INSTITUTO POLITÉCNICO NACIONAL

Escuelea Superior de Cómputo

Unit: Cryptography

Índice

Índice de Figuras

Índice de Tablas

Chapter 1

Introduction to security and cryptography

The communication between two entities is usually done through an insecure channel, which means that the information exchanged does not have any guarantee of privacy or integrity. Therefore, exchange of information should be provided by mechanisms or services that provide security. These security mechanisms or services are built using cryptographic tools.

In our Information Age, the need for protecting information is more pronounced than ever. Secure communication for the sensitive information is not only compelling for military or government institutions but also for the business sector and private individuals.

As the world becomes more connected, the dependency on electronic services has become more pronounced. In order to protect valuable data in computer and communication systems from unauthorized disclosure and modification, reliable non-interceptable means for data storage and transmission must be adopted.

As society has evolved, the need for more sophisticated methods of protecting data has increased. Now, with the information era at hand, the need is more pronounced than ever. As the world becomes more connected, the demand for information and electronic services is growing, and with the increased demand comes increased dependency on electronic systems. Already the exchange of sensitive information, such as credit card numbers, over the Internet is common practice. Protecting data and electronic systems is crucial to our way of living.

Figure 1 shows a hierarchical six-layer model for information security applications. Let us analyze that figure from a top-down point of view. On layer 6, several popular security applications have been listed such as: secure e-mail, digital cash, e-commerce, etc. Those applications depend on the implementation in layer 5 of secure authentication protocols like SSL/TLS, IPSec, IEEE 802.11, etc. However, those protocols cannot be put in place without implementing layer 4, which consists on customary security services such as: authentication, integrity, non-repudiation and confidentiahty. The underlying infrastructure for such security services is supported by the two pair of cryptographic primitives depicted in layer 3, namely, encryption/decryption and digital signature/verification. Both pair of cryptographic primitives can be implemented by the combination of public-key and private key cryptographic algorithms, such as the ones listed in layer 2. Finally, in order to obtain a high perfor-

mance from the cryptographic algorithms of layer 1, it is indispensable to have an efflftcient implementation of arithmetic operations such as, addition, subtraction, multiplication, exponentiation, etc.

imagen

In our course are addressed the layers 1, 2, 3 y 4.

1.0.1 Cryptography

History is filled with examples where people tried to keep information secret from adversaries. Kings and generals communicated with their troops using basic methods to prevent the enemy from learning sensitive military information. Encrypting the communication can protect it from prying eyes.

For thousands of years we have been inventing codes, and when those were broken we set about inventing better codes. *The search for the unbreakable code continues!*

The techniques needed to protect data belong to the field of cryptography. Actually, the subject has three names, cryptography, cryptology and cryptanalysis.

1.0.2 Cryptology

Is the all-inclusive term for the study of communication over nonsecure channels, and related problems. The process of designing systems to do this is called cryptography. Cryptanalysis deals with breaking such systems.

Since a dictionary **Kriptos=hide** and **graphos=write**, cryptography is the art to write in an enigmatic mode.

The study of mathematical techniques related to the aspects of information security, such as confidentiality, integrity and availability of the data, authentication of entity and origin, it doesn't includes only the media to provide information security, but to a set of techniques. Menezes & Vanstone.

Modern cryptography is a field that draws heavily upon mathematics, computer science, and cleverness.

1.1 Cryptographic services

1.1.1 Secure Communications

In the basic communication scenario, depicted in Figure 1, there are two parties, we'll call them Alice and Bob, who want to communicate with each other. A third party, Candy, is a potential eavesdropper.

When Alice wants to send a message, called the plaintext, to Bob, she encrypts it using a method prearranged with Bob. Usually, the encryption method is assumed to be known to Candy; what keeps the message secret is a key. When Bob receives the encrypted message, called the ciphertext, he changes it back to the plaintext using a decryption key.

Candy could have one of the following goals:

1. Read the message.
2. Find the key and thus read all messages encrypted with that key.
3. Corrupt Alice's message into another message in such a way that Bob will think Alice sent the altered message.
4. Masquerade as Alice, and thus communicate with Bob even though Bob believes he is communicating with Alice.

imagen

Communication Scenario for Cryptography. Which case we're in depends on how evil Candy is. Cases (3) and (4) relate to issues of integrity and authentication, respectively.

A more active and malicious adversary, corresponding to cases(3) and (4), is sometimes called Mallory in the literature. More passive observers (as in cases (1) and (2)) are sometimes named Oscar.

1.2 Attacks

1.2.1 Possible Attacks

There are four main types of attack that Candy might be able to use. The differences among these types of attacks are the amounts of information Candy has available to her when trying to determine the key. The four attacks are as follows:

1. Ciphertext only: Candy has only a copy of the ciphertext
2. Known Plaintext: Candy has copy a ciphertext and the corresponding plaintext. For example, suppose Candy intercepts an encrypted press release, then sees the decrypted release the next day. If she can deduce the decryption key, and if Alice doesn't change the key, Candy can read all future messages. Or, if Alice always starts her messages with "Dear Bob," then Candy has a small piece of ciphertext and corresponding plaintext. For many weak cryptosystems, this suffices to find the key. Even for stronger systems such as the German Enigma machine used in World War II, this amount of information has been useful.
3. Chosen plaintext: Candy gains temporary access to the encryption machine. She cannot open it to find the key; however, she can encrypt a large number of suitably chosen plaintexts and try to use the resulting ciphertexts to deduce the key.

4. Chosen ciphertext: Candy obtains temporary access to the decryption machine, uses it to "decrypt" several strings of symbols, and tries to use the results to deduce the key.

A chosen plaintext attack could happens as follows. You want to identify an aiplane as friend or foe. Send a random message to the plane, which encrypts the message automatically and sends it back. Only a friendly airplane is assumed to have the correct key. Compare the message from the plane with the correctly encrypted message. If the match, the plane is friendly. Ig not, it's the enemy. However, the enemy can send a large number of chosen messages to one of your planes and look at the resulting ciphertexts. If this allows them to deduce the key, the enemy can equip their plane so they can masquerade as friendly.

