HONEY CHATTING: A NOVEL INSTANT MESSAGING SYSTEM ROBUST TO EAVESDROPPING OVER COMMUNICATION

Joo-Im Kim, Ji Won Yoon*

Center for Information Security Technologies (CIST), Korea University, {jooimkim, jiwon_yoon}@korea.ac.kr

INTRODUCTION

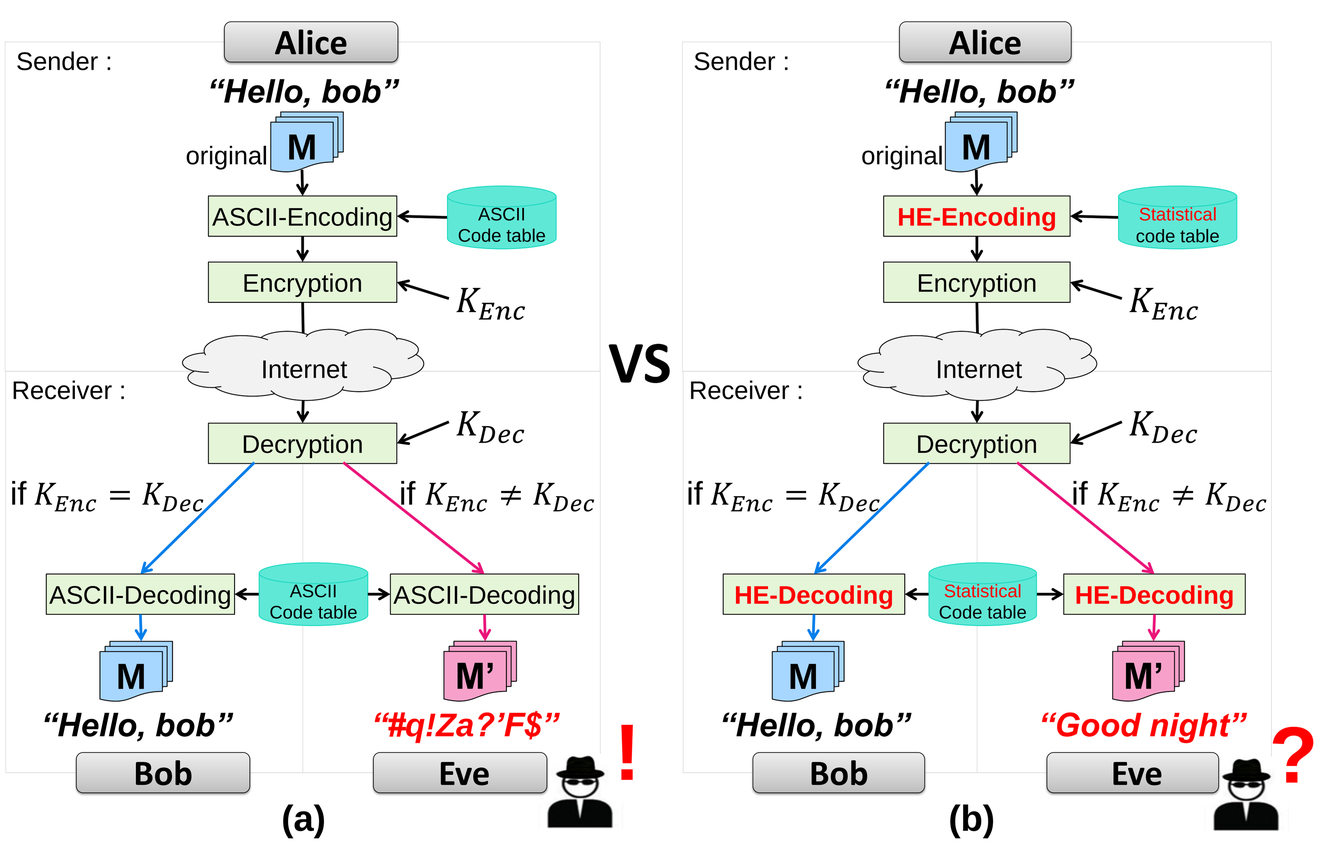
Secure Chatting

- One of methods to strengthen the security of Instant Messaging system is the message encryption.
- However, the key used for encryption has potential vulnerability to be cracked by a brute-force attacker if the key size is not long enough.
- So, we introduce a new concept of secure chatting by applying Honey Encryption which makes decrypted texts with wrong keys hard to be distinguished from the decrypted text with a real key.

