**Appendix I:**

# Research Directions of 2020 CCF-Tencent Rhino-Bird Young Faculty Open Research Fund

# Project Topics

**(The following topics are not limited to the given suggested research ds and can vary based on the researcher's background and interest.)**

## 1. Machine Learning

### 1.1 Constrained Machine Learning

Constrained machine learning has great practical value, and has therefore attracted increasing attention in academia. Machine learning algorithms often encounter limited amounts of labeled data, constrained training resources, or a lack of explicit negative definitions during application. Numerous studies have been done, however, the constrained machine learning problem still presents a challenge.

Research is encouraged to focus on the application of image classification problems.

**Suggested research topics：**

1) One shot learning, Semi-supervised Learning, Few shot learning;
2) Positive-unlabeled Learning, Open set learning;
3) Efficient learning on limited GPU resources).

Back to Contents

### 1.2 Research on GNN Algorithm and its Application

The Graph Neural Network (GNN) has proven its powerful graph data learning ability in recent years by naturally integrating node information and topology. These advantages indicate that GNN may have great potential in the application of social network-based recommendation systems. Data from a social network-based recommendation system can be represented as user-user social graphs and user-item behavioral interaction diagrams. The key is to understand the potential relationship between the user and the item. However, there are still many challenges in building a social network-based recommendation system using GNN algorithms. For example, the design of GNN algorithms for multi-source heterogeneous graphs, the graph sampling algorithm, training and inference on GNN for large-scale graphs, and so on.

**Suggested research topics:**

1) GNN algorithm optimization: It considers social influence, friend sampling methods, etc;
2) Research on heterogeneous GNN algorithms;

3) Large-scale GNN algorithm training and inference.

Back to Contents

## 1.3 Key Technologies of Machine Learning in Material Informatics and Design

Machine learning and artificial intelligence have been very successful in the fields of image processing, NLP and security, and offer promising solutions in chemistry, biology and pharmaceutical research. In materials science, simulations of functional material and material properties have posed fundamental problems. Recently, material scientists started using machine learning technologies to help save time and costs in material research. Some techniques have become promising keys to solving critical problems.

**Suggested research topics:**

1) By using machine learning technologies, create new strategies to predict and design functional material based on big experimental/simulation data (applicants should try to obtain open source data or generate data using a conventional simulation package. Tencent's Quantum Lab offers help and guidance);
2) Improve and accelerate conventional material simulation methods by applying machine learning strategies on material modeling on a microscope/atomic scale (the applicant should try to generate training data by using a conventional simulation package. Tencent Quantum Lab offers help and guidance);
3) Construct a material informatics database on Tencent Cloud, and interface with a high throughput HPC calculation stream.

Back to Contents

## 1.4 Machine Learning Research in the Energy Industry

Currently, the energy industry is facing not only the issue of production capacity, but also the issue of energy conservation and emission reduction. Given the vast amount of data accumulated by the energy industry, machine learning is a potentially viable technology to address these issues. Therefore, this project aims to promote the innovative application of machine learning in the energy industry, for purposes such as energy conversion and utilization. You are encouraged to use public data sets such as the UCI Machine Learning Repository.

**Suggested research topics:**

1) Machine learning-based power load forecasting;
2) Machine learning-based boiler combustion control;

3) Machine learning-based forecast of maximum production capacity;
4) Other innovation directions.

Back to Contents

## 1.5 Reinforcement Learning and Game Theory for Multi-Agent Cooperation and Confrontation

Multi-agent cooperation and confrontation is one of the major problems in Artificial Intelligence, with its wide applications in quantitative investment, distributed control, autonomous driving, etc. Game environments, such as board games and mobile games, have been common multi-agent research benchmarks within the global AI research community. In this project, we investigate the problem of multi-agent cooperation and confrontation via reinforcement learning and game theory, and we use games as testing fields. Also, we provide supporting programming interfaces and computing resources through our self-developed AI research platform, "Tencent AI Arena".

