Kapitel 2
Informationstheoretische Ansätze

Im vorigen Kapitel haben wir anhand einiger Beispiele gesehen, wie die statistischen Eigenschaften eines Kryptogramms zu Informationen über den Klartext und/oder den Schlüssel führen können. Dabei wurde klar, daß gewisse Verfahren trotz auf den ersten Blick gewaltiger Komplexität bereits für relativ kurze Nachrichten unsicher sind.

Bei der Wahl eines Kryptosystems interessiert vor allem die umgekehrte Frage: Wie *sicher* ist das System? Es hieße, den Gegner möglicherweise gewaltig zu unterschätzen, wenn man dabei nur die Sicherheit gegenüber den Angriffsmöglichkeiten betrachtet, die man selbst kennt: Ein ernstzunehmender Gegner wird häufig über Methoden verfügen, die nicht in der offenen Literatur dokumentiert sind.

In diesem Kapitel wollen wir daher untersuchen, wieviel Information über Schlüssel und Klartext in einem Kryptogramm enthalten ist – unabhängig davon, ob ein Gegner in der Lage ist, diese Information mit realistischem Aufwand nutzbar zu machen.

Man kann das auch so formulieren, daß wir den Gegner deutlich *über*schätzen, indem wir annehmen, daß ihm unbegrenzte Resourcen zur Verfügung stehen; er kann also *jeden* endlichen Algorithmus ausführen, unabhängig von physikalischen und sonstigen Einschränkungen. Insbesondere könnte er also eine Substitutionschiffre einfach dadurch lösen, daß er alle 26! Möglichkeiten ausprobiert; sofern nur in einem Fall sinnvoller Klartext entsteht, hat er die Lösung gefunden. Es ist klar, daß von einem realen Gegner keine größere Gefahr ausgeht, als von diesem idealisierten Gegner; ein Kryptosystem ist also sicher, wenn es gegen einen solchen Gegner sicher ist.

Um diese Sicherheit abschätzen zu können, müssen wir zunächst klären, was wir unter der in einem Klartext oder Kryptogramm enthaltene Information verstehen wollen; damit beschäftigt sich der nächste Paragraph.

§1: Die Entropie einer Quelle

Gegeben sei eine Nachricht aus N Buchstaben $x_1, \ldots, x_N$, die einem Alphabet A entstammen. Das Alphabet A kann dabei beispielsweise wie im vorigen Kapitel aus den 26 Buchstaben von „A“ bis „Z“ bestehen, es könnte sich aber auch um die 256 möglichen ASCII-Codes handeln oder um die Codewörter eines der alten Telegrammcodes für Geschäftsleute.

Wieviel Information enthält diese Nachricht für den Empfänger? Diese Information hängt offensichtlich ab von seinem Wissen über die Eigenschaften der Quelle: Bei einer Quelle, die deutschen Klartext sendet, wird er beispielsweise nach einem „C“ ein „H“ erwarten, so daß dieses, wenn es wirklich kommt, seine Information über die Nachricht nicht wesentlich erhöht; ein „I“ dagegen wäre unerwartet und damit informativ.

In den nächsten drei Abschnitten wollen wir dies genauer untersuchen für drei verschiedene statistische Modelle für die deutsche Sprache oder irgendwelchen anderen Klartext.

Als *Informationsgehalt* der Nachricht bezeichnen wir dabei die minimale Anzahl von Bits, die der Empfänger braucht, um die Nachricht zu speichern. Diese Anzahl hängt natürlich davon ab, welches Vorwissen er über die Nachricht hat; je mehr er darüber weiß, desto effizienter kann er sie abspeichern.

a) Ohne Vorwissen über die Quelle

Falls der Empfänger nichts über die statistische Verteilung der Buchstaben weiß, hat jeder Buchstabe dieselbe Wahrscheinlichkeit und enthält somit dieselbe Information. Es gibt daher $|A|^N$ verschiedene Möglichkeiten, von denen jede gleich wahrscheinlich ist.

Um einen einzigen Buchstaben binär zu speichern, brauchen wir $\lceil \log_2 |A| \rceil$ bit, wobei $\lceil x \rceil$ für die kleinste ganze Zahl $\geq x$ steht. Im Falle eines Alphabets aus 26 Buchstaben wären dies wegen

$$\log_2 26 \approx 4{,}70044$$

als fünf Bit.

Das ist aber noch nicht optimal: Wenn wir etwa die Buchstaben der Nachricht in Siebenergruppen zusammenfassen, brauchen wir pro Gruppe nicht 35, sondern nur 33 Bit, denn

$$7 \log_2 26 \approx 32{,}90308\,,$$

so daß pro Buchstabe nur noch

$$\frac{33}{7} \approx 4{,}71428$$

Bit benötigt werden. Mit noch größeren Gruppen kämen wir noch näher an $\log_2 26$ heran, und da es in diesem Kapitel nur ums Prinzip geht, nicht um Praktikabilität, wollen wir diese Zahl als Informationsgehalt pro Buchstabe betrachten.

Allgemein bezeichnen wir also den Zweierlogarithmus $N \log_2 |A|$ von $|A|^N$ als Informationsgehalt der Nachricht und $\log_2 |A|$ als den Informationsgehalt pro Buchstabe.

b) Mit Vorwissen über die Häufigkeitsverteilung

Falls der Empfänger die Häufigkeitsverteilung der Buchstaben kennt, weiß er bereits, wie oft jeder Buchstabe vorkommt: Ist p_i die Häufigkeit des i-ten Buchstaben, so kommt dieser $p_i N$-mal vor. Damit sind also alle Buchstaben der Nachricht bekannt; es geht „nur“ noch um ihre Reihenfolge. Im Prinzip gibt es $N!$ Möglichkeiten, die Reihenfolge von N Buchstaben festzulegen; hiervon sind allerdings alle diejenigen uninteressant, die nur Buchstaben miteinander vertauschen, die ohnehin schon gleich sind. Für den i-ten Buchstaben, der $p_i N$-mal vorkommt, sind dies $(p_i N)!$ Stück; tatsächlich gibt es also nur

$$\frac{N!}{(p_1 N)! \cdots (p_m N)!}, \qquad m = |A|$$

Möglichkeiten, die aus Sicht des Empfängers wieder allesamt gleich wahrscheinlich sind. Der Informationsgehalt der Nachricht ist deshalb gleich dem Zweierlogarithmus dieser Zahl.

Um diesen Logarithmus weiter auszurechnen, erinnern wir uns an die STIRLINGsche Formel, wonach für eine natürlich Zahl n gilt

$$\ln n! = n \ln n - n + \frac{\ln n}{2} + O(1)\,,$$

wobei $O(1)$ bedeuten soll, daß dieser Term für $n \to \infty$ beschränkt bleibt. Da Logarithmen zu verschiedenen Basen proportional zueinander sind, ist auch

$$\log_2 n! = n \log_2 n - \frac{n}{\ln 2} + \frac{\log_2 n}{2} + O(1)\,,$$

und der Informationsgehalt der Nachricht ist

$$\begin{aligned}
&N \log_2 N - (p_1 N) \log_2 (p_1 N) - \cdots - (p_m N) \log_2 (p_m N) \\
&- \frac{N}{\ln 2} + \frac{p_1 N}{\ln 2} + \cdots + \frac{p_m N}{\ln 2} \\
&+ \frac{\log_2 N}{2} - \frac{\log_2 (p_1 N)}{2} - \cdots - \frac{\log_2 (p_m N)}{2} + O(1) \\
&= N \log_2 N (1 - p_1 - \cdots - p_r) - N (p_1 \log_2 p_1 + \cdots + p_m \log_2 p_m) \\
&- \frac{N(1 - p_1 - \cdots - p_m)}{\ln 2} \\
&+ \frac{1 - m}{2} \log_2 N + O(1) \\
&= -N \sum_{i=1}^{m} p_i \log_2 p_i + \frac{1 - n}{2} \log_2 N + O(1)\,,
\end{aligned}$$

denn

$$1 - \sum_{i=1}^{m} p_i = 0\,.$$

Für $N \to \infty$ ist die Information pro Buchstabe also gleich

$$-\sum_{i=1}^{m} p_i \log_2 p_i\,.$$

Diese Zahl bezeichnen wir als die *Entropie* der Quelle. Sie sollte im allgemeinen kleiner sein als die oben berechneten $\log_2 m$ Bit. Für die 26 Buchstaben der deutschen Sprache mit ihren Häufigkeiten gemäß der Auszählung von *Dr. Katzenbergers Badereise* ist in der Tat

$$-\sum_{i=1}^{26} p_i \log_2 p_i \approx 4{,}04088 < \log_2 26 \approx 4{,}70044\,,$$

und auch allgemein haben wir das folgende

Lemma: Für m Zahlen $p_1, \ldots, p_m \in [0,1]$ mit $\sum_{i=1}^m p_i = 1$ gilt stets

$$0 \le -\sum_{i=1}^{m} p_i \log_2 p_i \le \log m\,;$$

dabei steht rechts genau dann ein Gleichheitszeichen, wenn alle p_i gleich $1/m$ sind.

