## Probability as a Tool for Codebreaking

In our exploration of cryptography, we want to exploit the "structure" of plaintext messages. To this end, let us first consider the statistical "structure" of english plaintext through the following experiment. Let us generate at random 4-letter words, with an *equal* probability for each letter. The following is a list of 100 such words:

```
YCVT CWCV KKDA TGKW HFRR VGOJ LRIT RZEU ZGWK DMGB
XOBQ UJQK ISEG SOJM PNIG SGIP AIDX RMYC ZAYY OSYZ
TLOG IPUM UKOO TVSB DYZW ODWA LMUI GREY FCMV LEXK
TDSE XWMK ESUT VYLL NULO XWLB MGBW VDVU QTBP YMVC
IIPX MRCY WKAE XDJF NCXV BSOZ ALIJ NMOZ RYRA TWDV
EXJI TRLL UHLT YZND WYCQ RCMA DOPD BLJU SVVW KQSA
REQT PWUW VRAN NIMA PWJJ XVIF LVFR LVUB KARM MAAX
XSNK HCET GGAE LUUP FYIA VTAB WXHE YIJG ITTV PARE
JCEY PGEY BFEQ PVWX BLXO WFUJ MFQA UDUF AHDQ KSOY
BMBL EWRE NZGH OPOQ IQMP VFDI LCEV OJUZ HQJD OVQJ
```

Notice that very few of these are actual English words. Also, a quick glance will reveal that there are more Q's and Z's than one would expect. In a typical English text of 1000 letters, the frequencies of the letters of the alphabet are given in Table 1. The relative frequencies are shown as a histogram in Figure 1.

| letter | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| frequency | 73 | 9 | 30 | 44 | 130 | 28 | 16 | 35 | 74 | 2 | 3 | 35 | 25 | 78 | 74 | 27 | 3 | 77 | 63 | 93 | 27 | 13 | 16 | 5 | 19 | 1 |

Table 1: Frequencies of each letter found in a typical text of 1000 English letters.

We can exploit this distribution of letters in our generation of random 4-letter words by selecting letters randomly based on these frequencies. In other words, we can randomly generate letters by spinning the wheel in Figure 2. The following is a list of 100 words generated in this way:

```
ECHO YATE DRMI EXVT EWTS DATL YRAN CTTT FANC ESEY
VTOP IRPE YARE QAMN XSAI ELWP PYRT SRBT ASXN SIEE
MONT OEDA LIYR NAOR HRDA HRYE CRVR CGAN MSEH PRTI
UANN NSOE EMCT WDVT PDAE UAAA HRNN OFEF HGIV NFAO
EHDD ERNP ETTP ENAO MOOC OIEE QOTO EEOD LEFD MIDE
EETT NMAP NOBD ENNC NSSV LOII PTTR CGNH CNIE VTOR
EEEE TLAV EJIH OOAN DSAI IFMF IRTS DVFT SRJH DTFS
OCRR EIIC REAR DPUM EHTE EIIN WTBA HLAO FWDH TRNU
AESO IOCH HTCP UHER MAEH AEIT NODT LZDI WEYR AWCU
RAET LPHI IRDJ MGOC RDEO ELIS NDIE NIOP SOCO ECWN
```

As we can observe these words are a little closer to usual english words. But we can do even better by taking into account more intricate statistics on english plaintext. For instance a table of frequencies of initial letters, final letters, and transitions in english letter pairs.

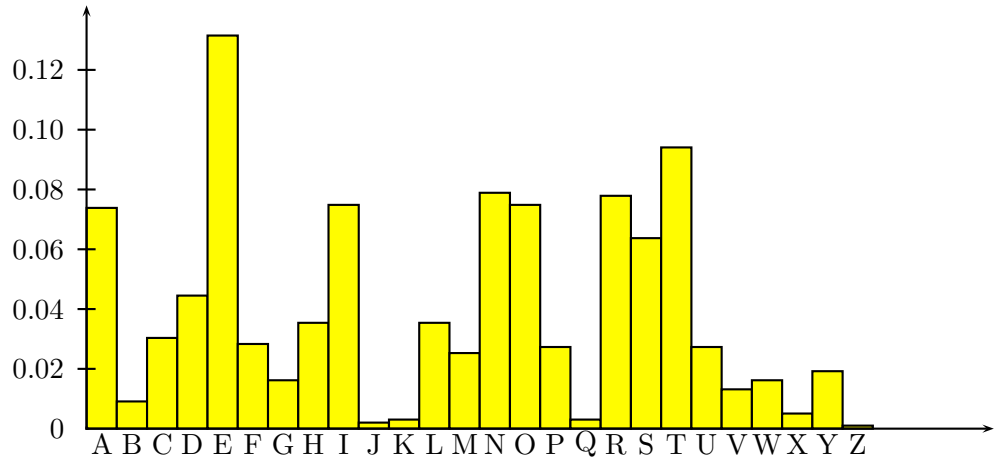


Figure 1: Histogram of relative frequencies for each letter in the English language.