**Suggested research topics:**

1) Reinforcement Learning and Game Theory for Perfect-Information Multi-Agent Cooperation and Confrontation;
2) Reinforcement Learning and Game Theory for Imperfect-Information Multi-Agent Cooperation and Confrontation.

Back to Contents

## 1.6 Several Challenges in Conversion Rate Prediction Scenarios

Conversion rate prediction is a major link in the online advertisement delivery technology chain. Diverse conversion targets and pervasive conversion data delays bring many challenges to the prediction process. This project aims to find effective solutions that more accurately predict the conversion rate.

**Suggested research topics:**

1) Design practical multi-objective/multi-task learning algorithms;
2) Model conversion delays or improve prediction models in a targeted manner to reduce the adverse effects of conversion delays;
3) Explore a unified method to develop user conversion interests and integrate diverse user conversion behaviors;
4) Few-shot learning: In a sparse training data scenario, high prediction accuracy can still be guaranteed.

Back to Contents

## 1.7 Research on Machine Learning-Based Code Intelligent Assistive Technology

Machine learning theory represented by deep learning has enabled many successful software development practices. For example, automatic code completion and intelligent prompting can improve software development efficiency; software defect detection and automatic repair can improve software quality; clone detection and code watermarking can ensure legal and efficient code reuse. Code intelligent assistive technology is still a popular research field. In particular, finding ways to use this technology to better assist programmers in code development in the case of large data scale (such as massive code bases) is critical to industrial practice.

**Suggested research topics:**

1) Research on the quality of software development based on machine learning theories like deep learning, for example, code quality assessment, defect prediction, and automatic repair methods;
2) Assisted software development research based on machine learning, knowledge inference and NLP techniques, such as automatic code completion, intelligent prompting, and automatic annotation generation mechanisms;
3) Research on software property protection and traceability, such as detection of code clones and code reuse tracking.

Back to Contents

## 1.8 Deep Learning for Software Security

With the increasing complexity of software, vulnerability mining faces new opportunities and challenges when applied to large-scale source code and binary programs. This research aims to apply deep learning related technologies (such as NLP, GNN, DRL, etc.) to software security. The findings could greatly improve reverse engineering, fuzzing, and bug hunting research.

**Suggested research topics:**

1) Deep learning for binary code composition analysis, such as ISA, compiler, optimization, programmer classification;
2) Deep learning for code generation and code translation, which can be used in interpreter fuzzing or decompilation area;
3) Deep learning for security analysis, such as code similarity detection, API misuse, vulnerability searching, etc.

Back to Contents

## 1.9 Large Batch Size Assessment and Convergence Study under a Given Model and Data Set

In machine learning training scenarios, multiple machines and multiple GPUs are often used to accelerate the training and improve iteration efficiency. However, this causes the issue of batch size (BS) convergence, which can lead to a decrease in convergence precision or no convergence. This project will study how to scientifically assess the reasonable BS range under a given data set and model. It will also examine how to effectively maintain linear convergence during the single-GPU to multi-GPU expansion after an assessment. Tencent will provide its Jizhi Machine Learning Platform for the collaboration to verify the experiment results, and researchers will have an opportunity to implement the platform in a live environment.

**Suggested research topics:**

1) Implement a complete set of quantitative assessment methods for large BS convergence on mainstream open source models. Assess the most reasonable BS value using scientific assessment methods suitable for business practices while maintaining linear convergence in a multi-GPU expansion.

Back to Contents

# 2. Digital Image Processing and Computer Vision

## 2.1 Contrastive Unsupervised Representation Learning

Unsupervised representation learning has received increasing attention from the community. Recently, several methods based on contrastive learning have considerably narrowed the performance gap between unsupervised and supervised learning in vision-related tasks and show their potential practical value. However, it is still a challenging task to further improve the performance of such methods and to apply them in many vision tasks.