Der *Beweis* erfolgt durch vollständige Induktion nach m.

Der Fall $m = 1$ ist trivial, denn dann muß $p_1 = 1$ sein, und die Behauptung gilt, da sowohl $\log_2 p_1$ also auch $\log_2 m$ verschwinden.

Für $m > 1$ können wir uns sofort auf den Fall $m = 1$ berufen, falls eine der Zahlen p_i gleich eins sein sollt. Interessant ist daher nur der Fall, daß mindestens ein p_i von eins verschieden ist, und dies sei o.B.d.A. das letzte, also p_m.

Wir betrachten die Summe $-\sum_{i=1}^m p_i \log_2 p_i$ zunächst nur unter der Nebenbedingung $p_m = q < 1$. Dann ist

$$\sum_{i=1}^{m-1} p_i = 1 - q\,,$$

mit

$$p_i^* = \frac{p_i}{1-q} \quad \text{ist also} \quad \sum_{i=1}^{m-1} p_i^* = 1\,.$$

Darau können wir die Induktionsannahme anwenden:

$$0 \le -\sum p_i^* \log_2 p_i^* \le \log_2(m-1)$$

mit Gleichheitszeichen rechts genau dann, wenn alle $p_i^* = \frac{1}{1-m}$ sind. Dies müssen wir mit $\sum_{i=1}^m p_i \log_2 p_i$ in Verbindung bringen; setzen wir also die Definition von p_i^* ein:

$$\begin{aligned}
\sum_{i=1}^{m-1} p_i^* \log_2 p_i^* &= \frac{1}{1-q}\sum_{i=1}^{m-1} p_i\big((\log_2 p_i - \log_2(1-q)\big)\\
&= \frac{1}{1-q}\sum_{i=1}^{m-1} p_i \log_2 p_i - \frac{1}{1-q}\sum_{i=1}^{m-1} p_i \log_2(1-q)\\
&= \frac{1}{1-q}\sum_{i=1}^{m-1} p_i \log_2 p_i - \frac{1}{1-q}(1-q)\log_2(1-q)\\
&= \frac{1}{1-q}\sum_{i=1}^{m-1} p_i \log_2 p_i - \log_2(1-q)\,.
\end{aligned}$$

Somit ist die hier interessante Summe

$$-\sum_{i=1}^{n} p_i \log p_i = -(1-q)\sum_{i=1}^{m-1} p_i^* \log_2 p_i^* - (1-q)\log_2(1-q) - q\log_2 q\,.$$

Die negative Summe über die p_i^* wird nach Induktionsannahme genau dann maximal, wenn alle $p_i^* = \frac{1}{m-1}$ sind; unter der Nebenbedinung $p_m = q$ ist der gesamte Ausdruck also höchstens gleich

$$(1-q)\log_2(m-1) - (1-q)\log_2(1-q) - q\log_2 q\,.$$

Das müssen wir nun noch noch in Abhängigkeit von q maximieren; die Ableitung nach q ist

$$\begin{aligned}
&-\log_2(m-1) + \log_2(1-q) + \frac{1-q}{1-q} - \log_2 q - \frac{q}{q}\\
= &-\log_2(m-1) + \log_2(1-q) - \log_2 q = \log_2 \frac{1-q}{(m-1)q}\,,
\end{aligned}$$

und dies wird genau dann gleich null, wenn der Bruch eins ist, d.h. für

$$1 - q = (m-1)q \qquad \text{oder} \qquad q = \frac{1}{m}\,.$$

In diesem Fall sind auch alle $p_i = 1/m$, und damit ist das Lemma bewiesen.

■

3) Mit Vorwissen über die Kontakthäufigkeiten

Wie wir bei der Kryptoanalyse von allgemeinen monoalphabetischen Substitutionen gesehen haben, ergeben die Kontakthäufigkeiten ein deutlich klareres Bild als die bloßen Buchstabenhäufigkeiten; auch dies läßt sich informationstheoretisch nachvollziehen:

Wir bezeichnen die Wahrscheinlichkeit dafür, daß der Buchstabe y auf den Buchstaben x folgt mit $p(x|y)$ und die Wahrscheinlichkeit für das Auftreten von x mit $p(x)$. Die Nachricht $x_1x_2\ldots x_N$ hat dann die Wahrscheinlichkeit

$$p(x_1\ldots x_N) = p(x_1)p(x_2|x_1)\cdots p(x_N|x_{N-1})\,.$$

Wir betrachten diese Nachricht formal als *einen* Buchstaben **x** aus dem Alphabet A^N und wissen dann aus dem vorigen Abschnitt, daß zur Übertragung eines solchen Buchstaben im Mittel

$$H_N = -\sum_{\mathbf{x}\in A^N} p(\mathbf{x})\log_2 p(\mathbf{x})$$

Bit notwendig sind. Dabei ist H_1 gerade die im vorigen Abschnitt berechnete Entropie für das Alphabet A.

Versuchen wir, H_N aus H_{N-1} zu berechnen! Dazu schreiben wir

$$\mathbf{x} = \mathbf{y}x_N \quad \text{mit} \quad \mathbf{y}\in A^{N-1}\,;$$

dann ist

$$p(\mathbf{x}) = p(x_1\ldots x_N) = p(\mathbf{y})p(x_N|x_{N-1})$$

und

$$\log_2 p(\mathbf{x}) = \log_2 p(\mathbf{y}) + \log_2 p(x_N|x_{N-1})\,,$$

also

$$p(\mathbf{x})\log_2 p(\mathbf{x}) = p(\mathbf{y})p(x_N|x_{N-1})\big(\log_2 p(\mathbf{y}) + \log_2 p(x_N|x_{N-1})\big)$$

und

$$\begin{aligned} H_N &= -\sum_{\mathbf{x}\in A^N} p(\mathbf{x})\log_2 p(\mathbf{x}) \\ &= -\sum_{\mathbf{x}\in A^N} p(\mathbf{y})\log_2 p(\mathbf{y})\cdot p(x_N|x_{N-1}) - \sum_{\mathbf{x}\in A^N} p(\mathbf{x})\log_2 p(x_N|x_{N-1})\,. \end{aligned}$$

Dabei ist

$$\begin{aligned} &-\sum_{\mathbf{x}\in A^N} p(\mathbf{y})\log_2 p(\mathbf{y})\cdot p(x_N|x_{N-1}) \\ &= -\sum_{\mathbf{y}\in A^N} p(\mathbf{y})\log_2 p(\mathbf{y}) \sum_{x_N\in A} p(x_N|x_{N-1})\,. \end{aligned}$$

Für jedes $x_{N-1}\in A$ und damit auch für jedes $\mathbf{y}\in A^{N-1}$ ist

$$\sum_{x_N\in A} p(x_N|x_{N-1}) = 1\,,$$

denn irgendein Buchstabe muß schließlich auf x_{N-1} folgen. Also ist

$$-\sum_{\mathbf{x}\in A^N} p(\mathbf{y})\log_2 p(\mathbf{y})\cdot p(x_N|x_{N-1}) = H_{N-1}$$

und

$$H_N = H_{N-1} + \Delta_N \quad \text{mit} \quad \Delta_N = -\sum_{\mathbf{x}\in A^N} p(\mathbf{x})\log_2 p(x_N|x_{N-1})\,.$$

Für $N = 2$ ist also

$$\Delta_2 = -\sum_{x_1x_2\in A^2} p(x_1x_2)\log_2 p(x_2|x_1)\,.$$

Für $N > 2$ ist nach Definition

$$p(x_1\ldots x_N) = p(x_1)p(x_2|x_1)\cdots p(x_N|x_{N-1})$$

und

$$\begin{aligned} \sum_{x_1\in A} p(x_1)p(x_2|x_1)\cdots p(x_N|x_{N-1}) &= \sum_{x_1\in A} p(x_1x_2)p(x_3|x_2)\cdots p(x_N|x_{N-1}) \\ &= p(x_2)p(x_3|x_2)\cdots p(x_N|x_{N-1})\,, \end{aligned}$$

denn

$$\sum_{x_1\in A} p(x_1x_2) = p(x_2)\,.$$

Also ist $\Delta_N = \Delta_{N-1}$, d.h. $\Delta_N = \Delta_2$ für alle $N \geq 2$.

