Using the Entropy of N-Grams to Evaluate the Authenticity of Substitution Ciphers and Z340 in Particular

Tom S Juzek
Saarland University
Campus A2.2, R1.25
66123 Saarbrücken, Germany
tom.juzek@uni-saarland.de

Abstract

The present paper uses information theoretic entropy as a means to evaluate the authenticity of homophonic substitution ciphers. We motivate the use of entropy on n-grams and then validate its applicability, by using it on various true ciphers and pseudo-ciphers. Differences in entropy allow us to apply further formal analyses, e.g. support-vector machines, in order to make predictions about a potential cipher's status. We train several support-vector machines and validate them. We then apply the models to two classic ciphers, the Zodiac Killer's first major cipher, z408, which has been solved, and his second cipher, z340, which remains unsolved. The models correctly identify z408 as a substitution cipher. z340 is classified as an advanced cipher or pseudo-cipher.

1 Introduction

For unsolved ciphers, it is often difficult to determine which type of cipher is applicable. A seemingly encrypted text might be a simple homophonic substitution cipher, a transposition cipher, a Vigenère cipher, and the like, or it might not even be a *bona fide* cipher at all. Even if one suspects that a cipher is a simple homophonic substitution cipher, then there are only a few known analyses that count as formal evidence, such as e.g. Ravi and Knight (2008), Corlett and Penn (2010), and Ravi and Knight (2011).

The present paper presents another formal measure that can be used to detect underlying language-like information in a possible homophonic substitution cipher (in the following shortened to just *substitution cipher*). The measure quantifies differences in (information theoretic) entropy between a cipher and a meaningless baseline. Critically, in contrast to previous analyses, the measure can be used to train classifiers like support-vector machines or k-means clusters to make predictions about the status of a text.

The present paper is structured as follows: In Section 2, we review common types of substitution ciphers, including one-to-one ciphers, one-to-many ciphers, and many-to-many ciphers. Our focus is on one-to-many ciphers and many-to-many ciphers.

The Zodiac Killer's second major cipher, z340, is often assumed to be a substitution cipher, but it is generally accepted that it has not been solved yet. Section 3 briefly introduces z340 and compares it to the Zodiac Killer's first major cipher, z408, which has been deciphered, cf. Graysmith (1987).

We give basic analyses of z340 and z408, viz. character frequency analyses and n-gram analyses, in Section 4.1. We compare the two ciphers to a pseudo-cipher that we have created, to illustrate the limitations of character frequency analyses. N-gram analyses, on the other hand, give useful hints about the authenticity of a cipher. However, even more powerful measures are needed to give more definite insights.

For this, we use the information theoretical measure of entropy, cf. Shannon (1948). We calculate entropy based on bigrams and trigrams and we apply the measure to various texts and ciphers. To be able to train a classifier model, some sort of baseline is needed and we establish such baselines for plain texts and ciphers. All this is done in Section 4.2. In Section 4.3, we then use the results to train a support-vector machine, which we validate on further test data. The classifiers give very good results for plain texts and one-to-many substitution ciphers and good results for many-to-many substitution ciphers. The models correctly classify z408, the Zodiac Killer's solved cipher, as a substitution cipher. z340, the unsolved cipher, is classified as an advanced cipher or pseudo-cipher.

Proceedings of the 2nd International Conference on Historical Cryptology, pages 117-125
Mons, Belgium, 23-26 June, 2019

Section 5 discusses a few issues with the measure and Section 6 concludes the paper.

2 Substitution ciphers

Common substitution ciphers include one-to-one ciphers, one-to-many ciphers, and many-to-many ciphers. For an introduction to ciphers (and cryptography in general), see e.g. Hoffstein et al. (2008). In the following, we use "letter" to refer to a unit of an unencrypted plain text, "symbol" for a unit of a cipher, and "character" for both.