An example of a known plaintext attack reportedly happened in World War II In the Sahara Desert. An isolated German outpost every day sent an identical message saying that there was nothing new to report, but of course it was encrypted with the key being used that day. So each day the Allies had a piaintext-ciphertext pair that was extremely useful in determining the key. In fact, during the Sahara campaign, General Montgomery was carefully directed around the outpost so that the transmissions would not be stopped.

One of the most im portant assumptions in modern cryptography is Kerckhoff's principle: In assessing the security of a cryptosystem, one should always assume the enemy knows the method being used. This principle was enunciated by Auguste Kerclchoffs in 1883 in his classic treatise La Cnjptographie Militaire. The enemy can obtain this information in many ways. For example, encryption/decryption machines can be captured and analyzed. Or people can defect or be captured. The security of the system should therefore be based on the key and not on the obscurity of the algorithm used. Consequently, we always asume that Candy has knowledge os the algorithm that is used to perform encryption.

Chapter 2

Overwiew of Cryptology (and This Book)

This first chapter introduce us on the real meaning of Cryptography, this science is an old art, who the ancient Egypt used.
Since the beginig of the use of this art, another cultures start to use it as a method to hide the most important secret. Actually there are some old documents who are in anciente Greece, or the famous Caesar cipher in the ancient Rome.

imagen

We have to make an important statment the correct term is CRYPTOLOGY and not cryptography. The cryptography is the field.
Therefore cryptology split into two main branches :

- **Cryptography**: This is the science of secret writing with the goal of hiding the real meaning of a message.
- **Cryptanalysis**: This is the science and sometimes art of breaking cryptosystems. You might think that code breaking is for the intelligence community or perhaps organized crime, and should not be included in a serious classification of a scientific discipline. However, most cryptanalysis is done by respectable researchers in academia nowadays. Cryptanalysis is of central importance for modern cryptosystems: without people who try to break our crypto methods, we will never know whether they are really secure or not.
  Cryptanalysis is the only way to assure that a cryptosystem is secure.

Cryptography have their own braches.

2.1 Symmetric Algorithms

Are what many people assume cryptography is about: Two parties have an encryption and decryption method for which they share a secret key. All cryptography from ancient times until 1976 was exclusively based on symmetric methods. Symmetric ciphers are still in widespread use, especially for data encryption and integrity check of messages.

2.2 Asymmetric (Public-Key) Algorithms

In 1976 an entirely different type of cipher was introduced by Whitfield Diffie, Martin Hellman and Ralph Merkle. In public-key cryptography, a user possesses a secret key as in symmetric cryptography but also a public key. Asymmetric algorithms can be used for applications such as digital signatures and key establishment, and also for classical data encryption.

2.3 Cryptographic Protocols

Roughly speaking, crypto protocols deal with the application of cryptographic algorithms. Symmetric and asymmetric algorithms can be viewed as building blocks with which applications such as secure Internet communication can be realized. The Transport Layer Security (TLS) scheme, which is used in every Web browser, is an example of a cryptographic protocol.

In the real world the use of this algorithms is mixed, i mean you use more than one to work, and this mixtures have a name "hybrid schemes".

2.4 Symmetric Cryptography

This section deals with the concepts of symmetric ciphers and it introduces the historic substitution cipher. Using the substitution cipher as an example, we will learn the difference between brute-force and analytical attacks.

2.5 Basics

Symmetric cryptographic schemes are also referred to as symmetric-key, secret key and single-key schemes or algorithms, There are two users, Alice and Bob, who want to communicate over an insecure channel (channel might sound a bit abstract but it is just a general term for the communication link: This can be the Internet, a stretch of air in the case of mobile phones or wireless LAN communica-tion, or any other communication media you can think of) The actual problem starts with the bad guy Oscar, who has acces to the channel, for instance, by hacking into an Internet router or by listening to the radio signals of a Wi-Fi communication. This kins of stuffs is called eavesdropping. Obviously, there are many situations in which Alice and Bob would prefer to communicate without Oscar listening. For instance, if Alice and Bob represent two offices of a car manufacturer, and they are transmitting documents containing the business strategy for the introduction of new car models in the next few years, these documents should not get into the hands of their competitors, or of foreign intelligence agencies for that matter.

imagen

In this situation, symmetric cryptography offers a powerful solution: Alice encrypts her

message x using a symmetric algorithm, yielding the ciphertext y. Bob receives the ciphertext and decrypts the message. Decryption is, thus, the inverse process of encryption. What is the advantage? If we have a strong encryption algorithm, the ciphertext will look like random bits to Oscar and will contain no information whatsoever that is useful to him.

imagen

As we can see in the second image, there is the explanation of what we just reed, this method work nicely with WPA (Wi-Fi Protected Access).

2.6 Simple Symmetric Encryption: The Substitution

This is one of the simplest methods for encrypting text, SUBSTITION, in the history of the man, this is an old technic whow has been used many times, and is a good way to show us the basic cryptography.

imagen

2.7 Brute-Force

We hace the one attack, who use the brute-force this means that the attacker has the ciphertext now he would try to descovered the message using all the keys, until he found the right key. In the real world maybe you can use this but the most common things to happen is that you can't find the right or maybe you just have false results.

2.8 Frequency Analysis

The major wakness of the cipher os that each plaintext symbol always maps to the same ciphertext symbol. That means that the statistical properties of the plaintext are preserved in the ciphertext. Like we saw in the golden bug.

Chapter 3

Cryptanalysis

imagen

imagen

In this section I talk about the cryptanalysis and advices about key lengths of symmetric ciphers and different ways of attacking the algorithms.

3.1 General Thoughts on Breaking Cryptosystems

When we talk about some technical background what breaking ciphers.the people say that code breaking has to do with heavy mathematics, smart people and big computers. The computers who ejected the attacks in the World War II was so smart for this age, The size of this computer is around a room at the time goes by the people found another technicals of breaking cryptosystems.

imagen

You can see an image that contents another ways to breaking crypto systems in the real life.

