Context Encryption using Natural Language Processing

Babu Priyavrat
Amdocs
Kuala Lumpur, Malaysia

Abstract

With the advancement in Natural Language processing, it is now possible to devise schemes to change the context of sentence, paragraph or article so that the original context is hidden. Such schemes would be required in post-quantum world. I propose a novel method of Context Encryption which will contextually encrypt the sentence. The proposed model adds another layer to the encryption model of PGP. The proposed system will be using Stanford NLP Parser to get the structure of sentence and replace it with the sentence of similar structure.

1. Introduction

Since ancient times, people are trying to hide the information to keep it secret. In 20th century, the encryption took a whole new level with the advent of computer science. Various symmetric and asymmetric encryption schemes were developed like RSA. In theory, every encryption key is breakable. The strength of encryption schemes lies in the computation time required to find the key to decrypt. The longer the time to break the key, the harder is to take the break encryption method.

However, with increasing computation speed and development of new quantum computers like D-wave systems, the barrier to compute the decryption key will be lowered. There is another way to hide the key: to prevent the hacker to know the message is encrypted.

Context encryption would be the means by which we can encrypt the context of sentence or paragraph and still make it human readable. A context can be derived from nouns, verbs or adjectives in the text.

2. Principles and Approaches suggested

2.1 Structural integrity

With the advancement of NLP, it is possible to determine the semantics of the sentence. Thus, if semantics of the sentence can be changed without realizing the inherent change in the structure, it can be encrypted semantically.

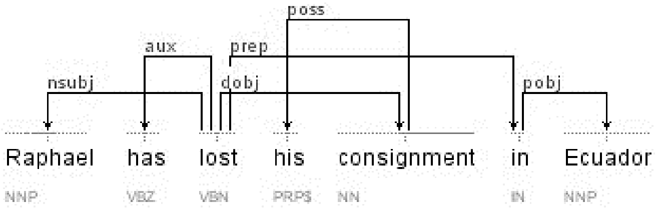
A known database of sentence can be acquired using search engine which picks up sentences and paragraphs from public domains like News, educational websites, etc.

It replaces the sentence in the private message with a sentence in the public domain.

For example:
Secret agent in Ecuador delivers a text message.

"Raphael has lost his consignment in Ecuador."

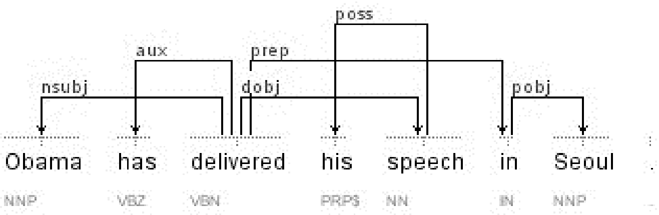
Using Stanford's typed dependencies, we can determine the structure of the sentence.



The above dependency is generated by typed dependencies Viewer [2]
Abbreviations can be found in Section in section 5.

Once the structure is extracted, it is contextually encrypted using the information available in the public domain. To maintain the structural integrity, only those sentences having the same typed dependencies are constructed.

The original sentence in public context cloud can be 'Obama delivered his speech in Seoul'. However, to have same typed dependencies as original sentence, the sentence structure from public context is modified. This process is called syntactic transformation which is further discussed in section 3.3.



The above dependency is generated by typed dependencies Viewer [2]

Abbreviations can be found in Section in section 5.

2.2 Cohesiveness

This aspect of encryption is to make believe human reader that the encrypted text is not encrypted further and make them intelligible to human users. It has to make human readers think that phrases in paragraph or passage are related to the same context. It is measured by coherence. Implementing this principle of encryption is more challenging than usual as it requires the selection of the correct public text available.

To find the correct public context, it requires the analysis of various texts in email, gather typed dependencies and scan the public contextual database or cloud to find the paragraphs or sentences which have the same number of sentences and have similar type dependencies. Thus, to establish coherence, Text Tiling is used which automatically segregate the message into sub-topics. Any public context chosen for context encryption should have similar coherence levels.

To understand the complexity, let's take an example of e-mail that a friend sends an e-mail asking about the job. She might want to encrypt contextually certain segments of the message.

> Hi Drew,
> I hope all is well with you on the East Coast! It has been pretty busy these past few years at my firm in San Francisco, but I'm having a blast and learning a lot.
>
> A favor to ask of you: I'm in the process of exploring new career opportunities and am wondering if you're able to connect me to any interesting people or companies.
>
> As you know from our work together, I've developed an interest in the area of socially responsible business. In particular, I'm hoping to combine my experience in finance with my passion of sustainable business and socially responsible investing. Are you still involved with the Green Party in New York? Can you point me to anyone in your network who might be able to help? Thank you!

