dx.doi.org/10.14722/madweb.2025.23020
www.ndss-symposium.org

II. Background

Password Strength. This metric measures the difficulty of guessing or cracking a password. However, there is no consensus on how to quantify such difficulty. Some experts use dictionary-based methods, while others rely on probabilistic deep-learning based approaches. Nonetheless, certain lower-bound measures are widely accepted, such as the minimum number of characters required to prevent brute-force attacks.

Shannon's entropy (Eq. 1) quantifies the uncertainty or information content associated with a random variable.

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{1}$$

In the context of password strength, it measures how unpredictable or "random" a password is. The higher the entropy, the more difficult the password is to guess, making it stronger. This metric estimates the minimum number of bits required to encode the password. However, the entropy measure for passwords is often given in the form of $\log_2(R^L)$, where R represents the total number of characters in the character sets used in the password, and L is the length of the password.

The limitation of $\log_2(R^L)$ is that it assumes all R characters are equally likely, disregarding the dependency between characters in human-chosen passwords. This approach oversimplifies, as passwords like `8zm!U6gNL%` and `p4$sW0rd1!` are calculated to have the same entropy. However, such an assumption is unrealistic, as the latter password is essentially the word *password* with predictable transformations.

Password Composition Policies (PCP). PCP is set of rules, often in natural language, that only allows the user to sample from a password space that is assumed to be resistant to guessing attacks. The *National Cyber Security Centre (NCSC)* [10] in the UK and the *National Institute of Standards and Technology (NIST)* [9] in the USA are leading authorities in cyber-security. Their recommendations, particularly on password policies, should serve as standards or benchmarks in the field. Specifically, the NCSC advocates for the use of three random words [13] and advises against mandatory password complexity requirements [14]. This approach, detailed in [15], is based on the principle that a sequence of three random words is both easier for users to remember and sufficiently complex to deter unauthorised access, striking a balance between security and user convenience. This three word and structure-based PCP is conditioned on the following:

- **Length**. To satisfy the minimum length requirements.
- **Impact**. Easy to adopt and understand.
- **Novelty**. Unlikely to find it in a breached database.
- **Usability**. More likely to be remembered.

NIST's [16] PCP is as follows. Passwords should have a minimum length of 8 characters, but systems should support more than 64 characters. It is advised to compare the selected password against prior breach data, dictionary entries, and repetitive patterns. Context-specific terms should also be considered. Additionally, users should be aided with strength meters to gauge password robustness. While rate limiting should be enforced during authentication attempts, use of specific composition rules is discouraged. OWASP [17] follows the recommendations set by NIST.

passwordNinja Dataset [18] is a human and machine labelled datasets that has decomposed password strings into its finer constituents. For example, the password `0rwell1984` would be structured as follows.

$$\underbrace{\texttt{0rwell}}_{\text{word } (w)}\underbrace{\texttt{1984}}_{\text{date}}$$

The passwordNinja [18] dataset is composed of a complete hotmail and LizardSquad [19] breached passwords from 2009 and 2015 respectively. The hotmail dataset, which leaked in 2009 [20], comprises 8929 records of passwords. Public sources indicate that these passwords were procured through phishing. Abundant evidence within the dataset supports this claim. For instance, we observe numerous passwords with inverted capitalisation and similar passwords that contain typos, suggesting that users might have unintentionally activated `CAPS LOCK`. The LizardSquad dataset contains 11 808 records.

III. Password Composition Policy

There is a lack of consensus on password composition policy amongst industry experts, academics and government agencies. This has created some confusion which is evident by the fact that most web applications do not follow any best practices [1]–[8]. This raises the question which PCP is the *right* one? Specifically, should we have a structure-based policy as recommended by UK's NCSC?

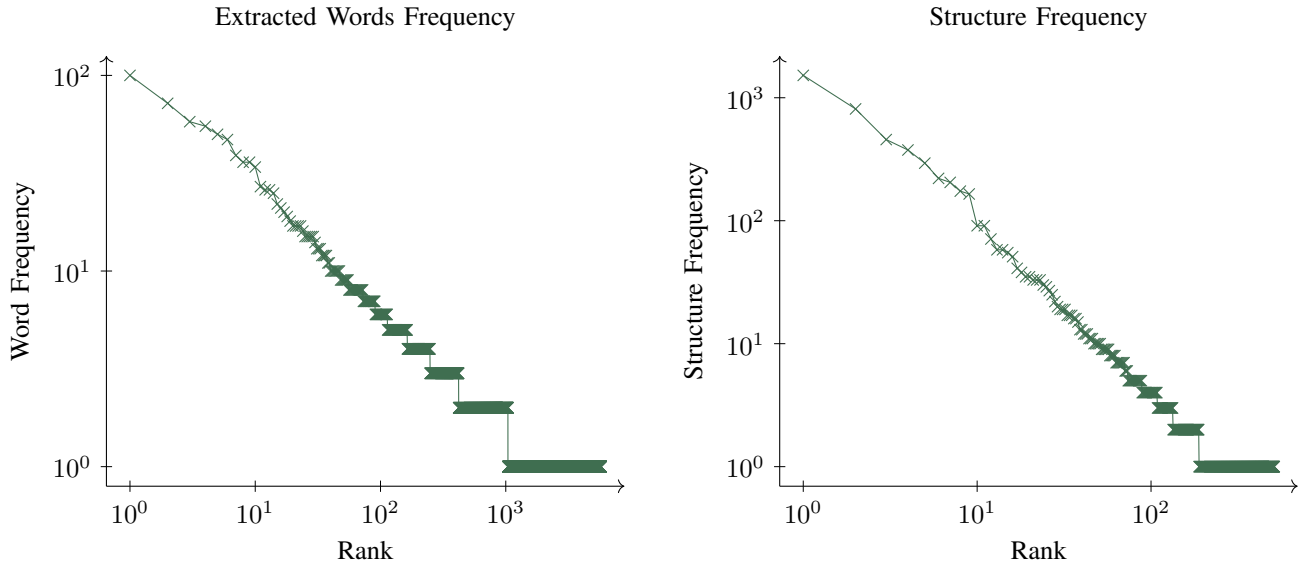
To answer this question we pay particular attention to the NCSC's PCP of three random words. We first develop a mathematical model that incorporates password structures, this is novel and enabled by the passwordNinja dataset.