3.1.1 Classical Cryptanalysis

The classical cryptanalysis is a science of recovering the plaintext form cypher text and recovering the key. The cryptanalysis can be divide into analytical attacks, which exploit internal structure of the encryption and brute force attacks, when treat te encryption algorithm as a black box.

3.1.2 Implementation Attacks

Can be used to obtain a secret key, by measuring the electrical power consumption of a processor with operates on te secret key, this power can be used to recover the key by applying signal processing techniques, at implementation attacks are mostly relevant against cryptosystems to which an attacker has physical access, such as smart cards.

3.1.3 Social Engineering Attacks

Bribing, blackmailing, tricking or classical espionage can be used to obtain a secret key by involving humans. For instance, forcing someone to reveal his/her secret key. This list of attacks against cryptographic system is certainly not exhaustive. For instance, buffer overflow attacks or malware can also reveal secret keys in software systems.

3.2 How Many Key Bits Are Enough?

In the 90's was much public discussion about the key length of ciphers.

1. he discussion of key lengths for symmetric crypto algorithms is only relevant if a brute-force attack is the best known attack. Is an analytical attack that works a large key space does not help at all.

2. The key lengths for symmetric and asymmetric key provides roughly the same security as a key.

In the next table you can see the estimated time for successful brute force attacks on different key length.

tabla

Foretelling the Future Of course, predicting the future tends to be tricky: We can't really foresee new technical or theoretical developments with certainty. As you can imagine, it is very hard to know what kinds of computers will be available in the year 2030 .

Now, these are the cost of breaking cyphersystems.
The cost for breaking the cipher will be $500,000 in 18 months (since we only have to buy half as many computers), $250,000 in 3 years, $125,000 in 4.5 years, and so on.

After 10 iterations of computer power doubling, we can do 210 = 1024 as many computations for the same money we would need to spend today. Stated diffently, we only need to spend about 1/1000th of today's money to do the same computation. In the example above that means that we can break cipher X in 15 years within one month at a cost of about $\$1{,}000{,}000/1024 \approx \$1000$.

Chapter 4

The Anonymous Codebreaker

For centuries, the simple monoalphabetic substitution cipher had been sufficient to ensure secrecy. The subsequent development of frequency analysis, first in the Arab world and then in Europe, destroyed its security. The tragic execution of Mary Queen of Scots was a dramatic illustration of the weaknesses of monoalphabetic substitution, and in the battle between cryptographers and cryptanalysts it was clear that the cryptanalysts had gained the upper hand. Anybody sending an encrypted message had to accept that an expert enemy codebreaker might intercept and decipher their most precious secrets. The burden was clearly on the cryptographers to concoct a new, stronger cipher, something that could outwit the cryptanalysts. Although this cipher would not emerge until the end of the sixteenth century, its origins can be traced back to the fifteenth-century Florentine polymath Leon Battista Alberti. Born in 1404, Alberti was one of the leading figures of the Renaissance—a painter, composer, poet and philosopher, as well as the author of the first scientific analysis of perspective, a treatise on the housefly and a funeral oration for his dog. He is probably best known as an architect, having designed Rome's first Trevi Fountain and having written De re aedificatoria, the first printed book on architecture, which acted as a catalyst for the transition from Gothic to Renaissance design. Sometime in the 1460s, Alberti was wandering through the gardens of the Vatican when he bumped into his friend Leonardo Dato, the pontifical secretary, who began chatting to him about some of the finer points of cryptography. This casual conversation prompted Alberti to write an essay on the subject, outlining what he believed to be a new form of cipher. At the time, all substitution ciphers required a single cipher alphabet for encrypting each message. However, Alberti proposed using two or more cipher alphabets and switching between them during encipherment, thereby confusing potential cryptanalysts. Plain alphabet: $a-b-c-d-e-f-g-h-i-j-k-l-m-n-o-p-q-r-s-t-u-v-w-x-y-z$

Cipher alphabet 1: $F-Z-B-V-K-I-X-A-Y-M-E-P-L-S-D-H-J-O-R-G-N-Q-C-U-T-W$

Cipher alphabet 2: $G-O-X-B-F-W-T-H-Q-I-L-A-P-Z-J-D-E-S-V-Y-C-R-K-U-H-N$

For example, here we have two possible cipher alphabets, and we could encrypt a message by alternating between them. To encrypt the message hello, we would encrypt the first letter according to the first cipher alphabet, so that h becomes A, but we would encrypt

the second letter according to the second cipher alphabet, so that e becomes F. To encrypt the third letter we return to the first cipher alphabet, and to encrypt the fourth letter we return to the second alphabet. This means that the first l is enciphered as P, but the second l is enciphered as A. The fi- nal letter, o, is enciphered according to the first cipher alphabet and becomes D. The complete ciphertext reads AFPAD. The crucial advantage of Alberti's system is that the same letter in the plaintext does not necessarily appear as the same letter in the ciphertext, so the repeated l in hello is enciphered differently in each case. Similarly, the repeated A in the ciphertext represents a different plaintext letter in each case, first h and then l. Blaise de Vigenère, a French diplomat born in 1523, became acquainted with the writings of Alberti when, at the age of twenty-six, he was sent to Rome on a two-year mission. To start with, his interest in cryptography was purely practical and was linked to his work. Then, at the age of thirty-nine, Vigenère decided that he had accumulated enough money to be able to abandon his career and concentrate on a life of study. It was only then that he examined Alberti's idea and turned it into a coherent and powerful new cipher, now known as the Vigenère cipher. The strength of the Vigenère cipher lies in its use of not one or two but twenty-six distinct cipher alphabets to encrypt a message. The first step in encipherment is to draw up a so-called Vigenère square, as shown in Table 3.

