iIS 資訊科學簡訊 Update

Newsletter of the Institute of Information Science, Academia Sinica, Taiwan

2016

Message from the Director

The Institute of Information Science (IIS) was established in 1982. We currently have 38 full-time research faculty, 30 post-doctoral research fellows, and slightly more than 300 research associates and specialists. Our research is conducted in eight specialized laboratories: Bioinformatics, Computer Systems, Information Processing and Discovery (iPAD), Multimedia Technology, Natural Language and Knowledge Processing, Network Systems and Services, Programming Languages and Formal Methods, and Computation Theory and Algorithms.

IIS is not a degree-granting institution, with two important exceptions. In 2003 a Ph.D. program in bioinformatics was established under the auspices of Academia Sinica's Taiwan International Graduate Program; it has already enrolled more than 50 students. And in 2014 a doctoral program in Social Networks and Human-Centered Computing was inaugurated. It has already enrolled more than 20 students.

Many of our research fellows hold joint faculty appointments at top universities in Taiwan. This allows our institution to play a very significant role in training and fostering advanced research talent in the IT industry and in academia in Taiwan.

First of all, I would like to congratulate Dr. Huai-Kuang Tsai and Dr. De-Nian Yang on their promotions; each has become a Research Fellow.

In the thirty-four years since the founding of the Institute of Information Science, our faculty members have steadfastly maintained their dedication to research in their areas of expertise. We are proud that their efforts have been recognized both domestically and internationally through numerous awards, such as the Academia Sinica Research Award for Junior Research Investigators, the Academia Sinica Career Development Award, the Ta-You Wu Memorial Award from the Ministry of Science and Technology, and the Phase 1 Award of the WSDM Cup Challenge.

In this issue of IIS Update, the Developing Story section features two articles: Dr. Ting-Yi Sung introduces her research project on bioinformatics for proteomics, which plays a crucial role in the internationally collaborative human proteome project; and Dr. Yuan-Hao Chang introduces his research on optimizing space utilization of embedded file systems, which saves space by up to 64% by using a dynamic tail packing scheme. The Lab Profile introduces the Computation Theory and Algorithms Laboratory, with special focus on Dr. Kai-Min Chung's research on classical and (post-) quantum theoretical cryptography. In the Great Idea section, Dr. Huai-Kuang Tsai presents how bioinformatics approaches are used for analyzing transcription factor binding properties. The Spotlight section presents Dr. Wei-Yun Ma's "Natural Language Processing – New Developer for Next Generation."

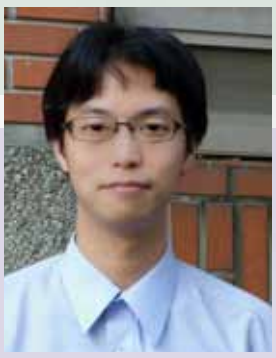
As always, your valuable comments and feedback on this newsletter are much appreciated. And you are welcome to visit the IIS at any time.

Honors and Awards

Dr. **Mi-Yen Yeh**'s jointly supervised team NTU_TriPartite winning Phase 1 of the WSDM Cup Challenge.

Dr. **Huai-Kuang Tsai** being promoted to Research Fellow, effective November 16th, 2015.

Dr. **Kai-Min Chung** receiving the 2016 Academia Sinica Career Development Award, and being promoted to Associate Research Fellow, effective March 19th, 2015.



Dr. **D. N. Yang** receiving the 2015 Academia Sinica Research Award for Junior Research Investigators.

Dr. **Lin-shan Lee** receiving the 2015 Presidential Science Prize and being Academician of Academia Sinica.

Dr. **Yuan-Hao Chang** receiving the 2015 Ta-You Wu Memorial Award from Ministry of Science and Technology.

Distinguished Lecture Series

October 2016	Revisiting Control/Data Plane Separation in Software Defined Networking **Giuseppe Bianchi** Professor, University of Roma Tor Vergata Networking, wireless networks, network security

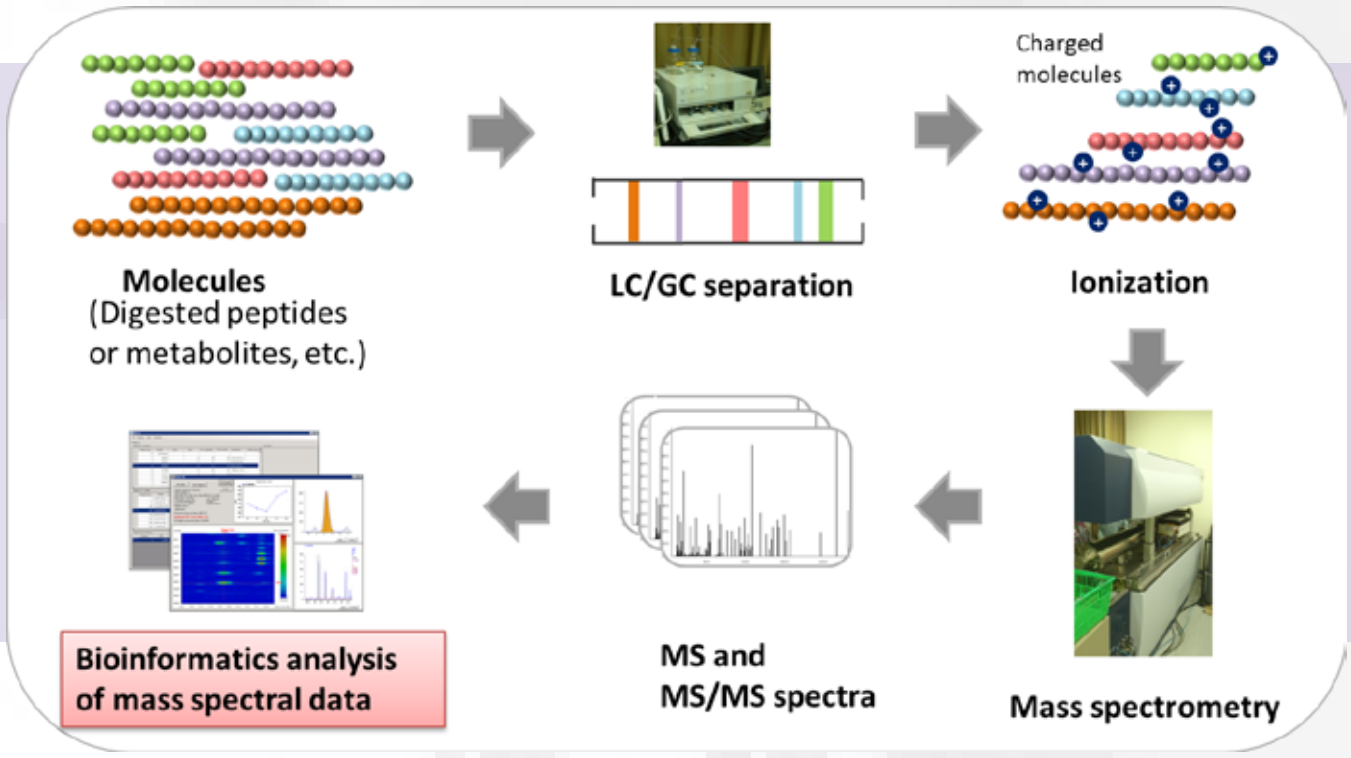
Bioinformatics for Proteomics Plays a Crucial Role in Human Proteome Project