Damit ist gezeigt:

Lemma: $H_N = H_1 + (N-1)H_\infty$, wobei H_1 die Entropie ist und

$$H_\infty = -\sum_{x_1x_2\in A^2} p(x_1x_2)\log_2 p(x_2|x_1)\,.$$

■

Asymptotisch, für $N\to\infty$, ist also der Informationsgehalt pro Buchstabe gleich H_∞.

Auf der Basis von JEAN PAULS Roman *Dr. Katzenbergers Badereise* erhält man

$$H_2 \approx 7{,}43853 \text{ Bit}\,.$$

Da wir

$$H_1 \approx 4{,}04088 \text{ Bit}$$

bereits kennen, liefert das Lemma

$$H_\infty \approx 3{,}39765 \text{ Bit/Buchstabe}\,.$$

Schließlich kann man auch Trigrammhäufigkeiten und so weiter im statistischen Modell berücksichtigen; hierfür habe ich nur die Entropie H_n pro n-Gramm und, als ein n-tel davon, die pro Buchstabe ausgerechnet:

n	H_n	H_n/n
1	4,04088	4,04088
2	7,43853	3,71926
3	10,37095	3,45698
4	12,81915	3,20479

Man schätzt, daß bei voller Ausnutzung aller statistischer Strukturen der Sprache schließlich noch etwa 30% der $\log_2 26$ Bit/Buchstabe notwendig sind.

§2: Kryptoanalyse durch den Bayesschen Gegner

a) Die Grundidee

Gegeben seien ein Alphabet A (über das wir im Augenblick keine weitere Annahmen machen) und ein Kryptosystem

$$\{T_s\colon A\to A \mid s\in S\}\,.$$

Dabei sind die Abbildungen $T_s\colon A\to A$ Verschlüsselungsverfahren, die von einem Schlüssel $s\in S$ abhängen.

Bei der Einschätzung der Sicherheit dieses Systems gehen wir, aus den im vorigen Kapitel genannten Gründen, davon aus, daß der Gegner das Kryptosystem kennt; wir beurteilen die als nur auf der Grundlage des Schlüssels.

Ein Gegner muß zur Entschlüsselung einer Nachricht also „nur“ noch den Schlüssel finden; ein gutes Kryptosystem versucht, ihm dieses so schwer wie möglich zu machen.

Wie eingangs erwähnt, betrachten wir in diesem Paragraphen einen „idealisierten“ Gegner, der über unbegrenzte Rechenmöglichkeiten verfügt, so daß er jeden endlichen Algorithmus, unabhängig von der Zahl der notwendigen Rechenschritte, in kurzer Zeit durchführen kann; diesen Gegner bezeichnet man in der Kryptologie als den BAYESschen Gegner.

Sein Name ist abgeleitet von dem englischen Theologen THOMAS BAYES (1702–1761), der als erster das Entscheidungsprinzip formulierte, wonach unter mehreren möglichen Hypothesen diejenige als wahr anzusehen sei, die vor dem Hintergrund seines vorhandenen Wissens die größte Wahrscheinlichkeit hat; auf die Kryptanalyse angewandt bedeutet dies, daß man sich unter allen möglichen Schlüsseln für den entscheidet, bei dem Klartext entsteht, der gemäß dem jeweiligen statistischen Modell für Klartexte die größte Wahrscheinlichkeit hat.

b) Einführendes Beispiel

Um zu sehen, wie das funktioniert, beobachten wir den BAYESschen Gegner bei der Entschlüsselung einer CAESAR-Chiffre – bei komplizierteren Chiffren macht sich zu schnell bemerkbar, daß er deutlich schneller rechnen kann als wir.

Angenommen, er empfängt das schon im letzten Kapitel betrachtete Kryptogramm QJULIILPPRUJHQJUDXH Q .

Unter der Annahme, daß es sich hier um eine CAESAR-Chiffre handelt, ist es nicht nur für den BAYESschen Gegner, sondern auch für uns leicht möglich, alle 26 möglichen Klartexte sowie deren jeweiligen Wahrscheinlichkeiten zu berechnen. Zu jedem dieser Klartexte berechnen wir dann dessen Wahrscheinlichkeit in der deutschen Sprache.

Um zu sehen, wie viele Chiffretextbuchstaben zur Entschlüsselung notwendig sind, betrachten wir nicht das gesamte Kryptogramm auf einmal, sondern wir betrachten in der i-ten Zeile der folgenden Tabelle nur die ersten i Buchstaben. Für diese ist jeweils die wahrscheinlichstse Entschlüsselung angegeben, zunächst auf der Basis der Buchstabenwahrscheinlichkeiten, dann auf Basis der Kontaktwahrscheinlichkeiten. Die angegebenen Wahrscheinlichkeiten sind der Übersichtlichkeit halber nicht die absoluten Klartextwahrscheinlichkeiten der jeweiligen Entschlüsselungen, sondern die relativen: Wir dividieren also durch die Summe der Wahrscheinlichkeiten der 26 möglichen Klartexte. Das Ergebnis ist in Prozent angegeben:

E	18,498	E	18,498
RE	25,379	ER	54,251
NAT	25,832	ERK	40,153
NATE	71,640	NATE	77,103
ANGRI	50,664	ANGRI	58,662
NATEV S	58,718	ANGRI F	95,260
NATEV SS	78,445	ANGRI FF	99,281
NATEV SSV	33,482	ANGRI FFI	99,311
ANGRI FFIM	42,358	ANGRI FFIM	99,974
ANGRI FFIMM	31,116	ANGRI FFIMM	99,997
FSLWN KKNRR T	35,311	ANGRI FFIMM O	100,000
ANGRI FFIMM OR	43,708	ANGRI FFIMM OR	100,000
ANGRI FFIMM ORG	33,840	ANGRI FFIMM ORG	100,000
ANGRI FFIMM ORGE	65,818	ANGRI FFIMM ORGE	100,000
ANGRI FFIMM ORGEN	84,217	ANGRI FFIMM ORGEN	100,000
ANGRI FFIMM ORGEN G	73,487	ANGRI FFIMM ORGEN G	100,000
ANGRI FFIMM ORGEN GR	51,676	ANGRI FFIMM ORGEN GR	100,000
NATEV SSVZZ BETRA TEN	62,256	ANGRI FFIMM ORGEN GRA	100,000
NATEV SSVZZ BETRA TENH	66,592	ANGRI FFIM MORGENG RAU	100,000
ANGRI FFIMM ORGEN GRAUE	57,125	ANGRI FFIMM ORGEN GRAUE	100,000
ANGRI FFIMM ORGEN GRAUE N	70,387	ANGRI FFIMM ORGEN GRAUE N	100,000

Mit Buchstabenhäufigkeiten allein schwankt der BAYESsche Gegner also noch ziemlich lange zwischen zwei Schlüsseln, beachtet er auch Kontakthäufigkeiten, hat er schon nach fünf Buchstaben die richtige Entschlüsselnung und ab dem sechsten Buchstaben weiß er das mit einer Wahrscheinlichkeit von über 95%.

Als nächstes Kryptogramm betrachten wir PWDUYFSFQDXJ . Hier ergeben sich folgende wahrscheinlichste Entschlüsselungen:

E	18,498	E	18,498
EL	20,611	EL	27,369
ELS	36,252	ELS	61,022
ZGNE	20,524	MTAR	27,712
ZGNEI	51,027	MTARV	32,544
NUBSW D	25,059	DKRIM T	46,246
ELSJN UH	42,422	KRYPT AN	63,235
ELSJN UHU	44,039	KRYPT ANA	72,856
DKRIM TGTE	81,297	KRYPT ANAL	53,404
DKRIM TGTER	85,742	DKRIM TGTER	99,705
DKRIM TGTER L	87,118	DKRIM TGTER L	98,089
ZGNEI PCPAN HT	75,437	KRYPT ANALY SE	99,797

Hier gelingt die Entschlüsselung also nur mittels Kontakthäufigkeiten, und auch da nur dank des guten Doktors STRYKIUS, der dafür sorgt, daß das Digramm „RY“ in *Dr. Katzenbergers Badereise* mit positiver Wahrscheinlichkeit auftritt: Bei vielen anderen Referenztexten tritt diese Buchstabenkombination nie auf, so daß die Wahrscheinlichkeit eines jeden Worts, daß sie enthält auf null gesetzt würde.

Man muß freilich beachten, daß der ideale BAYESsche Gegner nicht mit der Auszählung eines einzelnen Texts arbeitet, sondern *die* korrekte Wahrscheinlichkeitsverteilung kennt – wie immer diese auch definiert sein mag.