Our Work

- We develop a Instant Messaging system(Honey Chatting) robust to eavesdropping by using the basic idea of Honey Encryption.
- In our system, we generate plausible-looking but fake plaintexts by using statistical encoding/decoding scheme to confuse the brute-force attacker.
- Through simple experiments, we show the difference between a real message and fake messages by calculating the entropy of texts in the decrypted message.

STRUCTURE SUMMARY



Overall Procedure

1) The sender's message **M** is encoded using the code table and encrypted with K_{Enc}.
2) It passes through the communication channel such as Internet.
3) The receiver decrypt it with K_{Enc} and decode it using the same code table.
   If $K_{Enc} = K_{Dec}$, the receiver can obtain a true message in both cases.
   Else If $K_{Enc} \neq K_{Dec}$, **M'** is become **false message in (a)** or **plausible-looking fake message in (b)**.

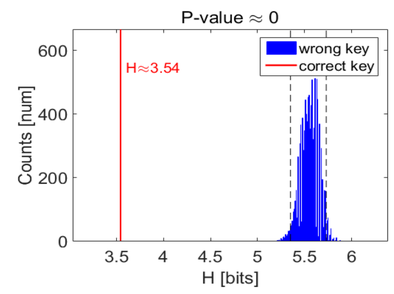
➔ **Therefore, Eve(brute-force attacker) would confuse to find real message.**

❖ Here, the **Statistical code table** is made from the statistical coding scheme using text corpus in advance, and the sender and the receiver share it.

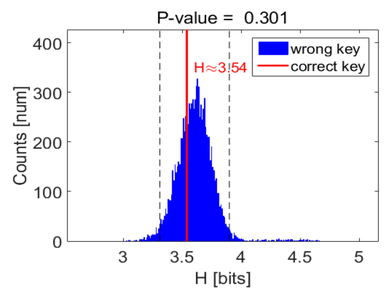
EXPERIMENT

The Difference of Entropy

- We conduct a significance test(hypothesis test) to show difference between decrypted text with wrong and real key when applying HE scheme.
- In (a), P-value is significantly small which means there are clear distinction between **M** and **Ms**.
- In (b), moderately large P-value shows that observed data **M** is agreed with **Ms**. It means that **M** is similar with **Ms**, so the brute-force attacker could not notice his success.



(a) ASCII (L=50)



(b) HE (L=50)