Ting-Yi Sung

Project Coordinator

Proteins are final products of genes and perform functions in living organisms. In the area of biomedical research, proteins are prominent drug targets. Thus, in the post-genomics era, proteomics has been gaining ever-increasing attention. In "Mass spectrometry-based proteomics," Aebersold and Mann (*Nature* 2003) proclaimed mass spectrometry as an indispensable tool for proteomics. Later in 2005, Ong and Mann published "Mass spectrometry-based proteomics turns quantitative" in *Nature Chemical Biology*. Currently, liquid chromatography (LC) coupled with mass spectrometry (MS) technology has been widely used in proteomics research, especially in biomarker discovery and cancer research. In LC-MS experiments, proteins are digested into peptides, separated by LC, and then analyzed by MS. Proteins differentially expressed in different bio-samples can be determined by analyzing the acquired large-scale mass spectra. However, analysis of mass spectral data is challenging for a variety of reasons, including different fragmentation modes in mass spectrometry, coeluting, quality of samples and sample complexity, and noise in the mass spectra. In 2003, in collaboration with Dr. Yu-Ju Chen of Academia Sinica's Institute of Chemistry, we began to develop computation methods and automated tools for mass spectrometry-based quantitative proteomics. We have published three quantitation tools available for download: Multi-Q, MaXIC-Q, and IDEAL-Q.

At the advent of the proteomics age, the Human Proteome Organization initiated the Chromosome-centric Human Proteome Project (C-HPP), similar to Human Genome Project, in which an international collaborative effort has been organized, with 25 working groups — one per chromosome. Taiwan's team, led by Dr. Yu-Ju Chen, is responsible for chromosome 4. The project is aimed at discovering and characterizing all human proteins encoded from genes for the purpose of filling the gap between genomics and proteomics. Since 2013, the main theme of this project has been to experimentally discover missing proteins, which have not been detected by MS or antibody experiments, i.e., those that lack of experiment evidence at the protein level. These proteins are missing for various reasons, such as low abundance, expression in transient states or rare samples, and unfavorable cleavage sites for MS experiments. In order to detect missing proteins, Dr. Chen conducted LC-MS/MS experiments on 11 non-small cell lung cancer cell lines. By using existing database sequence search engines (e.g., Mascot), proteins can be confidently reported from acquired LC-MS/MS spectra. However, because most search engines do not report false discovery rate (FDR) at the protein level, those confidently reported proteins may not be really identified. Rigorous analysis of search engine results is essential to claim missing proteins being detected. Thus, our lab, which has bioinformatics



Workflow of mass spectrometry-based proteomics

expertise, was responsible for this critical task. Based on the reported proteins from the search engine, we first used PeptideShaker, an existing tool, to exclude proteins/peptides with FDR higher than 1%, which is a required criterion in C-HPP. Second, since proteins reported by search engines are based on inference from peptides identified from MS/MS spectra, ambiguity in protein inference is inevitable. To avoid ambiguity in protein inference, we further excluded identified proteins without any identified peptide that belongs to only a single protein in the entire human protein database. Up to this stage, we have confidently identified 7702 proteins, with 66% being membrane proteins. Third, these proteins were compared with the missing protein list provided by the project consortium to find if any missing proteins may have been detected. To confirm whether missing proteins were indeed detected, we performed two critical investigations for further filtering: (1) checking whether the identified peptides used to infer these proteins had been previously detected and deposited in PeptideAtlas, a huge MS/MS spectra repository and recommended for use in this project; (2) checking whether these identified peptides could be derived from a single amino acid variation of a peptide in another protein. By performing the above confirmations, we successfully detected 178 missing proteins, including 74 membrane proteins. Dr. Chen's lab used the multiple reaction monitoring MS technique on eight synthetic peptides to confirm our rigorous workflow to determine detected missing proteins. This work was published in the 2015 special issue of C-HPP in the *Journal of Proteome Research* (*JPR*).

Furthermore, from the perspective of bioinformatics, we attempted to address one of the priority questions for the C-HPP: which of the missing proteins are unlikely to be detectable by even an "ideal" shotgun MS/MS analysis of the human proteome. Thus we performed three *in silico* digestions by commonly used trypsin, Lys-C and both on the human proteins to generate all *in silico* fully digested peptides. With these presumed peptides, we found 145 proteins, including 77 missing proteins, containing no unique peptide, consisting of only shared peptides. These missing proteins were hard to confirm in shotgun proteomics experiments. In addition, missing proteins with high sequence similarity, even up to 100% similarity, are also hard to identify. We also noted that among all missing proteins with evidence at the transcript level, G protein-coupled receptors and olfactory receptors, based on InterPro classification, were the largest families of proteins that exhibited more frequent variants, and thus are hard to identify. In order to identify the above-mentioned types of missing proteins, new MS experiment designs and improved identification methods are needed. This work was published in *JPR* in 2015.

Proteomics research is gaining ever-increasing attention. Computing and analyzing big data acquired from mass spectrometry requires more IT expertise. IT experts are welcome to join the facinating area of bioinformatics for proteomics.

Distinguished Lecture Series

Technology Considerations in Computer Architecture

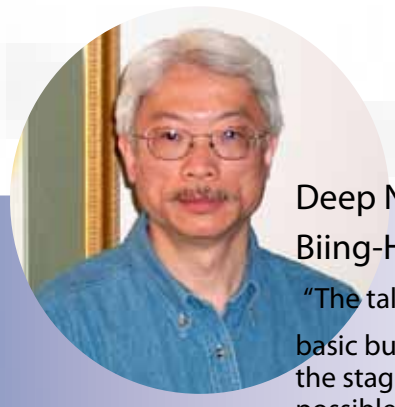
Jean-Luc Gaudiot —September 5, 2016

"….. FPGAs are extremely useful in mobile embedded systems where computing power and energy considerations are major concerns….."

Building Algorithms for the Next Generation Route Planner

Dorothea Wagner — May 17, 2016

"Nowadays, route-planning systems are among the most frequently used information systems. The algorithmic core problem of such systems is the classical shortest-paths problem, which can be solved by Disjkstra's algorithm in almost linear time. However, Dijkstra's algorithm still takes a few seconds in continental-sized graphs, which is too slow for practical scenarios....."



Deep Neural Networks – A Developmental Perspective

Biing-Hwang (Fred) Juang — December 22, 2015

"The talk is given from a developmental perspective with a comprehensive view, from the very basic but oft-forgotten principle of statistical pattern recognition and decision theory, through the stages of problems that are encountered during system design, to key ideas that may lead to possible new advances toward deep learning....."

Building Computing Machines That Sense, Adapt and Approximate

Rajesh K. Gupta — October 27, 2015

"Computing machines today are largely ignorant of the variability in the behavior of underlying components from device to device, chip to chip, and their wear over time — save for thermal sensing in limited energy/power constrained applications....."

Differential Privacy: Theoretical and Practical Challenges

Salil Vadhan — November 10, 2015

"Addressing pressing privacy problems in a variety of domains has attracted the interest of scholars from many fields, including statistics, database management, medical informatics, law, social science, computer security, and programming languages....."

Joint Software-Defined Application-Network Control Plane for Next Generation Real-Time Applications

Klara Nahrstedt — May 25, 2015

"I will argue for a joint software-defined application-network control plane to assist next-generation, real-time applications such as telepresence and teleimmersion. Furthermore, I will discuss OpenSession, the new "Northbound" application-network control plane for multi-stream and multi-site real-time applications, which represents the interaction between the application-level session controller and a Software-Defined Network (SDN) controller. ...."