Es ist natürlich kein Zufall, das der BAYESsche Gegner einbuchstabige Nachrichten immer als „E“ entschlüsselt: Für einbuchstabige Texte gibt es nach beiden Wahrscheinlichkeitsmodellen keine bessere Alternative. Entsprechend einfach werden auch Wörter erkannt, die mit „E“ beginnen und viele häufige Buchstaben und Buchstabenkombinationen enthalten

wie beispielsweise im Kryptogramm JSIQNHMPJNY ,
wo beide Modelle sofort den richtigen Schlüssel finden:

E	18,498	E	18,498
EN	43,447	EN	66,315
END	49,997	END	96,907
ENDL	60,346	ENDL	93,935
ENDLI	89,043	ENDLI	99,306
ENDLI C	89,309	ENDLI C	99,995
ENDLI CH	96,787	ENDLI CH	100,000
ENDLI CHK	92,988	ENDLI CHK	100,000
ENDLI CHKE	98,564	ENDLI CHKE	100,000
ENDLI CHKEI	99,529	ENDLI CHKEI	100,000
ENDLI CHKEI T	99,939	ENDLI CHKEI T	100,000

Auch nicht zu exotische Substantive mit Artikel sind meist problemlos:
IJWYNXHM führt zu

E	18,498	E	18,498
DE	22,057	DE	38,733
DER	38,490	DER	87,654
DERT	51,481	DERT	96,107
DERTI	75,471	DERTI	99,393
DERTI S	88,877	DERTI S	99,717
DERTI SC	87,934	DERTI SC	99,999
DERTI SCH	90,031	DERTI SCH	100,000

und PUQEFGQTXQ zu

E	18,498	E	18,498
IN	23,056	CH	40,135
DIE	41,643	DIE	43,465
DIES	64,018	DIES	70,969
DIEST	69,681	DIEST	90,466
DIEST U	53,199	DIEST U	76,378
DIEST UE	83,884	DIEST UE	91,854
DIEST UEH	87,958	DIEST UEH	97,642
DIEST UEHL	96,798	DIEST UEHL	99,523
DIEST UEHLE	99,372	DIEST UEHLE	99,978

Nur der sächliche Artikel ist etwas problematischer: EBTIBVT
wird entschlüsselt als

E	18,498	E	18,498
HE	29,769	HE	28,458
EBT	20,051	DAS	47,768
EBTI	34,481	DASH	32,720
HEWLE	45,218	DASHA	75,568
DASHA U	47,297	DASHA U	98,212
DASHA US	57,475	DASHA US	99,609

Hier gibt es also Schwierigkeiten mit den Buchstabenhäufigkeiten allein, da der Text aus eher nicht so häufigen Buchstaben zusammengesetzt ist. Dieses Problem tritt auch bei anderen Texten auf wie etwa HAQQNAA mit

E	18,498	E	18,498
LE	20,611	UN	25,577
LEU	29,391	UND	69,721
LEUU	24,785	UNDD	88,852
LEUUR	44,254	UNDDA	97,663
LEUUR E	68,174	UNDDA N	99,279
LEUUR EE	81,555	UNDDA NN	99,890

und selbst bei bei so zentralen deutschen Wörtern wieYRVGX HYGHE mit

E	18,498	E	18,498
LE	20,611	UN	25,577
HAE	33,485	LEI	39,068
LEIT	54,851	LEIT	71,632
UNRCT	43,649	LEITK	41,025
UNRCT D	51,689	VOSDU E	56,410
UNRCT DU	59,272	LEITK UL	71,093
UNRCT DUC	47,724	LEITK ULT	98,296
UNRCT DUCD	48,308	LEITK ULTU	91,777
UNRCT DUCDA	50,558	LEITK ULTUR	97,113

oder FNHZNTRA mit

E	18,498	E	18,498
EM	15,918	EM	21,504
LTN	19,493	CKE	37,812
AICU	20,640	SAUM	70,182
AICUI	28,589	SAUMA	97,483
AICUI O	29,941	SAUMA G	98,658
SAUMA GE	52,561	SAUMA GE	99,995
SAUMA GEN	78,116	SAUMA GEN	100,000

Zusammenfassend können wir festhalten, daß dem BAYESschen Gegner zumindest mit Kontakthäufigkeiten bereits wenige Buchstaben zur korrekten Entschlüsselung reichen. Beachtet man noch, daß sich weder ein BAYESsche Gegner noch ein realer Kryptanalytiker auf Kontakthäufigkeiten beschränken muß, sondern auch Sprachkenntnisse einsetzt, die sich auf Trigramme, Worte und so weiter beziehen, kommen sie mit noch weniger Buchstaben aus, als obige Beispiele nahelegen. Während des ersten Weltkriegs aktive Militärkryptographen gehen davon aus, daß man nur knapp zwei Buchstaben pro Alphabet (das heißt also pro CAESAR-Chiffre) benötigt, um eine VIGENÈRE-Chiffre zu brechen.

Als Kuriositär am Rande sei noch erwähnt, daß der BAYESsche Gegner auch ohne Chiffretext bereits wahrscheinlichste „Entschlüsselungen“ findet: Falls er nur mit Buchstabenhäufigkeiten arbeitet, ist das natürlich eine Folge von lauter „E“s, schon bei Kontakthäufigkeiten ergeben sich aber interessantetere Ergebnisse, die hier bis zur Nachrichtenlänge acht aufgeführt sind:

n	Text	Wahrscheinlichkeit
1	e	0.184945672750
2	en	0,042051665485
3	ich	0,011662006378
3	ich	0,011662006378
4	ende	0,003537155688
5	eiche	0,000825895113
6	endich	0,000201549003
7	endende	0,000067649431
8	eichende	0,000015795555

c) Der allgemeine Ansatz

Der BAYESsche Gegner braucht als erstes Wissen oder zumindest Annahmen über den Inhalt der Nachrichten: In den obigen Beispielen handelte es sich um deutschen Klartext, verschlüsselt auf der Grundlage von 26 Buchstaben ohne Zwischenräume und Satzzeichen; das ist natürlich nicht mehr der typische Fall für heutige Kryptanalyse: Hier geht es eher um Dokumente von Textverarbeitungs-, Tabellenkalkulations- oder Datenbankprogrammen oder um ausführbare Programme für ein bekanntes Betriebssystem oder ähnliches.

In jedem Fall verschafft es sich als erstes durch Auszählen eine Wahrscheinlichkeitsverteilung für die möglichen Klar„texte“ der Länge N und ordnet dadurch jedem solchen Klartext W eine

Klartextwahrscheinlichkeit $p_K(W)$

zu; diese kann beispielsweise so wie im letzten Paragraphen über Buchstabenhäufigkeiten oder über Kontakthäufigkeiten definiert sein, aber auch viele andere Ansätze sind möglich.

Außerdem ordnet er jedem Schlüssel s aud dem Schlüsselraum S eine

Schlüsselwahrscheinlichkeit $p_S(s)$

zu. Bei einem gut gemanagten Kryptosystem sollten alle Schlüssel mit gleicher Wahrscheinlichkeit auftreten, so daß

$$p_S(s) = \frac{1}{\#S} \quad \text{für alle } s \in S$$

ist, aber in der Praxis kann der Gegner oft große Vorteile aus der Tatsache ziehen, daß dies nicht der Fall ist. (Wie zufällig sind Ihre Paßwörter?)

Falls nun ein Chiffretext $C \in A^N$ aufgefangen wird, muß der BAYESsche Gegner berechnen, wie wahrscheinlich jeder Schlüssel $s \in S$ vor dem Hintergrund dieses Chiffretexts ist. Dazu müssen bedingte Wahrscheinlichkeiten berechnet werden, also muß man zunächst die Wahrscheinlichkeit des Chiffretexts $C \in A^N$ kennen.

C entsteht aus einem Klartext $W \in A^N$ durch Verschlüsselung mit einem Schlüssel $s \in S$; da der Schlüssel vom Klartext unabhängig sein sollte, ist die Wahrscheinlichkeit für das Zusammentreffen von Klartext W und Schlüssel s gleich

$$p_K(W) \cdot p_S(s) .$$

Die Wahrscheinlichkeit, einen bestimmten Chiffretext C zu empfangen, ist also *a priori* gleich

$$p_{Ch}(C) = \sum_{\substack{(W,s) \in A^N \times S \\ T_s(W)=C}} p_K(W) \cdot p_S(s) .$$

Die Wahrscheinlichkeit, daß C auftritt und mit Schlüssel $s \in S$ verschlüsselt wurde, ist entsprechend gleich

$$p_{Ch,S}(C,s) = \sum_{\substack{W \in A^N \\ T_s(W)=C}} p_K(W) \cdot p_S(s) ,$$

wobei die Summe hier im allgemeinen wegen der Injektivität von T_s nur einen Summanden hat. (T_s muß nicht injektiv auf A^N sein; es muß nur injektiv auf der Teilmenge der tatsächlich vorkommenden Nachrichten sein.)

