Kasiski's Test: Couldn't the Repetitions be by Accident?

KLAUS POMMERENING

Abstract: In searching for repetitions in a periodic polyalphabetic ciphertext, we usually find several true (causal) repetitions that give information about the period. But we also find some accidental repetitions at distances unrelated to the period which may mislead the cryptanalyst. A simple formula shows that these accidents are rather unlikely.

Keywords: Kasiski, repetition, Birthday Paradox.

Address correspondence to Klaus Pommerening, Institut für Medizinische Biometrie, Epidemiologie und Informatik der Johannes-Gutenberg-Universität, D-55101 Mainz, Germany. E-mail: pommerening@imbei.uni-mainz.de.

Repetitions in a Polyalphabetic Ciphertext

Kasiski's method finds the period of a polyalphabetic cipher in the following way: If a string of characters repeatedly appears in the

ciphertext, assume that the distance between the occurences is a multiple of the period. Find as many repetitions as possible and calculate the greatest common divisor of the distances. This gives the period or a small multiple of it.

For the historic context of this method see [3]; Babbage had invented the method ten years earlier than Kasiski but never published his results, see [7].

Kasiski's method is based on the following observations [4, Section 14]:

1. If a plaintext is encrypted by distinct alphabets that cyclically repeat with a period of l, and if a certain sequence of letters occurs k times in the text, then it will be encrypted with the same sequence of alphabets k/l times in the mean.

2. If a repeating sequence of letters is encrypted by the same sequence of alphabets, then the ciphertext contains a repeated pattern; the distance of the two occurences is a multiple of the period l.

3. Not every repeated pattern in the ciphertext arises in this way; but the probability of an accidental repetition is noticeably smaller.

Because of observation 3 the cryptanalyst has to omit some of the distances—by intuition, but essentially by trial and error. Therefore an obvious and natural question is: Is the probability of an accidental repetition really much smaller, as stated in 3?

The answer is a simple exercise in probability theory, a corollary of the Birthday Paradox. In spite of its simplicity, there seems to be no explicit

reference to this result in the cryptologic literature in the context of Kasiski's method.

The goal of this paper is to show that elementary calculus may give a satisfying answer to the question in the title. The intermediate results might be improved by refined theoretic considerations. There is room for experimental mathematics as well. The final section discusses some open problems that make up suitable undergraduate projects.

Note. The Birthday Paradox also has other applications in cryptology, the most renowned is to hash functions: the Birthday Paradox tells how long the hashes should be in order to avoid collisions (= repetitions), see [5, Sections 9.5 and 9.7] [6, Section 7.4] [8, Section 7.3]. For statistical applications see [2, Chapter II, Section 3].

Counting Repetitions

In several situations we want to know the probability that certain data agree or that certain events repeat. Here are three sample questions:

- What is the probability that at least two of a group of people meeting accidentally in the same room share their birthdays?
- What is the probability that at least two of r randomly and independently chosen character strings of length t are the same?
- Draw r balls from an urn containing N distinct balls (with replacement). What is the probability that you get at least one of the balls twice?

Let us calculate the probability in the urn experiment. There are N possible events of which we observe r (with possible repetitions).

- The probability that the first event is a repetition is 0.
- Therefore the probability that the first event *is not* a repetition is $1 = \frac{N}{N}$.
- The probability that the second event is not a repetition is $\frac{N-1}{N}$.
- The probability that **then also** the third event is not a repetition is $\frac{N-2}{N}$. (There are $N-2$ choices left that don't give a repetition.)
- The general case: If there was no repetition among the first $r-1$ events, then the probability is $\frac{N-r+1}{N}$ that also the r-th event is not a repetition.

From this we get the following well-known result [2, chapter II, Section 3]:

Theorem 1 *The probability of a repetition in a sequence of r independent events from a set of N is*

$$K(N,r) = 1 - Q(N,r)$$

where

$$Q(N,r) = \frac{N \cdot (N-1) \cdots (N-r+1)}{N^r} = [1 - \frac{1}{N}] \cdots [1 - \frac{r-1}{N}].$$

Applications

Birthdays: For $N \approx 365.22$, $r = 23$, we have $Q(N, r) \approx 0.493$, therefore the probability of a coincidence is ≈ 0.507. *If there are 23 people in the same room, the probability that two of them share their birthdays, is greater then* $\frac{1}{2}$. From this observation the Birthday Paradox got its name.