A. Empirical Evaluation

To analyse how password structure affects security, we draw on the passwordNinja [18] dataset. This dataset provides something previously unavailable to researchers: individually labelled passwords that reveal their structural composition. This labelling is essential for our analysis, as it allows us to systematically evaluate how different password structures influence the security strength.

Our analysis of password datasets reveals several key vulnerabilities in how users choose passwords. First, users select words from a highly biased distribution, making passwords more predictable (Figure 1a). Second, commonly used password structures contain patterns that attackers can exploit using techniques such as Probabilistic Context-free Grammar (`PCFG` [22]) or mask attack mode in Hashcat [21].

As shown in Table I, single words (w) and names (n) dominate password structures, making them vulnerable to dictionary attacks. The effectiveness of these attacks varies by hash function; our tests using consumer hardware demonstrate significantly different cracking times across hash functions. A newer dataset (LizardSquad [19] 2015) shows similar structural patterns (More detail in Table III following Reference section.)



(a) Unique Words

(b) Structures

Figure 1: Charts showing the log-log plot of word and structure against their respective rank for the hotmail dataset. Words are from extracted unique passwords. The distribution of both words and structure exhibit a power law similar to Zipf's law.

However, passwords containing multiple words (such as w^4 structures) demonstrate much stronger security, resisting hybrid attacks even when hashed with MD5. Yet this strength depends critically on word choice. When users select related words or popular phrases (such as song lyrics), passwords become more predictable. For example, the phrase "`mira que eres canalla`" is vulnerable even though it is made up of four words. This is because it originates from Mexican folk music.

To maximise security, users should select words with minimal semantic or probabilistic relationships in their sequence. This prevents attackers from using language models to predict subsequent words in multi-word passwords or passphrases.

Our analysis of the passwordNinja datasets supports the NCSC guidelines. Testing with `zxcvbn` [11] (Table II) confirms that multi-word passwords do improve the security of human-chosen and memorable password.

B. Theoretical Evaluation

1) Password Structure Probability: Let S denote the structure of a password, where $S(p) = (e_1, e_2, \ldots, e_n)$ and each $e_i \in \Sigma$ (e.g., $\Sigma = \{w, n, d, s, \ldots\}$). The probability of a structure S is given by $P(S) = P(e_1, e_2, \ldots, e_n)$. Assuming dependence between the elements in the structure, this can be expressed as:

$$P(S) = P(e_1) \prod_{i=2}^{n} P(e_i \mid e_{i-1}) \tag{2}$$

2) Combined Relationship Probability: If e_i represents the type of the element, then let $\hat{e_i}$ represent the instance of the element, e.g. $\hat{e} \rightarrow$ *password* if $e = w$ (i.e. a word). We can represent both structure and instance relationships using a joint distribution. For a password p with elements $(e_1, ..., e_n)$, we can define:

$$P(p) = P(S(p),\, E(p)) \tag{3}$$
$$= P(S(p))\, P(E(p) \mid S(p)) \tag{4}$$

Where $E(p)$ represents the specific element instances.

$$P(E(p) \mid S(p)) = P(\hat{e}_1) \prod_{i=2}^{n} P(\hat{e}_i \mid \hat{e}_{i-1}, S(p)) \tag{5}$$

The final equation simplified is:

$$P(p) = P(e_1)P(\hat{e}_1) \prod_{i=2}^{n} [P(e_i \mid e_{i-1})P(\hat{e}_i \mid \hat{e}_{i-1}, S(p))] \tag{6}$$

3) Incorporating NCSC's Recommendation: Aligning with the NCSC's recommendation of using at least three random words, we define a specific structure S^* where $S^* = (w,\ w,\ w)$. We assume a fixed structure of just three words, thus $P(S^*) = 1$. Using Eq. 4, we are left with $P(p \mid S^*) = P(E(p) \mid S(p))$. We also assume that the words are chosen independently. Thus, we have:

$$P(p \mid S^*) = \prod_{i=1}^{3} P(w_i) \tag{7}$$

C. Non-uniform NCSC PCP

In practice humans do not select words at random and the passwordNinja [18] dataset shows that the selection of words follows a power-law like distribution. This has significant implications for the NCSC's recommendation of using at least three random words for password composition. Below is a mathematical framework to model this scenario and assess the impact on the password composition policy.