Die eigentlich interessante Wahrscheinlichkeit, die Wahrscheinlichkeit, daß ein gegebener Chiffretext C durch Verschlüsselung mit $s \in S$ entstanden ist, berechnet sich nun als

$$p_{S|Ch}(s|C) = \frac{p_{Ch,S}(C,s)}{p_{Ch}(C)} .$$

Der BAYESsche Gegner entscheidet sich bei seinem Entschlüsselungsversuch für einen Schlüssel, der diese bedingte Wahrscheinlichkeit maximal macht. Um einen solchen Schlüssel zu bestimmen, muß er $p_{S|Ch}(s|C)$ möglicherweise für alle Schlüssel $s \in S$ berechnen, jedoch ist dies angesichts seiner unbegrenzten Rechenfähigkeit kein Problem.

Solange es sich nur um eine Nachricht dreht, wird der BAYESsche Gegner nicht in erster Linie am Schlüssel interessiert sein, sondern am wahrscheinlichsten Klartext. Hier ist entsprechend die Wahrscheinlichkeit für das Zusammentreffen von Klartext W und Chiffretext C gleich

$$p_{Ch,K}(C,W) = \sum_{\substack{s \in S \\ T_s(W)=C}} p_K(W) \cdot p_S(s)$$

und die bedingte Wahrscheinlichkeit für Klartext W, nachdem Chiffretext C aufgefangen wurde, ist

$$p_{K|Ch}(W|C) = \frac{p_{Ch,K}(C,W)}{p_{Ch}(C)} .$$

Definition: Eine BAYESsche Entscheidungsfunktion ist eine Familie von Abbildungen

$$\delta_N : A^N \to A^N$$

mit der Eigenschaft, daß für jeden Chiffretext $C \in A^N$ gilt:

$$p_{K|Ch}(\delta(C)|C) = \max_{W \in A^N} p_{K|Ch}(W|C) .$$

Weniger formal ausgedrückt: Für jeden Chiffretext $C \in A^N$ ist $\delta(C)$ ein Klartext mit höchstmöglicher Wahrscheinlichkeit – idealerweise *der* Klartext mit höchstmöglicher Wahrscheinlichkeit, allerdings könnte es auch mehrere Klartexte mit gleicher Wahrscheinlichkeit geben, so daß $\delta(C)$ im allgemeinen durch die obige Definition nicht eindeutig bestimmt ist. Trotzdem ist klar, daß man nichts besseres tun kann, als mit einer solchen BAYESschen Entscheidungsfunktion zu arbeiten; wenn dies zu nichts führt, ist ein Kryptosystem sicher.

d) Perfekte Sicherheit

Am allersichersten ist ein Kryptosystem, wenn der BAYESsche Gegner aus dem aufgefangenen Chiffretext C *überhaupt keine* Information gewinnen kann, die über seine ursprünglich gewählte Wahrscheinlichkeitsfunktion p_K hinausgeht, wenn also die bedingten Wahrscheinlichkeiten, die er in Kenntnis von C errechnet, gleich den ursprünglichen Wahrscheinlichkeiten sind:

Definition: Ein Kryptosystem $\{T_s \mid s \in S\}$ über dem Alphabet A hat *perfekte Sicherheit* für Nachrichten der Länge N, wenn für jeden Chiffretext $C \in A^N$ und jeden Klartext $W \in A^N$ gilt:

$$p_{K|Ch}(W|C) = p_K(W) .$$

Wir kennen bereits ein System mit perfekter Sicherheit zumindest für Nachrichten der Länge eins, nämlich die klassischen CAESAR-Substitutionen: Da hier, je nach Schlüssel, jeder Buchstabe in jeden anderen verschlüsselt werden kann, gibt der erste Buchstabe des Chiffretexts keinerlei Information über den Klartext.

Dies läßt sich ausbauen zu einem System mit perfekter Sicherheit für Nachrichten der Länge N, indem wir einfach für jeden der N Buchstaben der Nachricht ein neues CAESAR-System wählen, d.h. wir nehmen eine polyalphabetische VIGENÈRE-Chiffre mit einem Schlüssel der Länge N. Wieder läßt sich je nach Schlüssel *jeder* Klartext der Länge N in einen gegebenen Chiffretext $C \in A^N$ transformieren; bei zufällig gewählten Schlüsseln gewinnt der Gegner also keinerlei Information über den Klartext.

Solche Systeme wurden unter dem Namen *one time pad* beispielsweise während des zweiten Weltkriegs für die Kommunikation zwischen Widerstandskämpfern im besetzten Frankreich und London benutzt: Eine Maschine in London produzierte zufällige Buchstabenfolgen und druckte diese in zwei Exemplaren, die jeweils zu einem Block (*engl.* pad) zusammengebunden wurden. Der Sender benutzte zur Verschlüsselung seiner ersten Nachricht das erste Blatt, warf dieses dann weg und benutzte für die zweite Nachricht das zweite Blatt *usw.;* bei nicht zu vielen Nachrichten und der Möglichkeit, regelmäßig Blöcke per Kurier nach Frankreich zu bringen, war dies ein perfektes Verfahren. Mit Magnetbändern anstelle von Blöcken wird es heute noch benutzt zur Sicherung des „roten Telephons“ zwischen Washington und Moskau.

Für große Kommunikationsnetzwerke, etwa im Bankenbereich, ist ein solches Verfahren nicht praktikabel, da jeder Teilnehmer mit jedem anderen vorher riesige Schlüssel austauschen müßte. Auch ist die Erzeugung *wirklich* zufälliger Buchstaben- oder Zahlenfolgen nicht ganz einfach; die typischen Zufallsgeneratoren der gängigen Betriebssysteme und Compiler sind kryptographisch extrem schwach und können für solche Zwecke nicht verwendet werden.

Ideal wäre deshalb perfekte Sicherheit mit deutlich kleineren Schlüssellängen, aber dies ist leider nur sehr eingeschränkt möglich:

Satz: Falls ein Kryptosystem $\{T_s \mid s \in S\}$ perfekte Sicherheit für Nachrichten der Länge N hat, ist die Anzahl der Schlüssel $s \in S$ mindestens gleich der Anzahl der möglichen (d.h. mit Wahrscheinlichkeit $p_K(W) > 0$ auftretenden) Klartexte der Länge N.

Beweis: K^+ sei die Menge aller Klartexte aus A^N, die mit positiver Wahrscheinlichkeit auftreten, und C^+ sei die Menge der mit positiver Wahrscheinlichkeit auftretenden Chiffretexte. Für jedes feste $W \in K^+$ und jedes $C \in C^+$ muß es dann einen Schlüsses $s \in S$ geben, so daß $T_s(W) = C$ ist; denn sonst wäre $p_{K|Ch}(W|C) = 0$, aber $p_K(W) > 0$; der Gegner könnte also aus dem Auftreten von C Information gewinnen. Damit muß es mindestens so viele Schlüssel geben, wie es Chiffretexte in C^+ gibt. Für jeden Schlüssel $s \in S$ ist aber T_s eine injektive Abbildung von K^+ nach C^+, d.h. es muß erst recht so viele Schlüssel geben, wie es mögliche Klartexte gibt. ■

e) Die Mehrdeutigkeit eines Schlüssels

In der Praxis ist es nicht unbedingt notwendig, daß der Gegner aus einem aufgefangenen Chiffretext *überhaupt keine* Information ziehen kann; es reicht, wenn er *nicht genügend* Information bekommt.

Als nächstes wollen wir daher abschätzen, wieviel Information der (besonders gefährliche) BAYESsche Gegner aus einem Chiffretext ziehen kann. Besonders wichtig ist der Schutz des Schlüssels s, denn sobald dieser bekannt ist, können bis zum nächsten Schlüsselwechsel alle Nachrichten entschlüsselt werden.

Die Information, die der Gegner zur Rekonstruktion des Schlüssels gewinnen muß, läßt sich nach den Vorarbeiten des letzten Paragraphen leicht quantifizieren: Es ist die Entropie

$$H_S = -\sum_{s\in S} p_S(s) \log_2 p_S(s) .$$

Das Lemma über die Maximalität der Entropie bei Gleichverteilung zeigt also noch einmal die eigentlich auch so selbstverständliche Tatsache, daß ein Kryptosystem umso sicherer ist, je mehr die Wahrscheinlichkeitsverteilung der Schlüssel einer Gleichverteilung entspricht.

Der BAYESsche Gegner kennt die Entropie

$$H_{Ch} = -\sum_{C \in A^N} p_{Ch}(C) \log_2 p_{Ch}(C)$$

der Menge aller Chiffretexte der Länge N; nachdem er eine Nachricht C aufgefangen hat, hat er im Mittel diese Information gewonnen.