In a one-to-one cipher, any plain text letter is mapped to exactly one cipher symbol, e.g. "A" to "*X*", "B" to "ϒ", etc. Let l be a letter of the set of letters **L**, s a symbol in the set of symbols **S**. This is expressed through the relation r in Equation 1: Any given letter l maps to exactly one symbol s; and any given symbol s maps to exactly one letter l.

In a one-to-many cipher, some or all plain text letters are mapped to more than one cipher symbol, e.g. "A" might be mapped to both "*X*" and "Ω", "B" might be mapped to "ϒ" and "Ψ", etc. Here, too, each cipher symbol is used for one plain text letter only. In our example, "Ω" is always "A", etc. One-to-many ciphers are expressed through the relation in Equation 2: Any given letter l maps to at least one symbol s; but any given symbol s maps to exactly one letter l.

In a many-to-many cipher, there is no restriction of having exactly one mapping back from symbol to letter. Each plain text letter can map to more than one cipher symbol and vice versa. "A" might be mapped to both "*X*" and "Ω", "B" might be mapped to "ϒ", but also "Ω", etc. "Ω" now represents both "A" and "B". Many-to-many ciphers are expressed in Equation 3: Any given letter l maps to at least one symbol s; and any given symbol s maps to at least one letter l.

$$r(l) \mapsto \exists! s \;\; s \in \mathbf{S} \quad \textit{and} \quad r(s) \mapsto \exists! l \;\; l \in \mathbf{L} \quad (1)$$

$$r(l) \mapsto \exists s \;\; s \in \mathbf{S} \quad \textit{and} \quad r(s) \mapsto \exists! l \;\; l \in \mathbf{L} \quad (2)$$

$$r(l) \mapsto \exists s \;\; s \in \mathbf{S} \quad \textit{and} \quad r(s) \mapsto \exists l \;\; l \in \mathbf{L} \quad (3)$$

Table 1 gives an example of a many-to-many cipher. Note how e.g. "S" maps to both "10" and "57" and how "31" represents both "Y" and "G".

Figure 1: The beginning of the z340 cipher.

The text is the beginning of *The Confession*, the first letter attributed to the Zodiac Killer, cf. Graysmith (1987). The letter can be found on Wikisource (The Wikimedia Foundation, 2018b). *The Confession* also features as one of the texts used in Section 4.2.

One-to-one substitution ciphers are relatively easily detected – unless some other trick is employed. The focus of the present paper is on one-to-many ciphers and many-to-many ciphers. Such ciphers come in degrees. While weak instances can be rather easy to decrypt, strong instances can be very hard, sometimes virtually impossible to decrypt (see below for details).

3 z340

The Zodiac Killer's second major cipher, z340, is often assumed to be a substitution cipher. However, even 50 years after its appearance, a generally accepted solution is lacking. Figure 1 illustrates the beginning of the cipher and the full cipher can be found on Wikisource (The Wikimedia Foundation, 2018b). This contrasts to the Zodiac Killer's first major cipher, z408. z408 is generally assumed to have been solved by Donald and Bettye Harden. Parts of the cipher and the full solution can also be found on Wikisource (The Wikimedia Foundation, 2018b). Zodiackillerciphers.com (2012) contains the full cipher and the Harden solution.

4 Analyses

4.1 Character frequencies and n-gram frequencies

To gain first insights into a potential substitution cipher, there are two basic analyses: A character frequency analysis and an n-gram analysis. Character frequencies provide the number of occurrences per character of a plain text or cipher. N-gram analyses of characters give the number of occurrences per n-gram (bigram, trigram, etc.) in a text – see e.g. Jurafsky (2019) for an introduction to n-grams.

Plain text:	S	H	E	W	A	S	Y	O	U	N	G	A	N	D	B	E	A	U	T	I	F
Encrypted:	10	21	48	40	57	47	31	53	51	28	31	7	44	26	6	25	4	30	1	8	29

Table 1: An example of a many-to-many substitution cipher. The top row is the plain text, below it is the encrypted message, using the encryption algorithm described in Section 4.2.

