

Algebra of Cryptographic Collisions

Azka Ghaffar*, Julian Thiele-Pissini[†], Christopher Briscoe[‡],
Albert Carlson[§], Sai Ranganath Mikkilineni[¶], and Mandeep Singh^{||}.

*^{†||}Independent Researcher

[‡]Department Chair of Mathematics and Physics, Emet Classical Academy, New York, NY, USA

[§]Chair for Entropy and Encryption, Quantum Security Alliance, Kyle, TX, USA

[¶]College of Business, Delaware State University, Dover, DE, USA

Email: *azkaghaffar11@gmail.com, [†]jctp277@gmail.com, [‡]Cbriscoe@emetclassicalacademy.org,

[§]acarlson2@nu.edu, [¶]SMikkilineni@desu.edu, ^{||}mandeep.singh@gmail.com

Abstract—Building on the collision-based cryptanalysis concepts introduced in a recent study, this work provides a more detailed mathematical treatment by extending the analysis of the Birthday Problem beyond the idealized uniform case. We examine anticipated collisions under a general probability distribution, deriving expressions for the expected collision count and its variance as functions of the sample size, analogous to the number of people in a room in the classical Birthday Problem. By including the lowest-order corrections, we can make practical predictions about the underlying probability mass function, thereby refining our ability to characterize encryption outputs and infer the structural properties of the plaintext or encryption scheme. This deeper probabilistic framework strengthens the theoretical foundation for identifying ciphertext patterns and supports practical applications in cryptography, stylometry, and document authentication.

Keywords—Collision Curves, Y-intercept, Cryptanalysis, Cryptography, Linguistic Analysis, Probability Mass Function, Birthday Paradox

I. INTRODUCTION

Recent studies [1] have explored the slope of collision curves when examining increasing volumes of encrypted text, with a particular focus on an inevitable saturation of meta-characters given enough data and classification of meta-languages based on this slope, or equivalently, the y-intercept via the Legendre transform, of the curve. Feistel’s [2] assertion that all ciphers are fundamentally substitution ciphers implies that encryption alone does not remove plaintext patterns. While randomization can obscure these patterns, it may be imperfect or introduce new vulnerabilities due to its pseudo-random nature. Patterns in ciphertext enable attackers to utilize language statistics to estimate the plaintext probability mass function (PMF) and associate it with an alphabet, thereby aiding their cryptanalysis.

Since Kerckhoffs [3], cryptographers have assumed a “worst case” scenario in which an attacker knows the encryption algorithm, plaintext language, randomization methods, and block size. While such complete knowledge is theoretical, identifying the PMF at equilibrium enables comparison with known languages over various block sizes. Block treatment forms a “meta-language” of meta-characters [4], [5], and matching PMFs reveals possible block sizes. At integer multi-

ples of the block size, the ciphertext PMF matches the plaintext if randomization is absent or broken.

The Law of Large Numbers [6] ensures accuracy over time, but many languages can be ruled out early in the cryptanalysis. Each cipher or plain text block sample comes from a constituent set at least as large as the unique sample set. The first appearance of an element confirms its existence, and its subsequent repeated appearances are referred to as “collisions.” Counting collisions quickly yields information about the underlying plaintext and its redundancy [4], thereby reducing language matching to a single-metric comparison. The Birthday Problem illustrates the probability of a collision in a sample of size N drawn from a uniform random variable with constituent set size k :

$$p(N, k) = 1 - \frac{k!}{k^N (k - N)!}. \quad (1)$$

The expected number of collisions, rather than a single experimental outcome, is the focus, accounting for variance in the observed data. Here, we focus on expected collision counts and variance for arbitrary PMFs. Using indicator variables, formulas for the expected number of collisions and variance are obtained for both uniform and non-uniform PMFs as functions of N and k . Because language is non-uniform [7]–[9], collisions occur sooner than in uniform cases. We acknowledge that the non-uniform PMF Birthday Problem has been examined in prior work [10]–[14], and our work contributes to the expansion of the existing literature.

The method is computationally simple, and “language” is generalized to include dialects, author-specific styles [15], or classified document types, treating any finite set with a defined PMF as a unique language for analysis. For large N , calculations become challenging. This study restructures and approximates the expected value and variance formulas to enable simplified analysis in terms of N , k , and PMF moments [16]. Substitution ciphers do not alter k or the PMF moments, only relabeling elements of k [4], permitting analysis of plaintext and ciphertext on equal footing.