Um daraus Informationen über den Schlüssel zu erhalten, berechnet er zunächst die Entropie der Menge aller Paare (C, s):

$$\begin{aligned}
H_{Ch,S} &= -\sum_{(C,s) \in A^N \times S} p_{Ch,S}(C, s) \log_2 p_{Ch,S}(C, s) \\
&= -\sum_{(C,s) \in A^N \times S} p_{Ch,S}(C, s)\bigl(\log_2 p_{Ch}(C) + \log_2 p_{S|Ch}(s|C)\bigr) \\
&= -\sum_{(C,s) \in A^N \times S} p_{Ch,S}(C, s) \log_2 p_{Ch}(C) \\
&\quad\; -\sum_{(C,s) \in A^N \times S} p_{Ch,S}(C, s) \log_2 p_{S|Ch}(s|C) \\
&= -\sum_{C \in A^N} p_{Ch}(C) \log_2 p_{Ch}(C) \\
&\quad\; -\sum_{(C,s) \in A^N \times S} p_{Ch,S}(C, s) \log_2 p_{S|Ch}(s|C) \\
&= H_{Ch} - \sum_{(C,s) \in A^N \times S} p_{Ch,S}(C, s) \log_2 p_{S|Ch}(s|C)\,.
\end{aligned}$$

Die letzte Summe in der letzten Zeile bezeichnen wir als *bedingte Entropie* oder *Mehrdeutigkeit*

$$H_{S|Ch} \underset{\text{def}}{=} -\sum_{(C,s) \in A^N \times S} p_{Ch,S}(C, s) \log_2 p_{S|Ch}(s|C)$$

des Schlüssels bei vorliegendem Chiffretext; diese Information muß sich der Gegner verschaffen, um den Schlüssel zu bestimmen. Sobald $H_{S|Ch} = 0$ ist, kennt er den Schlüssel, denn da in der obigen Summe alle Summanden kleiner oder gleich null sind, kann $H_{S|Ch}$ nur dann gleich null sein, wenn jeder einzelne Summand verschwindet, wenn also für jeden Schlüssel $s \in S$ entweder $p(C, s) = 0$ ist, womit der Schlüssel mit Sicherheit ausgeschlossen ist, oder $\log_2 p(s|C) = 0$, d.h. $p(s|C) = 1$, womit der Schlüssel s mit Sicherheit richtig ist.

Um die Information abzuschätzen, die der Gegner maximal aus einem Chiffretext gewinnen kann, müssen wir also $H_{S|Ch}$ berechnen. Für ein beliebiges Kryptosystem und beliebige Wahrscheinlichkeitsverteilung der Klartexte können wir offensichtlich nichts konkreteres hinschreiben als die definierende Summe. Die für die gesamte Informationstheorie grundlegende Idee von CLAUDE SHANNON war es, spezifischere Aussagen nicht über ein spezielles, sondern über das *durchschnittliche* Kryptosystem zu machen; diese Aussagen sollten dann *im wesentlichen* für die meisten Kryptosysteme gelten.

f) Randomisierung

Um von Durchschnitten reden zu können, müssen wir zunächst einige Annahmen machen über das Verhalten der Klartexte. Wir betrachten wieder Klartexte über einem Alphabet A aus n Buchstaben; für $x \in A$ sei $p(x)$ die Wahrscheinlichkeit, mit der der Buchstabe x auftritt, und für $(x, y) \in A^2$ sei $p(x|y)$ die bedingte Wahrscheinlichkeit dafür, daß er auf ein y folgt.

Wir hatten bislang immer angenommen, daß etwa die Buchstabenhäufigkeiten, die wir durch Auszählen eines hinreichend langen Texts erhalten, in hinreichend guter Näherung mit denen übereinstimmen, die wir in einem gegebenen Klartext vorfinden. Dies müssen wir für die folgenden Überlegungen etwas präziser fassen:

Definition: Für einen gegebenen Klartext $W = w_1 \dots w_N \in A^N$ sei

$$m_x(W) = \#\{i \le N \mid w_i = x\}$$

und

$$m_{xy}(W) = \#\{i \le N-1 \mid w_i = x \text{ und } w_{i+1} = y\}\,.$$

$m_x(W)$ und $m_{xy}(W)$ zählen also, wie oft der Buchstabe x *bzw.* das Buchstabenpaar xy in W vorkommen. Idealerweise sollte also für einen hinreichend langen Text W und zwei Buchstaben $x, y \in A$ gelten

$$p(x) \approx \frac{m_x(W)}{N} \qquad \text{und} \qquad p(x|y) \approx \frac{m_{yx}(W)}{m_y(W)} .$$

Exakt kann das natürlich schon aus zahlentheoretischen Gründen nicht gelten, aber wir wollen verlangen, daß die Abweichung für große N mit sehr kleiner Wahrscheinlichkeit größer als eine vorgebbare Schranke δ ist:

Definition: Eine Quelle für Klartexte heißt *ergodisch,* wenn es zu zwei beliebig vorgebbaren reellen Zahlen $\delta, \varepsilon > 0$ stets eine natürliche Zahl N_0 gibt, so daß für alle $N \geq N_0$ und alle $x \in A$ für Texte $W \in A^N$ gilt

$$p\left(\left|\frac{m_x(W)}{N} - p(x)\right| > \delta\right) < \varepsilon .$$

Unsere Ansätze zur Kryptoanalyse klassischer Systeme beruhten somit stets darauf, daß wir annahne, daß die deutsche Sprache eine ergodische Quelle für Klartexte ist; der Erfolg bei unseren Entschlüsselungsversuchen spricht dafür, daß diese Annahme nicht garzu falsch sein sollte. In der Tat zeigt man in der Stochastik, daß jede Quelle, bei der es zu jedem Buchstabenpaar (x, y) einen Text gibt, in dem y irgendwo hinter x steht, ergodisch ist. Da wohl niemand bezweifelt, daß es solche Texte zu jedem Buchstabenpaar gibt (für (q, y) etwa **q**uer durch **Zy**pern), kann so die Ergodizität auch bewiesen werden.

Wir wollen im folgenden Aussagen machen über die *durchschnittliche* ergodische Quelle. Es ist klar, daß wir keine Chance haben, alle ergodischen Quellen zu bestimmen und dann einen Mittelwert zu bilden; der folgende Satz von SHANNON zeigt aber, daß wir durch Zusammenfassen der Buchstaben zu hinreichend langen Wörtern erreichen können, daß die Wahrscheinlichkeitsverteilung sehr einfach ist: Die Wörter zerfallen in zwei Klassen S und O, so daß Wörter aus R (wie *rar*) praktisch nie vorkommen, während die restlichen Wörter (aus O wie *oft*) alle praktisch dieselbe Wahrscheinlichkeit haben. Aus technischen Gründen arbeiten wir nicht mit Wahrscheinlichkeiten, sondern mit der Entropie:

Die buchstabenweise Entropie der Quelle ist, wenn wir wie in §1, *3)* von Kontakthäufigkeiten ausgehen, gleich

$$H = -\sum_{x \in A} \sum_{y \in A} p(x) p(y|x) \log_2 p(y|x) ,$$

und wir fordern, daß sich diese Entropie gleichmäßig auf die $W \in O$ verteilen soll:

Satz: Für eine ergodische Quelle gibt es zu zwei beliebig vorgebbaren reellen Zahlen $\varepsilon, \eta > 0$ stets ein $N_0 \in \mathbb{N}$, so daß A^N für $N \geq N_0$ als Vereinigung zweier Teilmengen O und S geschrieben werden kann mit den Eigenschaften

1.) $p(W \in R) < \varepsilon$

2.) $\left|\frac{-\log_2 p(W)}{N} - H\right| < \eta \quad$ für alle $\quad W \in O$.