Character strings: Consider strings over the alphabet {A,...,Z}. Choose r strings of length t randomly and independently: This makes $N = 26^t$ possible events. The probability that at least two strings are identical is $K(26^t, r)$. For $r = 100, 300, 1000, 5000$ let these probabilities be p_t, q_t, r_t, s_t, respectively. Direct calculation from Theorem 1—with the help of a small computer program—gives Table 1. The table shows for example, that for $r = 1000$ there is more than a 60% chance that we find two identical four letter strings; but two identical five letter strings are rather unlikely (probability $< 5\%$).

$t \rightarrow$	1	2	3	4	5	6	7	$r \downarrow$
p_t	1	**1.000**	**0.246**	0.011	0.00042			100
q_t	1	1.000	**0.923**	**0.094**	0.0038	0.00015		300
r_t	1	1	1.000	**0.665**	**0.041**	0.0016		1000
s_t	1	1	1.000	1.000	**0.651**	**0.040**	0.0016	5000

Table 1. Probabilities for repetitions of strings. Entries $< 10^{-4}$ are omitted. Values given as 1 are exact, values given as 1.000 are rounded off. In each row the cut point "50% probability" lies between the two entries in boldface.

Bounds for the Number of Repetitions

The formula in Theorem 1 is awkward for manual calculation; it also gives no direct idea of the order of magnitude of the probability. Fortunately, using some elementary calculus, we find convenient bounds that also show the behaviour for large values of the parameters. First we derive an upper bound for the number $K(N,r)$ of repetitions:

- The probability that the i-th event is a repetition is $\leq \frac{i-1}{N}$, because there were only $i-1$ events before.
- Therefore the probability that up to the r-th event there is a repetition is

$$K(N,r) \leq \frac{0}{N} + \cdots + \frac{i-1}{N} + \cdots + \frac{r-1}{N} = \frac{r(r-1)}{2N}.$$

From this we get the right inequalities of Theorem 2.

Theorem 2 (i) *The probability $K(N,r)$ of a repetition is bounded by*

$$1 - e^{-\frac{r(r-1)}{2N}} \leq K(N,r) \leq \frac{r(r-1)}{2N}.$$

(ii) *If $r \leq \sqrt{2N}$, then we have*

$$(1 - \frac{1}{e}) \cdot \frac{r(r-1)}{2N} \leq K(N,r) \leq \frac{r(r-1)}{2N}.$$

or, somewhat weaker,

$$0.3 \cdot \frac{r(r-1)}{N} \leq K(N,r) \leq 0.5 \cdot \frac{r(r-1)}{N}.$$

(iii) *If* $r \leq \sqrt{N}$, *then* $K(N,r) < \frac{1}{2}$.

(iv) *If* $r \geq 1 + \sqrt{2N \ln 2}$, *then* $K(N,r) > \frac{1}{2}$.

Proof. The left inequality in (i) follows from the inequality $1 - x \leq e^{-x}$ for $x \in \mathbb{R}$, hence

$$Q(N,r) \leq e^{-\frac{1}{N}} \cdots e^{-\frac{r-1}{N}} \leq e^{-\frac{r(r-1)}{2N}},$$

and $K(N,r) = 1 - Q(N,r)$.

The lower bound in (ii) follows from the inequality $1 - e^{-x} \geq (1 - \frac{1}{e})x$ in the real interval $0 \leq x \leq 1$; and this is true because the function $f(x) = 1 - e^{-x}$ is concave ($\cap$-shaped), $g(x) = (1 - \frac{1}{e}) \cdot x$ is linear, and $f(0) = g(0)$, $f(1) = g(1)$.

For (iii) the upper bound simplifies to $K(N,r) < \frac{r^2}{2N} \leq \frac{N}{2N} = \frac{1}{2}$.

In (iv) we have $r(r-1) > 2N \ln 2$. Therefore the left hand side of (i) is $> \frac{1}{2}$. $\diamond$

Theorem 2 (iii) and (iv) together give the rule of thumb that appears in many cryptography textbooks, see [6, Section 7.4] [8, Section 7.3]:

The cut point "50% probability" for repetitions is close to $r = \sqrt{N}$.

More exactly it is between $\sqrt{N}$ and $1 + 1.18\sqrt{N}$. As a special case of Theorem 2 (iii) with $N = n^t$ we immediately get

Theorem 3 *For r random character strings of length t over an alphabet of n characters with $r \leq n^{t/2}$ the probability of a repetition is less than $\frac{1}{2}$.*

The Probability of Accidental Repetitions

Now we apply this to the substrings of a random character string of length r (over an n letter alphabet), where "random" means that each character is chosen independently and with probability $\frac{1}{n}$. We abandon the exact mathematical reasoning and make the **simplifying assumption** that the substrings are stochastically independent; this is clearly not perfectly correct, because the substrings overlap—but see the discussion in the final section. We also neglect the fact that a string of length r has only $r - t + 1$ substrings of length t. Then the probability that a repetition of length t occurs is (approximately) $K(n^t, r)$, and Table 1 above illustrates the order of magnitude of these numbers (when $n = 26$).