In this study, we examine earlier phases of the collision curves through careful stochastic analysis, aiming to classify meta-languages more efficiently and effectively. The results serve as a theoretical framework for incremental improvements

over previous results [1], reducing the amount of data required to reveal the context and content of the ciphertext. Additional cryptographic techniques, such as cipher-block chaining (CBC) [17], are not considered here but will be addressed in future work. Similarly, real languages involve conditional probabilities which may be incorporated recursively or perturbatively in future studies [18] [19] [20].

II. ANALYSIS

We define the anticipated number of collisions, A , to be a function of both the expected number of collisions, E , and its associated variance, V . This paper presents our analysis of this function through four theorems, including exact and approximate equations for the expectation values and variances of uniform and general distributions. The approximations require the ratio of sample size to constituent set size, $N/k \ll 1$, and break down if the PMF is sharply peaked over a small number of elements, effectively reducing k and increasing the ratio. Many cases also involve $N \gg 1$, as collisions are rare in real meta-languages and a large sample set is often needed to encounter the first collision. The first result gives an exact equation for the expected number of collisions, including the variance, for a uniform distribution.

Theorem 1. *The anticipated number of collisions, $A(N, k)$, including variance around the expected value, between elements of a sample set with fixed sample set size, N , drawn from a uniformly distributed, independent random variable with constituent set size, k , is given by*

$$A(N, k) = E(N, k) \pm \sqrt{V(N, k)} \quad (2)$$

where $E(N, k)$ is given by

$$E(N, k) = \sum_{\alpha=2}^N \frac{(-1)^\alpha}{k^{\alpha-1}} \binom{N}{\alpha} \quad (3)$$

and $V(N, k)$ is given by

$$V(N, k) = \sum_{\alpha=2}^N C(N, \alpha) \left(\frac{1}{k^{\alpha-1}} \right) \quad (4)$$

with coefficients $C(N, \alpha)$ given by

$$C(N, \alpha) = (-1)^\alpha (1 - 2^\alpha) \binom{N}{\alpha} - (-1)^\alpha \left(2^{\alpha+1} \binom{N}{\alpha+1} - \binom{2N}{\alpha+1} \right). \quad (5)$$

Our second result applies Stirling's Approximation to Theorem 1, yielding an approximate expression for the anticipated

number of collisions.

Theorem 2. *The anticipated number of collisions, $A(N, k)$, including variance around the expected value, between elements of a sample set with fixed sample set size, N , drawn from a uniformly distributed, independent random variable with constituent set size, k , can be approximated by*

$$A(N, k) = E(N, k) \pm \sqrt{V(N, k)} \quad (6)$$

where $E(N, k)$ is given by

$$E(N, k) \approx \sum_{\alpha=2}^{\phi} \frac{(-1)^\alpha}{k^{\alpha-1}} \frac{N^\alpha}{\alpha!} \quad (7)$$

and $V(N, k)$ is given by

$$V(N, k) \approx \sum_{\alpha=2}^{\phi} \frac{(-1)^\alpha}{k^{\alpha-1} \alpha!} N^\alpha (1 + \alpha 2^{\alpha-1} - 2^\alpha), \quad (8)$$

where ϕ is the chosen upper bound of the summation, that is, the chosen order of the approximation.

Displayed to order $\phi = 3$ and valid for $k \gg N \gg 1$ we have:

$$A(N, k) \approx \frac{N^2}{2k} - \frac{N^3}{6k^2} \pm \sqrt{\frac{N^2}{2k} - \frac{5N^3}{6k^2}}. \quad (9)$$

Although the uniform distribution approximation is helpful in some cases, it is a special case of the more general non-uniform PMF. Our third and fourth results provide exact and approximate expressions, respectively, for the anticipated number of collisions for a non-uniform PMF.