Table I: The table presents the top 25 labelled password structures from the passwordNinja [18] dataset, accounting for 5 032 (69.5%) of the labelled dataset. Hashcat [21] benchmarks indicate the performance of consumer hardware in a hybrid attack. The dictionary size for $\boldsymbol{w}$ (words) and $\boldsymbol{n}$ (numbers) is set at 5×10^5. `zxcvbn` [11] serves as a password strength meter, with scores ranging from 0 to 4, and guess estimates are log based.

	Frequency		Hashcat using Nvidia RTX 4090				`zxcvbn` Mean	
Structure *log* **perm'**	**Count**	**%**	**MD5**	**PBKDF2-sha256**	**scrypt**	**bcrypt-sha512**	**Score**	**Guesses**
w	1520	21.0	~ 0	0.1 s	1 min	4 min	1.3	5.36
n	811	11.2	~ 0	0.1 s	1 min	4 min	1.0	4.72
nd	58	0.8	~ 0	0.6 s	12 min	42 min	1.4	5.68
wd	91	1.3	~ 0	0.6 s	12 min	42 min	1.5	5.97
nl	33	0.5	~ 0	1.5 s	30 min	2 h	1.2	5.09
wdd	294	4.1	~ 0	5.6 s	2 h	7 h	1.8	6.69
ndd	221	3.1	~ 0	5.6 s	2 h	7 h	1.6	6.33
nddd	71	1.0	~ 0	56.4 s	19 h	3 day	1.7	6.37
wddd	91	1.3	~ 0	56.4 s	19 h	3 day	1.9	6.84
wsdd	29	0.4	~ 0	3 min	3 day	9 day	2.0	7.28
ddddw	35	0.5	~ 0	9 min	8 day	29 day	2.4	7.95
wdddd	174	2.4	~ 0	9 min	8 day	29 day	2.3	7.59
ndddd	205	2.8	~ 0	9 min	8 day	29 day	2.0	6.90
nn	376	5.2	1.5 s	8 h	1 yr	4 yr	2.0	6.91
ww	457	6.3	1.5 s	8 h	1 yr	4 yr	2.0	7.04
nw	30	0.4	1.5 s	8 h	1 yr	4 yr	2.2	7.51
nddddd	51	0.7	3.0 s	16 h	2 yr	8 yr	2.5	8.05
wdddddd	33	0.5	3.0 s	16 h	2 yr	8 yr	2.8	8.84
wwdd	58	0.8	3 min	33 day	111 yr	402 yr	2.9	8.97
nndd	38	0.5	3 min	33 day	111 yr	402 yr	3.2	9.62
wwddd	33	0.5	4 h	9 yr	1.11×10^4 yr	4.02×10^4 yr	3.3	9.81
nwn	41	0.6	9 day	447 yr	5.56×10^5 yr	2.01×10^6 yr	3.0	8.88
www	165	2.3	9 day	447 yr	5.56×10^5 yr	2.01×10^6 yr	3.0	9.26
nww	35	0.5	9 day	447 yr	5.56×10^5 yr	2.01×10^6 yr	2.7	8.18
nnn	27	0.4	9 day	447 yr	5.56×10^5 yr	2.01×10^6 yr	3.3	9.95
wwww	55	0.8	171 day	8.70×10^3 yr	1.08×10^7 yr	3.91×10^7 yr	3.5	11.24

Table II: Top 10 password structures for each `zxcvbn` [11] score of 3 an 4. These are the password structures that are categorised in the most secure bin (4) by the `zxcvbn` algorithm; Avg Guesses is the `zxcvbn guesses_log10`.

Structure		**Count**	**Avg Guesses**	**Avg Length**
zxcvbn score = 4				
r2	$r2$	160	12.6	14
www	w^3	57	12.1	14
ww	w^2	48	12.0	14
wwww	w^4	34	13.1	15
w	w	25	11.3	12
nn	n^2	24	11.2	13
wdddd	wd^4	19	11.6	12
wwdddd	w^2d^4	16	11.3	14
nndd	n^2d^2	15	11.5	13
ndddd	nd^4	14	10.8	12
zxcvbn score = 3				
r2	$r2$	185	8.9	10
w	w	100	8.8	10
ww	w^2	86	8.8	11
nn	n^2	77	8.8	11
www	w^3	62	9.1	11
wdddd	wd^4	44	8.8	11
ndddd	nd^4	43	8.7	10
wdd	wd^2	38	9.0	10
wwdd	w^2d^2	31	9.1	11
n	n	26	8.9	10

1) Word Frequency Distribution: Let w denote a word from the vocabulary $\mathcal{W}$. Under Zipf's law, the probability $P(w)$ of selecting the r-th most frequent word is given by:

$$P(w_r) = \frac{1/r^s}{\sum_{k=1}^{|\mathcal{W}|} 1/k^s} \tag{8}$$

where s is the exponent characterizing the distribution (typically $s \approx 1$ for natural languages).