Research works are encouraged to focus on the application of contrastive unsupervised representation learning in face recognition and may be conducted in combination with other research frontiers, including but not limited to graph neural networks and meta-learning.

**Suggested research topics:**

1) Unsupervised Domain Adaption for Face Recognition;
2) Unsupervised Large-scale Pre-training for Face Recognition;
3) Unsupervised Multimodal Representation Learning for Face Recognition;
4) Unsupervised Few-shot Learning.

Back to Contents

## 2.2 Visual-to-Audio and Spatial Audio Generation

This project aims to fully explore the correlation between vision and sound. It proposes to generate sound from visual input, thus further producing spatial audio effects that match changes in video image. Such features can be used in movies and virtual reality scenes (automatically generating sounds for virtual scenes) or to provide reference information on images or videos for people with visual impairments. This project intends to use the published video data set containing various sounds to evaluate models (ambient sounds in various daily scenes and sounds from people/animals). The goal is to generate realistic sound that is synchronized with the visual input and which has a good spatial effect.

**Suggested research topics:**

1) Establish an end-to-end model which has visual information and single-channel sound input;
2) Establish an end-to-end model which has visual information and single-channel audio input and high-end spatial audio output;
3) Study the target detection and scene recognition of visual scenes and establish a mapping of target semantics for sound;
4) Explore the target detection of the audio-associated visual scenes and extract spatial and depth information to obtain a target trajectory, thereby generating spatial audio effects.

Back to Contents

## 2.3 Long-range Video Representation Learning

Video representation learning is an important task in video understanding. Existing 3D CNNs based methods mainly focus on learning information from the clip video. Firstly, clip videos can not extract effective inforamation with temporal evolution of long range video. Also, most existing long-range video modeling methods sample a set of frames from the whole video and then fuse them to form the global features. Although this method can capture the global information, they fail to keep up good performance in fine-grained tasks because of simple fusion methods. Directly applying 3D CNN to model long range video information is

impractical due to the intensive computational cost and memory consumption in the 3D convolutions. This project will explore ways to learn long-range video representation both efficiently and effectively, and apply it to video-related computer vision tasks.

**Suggested research topics:**

1) Propose an efficient method for feature extraction on long-range video based on its context;
2) Test and analyze your method on open large-scale video recognition benchmark datasets.

Back to Contents

## 2.4 AI-Enabled Traffic Video Analytics, Modeling, and Applications

Real-time traffic videos and images are the most direct data sources for sensing traffic dynamics. This project will advance video analysis technologies and applications to extract traffic information from videos/images, analyze traffic flows, comprehensively model traffic dynamics in real time, analyze driving behavior, detect accidents, monitor road conditions, and more. Videos/images are collected from either roadside devices or mobile devices on vehicle.

**Suggested research topics:**

1) Machine learning for video/image object detection, classification, and tracking, e.g., vehicle/pedestrian classification and traffic flow analysis;
2) Video/image object positioning, scene understanding, feature extraction, accident/anomaly detection, driving behavior modeling, road condition monitoring, etc.;
3) Techniques for AI model compression, acceleration, transfer and deployment on lightweight mobile devices.

Back to Contents

## 2.5 Security Analysis of Fingerprint and Face Authentication Algorithm

At present, face payment and fingerprint payment have become popular in many scenarios. To ensure the safety of user funds, manufacturers have been working hard to improve the security of their products at the hardware level and machine learning algorithm level. However, the existing biometric authentication algorithms are still under new threats such as adversarial attack. It is important to find ways to test and improve the security of face and fingerprint authentication algorithms, as this will help to ensure the security of Tencent's financial payment

products. In order to improve research efficiency, our team will provide test devices, pre-trained image data sets, and other technical guidance concerning security.

**Suggested research directions:**

1) How to attack the black box face liveness detection and authentication algorithms in the physical world;
2) How to improve the robustness of biometric authentication algorithms and models.