For example, if the cipher text was simply "DADAISM", then the character frequencies are as follows: 'A':2, 'D':2, 'I':1, 'M':1, 'S':1. The bigram counts are as follows: 'DA':2, 'AD':1, 'AI':1, 'IS':1, 'SM':1.[1] The trigrams are: 'DAD':1, 'ADA':1, 'DAI':1, 'AIS':1, 'ISM':1.

However, character frequency analyses are only of limited value when one wishes to assess the authenticity of substitution ciphers. When using character frequencies, it is difficult to distinguish a semi-random string that observes common character frequencies and that is then encrypted using a true cipher. However, differences will show up in an n-gram analysis. This is illustrated in Figure 2. Figure 2 (left) plots the character frequencies of a true cipher, z408, and those of a comparable pseudo-cipher, which is based on a semi-random string. The pseudo-cipher is comparable in length and has a similar symbol set to z408. For details regarding the encryption algorithm, see Section 4.2.

Character frequencies of both ciphers are plausible. However, as Figure 2 (right) illustrates, the bigram frequencies reveal differences – while the bigram frequencies of z408 are plausible, the bigram frequencies for the pseudo-cipher fall "flat", i.e. the pseudo-cipher seems to lack a plausible bigram count.

It has been noted that the Zodiac Killer's unsolved z340 also seems to lack plausible n-gram counts (Knight, 2013, p. 91). Figure 3 gives the character frequencies and bigram frequencies for z340. Note how z340 rather resembles the pseudo-cipher than the true cipher.

The n-gram analysis of z340 indicates that it might not be a *bona fide* cipher. However, a more formal analysis is needed for more conclusive evidence. We provide such evidence by analysing the entropy of various texts, including true ciphers and pseudo-ciphers, and then training several support-vector machines on the results.

[1] And '^D':1 and 'M$':1 if one wishes to account for beginning of line ("^") and end of line ("$"). In the following, we will not include those two.

4.2 Entropy as a measure of authenticity

Entropy in information theory, as introduced by Shannon (1948), is a measure of order of a system and can be applied to various levels of a linguistic sequence, including characters, n-grams, words, multiple words, and entire sentences. The general formula for entropy, H, is given in Equation 4. In our case, F is the frequency of the respective bi- and trigrams.

$$H = -\sum_i F_i \, log(F_i) \tag{4}$$

For instance, the bigram entropy for "DADAISM" is 0.68, the entropy for "IADSDMA" is 0.78. However, it can be difficult to make sense of the values, especially when it is compared across sequences with different symbol sets and of different lengths. Consider the bigram entropy for the pseudo-cipher from above: It comes out at 2.52 – but it is not clear what this exactly means.

Thus, for ease of interpretation, we compare a sequence's entropy, H_s, to the entropy of a meaningless baseline, H_b. For any sequence, we pseudo-randomly shuffle the sequence in question and use the shuffle as the baseline. To avoid distortions of an unlucky shuffle, we take 1000 shuffles and average their entropy values. This is $\overline{H}_b$. The absolute difference between H_s and $\overline{H}_b$ is what we abbreviate by $|\Delta_H|$. Creating the pseudo-random baselines and calculating $|\Delta_H|$ is done with a script that we wrote in Python (Python Software Foundation, 2018) (all scripts can be found on GitHub[2]. $|\Delta_H|$ is easier to interpret: A value of 0 means that the sequence lacks any order, just like the pseudo-random baselines. The greater the value, the more ordered a sequence is. Accordingly, the $|\Delta_H|$ for "IADSDMA" is 0.0, but the $|\Delta_H|$ for "DADAISM" is 0.1. $|\Delta_H|$ for the pseudo-cipher from above is 0.0, but $|\Delta_H|$ for the true cipher z408 comes out at 0.06.

However, even a $|\Delta_H|$ can be hard to interpret. What does a $|\Delta_H|$ of 0.06 mean? Some contextualisation is needed and in order to provide it, we

[2] https://github.com/superpumpie/z340_2.

