MATH 587/CSCE 557 - SUMMARY OF CLASS, 2/15/07

Class Feb 8 was review for the 1st midterm. Class Feb 13 was the 1st midterm. Today we continue our study of frequency analysis.

We've been considering three kinds of attack. For the ciphertext-only attacks on shift ciphers and affine ciphers we saw the benefit of frequency analysis. For shift ciphers the most frequent letter in the ciphertext usually stands for E and assuming that gives you the key. Failing that, you assume it stands for the second most frequent letter T, and so on. For affine ciphers you need to make two assumptions in order to solve for a, b in the encryption function $e_k(x) = ax+b \pmod{26}$. Usually you assume that the most frequent letter in the ciphertext stands for E and the second most frequent for T - failing that try other likely combinations.

Today we use frequency analysis on the remaining type of monoalphabetic cipher, the general substitution cipher where any permutation of the letters could be the key. Suppose the ciphertext is as follows:

LIVITCSWPIYVEWHEVSRIQMXLEYVEOIEWHRXEXIPFEMVEWHKV
STYLXZIXLIKIIXPIJVSZEYPERRGERIMWQLMGLMXQERIWGPSRIH
MXQEREKIETXMJTPRGEVEKEITREWHEXXLEXXMZITWAWSQWXS
WEXTVEPMRXRSJGSTVRIEYVIEXCVMUIMWERGMIWXMJMGCSM
WXSJOMIQXLIVIQIVIXQSVSTWHKPEGARCSXRWIEVSWIIBXVIZMX
FSJXLIKEGAEWHEPSWYSWIWIEVXLISXLIVXLIRGEPIRQIVIIBGIIHM
WYPFLEVHEWHYPSRRFQMXLEPPXLIECCIEVEWGISJKTVWMRLIH
YSPHXLIQIMYLXSJXLIMWRIGXQEROIVFVIZEVAEKPIEWHXEAMWY
EPPXLMWYRMWXSGSWRMHIVEXMSWMGSTPHLEVHPFKPEZINTC
MXIVJSVLMRSCMWMSWVIRCIGXMWYMX

Counting, we find that I is the most common single letter, that XL is the most common digram (meaning pair of successive letters), and XLI the most common trigram (three successive letters) in the ciphertext. Since in English E is the most common letter, TH is the most common digram, and THE the most common trigram, we guess that I stands for E, X stands for T, and L stands for H. Indeed X is the 2nd most frequent letter in the ciphertext. We guess that the 3rd most frequent letter E in the ciphertext stands for the 3rd most frequent letter in English, A. Let us write our guesses in lower case. So far we have:

heVeTCSWPeYVaWHaVSReQMthaYVaOeaWHRtatePFaMVaWHKVS
TYhtZetheKeetPeJVSZaYPaRRGaReMWQhMGhMtQaReWGPSReHM
tQaRaKeaTtMJTPRGaVaKaeTRaWHatthattMZeTWAWSQWtSWatTVaP
MRtRSJGSTVReaYVeatCVMUeMWaRGMeWtMJMGCSMWtSJOMeQ
theVeQeVetQSVSTWHKPaGARCStRWeaVSWeeBtVeZMtFSJtheKaG
AaWHaPSWYSWeWeaVtheStheVtheRGaPeRQeVeeBGeeHMWYPFh
aVHaWHYPSRRFQMthaPPtheaCCeaVaWGeSJKTVWMRheHYSPHth
eQeMYhtSJtheMWReGtQaROeVFVeZaVAaKPeaWHtaAMWYaPPthMW
YRMWtSGSWRMHeVatMSWMGSTPHhaVHPFKPaZeNTCMteVJSVhM
RSCMWMSWVeRCeGtMWY

Now what? You see things that somewhat confirm your guesses, like 'that', and things that suggest new guesses. For instance, 'Rtate' might be 'state' and 'atthattMZe' could be 'at that time', and 'heVe' might be 'here'. This suggests that R,M,Z,V stand for S,I,M,R respectively. Putting these in gives:

hereTCSWPeYraWHarSseQithaYraOeaWHstatePFairaWHKrSTYhtme
theKeetPeJrSmaYPassGaseiWQhiGhitQaseWGPSseHitQasaKeaTtiJTP
sGaraKaeTsaWHatthattimeTWAWSQWtSWatTraPistsSJGSTrseaYreat
CriUeiWasGieWtiJiGCSiWtSJOieQthereQeretQSrSTWHKPaGAsCStsW
earSWeeBtremitFSJtheKaGAaWHaPSWYSWeWeartheStherthesGaPes
QereeBGeeHiWYPFharHaWHYPSssFQithaPPtheaCCearaWGeSJKTrW
isheHYSPHtheQeiYhtSJtheiWseGtQasOerFremarAaKPeaWHtaAiWYaPP
thiWYsiWtSGSWsiHeratiSWiGSTPHharHPFKPameNTCiterJSrhisSCiWiS
WresCeGtiWYit

The way I proceeded now is to guess that remarAaKPe is remarkable so A,K,P stand for K,B,L respectively. Then I guessed the next part is 'taking all things into consideration', which yielded a lot of information and then I picked off the other letters easily. I didn't have to take back any guess. The plaintext (with spaces inserted) comes out to be:

hereupon legrand arose with a grave and stately air and brought me the beetle from a glass case in which it was enclosed it was a beautiful scarabaeus and at that time unknown to naturalists of course a great prize in a scientific point of view there were two round black spots near one extremity of the back and along one near the other the scales were exceedingly hard and glossy with a llthe appearance of burnished gold the weight of the insect was very remarkable and taking all things into consideration i could hardly blame jupiter for his opinion respecting it

A Google search identifies this as part of Edgar Allan Poe's 'The Gold Bug'. This and Conan Doyle's tale 'The Adventure of the Dancing Men' are the two most famous fictional instances of frequency analysis being used to break a substitution cipher.

Notes: using an editor simplifies the process. I used one that replaces all instances of one letter with another, e.g. I by e (turn off the 'ignore case' option!) The final decryption key sends A,B,...,Z to k,x,p,q,a,y,c,d,e,f,b,h,i,j,v,l,w,s,o,u,z,r,n,t,g,m respectively (note that no D appears in the ciphertext but it has to correspond to q since that's the only letter remaining).