Back to Contents

## 2.6 Research on the Secure and Reliable New-Generation Face Recognition Technology

Face recognition technology based on deep neural networks has achieved great success in recent years. It has been used successfully in many mission-critical scenarios, including finance, security, social networks and e-commerce, etc. However, there are serious security risks in deep neural networks, such as the vulnerability to adversarial attacks and backdoor attacks. Therefore, it is urgent to develop the secure and reliable new-generation face recognition technology.

**Suggested research topics:**

1) Research on effective and efficient attack methods;
2) Research on effective and efficient defense methods;
3) Developing the new-generation face recognition system that is secure and reliable in practical scenarios.

Back to Contents

## 2.7 Quality Assessment and Enhancement for Computer-generated Images

The image quality of games is one of the main determinants of player experience. The video streaming-based cloud gaming solution has now become mainstream. However, due to the limitations of computing power, bandwidth, and display devices, game images will be damaged throughout the encoding/decoding-transmission-display pipeline. Therefore, a reliable perceptual experience assessment model and methods to restore and enhance distorted game images is critical to measure, monitor, and optimize cloud gaming solution processes, including encoding/decoding, transmission, and terminal display device adaptation.

Image/video quality assessment, and studies on image enhancement and super-resolution for optimizing the subjective experience have been hot research topics of industry and academia for many years. However, most of the current research focuses on natural content and there has been little research conducted for CG content, particularly game images. As CG content, games images are quite different from natural images in terms of data distribution and human perception. Such research is of great significance to computer graphics, computer vision, encoding/decoding and image processing.

**Suggested research topics:**

1) Quality assessment model for CG content;
2) Study of image de-noising, enhancement, and super-resolution technologies for CG content.

Back to Contents

## 2.8 Automatic Special Effects Editing for Game Videos

Content producers often conduct secondary video processing to make videos more expressive and attract more users. They will select highlights, then add special effects and use montages to make the finalized videos more rhythmic and interesting. Using algorithms for automatic video editing and special effects processing is an important and challenging machine production issue.

This project focuses on automatic video editing for game video scenarios. It is expected to analyze and understand game videos and background music through computer vision and audio processing algorithms. Montages (including slow motion, playback, pause, play in reverse, zoom in, filter addition, panning, etc.) are added to videos using computational photography to make game videos appear more vivid.

**Suggested research topics:**

1) Automatic video editing technology based on video content understanding: This technology is to automatically add montages to enrich video expression techniques through the analysis and understanding of game video content;
2) Automatic video editing technology based on audio rhythm analysis: This technology automatically selects appropriate video clips for audio segments and adds montages through the analysis of audio rhythm and emotions so that the picture rhythm is consistent with the audio rhythm;
3) Extract semantic label information from game video content and audio content.

Back to Contents

## 2.9 Learning-based Motion Synthesis and Character Control in Games

A common approach to character animation development in games is using a finite state machine that transits between motion clips. It allows game developers to precisely control the behavior of a character and obtain believable animations and definite behaviors that rapidly respond to player input. In recent years, R&D teams have tried using Motion Matching, a new method that searches for the most appropriate sequence from a motion database every few frames, and which has shown promising results in terms of interaction responsiveness and visual quality. It can be a potential alternative to finite state machine, but is still difficult to be applied to large motion databases due to its memory consumption and computation cost, so practical applications are limited.

For this project, teams would like to work together to search for a locomotion animation synthesis solution which achieves an animation quality that is comparable to Motion Capture. This would respond in the shortest time possible, and avoid issues like sliding, shaking, or loss of motion details, and could cover multiple categories of locomotion (walking, running, jumping, climbing, backward movement, crawling, etc. It would be able to sense and adapt to virtual environments such as undulation, obstacles, and common structures.

**Suggested Research topics:**

Locomotion synthesis with user instructions and environment as input by learning from a motion capture dataset.