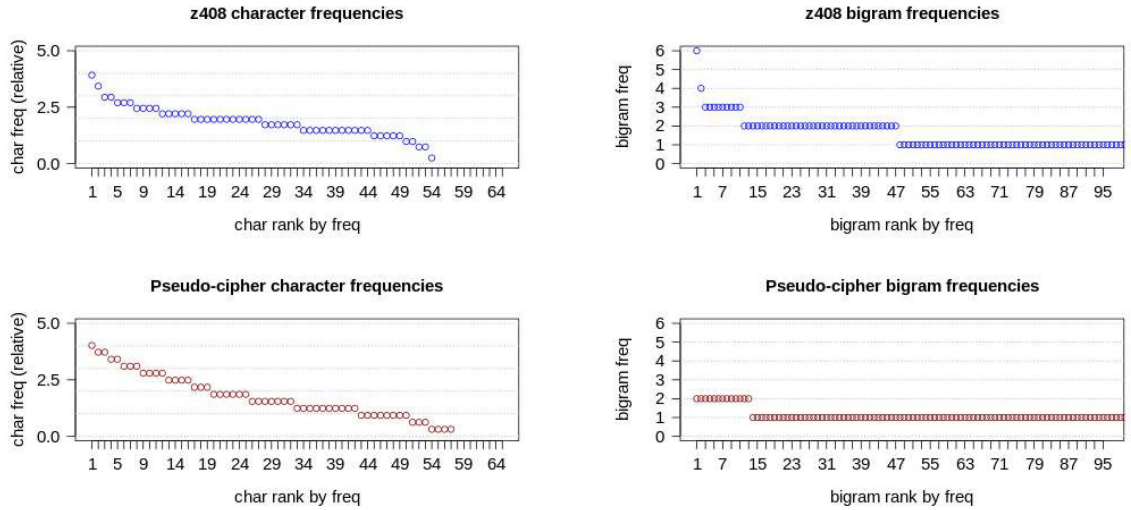

Figure 2: Left: The relative character frequencies, y-axis, for the true cipher z408 (top) and for a pseudo-cipher based on a semi-random string (bottom). Items are ordered on the x-axis by their frequencies in descending order. Right: The bigram frequencies, y-axis, for the same two ciphers, again in descending order as per rank.

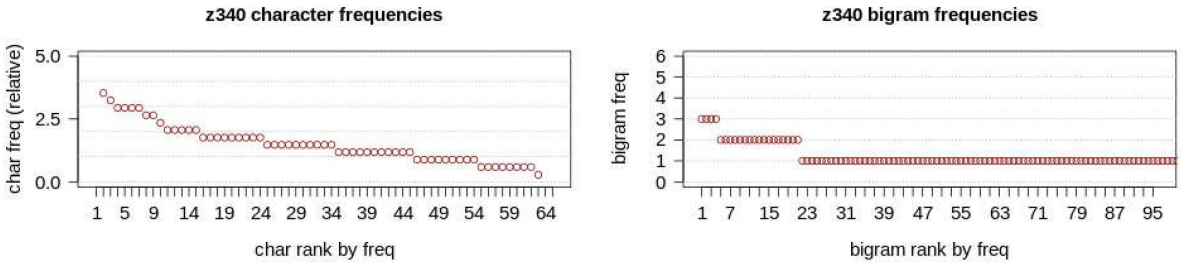

Figure 3: Left: The relative character frequencies, y-axis, for the unsolved z340 cipher. Items are ordered on the x-axis by their frequencies in descending order. Right: The bigram frequencies, y-axis, for the z340 cipher, again in descending order as per rank.

analyse various plain texts vs semi-random strings and various true ciphers vs pseudo-ciphers. We begin with a training set, which we then use to train several support-vector machines. We later validate the models with further test data sets.