New Advances in Forensic Identification

Henry Lee — March 23, 2015

"Collection of fingernails is painless, harmless, and convenient. Fingernails can also be found on badly decomposed bodies and body parts, especially in catastrophic incidents. Not only mitochondrial DNA, but also nuclear DNA has been successfully analyzed from fingernail fragments. Fingernail patterns and physical features were extracted by image processing. Features of size, length, and width of fingernails were examined....."

OpenCloud: A Value-added Cloud for Network Operators

Larry Peterson — April 8, 2015

"Network operators are migrating away from purpose-built hardware appliances and moving toward infrastructure that exploits virtualized commodity servers and SDN at the very edge of the Internet, a practice called Network Functions Virtualization (NFV). This talk puts forward a vision for a value-added cloud that demonstrates how network operators can take advantage of cloud technology. It also describes a prototype, OpenCloud, that we are building with Internet2....."

Keeping a Crowd Safe: On the Complexity of Parameterized Verification

Javier Esparza — March 2, 2015

"Many computer systems consist of an arbitrary large number of identical components — a "crowd" — communicating by some means. Examples include distributed algorithms, communication and network protocols, and computer models of biochemical systems. The safety problem for these systems consists of checking that, whatever the size of the crowd, no individual can reach a dangerous state....."

Deep Learning Workshop

December 21–22, 2015

Topics

Deep Neural Networks – A Developmental Perspective

Deep Learning in Computer Vision

Deep Learning in Speech

Frontiers of Communications and Networking Workshop 2015

November 6, 2015

To provide a platform for scholars in the field of communications and networking in Taiwan to interact, collaborate, and share new and trending ideas with each other annually

Topics

Multiuser MIMO Systems: From Rate Adaptation to User Selection ● Some Communication Issues in IoT ● Green Resource Management for Distributed Antenna Systems ● Asynchronous Quorum-based Channel Hopping Schemes for Cognitive Radio Networks ● Enabling Internet of Vehicles with Vehicular Visible Light Communications ● Sharing Experiences on International Academic Services and Research ● Maintaining Competitiveness in the International Academic and Industrial Community

2015 IR Workshop

New Trends and Technologies of Cross-discipline NLP and IR

December 18, 2015

Topics

Jointly Modeling Topics, Events and User Interests on Twitter

A Community-based Method for Valence-Arousal Prediction of Affective Words

Active Learning by Learning

Capture the Great Moment of Social Network Efficiently and Effectively

A Novel Social Influence Model based on Multiple States and Negative Social Influences

Using Non-Verbal Information to Augment Designs of Language-based Interactions

Information Cup of Table Tennis on June 23-24, 2016

Open House on October 31, 2016

Information Cup of Basketball on June 20-21, 2016

Information Cup of Badminton on June 22-23, 2016

Classical and (Post-)Quantum Theoretical Cryptography

Computation Theory and Algorithms Laboratory

Kai-Min Chung
Associate Research Fellow

The Computation Theory and Algorithms Laboratory's cryptography group performs research on theoretical cryptography, from various classical cryptographic topics to (post-)quantum cryptography. Over the past decades, cryptography has evolved far beyond its original goal of secure communication, and is allowing us to realize more and more complex tasks with desired security. With the development of the Internet, cryptography has become ubiquitous, such as with e-mail, mobile phones, SSL, e-commerce, and e-voting. It has permeated everyday life and is heavily used by many applications.

To give a sense of recent exciting developments, this article focuses on the fundamental primitive of encryption schemes and introduces two fascinating strengthened notions of it that allow computation over encrypted data in different forms. These two notions — fully homomorphic encryptions (FHE) and functional encryptions (FE) — have been achieved and even defined only within the past ten years, and they remain very active research topics. We will introduce both notions by their natural applications and provide some physical analogies to help readers to understand the notions intuitively. Finally, we briefly introduce some of our group's recent research focuses and research activities.

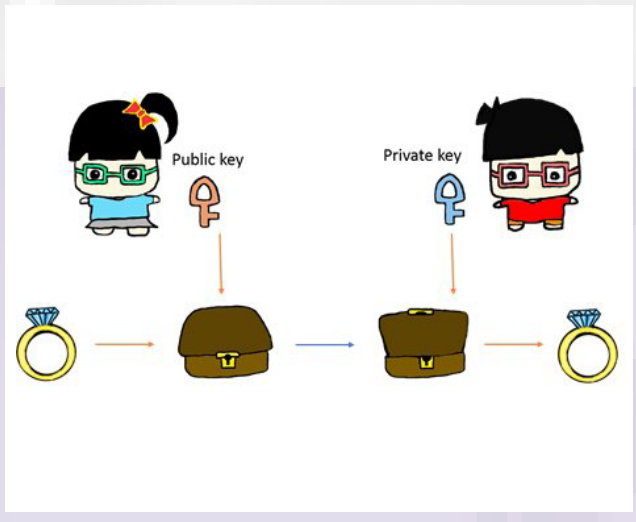
Historically, cryptography emerged from the need of the military/government to have a means of secure communication between military forces or government agencies (such as depicted in the movie The Imitation Game [1]). For example, consider that Alice wants to send a private message (m) to Bob, who is far away, and that the information transmitted may be listened to by an adversarial eavesdropper, Eve. Clearly, Alice cannot send her message m directly, otherwise it will be learned Eve. The goal of secure communication is to ensure that Bob can correctly receive m from Alice while Eve learns nothing about the message m. For example, in online shopping, credit card numbers and security codes must be transmitted to the bank and secure communication is needed to protect the the information against eavesdroppers.

A Public-key encryption scheme (PKE) is a cryptographic primitive to achieve this goal. A PKE consists of three algorithms (KeyGen, Enc, Dec). One can use the key generation algorithm KeyGen to generate a pair of (randomized) keys, called public key (pk) and secret key (sk). Anyone who knows pk can use the encryption algorithm to encrypt a message m to produce a ciphertext ct = $Enc_{pk}(m)$ such that whoever holds sk can decrypt ct to recover m = $Dec_{sk}(ct)$; but those without sk can learn nothing about the underlying message m. Given such PKE, Bob can generate (pk, sk), sending pk to Alice while keeping sk private. Alice can then encrypt m and send the ciphertext (ct) to Bob, who decrypts ct to recover m. Eve, who may learn ct but does not know sk, is guaranteed to learn nothing about m, as desired.

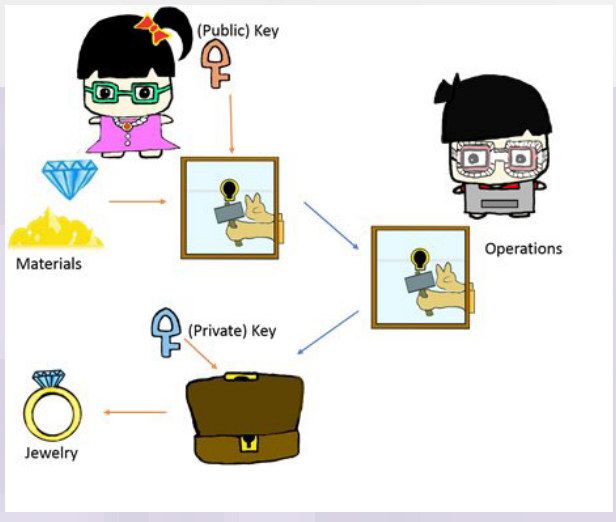
We note that formalizing what it means by "Eve learns nothing about m from ct" is in fact highly non-trivial, and one of the main reasons that Shafi Goldwasser and Silvio Micali received the Turing Award in 2012 (Interested readers are encouraged to look at the following online material: [2,3,4]). We will keep our discussion on an informal and intuitive level, and provide a physical analogy (see Figure 1): One can think of the secret key as a physical key, and the public key as a mold that can be used to produce an unbreakable box (the ciphertext) that can only be opened by sk. Bob can send Alice the mold, who can produce a box to store what she wants to send to Bob (say, jewelry), lock the box, and send it to Bob, who can unlock the box to obtain the jewelry by using his key.

