

SitePassword Modulo Bias

Alan H. Karp

July 10, 2024

Abstract

SitePassword hashes a super password, a per site nickname, and a per site userid to produce a site password. The bytes produced by the hash need to be converted to a string of characters selected from the alphabet supported by the web site associated with that user account. In some situations, *modulo bias* reduces the entropy of the site password. This note shows that the effect is modest for the range of values SitePassword supports.

1 Introduction

The original SitePassword algorithm used each byte of the hash produced by PBKDF2¹ to compute the i -th character of the site password using $c_i = A[b_i \bmod |A|]$, where A is an array of characters representing the alphabet, b_i is the byte value, and $|A|$ is the size of the alphabet. If the size of the alphabet evenly divides $2^8 = 256$, every character in the alphabet has the same probability of being selected. If it doesn't, then some characters are more likely to be selected. This *modulo bias* reduces the entropy of the resulting string.

This document shows that modulo bias has only a modest effect on the guessability of the site password it generates.

2 Recommended Approaches

The recommended approach to eliminating modulo bias is *rejection sampling*. Simply ignore any byte value index outside the range of the alphabet. In other words, don't use any $b_i > |A|$. Theoretically, this approach could require an unbounded number of bytes, but the number of bytes needed is quite modest in practice. The recommended approach to mitigating modulo bias is to use a larger range of random numbers, say two bytes instead of one. There is still bias, but the effect on guessability is substantially smaller.

Unfortunately, usability considerations require that SitePassword always succeed in computing a password, which means it must be conservative in how many bytes

¹Password Based Key Derivation Function

it computes with rejection sampling. While it's possible to estimate how many extra bytes are needed to achieve a given probability of success, there are no guarantees. A deterministic algorithm, such as one with modulo bias, meets the requirements.

A second problem with rejection sampling is that it makes offline guessing of the super password easier. The runtime of PBKDF2 depends on the number of iterations and the key size. SitePassword might choose to compute 10 or 20 times as many bytes as are typically needed. An adversary might choose to compute only 2 times as many and give up on the very small percentage that need more. This 5-10 times advantage in guess rate is significant.

One workaround would be for SitePassword to first compute a key with a modest number of bytes, say 2 times the size of the site password. If more bytes are needed, SitePassword could redo the calculation with 4 times as many, *etc.* Unfortunately, the fixed overhead of each run and the wasted computation make it hard to meet the latency requirement.

3 Quantifying Modulo Bias

There will be no need to use extra bytes to construct the site password if modulo bias doesn't weaken the site password very much. One way to quantify that weakening is by looking at the entropy of the site password both with and without modulo bias. Without bias, the entropy of an L -character site password is

$$E_0 = L \log_2 |A| \quad (1)$$

The easiest way to understand modulo bias is to repeat the alphabet until the result is has more elements than the largest possible random value R and truncate to size R . If $|A|$ doesn't evenly divide R , some characters will appear one more time than others. In those cases, the number of times a character appears is either

$$n_1 = \lfloor R/|A| \rfloor \quad \text{or} \quad n_2 = \lceil R/|A| \rceil, \quad (2)$$

so that $n_1 = n_2$ if $|A|$ divides R or $n_2 = n_1 + 1$ otherwise. The probabilities that one of those characters is chosen is,

$$p_1 = n_1/R \quad \text{or} \quad p_2 = n_2/R. \quad (3)$$

The entropy of selecting a single character from each subset is then

$$E_1 = -(|A| - R \bmod |A|)p_1 \log_2 p_1 \quad \text{or} \quad E_2 = -(R \bmod |A|)p_2 \log_2 p_2. \quad (4)$$

Here $R \bmod |A|$ is the number of characters that appear an extra time. Note that if $|A|$ divides R , $E_2 = 0$ and $E_1 = E_0$. The entropy of an L -character site password with modulo bias is $E_b = L(E_1 + E_2)$.

We can now plug in some numbers. SitePassword uses $R = 256$ and allows alphabets in the range $10 \leq |A| \leq 96$. With the default values of $|A| = 62$ and $L = 12$, $E_0 - E_b = .05$ which requires 96% of the 2^{71} guesses needed if there was no modulo bias. The worst case in the allowed range is at $|A| = 96$ where $E_0 - E_b = 0.28$,

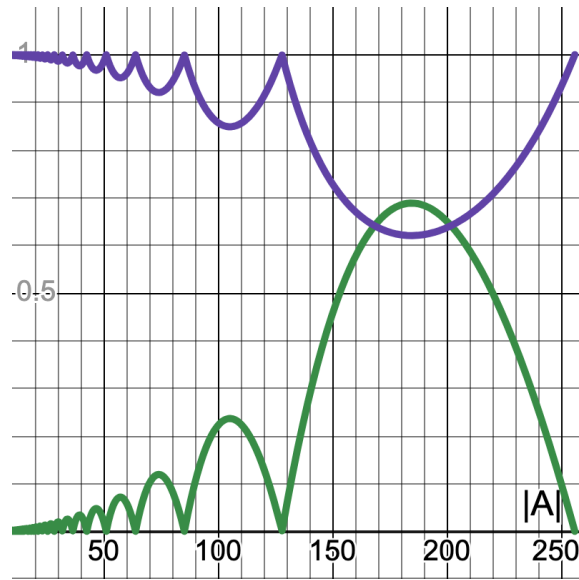


Figure 1: Modulo bias for SitePassword for an L -character site password. The green line is the entropy loss; the purple line the reduction in the number of guesses. Only values of $10 \leq |A| \leq 96$ are relevant to SitePassword.

which translates into 82% as many guesses. The greater loss in guesses is offset by the larger alphabet size leaving the entropy with modulo bias greater than 78 bits. Figure 1 shows the entropy loss, $E_0 - E_b$, in green and the ratio of the number of guesses needed, $2^{-(E_0 - E_b)}$, in purple for $R = 256$ and $L = 12$.

4 Conclusions

There are cases where a small modulo bias can be a serious problem, such as in choosing a nonce for a stream cypher. That is not the case for SitePassword. Even the most extreme modulo bias reduces the entropy of the site password by less than one bit. Even so, the SitePassword algorithm was changed to use 2 bytes per character, which makes the modulo bias considerably smaller.