

Breaking Classical Ciphers

Agenda for Today

- More breaking of monoalphabetic substitution ciphers

Breaking Monoalphabetic Substitution Ciphers

Monoalphabetic Substitution Ciphers

- Assign to each **plaintext** letter a different **coding** letter to ensure correct decryption.
- 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
- u v w c a b x y z q p o r s t e f g l m n d h i j k
- Today we look into monoalphabetic substitution ciphers (plaintext)
- MTCUJ HA OTTP ZSMT RTSTUOEYUVAMZW
LNVLMZMNMZTS WZEYAGL (ciphertext)

Frequency of Letters

- Letters in order of frequency (highest to lowest):
- E T A O N R I S H D L F C M U G Y P W B V K X J Q Z
- The letters can be grouped further by their frequencies:
 - Very Common: E
 - Common: T
 - Next most common: A O N R I S
 - Less Common: H
 - Less Common Still: D L F C M U
 - Rare: V K X J Q Z

Frequency of Digrams

- Most common pairs of letters (digrams), in order of frequency:
- TH HE AN RE ER IN ON AT ND ST ES EN OF
TE ED OR TI HI AS TO AR OU IS IT LE NT RI
SE HA AL DE EA NE RO OM IO WE VE TA
TR CO ME NG MA CE RA IC NS UT US BE
UN CH WA SI LA AD LI RT CA NC SO NC SO
LL UR EL RS EM AC IM PR TT OT WI EC
- The most common words in English are:
- THE OF AND TO IN A IS THAT FOR IT BY
ARE BE WAS AS HE WITH HIS

Ciphertext from a Monoalphabetic Substitution Cipher

- EZNYBWNFPEZNDWEFOGQMPBODQUGYNDWDOFUYKDSQZZOOZKYZZJEMZFUWYFD
SXFLOWDSQLCSYNUYWBNLOKZGDWEPOYBYWDSQDQZONYWMUZEZNYBWMZED
OZDNMUZOZYNDBOZDMEZDSPADFMKYMGYWEZNYBWMUZPOGFPEYWBMUZPOG
DWEFOGQMPBODQUGDWEQZFWMSGMUZOZUDNCZZWFPWNYEZODCSZOZNZDOFU
EPWZJUUFUCOYWBNMPBZMUZOTPOZMUDWPWZPAMUZNLCXZFMNXPLOWDSNFLO
OZWMSGZHYNMAPOZDFUPAMUZYWEYKYELDSOZDNCLMWPWZZWFPLODBZ
MUZYWMZODFMYPWPAMUZEYNFYQSYWZN
- NPTZPAMUZNZXFLOWDSNZTQUDNYIZPWSGMUZMUZPOZMYFDSNDQZFMNPAZDFUN
LCXZFMNPTZPWSGMUZQODFMYFDSEZNYBWNFPEZNDWEFOGQMPBODQUGQOPKYE
ZNDAPOLTAPOUYBURLDSYMGQDQZONPACPMUDMUZPOZMYFDSWDWEDQODFMYFD
SWDMLOZJUUFUCOYEBZTPOZMUDWPWZPAMUZNZDOZDNMUZNFQZPAMUZXPLO
WDSYNJYEZOQDQZONZTQUDNYIYWBMUZDSBZCODYFDWEBZPTZMOYFDNQZFMNP
ADWGPAMUZDOZDNJYSSCZFPWNYEZOZEJYMUQDQZONMPLFUYWBTPOZMUDWPW
ZPAMUZDOZDNCZYWBZNQZFYDSSGJZSFPTZ
- MUZXFLOWDSYNCZYWBEYNMOYCLMZEMPZWBYWZZONTDMUZTDMYFYDWNFPTQ
LMZONFYZWMYNMNDWEPMUZOYW ZELFDMYPWYWELNMOGDWEBPKZOWTZWM

Our Task Today

- We shall break this cipher.
- We were told that the original plaintext is in English and was encrypted by a one-to-one mapping from the English alphabet to itself.
- We can see that spaces between words and punctuation were deleted.