Theorem 3 immediately gives: For a random character string of length $r \leq n^{t/2}$ (over an n letter alphabet) the probability of a repetition of length t is $< \frac{1}{2}$. That means: *For random strings up to a length of $n^{t/2}$ a repetition of any substring of length t is fairly unlikely.* Or to express it conversely:

For random strings of length r a repetition of any substring of length t is rather unlikely ($< 50\%$) as long as

$$\text{(A)} \qquad t \geq 2 \cdot \frac{\log r}{\log n}\,.$$

For $n = 26$ the bound (A) is approximately $t \geq 1.413 \cdot \log r$ (logarithm in base 10).

This is the main answer to the title question. For a non-mathematician maybe we would express it as follows:

- For texts of length 100, accidential repetitions of length 3 or more are rather unlikely; Table 1 gives the more exact result that the probability is $< 25\%$.
- For texts of length 300, accidential repetitions of length 4 or more are rather unlikely (Table 1: probability $< 10\%$), but at least one accidental repetition of length 3 occurs with high probability (Table 1: $> 90\%$).

And so on—use formula (A), Table 1, or Theorem 2.

One might wish to derive more statistical results on the probabilities of repetitions. However the simple statements given here are sufficient as a justification for Kasiski's method; in particular considering the cut point "50%" seems adequate for the cryptanalyst, even if this simplistic view is somewhat unsatisfactory for the mathematician.

Kasiski's Test

When the cryptanalyst carries out Kasiski's test he doesn't examine a random text. In order to apply the results of the preceding section we have to make one **further simplifying assumption**: a polyalphabetic ciphertext behaves randomly except for the effect of the period. Now when we find a repetition of length t, and t is as least as large as in (A), then we are pretty sure that we have found a true (or causal) repetition and the period is a divisor of the distance. The smaller t is, the more we are prepared to reject some repetitions; again Table 1 gives more precise hints for ciphertexts of lenghts 100, 300, 1000, or 5000. If we find a "long" repetition, we may assume with extremely high probability that it is a causal repetition.

Discussion

Are the theoretical results above exact enough for Kasiski's test in view of the simplyfing assumptions that we had to make? Here we give only some coarse empirical results, leaving room for more elaborate investigations.

1. May we really apply the theorems and the resulting Table 1 to the substrings of a long character string? to get empirical evidence Wwe generated 100 random texts of lengths 100 and 300 each, 26 random texts of lengths 1000, 5000 each, over the 26 character alphabet, and found no remarkable deviations from the theoretical results derived for independent strings.

2. Is the number of accidental repetitions in a polyalphabetic ciphertext really as low as in a random text? We encrypted 100 English plaintexts of length 300 with keys of lengths 6, 10, and 17 each (with mixed alphabets by the way). Here we found small deviations: The ciphertexts seem to have *fewer* accidental repetitions than random texts, see figures 1 and 2 for key lenghts 6 and 17. A partial explanation is given below.

These simulation results confirm that the formulas in this paper apply to polyalphabetic ciphertexts with negligeable deviations.

Here are some observations and heuristic arguments that could merit some further investigations:

- Why is the number of accidental repetitions in item 2 smaller than in random strings? One major effect is: there can be no accidental repetitions whose distance is a multiple of l, the period of the cipher; each such repetition must be causal since ciphertext and key conform. Therefore we expect that the number of repetitions (for any length) is smaller by $1/l$. However this is not yet the complete truth: Non-accidental, but "false", repetitions may arise in some other ways, as shown in [1, Section 17.4]: when the key contains a repeated substring—as in "seventyseven"— or when key and plaintext contain the same word, for example if the key for a military text contains the word "division". It seems hard to adequately adjust a general model to fit these observations. But unfortunately in exceptional cases these effects can lead to annoying *long* "accidental" repetitions, not

predicted by the estimates above.

- How many causal repetitions can we expect? This depends on the statistics of the plaintext language. Possible approaches are:
  - *Simulation.* In the experiment of item 2 we also counted the causal repetitions and found significantly more causal than accidental repetitions. See Figures 1 and 2 for repetitions of length 3.
  - Start with *trigram freqencies* and calculate the resulting probabilities for repetitions of length three in a ciphertext, depending on the key length, under suitable simplifying assumptions.
  - Model the language by a *Markov source* of low order and derive the relevant probabilities.
- Consider the distribution of the number of repetitions of a fixed length—in random texts, or accidental or causal repetitions in ciphertexts. They all seem to follow a Poisson distribution. Determine the parameters.