For a password p consisting of n words $(w_1, w_2, \ldots, w_n)$ selected independently according to Zipf's law, the structure probability $P(S)$ where $S = (w, w, \ldots, w)$ is:

$$P(S) = \prod_{i=1}^{n} P(w_i) = \prod_{i=1}^{n} \frac{1/r_i^s}{\sum_{k=1}^{|\mathcal{W}|} 1/k^s} \tag{9}$$

where r_i is the rank of the i-th word.

2) Entropy of the Password Distribution under Zipf's Law: The entropy H of the password distribution. Substituting $P(p)$ with the structure probability:

$$H = -\sum_{w_i \in \mathcal{W}}^{n} \left(\prod_{i=1}^{n} \frac{1/r_i^s}{\sum_{k=1}^{|\mathcal{W}|} 1/k^s} \right) \log \left(\prod_{i=1}^{n} \frac{1/r_i^s}{\sum_{k=1}^{|\mathcal{W}|} 1/k^s} \right) \tag{10}$$

Simplifying the logarithm:

$$H = -\sum_{i=1}^{n} \sum_{w_i \in \mathcal{W}} P(w_i) \log P(w_i) = n \cdot H_{\text{word}} \tag{11}$$

where H_{word} is the entropy per word. We will now use this to draw a comparison with the uniform distribution, which we take to be the theoretical version of the NCSC's PCP.

3) Impact on NCSC's Recommendation: To align the password policy with enhanced entropy under Zipf's law, we need to determine the minimum number of words n required to achieve a desired entropy H_{desired}. This can be formulated as:

$$n \geq \frac{H_{\text{desired}}}{H_{\text{word}}}$$

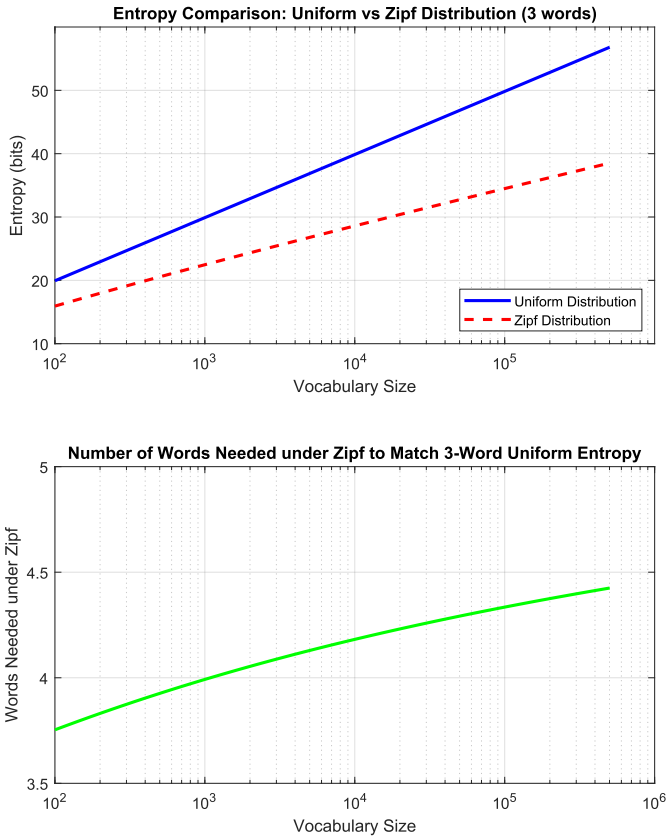


Figure 2: Chart showing the equivalent number of words required in practice to match the theoretical complexity of the National Cyber Security Centre's (NCSC) password composition policy. The NCSC recommends at least three random words, using our model we show that in practice a minium of 4 words is more secure.

Given that H_{word} is lower under Zipf's law compared to a uniform distribution (due to the presence of highly probable words), a larger n may be required to achieve the same entropy.

Under a uniform distribution where each word has $P(w) = \frac{1}{|\mathcal{W}|}$, the entropy per word is:

$$H_{\text{word}}^{\text{uniform}} = \log |\mathcal{W}|$$

Comparing with Zipf's law:

$$H_{\text{word}}^{\text{Zipf}} < H_{\text{word}}^{\text{uniform}}$$

Zipf's law requires more words than a uniform distribution to achieve the same entropy level. As Figure 2 demonstrates, when we increase the vocabulary size, the entropy difference between uniform and Zipf's distributions also increases. This comparison leads us to conclude that we need at least 4-5 words to match the entropy that NCSC's theoretical PCP describes. Moreover, this minimum word requirement grows as the vocabulary expands.