Back to Contents

# 3. Knowledge Graph and Natural Language Processing

## 3.1 Exploration of In-depth Text Comprehension Technology

Text understanding techniques have been widely applied in scenarios like web searching, personalized recommendation, advertisements, and chat bots, to analyze the structure of natural language text. With the rise of deep learning methods in recent years, great progress has been made in text understanding. However, the ability of AI system to deeply understand the semantics of text is still very limited. The subject of this project is to study and explore deep text understanding techniques based on semantic analysis and knowledge reasoning, including but not limited to:

1) models for deep semantic analysis;
2) new text understanding model;
3) text understanding model with common sense and external knowledge

introduced;
4) fine-grained named entity recognition and its semantic analysis;
5) representation, construction of knowledge graph and reasoning;
6) the application of text understanding techniques in real scenario.

Back to Contents

## 3.2 Construction and Application of Knowledge Graphs in the Information Security Field

Deep learning models in the open domain have bad interpretability and lack priori knowledge. A knowledge graph is an editable structured knowledge carrier. If this knowledge can be integrated into the neural networks, it will effectively improve the interpretability and editability of the networks. Meanwhile, the networks can be customized according to the specific task.

Nowadays, academia and the industry are all studying the knowledge graph and have made significant achievements. However, corporations are not willing to disclose their knowledge graphs, resulting in fewer available high quality knowledge graphs.

Moreover, in terms of the information security field, domain knowledge graph will help to improve the effects of information security. Therefore, it is necessary to build a domain knowledge map of information security.

**Suggested research topics:**

1) Building a million-level knowledge graph in the field of information security;
2) Integrating the knowledge graph into neural networks to improve the interpretability of the networks;
3) Applying the knowledge graph to the information security field.

Back to Contents

## 3.3 Medical Natural Language Understanding

Medical Natural Language Processing faces problems such as oralization of patients, difficulty in data labeling and structure of clinical electronic medical records. We hope to use natural language understanding technologies such as medical knowledge graphs and medical language modeling to improve functions in Tencent's Healthcare Mini Program, including intelligent diagnose, healthcare assistance, online consultation, and more.

**Suggested research topics:**

1) Medical language modeling and knowledge distillation;

2) Long text classification, reading comprehension, summary generation, text matching;
3) Dialogue system and question-answering system;
4) Information extraction, commonsense question answering.

Back to Contents

## 3.4 Multi-Mode Medical Knowledge Graph

The information-based development of medical data has produced a large amount of multimodal data, including text, image, and time series data. This data contains a lot of knowledge that has not been well-explored and utilized. A knowledge graph is an expressive and expandable mode of knowledge representation which considers both human cognition and automatic machine processing. If the knowledge contained in multimodal medical data can be extracted and represented in knowledge graph, then industry applications (such as intelligent recommendation, dialogue system based on knowledge graph etc.) can be better supported.

**Suggested research topics:**

1) Mining knowledge from large multimodal medical data and construct knowledge graph;
2) Apply multimodal knowledge graph to practical medical scenarios, including dialog system and medical publication recommendations.

Back to Contents

## 3.5 AI Technology Research for K-12 Education

With the continuous development of online education, the application of computer technology in education has gradually evolved from the early "education + Internet" to the current "education +AI". The use of AI technology for education has given more attention than ever to how to improve the efficiency and effectiveness of education, thereby to achieve equity in education. This project focuses on AI technology research in K-12 education.

**Suggested research topics:**

1) Basic NLP issues in education include but are not limited to lexical, syntactic and text analysis and subject-related semantic understanding (such as mathematical formula understanding);
2) The construction of a discipline atlas for education and the technology of automatic association between the discipline atlas and educational resources (exercises, teaching plans, teaching videos, etc.);

3) Adaptive learning techniques include but are not limited to knowledge tracking and recommendation of examination questions and learning materials;
4) Automatic correction techniques include but are not limited to the correction of compositions and answers to math questions.