As a first step, we analyse 32 plain texts and 32 semi-random strings. The majority of those are snippets from the Top 100 books on Project Gutenberg (2018), others were extracted from various sources, incl. Wikipedia (The Wikimedia Foundation, 2018a) and Wikisource (The Wikimedia Foundation, 2018b). All data sets, whose length varies between 255 and 459 characters, can be found in the above mentioned GitHub repository, including their sources. The semi-random texts, with a length of 255 to 425 characters, were created with a script that we wrote in Python (Python Software Foundation, 2018). The texts are semi-random in the sense that they observe English letter frequency. The corresponding $|\Delta_H|$'s for bigrams and trigrams are illustrated in Figure 4 (left).

We encrypt all true texts and pseudo-texts with an algorithm modelled after the encryption method used for z408, also using one of our Python scripts. Each letter is, partly depending on its frequency, pseudo-randomly mapped to one to five unique symbols, resulting in one-to-many ciphers with symbol sets of 52 to 66 symbols. The encryption script can also be found in the above mentioned GitHub repository. The $|\Delta_H|$'s in entropy for the one-to-many ciphers are illustrated in Figure 4 (right).

We also create many-to-many ciphers. The encryption algorithm for this has two layers. First, similar to the algorithm above, each letter is, partly depending on its frequency, pseudo-randomly mapped to one to five symbols. This is the first encryption layer. Then, the first layer is encrypted again: Each first layer symbol is pseudo-randomly mapped to one to four second layer symbols. The second mapping is not unique, in the sense that most second layer symbols map to more than one first layer symbol, resulting in a many-to-many cipher. The ciphers have second layer symbol sets of 60 to 64 symbols, which is similar in size to the symbol sets of the one-to-many ciphers.

The $|\Delta_H|$'s for the many-to-many ciphers are illustrated in Figure 5 (left). As Figure 5 (left) indicates, the many-to-many encryption algorithm is a lot stronger than the one-to-many algorithm. True many-to-many ciphers and pseudo-ciphers overlap to some degree, illustrating that a very strong many-to-many cipher might be indistinguishable from a pseudo-cipher.

In a last step, we add the two ciphers by the Zodiac Killer to the picture. Their $|\Delta_H|$'s are illustrated in Figure 5 (right), including a comparison with our other ciphers. Z340 sits among the pseudo-ciphers. And compared to our encryption algorithms, z408 uses a fairly weak encryption technique. The latter is not a surprise. In their selection of symbols, humans are biased. For instance, if the mappings for "A" are "*X*" and "Ω", then a human might tend to choose "*X*" if e.g. "A" precedes an "N" but choose "Ω" if "A" precedes a "T". Our pseudo-random encryption algorithm has no such biases.

4.3 Training and testing support-vector machines for classification

We use the results from above to train three support-vector machines (*SVMs*), cf. Ben-Hur et al. (2002). One SVM for the plain texts vs the semi-random strings, one for the true one-to-many ciphers vs the one-to-many pseudo-ciphers, and one for the true many-to-many ciphers vs the many-to-many pseudo-ciphers. The one-to-many SVM and the many-to-many SVM are illustrated in Figure 6.

In a second step, we validate the SVMs on further test data sets. We use another 32 true plain texts and 32 semi-random strings, 32 true one-to-many ciphers and 32 one-to-many pseudo-ciphers, and 32 true many-to-many ciphers and 32 many-to-many pseudo-ciphers. This gives us a 50-50 training-testing split. The texts and ciphers were obtained in a similar fashion as described in Section 4.2. The results of the testing phase are given in Table 2.

5 Discussion

The SVM for plain texts makes excellent predictions, the one-to-many SVM also makes very good predictions. The many-to-many still makes good predictions, but with a somewhat lower accuracy than the other models, presumably because the encryption technique is rather strong.