imagen

a plaintext alphabet followed by twenty-six cipher alphabets, each shifted by one letter with respect to the previous alphabet. Hence, row 1 represents a cipher alphabet with a Caesar shift of 1, which means that it could be used to implement a Caesar shift cipher in which every letter of the plaintext is replaced by the letter one place further on in the alphabet. Similarly, row 2 represents a cipher alphabet with a Caesar shift of 2, and so on. The top row of the square, in lowercase, represents the plaintext letters. You could encipher each plaintext letter according to any one of the twenty-six cipher alphabets. For example, if cipher alphabet number 2 is used, then the letter a is enciphered as C, but if cipher alphabet number 12 is used, then a is enciphered as M. If the sender were to use just one of the cipher alphabets to encipher an entire message, this would effectively be a simple Caesar cipher, which would be a very weak form of encryption, easily deciphered by an enemy interceptor. However, in the Vigenère cipher a different row of the Vigenère square (a different cipher alphabet) is used to encrypt different letters of the message. In other words, the sender might encrypt the first letter according to row 5, the second according to row 14, the third according to row 21, and so on. To unscramble the message, the intended receiver needs to know which row of the Vigenère square has been used to encipher each letter, so there must be an agreed system of switching between rows. This is achieved by using a keyword. To illustrate how a keyword is used with the Vigenère square to encrypt a sample message, let us encipher divert troops to east ridge, using the keyword WHITE. First of all, the keyword is spelled out above the message and repeated over and over again, so that each letter in the message is associated with a letter from the keyword, as shown on page 56. The ciphertext is then generated as follows. To encrypt the first letter, d, begin by identifying the key letter above it, W, which in turn de- fines a particular row in the Vigenère square. The row beginning with W, row 22, is the cipher alphabet that will be used to find the substitute letter for the plaintext d. We look to see where the column headed by d intersects the row beginning with W, which turns out to be at the letter Z. Consequently he letter d in the plaintext is represented by Z in the ciphertext.

tabla

Keyword: $W-H-I-T-E-W-H-I-T-E-W-H-I-T-E-W-H-I-T-E-W-H-I$

Plaintext divert t roops to eastridge

Ciphertext: $Z-P-D-X-V-P-A-Z-H-S-L-Z-B-H-I-W-Z-B-K-M-Z-N-M$

To encipher the second letter of the message, i, the process is repeated. The key letter above i is H, so it is encrypted via a different row in the Vigenère square: the H row (row 7), which is a new cipher alphabet. To encrypt i, we look to see where the column headed by i intersects the row beginning with H, which turns out to be at the letter P. Consequently, the letter i in the plaintext is represented by P in the ciphertext. Each letter of the keyword indicates a particular cipher alphabet within the Vigenère square, and because the keyword contains five letters, the sender encrypts the message by cycling through five rows of the Vigenère square. The fifth letter of the message is enciphered according to the fifth letter of the keyword, E, but to encipher the sixth letter of the message we have to return to the first letter of the keyword. A longer keyword, or perhaps a keyphrase, would bring more rows into the encryption process and increase the complexity of the cipher. Table 4 shows a Vigenère square, highlighting the five rows (i.e., the five cipher alphabets) defined by the keyword WHITE. The great advantage of the Vigenère cipher is that it is invulnerable to the frequency analysis described in Chapter 1. For example, a cryptanalyst applying frequency analysis to a piece of ciphertext would usually begin by identifying the most common letter in the ciphertext, which in the case above is Z, and then assume that this represents the most common letter in English, e.In fact, the letter Z represents three different letters, d, r and s, but not e.This is clearly a problem for the cryptanalyst.The fact that a letter that appears several times in the ciphertext can represent a different plaintext letter on each occasion generates tremendous ambiguity for the cryptanalyst. Equally confusing is the fact that a letter that appears several times in the plaintext can be represented by different letters in the ciphertext. For example, the letter o is repeated in troops, but it is substituted by two different letters—the oo is enciphered as HS.

tabla

As well as being invulnerable to frequency analysis, the Vigenère cipher has an enormous number of keys. The sender and receiver can agree on any word in the dictionary or any combination of words, or even fabricate words. A cryptanalyst would be unable to crack the message by searching all possible keys because the number of options is simply too great. The traditional forms of substitution cipher, those that existed before the Vigenère cipher, were called monoalphabetic substitution ciphers because they used only one cipher alphabet per message. In contrast, the Vigenère cipher belongs to a class known as polyalphabetic because it employs several cipher alphabets per message. In 1586 Vigenère published his work in A Treatise on Secret Writing. Although some people continued to use traditional ciphers (Appendix D), use of the Vigenère cipher spread during the seventeenth and eighteenth centuries, and the arrival of the telegraph in the nineteenth century suddenly made

it popular within the business community. The polyalphabetic Vigenère cipher was clearly the best way to ensure secrecy for important business communications that were transmitted via a telegraph operator, who would otherwise be able to read the contents of the message. The cipher was considered unbreakable, and became known as le chiffre indéchiffrable, the uncrackable cipher. Cryptographers had, for the time being at least, a clear lead over the cryptanalysts.

Chapter 5

Substitution Ciphers

One of the more popular cryptosystems is the substitution cipher. It is commonly used in the puzzle section of the weekend newspapers, for example. The principle is simple: Each letter in the alphabet Is replaced by another (or possibly the same) letter. More precisely, a permutation of the alphabet is chosen and applied to the plaintext. In the puzzle pages, the spaces between the words are usually preserved, which is a big advantage to the solver, since knowledge of word structure becomes very useful. However, to increase security it is better to omit the spaces.

The shift and affine ciphers are examples of substitution ciphers. The Vigenere and Hill ciphers (see Sections 2.3 and 2.7) are not, since they permute blocks of letters rather than one letter at a time.

Everyone “knows" that substitution ciphers can be broken by frequency counts. However, the process is more complicated than one might expect.

Consider the following example. Thomas Jefferson has a potentially treasonous message that he wants to send to Ben Franklin. Clearly he does not want the British to read the text if they intercept it, so he encrypts using a substitution cipher. Fortunately, Ben Franklin knows the permutation being used, so he can simply reverse the permutation to obtain the original message (of course, Franklin was quite clever, so perhaps he could have decrypted it without previously knowing the key).

Now suppose we are working for the Government Code and Cypher School in England back in 1776 and are given the following intercepted message to decrypt.