Frequency of Letters: Comparison

- In message In English

- Z: 13.2 e: 12.7
- D: 8.6 t: 9.1
- M: 7.6 a: 8.2
- Y: 7.4 i: 7.0
- O: 7.2 n: 6.7
- W: 7.1 o: 6.3
- P: 6.8 h: 6.1
- N: 6.4 r: 6.0
- U: 5.5 d: 4.3
- F: 4.8 q: 4.3
- E: 3.8 l: 4.0
- S: 3.2 c: 2.8
- Q: 3.1 u: 2.8

- In message In English

- B: 2.8 m: 2.4
- L: 2.4 w: 2.3
- G: 2.1 f: 2.2
- A: 1.7 s: 2.2
- C: 1.6 g: 2.0
- T: 1.6 y: 2.0
- J: 0.8 p: 1.9
- K: 0.8 b: 1.5
- X: 0.8 v: 1.0
- I: 0.2 k: 0.8
- H: 0.1 j: 0.2
- R: 0.1 x: 0.1
- V: 0.0 z: 0.1

Most Common Digraphs

- In message:
 - MU, UZ, OZ, DS, ZO, YW, DW, ZN, ZD, WE
- In English:
 - th, he, in, er, ed, an, nd, ar, re, en
- Observations:
 - from single frequency, clearly $Z \leftrightarrow e$, but we are not sure about
 - $D \leftrightarrow t$ and $M \leftrightarrow a$ or $D \leftrightarrow a$ and $M \leftrightarrow t$
 - from the first two digrams, we can guess that $M \leftrightarrow t$ and $U \leftrightarrow h$
- We should first guess (this could be wrong):
 - $D \leftrightarrow a$ and $M \leftrightarrow t$ and $U \leftrightarrow h$ and $Z \leftrightarrow e$

After the First Guess

- EeNYBWNFP EeNaWEFOGQtPBOaQhGYNaWaOFhYKaSQeeOOeKYeJeEteFhWYFaSXPLOWaSQLCSYNhYWBNLOKeGaWEPOYBYWaSQaQeONYWtheEeNYBWateEaOeaNtheOeYNaBOea tEeaSPAaFtYKYtGYWEeNYBWthePOGFPEYWBthePOG
aWEFOGQtPBOaQhGaWEOfFeWtSGtheOehaNcEeWFPWNYEeOaCSeOeNeaOFhEPWeJhYFhC
OYWBNTPBetheOTPOethaWPWePAtheNLCXeFtNXPLOWaSNFLOOeWtSGeHYNTAPOeaFhPAth
heYWEYKYELaSaOeaNCLtWPWeeWFPLOaBe theYWteOaFtYPWPAtheEYNFYQSYWeN
- NPTePAtheNeXPLOWaSNeTQhaNYIePWSGthethePOetYFaSaNQeFtNP AeaFhNLCXeFtNPTePW
SGtheQOaFtYFaSEeNYBWNFP EeNaWEFOGQtPBOaQhGQOPKYEeNaAPOLTAPOhYBhRLaS
YtGQaQeONPACPthathePOetYFaSaWEaQOaFtYFaSWatLOeJhYFhCOYEBeTPOethaWPWePAth
eNeaOeaNtheNFPQePAtheXPLOWaSYNJYEeOQaQeONeTQhaNYIYWBtheaSBeCOaYFaWEBe
PTetOYFaNQeFtNPAAWGPAtheaOeaNJYSSCeFPWNYEeOeEJYthQaQeONtPLFhYWBTPOetha
WPWePAtheaOeaNCeYWBEnQeFYaSSGJeSFPTe
- theXPLOWaSYNCeYWB EYNtOYCLteEtPeWBYWeeONTatheTatYFYaWNFPTQLteONFYeWtY
NtNaWEPtheOYW eELFatYPWYWELNtOGaWEBPKeOWTeWt
- **What can we say from this text? What should we do next?**

What Should We Do Next?

- It is hard for us to extract more information from the text in the previous page.
- We have to go back and look at the statistics of single letters and digrams.
 - From single letter frequency, very likely $O \leftrightarrow \{i, n, o, h, r\}$
 - Since $Z \leftrightarrow e$, we have $OZ \leftrightarrow *e$
 - From the digram frequency, very likely $OZ \leftrightarrow re$
 - Combining the two above, it is very likely $O \leftrightarrow r$
- Let us try the second guess $O \leftrightarrow r$