Fully Homomorphic Encryption (FHE)

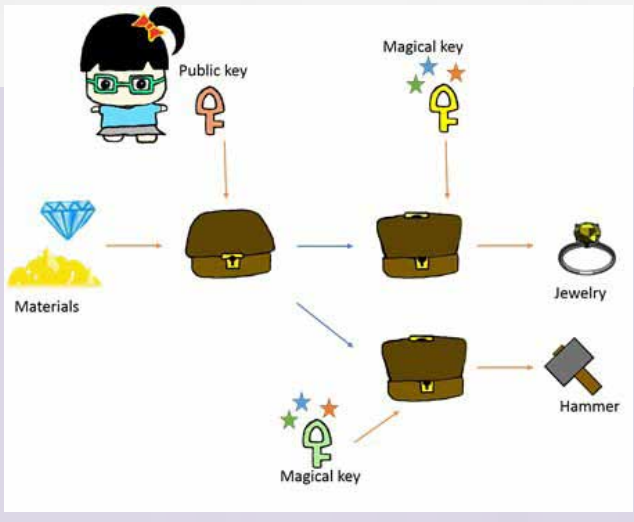
In the era of cloud computing, it is common that we delegate our data and computation to a cloud server (e.g., Amazon EC2). However, when the data is sensitive (e.g., personal private data, hospital medical records), delegation to a potential untrusted server may compromise privacy. In an abstract scenario, consider that a client wants to delegate computation of a function f on his data m (e.g., statistical analysis of medical records) to a server. Can this be done in a way that the server learns nothing about the data m?



Public-key encryption: Key and mold. (Fig. 1)



Fully homomorphic encryption: Alice's jewelry store. (Fig. 2)



Functional encryption: Alice's magical key. (Fig. 3)

Paradoxical as it may sound, this task can be achieved by fully homomorphic encryption, an encryption scheme that allows computations to be performed over encrypted data. More precisely, it is a PKC with an additional evaluation algorithm Eval, which can take an encryption ct = $Enc_{pk}(m)$ of m and a function f as input, and produce an encryption ct' = $Enc_{pk}(f(m))$ of f(m) as output. Note that without secret key sk, one can still learn nothing from ct and ct'. This amazing primitive was first constructed by Craig Gentry in 2009 [5]. Using FHE, the client can send encryption ct = $Enc_{pk}(m)$ of his message m to the server. The server can use Eval to evaluate f on ct to produce ct' = $Enc_{pk}(f(m))$, and send ct' back to the client, who can decrypt to learn f(m). Since the server does not know sk, it learns nothing about m.

As a physical analogy (see Figure 2), we can think of the ciphertext as an unbreakable glovebox that allows one to manipulate objects inside through the gloves. Thus, Alice (the client) can put raw material inside the glovebox and send it to Bob (the server), who can assemble the material into jewelry inside the glovebox and send it back to Alice, who can unlock the box to get the jewelry.

Functional Encryption (FE).

While FHE allows any computation on en-crypted data to be performed, the com-putation result remains encrypted in a ciphertext. Functional encryption (FE) allows delicate access control, such that different users can learn different information from encrypted data. For example, consider that a hospital maintains an encrypted medical da-tabase ct = $Enc_{pk}(m)$ and authorizes each department to perform different statistical analysis on the database but does not allow each department to learn any additional information from the sensitive medical database. Functional encryption allows the hospital to produce a different functional key $sk[f_i]$ for each department D_i such that given ct and $sk[f_i]$, D_i can learn $f_i(m)$ from ct but nothing else (Think of fi as the statistical analysis Di wants to perform on the database m). In terms of physical analogy (see Figure 3), now Alice can produce different "magic keys" that can open a box of raw materials for different products made of those materials.

Our Research: Computation Model in Cryptography

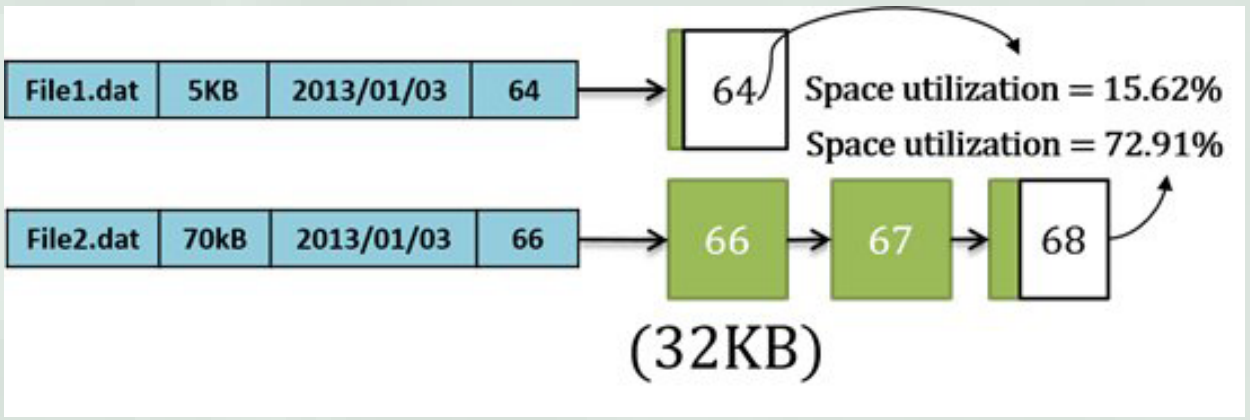
Observe that compu-tation is a common theme in the two amazing primitives FHE and FE; it also plays a central role in cryptography, such as in secure computation and ob-fuscation. But how do we represent/describe a computation, e.g., the sta-tistical analysis example mentioned above? In computer science, a natural thought is to write a program in a cer-tain language, such as C++ or Python, to implement the computation. However, in cryptography, computation is typically represented as a digital circuit (i.e., half adder in Figure 4) composed of logical gates such as AND, OR, and NOT. The advantage of circuits is its simplicity for cryptographic design, since it can reduce to design the solutions for gates and how to compose them (somewhat oversimplifying). However, unlike most programming languages, circuits cannot express for-loops and conditional branches, which

(cont'd on page 16)

Optimizing Space Utilization of Embedded File Systems

Dynamic Tail Packing Scheme Saves Space up to 64%

Yuan-Hao Chang
Associate Research Fellow



The problem of low space utilization (Fig. 1).

Embedded/mobile computing systems are usually battery-powered devices and are widely adopted in various application domains. Due to cost and energy considerations, they usually have limited computing power, RAM space, and storage capacity. However, recent improvements in embedded/mobile computing systems' computing ability have allowed some embedded computing systems to adopt embedded file systems to simplify the complexity of managing their data . For example, Android file systems no longer use a log-based file system (e.g., yaffs2). Instead of utilizing a log-based file system, some embedded/mobile computing systems, such as mobile phones and embedded consumer electronics, manage their data in flash storage devices with a (simple) file system. However, existing file systems usually allocate storage space in the unit of a cluster, whose size often reaches several kilobytes. This leads to low space utilization in the storing of the tail data of (small) files, and becomes a critical issue in the design of embedded storage systems (i.e., storage systems in embedded computing systems), which usually have a limited storage capacity.

