Assignment 1

Policy: You may discuss problems with others, but all written work submitted must be your own. *You may not copy nor solicit solutions, complete or partial, from any source.*

1. Alice and Bob want to write encrypted messages to a diary so that after decrypting the message they will know who wrote which message. They decide on the following method:
   1. All messages of Alice will start with n 0s; whereas
   2. All messages of Bob will end with n 0s; and
   3. No one will write the message containing all 0s.

   So if Alice wants to write a message m to the diary, she will encrypt the message $0^n||m$ where 0^n is a string of n 0s, and $||$ denotes concatenation. Likewise, Bob's messages will be of the form $m||0^n$. Assume that m is also of length n and $m \neq 0^n$. Note that with this encoding, each string that Alice and dBob write in the diary is of length $2n$ and it is never 0^{2n}.

   To encrypt their messages Alice and Bob agree to use the "one-time" pad and jointly select a random key k of length $2n$ which they will use to encrypt and write their messages to the diary.

   Show how to decrypt all the messages in the diary without knowing the key k as soon as both Alice and Bob written one string each in the diary. Also, show how to recover the key k.

   Answer: Let m_a be Alice's message, so it will have the form

   $$m_a = 00\ldots0m_{a1}m_{a2}\ldots m_{an}.$$

   On the other hand, the Bob message m_b will have the form

   $$m_b = m_{b1}m_{b2}\ldots m_{bn}00\ldots0.$$

   Let $k = k_1k_2\ldots k_nk_{n+1}\ldots k_{2n}$ be the key. Let $c_a = k \oplus m_a$ and $c_b = k \oplus m_b$ be the encryptions. Adding these together gives

   $$c_a \oplus c_b = (k \oplus m_a) \oplus (k \oplus m_b) = (k \oplus k) \oplus (m_a \oplus m_b) = m_a \oplus m_b$$

   This will be completely readable, as it will be Bob's message followed by Alice's message by the construction:

   $$m_a \oplus m_b = m_{b1}m_{b2}\ldots m_{bn}m_{a1}m_{a2}\ldots m_{an}.$$

   As for the key, since the first half of m_a is 0^n, the first half of c_a will be $k_1\ldots k_n$. Similarly, the second half of c_b will be $k_{n+1}\ldots k_{2n}$. If Eve can guess correctly who wrote the first message and who wrote the second message, then Eve can figure out the key. It would be both $m_a + c_a$ and $m_b + c_b$. Otherwise, there would be a second possibility, the one obtained under the (incorrect) guess that c_b was sent by Alice and c_a was sent by Bob. In this case, the key would seem to be $m_a + c_b = m_b + c_a$. As soon as somebody sends another different message, the key should be clear (half of it will be appear in the cipher again).

2. Two 8-bit strings m_1 and m_2 were sent after each being encrypted with a "one-time" pad using the same 8-bit key k. We see the resulting ciphers $c_1 = 10010010$ and $c_2 = 10110000$. We are pretty sure that either m_1 and m_2 are the ASCII binary codes for either 'M' (01001101) and 'o' (01101111) respectively; or 'I' (01001001) and 'l' (01101100) respectively. They were sending the letters of the abbreviation of Missouri (Mo) or Illinois (Il). Which is correct? What was the key k?

   Answer:

   $$c_1 + c_2 = (m_1 \oplus k) + (m_2 \oplus k) = (m_1 \oplus m_2) + (k \oplus k) = (m_1 \oplus m_2)$$

   So we can compare

   $$c_1 + c_2 = 10010010 \oplus 10110000 = 00100010$$

   to the two possiblities. For 'M' and 'o':

   $$m_1 + m_2 = 01001101 \oplus 01101111 = 00100010,$$

   well that's it. Of course it would be Missouri! Let's double check that it can't be the other possiblity:

   $$m_1 + m_2 = 01001001 \oplus 01101100 = 10100101.$$

   Since $c_1 = m_1 \oplus k$, we have $k = c_1 \oplus m_1$, so

   $$k = c_1 \oplus m_1 = 10010010 \oplus 01001101 = 11011111$$

3. When using the one-time pad with the key $k = 0^n$, it follows that $Enc_k(m) = m$, and the message is effectively sent in the clear! It has therefore been suggested to improve the one-time pad by only encrypting with a key $k \neq 0^n$ (i.e. to have Gen choose k uniformly at random from the set of non-zero keys of length n). Is this an improvement? In particular, is it still perfectly secret? Prove your answer.

   Answer: The message space has size $|M| = 2^n$. If we remove this one choice of key, $|K| = 2^n - 1 < |M|$, so it is not perfectly secret by Shannon's Theorem. Specifically here, for any possible message m_0, $(P[m = m_0] > 0)$

   $$P_{k,m}[m = m_0 | \text{Enc}_k(m_0) = m_0] = 0,$$

   so it's not Shannon secret (nor perfectly secret).

4. We will consider how secure the two historical ciphers in the text are. Our message space will be strings of the English alphabet.

   (a) Prove that if only a single character is encrypted, then the shift cipher is perfectly secret.

   Answer: Let's represent the letters as $\{0, 1, \ldots, 25\}$. Let m be a message and c be a cipher, which are each numbers in this set. Then the key k is also an integer in this set, which we will choose uniformly at random. The encryption is addition mod 26.

   $$P_k[\mathrm{Enc}_k(m) = c] = P_k[m + k = c \bmod 26] = P_k[k = c - m \bmod 26] = 1/26.$$

   Since m was arbitrary and this probability is a constant, the encryption is perfectly secret.