Theorem 3. *The anticipated number of collisions, $A(N, k)$, including variance around the expected value, between elements of a sample set with fixed sample set size, N , drawn from a uniformly distributed, independent random variable with constituent set size, k , with a general distribution with PMF for the possible outcomes $\vec{p} = [p_1, p_2, \dots, p_{k-1}, p_k]$, is given by*

$$A(N, \vec{p}) = E(N, \vec{p}) \pm \sqrt{V(N, \vec{p})} \quad (10)$$

where $E(N, \vec{p})$ is given by

$$E(N, \vec{p}) = \sum_{\alpha=2}^N (-1)^\alpha \binom{N}{\alpha} \langle p^{\alpha-1} \rangle. \quad (11)$$

and $V(N, \vec{p})$ is given by

$$\begin{aligned} V(N, \vec{p}) = & \sum_{\alpha=2}^{N-1} (-1)^\alpha \binom{N}{\alpha} (1 - 2^\alpha + 2\alpha) \langle p^{\alpha-1} \rangle \\ & + \sum_{\alpha=3}^N \left(\frac{1 - (-1)^\alpha}{2} \right) \left(\binom{N}{\alpha+1} \binom{\alpha+1}{\frac{\alpha+1}{2}} - \binom{N}{\frac{\alpha+1}{2}}^2 \right) \langle p^{\frac{\alpha-1}{2}} \rangle^2 \\ & + 2 \sum_{\alpha=4}^N \sum_{\beta=2}^T (-1)^\alpha \left(\binom{N}{\beta} \binom{N}{\alpha+1-\beta} - \binom{N}{\alpha-1} \binom{\alpha-1}{\beta} \right) \langle p^{\beta-1} \rangle \langle p^{\alpha-\beta} \rangle \\ & + (-1)^N (2N + 1 - 2^N) \langle p^{N-1} \rangle - \sum_{b=R+1}^N \binom{N}{b}^2 \langle p^{b-1} \rangle^2 \\ & - 2 \sum_{d=2}^{N-1} (-1)^d \binom{N}{d} \sum_{f=N-d+1}^N (-1)^f \binom{N}{f} \langle p^{d-1} \rangle \langle p^{f-1} \rangle, \end{aligned} \quad (12)$$

where

$$\begin{aligned} T &= \begin{cases} \frac{c-1}{2} & c \text{ is odd} \\ \frac{c}{2} - 1 & c \text{ is even,} \end{cases} \text{ and} \\ R &= \begin{cases} \frac{N}{2} & N \text{ is even} \\ \frac{N+1}{2} & N \text{ is odd.} \end{cases} \end{aligned} \quad (13)$$

Applying Stirling's approximation to the previous result yields an approximate expression for the anticipated number of collisions, which provides our final main result.

Theorem 4. *The anticipated number of collisions, $A(N, k)$, including variance around the expected value, between elements of a sample set with fixed sample set size, N , drawn from a uniformly distributed, independent random variable with constituent set size, k , with a general distribution with PMF for the possible outcomes $\vec{p} = [p_1, p_2, \dots, p_{k-1}, p_k]$, can be approximated by*

$$A(N, \vec{p}) = E(N, \vec{p}) \pm \sqrt{V(N, \vec{p})} \quad (14)$$

where $E(N, \vec{p})$ is given by

$$E(N, \vec{p}) \approx \sum_{\alpha=2}^{\phi} (-1)^{\alpha} \frac{N^{\alpha}}{\alpha!} \langle p^{\alpha-1} \rangle \quad (15)$$

and $V(N, \vec{p})$ is given by

$$\begin{aligned} V(N, \vec{p}) &\approx \sum_{\alpha=2}^{\phi} (-1)^{\alpha} \frac{N^{\alpha}}{\alpha!} (1 - 2^{\alpha} + 2\alpha) \langle p^{\alpha-1} \rangle \\ &+ \sum_{\alpha=3}^{\phi} \left(\frac{(-1)^{\alpha} - 1}{2} \right) \frac{N^{\alpha}}{((\frac{\alpha-1}{2})!)^2} \langle p^{\frac{\alpha-1}{2}} \rangle^2 \\ &+ 2 \sum_{\alpha=4}^{\phi} \sum_{\beta=2}^T (-1)^{\alpha} \frac{N^{\alpha}}{(\alpha - \beta)!(\beta - 1)!} \langle p^{\beta-1} \rangle \langle p^{\alpha-\beta} \rangle, \end{aligned} \quad (16)$$

with T as defined in Theorem 3 and where ϕ is the chosen upper bound of the summation, that is, the chosen order of the approximation.