After the Second Guess

- EeNYBWNFP EeNa WEFrGQtPBraQhGYNaWarFhYKaSQeerreKYeJeEteFhWYFaSXPLrWaSQLC
SYNhYWB N LrKeGaWEPrYBYWaSQaQerNYWtheEeNYBWateEareaNthereYNaBreatEeaSPAaFt
YKYtGYWEeNYBWthePrGFPEYWBthePrG
aWEFrGQtPBraQhGaWEreFeWtSGtherehaNCeeWFPWNYEeraCSereNearFhEPWeJhYFhCrYWB
NtPBetherTPrethaWPWePAtheNLCXeFtNXPLrWaSNFLrreWtSGeHYNtAPreaFhPAtheYWEYKY
ELaSareaNCLtWPWeeWFPLraBe theYWteraFtYPWPAtheEYNFYQSYWeN
- NPTePAtheNeXPLrWaSNeTQhaNYIePWSGthethePretYFaSaNQeFtNPAeaFhNLCXeFtNPTePWS
GtheQraFtYFaSEeNYBWNFP EeNa WEFrGQtPBraQhGQrPKYEeNaAPrLTAPrhYBhRLaSYtGQa
QerNPACPthathePretYFaSaWEaQraFtYFaSWatLreJhYFhCrYEBEtprethaWPWePAtheNeareaNthe
NFPQePAtheXPLrWaSYNJYEerQaQerNeTQhaNYIYWBtheaSBeCraYFaWEBEPTetrYFaNQeFtN
PAaWGPAtheareaNJYSSCeFPWNYEereEJYthQaQerNtPLFhYWBTPrethaWPWePAtheareaNCe
YWBENQeFYaSSGJeSFPTe
- theXPLrWaSYNCeYWB EYNtrYCLteEtPeWBYWeerNTatheTatYFYaWNFP TQLterNFYeWtYNtN
aWEptherYW eELFatYPWYWELNtrGaWEBPKerWTeWt
- What can we derive from this text?

What Should We Do after the Second Guess

- Again, we have to look at the statistics.
- We have so far guessed the following:
 - $D \leftrightarrow a$, $M \leftrightarrow t$, $U \leftrightarrow h$, $Z \leftrightarrow e$, $O \leftrightarrow r$
- The next is to look at Y.
 - It is very likely that $Y \leftrightarrow \{i, n, o\}$ from the single letter frequency
 - From the digram statistics, $YW \leftrightarrow \{in, nd\}$ very likely.
 - By single letter statistics, very likely $YW \leftrightarrow in$
- Let us make the 3rd guess: $YW \leftrightarrow in$

After the 3rd Guess

- EeNiBnNFPEeNanEFrGQtPBraQhGiNanarFhiKaSQeerreKieJeEteFhniFaSXPLrnaSQLCSiNhinBN
LrKeGanEPriBinaSQaQerNintheEeNiBnateEareaNthereiNaBreatEeaSPAaFtiKitGinEeNiBnthePrGF
PEinBthePrG
anEFrGQtPBraQhGanEreFentSGtherehaNceenFPnNiEraCSereNearFhEPneJhiFhCrimBNtPBetherT
PrethanPnePAtheNLCXeFtNXPLrnaSNFLrrentSGeHiNtAPreaFhPAtheinEiKiELaSareaNCLtnPnee
nFPLraBe theinteraFtiPnPAtheEiNFiQSineN
- NPTePAtheNeXPLrnaSNeTQhaNiIePnSGthethePretiFaSaNQeFtNPAeaFhNLCXeFtNPTePnSGthe
QraFtiFaSEeNiBnNFPEeNanEFrGQtPBraQhGQrPKiEeNaAPrLTAPrhiBhRLaSitGQaQerNPACPth
athePretiFaSanEaQraFtiFaSnatLreJhiFhCriEBetPrethanPnePAtheNeareaNtheNFPQePAtheXPLrna
SiNjiEerQaQerNeTQhaNiIinBtheaSBeCraiFanEBetTetriFaNQeFtNPAanGPAtheareaNjiSSCeFPn
NiEereEJithQaQerNtPLFhinBTPrethanPnePAtheareaNceinBeNQeFiaSSGJeSFPTe
- theXPLrnaSiNceinBEiNtriCLteEtPenBineerNTatheTatiFianNFPTQLterNFientiNtNanEPtherin
eELFatiPninELNtrGanEBPKernTent
- **What can we derive from this text?**