LWNSOZBNWVWBAYBNVBSQWVUOHWDIZWRBBNPBPOOUWRPAWXAW
PBWZWMYPOBNPBBNWJPAWWRZSLWZQJBNVIAXAWPBSALIBNXWA
BPIRYRPOIWRPQOWAIENBVBNPBPUSREBNWVWPAWOIHWOIQWAB
JPRZBNWFYAVYIBSHNPFFIRWVVBNPBBSVWXYAWBNWVWAIENBV
ESDWARUWRBVPAWIRVBIBYBWZPUSREUWRZWAIDIREBHWIATYV
BFSLWAVHASUBNWXSRVWRBSHBNWESDWARWZBNPBLNRWWDWAPR
JHSAUSHESDWARUWRBQWXSUWVZWVBAYXBIDWSHBNWVWWRZVIB
IVBNVAIENBSHBNWFWSFOWBSPOBWASABSPQSOIVNIBPRZBSIR
VBIBYBWRWLESDWARUWRBOPJIREIBVHSYRZPBISRSRVYXNFAI

RXIFOOTPRZSAEPRIKIREIBVFSLWAVIRVYXNHSAUPVBSVWWUU
SVBOICWOJBSWHHWXBBNWIAVPHWBJPRZNPFFIRWW

A frequency count yields the following (there are 520 letters in the text):

tabla

The approximate frequencies of letters in English were given in Section 2.3. We repeat some of the data here in Table 2.2. This allows us to guess with reasonable confidence that W represents e (though B is another possibility). But what about the other letters? We can guess that B , R, S, I, V, A, P, N , with maybe an exception or two, are probably the same as t, a, o, i, n, s, h, r in some order. But a simple frequency count is not enough to decide which is which. What we need to do now is look at digrams, or pairs of letters. We organize our results in Table 2.3 (we only use the most frequent letters here, though it would be better to include all).

tabla

The entry 1 in the IV row and N column means that the combination W N appears 1 time in the text. The entry 14 in the N row and W column means that AHV appears 14 times.

tabla

We have already decided that W = e, but if we had extended the table to include low-frequency letters, we would see that W contacts many of these letters, too, which is another characteristic of e. This helps to confirm our guess. The vowels a, i, o tend to avoid each other. If we look at the R row, we see that R does not precede S, I, A, N very often. But a look at the R column shows that R follows S, I, A fairly often. So we suspect that R is not one of a, i, o. V and N are out because they would require a, i, or o to precede W = e quite often, which is unlikely. Continuing, we see that at the most likely possibilities for a, i, o are S, I, P in some order. The letter n has the property that around 80% of the letters that precede it ore vowels. Since we already have identified W, S, I, P as vowels, we see that R and A are the most likely candidates. We'll have to wait to see which is correct. The letter h often appears before e and rarely after it. This tells us that N = h. The most common digram is th. Therefore, B = t. Among the frequent letters, r and s remain, and they should equal V and one of A, R. Since r pairs more with vowels and s pairs more with consonants, we see th at V must be s and r is represented by either A or R. The combination rn should appear more than nr, and A R is more frequent than RA, so our guess is that A = r and R = n. We can continue the analysis and determine that S — o (note that to is much more common than ot), I = i, and P — a are the most likely choices. We have therefore determined reasonable guesses for 382 of the 520 characters in the text:

tabla

At this point, knowledge of the language, middle-level frequencies (i, d,...), and educated guesses can be used to fill in the remaining letters. For example, in the first line a good guess

is that Y = u since then the word truths appears. Of course, there is a lot of guesswork, and various hypotheses need to be tested until one works. Since the preceding should give the spirit of the method, we slop the remaining details. The decrypted message, with spaces (but not punctuation) added, is as follows (the text is from the middle of the Declaration of Independence):

we hold these truths to be self evident that all m en are created equal that they are endowed by their creator with certain unalienable right) that among these are life liberty and the pursuit of happiness that to secure these rights governments are instituted among men deriving their just powers from the consent of the governed that whenever any form of government becomes destructive of these ends it is the right of the people to alter or to abolish it and to institute new government laying its foundation on such principles and organizing its powers in such form as to seem most likely to effect their safety and happiness

Chapter 6

Divisibility

Number theory is concerned with the properties of the integers. One of the most important is divisibility.

Definition Let a and b be integers with $a \neq 0$ We say that a divides b, if there is an integer k such that $b = ak$. This is denoted by $a|b$. Another way to express this is that b is a multiple of a.

Examples, $3|15$, $^{-}15|60$, $7|18$ (does not divide).

The following properties of divisibility are useful.

Proposition . Let a,b,c represent integers.

1. For every $a \neq 0$,$a \vee 0$ and $a \vee a$. Also, $1|b$ for every b.
2. If $a \vee b$ and $b \vee c$, then $a \vee c$.
3. If $a \vee b$ and $a \vee c$, then $a \vee (sb + tc)$ for all integers s and t;

Proof. Since $0 = a.0$,we may take $k = 0$ in the definition to obtain $a \vee 0$.
Since $a = a.1$, we take $k = 1$ to prove $a \vee a$. Since $b = b.1$, we have $1 \vee b$. This proves (1). In (2), there exist k and t such that $b = ak$ and $c = bl$ Therefore, $c = (kl)a$, so $a \vee c$. For (3), write $b = ak1$ and $c = ak2$. Then $sb + tc = a(ski + tfcj)$, so $a \vee sb + tc$.

For example, take $a = 2$ in part (2). Then $2 \vee b$ simply means that b is even. The statem ent in the proposition says that c, which is a multiple of the even number b, must also be even (that is, a multiple of $a = 2$).

6.1 Prime Numbers

A number $p > 1$ that is divisible only by 1 and itself is called a **prime number**. The first few primes are 2,3,5,7,11,13,17,... An integer $n > 1$ that is not prime is called **composite**, which means that n must expressible as a product ab of integers with $1 < a$, $b < n$. A fact, known already to Euclid, is that there are infinitely many prime numbers. A more precise

statement is the following, proved in 1896.

Prime Number Theorem
Let $\pi(x)$ be the number of primes less tha x. Then $\pi(x) \approx \frac{x}{Inx'}$.