The equations for a general distribution reduce to the equations for the uniform distribution in the limit of equal probabilities.

Displayed to order $\phi = 3$ and valid for $\frac{1}{\langle p \rangle} \gg N \gg 1$ we have:

$$\begin{aligned} A(N, \vec{p}) &\approx \frac{N^2}{2} \langle p \rangle - \frac{N^3}{6} \langle p^2 \rangle \\ &\pm \sqrt{\frac{N^2}{2} \langle p \rangle + \frac{N^3(\langle p^2 \rangle - 6 \langle p \rangle^2)}{6}}. \end{aligned} \quad (17)$$

The results display Poisson behavior in the early phases of analyzing collisions, where expected value and variance are equal:

$$E(N, \vec{p}) \approx \frac{N^2}{2} \langle p \rangle \approx V(N, \vec{p}) \quad (18)$$

This is not unexpected, as collisions are typically rare events, for most meta-languages, well in advance of the steady state analysis of [1]. Arbitrary accuracy may, of course, be achieved by calculating to higher orders of ϕ as N gets larger or as the problem demands greater precision.

III. IMPLICATIONS

It is crucial to examine how shifting from a uniform PMF to a general one affects the expected number of collisions. Many cryptographic studies assume approximate uniformity to apply standard results from the Birthday Problem. Consider an exponentially distributed random variable with probabilities $p_i \propto e^{-\nu i}$ for $i = 0, \dots, k-1$. For large k , the continuous limit can be applied ($k \rightarrow \infty, PMF \rightarrow PDF$). The exponential distribution then reflects a state of maximal entropy with expected index $\langle i \rangle = 1/\nu$:

$$p_i = \nu e^{-\nu i}, \quad \langle p^n \rangle^m = \frac{\nu^{nm}}{(n+1)^m}. \quad (19)$$

This form approximates the single-letter PMF of English [4] and can extend to larger block sizes. While single-letter substitution ciphers do not require collision analysis, larger block sizes do. For instance, let $\nu = 2^9/k$, so that $i = k/2^8$ corresponds to a probability drop by a factor $e^{-2} \approx 0.135$. Then $\langle p \rangle = \nu/2 = 2^8/k$. Using the lowest-order estimate ($\alpha = 2$) from Theorem 4, the expected number of collisions is

$$\frac{N^2 \langle p \rangle}{2} = \frac{N^2 \nu}{4} = 2^8 \frac{N^2}{2k}, \quad (20)$$

compared to $\frac{N^2}{2k}$ for a uniform distribution, and solving for N to reach one collision gives

$$N = \sqrt{\frac{4}{\nu}} = \frac{\sqrt{2k}}{2^4}, \quad (21)$$

16 times smaller than the uniform case. For $k = 2^{31}$, the first collision occurs at $N = 2^{12} = 4096$ instead of $2^{16} = 65536$, with $2^8 = 256$ expected collisions at $N = 2^{16}$.

These parameters illustrate the effect of a non-uniform PMF; other values of ν or PMF forms, such as a power-law [21], may also apply. A complete functional form is unnecessary, as knowledge of element probabilities suffices to compute any moments [16] and perform collision analysis. Deviating from uniformity increases the expected number of collisions, reflecting the reduced entropy of the underlying language or system.

For completeness, assuming an exponential distribution with arbitrary ν , the anticipated number of collisions is

$$A(N, \vec{p}) \approx \frac{N^2 \nu}{4} - \frac{N^3 \nu^2}{18} \pm \sqrt{\frac{N^2 \nu}{4} - \frac{7N^3 \nu^2}{36}}. \quad (22)$$

IV. DERIVATIONS

The proof of the equations for the expected number of collisions follows Siegrist's method [22], summarized here.

A similar approach is discussed in the context of the formulas for variance. However, a more detailed set of proofs will be provided in a subsequent paper due to the significant level of complexity.

Let Y_i count the number of times that a particular meta-character appears in the text, with i ranging from 1 to k . If the i^{th} meta-character has not appeared, $Y_i = 0$. Let X_i be an indicator variable such that $X_i = 1 \times (Y_i = 0)$. Thus, X_i will equal 1 if $Y_i = 0$, and will equal 0 if $Y_i > 0$. This variable will therefore indicate whether or not a meta-character has appeared in the text, disregarding repetition. Each X_i is considered statistically independent from all other X_i 's in this analysis.