Figures 1 and 2 show a selection of typical simulation results. The x-axis represents the number of repetitions of length 3 in one text; note that one long repetition of length $t \geq 3$ counts as $t - 2$ repetitions of length 3. The y-value shows how often exactly x repetitions occured in 100 texts (all of length 300). The fat gray line gives this frequency for random texts and serves as reference, it is the same in both diagrams. The thin gray line

gives the frequency of x accidental repetitions, the black line the frequency of x causal repetitions in the polyalphabetic ciphertexts.

About the Author

Klaus Pommerening works as a professor in Medical Informatics and tries to build a cryptographic infrastructure for the German medical research networks. He has taught mathematics at the universities of Heidelberg and Mainz Germany after graduating from the Freie Universität Berlin and receiving his PhD from the University of Mainz. Home page: `http://www.staff.uni-mainz.de/pommeren/`.

References

1. Bauer, F. L. 1997. *Decrypted Secrets; Methods and Maxims of Cryptology*. Berlin: Springer.

2. Feller, W. 1957. *An Introduction to Probability Theory and Its Applications*. Volume I. New York: Wiley.

3. Kahn, D. 1967. *The Codebreakers*. New York: Macmillan.

4. Kullback, S. 1976. *Statistical Methods in Cryptanalysis*. Laguna Hills: Aegean Park Press.

5. Menezes, A. J., van Oorschot, P. C., Vanstone, S. A. 1997. *Handbook of Applied Cryptography*. Boca Raton: CRC Press.

6. Schneier, B. 1996. *Applied Cryptography*. New York: John Wiley.

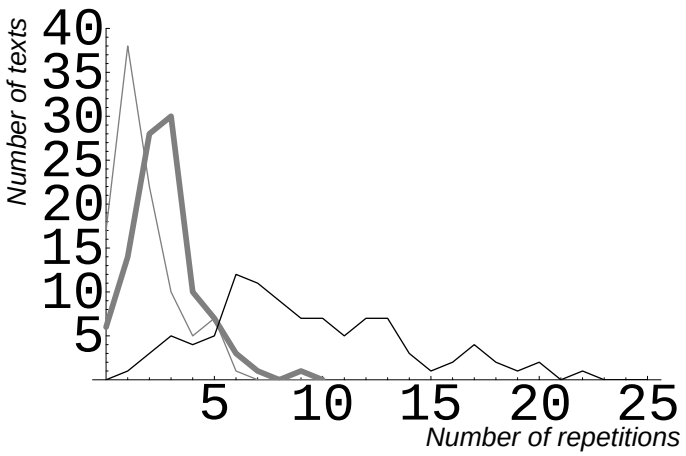

Figure 1. Distribution of the number of repetitions in polyalphabetic ciphertexts, key length 6. x-axis: number of repetitions of length 3, y-axis: number of occurences of x repetitions. Fat gray line: random texts, thin gray line: accidental repetitions, black line: causal repetitions; one count of 31 causal repetitions falls outside the picture.

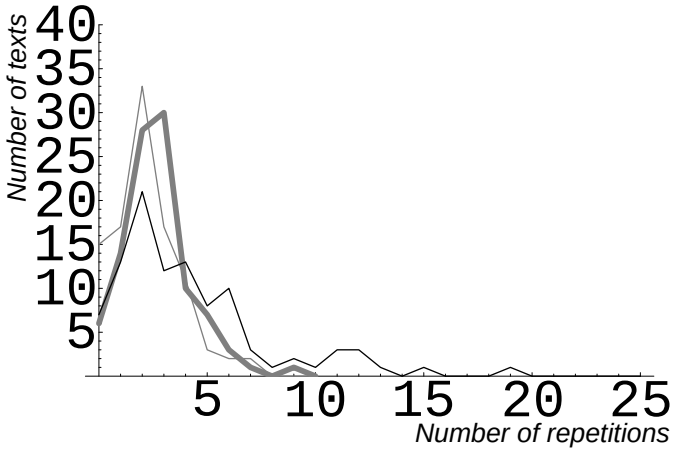

Figure 2. Distribution of the number of repetitions, key length 17.

7. Singh, S. 1999. *The Code Book*. London: Fourth Estate.

8. Stinson, D. R. 1995. *Cryptography - Theory and Practice*. Boca Raton: CRC Press.