| | | | | | | |
|---|---|---|---|---|---|---|
| A 1802 | E 410 | I 922 | M 578 | Q 31 | U 224 | |
| B 757 | F 666 | J 95 | N 401 | R 513 | V 100 | Y 126 |
| C 918 | G 293 | K 88 | O 1176 | S 1213 | W 833 | Z 6 |
| D 459 | H 636 | L 348 | P 768 | T 2614 | X 10 | |

Table 2: Frequencies of each letter appearing as the initial letter of 16,410 words of newspaper text.

To compute Table 4, a typical example of english text was chosen. Then the first row of the table was obtained by recording, for each of 10,000 occurrences of the letter A, the letter that immediately followed it. Thus the entry 469 in the column indexed by D means that in this sample of 10.000 occurrences of the letter A the letter D was observed to immediately follow A exactly 469 times. The same procedure was repeated for each of the letters of the alphabet. We see that in 10.000 occurrences of Q, the letter U followed it all 10.000 times (not surprising!). As we will see in what follows, to get the conditional probability $P[\text{next letter} = Y | \text{preceeding letter} = X]$ you simply look at the entry in row $X$ and column $Y$ and divide by 10,000. Thus

$$P[\text{next letter} = \text{E} \mid \text{preceeding letter} = \text{R} \,] \; = 2795/10,000 = .27$$

The following list of four-letter words was constructed by selecting the first letter of each word based on the frequencies given in Table 2. Each of the next three letters was chosen according to the previous letter using Table 4.

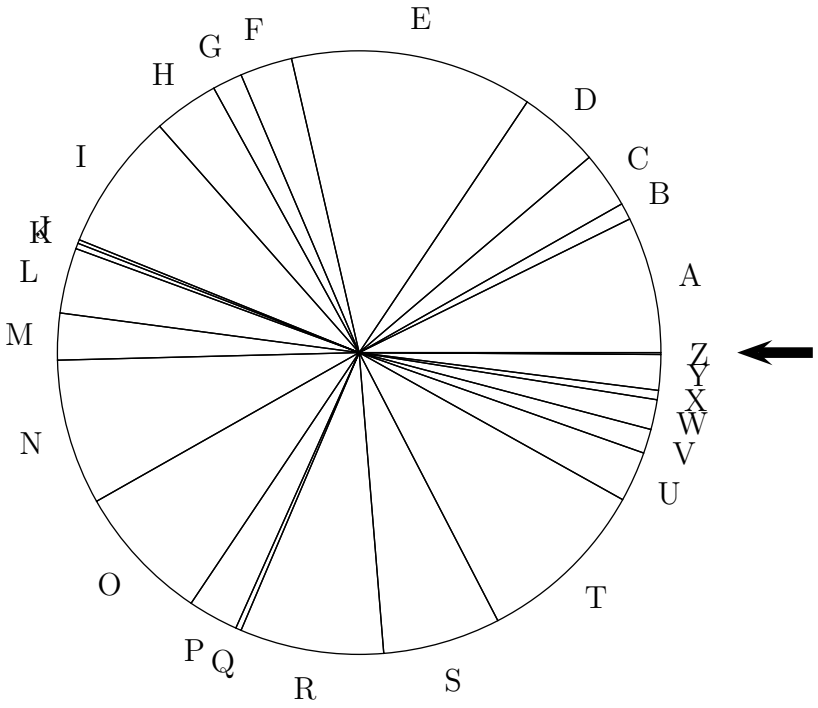


Figure 2: Fortune wheel simulating relative frequencies for each letter in the English language.

| | | | | | | |
|---|---|---|---|---|---|---|
| A 480 | E 3325 | I 72 | M 220 | Q 1 | U 29 | |
| B 25 | F 744 | J 6 | N 1592 | R 906 | V 15 | Y 903 |
| C 107 | G 463 | K 148 | O 745 | S 2077 | W 166 | Z 5 |
| D 1649 | H 407 | L 599 | P 84 | T 1587 | X 34 | |