To maintain or exceed the entropy level recommended by NCSC using three random words under a Zipfian distribution, the policy should be adjusted to:

$$n_{\text{adjusted}} \geq \frac{3 \cdot \log |\mathcal{W}|}{H_{\text{word}}^{\text{Zipf}}}$$

This ensures that the password composition maintains sufficient entropy despite the skewed word frequency distribution.

4) Adjusted Password Structure Probability: If we adjust the structure to use n words to counteract the lowered entropy per word, the structure probability becomes:

$$P(S_n) = \prod_{i=1}^{n} P(w_i) = \left(\frac{1}{\sum_{k=1}^{|\mathcal{W}|} 1/k^s} \right)^n \prod_{i=1}^{n} \frac{1}{r_i^s} \tag{12}$$

Ensuring that n is sufficiently large to maintain high entropy despite the non-uniform word distribution.

IV. Discussion

Password Composition Policy is a culmination of our understanding of password security, it takes into account password complexity (or strength), password guessing algorithms, and user behaviour. Whilst many policies exist such as minimum number of characters, the NCSC recommendation of three random words is the only one that proposes a *structure* policy. This paper has then analysed this policy and its necessity. We find that a structure-based policy in addition to other existing policies does improve password strength.

The passwordNinja dataset showed that the expected complexity of password structures that are unrestricted (i.e. no recommendation or policy) is weak. In order for a system or web admin to increase the expected complexity of their dataset then they can recommend their users to adopt complex password structures, such as multiple words.

Using a theoretical model we show that in practice a user would need 4 to 5 words to mitigate the lack of randomness in choosing a word. Our recommendation is then a five word password structure policy to offset non-uniform word selection.

A. Ethical Considerations

In handling sensitive password data, we enforce strict ethical safeguards. We limit our analysis to the well-established SecLists datasets [23], which researchers have extensively used in prior password-guessing literature. This choice allows us to:

- Avoid releasing any new compromised data.
- Ensure complete removal of personally identifiable information (PII)—No PII is processed or stored.
- Build upon an established ethical framework in password research. The datasets used in this study have been regularly used and vetted in prior research.

B. Limitations

The passwordNinja [18], while valuable, represents a specific snapshot in time (2009-2015) and primarily includes English and Spanish-language passwords. This could limit the generalisability of our findings. However, even between 2009 and 2015, there were insignificant differences in the characteristics of the data. This leads us to believe that patterns observed in the older datasets would also be present in more recent ones.

The evaluation of our theoretical model could, in the future, account for semantic relationships between instances of elements. Although we accounted for this in our model, we did not consider these semantic relationships in the final

evaluation. Future research could explore how these semantic connections influence the overall system behaviour, particularly in edge cases where traditional evaluation metrics may fall short. However, this limitation does not diminish the significance of our findings. Moreover, the theoretical model could evaluate other types of password structures beyond three words.

Despite these limitations, our findings provide strong evidence supporting the need for adjusted password composition policies that account for real-world user behaviour while maintaining security requirements. The proposed five-word minimum represents a practical enhancement to existing guidelines that balances security needs with human factors in password creation.

V. Related Work

Despite extensive research in password complexity measures and meters, **recommendations** for password policies remain scarce. Bonk *et al.* [24] offer guidance on constructing passphrases, suggesting longer sequences, such as seven words or more. The study found that long passphrases (7+ words) achieved login success rates of 74-86% with average login times of 25-30 seconds, suggesting reasonable usability for high-security accounts accessed daily. However, the paper mainly focussed on usability. Our recommendation of five words provides a more quantitative reasoning behind its choice. Gerlitz *et al.* [8] provided a comprehensive analysis of password composition policies in Germany.

Early password modelling techniques, such as the Markov n-gram model [25], focused on individual characters, positing that *"the distribution of letters in easy-to-remember passwords likely mirrors the letter distribution in the user's native language."* We have validated this hypothesis at the word level. The observation that letter distribution is not uniform holds true for words, exhibiting a power-law distribution akin to Zipf's law. Modelling passwords based on Probabilistic Context-Free Grammars (`PCFG` [22]) was the next innovation in password modelling. Our insights into empirical password structures can refine `PCFG` modelling. Incorporating W_n and N_n as variables for words and names respectively enhances the computational viability of modelling more intricate structures. Subsequent advancements have leveraged machine learning techniques in password modelling, including neural networks [26], GANs [27], RNNs [28], and Transformers [29], among others.

Historical analyses of password characteristics have focused on *length*, *count*, and *character-level structures* [30]–[37].

`zxcvbn` [11] employs dictionaries and bespoke rules to gauge password strength on a scale from 0 to 4. However, `zxcvbn` has its limitations; for example, it incorrectly parses the password `gomythsun` as `go myths un`, impacting its strength rating. Additionally, it struggles with transformations like `numbr` from `number`. Wang *et al.* [38] delve deeper into evaluating password strength meters. Their endorsement of `zxcvbn` and Markov n-gram models informed our method.

In an analysis of Chinese passwords [31], Li *et al.* conduct a broad analysis on the syntax of passwords. Our research extends their work by digging deeper into the password structures. For instance, Li *et al.* identified $LLLLLL$ as the predominant structure within the Rockyou dataset, typically representing a word as we showed in our results. Further structural analysis on various password datasets is presented in [39].