Each meta-character has a unique probability, p_i , such that the sum over all $p_i = 1$. We determine the expected value of X_i to be

$$E(X_i) = 1 * (P(X_i) = 1) + 0 * (P(X_i) = 0) = (1 - p_i)^N. \quad (23)$$

Now define a variable U that counts all of the meta-characters that have not been encountered, thus,

$$U = \sum_{i=1}^k X_i. \quad (24)$$

Hence, $E(U)$ will be the sum of each $E(X_i)$ as defined above. Next, we define V to be the number of meta-characters that have been encountered, $V = k - U$, and therefore, the expected number of characters we have encountered is

$$E(V) = k - \sum_{i=1}^k (1 - p_i)^N. \quad (25)$$

The definition of a collision is the occurrence of encountering a meta-character more than once. We define a variable C to count collisions of meta-characters as a function of N intercepted meta-characters. Then $C = N - V = N - k + U$ since we must subtract the first time we encounter a given meta-character, due to the initial encounter not being a collision by definition. Therefore, $E(C) = N - k + E(U)$ and thus

$$E(C) = E(N, \vec{p}) = N - k + \sum_{i=1}^k (1 - p_i)^N. \quad (26)$$

Applying the binomial expansion, we have:

$$E(N, k, \vec{p}) = N - k + \sum_{i=1}^k \sum_{\alpha=0}^N (-1)^\alpha (p_i)^\alpha \binom{N}{\alpha} \quad (27)$$

which reduces to

$$E(N, k, \vec{p}) = \sum_{i=1}^k \sum_{\alpha=2}^N (-1)^\alpha (p_i)^\alpha \binom{N}{\alpha}. \quad (28)$$

We use the following notation for the various moments of

the PMF:

$$\langle p^{\alpha-1} \rangle = \sum_{i=1}^k (p_i)^\alpha, \quad (29)$$

which allows us to rewrite Eq. 28 as

$$E(N, \vec{p}) = \sum_{\alpha=2}^N (-1)^\alpha \langle p^{\alpha-1} \rangle \binom{N}{\alpha}. \quad (30)$$

In the limit of large N , we asymptotically approach a collision curve of slope unity, which is clear when examining Eq. 26 and makes intuitive sense, as we will have encountered (nearly) all meta-characters given a large enough sample set size, clearly verifying the efficacy of the methods applied in [1].

Non-uniform probability distributions can take arbitrary forms. Probability moments are maximized when one of the k elements has probability 1 (others 0), minimizing entropy, and minimized for the uniform PMF, which maximizes entropy. Realistic languages are usually closer to uniform, with moments of order γ , $\langle p^\gamma \rangle \propto \frac{1}{k^\gamma}$. For small k , direct use of the original equations is simpler, but for $k \gg N \gg 1$, Theorem 3 is more efficient for approximating expectation and variance. Considering only the first three sums of the variance in Theorem 3 and using the uniform PMF as a rough moment approximation, each term will be proportional to:

$$\frac{N^\alpha}{\alpha! k^{\alpha-1}} = \frac{k}{\alpha!} \left(\frac{N}{k} \right)^\alpha \quad (31)$$

For $k = 10,000$ and $N = 1000$, terms decrease by roughly a factor of 10α , so accurate approximations follow from just the first few α .

If k is large but most probabilities are near zero, these approximations fail, and almost all blocks are collisions. Theorem 3 still holds, but all sums must be computed. In the extreme where all but one meta-character has zero probability, k is effectively reduced, entropy is minimized, and the message becomes trivial, similar to the uniform PMF case.

For many relevant scenarios, the expected number of collisions may therefore be calculated with arbitrary accuracy. To the lowest order, the expected number of collisions is then:

$$E(N, \vec{p}) \approx \frac{N(N-1)}{2} \langle p \rangle \approx \frac{N^2}{2} \langle p \rangle. \quad (32)$$

In the case of uniform distribution, $p_i = 1/k$, but this occurs only in the instance of maximal entropy and does not represent an actual meta-language [9]. However, we can easily determine the expected number of collisions in this situation.

$$E(N, k) = \sum_{\alpha=2}^N (-1)^\alpha \left(\frac{1}{k^{\alpha-1}} \right) \binom{N}{\alpha}. \quad (33)$$

The expected number of collisions can be estimated with arbitrary precision simply by expanding in increasing orders of k^α . To lowest order ($\alpha = 2$) we have:

$$E(N, k) \approx \frac{N(N-1)}{2k} \approx \frac{N^2}{2k}, \quad (34)$$

where the second approximation holds for $N^2 \gg N$.