Table 3: Frequencies of each letter appearing as the final letter of 16,410 words of newspaper text.

```
KETO RMER UEAN TINT RERE SERN ALEC ILET SAIA AMAN
FRAT SERE MERT ASTE WEDI OORL THTI ANIS PERS WSER
ORET FLIT NDIN ESER NDEN MEEH TONT ITOR CESE RORI
AIAT IGOC BALI HENE WEIA JOIE ONIN ECHE SETE OUND
SIAR ONTO ACOR CITO THAR RESS BESE RDED FERE FFFO
OPER TESE TARS CALE KNTT AUTO INEA MESE THTO SISP
WNES ARED SPOO INTS ANTH TSHE ASIN RTER SEUN TENS
TEEL ATHE ASTO GLOR HRER HEDA GANE MATH COFT TAST
STHE UNAE ATIN ATST BETH IATO EANE TENT BERE VECT
BANT VENE ATHE ITHO CURE CEEA TEST ONTA MATS ONSE
```

It is clear that we can extend this mechanism to account for final letter frequencies, triple letter frequencies, etc. We end by listing 100 four-letter words that were randomly selected by the same

mechanism outlined above, except that the third and fourth letters were chosen according to the two previous letters and a table of triple letter statistics (not shown).

FADD WERA ASSI ENEW THEL DIAL UTCH IESP PHES TOCK
ORYS ARGA THAR HESU ATED WAYE COMI ASDA INTO OUTO
FACT DISP INTE TEDF CALU WATE TOTO WERE CTEX ANUA
UNSP ATIS GAME TERF STWH TRUC GRIF ROBJ ESYO OREQ
SENA TERS TEMP INAT PTEN ALET HEDI PRIS ITSA ONSW
SIBE DINT BERM DINE ALON HESC INES BREA ONCL STRI
CADD UTGI FILL ORPR BERS TERO NETW ATHE SHOR IESO
WSIN BERE TOCD ITES AILE RIES TORK OTON PART TABL
CARE ANDF ETWI OPTI FECK ANDI CONI PROB STNE GHTH
ANDE CEGA RAND ASST STCA TODE ONOM UNSN SINE DERO

