**CS311 Yoshii – Week 14 (Notes-14B) Encryption for Security – Big Data Analysis**

**Questions from this lecture will be on the final exam.**

The following are new topics for CS311. **Cyber security** and **Big Data** have become a very important topic these days.

**Cybersecurity**

The field of cybersecurity is a new and rapidly developing field. Open job listings for cybersecurity positions throughout the nation **rose 73 percent** in the five years through 2012, 3.5 times faster than postings for computer jobs as a whole, according to Boston-based Burning Glass, a labor market analytics firm that collects data from more than 22,000 online jobs sites. Furthermore, the data show “employers literally just posting and reposting” their positions, suggesting that there are not enough qualified professionals to fill those positions. There were **64,383 jobs** related to cybersecurity listed for the twelve months through April, about 3 percent of all information technology positions, according to the company.

San Diego has been identified as a Cybersecurity hub with the development of Securing our eCity program, Stop Think Connect campaign and the Cyber Hive collaboration program. The San Diego Labor statistics reflect an above average increase, 12.3%, in cybersecurity related jobs projected for 2013-2020.

**Cybersecurity involves: 1) communication safety, 2) prevention of intrusion, 3) penetration testing (ethical hacking) , and 4) detection of intrusion and recovery.**

This lecture will introduce you to **(1)** the world of **encryption** so that you will become interested in security issues.

With hashing you converted one number (e.g. social security) into another number (the slot).

**Math plays an important role** in this conversion.

In encryption, you also use math to convert a number into another number. But it is more complex because you will **have to be able to get back the original from the encoded information.**

For example, if I convert every number by doing Num mod 2, the result will be 1 in many cases.

From 1, I cannot get back Num. Math is fascinating.

**Cryptography/Encoded Messages:**

For cryptography, you need **a lock (encryption tool)** and **a key (decryption tool).**

The more mathematically complex the scheme is, the more secure the system is. For the math behind cryptography, take MATH523 and CS538.

The lock is referred to as the **encryption key (EK)** while the key that opens a lock is referred to as the **decryption key (DK**)**. It is confusing because both are called keys. So, from here on, let me refer to them as simply EK and DK.**

**Basic Cryptography Algorihms: (We will go over 4 methods)**

1. S**ymmetric-Key Encryption (Good between two close friends)**:

**EK and DK are shared by close friends.**

1. Computer A and Computer B both have EK ***where DK can be figured out easily from EK.***
2. Computer A encrypts a message using EK and sends it.
3. Computer B decrypts the message using DK.
4. Computer B encrypts a message using EK and sends it.
5. Computer A decrypts the message using DK.

**Example:** EK is to convert every letter by the letter that occurs next in the alphabet. (A -> B). DK is then to do the reverse (B -> A)

**\*Inter\* Think of another symmetric EK-DK example. If you had to use it from today to send documents between you and your friend, how will you make sure the method stays between just you two?**

**Analogy = “Let’s share the lock and the key”:**

**Mary C:\Users\rtuuser\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\A9392OR9\MC900389750[1].wmf C:\Users\rtuuser\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\A9392OR9\MC900389750[1].wmfJohn**

* 1. Mary first puts the secret message in a box, and locks the box using a padlock to which she has a key.
  2. Mary sends the box to John through regular mail.
  3. John uses a **copy of Mary's key** **to open the box**, and reads the message.
  4. John uses the **same padlock** **to send his secret reply**.

***Advantage:***  The mathematics involved in encoding and decoding is simple.

***Problem:***

* *This algorithm works well between two people who know each other well but if you are communicating with numerous people, how do you share EK and DK with everyone?*
* *Also, even between friends, how do you get EK/DK to your friend securely?*

**Note that since it is easy to figure out DK from EK, you should not give out EK publicly. It would be easier if you could give out EK publicly.**

**Data Encryption Standard** (**DES**) was the first major symmetric algorithm developed for computers in the United States:

**2. Public/Asymmetric-Key Encryption (All my friends have to use this lock):**

**Your friends know how to encode. Only you know how to decode.**

Each user has a pair of keys – a **public EK** and a **private DK**. The publicly EK is widely distributed to your friends, while the private DK is known only to you. Your friends use your public EK to send messages to you.