Zum *Beweis* wählen wir zunächst willkürlich eine natürliche Zahl N und eine reelle Zahl $\delta > 0$ und setzen dann

$$O \underset{\text{def}}{=} \left\{ W \in A^N \;\middle|\; \begin{array}{l} p(W) > 0 \text{ und für alle } x, y \in A \text{ ist} \\ \left|m_{xy}(W) - Np(x)p(y|x)\right| < N\delta \end{array} \right\}$$

und $R = A^N \setminus O$. Für $W \in O$ können wir dann $m_{xy}(W)$ schreiben als

$$m_{xy}(W) = Np(x)p(y|x) + N\delta_{xy} \qquad \text{mit} \qquad \left|\delta_{xy}\right| < \delta$$

und erhalten damit für $W = w_1 \dots w_N$

$$\begin{aligned} p(W) &= p(w_1) \prod_{i=2}^{N} p(w_i|w_{i-1}) \\ &= p(w_1) \prod_{x \in A} \prod_{y \in A} p(y|x)^{m_{xy}(W)} \\ &= p(w_1) \prod_{x \in A} \prod_{y \in A} p(y|x)^{Np(x)p(y|x) + N\delta_{xy}} . \end{aligned}$$

Logarithmieren führt auf

$$\begin{aligned} -\log_2 p(w) &= -\log_2 p(w_1) - N \sum_{x\in A}\sum_{y\in A} p(x)p(y|x)\log_2 p(y|x) \\ &\quad - N\sum_{x\in A}\sum_{y\in A} \delta_{xy}\log_2 p(y|x) \\ &= -\log p(w_1) + NH - N\sum_{x\in A}\sum_{y\in A}\delta_{xy}\log_2 p(y|x) \end{aligned}$$

und somit zur Ungleichung

$$\begin{aligned} \left|\frac{-\log_2 p(W)}{N} - H\right| &= \left|\frac{-\log_2 p(W)}{N} - \sum_{x\in A}\sum_{y\in A}\delta_{xy}\log_2 p(y|x)\right| \\ &< \left|\frac{-\log_2 p(W)}{N} + \delta\sum_{x\in A}\sum_{y\in A} -\log_2 p(y|x)\right| . \end{aligned}$$

Wählt man N hinreichend groß und δ hinreichend klein, so läßt sich dieser Ausdruck offensichtlich kleiner als jedes vorgegebene $\eta > 0$ machen, so daß die Eigenschaft 2.) erfüllt ist.

Für die Eigenschaft 1.) müssen wir nachrechnen, mit welcher Wahrscheinlichkeit eine Nachricht $W \in A^N$ in R liegt. Sie liegt genau dann in R, wenn sie nicht in O liegt, für mindestens ein Buchstabenpaar (x, y) muß also

$$\big|m_{xy}(W) - Np(x)p(y|x)\big| \geq N\delta$$

sein. Die Wahrscheinlichkeit hierfür ist sicherlich nicht größer als die Summe über alle Paare (x, y) für die entsprechenden Wahrscheinlichkeiten, d.h.

$$p(W \in S) \leq \sum_{x\in A}\sum_{y\in A} p\left(\big|m_{xy}(W) - Np(x)p(y|x)\big| \geq N\delta\right) .$$

Wegen der Ergodizität der Quelle können wir für jedes $\widetilde{\varepsilon} > 0$ ein $N_1 \in \mathbb{N}$ finden, so daß für $N \geq N_1$ gilt

$$p\left(\left|\frac{m_x(W)}{N} - p(x)\right| > \frac{\delta}{2}\right) < \widetilde{\varepsilon}$$

oder, was dasselbe ist,

$$p\big(|m_x(W) - Np(x)| > \frac{\delta}{2}N\big) < \widetilde{\varepsilon} .$$

Für große N wird, wieder wegen der Ergodizität der Quelle, auch $m_x(W)$ groß; falls wir N so groß wählen, daß wir die Wahrscheinlichkeit, daß $m_x(W) < N_1$ ist, vernachlässigen können, gilt also auch

$$p\left(\left|\frac{m_{xy}(W)}{m_x(W)} - p(y|x)\right| > \frac{\delta}{2}\right) < \widetilde{\varepsilon} .$$

Die komplementäre Wahrscheinlichkeit dafür, daß

$$|m_x(W) - Np(x)| < \frac{\delta}{2}N$$

bzw.

$$\big|m_{xy}(W) - m_x(W)p(y|x)\big| < \frac{\delta}{2}m_x(W) \leq \frac{\delta}{2}N$$

ist, beträgt jeweils mindestens $1 - \widetilde{\varepsilon}$; die Wahrscheinlichkeit dafür, daß beides eintritt, ist also mindestens

$$(1 - \widetilde{\varepsilon})^2 > 1 - 2\widetilde{\varepsilon} .$$

Falls die erste der beiden Ungleichungen erfüllt ist, gilt erst recht

$$|m_x(W)p(y|x) - Np(x)p(y|x)| < \frac{\delta}{2}Np(y|x) \leq \frac{\delta}{2}N$$

und damit nach der Dreiecksungleichung zusammen mit der zweiten Ungleichung

$$\big|m_{xy}(W) - Np(x)p(y|x)\big| < \frac{\delta}{2}N + \frac{\delta}{2}N = N\delta .$$

Diese Ungleichung ist somit mindestens mit Wahrscheinlichkeit $1 - 2\widetilde{\varepsilon}$ erfüllt; die komplementäre Wahrscheinlichkeit ist

$$p\big(\big|m_{xy}(W) - Np(x)p(y|x)\big| > N\delta\big) < 2\widetilde{\varepsilon} .$$

Dies können wir in die oben abgeleitete Formel

$$p(W \in R) \leq \sum_{x\in A}\sum_{y\in A} p\left(\big|m_{xy}(W) - Np(x)p(y|x)\big| \geq N\delta\right)$$

einsetzen und erhalten, da es mit $m = \#A$ gerade m^2 Buchstabenpaare gibt,

$$p(W \in R) < 2m^2\tilde{\varepsilon}\,.$$

Da m eine Konstante ist, müssen wir nun nur $\tilde{\varepsilon}$ hinreichend klein wählen und erhalten dann, daß diese Wahrscheinlichkeit höchstens gleich ε ist, wie behauptet.

■

Die Mengen O und R aus dem gerade bewiesenen Satz lassen sich für eine gegebene Quelle natürlich nur schwer konstruieren, denn im allgemeinen muß N doch schon sehr groß gewählt werden, damit dieser Satz gilt. Wir können allerdings trotzdem Aussagen über die Größe von O und R machen: Für kleine Werte von ε und η machen wir keinen großen Fehler, wenn wir davon ausgehen, daß alle Nachrichten aus O mit *exakt* derselben Wahrscheinlichkeit vorkommen, und daß Nachrichten aus S *nie* vorkommen. Dann ist für $W \in A^N$

$$p(W) = \begin{cases} 1/\#O & \text{falls } W \in O \\ 0 & \text{falls } W \in S \end{cases},$$

und die Entropie der Quelle ist

$$-\sum_{W\in A^N} p(W)\log_2 p(W) = \sum_{W\in O} \frac{1}{\#O}\log_2 \#O = \log_2 \#O\,.$$

Die Entropie pro Buchstabe des Alphabets A ist damit gleich

$$\frac{\log_2 \#O}{N}\,.$$

Diese Entropie pro Buchstabe können wir berechnen; für die Quellen aus §1 haben wir dies dort bereits getan: Geht man nur von den Buchstabenhäufigkeiten aus, ist sie gleich 4,04088 Bit, und bei Berücksichtigung der Kontakthäufigkeiten gleich

$$\frac{4{,}04088 + 3{,}39765(N-1)}{N}\ \text{Bit}\,.$$

Also ist im ersten Fall

$$\#O \approx 2^{4{,}04088N}$$

und im zweiten

$$\#O \approx 2^{4{,}04088+3{,}39765(N-1)}\,.$$

In A^N gibt es insgesamt

$$26^N \approx 2^{4{,}7044N}$$

Nachrichten; das Verhältnis $\#O$ zu $\#A^N$ läßt sich daher schreiben als

$$\frac{\#O}{\#A^N} = 2^{-R_N}\,,$$

wobei im ersten Fall

$$R_N \approx (4{,}7044 - 4{,}04088)N = 0{,}66352N$$

ist und im zweiten

$$R_N \approx 4{,}7044N - 4{,}04088 + 3{,}39765(N-1) = 1{,}30675N - 0{,}64323\,.$$

Damit können wir uns daran machen, die Mehrdeutigkeit eines Schlüssels für diese (und viele andere) Quellen abzuschätzen: Wir ersetzen die jeweilige Quelle durch ein „durchschnittliche“ ergodische Quelle derselben Entropie und rechnen mit dieser in der Hoffnung, daß wir dadurch keinen allzu großen Fehler machen. Außerdem nehmen wir an, daß N hinreichend groß sei, so daß wir den obigen Satz anwenden können.