What Should We Do after the 3rd Guess

- Again, we have to look at the statistics.
- We have so far guessed the following:
 - $D \leftrightarrow a$, $M \leftrightarrow t$, $U \leftrightarrow h$, $Z \leftrightarrow e$, $O \leftrightarrow r$, $Y \leftrightarrow i$, $W \leftrightarrow n$
- The next is to look at P.
 - It is very likely that $P \leftrightarrow o$ from the single letter frequency
- Let us make the 4th guess: $P \leftrightarrow o$

After the Fourth Guess

- EeNiBnNFoEeNanEFrGQtoBraQhGiNanarFhiKaSQeerreKieJeEteFhniFaSXoLrnaSQLCSiNhinBN
LrKeGanEoriBinaSQaQerNintheEeNiBnateEareaNthereiNaBreatEeaSoAaFtiKitGinEeNiBntheorGF
oEinBtheorG
anEFrGQtoBraQhGanEreFentSGtherehaNceenFonNiEeraCSereNearFhEoneJhiFhCrimBNtoBetherT
orethanoneoAtheNLCXeFtNXoLrnaSNFLrrentSGeHiNtAoreaFhoAtheinEiKiELaSareaNCLtnoneen
FoLraBe **theinteraFtion**oAtheEiNFfiQSineN
- NoTeoAtheNeXoLrnaSNeTQhaNiLeonSGthetheoretiFaSaNQeFtNoAeaFhNLCXeFtNoTeonSGtheQ
raFtiFaSEeNiBnNFoEeNanEFrGQtoBraQhGQroKiEeNaAorLTAorhiBhRLaSitGQaQerNoACothath
eoretiFaSanEaQraFtiFaSnatLreJhiFhCriEBEtoethanoneoAtheNeareaNtheNFoQeoAtheXoLrnaSiNJ
iEerQaQerNeTQhaNiinBtheaSBeCraifanEBeoTetriFaNQeFtNoAanGoAtheareaNJiSSCeFonNiEere
EJithQaQerNtoLFhinBTorethanoneoAtheareaNceinBeNQeFiaSSGJeSFoTe
- theXoLrnaSiNceinBEiNtriCLteEtoenBineerNTatheTatiFianNFoTQLterNFientiNtNanEotherin
eELFationinELNtrGanEBoKernTent
- **What should we do next?**
 - **the interaFtion** should be one word. So $F \leftrightarrow c$
 - **oAthe** is likely “of the”. So $A \leftrightarrow f$
- Let us try the fifth guess: $F \leftrightarrow c$ and $A \leftrightarrow f$

After the Fifth Guess

- EeNiBnNcoEeNanEcrGQtoBraQhGiNanarchiKaSQeerreKieJeEtechnicaSXoLrnaSQLCSiNhinBNLrKeGanEoriBinaSQaQerNintheEeNiBnateEareaNthereiNaBreatEeaSofactiKitGinEeNiBntheorGcoEi nBtheorG
anEcrGQtoBraQhGanErecentSGtherehaNceenconNiEraCSereNearchEoneJhichCrinBNtoBetherto
rethanoneoftheNLCXectNXoLrnaSNcLrrentSGeHiNtforeachoftheinEiKiELaSareaNCLtnoneencoLr
aBe the interaction of the EiNciQSineN
- NoTeoftheNeXoLrnaSNeTQhaNiLeonSGthetheoreticaSaNQectNofeachNLCXectNoTeonSGtheQract
icaSEeNiBnNcoEeNanEcrGQtoBraQhGQroKiEeNaforLTforhiBhRLaSitGQaQerNofCothattheoretica
SanEaQracticaSnatLreJhichCriEBetoethanoneoftheNeareaNtheNcoQeoftheXoLrnaSiNjiEerQaQer
NeTQhaNiLinBtheaSBeCraicanEBeoTetricaNQectNofanGoftheareaNJiSSCeconNiEereEJithQaQerN
toLchinBTorethanoneoftheareaNceinBeNQeciaSSGJeScoTe
- theXoLrnaSiNceinBEiNtriCLteEtoenBineerNTatheTaticianNcoTQLterNcientiNtNanEotherin
eELcationinELNtrGanEBoKernTent
- theorG should be “theory”. So $G \leftrightarrow y$
- technicaS should be “tecnical”. So $S \leftrightarrow l$
- Let us make the sixth guess: $S \leftrightarrow l$ and $G \leftrightarrow y$