Most file systems in embedded computing systems allocate space for storage of a file in the unit of a cluster — no matter how small the file. Well-known examples are FAT and ext3/ext4 file systems; this is because FAT is simple enough to be used in resource-limited embedded systems and because ext4 is the default file system of Linux operating systems, which are widely used in many embedded systems. However, the cluster-based allocation adopted in many embedded file systems seriously decreases the space available for storage, especially in applications such as sensor nodes and control systems that need to store small data files. This situation is exacerbated in some file systems when their cluster size is increased in proportion to the storage capacity. For example, the cluster size of FAT32 is proportional to the storage capacity, and reaches 32KB when the storage capacity is larger than 32GB. As a result, embedded file systems have low space utilization in the large size of a basic allocation unit because the size of files is usually very small. For example, when the cluster size is 32KB, a 5KB small file would result in the waste of 27KB of space, while a 70KB file would waste 26KB of space, because the file system still allocates one cluster for the small tail of the file (see Figure 1 for details), where the tail of a file indicates the file's last part that cannot completely fill a cluster.

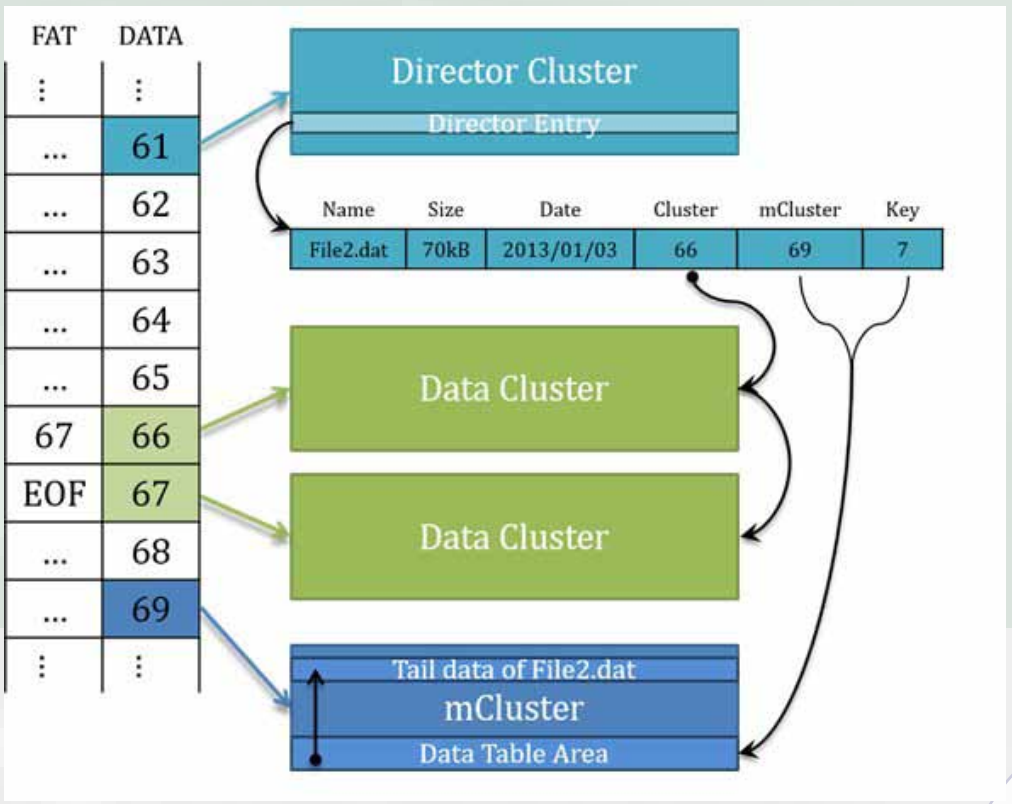
To solve this space-utilization problem, we propose a dynamic tail packing (DTP) scheme to resolve the space utilization issue of embedded file systems (i.e., the file systems used in embedded computing systems). It has two main objectives: (1) to optimize the space utilization of storage systems by packing the tail data of files together with limited performance overhead, and (2) to minimize the internal/external fragmentation issues that exist in the existing tail packing techniques. In order to achieve these objectives, the proposed DTP scheme defines a new type of cluster, called a micro cluster or mCluster, and dynamically packs the tail data of files in mClusters. An mCluster is divided into two parts, the data area and the data table. The data area is used to store the tail data of files, and the data table is used to maintain the start address and the size of the tail data

stored in the mCluster. The status and information of mClusters are maintained in the mCluster Allocation Tables (mCATs) to facilitate the management and space allocation of mClusters. With mClusters and mCATs, the proposed DTP scheme adopts a mechanism, called mCluster-based management, to manage mClusters and allocate (free) space from mClusters for the tail data of files at the byte level. This is different from the existing solution that allocates free space at the sub-cluster level; thus, the internal fragmentation issue can be minimized in mClusters. At the same time, in consideration of the growing/shrinking file sizes at run time, the mCluster-based management also adopts a dynamic tail chain to dynamically pack and distribute the tail data of the same file in multiple mClusters without the overheads on excessively copying the existing tail data of files among mClusters. Thus, the external fragmentation and excessive tail-data copies in existing tail packing techniques can be minimized in the proposed DTP scheme.

Figure 2 shows an example of the proposed DTP scheme that is integrated into the FAT file system. Suppose that the cluster size is 32KB and the file File1.dat is 70KB. The non-tail data of File1.dat are stored in Clusters 64 and 66, with its tail data in Cluster 69, which is an mCluster. In order to support the DTP scheme, the directory entry used to store the attributes of a file includes two new fields (i.e., mCluster and Key) to indicate where the tail data of its corresponding file are stored. The mCluster field indicates the mCluster that stores the tail data of its corresponding file, and the Key field is the key value to search the data table of the mCluster so as to derive the address of the tail data of its corresponding file. For example, the directory entry for file File1.dat indicates that the tail data of File1.dat can be found by using 7 as the key value to look up the information from the data table of mCluster 69. Note that mCluster 69 can be used to store/pack the tail data of multiple files and that the space allocation unit in mClusters is one byte instead of one sub-cluster.

To evaluate the capability of the proposed DTP scheme, in terms of space utilization and access performance, this DTP scheme was implemented and integrated in the FAT file system whose cluster size was 32KB. All the experiments were conducted on a platform with the Linux operating system. The results show that our DTP scheme can save up to 64% of the space required by the file system without the proposed scheme, and only needs the storage space similar to the original file size.