Back to Contents

# 4. Speech Signal Processing and Speech Synthesis

## 4.1 Objective Performance Assessment of Speech Processing Algorithms for Real-time Communication

For real-time voice communication applications, algorithms for de-reverberation, noise cancellation and packet loss compensation need to be deployed to cope with quality degradation due to environmental impacts or network loss. In the process of meticulously measuring algorithm effectiveness and performance, as an alternative and supplement to subjective evaluation methods, objective methods or solutions should be used to better understand users' subjective feelings, and make it easier to evaluate and improve the repeatability of evaluation results. This project will conduct cutting-edge research and technological innovation. Tencent will also provide a variety of practice scenarios for it.

**Suggested research topics:**

1) Study how to measure the effects of single-channel/multi-channel de-reverberation or noise suppression algorithms and determine the performance of de-reverberation/noise suppression in improving sound quality and intelligibility using single or compound indicators and a reproducible evaluation scheme;
2) Study how to objectively evaluate the effectiveness of echo cancellation algorithms and determine the algorithm processing performance using single or compound indicators and reproducible evaluation schemes;
3) Study how to model without reference and assess the quality of one-way speech after algorithm processing or packet loss compensation so as to provide new solutions to predict subjective user assessment results.

Back to Contents

## 4.2 Moving towards Multi-speaker Multi-style High-controllable Universal Neural Speech Synthesis

Most of today's speech synthesis systems only have a limited number of speakers and styles. When business applications require particular voices and styles, people often need to conduct data acquisition, annotation, and model training from scratch. This research aims to expand the scale of data used for training speech synthesis systems. It also explores the model structure and training criteria that can effectively accept all kinds of control signals to build a universal speech synthesis system covering all kinds of tones and styles. This system can flexibly synthesize various kinds of tones such as reading aloud, having a conversation, and speaking moodily or coquettishly.

**Suggested research topics:**

1) Scale the current training data used in synthetic systems and use as much data as possible, such as speech recognition and online audiobook data;
2) Study a neural network structure that can effectively receive control signals;
3) Study the training criteria using various styles to control and monitor signals.

Back to Contents

# 5. Cryptography

## 5.1 Client Key Protection Mechanism Based on National Secret Algorithm System

For cryptography, a very basic scenario is how to ensure the secure storage of keys on the client side. General key embedding codes and file encryption storage solutions have the risk of leakage. A more extreme approach is to use a hardware KEY solution, but in mobile device scenarios or popular applet scenarios, it is unlikely that users will carry the hardware KEY with them. This project seeks to combine better mechanisms based on the national secret algorithm system, and on the premise that the functions of key encryption and decryption and signature verification can be achieved normally, there is a unique solution for key storage and protection. The scheme further protects the key.

**Suggested research topics:**

1) Isolate key data through the software sandbox mechanism;

2) Based on homomorphic encryption and zero-knowledge certification, encryption and decryption can be achieved without leaking key data to memory, and signature verification;
3) On the basis of collaborative signature, further strengthen the mechanism of protecting the local partial cutting key.

Back to Contents

## 5.2 Privacy Preserving Data for Multiparty Collaboration

Many industries try to collaborate by using a nationwide open sharing platform with industry digitalization, especially through the use of blockchain. A DLT-based platform implicitly assumes complete access to whole data set which can be used to extract user activity details. This is a major privacy concern in most of cases, which could pose a great threat to user privacy and benefit. Sharing of private data is prevented because of security concerns. As you know, cryptography, which is used to preserve data privacy, is popular around the world. And the solutions like zero-knowledge proof, secure multiparty computation, homomorphic encryption and public key encryption with equality test are proposed. For this research, we propose some efficient and safe solutions for some special scenarios with cryptography and other technology. The utility is able to perform real time demand management during multi-party collaboration, without knowing the actual value of each party's data.

**Suggested research topics:**

1) Ciphertext equality test instead of public key encryption in the cloud storage;
2) Preserve transaction privacy based on blockchain with secure multiparty computation, which can be used for multiparty collaboration on business risk control;
3) Do trusted computing and deep learning with privacy preserving data under multiparty collaboration, which can be for precise advertising.