According to the one-to-many model and the many-to-many model, z408 is classified as a true substitution cipher. z340, on the other hand, is

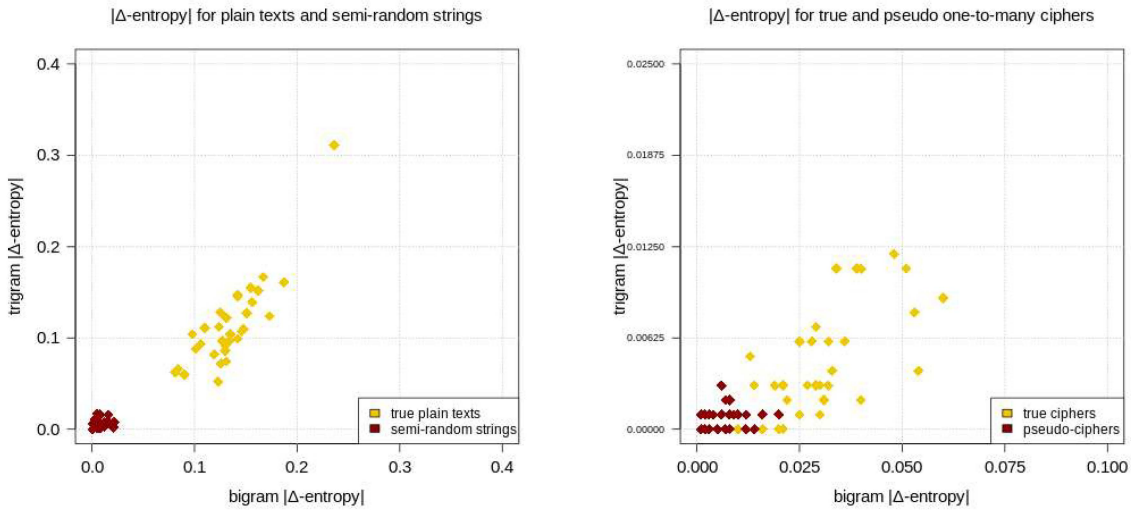

Figure 4: Left: The bigram $|\Delta_H|$'s, on the x-axis, and the trigram $|\Delta_H|$'s, on the y-axis, for the 32 plain texts and 32 semi-random texts in our training set. Right: The $|\Delta_H|$'s for the 32 true one-to-many ciphers and the 32 one-to-many pseudo-ciphers in our training set.

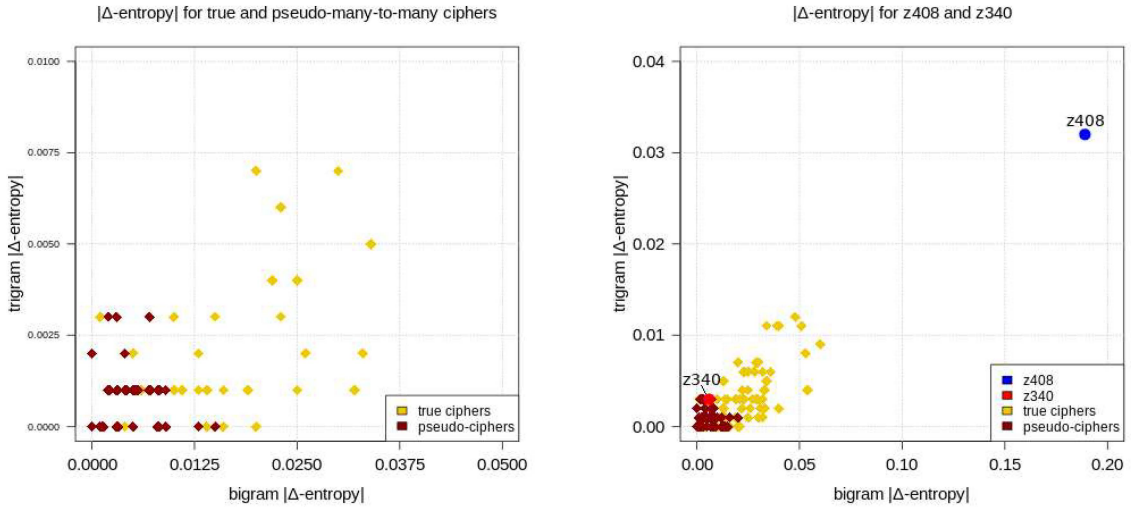

Figure 5: Left: The bigram $|\Delta_H|$'s, on the x-axis, and the trigram $|\Delta_H|$'s, on the y-axis, for the 32 true many-to-many ciphers and the 32 one-to-many pseudo-ciphers in our training set. Right: The $|\Delta_H|$'s for z408 and z340, with the results for the other ciphers for contextualisation.