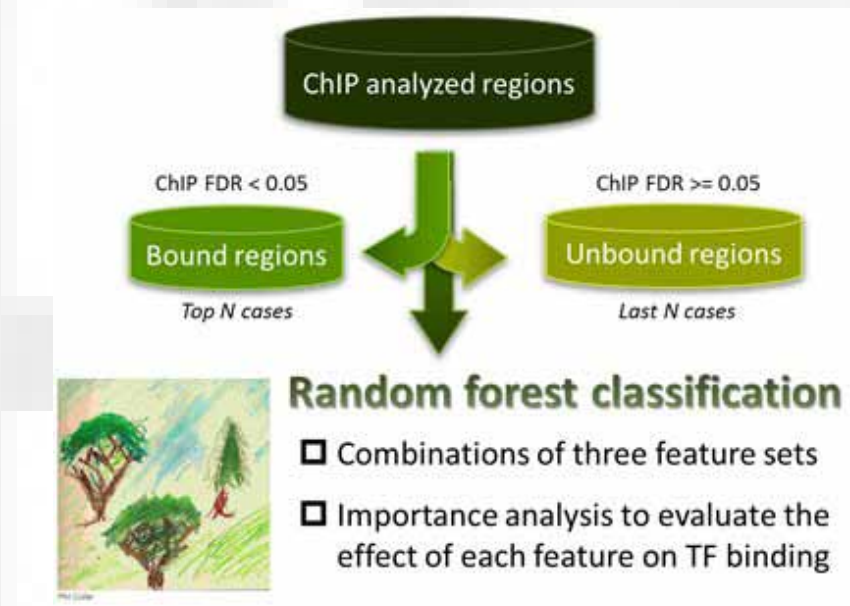
In conclusion, a dynamic tail packing scheme is introduced to optimize the space utilization of embedded file systems. In particular, an mCluster-based management is proposed to dynamically pack tail data of files in mClusters with minimized overheads and space fragmentation. At the same time, by considering the growing/shrinking of file sizes at run time, a dynamic tail chain is also put forward to dynamically pack and distribute the tail data of the same file in multiple mClusters without the overheads from excessively copying the existing tail data of files among mClusters. The proposed scheme was implemented in the file system of Linux operating systems. A series of experiments on copying/handling different types of real files was conducted to evaluate the capability of the proposed scheme, and the results showed that the proposed scheme can save up to 64% of the space required by the original file system with limited performance degradation.



A working example of dynamic tail packing (Fig. 2).

Bioinformatics Approach for Transcription Factor Binding Properties

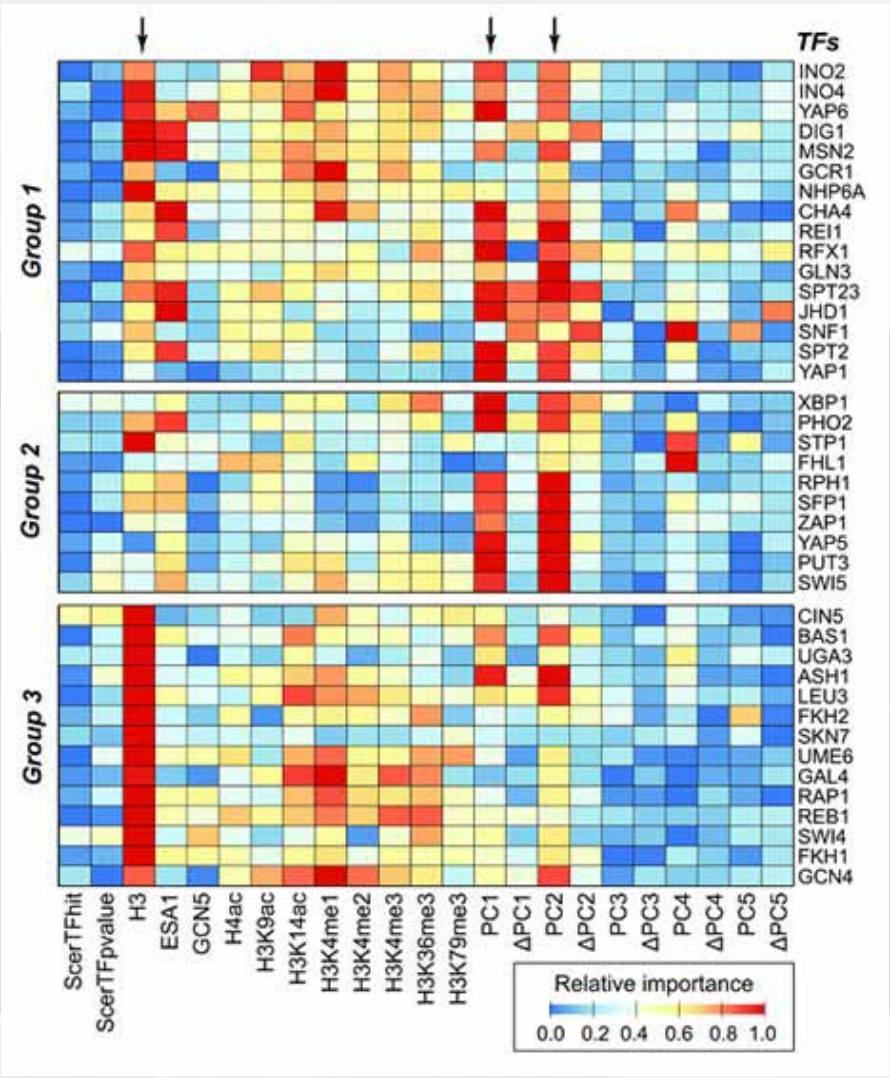
Huai-Kuang Tsai
Research Fellow



A random forest classifier for transcription factor binding properties. (Fig. 1).

One of the central questions in molecular genetics regards the mechanisms of transcriptional regulation, particularly how transcription factors (TFs) regulate expression of target genes with specific TF binding sites (TFBSs). Identifying TFBSs would permit a more comprehensive and quantitative mapping of the regulatory mechanisms within cells. Unfortunately, TFBSs are usually short (~5–20 bp) and degenerate, making it difficult to accurately identify TFBSs. With the advancement of biological techniques, there are widely applicable methods to identify TFBSs experimentally, such as Electrophoretic Mobility Shift Assay (EMSA) , Systematic Evolution of Ligands by Exponential Enrichment (SELEX), Chromatin Immunoprecipitation (ChIP) assays, and Protein Binding Microarrays (PBMs). Experimental methods provide in vivo evidence of TF binding or in vitro measurement of affinity of TF–DNA interactions. However, large-scale and precise prediction of TFBSs remains one of the greatest challenges due to high cost and low time efficiency. Genome-wide TFBS identification thus requires the complementation of bioinformatics.

A large number of studies have developed computational methods for TFBS prediction that examine the presence of sequence motifs, usually simplified as a motif-discovery problem, which involves seeking a sequence motif from a vast array of biological data, such as the promoter sequences of target genes. The motif is typically modeled as a position weight matrix (PWM). PWMs can be used to infer the binding strength of sequences based on the number of known TFBSs or potential binding sequences. Motif-discovery methods for TFBS identification usually follow an enumerative or probabilistic approach. Enumerative approaches investigate the occurrence frequency of all strings, and generate a PWM composed of over-represented strings.



The relative importance of three kinds of features, including sequence motif, chromatin state, and DNA structure, for predicting binding regions of different TFs. Arrowheads indicate the most important features for TFs. (Fig. 2).

Alternatively, probabilistic approaches conduct a multiple sequence alignment of input sequences and simultaneously optimize PWM parameters using machine learning methods, such as Expectation-Maximization algorithm and Gibbs sampling.