In the sense that ratio $\frac{x}{In(x)} \to 1 as x \to \infty$. $\pi(x)/x$

We won't prove this here; its proof would lead us too far away from our cryptographic goals. In various applications, we'll need large primes, say of around 100 digits. We can estimate the number of 100-digit primes as follows:
$n = p_1^a p_2^a ... p_s^a = q_1^b q_2^b ... q_q^b$

So there are certainly enough such primes. Later, we'll discuss how to find them.

Prime numbers are the building blocks of the integers. Every positive integer has a unique representation as a product of prime numbers raised to different powers. For example, 504 and 1125 have the following factorizations

$504 = 23327, 1125 = 3253$

Moreover, these factorizations are unique, except for reordering the factors. For example, if we factor 504 into primes, then we will always obtain- three factors of 2, two factors of 3, and one factor of 7. Anyone who obtains the prime 41 as a factor has made a mistake.

Theorem Every positive integer is a product of primes. This factorization into primes is unique, up to reordering the factors.

Proof. There ia a small technicality that must be dealt with before we begin. When dealing with products, it is convenient to make the convention that an empty product equals 1. This is similar to the convention that $x = 1$. Therefore, the positive integer 1 is a product of primes, namely the empty product. Also, each prime is regarded as a one factor product of primes.

Suppose there exist positive integers that are not products of primes. Let n be the smallest such integer. Then n cannot be 1 (= the empty product), or a prime = a one factor product), so n must be composite. Therefore, $n = ab$ with $1 < a, b < n$. Since n is the smallest positive integer that is not a product of primes, both a and b are products of primes. But a product of primes times a product of primes is a product of primes, so $n = ab$ is a product of primes. This contradiction shows that the set of integers that are not products of primes must be the empty set. Therefore, every positive integer is a product of primes.

The uniqueness of the factorization is more difficult to prove. We need the following very important property of primes.

Lemma If p is a prime and p divides a product of integers ab, then either $p \vee a$ or $p \vee b$. More generally, if a prime p divides a product $ab...z$, then p must divide one of the factors $a, b, ..., z$.

For example, when $p = 2$, this says that if a product of two integers is even then one of the two integers must be even. The proof of the lemma will be given at the end of this section, after we discuss the Euclidean algorithm.

Continuing with the proof of the theorem, suppose th at an integer n can be written as a product of primes in two different ways;

where $pi, ..., ps$ and $qi, ., ., qt$ are primes, and the exponents a,- and bj are nonzero. If a prime occurs in both factorizations, divide both sides by it to obtain a shorter relation. Continuing in this way, we may assume that none of the primes $pi, ..., ps$ Occur among the qj's. Take a prime that occurs on the left side, say p. Since pt divides n, which equals $q1q1...q1q2...qt$, the lemma says that $p1$ must divide one of the factors qj. Since qj is prime, $pi = pj$. This contradicts the assumption that pi does not occur among the qj's. Therefore, an integer cannot have two distinct factorizations, as claimed.

Chapter 7

Finite Fields

7.1 Division

We can easily add, subtract, and multiply polynomials in $ZP[X]$, but division is a little more subtle. Let's look a t an example. The polynomial $X^8 + X^4 + X^3 + X + 1$ is irreducible in $ZzlXJ$ (although there are faster methods, one way to show it is irreducible is to divide it by all polynomials of smaller degree in $Z2[X']$). Consider the field

$GF(28) = Z2[X](modX^8 + X^4 + X^3 + X + 1)$

Since $X^7 + X^6 + X^3 + X + 1$ is not 0, it should have an inverse. The inverse is found using the analog of the extended Euclidean algorithm. First, perform the gcd calculation for gcd($X^7 + X^6 + X^3 + X + 1, X^8 + X^4 + X^3 + X + 1$).

The procedure (remainder $\rightarrow$ divisor $\rightarrow$ dividend $\rightarrow$ ignore) is the same as for integers:

$X^8 + X^4 + X^3 + X + 1 = (X + 1)(X^7 + X^6 + X^3 + X - 1) - (X^6 + X^2 + X)$

$X^7 + X^6 + X^3 + X + 1 = (X + 1)(X^6 + X^2 + X) + 1$

The last remainder is 1, which tells us that the "greatest common divisor" of $X^7 + X^6 + X^3 + X + 1$ and $X^8 + X^4 + A^3 + X + 1$ is 1. Of course, this must be the case, since $X^8 + X^4 + X^3 + X + 1$ is irreducible, so its only factors are 1 and itself.

Now work back through the calculation to express 1 as a linear combination of $X^7 + X^6 + X^3 + X + 1$ and $X^8 + X^4 + X^3 + X + 1$ (or use the formulas for the extended Euclidean algorithm). Recall th a t in each step we take the Inst unused remainder and replace it by the dividend minus the quotient times the divisor; since we are working mod 2, the minus signs disappear.

$1 = (X^7 + X^6 + X^3 + X - 1) + (X + 1)(X + X2 - I - X)$
$1 = (X^7 + X^6 + X^3 + X - 1) + (X + 1)(X^8 + X^4 + X^3 + X + 1) + (X + 1)(X^7 + X^6 + X^3 + X - 1)$
$1 = (1 + (X + 1)^2(X^7 + X^6 + X^3 + X - 1) + (X + 1)(X^8 + X^4 + X^3 + X + 1)$
$1 = (X^2)(X^7 + X^6 + X^3 + X + 1) + (X + 1)(X^8 + X^4 + X^3 + X + 1)$

Therefore,

$1 = (X^2)(X^7 + X^6 + X^3 + X + 1) + (X + 1)(X^8 + X^4 + X^3 + X + 1)$

Reducing mod $X^8 + X^4 + X^3 + X + 1$, we obtain

$(X^2)(X^7 + X^6 + X^3 + X + 1) = 1(modX^8 + X^4 + X^3 + X + 1)$

which means th at X^2 is the multiplicative inverse of $X^7 + X^6 + X^3 + X - 1$. Whenever we need to divide by $X^8 + X^4 + X^3 + X + 1$, we can instead multiply by X^2. This is the analog of what we did when working with the usual integers mod p.