Using similar methods, the variance in the expected number of collisions can be computed. We define Y_i , X_i , p_i , N , V , and U as done above. Recall that the number of collisions, C , will be equal to $N - k + U$. Thus, since C and U differ only by constants, we note that $\text{var}(C) = \text{var}(U) = E(U^2) - E(U)^2$.

Since $U = \sum_{i=1}^k X_i$, we have that

$$E(U) = \sum_{i=1}^k (1 - p_i)^k. \quad (35)$$

Next, we calculate:

$$\begin{aligned} E(U^2) &= E\left(\left(\sum_{i=1}^k X_i\right)^2\right) \\ &= E\left(\sum_{i=1}^k X_i \sum_{j=1}^k X_j\right) \\ &= E\left(\sum_{i=1}^k \sum_{j \neq i} X_i X_j + \sum_{i=1}^k X_i^2\right) \\ &= \sum_{i=1}^k \sum_{j \neq i} E(X_i X_j) + \sum_{i=1}^k E(X_i^2). \end{aligned} \quad (36)$$

We note that X_i can only take on the value of zero or one, and therefore $E(X_i^2) = E(X_i) = (1 - p_i)^k$. We also recognize that $E(X_i X_j)$ will be non-zero only when both X_i and X_j are equal to one. Thus, since X_i and X_j are taken to be statistically independent in this analysis, the probability that we have not encountered the i^{th} or j^{th} meta-character is $(1 - (p_i + p_j))^N$. As such, we have,

$$E(U^2) = \sum_{i=1}^k \sum_{j \neq i} (1 - (p_i + p_j))^N + \sum_{i=1}^k (1 - p_i)^N. \quad (37)$$

Combining our results from Equations 35 and 37, we have found that

$$\begin{aligned} V(N, k, \vec{p}) &= \text{var}(U) = \sum_{i=1}^k \sum_{j \neq i} (1 - (p_i + p_j))^N \\ &\quad + \sum_{i=1}^k (1 - p_i)^N - \left(\sum_{i=1}^k (1 - p_i)^k\right)^2. \end{aligned} \quad (38)$$

We can easily obtain a similar result for the variance with a uniform distribution.

$$\begin{aligned} V(N, k) &= k(k-1) \left(1 - \frac{2}{k}\right)^N + k \left(1 - \frac{1}{k}\right)^N \\ &\quad - k^2 \left(1 - \frac{1}{k}\right)^{2N}. \end{aligned} \quad (39)$$

V. CONCLUSIONS AND FUTURE WORK

This study has thoroughly investigated an approach to ascertaining information about an underlying meta-language simply by counting collisions and studying their mean and variance. The theoretical framework presented enables incremental improvements to available tools [1] to eliminate possible languages and encryption schemes, thus reducing the amount of ciphertext required to reveal the underlying structure of the plaintext [4]. A Poisson distribution governs the early phases of the collision curves, and a mathematical framework is developed to provide additional terms for increasing accuracy. A linear relationship is achieved when collisions are saturated, and the analysis enters a steady state. Subsequent work will explore simulations leveraging these tools to test the efficacy and limitations of these methods, as well as provide a more formal proof of the theorems presented.

Applications include exploiting poorly distributed random variables [5] in cryptographic attacks; for example, if k denotes the size of a hash set and the underlying distribution is exponential (Eq. 19) rather than uniform, collisions may increase, aiding malicious actions. Another application is to infer the underlying block size and PMF from collision counts. For encrypted files with unknown language, collision data can help identify or exclude possible PMFs [23], revealing characteristics of the data (e.g., technical or military). Larger block sizes often render many meta-language elements negligible [4], [5], allowing the use of an adequate set size, k_{eff} , to speed calculations. However, since expected collisions and variance depend on probability moments, negligible probabilities must remain insignificant relative to these moments, and N/k must remain small for valid approximations.