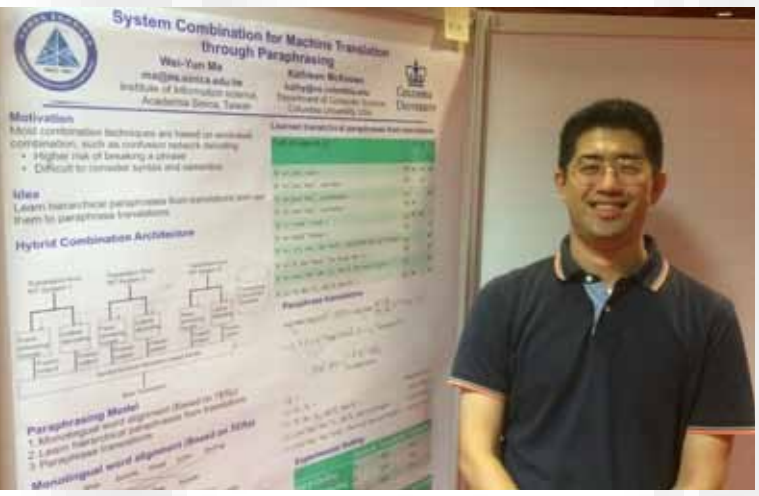
With the increase of available genome-wide data is greater evidence indicating that DNA sequence is not the only factor determining TF binding. In particular, studies have shown that a large portion of false positives in TFBS identification (i.e., motif occurrences which are not really bound by TFs) is due to chromatin inaccessibility. Furthermore, chromatin accessibility and TF binding affinity are found to be correlated. These observations reveal that chromatin accessibility could be a key factor that controls TF binding. Chromatin state feature and DNA structural property are the two main categories of genomic attributes associated with chromatin accessibility. Both chromatin state and DNA structural properties have been shown to be determinants of chromatin accessibility and consequently influence TF binding. Recently, certain TFBS identifications using chromatin state or DNA structural properties have been developed.

Although sequence motifs (SM), chromatin state (CS), and DNA structural (DS) properties have been used to predict TF binding sites, a predictive model that jointly considers CS and DS has not been developed to predict either TF-specific binding or general binding properties of TFs. Using budding yeast as model, we found that machine learning classifiers (random forest, as shown in Fig. 1) trained with either CS or DS features alone perform better in predicting TF-specific binding compared to SM-based classifiers. In addition, simultaneously considering CS and DS further improves the accuracy of the TF binding predictions, indicating the highly complementary nature of these two properties. The contributions of SM, CS, and DS features to binding site predictions differ greatly between TFs, allowing TF-specific predictions and potentially reflecting different TF binding mechanisms. In addition, a "TF-agnostic" predictive model based on three DNA "intrinsic properties" (in silico predicted nucleosome occupancy, major groove geometry, and dinucleotide free energy, see Fig. 2) that can be calculated from genomic sequences alone has performance that rivals the model incorporating experiment-derived data. This intrinsic property model allows prediction of binding regions not only across TFs but also across DNA-binding domain families with distinct structural folds. Furthermore, these predicted binding regions can help identify TF binding sites that have a significant impact on target gene expression. Because the intrinsic property model allows prediction of binding regions across DNA-binding domain families, it is TF agnostic and likely describes the general binding potential of TFs. Thus our findings suggest that it is feasible to establish a TF-agnostic model for identifying functional regulatory regions in potentially any sequenced genome.

Dr. Zing Tsung-Yeh Tsai, a postdoc at my lab; and Dr. Shin-Han Shiu, an associate professor of plant biology at Michigan State University, contributed to this work.

Natural Language Processing – New Developer for Next Generation

Dr. Wei-Yun Ma, Assistant Research Fellow



Presentation of my research in EMNLP conference.

Why did you choose the IIS and what led you to do research in this field?

I first joined the Institute of Information Science (IIS) in 2001 as a research assistant working on natural language processing (NLP). Then I went to the United States to pursue a doctorate and came back to this big family of IIS as an Assistant Researcher. When I was still a research assistant, I felt the enthusiasm and passion of doing research through listening to lectures and academic communication with scholars and researchers. It was from that time the seed of pursuing research as my lifelong career was planted in my heart. The topic of my Ph.D. study was NLP. Upon graduation, I happened to learn there was an opening in IIS. It was a once-in-a-lifetime chance for me. If I could host a lab, I would have a chance to realize many innovative ideas and thoughts. It will be very exciting! IIS is a renowned research institute. It provides a great environment for research in many ways, including (1) the freedom to choose research topics, (2) the presence of at least eight research teams of more than forty researchers provides a great chance for cooperation among different fields, (3) administrative support/grants for researchers, (4) the opportunity to network with domestic and international professors and researchers, and (5) no teaching obligations, so that researchers can concentrate on research. However, its TIGP programs give researchers a chance to teach.

Self portrait

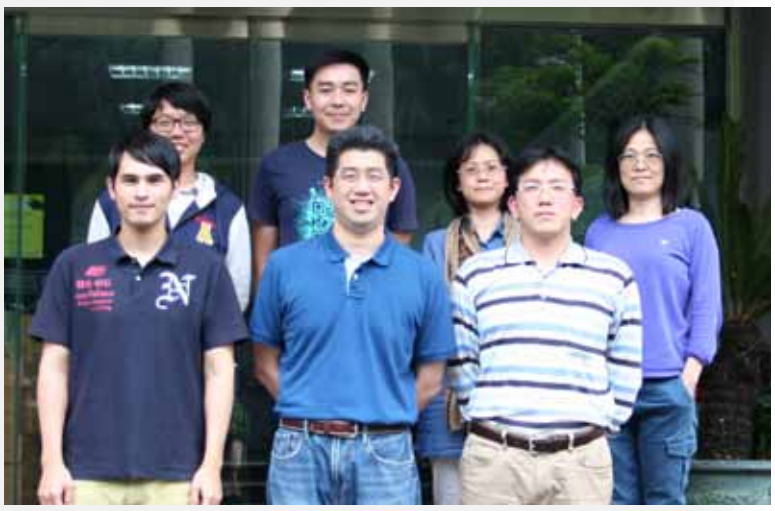
I started taking programming courses when I was a freshman in college. In my junior year, I worked as an intern at ITRI to design a voice command recognition system for the 8051 chip. It was the beginning of my research interest. Later, when I continued the study of voice recognition, I needed to use language models and was surprised to discover that NLP is a field for computer science! It was an eye-opening experience for me. Then I served Defense Industry Reserve Duty at CKIP in IIS for four years after I graduated from the master's program at NCTU. The many NLP studies I had conducted triggered my desire to further my professional skills abroad. Therefore, after my service was completed, I went to Columbia University in New York City, where I worked under Professor Kathleen Mackeown. As I continued to study NLP, I not only gained knowledge from many professors and world-renowned scholars but also learned from them about ways of doing research. While at Columbia, I was fortunate to have joined many intercollegiate grants, such as a multilingual QA program hosted by DARPA and machine translation by NSF. I had chances to participate in big research programs and to learn how they were planned and carried out. During that time, I had many chances to communicate and associate with other professionals from other schools. Before I finished my doctorate, I served as an intern at Microsoft, where I witnessed how a big company valued the importance of NLP research and how related skills have been put to use in services and products.

What are your main research topics and objectives?

Human beings use languages to record knowledge and communicate with each other. NLP is a technology to give computers the ability to handle human languages. Human language is very complicated for computers to understand, in part because there are many ambiguities. For example, when in Chinese we say "Wǒ kǎoshì dé le yādàn," the word "yādàn" here means zero rather than literally a duck's egg. If computers can distinguish such ambiguities, it means computers have understood human languages at some level.

For the past thirty years, to solve the ambiguities of different tasks, we have developed many NLP systems, such as automatic classification of Chinese unknown words, Chinese word identification systems, and sentence parsers. At the same time, in order to construct the infrastructure for Chinese language processing, we have developed part-of-speech tagged corpora, treebanks, Chinese lexical databases, etc. These human-annotated data, however, are relatively limited, especially in contrast to the amount of text data on the Internet. Therefore a

Ph.D. graduation.