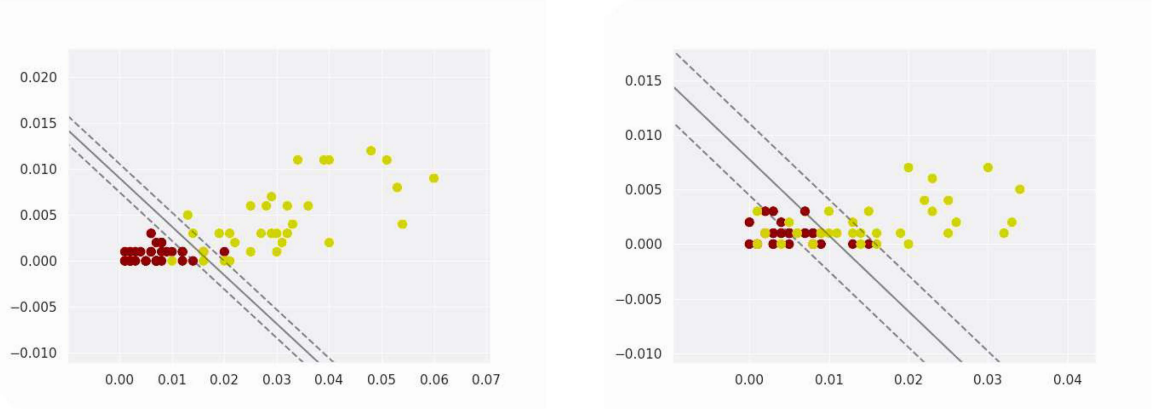

Figure 6: Left: The results of an SVM trained on the 64 one-to-many ciphers in our training data sets, 32 of which are true ciphers, the other 32 being pseudo-ciphers. The decision boundary is in grey. Items left of the boundary are predicted to be pseudo one-to-many ciphers, items right of it true one-to-many ciphers. Right: The same for the 64 many-to-many ciphers in our training set. Visualisation of the SVM was partly done with code from https://jakevdp.github.io.

text	predicted true	predicted pseudo
actual true	32	0
actual pseudo	0	32
$F_1 = 1.0$		

o-t-m	predicted true	predicted pseudo
actual true	30	2
actual pseudo	0	32
$F_1 = 0.96$		

m-t-m	predicted true	predicted pseudo
actual true	25	7
actual pseudo	6	26
$F_1 = 0.80$		

Table 2: Confusion matrices and F1-scores for the three support-vector machine models. Top: The model that classifies plain texts vs semi-random strings (*text*). Middle: The model that classifies true one-to-many ciphers vs one-to-many pseudo-ciphers (*o-t-m*). Bottom: The model that classifies true many-to-many ciphers vs many-to-many pseudo-ciphers (*m-t-m*).

extremely close to zero and both models predict that it is not a true substitution cipher. However, it should be noted that z340 sits relatively close to the decision boundary of the many-to-many model and that in some of the re-runs of the procedure, the many-to-many model places z340 right above the decision boundary.

Considering that z408 is a rather weak one-to-many cipher, we think that it is unlikely that the same author had been able to produce another substitution cipher, i.e. z340, such that its encryption mechanism became stronger by several orders of magnitude. We interpret this as evidence that z340 is either not a *bona fide* substitution cipher or uses a different, more sophisticated encryption mechanism altogether. For instance, it might be a transpose cipher or a Vigenère cipher.