7.2 GF (2^8)

Lliter in this book, we shall discuss Rijndael, which uses GF(2^8), so let's look a t this field a little more closely. We'll work mod the irreducible polynomial $X_s + X * 1 + X^3 + X + 1$, since th at is the one used by Rijndael. However, there are other irreducible polynomials of degree 8, and any one of them would lead to similar calculations. Every element can be represented uniquely as a polynomial

$_{b7}X^7 +_{b6} X^6 +_{b5} X^5 +_{b4} X^4 +_{b3} X^3 +_{b2} X^2 +_{b1} X +_{b0}$

where each bi is 0 or 1. The 8 bits $b_6b_5b_4b_3b_2b_1b_0$ represent a byte, so we can represent the elements of GF (2^8) as 8-bit bytes. For example, the polynomial $X^7 + X^6 + X^3 + X + 1$ becomes 11001011. Addition is the XOR of the bits:

$(X^7 + X^6 + X^3 + X + 1) + (X^4 + X^3 + 1)$
$\rightarrow 11001011 \oplus 00011001 = 11010010$
$\rightarrow X^7 + X^6 + X^3 + X + 1$

Multiplication is more subtle and does not have os easy an interpretation. That is because we are working mod the polynomial $X^8 + X^4 + X^3 + X + 1$, which we can represent by the 9 bits 100011011. First, let's multiply $X^7 + X^6 + X^3 + X + 1$ by X. With polynomials, we calculate

$(X^7 + X^6 + X^3 + X + 1)(X) = X^8 + X^7 + X^4 + X^3 + X$
$= (X^7 + X^6 + X^3 + X + 1) - (X^8 + X^7 + X^4 + X^3 + X)$
$= X^7 + X^6 + X^3 + X + 1(modX^8 + X^7 + X^4 + X^3 + X)$

The same operation with bits becomes

11001011 $\rightarrow$ 110010110 (shift left and append a 0)
$\rightarrow$ 110010110100011011 (subtract $X^8 + X^7 + X^4 + X^3 + X$)
= 010001101,

which corresponds to the preceding answer. In general, we can multiply by X by the following algorithm:

1. Shift left and append a 0 as the lost bit,
2. If the first bit is 0, stop.
3. If the first bit is 1, X O R with 100011011.

The reason we stop in step 2 is th at if the first bit is 0 then the polynomial still has degree less than 8 after we multiply by X , so it does not need to be reduced. To multiply by higher powers of X , multiply by X several times. For example, m ultiplication by X^3 can be done with three shifts and at most three XOR's. Multiplication by an arbitrary polynomial can be accomplished by multiplying by the various powers of X appearing in that polynomial, then adding (i.e., XORing) the results.

In summary, we see th at the fields operations of addition and multiplication in GF (2^8) can be carried out very efficiently. Similar considerations uppiy to any finite field.

The analogy between the integers mod a prime and polynomials mod an irreducible polynomial is quite remarkable. We summarize in the following.

integers $\longleftrightarrow$ $Z_p[X]$
prime number $q \longleftrightarrow$ irreducible $P(X)$ of degree n
$Z_q \longleftrightarrow Z_p[X]$ (mod $P(X)$)
field w ith q elements $\longleftrightarrow$ field with p n elements

Let GF$(p^n)^*$ denote the nonzero elements of GF(p^n). This set, which has $p^n - 1$ elements, is closed under ultiplication, just as the integers not congruent to 0 mod p are closed under multiplication. It can be shown that there is a generating polynomial g(X) such th at every element in GF $(p^n)^m$ can be expressed as a power of g(X). This also means th at the smallest exponent k such th at $g(X)^k \equiv 1$ is $p^n - 1$. This is the analog of a primitive root for primes. There are $0(p^n - 1)$ such generating polynomials, where 0 is Euler's function. An Interesting situation occurs when p = 2 and $2^n{}^-1$ is prime. In this case, every nonzero polynomial $f(X) \neq 1$ in GF(2^n) is a generating polynomial. (Remark, fo r those who know some group theory: The set GF $(2^n)^*$ is a group of prime order in this cose, so every element except the identity is a generator.

The discrete problem mod a prime, which we'll discuss in Chapter 7, has an analog for finite fields; namely, given h(x), find an integer k such thath $(X) = g(X)^k$ in GF(p^n). Finding such a k is believed to be very hard in most situations.

Chapter 8

Greatest Common Divisor

The greatest common divisor of a and b is the largest positive integer dividing both a and b and is denoted by either $gcd(a, b)$ or by (a, b). In this book, we shall use the first notation.

Examples: $gcd(6, 4) = 2, gcd(5, 7) = 1, gcd(24, GO) = 12$.

We say that a and b are relatively prime if gcd(a, b) = 1. There are two standard ways for finding the gcd:

1. If you can factor a and b into primes, do so, For each prime number, look at the powers that it appears in the factorizations of a and b. Take the smaller of the two. Put these prime powers together to get the gcd. This is easiest to understand by examples:

   $576 = 2^6 3^2, 135 = 3^3 * 5, gcd(576, 135) = 3^2 = 9$
   $gcd(2^5 3^4 7^2, 2^2 5^3 7) = 2^2 3^0 5^0 7^1 = 2^2 7 = 28$
   Note that if a prime does not appear in a factorization, then it cannot appear in the gcd.

2. Suppose a and b are large numbers, so it might not be easy to factor them. The gcd can be calculated by a procedure known as the Euclidean algorithm. It goes back to what everyone learned in grade school: division with remainder. Before giving a formal description of the algorithm, let's see some examples.

Example. Compute gcd(482, 1180).

Solution: Divide 482 into 1180. The quotient is 2 and the remainder is 216. Now divide the remainder 216 into 482. The quotient is 2 and the remainder is 50. Divide the remainder 50 into the previous remainder 216. The quotient is 4 and the remainder is 16. Continue this process of dividing the most recent remainder into the previous one. The last nonzero remainder is the gcd, which is 2 in this case:

$1180 = 2 * 482 + 216$

$482 = 2 * 216 + 50$
$216 = 4 * 50 + 16$
$50 = 3 * 16 + 2$
$16 = 8 * 2 + 0$

Notice how the numbers are shifted:

Remainder $\rightarrow$ *divisor* $\rightarrow$ *dividend* $\rightarrow$ *ignore*.