Nach Definition ist die Mehrdeutigkeit des Schlüssels nach Empfang einer Chiffretext-Nachricht C der Länge N gleich

$$H_{S|Ch} \underset{\text{def}}{=} -\sum_{(C,s)\,\in\, A^N\times S} p_{Ch,S}(C,s)\log_2 p_{S|Ch}(s|C)$$

Wenn wir davon ausgehen, daß jeder Schlüssel mit derselben Wahrscheinlichkeit auftritt, ist für jeden Schlüssel $s \in S$

$$p_S(s) = \frac{1}{\#S}\,.$$

und für einen Klartext $W \in A^N$ ist

$$p_K(W) = \begin{cases} 1/\#O & \text{falls } W \in O \\ 0 & \text{sonst} \end{cases};$$

eine Nachricht $C \in A^N$ kann also nur dann als Chiffretext auftreten, wenn es (mindestens) einen Schlüssel $s \in S$ gibt, so daß der zugehörige Klartext $W = T_s^{-1}(C)$ in O liegt. Wenn wir die Anzahl aller solcher

Schlüssel mit $N_S(C)$ bezeichnen, ist also die Wahrscheinlichkeit des Chiffretexts C gleich

$$p_{Ch}(C) = \frac{N_S(C)}{\#S} \cdot \frac{1}{\#O} = \frac{N_S(C)}{\#S \cdot \#O} ;$$

die Wahrscheinlichkeit für Chiffretext C verschlüsselt mit $s \in S$ ist

$$p_{Ch,S}(C,s) = \begin{cases} \frac{1}{\#O \cdot \#S} & \text{falls } T_s^{-1}(C) \in O \\ 0 & \text{sonst} \end{cases} ,$$

und damit ist die bedingte Wahrscheinlichkeit für den Schlüssel s nach Empfang des Chiffretexts C

$$p_{S|Ch}(s|C) = \frac{p_{Ch,S}(C,s)}{p_{Ch}(C)} = \frac{1}{N_S(C)} .$$

Somit ist

$$\begin{aligned} H_{S|Ch} &= - \sum_{(C,s) \in A^N \times S} p_{Ch,S}(C,s) \log_2 p_{S|Ch}(s|C) \\ &= - \sum_{C \in A^N} \sum_{\substack{s \in S \\ T_s^{-1}(C) \in O}} \frac{1}{\#O \cdot \#S} (-\log_2 N_S(C)) \\ &= \sum_{C \in A^N} \frac{N_S(C)}{\#O \cdot \#S} \log_2 N_S(C) . \end{aligned}$$

Diese Summe können wir nicht weiter ausrechnen, da wir die Zahlen $N_S(C)$ nicht kennen. Nun haben wir aber angenommen, daß wir ein *durchschnittliches* Kryptosystem haben, d.h. wir interessieren uns für den Mittelwert von $H_{S|Ch}$ über *alle* Kryptosysteme, und darüber können wir Aussagen machen:

Für einen zufällig gewählen Text $C \in A^N$ und einen zufällig gewählten Schlüssel $s \in S$ ist $T_s^{-1}(C) \in O$ mit Wahrscheinlichkeit

$$\frac{\#O}{\#A^N} = \frac{\#O}{n^N} = 2^{-R_N} ,$$

wobei die reelle Zahl R_N wie oben so gewählt wird, daß die rechtsstehende Gleichung gilt. Entsprechend ist die Wahrscheinlichkeit dafür, daß $T_s^{-1}(C)$ nicht in O liegt, gleich

$$1 - 2^{-R_N} .$$

Die Wahrscheinlichkeit dafür, daß $N_S(C)$ gleich einer festen Zahl m ist, können wir interpretieren als die Wahrscheinlichkeit dafür, daß für m Schlüssel s gilt $T_s^{-1}(C) \in O$, während für die restlichen $\#S - m$ Schlüssel gilt $T_s^{-1}(C) \notin O$. Falls die m Schlüssel vorgegeben sind, ist diese Wahrscheinlichkeit gleich

$$\left(2^{-R_N}\right)^m \left(1 - 2^{-R_N}\right)^{\#S-m} ,$$

und da es $\binom{\#S}{m}$ Möglichkeiten gibt, aus $\#S$ Schlüsseln m auszuwählen, ist die Wahrscheinlichkeit dafür, daß $N_S(C) = m$ ist, gleich

$$\binom{\#S}{m} \left(2^{-R_N}\right)^m \left(1 - 2^{-R_N}\right)^{\#S-m} ,$$

die Anzahl der Chiffretexte C mit $N_S(C) = m$ ist also gleich

$$\binom{\#S}{m} \left(2^{-R_N}\right)^m \left(1 - 2^{-R_N}\right)^{\#S-m} \#A^N .$$

Der durchschnittliche Wert von $H_{S|Ch}$ berechnet sich somit zu

$$\begin{aligned} &\sum_{m=0}^{\#S} \binom{\#S}{m} \left(2^{-R_N}\right)^m \left(1 - 2^{-R_N}\right)^{\#S-m} \#A^N \frac{m}{\#O\#S} \log_2 m \\ &= \frac{\#A^N}{\#O} \frac{1}{\#S} \sum_{m=0}^{\#S} \binom{\#S}{m} \left(2^{-R_N}\right)^m \left(1 - 2^{-R_N}\right)^{\#S-m} m \log_2 m \\ &= \frac{2^{R_N}}{\#S} \sum_{m=0}^{\#S} \binom{\#S}{m} \left(2^{-R_N}\right)^m \left(1 - 2^{-R_N}\right)^{\#S-m} m \log_2 m . \end{aligned}$$

Falls wir $\#S = 26$ setzen und von den beiden oben betrachteten Wahrscheinlichkeitsmodellen ausgehen, wenn wir also ein *zufälliges* Kryptosystem betrachten, das sich ungefähr wie das System der 26 CAESAR-Substitutionen verhält, läßt sich dies leicht berechnen; das Ergebnis ist für die beiden oben betrachteten Wahrscheinlichkeitsmodelle in den beiden folgenden Abbildungen dargestellt. Wie man sieht, ist die Mehrdeutigkeit des Schlüssels ab etwa zehn bis zwölf *bzw.* sechs bis acht Buchstaben Chiffretext praktisch gleich null, was auch ungefähr unseren Experimenten mit der CAESAR-Chiffre entspricht.

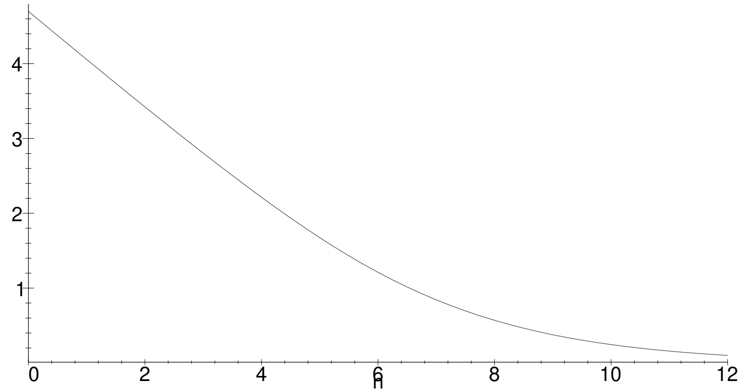

Abb. 1: Mehrdeutigkeit des Schlüssels via Buchstabenhäufigkeiten

Wie man den Kurven ansieht, nimmt der Mehrdeutigkeit des Schlüssels anfänglich etwa linear ab mit n. Wer die Binomialverteilung kennt, sieht auch leicht, warum dies so ist: Falls der Erwartungswert $2^{-R_N}\#S$ von $N_S(C)$ hinreichend groß ist, sind alle einigermaßen wahrscheinliche Werte in seiner näheren Umgebung, man kann daher ohne allzu großen Fehler $N_S(C)$ durch diesen Wert ersetzen und erhält

$$\begin{aligned} H_{S|Ch} &= \sum_{C\in A^N} \frac{N_S(C)}{\#O\cdot\#S}\log_2 N_S(C) \\ &\approx \#A^N\cdot\frac{2^{-R_N}\#S}{\#O\cdot\#S}(-R_N+\log_2\#S) \\ &= \log_2\#S - R_N\ . \end{aligned}$$

Anfänglich gewinnt man also aus N Buchstaben Chiffretext etwa R_N Bit des Schlüssels. Falls R_N wie in unseren Beispielen eine lineare Funktion von N ist, nimmt die Mehrdeutigkeit des Schlüssels also anfänglich linear ab.

Allgemein sieht man, daß der Gegner die Redundanz R_N der Nachrichtenquelle ausnutzt; je kleiner diese ist, desto weniger Information kann er gewinnen.

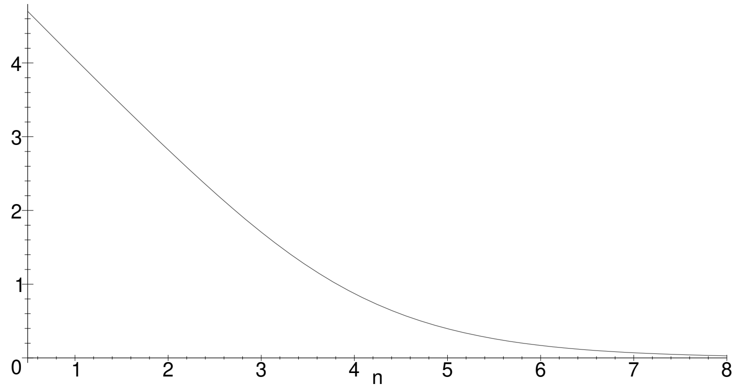

Abb. 2: Mehrdeutigkeit des Schlüssels via Kontakthäufigkeiten