Back to Contents

# 6. Database

## 6.1 Concurrent access control technology based on data state consistency

How many data anomalies are there in the transaction processing field of the database? What is the relationship between data anomalies, concurrent access control technology, and serial scheduling technology based on the dependency graph? Is there a model that can: 1) describe the nature of known data anomalies, 2) find more data anomalies，3) using a model to unify the description of all data anomalies and reveal the relationship between data anomalies and concurrent transactions (the previous technology is to describe each data anomaly case by case, scattered without logic)?

**Suggested research topics:**

1) Through in-depth study of the nature of data anomalies, find more data anomalies, and make formal proof;
2) Through in-depth study of existing technologies, such as dependency graphing, conflict serialization and so on, explore the advantages and disadvantages of these technologies, and confirm why these technologies can solve the issues of data anomalies;
3) Construct a new algorithm to achieve the goal of serialization;
4) Research on data anomaly, concurrent access control technology, and isolation level in distributed database system.

Back to Contents

# 7. Intelligent Transportation

## 7.1 A Traffic Information Processing Method Based on Multi-Source Information Fusion

Multi-source information fusion uses multi-sensor information collected from different time slots associated with the specific spatial zones for the data analysis following the predefined rules sets. It enables the consistent explanation and description of the measured objects. Since the new nationwide infrastructure building projects are being launched in China, the intelligent roads serving the intelligent connected vehicles are to be built across the country. The diverse sensors

are expected to be deployed along main roads and/or available in the intelligent connected vehicles. Such multi-source information enables us to access a large volume of real-time road traffic data (including speed, traffic flow, road occupancy and weather) that facilitates the accurate prediction of the variation in the road traffic. It allows us to conduct the traffic control and management in the more effective way. It should be a project with challenges and crucial impact that may assist the strategic decision-making and the timely upgrade in Tencent's products regarding Intelligent Travel or transportation.

**Suggested research topics:**

1) Study the multi-source information fusion system architecture based on the Internet of Vehicles;
2) Use multi-source information fusion to predict the traffic condition of backbone roads over 100 kilometers long;
3) Connect with Tencent's vehicle infrastructure cooperative platform and build and display a traffic information analysis system based on multi-source information fusion.

Back to Content

# 8. Blockchain and Big Data (WeBank)

## 8.1 Research on Transaction Mechanisms of Cross-chain Protocol

At present, there are many kinds of blockchain platforms, and the interfaces, protocols, and architectures of the platforms are all structured differently. To implement cross-chain interaction, a set of reliable cross-chain protocols is required. In addition to the interaction among heterogeneous block-chain platforms, cross-chain protocols need to implement atomic operations among the differently-structured platforms. Existing cross-chain protocols include hash time locked, relay, side chain, distributed key exchange, and the like, but most consider only the single scenario of asset exchange. In the case of more complicated scenarios such as data exchange and invocation of contract interface in a consortium blockchain, more theoretical and engineering exploration and breakthroughs are required to ensure the atomicity of cross-chain operations and the transactional nature of cross-chain access.

**Suggested research topics:**

This topic intends to study the cross-chain protocol transaction mechanism, including the design of inter-chain collaboration and off-/on-chain collaboration, in combination with the principle of distributed consensus technology, cryptography, game theory, and so on. One or more of the above points may be selected for

elaboration, and may be studied in combination with the FISCO BCOS and WeCross platforms.

Back to Contents

## 8.2 Research on Information Security for Blockchain Systems

A blockchain network needs to run on multiple networks, including a public network, a local area network, a cloud network, etc. The blockchain network is connected to various organizations and bears high-value data and transactions. Security risks may be encountered in the private key management of user organizations, a network layer of nodes, a storage layer, a computing layer (contract), a consensus algorithm, and a cryptography algorithm. This topic focuses on attack models for the consortium blockchain, including but not limited to network penetration, DDoS attacks, avatar attacks, access hijacking attacks, injection attacks, smart contract exploitation attacks, and malicious bookkeepers. One or more of the above points may be selected for elaboration, and may be studied in combination with the FISCO BCOS open-source consortium blockchain platform.