Here is another example:

$12345 = 1 * 11111 + 1234$ $11111 = 9 * 1234 + 5$ $1234 = 246 * 5 + 4$ $5 = 1 * 4 + 1$ $4 = 4 * 1 + 0$
Therefore, $gcd(12345, 11111) = 1$.

Using these examples as guidelines, we can now give a more formal description of the **Euclidean algorithm**. Suppose that a is greater than b. If not, switch a and b. The first step is to divide a by b, hence represent a in the form

$a = q_1 b + r_1$

If $r1 = 0$, then b divides a and the greatest common divisor is b. If $r1! = 0$, then continue by representing b in the form

$b = q_2 r_1 + r_2$

Continue in this way until the remainder that is zero, giving the following sequence of steps:

$a = q_1 b + r_1$ $b = q_2 r_1 + r_2$ $r_1 = q_3 r_2 + r_3$ ..... $r_k - 2 = q_k r_k - 1 + r_k$ $r_k - 1 = q_k + _1 r_k$

The conclusion is that

$gcd(a, b) = r_k$

There are two important aspects to this algorithm:

1. It does not require factorization of the numbers.
2. It is fast

For a proof that it actually computes the gcd, see Exercise 28.

The Euclidean algorithm allows us to prove the following fundamental result.

Theorem. Let a and b be two integers, with at least one of a, b nonzero, and let $d = gcd(a, b)$. Then there exist integers x, y such that $ax + by = d$. In particular, if a and b are relatively

prime, then there exist integers x, y, with $ax + by = 1$.

Proof. More generally, we'll show that if r_j is a remainder obtained during the Euclidean algorithm, then there are integers x_j, y_j such that $r_j = ax_j + by_j$. Start with $j = 1$, Taking $x_1 = 1$ and $y_1 = -q_1$, we find that $r_1 = ax_1 + by_1$. Similarly, $r_2 = a(-q_2) + b(1 + q_1q_2)$. Suppose we have $r_i = ax_i + by_i$, for all $i < j$.

Then

$r_j = r_{j-2} - q_j r_{j-1} = ax_{j-2} + by_{j-2} - q_j(ax_{j-1} + by_{j-1})$

Rearranging yields

$r_j = -a(x_{j-2} - q_j x_{j-1}) + b(y_{j-2} - q_j y_{j-1})$

Continuing, we obtain the result for all j, in particular for $j = k$. Since $r_k = gcd(a, b)$, we are done.

As a corollary, we deduce the lemma we needed during the proof of the uniqueness of factorization into primes.

Corollary. If p is a prime and p divides a product of integers ab, then either $p|a$ or $p|b$. More generally, if a prime p divides a product $ab\ldots z$, then p must divide one of the factors $a, b, ..., z$.

Chapter 9

Solving ax + by = d

We did not use the quotients in the Euclidean algorithm. Here is how we can use them. A very basic fact, proved in the last section, is that, given integers a and b, there are integers x and y such that:

$ax + by = gcd(a, b)$

How do we find x and y? Suppose we start by dividing a into b, so $b = q_1 a + r_i$, and then proceed as in the Euclidean algorithm. Let the successive quotients be $q_1, q_s, ...q_n$, so in the first example of Section 3.1., we have $q_1 = 2, q_2 = 2, q_3 = 4, q_4 = 3, q_s = 8$. Form the following sequences:

$x_0 = 0, x_1 = 1, x_j = -q_{j-1}x_{j-1} + x_{j-2}, y_0 = 1, y_1 = 0, y_j = -q_{j-1}y_{j-1} + y_{j-2}$.

Then

$ax_n + by_n = gcd(a, b)$

In the first example, we have the following calculation:

$x_0 = 0, x_1 = 1, x_2 = -2x_1 + x_0 = -2x_3 = -2x_2 + x_1 = 5x_4 = -4x_3 + x_2 = -22x_s = -3x_4 + x_3 = 71$

Similarly, we calculate $y_5 = -29$. An easy calculation shows that

$482 * 71 + 1180 * (-29) = 2 = gcd(482, 1180)$

Notice that we did not use the final quotient. If we had used it, we would have calculated $x_n + 1 = 590$, which is the original number 1180 divided by the gcd, namely 2. Similarly $y_n + 1 = 241$, is 482/2. The preceding method is often called the extended Euclidean algorithm. It will be used in the next section for solving certain congruences. For small numbers, there is another way to find x and y that does not involve as much book keeping with subscripts. Let's consider the example gcd (12345,11111) = 1 from the previous section. We'll use the numbers from that calculation. The idea is to work back through the remainders 1, 4, 5, 1234, and the original numbers 11111 and 12345, and eventually obtain the gcd 1 as a

combination of 12345 and 11111. From the line that revealed the gcd, we find

$1 = 5 - 1 * 4$

So we have 1 as a combination of the previous two remainders. Moving up one line, we write the remainder 4 as a combination of 1234 and 5, then substitute into the preceding equation:

$4 = 1234 - 246 * 5$

So,

$1 = 5 - 1 - 4 = 5 - 1 * (1234 - 246 * 5) = 247 * 5 - 1 * 1234$

We have now used the last two remainders from the gcd calculation. Write the last unused remainder, namely 5, as a combination of 11111 and 1234, then substitute into the preceding equation:

$1 = 247 * (11111 - 9 * 1234) - 1 * 1234 = 247 * 11111 - 2224 * 1234$

Finally, we substitute for 1234 to obtain

$1 = 247 * 11111 - 2224 * (12345 - 1 * 11111) = 2471 * 11111 - 2224 * 12345$

This yields the gcd 1 as a combination of 12345 and 11111, as desired. As long as the gcd calculation takes only a few steps, this procedure is quite easy to do by hand. But, in general, the previous method is better and adapts well to a computer.