Das *et al.* [32] explored password syntax, notably introducing analysis on word or phrase transformations. However, we disagree with their classification of sequential keys, alternate keys, and sequential alphabet as transformations, viewing these more as sequences. Riddle *et al.* [40] delve into the composition and semantics of passwords, with an emphasis on the psychological significance of words.

VI. Conclusion

We have developed a rigorous model that incorporates password structure and empirical user behaviour patterns, filling a critical gap in our understanding of password security. Through analysis of real-world password datasets, we demonstrate that structure-based password policy such as three-words (w, w, w) empirically and theoretically offer security advantages. However, the NCSC's previously recommended three-word password policy, while theoretically sound, fail to account for users' non-uniform word selection which follows a Zipf-like distribution. Our framework provides both theoretical foundations and practical guidelines for developing more robust authentication policies that reflect actual user behaviour rather than theoretical ideals.

We recommend extending the NCSC's password composition policy to require five words, which our models show provides comparable security to the theoretical three-word ideal even under real-world usage patterns. This recommendation maintains usability while providing demonstrably stronger protection.

Future work includes incorporating *memorability* into our optimisation problem. Memorability represents the cognitive overhead with respect to some password structure. Memorability would capture the expected structure given some policy.

Acknowledgement

The authors would like to thank *Nora A.* for her discussions, generous time, and support.

References

[1] S. Alroomi and F. Li, "Measuring Website Password Creation Policies At Scale," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '23. New York, NY, USA: Association for Computing Machinery, Nov. 2023, pp. 3108–3122, real_world_eval. [Online]. Available: https://dl.acm.org/doi/10.1145/3576915.3623156

[2] J. Lomchan, R. Wiangsripanawan, and S. F. Shahandashti, "The Comparison of Password Composition Policies Among US, German, and Thailand Samples," in *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Jun. 2023, pp. 213–218, real_world_eval. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10202141

[3] K. Lim, J. H. Kang, M. Dixson, H. Koo, and D. Kim, "Evaluating Password Composition Policy and Password Meters of Popular Websites," in *2023 IEEE Security and Privacy Workshops (SPW)*, May 2023, pp. 12–20, iSSN: 2770-8411. [Online]. Available: https://ieeexplore.ieee.org/document/10188654

[4] P. B. Maoneke and S. Flowerday, "Password Policies Adopted by South African Organizations: Influential Factors and Weaknesses," in *Information Security*, H. Venter, M. Loock, M. Coetzee, M. Eloff, and J. Eloff, Eds. Cham: Springer International Publishing, 2019, pp. 30–43, real_world_eval.

[5] S. Furnell, "Assessing password guidance and enforcement on leading websites," *Computer Fraud & Security*, vol. 2011, no. 12, pp. 10–18, Dec. 2011, real_world_eval. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361372311701233

[6] S. S. Woo, K. J. Jung, and B. J. Choi, "Survey on Current Password Composition Policies," *Review of KIISC*, vol. 28, no. 1, pp. 43–47, 2018, real_world_eval. [Online]. Available: https://koreascience.kr/article/JAKO201809253685093.page

[7] K. Lee, S. Sjöberg, and A. Narayanan, "Password policies of most top websites fail to follow best practices," in *Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security*, ser. SOUPS'22. USA: USENIX Association, Aug. 2022, pp. 561–580.

[8] E. Gerlitz, M. Häring, and M. Smith, "Please do not use !?_ or your License Plate Number: Analyzing Password Policies in German Companies," 2021, pp. 17–36. [Online]. Available: https://www.usenix.org/conference/soups2021/presentation/gerlitz

[9] "National institute of standards and technology," https://www.nist.gov, Feb. 2024, last Modified: 2024-02-07T09:49-05:00. [Online]. Available: https://www.nist.gov

[10] "National cyber security centre," https://www.ncsc.gov.uk, 2024. [Online]. Available: https://www.ncsc.gov.uk

[11] D. L. Wheeler, "zxcvbn: Low-Budget password strength estimation," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, Aug. 2016, pp. 157–173. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/wheeler

[12] D. Wang, X. Shan, Q. Dong, Y. Shen, and C. Jia, "No Single Silver Bullet: Measuring the Accuracy of Password Strength Meters," 2023, pp. 947–964. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/wang-ding-silver-bullet

[13] I. McCormack, "Three random words or #thinkrandom." [Online]. Available: https://www.ncsc.gov.uk/blog-post/three-random-words-or-thinkrandom-0

[14] NCSC, "Password policy: updating your approach," https://www.ncsc.gov.uk/collection/passwords/updating-your-approach, 2018. [Online]. Available: https://www.ncsc.gov.uk/collection/passwords/updating-your-approach

[15] R. Kate, "The logic behind three random words," Aug. 2021, https://www.ncsc.gov.uk/blog-post/the-logic-behind-three-random-words. [Online]. Available: https://www.ncsc.gov.uk/blog-post/the-logic-behind-three-random-words