Even for unencrypted data, collision counting can aid document authentication via stylometry, as authors' unique characteristics imply distinct PMFs [15], thereby assisting in verifying authenticity or detecting fraud [24]. Collision counting also suggests a phase transition: from out-of-equilibrium (collisions are rare and Poisson) to equilibrium (collisions saturate and PMF stabilizes), with N parameterizing time. This approach may inform statistical mechanics, for example, when particles equilibrate in a heat bath. Collision counting provides insight into steady-state probability distributions, which then further define entropy. Future work will explore this relationship to bridge the gap between information theory and physical systems, utilizing collision counts and entropy.

REFERENCES

- [1] M. Singh, A. Carlson and C. Briscoe, *Using the y-intercept in Collision Attacks for Encryption*, 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2025, pp. 00999-01006.
- [2] H. Feistel, *Cryptography and Computer Privacy*, Scientific American, 1973, 228, 15 - 20
- [3] A. Kerckhoffs, *La Cryptographie Militaire*, Journal des sciences militaires, 1883, IX, 5 - 83, 161 - 191
- [4] C. Shannon, *Communication Theory of Secrecy Systems*, Bell System Technical Journal, 28:656 - 715, 1949.
- [5] A. Carlson, *Set Theoretic Estimation Applied to the Information Content of Ciphers and Decryption*, PhD thesis, University of Idaho, 2012.

- [6] S. Ross, *A First Course in Probability*, MacMillan Publishing, Inc, New York, 1976.
- [7] D. T. Langendoen and P. Postal, *The Vastness of Natural Languages*, The Camelot Press, Ltd., Southampton, 1984
- [8] M.Lucks, *A Constraint Satisfaction Algorithm for the Automated Decryption of Simple Substitution Ciphers*, CRYPTO, 1988.
- [9] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc, New York, 2nd edition, 2005.
- [10] M. Dwass, *More Birthday Surprises*, J. Combinatorial Theory 7 (1969), 258–261
- [11] J. Klotz, *The Birthday Problem with Unequal Probabilities*, Department of Statistics, University of Wisconsin, Tech. Report, No. 59, 1979
- [12] G. C. Berresford, *The Uniformity Assumption in the Birthday Problem*, Math. Mag. 53 (1980), no. 5, 286–288.
- [13] S. Mase, *Approximations to the Birthday Problem with Unequal Occurrence Probabilities and their Application to the Aurname Problem in Japan*, Ann Inst Stat Math 44, 1992, 479–499
- [14] P. Rathie and P. Zörnig, *On the Birthday Problem: Some Generalizations and Applications*, International Journal of Mathematics and Mathematical Sciences, 2003 (60)
- [15] A. Morton, *Literary Detection*, Scribners, New York, 1978.
- [16] Moment. *Encyclopedia of Mathematics*. <http://encyclopediaofmath.org/index.php?title=Moment&oldid=55270>.
- [17] B. Schneier, *Applied Cryptography: Protocols, Algorithms, and Source Code in C*, John Wiley and Sons Inc., New York, 2nd edition, 1996.
- [18] Y. S. Han, Y. C. Park, and K. Choi *Recursive Markov Chains as a Stochastic Grammar*, AAAI Technical Report FS-92-04, 1992
- [19] K. Motegi and S. Woo, *A Note on the Exponentiation Approximation of the Birthday Paradox*, Communications in Statistics - Theory and Methods, 53(18), 6417–6426, 2024
- [20] J.E. Connett, *When are there too many collisions? Variants of the Birthday Problem*, Communications in Statistics - Theory and Methods, 53(12), 4487–4497, 2024
- [21] D. Powers, *Applications and Explanations of Zipf's Law*, NeM-LaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, pages 151 – 160, 1998.
- [22] Kyle Siegrist. *The Birthday Problem*. Random Services – Probability, Mathematical Statistics, and Stochastic Processes. <https://www.randomservices.org/random/urn/Birthday.html>, 2018.
- [23] R.L. Ott and M. Longnecker, *An Introduction to Statistical Methods and Data Analysis 7th edition*, Cengage Learning, 2016.
- [24] A. Carlson, R. Hiromoto and R. Wells, *Breaking Block and Product Ciphers Applied Across Byte Boundaries*, The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 733–736, 2011.