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 11 | 193 | 388 | 469 | 20 | 100 | 233 | 20 | 480 | 20 | 103 | 1052 | 281 | 1878 | 8 | 222 | 0 | 1180 | 1001 | 1574 | 137 | 212 | 57 | 26 | 312 | 23 |
| B | 932 | 57 | 16 | 8 | 3220 | 0 | 0 | 0 | 605 | 57 | 0 | 1243 | 49 | 0 | 965 | 0 | 0 | 662 | 229 | 49 | 727 | 16 | 0 | 0 | 1165 | 0 |
| C | 1202 | 0 | 196 | 4 | 1707 | 0 | 0 | 1277 | 761 | 0 | 324 | 369 | 15 | 11 | 2283 | 0 | 4 | 426 | 87 | 893 | 347 | 0 | 0 | 0 | 94 | 0 |
| D | 1044 | 20 | 26 | 218 | 3778 | 7 | 132 | 7 | 1803 | 33 | 0 | 125 | 178 | 53 | 733 | 0 | 7 | 324 | 495 | 13 | 601 | 99 | 40 | 0 | 264 | 0 |
| E | 660 | 36 | 433 | 1195 | 438 | 142 | 125 | 21 | 158 | 5 | 36 | 456 | 340 | 1382 | 40 | 192 | 34 | 1927 | 1231 | 404 | 48 | 215 | 205 | 152 | 121 | 4 |
| F | 838 | 0 | 0 | 0 | 1283 | 924 | 0 | 0 | 1608 | 0 | 0 | 299 | 9 | 9 | 2788 | 0 | 0 | 1215 | 26 | 496 | 462 | 0 | 0 | 0 | 43 | 0 |
| G | 1078 | 0 | 0 | 18 | 2393 | 0 | 177 | 1281 | 839 | 0 | 0 | 203 | 27 | 451 | 1140 | 0 | 0 | 1325 | 256 | 247 | 512 | 0 | 0 | 0 | 53 | 0 |
| H | 1770 | 5 | 14 | 8 | 5624 | 0 | 0 | 5 | 1168 | 0 | 0 | 16 | 16 | 38 | 786 | 0 | 0 | 153 | 27 | 233 | 85 | 0 | 11 | 0 | 41 | 0 |
| I | 380 | 82 | 767 | 459 | 437 | 129 | 280 | 2 | 16 | 0 | 50 | 567 | 297 | 2497 | 893 | 100 | 8 | 342 | 1194 | 1135 | 11 | 250 | 0 | 23 | 2 | 79 |
| J | 1259 | 0 | 0 | 0 | 1818 | 0 | 0 | 0 | 350 | 0 | 0 | 0 | 0 | 0 | 3147 | 0 | 0 | 70 | 0 | 0 | 3356 | 0 | 0 | 0 | 0 | 0 |
| K | 395 | 28 | 0 | 28 | 5283 | 28 | 0 | 198 | 1582 | 0 | 113 | 198 | 28 | 565 | 198 | 0 | 0 | 85 | 1102 | 28 | 28 | 0 | 0 | 0 | 113 | 0 |
| L | 1342 | 19 | 22 | 736 | 1918 | 105 | 108 | 0 | 1521 | 0 | 79 | 1413 | 82 | 4 | 778 | 41 | 0 | 34 | 389 | 254 | 269 | 56 | 11 | 0 | 819 | 0 |
| M | 1823 | 337 | 26 | 0 | 2976 | 10 | 0 | 0 | 1345 | 0 | 0 | 10 | 654 | 42 | 1246 | 722 | 0 | 26 | 244 | 5 | 337 | 5 | 0 | 0 | 192 | 0 |
| N | 550 | 4 | 621 | 1681 | 1212 | 102 | 1391 | 13 | 665 | 9 | 66 | 73 | 104 | 194 | 528 | 4 | 7 | 11 | 751 | 1641 | 124 | 68 | 18 | 2 | 157 | 4 |
| O | 85 | 101 | 162 | 231 | 37 | 1299 | 82 | 25 | 92 | 14 | 78 | 416 | 706 | 2191 | 222 | 292 | 0 | 1531 | 357 | 396 | 947 | 334 | 345 | 12 | 41 | 4 |
| P | 1358 | 0 | 6 | 0 | 1747 | 0 | 0 | 237 | 423 | 0 | 0 | 812 | 73 | 6 | 1511 | 581 | 0 | 2305 | 180 | 287 | 457 | 0 | 0 | 0 | 17 | 0 |
| Q | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10000 | 0 | 0 | 0 | 0 | 0 |
| R | 1026 | 33 | 172 | 282 | 2795 | 31 | 175 | 17 | 1181 | 0 | 205 | 164 | 303 | 325 | 1114 | 55 | 0 | 212 | 655 | 596 | 192 | 142 | 17 | 2 | 306 | 0 |
| S | 604 | 12 | 284 | 27 | 1795 | 24 | 0 | 561 | 1177 | 0 | 91 | 145 | 112 | 21 | 706 | 386 | 9 | 27 | 836 | 2484 | 579 | 0 | 39 | 0 | 81 | 0 |
| T | 619 | 3 | 36 | 2 | 1417 | 7 | 2 | 3511 | 1406 | 0 | 0 | 101 | 44 | 15 | 1228 | 3 | 0 | 479 | 418 | 213 | 195 | 5 | 88 | 0 | 203 | 5 |
| U | 344 | 415 | 491 | 243 | 434 | 52 | 382 | 10 | 258 | 0 | 14 | 1097 | 329 | 1518 | 19 | 386 | 0 | 1460 | 1221 | 1255 | 29 | 14 | 0 | 10 | 14 | 5 |
| V | 749 | 0 | 0 | 23 | 6013 | 0 | 0 | 0 | 2568 | 0 | 0 | 0 | 12 | 0 | 530 | 0 | 0 | 0 | 23 | 0 | 12 | 12 | 0 | 0 | 58 | 0 |
| W | 2290 | 8 | 0 | 32 | 1942 | 0 | 0 | 1422 | 2104 | 0 | 0 | 41 | 0 | 357 | 1292 | 0 | 0 | 106 | 366 | 16 | 0 | 0 | 0 | 0 | 24 | 0 |
| X | 672 | 0 | 1119 | 0 | 1269 | 0 | 0 | 75 | 1119 | 0 | 0 | 0 | 75 | 0 | 75 | 3507 | 0 | 0 | 0 | 1716 | 0 | 0 | 0 | 373 | 0 | 0 |
| Y | 586 | 34 | 103 | 69 | 2898 | 0 | 0 | 0 | 691 | 0 | 34 | 172 | 379 | 172 | 2208 | 310 | 0 | 310 | 1518 | 172 | 138 | 0 | 103 | 0 | 69 | 34 |
| Z | 2278 | 0 | 0 | 0 | 4557 | 0 | 0 | 0 | 2152 | 0 | 0 | 127 | 0 | 0 | 506 | 0 | 0 | 0 | 0 | 0 | 127 | 0 | 0 | 0 | 0 | 253 |

Table 4: English biletter conditional probabilities