The above mail is a sample letter provided by HAAS [1]

In the above example, she might want to hide the context of message highlighted in grey. Thus, this requires pairing of private context with typed dependencies. It is quite possible that none of the context is available for encryption. In that case, sentence from public context has to be transformed. This is later discussed in section 4.3.

2.3 Mutual exclusivity of public and private context

The private encrypted text has some context. The public context should be mutually exclusive of private context. The construction of the private and public context is explained later in section 4.1 and 4.2

One of the main reasons of mutual exclusivity is to ensure that private context can never be deciphered from public context.

2.4 Separating Private context and Data

The Data and private context can be sent in separate messages to improve the security. In diagram 2, contextually encrypted data and private context are shown together for simplicity.

The better way of securing information is to deliver the encrypted message first and deliver the private context only when the receiver opens the message. Such mechanisms will greatly enhance the security and privacy of data as anyone listening, capturing and cracking all encrypted data will never come to know whether it is contextually encrypted.

When the private context is captured, it is difficult to correlate with the message that was transmitted earlier.

2.5 Boundaries of Contextual Encryption

There are certain areas that contextual encryption has to be limited or not to be done. For example, in the above example, Drew cannot be encrypted as clearly as the sender is addressing Drew. There are other things which cannot be contextually encrypted like phone numbers, historical dates, etc.

Thus, it is important to give control to the users what they want to contextually encrypt. The example shown in section 3.1 is example where the user prefers to contextually only a part of the mail.

3. Proposed Steps in Context Encryption

3.1 Constructing the private context

A context is dictionary of words (excluding preposition and conjunction) which appears in any article, paragraph or sentence. Pronouns and adverb are usually avoided in making private context. Once the structure is evaluated, the process responsible for context encryption will identify the context of the current sentence by developing a dictionary of keywords

This context is called private context**.**

From the previous example,
Raphael, lost, consignment, Ecuador forms a private context.

The private context is stored in private context-typed dependencies (PCTD) map.

A PCTD map will have following attributes only:
1) Word position in the sentence
2) Word
For example, the PCTD map for the above context is illustrated below:

Raphael: 1, Lost: 3, Consignment: 5, Ecuador: 7

3.2 Getting the public context

All the public contexts are stored in Non-relational database like Mongo DB, which would be running on cloud.

'Public context cloud' as shown in Diagram 1 and Diagram 2 would require construction of public contexts by scanning news portal, books, etc. and indexation of public contexts of varying sizes according to typed dependencies.

Once the indexation is done, new API called getPublicContext API have to be written to acquire public context by mail client. This has to be done in stipulated time and within seconds after the mail is sent. Retrieving the correct public context which is mutually exclusive of private context will require an enormous amount of computing capacity. This can be achieved by querying Mongo DB using Hadoop and MapReduce system.

The public context is constructed by analyzing the document in public domain like webpages. While storing public context, following information need to be stored.

1) Context unique identifier
2) Word position in the sentence
3) Word
4) Grammatical attributes
5) Typed Dependencies
6) Sentence position

Each context is assigned a unique identifier. A random number generator picks up the number, and the corresponding context is picked up. This context is evaluated if it is mutually exclusive and similar typed dependencies of private context. The mutual exclusivity check is done only on Nouns and verbs present in the public context. If it is mutually exclusive, the following public context is used.

From the previous example,

Obama, deliver, speech, Seoul forms a valid public context for encryption as it is mutually exclusive with private context.

3.3 Syntactic transforming the public context

A sentence from public context might require some transformation to rebuild the original sentence. It is further illustrated by the below example:
For example, 'Obama delivered his speech in Seoul' is picked up from the public context cloud has the following typed dependencies tree.

```
(ROOT
  (S
    (NP (NNP Obama))
    (VP (VBD delivered)
      (NP (PRP$ his) (NN speech))
      (PP (IN in)
        (NP (NNP Seoul))))))
```

The above typed dependency tree is generated by Stanford parser [3]
In order to contextually encrypt the sentence using private context dependency tree, 'has' to be inserted.

Parent: S
Node: (VP (VBZ has)
Child: (VP (VBD delivered)
Operation: insertion

Once the tree structure is similar, the text generated is contextually encrypted text.