1. Computer A encrypts a message using publicly given EKB and sends it to B.
2. Computer B decrypts the message using its DKB.
3. Computer B encrypts a message using publicly given EKA and sends it to A.
4. Computer A decrypts the message using its DKA.

This is possible because calculating the private DK from the publicly available EK is **either impossible or prohibitively expensive.**

**Analogy = “Give me your lock for me to use”**

**May uses John’s lock John uses his secret key**

* Mary asks John to send his open padlock (EK) to her through regular mail. John keeps his key (DK) to himself.
* When Mary receives it she uses it to lock a box containing her message, and sends the locked box to John.
* John can then unlock the box with his key and read the message.
* To reply, John must similarly get Mary's open padlock (encoding key) to lock the box before sending it back to her.

***Advantage:*** You don’t have to worry about keeping the EK a secret. You can share it widely.

***Problems:***

* Computationally expensive.
* How do you know that the public EK you have received belongs to the person or entity claimed, and has not been tampered with or replaced by a malicious third party? Example: Does EKA really belong to Computer A? What if Computer Z says “This is EKA. Send messages to A using this.”

**Any publicly made available EK should be certified. Usual solution is PKI.**

**Public Key Infrastructure** (PKI) has a **certificate authority (CA)** that both issues and verifies the **digital certificates**. The digital certificate **certifies the ownership of a public EK**.

**RSA** is a widely used public-key algorithm. It uses **prime numbers** to generate keys. There are infinitely many prime numbers to choose from.

**How RSA Works:**

The public EK and private DK for the RSA algorithm are generated the following way:

1. Choose two distinct randomly picked prime numbers ***p*** and ***q***.
2. Compute ***n*** = *pq*.
3. Compute φ(*n*) = φ(*p*)φ(*q*) = (*p* − 1)(*q* − 1), where **φ is Euler's totient function**.
4. Choose an integer ***e*** such that 1 < *e* < φ(*n*) and [gcd](http://en.wikipedia.org/wiki/Greatest_common_divisor)(*e*, φ(*n*)) = 1

🡺The **public EK consists of *n* and *e*.**

1. Given ***d****\*e* ≡ 1 (mod φ(*n*)), solve for **d**.

🡺The **private DK consists of *n* and *d*,** where d must be kept secret.

*p*, *q*, and φ(*n*) must also be kept secret because they can be used to calculate *d*. **But from just n and e, it is impossible to figure out d.**

### How to Encrypt using *n* and *e*:

### The sender first turns message *M* into an integer *m*, such that 0 ≤ *m* < *n and then computes c:*

 c \equiv m^e \pmod{n} .

### How to Decrypt using n and d:

### The receiver converts c back to m :

 m \equiv c^d \pmod{n} .

**3. Using a Combination to Speed Up Things (Sending EK/DK securely):**

**May uses John’s lock John uses his secret key**

Our shared EK/DK

Public-key encryption described above takes a lot of computing, so most systems use a **combination of public-key and symmetric key encryption**. The following allows you to send the symmetric key to your friends securely:

1. Computer A creates symmetric (shared) EK/DK and sends it to Computer B using public-key encryption (i.e. using EKB).
2. Computers A and B can then communicate privately using symmetric-key encryption.
3. Once the session is finished, each computer discards the symmetric EK/DK used for that session.

**Analogy = “Send me the secret key using this lock”:**

* Mary asks John to send his public open padlock (EK) to her through regular mail. John keeps his DK to himself.
* When Mary receives it she uses it to lock a box containing the shared key, and sends the locked box to John.
* John can then unlock the box with his key and get the shared key for the future communications between them.