There are a few things to note about the presented measure. First, the measure is not absolute. Consider for instance Kryptos Passage 4 by Jim Sanborn, also available on Wikipedia (The Wikimedia Foundation, 2018a). Here are the first three lines of Kryptos 4:

NGHIJLMNQUVWXZKRYPTOSABCDEFGHIJL
OHIJLMNQUVWXZKRYPTOSABCDEFGHIJL
PIJLMNQUVWXZKRYPTOSABCDEFGHIJLM

The entropy for Passage 4 comes out as 1.55 and the $|\Delta_H|$ is 0.92. This, of course, does not mean that Passage 4 is a plain text or a substitution cipher. Upon visual inspection, it becomes immediately clear that neither is likely. This is a

limitation one has to keep in mind using entropy on potential ciphers.

Another issue is that there is no definite cut off point between strong many-to-many ciphers and pseudo-ciphers. Very strong many-to-many ciphers can become indistinguishable from pseudo-ciphers. However, this is a good reflection of the underlying reality: As the strength of the encryption mechanism increases, the probability of being able to make sense of it decreases. $|\Delta_H|$ reflects this inverse relationship.

Finally, while we use SVMs, other analyses can be used as well. For instance, one could use a k-means clustering analysis, cf. Lloyd (1982) and Forgy (1965), in addition to an SVM.

6 Conclusion

We have shown that using differences in information theoretical entropy can be used to evaluate the authenticity of substitution ciphers. We created 64 true ciphers and another 64 pseudo-ciphers and split those into training and test data sets. We then trained and tested support-vector machines on our data sets. The model for one-to-many ciphers makes very good predictions, the model for many-to-many ciphers makes decent predictions. We applied those two SVM models to the the Zodiac Killer's two major ciphers, z408 and z340. z408, which has been solved, is correctly predicted to be a real substitution cipher. z340, which remains unsolved, is predicted to not be a substitution cipher. We think that it is likely that z340 is either another type of cipher, e.g. a transpose cipher or a Vigenère cipher, or that it might not be a *bona fide* substitution cipher after all.

As a next step, it would be interesting to apply the measure to other, unsolved ciphers. Also, ideally, further measures could be developed for other types of ciphers, like transposition ciphers or Vigenère ciphers.

Acknowledgments

Many thanks to asetniop, Jarl, K. L., Sasja, and the anonymous reviewers for their valuable feedback.

References

Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik. 2002. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137.

Eric Corlett and Gerald Penn. 2010. An exact a* method for deciphering letter-substitution ciphers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1040–1047, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edward W. Forgy. 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769.

Robert Graysmith. 1987. *Zodiac*. Berkley, New York City, NY.

Jeffrey Hoffstein, Jill Pipher, and Joseph H. Silverman. 2008. *An Introduction to Mathematical Cryptography*. Springer Publishing Company, Incorporated, 1 edition.

Dan Jurafsky, 2019. *Lecture CS 124: From Languages to Information. Introduction to N-grams – lecture notes*. `http://web.stanford.edu/class/cs124/`.

Kevin Knight, 2013. *Decipherment Tutorial. Workshop at the 2013 Annual Meeting of the Association for Computational Linguistics – lecture notes*. `https://kevincrawfordknight.github.io/extra/acl-tutorial-13-decipher-final.pdf`.

Stuart P. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137.

Project Gutenberg, 2018. *Project Gutenberg*. `http://www.gutenberg.org`.

Python Software Foundation, 2018. *Python: A dynamic, open source programming language*. `https://www.python.org/`.

Sujith Ravi and Kevin Knight. 2008. Attacking decipherment problems optimally with low-order n-gram models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 812–819, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Bayesian inference for zodiac and other homophonic ciphers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 239–247, Stroudsburg, PA, USA. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 7.

The Wikimedia Foundation, 2018a. *Wikipedia, the free encyclopedia*. `https://wikipedia.org/`.

The Wikimedia Foundation, 2018b. *Wikisource, the free library*. `https://en.wikisource.org/wiki/Author:Zodiac_Killer#Letters`.

Zodiackillerciphers.com, 2012. *Annotated solution to the 408 cipher, based on the Harden worksheets.* `http://zodiackillerciphers.com/408/key.html#1`.