'Obama has delivered his speech in Seoul.'

Other kinds of syntactic transformation might be required to transform.
Few of them are listed below:

1. Topicalization
2. Passivization
3. Extraposition
4. Preposing
5. There-construction
6. Pro-normalization
7. Fronting
8. Locative alternation

3.4 Constructing contextually encrypted message

As the public context will have same typed dependencies tree, the PCTD map contains the position of element to be replaced by private context when the typed dependencies tree is traversed from left to right and no other attribute need to be present in other than word and word Position in PCTD map [8]

3.5 Encrypting the contextually encrypted message

Contextually encrypted message is further encrypted using PGP encryption. The contextually encrypted message is encrypted by using existing numeric encryption algorithms.

The mechanism is further illustrated in diagram 1.

3.5 Decrypting the contextually encrypted text

Once the contextually encrypted data, and private context is decrypted using a hash key as shown in diagram 2, contextually encrypted data is passed to Stanford NLP parser which generates typed dependencies tree. The original sentence is reconstructed by replacing the data in nodes of typed dependencies of contextually encrypted data with the data available in the private context.

In the example discussed, following substitutions will take place:

1. Obama →Raphael
2. Deliver → lose
3. Speech →Consignment
4. Seoul → Ecuador

Any man-in-middle attack will be unsuccessful as long as the receiver and sender machine are protected.

4. **Implications**

Adding one more layer of context encryption will provide added advantage to the users who are worried in the age of privacy violation, espionage and mass surveillance. With increasing bandwidth and computation power, overhead due to the context encryption seems affordable.

There are no long-term implications as the hacker or agencies will find a way to find the private context related to a particular message. But it would require exponential computational power to determine the private context of the message. Moreover, additional schemes of context encryption as discussed earlier will increase the difficulty to find the private context as the private context is delivered once when the receiver opens the message.