**4. Alternative - Doubling the Encryption to Avoid Sharing:**

Neither party needs to even touch the other party's padlock (or key) if we use a **double locking** system:

1. Computer A encrypts a message using his EKA and sends it to B. (one lock)
2. Computer B encrypts the message using his EKB and sends it back to A. (two locks)
3. Computer A decrypts the message using his DKA and sends it to B again. (one lock)
4. Computer B decrypts the message using his DKB to read it. (no lock)

**Analogy = “I will lock it until you can lock it yourself”**

**Mary uses her private EK John uses his private EK**

**Mary uses her DK John uses his DK**

1. Mary puts the secret message in a box, and locks the box using a padlock to which only she has a key. She then sends the box to John through regular mail.
2. When John receives the box, he adds his own padlock to the box, and sends it back to Mary.
3. When Mary receives the box with the two padlocks, she removes her padlock and sends it back to John again.
4. When John receives the box with only his padlock on it, John can then unlock the box with his key and read the message Mary.

***Problem:*** This is only possible if **commutative ciphers are used**. A commutative cipher is one in which the **order of encryption and decryption is interchangeable**. It is more mathematically complex.

**Let’s now turn out attention to things other than encryption.**

**Authentication of the Source/Digital Signature**

**Authentication** is used to verify that the information comes from a trusted source. Basically, if information is "authentic," you know who created it and you know that it has not been altered in any way since that person created it.

* Amessage is ***digitally*** ***signed with the sender's private key*** and can be verified by anyone who has access to the **sender's public key**.
* If anything is changed in the document after the digital signature is attached to it, it changes the value that the digital signature compares to, rendering the signature invalid.
* An analogy for digital signatures is the sealing of an envelope with a personal wax seal.

**Checksum and CRC to Check for Corrupted Data**

How do we ensure that the data has not been corrupted during transmission or encryption?

**Checksum and CRC** – compute a value based on the transmitted data and send it along with the data. Do the data and the attached value match?

**Review Questions:**

1. **Describe the differences between symmetric and asymmetric methods.**
2. **What does a digital certificate certify?**
3. **What are Checksum and CRC used for?**

**Big Data and the Internet of Things**

Computer Science has changed a lot in recent years due to an ability to collect, store and process a vast amount of data. In old days, Artificial Intelligence researchers used to say, it is impossible to do this or that because we cannot possibly collect and store everything about the world. But that is slowly changing.

Change 1:

**Moore's Law** states that the number of transistors on a [microprocessor](http://computer.howstuffworks.com/microprocessor.htm) continues to double every 18 months. By the year 2020 or 2030 will find the circuits on a microprocessor measured on an atomic scale. **Quantum computers** will use the power of atoms. Quantum computers have the potential to perform calculations significantly faster than any silicon-based computer. There is already a Quantum computer prototype. The computer itself is very small but the cooling system is huge. ***Please note that no matter how fast a computer becomes, still the time complexity of algorithms will be the same.***

Change 2:

**The Internet of Things** is a very popular concept these days. The idea is to embed a sensor/processor in all everyday objects so that the data from these objects can be sent via the Internet. This is a great way to collect a vast amount of data about everything and everybody. However, this will **invade the privacy** of people.

Change 3:

**Big Data** analysis is one of the most popular topics in computer science today. Once a vast amount of data is collected, how do you analyze the data quickly and accurately? Big Data analysis is important in being able to **do 1. Prediction, and 2. Identification/classification.**

Once we have the true Internet of Things (perhaps by2045), we will have a vast amount of data on everybody and everything. A fast computer will be used to discover interesting relations among things (i.e. find patterns). We will not even tell the computer what to discover. The computer will keep on finding patterns among things.

**Then, we will be able to predict many things before they happen.** For example, what kinds of people will get married at age 35? Predict at what age you will get married.

**Big Data Collection and Usage Example in Health:**

* A sensor in a tooth brush will send data about how often and how long you brush your teeth.
* A sensor in your coffee machine will send data about how often and how much you drink coffee.
* Sensors on your plate will send data about what kinds of food you eat and how much.
* A sensor in your shoes will send data about how often you walk and for how long.
  + [ collected data from billions of people ] + [ data bases of medical records ]

🡺 **Find patterns o**n what kinds of things are related to what illnesses.

End Results:

1. Prevention plans
2. Prediction on whether you will end up with a particular illness

What People Have Accomplished Already:

Sensors in medical devices in hospitals are used to collect data on patients.