CKIP team members.

crucial goal of the NLP field is figuring out how to utilize abundant but unlabeled raw data from the Internet. In recent years, deep learning frameworks have proven effective in many NLP applications. One fundamental research topic of deep learning on NLP is called word embedding, which is how to gather lexical knowledge from a huge cache of unlabeled text data. This kind of approach is very different from that of the traditional Chinese lexical database. Now we are aiming to improve this word embedding process by the incorporation of prior knowledge bases, such as E-HowNet, a traditional Chinese lexical database we have been developing over the past ten years. Our expectations are that the learned word embeddings are more suitable for reasoning about relationships between entities and that their meanings can also be clearly interpreted. This process is very similar to a child's learning process.

In addition, we are focusing on conceptual processing of Chinese documents. The design of knowledge-based language processing systems utilizes statistical, linguistic, and commonsense knowledge provided by our evolving knowledge bases to parse the conceptual structures of sentences and interpret the meanings of sentences. Knowledge-based language processing systems incorporate knowledge bases to form a learning system. Thus, language processing systems increase their processing power due to enhancement of the knowledge bases. Conversely, the knowledge bases are evolving due to the automatic knowledge extraction made by language processing systems.

What are your expectations, both personally and for IIS?

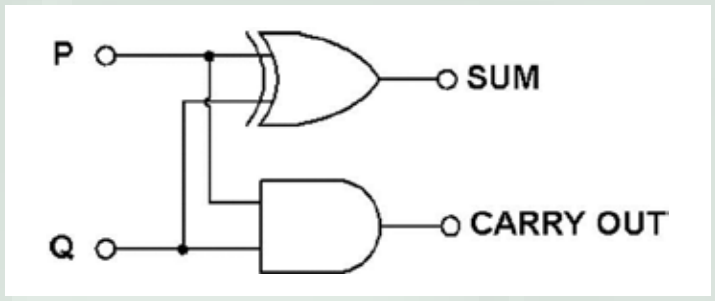
Big data and artificial intelligence applications are now fully embraced by many sectors, including industry and education. In United States, big companies — such as Google, Facebook, Microsoft and IBM, and even new software-innovation companies — have invested substantial resources in the development of a new generation of data-mining and artificial intelligence systems. IIS occupies a favorable position to take advantage of this trend. For example, we have multimedia information processing technology at all levels, and have gained great achievements in information theory, social networks, and bioinformatics. IIS has an excellent opportunity to achieve ground-breaking performance. Increases in exchanges and cooperation with colleagues at IIS may lead to more interdisciplinary study and research in the future. Moreover, building upon the foundation of what IIS already has, I can develop new language analytical systems to strengthen IIS and help IIS have more visibility in the world. I also expect what I develop can contribute to both industry and society.

What are your suggestions for students who would like to engage in research in the field of information science?

For research, I think the most important thing is passion. Only passion for a specific area can drive oneself to get things done, and even to spend numerous hours and sleepless nights in the field and still enjoy it. Information research offers a wide variety of topics. It is very important to find the specific field that you are interested in, you are passionate about, and is also suitable for your personality. Second, research demands critical thinking, a skill that most Taiwanese students lack, mistaking it for criticism when it is instead having a clear idea of the use of reason and logic to rigorously analyze things, find the causes or the nature of things, and see advantages and disadvantages. While students must maintain the spirit of suspicion and criticism, their goal should be to help themselves have more systematic thinking and active learning. Critical thinking can help us to have more curiosity and creativity. If you can find out what is unusual in what is taken for granted, you might hold the key to a breakthrough.

Last but not the least, I want to emphasize the importance of the ability of fast learning. Online courses, such as those in Coursera and YouTube, have provided a great deal of high-quality content and have lowered the barrier for everyone to gain knowledge. Making good use of these resources and maintaining a positive spirit of learning are very important. Fast learning of needed knowledge is a necessary skill in modern society. We should take advantage of its easy availability.

(cont'd from page 9)



Circuit of a half adder. (Fig. 4)

results in large des-cription size and poor efficiency. Recently, researchers (including our group) tackled this issue by designing cryptographic solutions for other computation models, such as the RAM (random-access machine) model, which can be viewed as an abstract model for C++ and Python, and thus can avoid the above-mentioned disadvantage.

Taking one step further, note that "cloud" programming such as MapReduce and GraphLab has emerged to tackle "big data" with massive parallelism, which, however, is not captured by the RAM model. Thus, we propose to consider a PRAM (parallel RAM) model to further capture the power of parallelism in MapReduce and GraphLab. We have initiated the study of cryptography for PRAM and designed cryptographic construction, such as functional encryption, secure computation, and obfuscation in the PRAM model.

Our Research: (Post-)Quantum Cryptography

(Post-)quantum cryptography studies how quantum computation affects cryptography. For example, it is known that quantum computation can solve the factoring problem efficiently. Thus, it can also break any cryptographic constructions that are based on the hardness of the factoring problem. In other words, quantum computation gives more power to the adversary to comprise security. The field of post-quantum cryptography tackles this problem by using new (potentially) quantum-secure computational assumptions, such as lattice-based assumptions, to design quantum-secure cryptographic constructions. On the other hand, quantum can also give more power to the honest party and allow cryptographic tasks that are impossible classically. A prominent example is quantum key-distribution (QKD), which is arguably the most important and close-to-practical application of quantum cryptography. Many countries — such as China, Japan, Switzerland, and the United States — have spent billions of dollars on building QKD networks, with several working prototypes already in existence [6]. QKD allows two parties (e.g., two military camps) to create their secret keys assuming authenticated classical communication. Such a task is impossible to achieve classically without quantum power.

From the above examples, we can see that quantum acts as a two-edged sword in cryptography. In fact, quantum cryptography is still in its infancy, and many new directions remain unexplored. Our group is generally interested in exploring such new directions. In particular, our recent focus is to understand the role of quantum information in post-quantum cryptography. Namely, we consider scenarios where the adversary has access to quantum side information (e.g., through a leakage attack), and develop techniques to ensure security against such adversaries.

Our Group Activities

We aim to create an inspiring environment for theory-prone students to gain exposure to research in theoretical computer science. Currently, we hold weekly seminars and host reading groups on select topics in cryptograph and related fields. We also invite international visitors to give talks about their research. Recently, in cooperation with Prof. Ho-Lin Chen of NTU and Prof. Chung-Shou Liao of NTHU, we organized Theory Days in Taiwan (http://theoryday.github.io), a one-day event with four hour-long talks on theoretical computer science by researchers from around the world. All activities were open to the public. The event has two releated mailing lists:

Theory-event-announcement:
http://ppt.cc/bpgJF
Theory-talk-announcement:
http://ppt.cc/phI5S

References

[1] The Imitation Game, Wikipedia, http://en.wikipedia.org/wiki/The_Imitation_Game
[2] Dan Boneh, Online cryptography class, Introduction to Cryptography, https://class.coursera.org/crypto-preview
[3] Jonathan Katz and Yehuda Lindell, Introduction to Modern Cryptography, Second Edition (Chapman & Hall/CRC Cryptography and Network Security Series), ISBN: 978-1466570269
[4] Salil P. Vadhan, Lecture note, Introduction to Cryptography (CS 127/CSCI E-127), http://people.seas.harvard.edu/~salil/cs127
[5] Craig Gentry, Fully homomorphic encryption using ideal lattices, STOC 2009
[6] Quantum key distribution, Wikipedia, https://en.wikipedia.org/wiki/Quantum_key_distribution

Institute of Information Science
Academia Sinica
128 Academia Road, Section 2, Nangang,
115, Taipei, Taiwan
tel.: +886-2-2788-3799
www.iis.sinica.edu.tw
Publisher: Director Wen-Lian Hsu
Editors: Anita Tien and Huey-Chyi Chris Tseng