After the 6th Guess

- EeNiBnNcoEeNanEcryQtoBraQhyiNanarchiKalQeerreKieJeEtechnicalXoLrnalQLCliNhinBNLrK eyanEoriBinalQaQerNintheEeNiBnateEareaNthereiNaBreatEealofactiKityinEeNiBn theory coEinB theory
anEcryQtoBraQhyanErecentlytherehaNceenconNiEeraClereNearchEoneJhichCrinBNtoBethertoret hanoneoftheNLCXectNXoLrnalNcLrrentlyeHiNtforeachoftheinEiKiELalareaNCLtnoneencoLraBe the interaction of the EiNciQlineN
- NoTeoftheNeXoLrnalNeTQhaNiIe only the theoretical aNQectNofeachNLCXectNoTe only the QracticalEeNiBnNcoEeNanEcryQtoBraQhyQroKiEeNaforLTforhiBhRLalityQaQerNofCothatheore ticalanEaQracticalnatLreJhichCriEBetoethanoneoftheNeareaNtheNcoQeoftheXoLrnaliNjiEerQaQ erNeTQhaNiIinBthealBeCraicaanEBeoTetricaNQectNofanyoftheareaNJillCeconNiEereEJithQaQerN toLchinBTorethanoneoftheareaNceinBeNQeciallyJelcoTe
- theXoLrnaliNceinBEiNtriCLteEtoenBineerNTatheTaticianNcoTQLterNcientiNtNanEotherin eELcationinELNtryanEBoKernTent
- anE should be “and”. So $E \leftrightarrow d$
- anarchiKal should be “an archival”. So $K \leftrightarrow v$
- Qractical should be “practical”. So $Q \leftrightarrow p$
- Let us do the 7th replacement $Q \leftrightarrow p$, $K \leftrightarrow v$, $E \leftrightarrow d$

After the 7th Guess

- deNiBnNcodeN and cryptoBrphy iN an archival peer revieJed technical
 XoLrnalpLCliNhinBNLrvey and oriBinal paper
 NinthedeNiBnatedareaNthereiNaBreatdealofactivityindeNiBn theory codinB theory and
 cryptoBrphy and recently
 therehaNceenconNideraClereNearchdoneJhichCrinBNtoBetherTorethanoneoftheNLCXectNXoLrnal
 NeLrrentlyeHiNtforeachoftheindividLalareaNCLtnoneencoLraBe the interaction of the diNciplineN
- NoTeoftheNeXoLrnalNeTphaNile only the theoretical aNpect
 NofeachNLCXectNoTeonlythepracticaldeNiBnNcodeNandcryptoBrphyprovideNaforLTforhiBhRL
 alitypaperNofCothatheoreticalandapracticalnatLreJhichCridBeTorethanoneoftheNeareaNtheNcopeof
 theXoLrnaliNJiderpaperNeTphaNi linBthealBeCraicandBeoTetricaNpectNofanyoftheareaNJillCecon
 NideredJithpaperNtoLchinBTorethanoneoftheareaNceinBeNpeciallyJelcoTe
- theXoLrnaliNceinBdiNtriCLtedtoenBineerNTatheTaticianNcoTpLterNcientiNtNandotherin
edLcation indLNtry and BovernTent
- Now everything is easy!

The Original Plaintext

- Designs, Codes and Cryptography is an archival peer reviewed technical journal publishing survey and original papers in the designated areas. There is a great deal of activity in design theory, coding theory and cryptography and recently there has been considerable research done which brings together more than one of the subjects. Journals currently exist for each of the individual areas but none encourage the interaction of the disciplines.
- Some of these journals emphasize only the theoretical aspects of each subject, some only the practical. Designs, Codes and Cryptography provides a forum for high quality papers of both a theoretical and a practical nature which bridge more than one of these areas. The scope of the journal is wider. Papers emphasizing the algebraic and geometric aspects of any of the areas will be considered with papers touching more than one of the areas being especially welcome.
- The journal is being distributed to engineers, mathematicians, computer scientists and other in education, industry and government.

The Secret Key for Encryption

- The secret key

- 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
- D C F E Z A B U Y X V S T W P Q R O N M L K J H G I

Summary of Part I

- In case of no spaces between words and no punctuation in the ciphertext, it may be hard to break a substitution cipher.
- It is very hard at the beginning. Statistics of single letter frequency and that of digrams should be used.
- A guess could be wrong. Avoid wrong guesses.
- Start and end of sentences are highest priority.