   (b) What is the largest message space $\mathcal{M} \subseteq \{a, b, \ldots, z\}^5$ (all 5-letter strings) you can find for which the mono-alphabetic substitution cipher provides perfect secrecy?

   Answer:

   If we take M to be all $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22$ strings with distinct letters, then an encryption of a message is equally likely to be any string with distinct letters. Let $m = m_1 m_2 m_3 m_4 m_5$ be a message and $c = c_1 c_2 c_3 c_4 c_5$ be a possible cipher (with both these strings, all characters of each are distinct). A key K will represent a bijection $K : A \to A$, where A represents the alphabet. Then

   $$P_k[Enc_k(m) = c] = P_k[K(m_1) = c_1, \ldots, K(m_5) = c_5].$$

   The probability that K maps these characters to these values is $21!/26!$. Since m was arbitrary and this probability is a constant, the encryption is perfectly secret.

5. Prove an analogue of Shannon's Theorem for the case of "almost perfect" secrecy. That is, let $\epsilon < 1$ be a constant and say that we only require that for any distribution over $\mathcal{M}$, any $m' \in \mathcal{M}$, and any $c \in C$;

   $$\big| P[m = m' \mid Enc_k(m) = c] - P[m = m'] \big| \leq \epsilon.$$

   Prove a lower bound on the size of the key space $\mathcal{K}$ relative to $\mathcal{M}$ for any encryption scheme that meets this definition.

   Hint 1: For simplicity, assume the uniform distribution over $\mathcal{M}$ and assume that for each k, Enc_k is deterministic. Note that the lower bound should agree with Shannon's Theorem for the case $\epsilon = 0$.

Hint 2: This one is more challenging, in particular because for a given message m and a cipher c, if there is a key k such that $Enc_k(m) = c$, this key is not necessarily unique. You may want to first see what bound you can arrive at when you assume that the key would be unique.

Answer: With a uniform distribution we have $P[m = m'] = 1/|M|$. Consider a particular message $m = m'$ and a key $k' \in K$, and let $c = Enc_{k'}(m')$. We want to find an expression or a bound for $P_{k,m}[m = m' \mid Enc_k(m) = c]$. Using Bayes' Theorem,

$$\begin{aligned}
P_{k,m}[m = m' \mid Enc_k(m) = c] &= \frac{P_{k,m}[Enc_k(m) = c \mid m = m']P_m[m = m']}{P_{k,m}[Enc_k(m) = c]} \\
&= \frac{P_k[Enc_k(m') = c]P[m = m']}{P_{k,m}[Enc_k(m) = c]} \\
&= \frac{P_k[Enc_k(m') = c]P[m = m']}{\sum_{m'' \in M} P_k[Enc_k(m'') = c]P[m = m'']} \\
&= \frac{P_k[Enc_k(m') = c]P[m = m']}{P[m = m'] \sum_{m'' \in M} P_k[Enc_k(m'') = c]} \\
&= \frac{P_k[Enc_k(m') = c]}{\sum_{m'' \in M} P_k[Enc_k(m'') = c]},
\end{aligned}$$

since the message distribution is uniform. For each message m (and this fixed c), define $K_m = \{k \in K : Enc_k(m) = c\}$. Then $P_k[Enc_k(m') = c] = |K_{m'}|/|K|$. Let x denote the sum in the denominator. Then

$$x = \sum_{m'' \in M} |K_{m''}|/|K| = (1/|K|) \sum_{m'' \in M} |K_{m''}|.$$

For a given key $k \in K$, it can be in at most one set $K_{m''}$. Otherwise, there would be two messages m_1 and m_2 such that $Enc_k(m_1) = c = Enc_k(m_2)$, violating a condition of a valid encryption scheme (we must be able to uniquely decrypt). Therefore,

$$\sum_{m'' \in M} |K_{m''}| \leq |K|,$$

so $x \leq 1$.

Now, overall we have $P_{k,m}[m = m' \mid Enc_k(m) = c] = \frac{|K_{m'}|}{x|K|}$. Suppose that $|K| < |M|$. Then this expression exceeds $1/|M|$, so we have

$$\big|P_{k,m}[m = m' \mid Enc_k(m) = c] - P[m = m']\big| = \Big|\frac{|K_{m'}|}{x|K|} - \frac{1}{|M|}\Big| = \frac{|K_{m'}|}{x|K|} - \frac{1}{|M|}.$$

Now using the bound from our hypothesis

$$\frac{|K_{m'}|}{x|K|} - \frac{1}{|M|} \leq \epsilon.$$

Isolating $|K|$ in this expression leads to

$$|K| \geq \frac{|K_{m'}||M|}{x(\epsilon|M|+1)}.$$

Since it was established at the beginning that $c = Enc_{k'}(m')$, we have $|K_{m'}| \geq 1$. As established earlier, $x \leq 1$. This means that

$$|K| \geq \frac{|K_{m'}||M|}{x(\epsilon|M|+1)} \geq \frac{|M|}{\epsilon|M|+1},$$

our lower bound under the assumption that $|K| < |M|$. Otherwise, $|K| \geq |M| \geq \frac{|M|}{\epsilon|M|+1}$, so in either case

$$|K| \geq \frac{|M|}{\epsilon|M|+1}.$$