* + [ collected data from patients] + [ progress of the patients ]

🡺 **Find patterns** on what kinds of things are related to fast recovery or complications

**Data Collection and Usage Example in Law Enforcement:**

Sensors embedded in a variety of objects on streets (traffic signals, street signs, roads, etc.) will send data about

* Location
* The amount of traffic (foot, bicycles, cars)
* The day and time
* How dark (visibility)
* How many stores are open versus closed
  + [ collected data from billions of locations ] + [ data bases of crimes]

🡺 **Find patterns** on what kinds of things are related to what crimes.

End Results:

1. Prevention plans
2. Prediction on where and what time certain types of crimes will occur today

What People Have Accomplished Already:

The above has already been done in Santa Cruz to predict locations of crimes. This has led to an increase in arrests and a decrease in crimes.

**Other Things People are Working On:**

Although we do not yet have sensors everywhere and we do not have a computer that will discover patterns among everything, people in Artificial Intelligence are working on the following **machine learning tasks**. In each case, we need to collect a vast amount of **training data (examples**). The computer will learn from the examples by finding a pattern.

**Text mining of medical records.** Medical records are often hand written doctor’s notes. Algorithms have been developed to extract out relevant information (i.e. symptoms, test results, medications, diagnosis, etc.) . It is much more difficult than you think because a variety of short-hand notations are used, and negations and questions are also included in the notes. To accomplish this task, **Bayesian statistics and probabilities are used. Math is important.**

Collect a vast number of medical records with words labeled by important categories -> Have the computer learn how to categorize words from these examples.

End result: feed a medical record -> The computer can extract out relevant information.

Usefulness: Extracted information can then be used to find patterns on what kinds of things are related to what illnesses.

**Analysis of facial expressions:** to be able to notice anger/frustration/stress, etc. even if the expression is for less than a second.

Collect a vast number of facial expressions each labeled by the expressed emotion -> Have the computer **find a pattern** to learn to recognize facial expressions.

End result: feed a facial expression -> The computer can determine the expressed emotion.

Usefulness: In law enforcement, the officers can “notice” a negative reaction immediately.

**Analysis of DNAs:** to beable to determine what part of the DNA is related to what human/animal feature.

Collect a vast amount of DNAs each labeled with an existence of a feature or not -> Have the computer **find a pattern**. (**Bio-informatics**)

End result: feed a DNA string -> The computer can determine the existence of a feature or not.

Usefulness: Is this person going to have a cancer?

**Analysis of MRI images:** to be able to “read” your mind (i.e. what image you are forming in your mind).

Collect a vast amount of MRI images each labeled by the item being visualized -> Have the computer **find a pattern**

End result: feed an MRI image -> The computer can determine what you are thinking of.

Usefulness: ??

**\*Inter\*: What other things do you want to collect? What is the end result? What is the usefulness?**

**Big Data Analysis to Find Patterns**

Heuristic (does not work perfectly) algorithms are being developed in Artificial Intelligence to discover patterns in a vast amount of data. In one method, collected data is expressed as points on a multi-dimensional graph. And when points **form a cluster**, they are related.

A simplest example:

Y axis - weight

X axis – calories per day

Z axis - blood pressure

Each patient is a point on the graph. Do you see clusters?

If a new patient with a given X and Y end up belonging to a cluster, you can predict Z.

There are other method involving **statistics and probabilities. Again, math is important**.

**====================================**

**Other Advances related to CS Algorithms**

**====================================**

Although the following are not related to CS311, other advances in computing will change your world by 2045.

* Nano machines – we will create tiny molecules with sensors that can be injected into a person.

e.g. a molecule will detect certain types of cells and inform you and even “kill” the cells

* 3D printers to create human organs – we will use a special 3D printer that uses human cells instead of plastic

to create human organs.

Creating these things require not just the engineering knowledge, **but creation of algorithms**.

e.g. algorithms for detecting types of cells

e.g. algorithms for reading designs and generating a 3D object

***That is why you have to become good at developing algorithms.***

***You cannot just turn other people’s algorithms into C++.***

**The End of the Note Section:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

***®Summarize here what you have learned in your own words and also write down your own thoughts/reactions/questions.***

***Email me now if you have any questions about what you read in this file.***