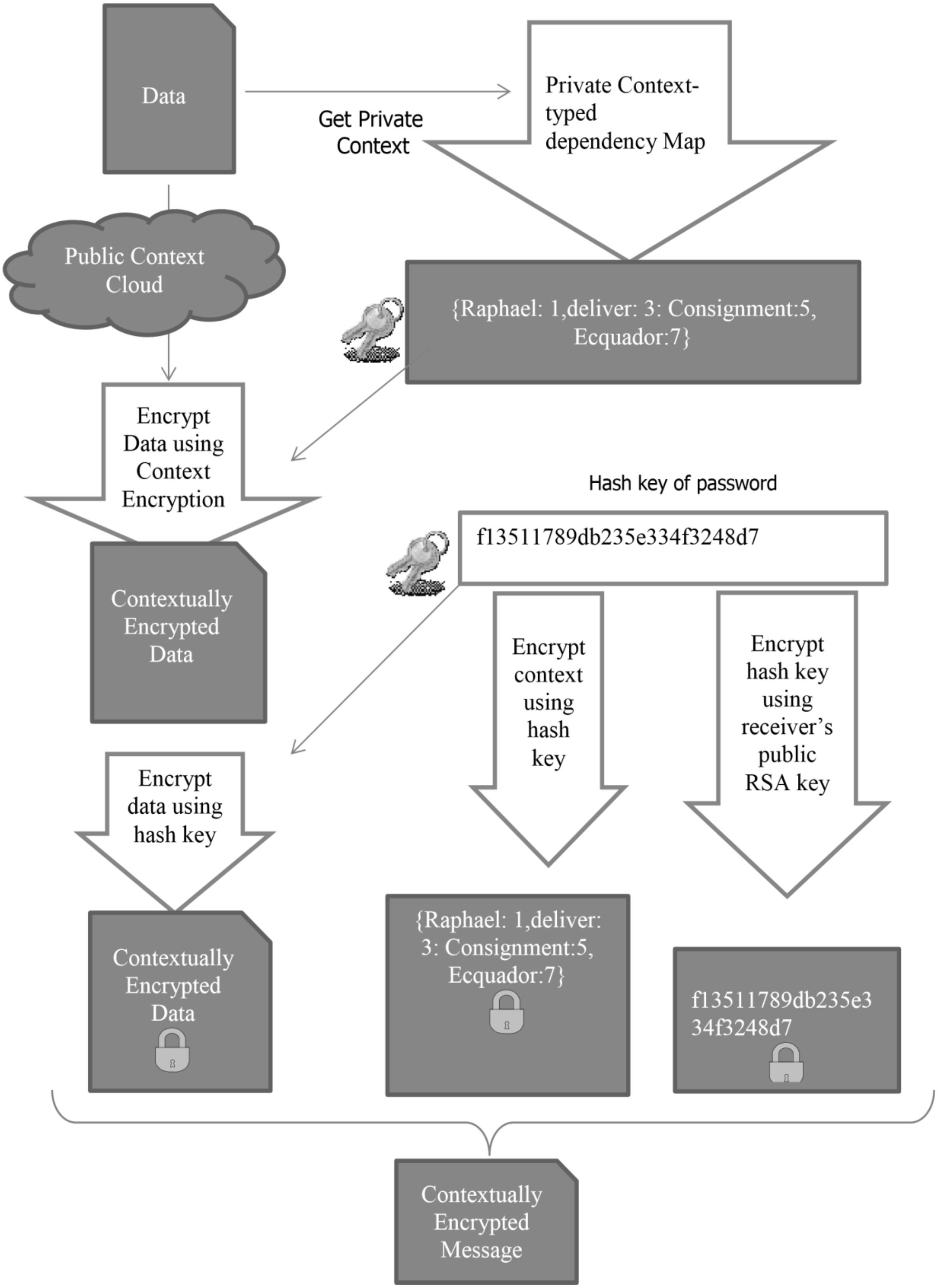

Data
Get Private Context
Private Context-typed dependency Map
Public Context Cloud
{Raphael: 1,deliver: 3: Consignment:5, Ecquador:7}
Encrypt Data using Context Encryption
Hash key of password
f13511789db235e334f3248d7
Contextually Encrypted Data
Encrypt context using hash key
Encrypt hash key using receiver's public RSA key
Encrypt data using hash key
Contextually Encrypted Data
{Raphael: 1,deliver: 3: Consignment:5, Ecquador:7}
f13511789db235e334f3248d7
Contextually Encrypted Message