Statistical Coding Scheme

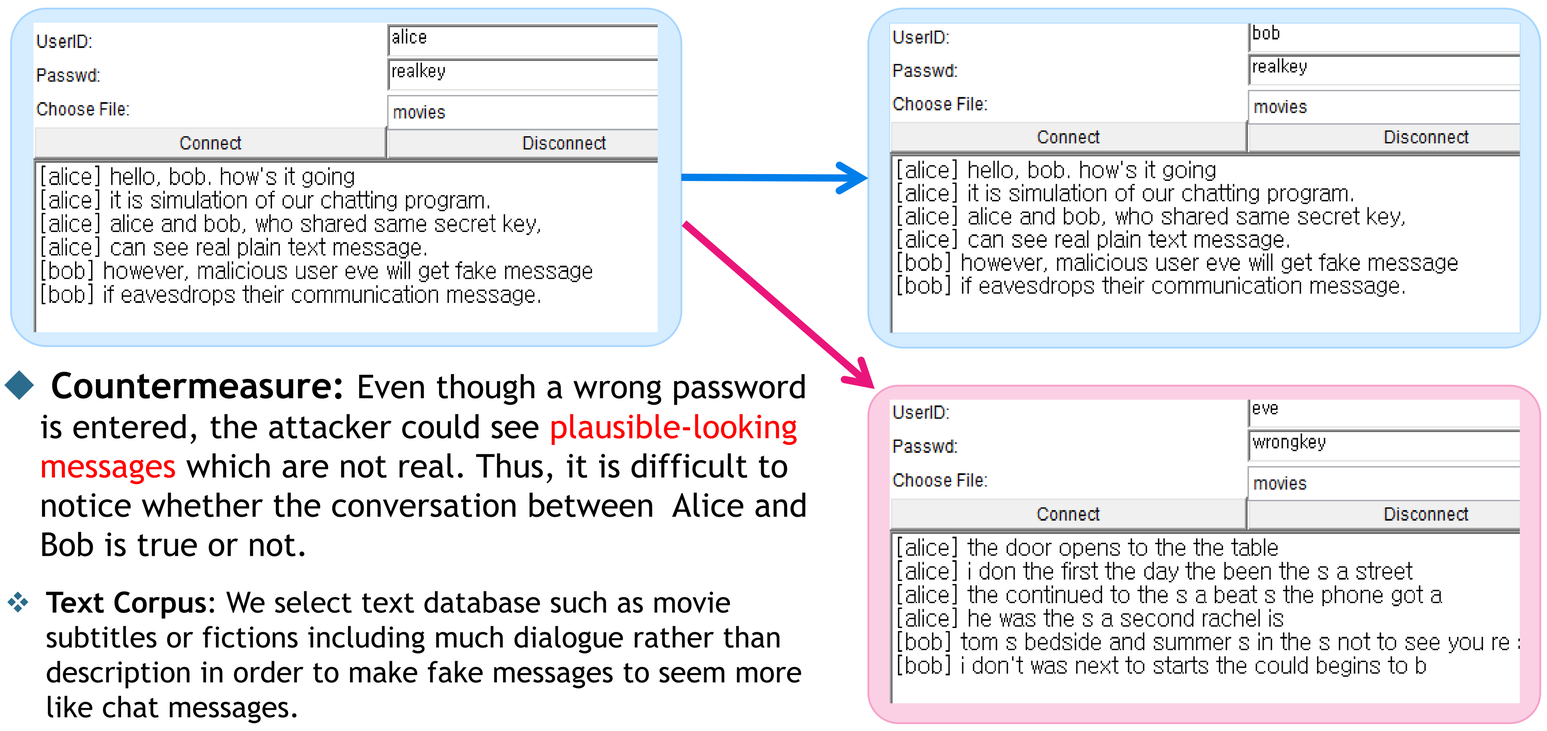
- Chat messages can be represented by N-gram language model, so we get the probability of consecutive characters in a sentence.
- We construct the cumulative massive function(CMF) based on the N-gram language model. CMF is used as statistical code table for HE-Encoding and HE-Decoding.
- The **CMF** for i-th character of message :

$$p_{\mathrm{cmf}}^{(\mathrm{i})}(c_k) = \sum_{k=0}^{S} \frac{p(x_i = c_k | \mathbf{x}_{i-1:i-n})}{\sum_{j=0}^{S} p(x_i = c_j | \mathbf{x}_{i-1:i-n})}$$

- S is the number of possible character set // *a~z, space, comma, period*
- n is the order of markov process. // *n=5 in our application*
- $p(x_i|x_{i-1:i-n})$ is the i-th character influenced by previous n-1 characters

HONEY CHATTING SIMULATION

Situation: While **Alice** and **Bob** enjoy chatting (share a real password), a malicious **Eve** is trying to eavesdrop their chat messages (try to enter wrong passwords).



Countermeasure: Even though a wrong password is entered, the attacker could see plausible-looking messages which are not real. Thus, it is difficult to notice whether the conversation between Alice and Bob is true or not.

❖ **Text Corpus**: We select text database such as movie subtitles or fictions including much dialogue rather than description in order to make fake messages to seem more like chat messages.

CONCLUSION

Summary

- There are many chatting systems, which enhance security with technology such as message encryption.
- But the key is fundamentally vulnerable to a brute-force attack.
- Through the approach in this paper, we could build a messaging system which is robust to eavesdropping.

Future Improvement

- For the practical use in real world, we need to consider the context and grammar of messages.
- Also, the available character set should be increased. Now 30 characters: letters(a-z), space, period, and comma.
- Consider other measures and experiment methods to prove the indistinguishability of decrypted messages.