[16] P. Grassi, M. Garcia, and J. Fenton, "Digital identity guidelines," Mar. 2020. [Online]. Available: https://csrc.nist.gov/pubs/sp/800/63/3/upd2/final

[17] OWASP, "OWASP Password Composition Policy Recommendation," 2024, pcp_owasp. [Online]. Available: https://cheatsheetseries.owasp.org/cheatsheets/Authentication_Cheat_Sheet.html

[18] S. Almasi, "Passwordninja: Human-labelled and real password datasets," 2024, human-annotated password dataset with linguistic transformations. [Online]. Available: https://github.com/sirvan3tr/passwordninja

[19] D. Miessler, "LizardSquad Dataset SecLists," 2015. [Online]. Available: https://github.com/danielmiessler/SecLists/blob/master/Passwords/Leaked-Databases/Lizard-Squad.txt

[20] B. Johnson and S. Francisco, "Hotmail password breach blamed on phishing attack," *The Guardian*, Oct. 2009. [Online]. Available: https://www.theguardian.com/technology/2009/oct/06/hotmail-phishing

[21] J. Steube, "Hashcat - advanced password recovery," Jun. 2023. [Online]. Available: https://hashcat.net

[22] M. Weir, S. Aggarwal, B. d. Medeiros, and B. Glodek, "Password Cracking Using Probabilistic Context-Free Grammars," in *2009 30th IEEE Symposium on Security and Privacy*, May 2009, pp. 391–405, iSSN: 2375-1207.

[23] D. Miessler, "danielmiessler/SecLists," Apr. 2024, original-date: 2012-02-19T01:30:18Z. [Online]. Available: https://github.com/danielmiessler/SecLists

[24] C. Bonk, Z. Parish, J. Thorpe, and A. Salehi-Abari, "Long passphrases: Potentials and limits," no. arXiv:2110.08971, Oct. 2021, arXiv:2110.08971 [cs]. [Online]. Available: http://arxiv.org/abs/2110.08971

[25] A. Narayanan and V. Shmatikov, "Fast dictionary attacks on passwords using time-space tradeoff," in *Proceedings of the 12th ACM conference on Computer and communications security*, ser. CCS '05. New York, NY, USA: Association for Computing Machinery, Nov. 2005, p. 364–372, https://www.cs.utexas.edu/~shmat/shmat_ccs05pwd.pdf. [Online]. Available: https://dl.acm.org/doi/10.1145/1102120.1102168

[26] L. de Castro, H. Lang, S. Liu, and C. Mata, "Modeling password guessing with neural networks," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:12251269

[27] B. Hitaj, P. Gasti, G. Ateniese, and F. Perez-Cruz, "Passgan: A deep learning approach for password guessing," in *Applied Cryptography and Network Security*, vol. 11464. Springer, 2019, p. 217–237, accepted: 2020-02-13T13:21:35Z. [Online]. Available: https://www.research-collection.ethz.ch/handle/20.500.11850/386747?locale-attribute=de

[28] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor, "Fast, lean, and accurate: Modeling password guessability using neural networks," 2016, p. 175–191. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/melicher

[29] M. Xu, J. Yu, X. Zhang, C. Wang, S. Zhang, H. Wu, and W. Han, "Improving real-world password guessing attacks via bi-directional transformers," 2023, p. 1001–1018. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/xu-ming

[30] J. Ma, W. Yang, M. Luo, and N. Li, "A study of probabilistic password models," in *2014 IEEE Symposium on Security and Privacy*, May 2014, p. 689–704, https://ieeexplore.ieee.org/document/6956595. [Online]. Available: https://ieeexplore.ieee.org/document/6956595

[31] Z. Li, W. Han, and W. Xu, "A large-scale empirical analysis of chinese web passwords," in *USENIX Security Symposium*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14461062

[32] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang, "The tangled web of password reuse," in *Proceedings 2014 Network and Distributed System Security Symposium*. San Diego, CA: Internet Society, 2014. [Online]. Available: https://www.ndss-symposium.org/ndss2014/programme/tangled-web-password-reuse/

[33] B. Ur, S. M. Segreti, L. Bauer, N. Christin, L. F. Cranor, S. Komanduri, D. Kurilova, M. L. Mazurek, W. Melicher, and R. Shay, "Measuring real-world accuracies and biases in modeling password guessability," 2015, p. 463–481, https://www.usenix.org/system/files/conference/usenixsecurity15/sec15-paper-ur.pdf. [Online]. Available: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/ur

[34] Y. Li, H. Wang, and K. Sun, "A study of personal information in human-chosen passwords and its security implications," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, Apr. 2016, p. 1–9, https://ieeexplore.ieee.org/document/7524583. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7524583

[35] S. Pearman, J. Thomas, P. E. Naeini, H. Habib, L. Bauer, N. Christin, L. F. Cranor, S. Egelman, and A. Forget, "Let's go in for a closer look: Observing passwords in their natural habitat," *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, p. 295–310, Oct. 2017, https://dl.acm.org/doi/10.1145/3133956.3133973. [Online]. Available: https://dl.acm.org/doi/10.1145/3133956.3133973

[36] D. Pasquini, M. Cianfriglia, G. Ateniese, and M. Bernaschi, "Reducing bias in modeling real-world password strength via deep learning and dynamic dictionaries," 2021, p. 821–838, https://www.usenix.org/conference/usenixsecurity21/presentation/pasquini. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/pasquini

[37] J. A. Cazier and B. D. Medlin, "Password security: An empirical investigation into e-commerce passwords and their crack times," *Information Systems Security*, vol. 15, no. 6, p. 45–55, Dec. 2006.