Diagram 1

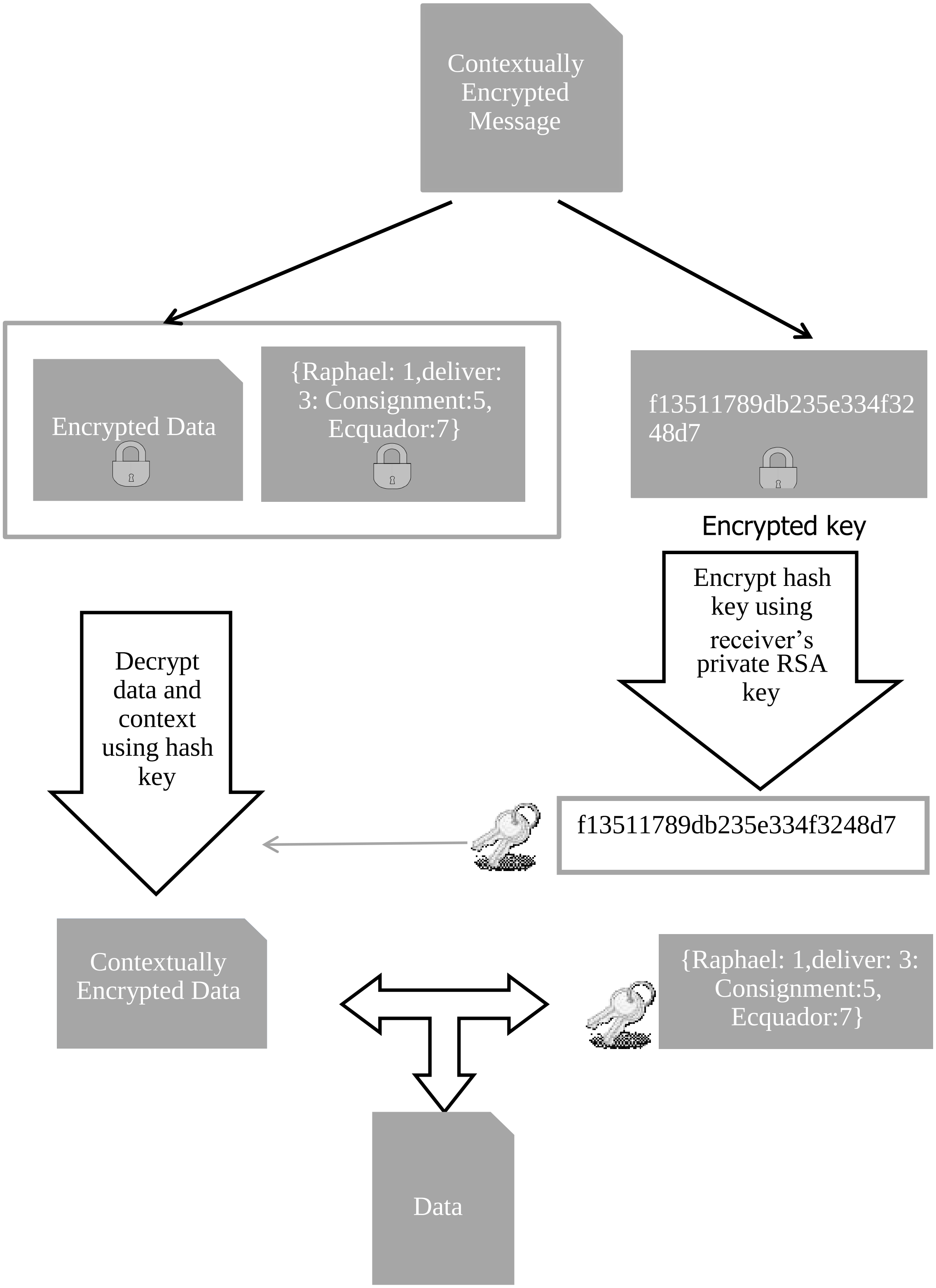

Contextually Encrypted Message
Encrypted Data
{Raphael: 1,deliver: 3: Consignment:5, Ecquador:7}
f13511789db235e334f3248d7
Encrypted key
Encrypt hash key using receiver's private RSA key
Decrypt data and context using hash key
f13511789db235e334f3248d7
Contextually Encrypted Data
{Raphael: 1,deliver: 3: Consignment:5, Ecquador:7}
Data

Diagram 2

5. Abbreviations

CC Coordinating conjunction
CD Cardinal Number
DT Determiner
EX Existential there
FW Foreign word
IN Preposition or subordinating conjunction
JJ Adjective
JJR Adjective, comparative
JJS Adjective, superlative
LS List Modal
NN Noun, singular or mass
NNS Noun, plural
NNP Proper noun, singular
NNPS Proper noun, plural
PDT Predeterminer
POS Possessive ending
PRP Personal pronoun
PRP$ Possessive pronoun RB Adverb
RBR Adverb, comparative
RBS Adverb, superlative
RP Particle
SYM Symbol
TO to
UH Interjection
VB Verb, base form
VBD Verb, past tense
VBG Verb, gerund or present participle
VBN Verb, past participle
VBP Verb, non3rd person singular present
VBZ Verb, 3rd person singular present
WDT Whdeterminer
WP Whpronoun
WP$ Possessive whpronoun
WRB Whadverbl ml
PGP Pretty Good Privacy
RSA Rivest Shamir Adleman
NLP Natural Language Processing
PTDC private context-typed dependencies

6. References

[1] E-mail Samples from HaaS, Berkeley, http://www.haas.berkeley.edu/groups/alumni/files/email-samples.pdf
[2] TyDevi, http://tydevi.sourceforge.net/
[3] Stanford NLP parser, http://nlp.stanford.edu:8080/parser/
[4] Text Encryption Algorithm Based on Natural Language Processing by Xianghe Jing, Yu Hao, Huaping Fei Zhijun Li, 2012 Fourth International Conference on Multimedia Information Networking and Security, IEEE
[5] Hearst, M. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics , 23 (1), pp. 33-64, March 1997. pdf
[6] Martín-López, Enrique; Enrique Martín-López, Anthony Laing, Thomas Lawson, Roberto Alvarez, Xiao-Qi Zhou & Jeremy L. O'Brien (12 October 2012). "Experimental realization of Shor's quantum factoring algorithm using qubit recycling". Nature Photonics
[7] Public key Encryption, Handbook of Applied cryptography, Alfred J. Menezes, Paul C. van Oorschot and Scott A. Vanstone
[8] An Optimized Natural Language Watermarking Algorithm Based on TMR, Peng Lu, Zhao Lu, Zili Zhou, Junzhong Gu, 2008, IEEE
[9] The official PGP user's guide, Philip R. Zimmermann
[10] Semantic sentence structure search engine, Nikita Gerasimov, Maxim Mozgovoy, Alexey Lagunov,
[11] Mongo DB ,MapReduce tutorial http://nosql.mypopescu.com/post/394779847/mongodb-tutorial-mapreduce
[12] MapReduce tutorial, http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html