MATH 587/CSCE 557 - SUMMARY OF CLASS, 3/1/07

Today we solved stage 4 of Simon Singh's Cipher Challenge. This is a Vigenere cipher and we focused on the most efficient way to obtain the keyword and hence the plaintext.

First of all, we went after the keyword length. The quickest way to guess it is to use Kasiski's attack. We found some repeated long strings - NUOCZGM occurs twice with separation 80; WXIZAYG occurs twice with separation 190; DOEOY occurs twice with separation 45. Since these are such long strings and the separations are all divisible by 5 and nothing larger, this suggests that the keyword has length 5. We'll get extra evidence for this when we go over to the next stage, namely that of frequency analysis for $\{y_1, y_6, y_{11}, ...\}$, for $\{y_2, y_7, y_{12}, ...\}$, and so on.

We split into groups to do frequency analysis for each of the above subsets of the ciphertext. For $\{y_1, y_6, y_{11}, ...\}$, we found that W occurred overwhelmingly most frequently (22 times versus S coming second, occurring only 12 times). This suggests that the 1st letter of the keyword shifts E to W, a shift of 18. Since 18 is S, we guess that the first letter of the keyword is S. At the same time, the frequencies found let us calculate that the index of coincidence of $\{y_1, y_6, y_{11}, ...\}$ is 0.08509, which is certainly large and so is extra evidence that 5 is the right keyword length.

Likewise, the most frequent letter of $\{y_2, y_7, y_{12}, ...\}$ was by far G. The guess that this is the encryption of E means that the second keyword letter is shift by 2 so is C. The index of coincidence of this subset of the ciphertext is 0.07196, again large.

Continuing the same way, the remaining three letters of the keyword come out as U,B,A and the corresponding indices of coincidence are $0.06756, 0.06872, 0.08775$, all large. It's interesting that the indices of coincidence are all larger than the usual 0.066 for English. All becomes clear when we subtract SCUBASCUBASC... from the ciphertext to get a plaintext SOUVENT POUR S'AMUSER LES HOMMES D'EQUIPAGE PRENNENT ..., which is French (typical text in French has an index of coincidence of about 0.074). It makes sense, so we have found the keyword and the plaintext.

I spoke a little about one-time pads where the encrypter and decrypter agree in advance on a very long random string of letters (longer than any plaintext to be enciphered), which will be the key. The ciphertext is created by adding the key to the plaintext. If used only once, the one-time pad has perfect secrecy, in that an eavesdropper who intercepts the ciphertext is no wiser, since the ciphertext is equally likely to have come from ANY plaintext. We'll talk later about pseudo-random sequences and linear feedback shift registers (LFSR's), which attempt to mimic the one-time pad. One-time pads are not practical unless we have some way to distribute the same key to encrypter and decrypter - this problem of key distribution leads to the introduction of public-key cryptography.

We began to discuss entropy, which will lead into topics like redundancy and language compression.