[38] D. Wang, X. Shan, Q. Dong, Y. Shen, and C. Jia, "No single silver bullet: Measuring the accuracy of password strength meters," 2023, p. 947–964. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/wang-ding-silver-bullet

[39] Y. Li, H. Wang, and K. Sun, "A study of personal information in human-chosen passwords and its security implications," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, Apr. 2016, p. 1–9. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7524583

[40] B. L. Riddle, M. S. Miron, and J. A. Semo, "Passwords in use in a university timesharing environment," *Computers & Security*, vol. 8, no. 7, p. 569–579, Nov. 1989.

Table III: The table presents the top 25 labelled password structures from the Lizard-Squad dataset. Hashcat [21] benchmarks indicate the performance of consumer hardware in a hybrid attack. The dictionary size for $\boldsymbol{w}$ (words) and $\boldsymbol{n}$ (numbers) is set at 5×10^5. `zxcvbn` [11] serves as a password strength meter, with scores ranging from 0 to 4, and guess estimates are log based.

	Frequency		Hashcat using Nvidia RTX 4090				`zxcvbn` Mean	
Structure *log* **perm'**	**Count**	**%**	**MD5**	**PBKDF2-sha256**	**scrypt**	**bcrypt-sha512**	**Score**	**Guesses**
w	594	5.5	~ 0	0.1 s	1 min	4 min	0.9	4.24
n	189	1.7	~ 0	0.1 s	1 min	4 min	0.8	4.13
wd	254	2.3	~ 0	0.6 s	12 min	42 min	1.2	5.06
nd	135	1.2	~ 0	0.6 s	12 min	42 min	1.5	5.71
wdd	667	6.1	~ 0	5.6 s	2 h	7 h	1.4	5.64
ndd	457	4.2	~ 0	5.6 s	2 h	7 h	1.4	5.85
wddd	736	6.8	~ 0	56.4 s	19 h	3 day	1.6	6.10
nddd	409	3.8	~ 0	56.4 s	19 h	3 day	1.6	6.04
wdddd	465	4.3	~ 0	9 min	8 day	29 day	1.7	6.41
ndddd	359	3.3	~ 0	9 min	8 day	29 day	1.7	6.18
nddddd	58	0.5	0.3 s	2 h	81 day	293 day	1.8	6.51
wddddd	84	0.8	0.3 s	2 h	81 day	293 day	2.0	6.89
nn	139	1.3	1.5 s	8 h	1 yr	4 yr	1.6	6.03
nw	26	0.2	1.5 s	8 h	1 yr	4 yr	1.8	6.42
ww	666	6.1	1.5 s	8 h	1 yr	4 yr	1.5	5.84
ndddddd	51	0.5	3.0 s	16 h	2 yr	8 yr	2.3	7.24
wddddd	52	0.5	3.0 s	16 h	2 yr	8 yr	1.9	6.92
nnd	73	0.7	15.2 s	3 day	11 yr	40 yr	2.4	7.82
wwd	231	2.1	15.2 s	3 day	11 yr	40 yr	2.1	7.18
wwdd	307	2.8	3 min	33 day	111 yr	402 yr	2.4	7.61
nndd	86	0.8	3 min	33 day	111 yr	402 yr	2.8	8.64
wwddd	233	2.1	25 min	326 day	1.11×10^3 yr	4.02×10^3 yr	2.5	7.93
nnddd	48	0.4	25 min	326 day	1.11×10^3 yr	4.02×10^3 yr	2.9	8.96
nndddd	28	0.3	4 h	9 yr	1.11×10^4 yr	4.02×10^4 yr	3.1	9.54
wwdddd	76	0.7	4 h	9 yr	1.11×10^4 yr	4.02×10^4 yr	2.8	8.39
www	210	1.9	9 day	447 yr	5.56×10^5 yr	2.01×10^6 yr	2.3	7.33
wwwd	42	0.4	88 day	4.47×10^3 yr	5.56×10^6 yr	2.01×10^7 yr	2.3	7.45
wwwdd	42	0.4	2 yr	4.47×10^4 yr	5.56×10^7 yr	2.01×10^8 yr	3.1	9.43
wwwddd	39	0.4	24 yr	4.47×10^5 yr	5.56×10^8 yr	2.01×10^9 yr	2.9	9.02
wwww	54	0.5	1.21×10^4 yr	2.24×10^8 yr	2.78×10^{11} yr	1.00×10^{12} yr	3.1	9.77