**Suggested research topics:**

Researchers can choose to research network penetration, DDoS attacks, avatar attacks, access hijacking attacks, injection attacks, smart contract exploit attacks, malicious bookkeepers, and other related attacks, and combine this with the FISCO BCOS open-source consortium blockchain platform.

Back to Contents

## 8.3 Research on Practical Scenariodriven Privacy Protection Mechanism

Privacy protection is increasingly important for sensitive information pertaining to individuals and institutions. Various technologies have been used in the blockchain architecture to study and explore the integration of privacy protection and blockchains in the industry, including zero-knowledge proof, homomorphic encryption, secure multi-party computation, and trusted hardware computing environments. However, these common cryptography technologies have large performance bottlenecks. Privacy protection has a very broad range, and involves privacy of the identity data of individuals or organizations, privacy of actions, privacy of statuses, and the like. The privacy protection requirements of different scenarios are also quite different, and efficient and feasible scenario-specific privacy protection solutions need to be designed for different scenario problems.

**Suggested research topics:**

This topic intends to study efficient and feasible privacy protection solutions for some specific scenarios, such as privacy protection in a machine learning scenario, a data mining scenario, and an edge computing scenario. One or more of the above points may be selected for elaboration, and may be studied in combination with the FISCO BCOS and WeDPR platforms.

Back to Contents

## 8.4 Research on Application of Decentralized Digital Identity

The technological and industry application of the distributed digital identity system has become one of the key research directions of the blockchain in recent years. The blockchain-based distributed digital identity system has been industrialized with the DID specification as the model (the open-source solution WeIdentity from WeBank), has multiple applications implemented, and achieves selective disclosure and zero-knowledge proof. This topic intends to analyze the application scenarios, related technical challenges (performance, storage, security, privacy, etc.), and technology and compliance risks of the distributed digital identity system; propose innovative scenarios, address the difficulties of traditional business areas when transitioning to distributed digital identities, or focus on technical challenges. A basic understanding of WeIDentity is preferred for this topic.

**Suggested research topics:**

This topic intends to analyze the application scenarios, related technical challenges (performance, storage, security, privacy, etc.), technology and compliance risks of the distributed digital identity system; propose innovative scenarios, address the difficulties of traditional business areas when transitioning to distributed digital identities, or focus on technical challenges.

Back to Contents

## 8.5 Research on Blockchain Applications for IoT

Blockchains and the Internet of Things (IoT) have been among the most popular topics in the industry in recent years. Using the advantages of blockchains to solve the challenges of IoT has become a focus in the industry. This problem has a very broad range, including the entire range from the wide area network, to edge computing, and to blockchains, and involves blockchain-based collaboration models between devices and between device vendors.

**Suggested research topics:**

Researchers may choose to address one or more problems related to IoT device and blockchain access protocol, device identification and management, route

addressing, privacy protection, message response models, and similar topics; or may propose a scenario (industry, home, traffic, energy, social governance, etc.), for which an integral closed-loop IoT and blockchain joint working scenario is designed, and business and technical challenges are identified, and a solution is identified.
Back to Contents

## 8.6 Credit and Risk Assessment of Small and Micro Business

Currently, the information of small and micro businesses is limited, and the existing credit model mainly relies on information such as enterprise tax records to establish the risk model.
In addition to tax information, there should be other relevant enterprise information that may help assess the credit and risk of an enterprise.
Therefore, it is necessary to study whether user risk assessment can be performed to discover trustworthy enterprises as much as possible, so as to provide financing loans in response to the government's policy of supporting small and micro enterprises.

**Suggested research topics:**

1) Legal documents, such as court announcements and written judgments, contain a great deal of risk information, and should be studied to summarize corresponding mechanisms for enterprise risk assessment;
2) The probability that an enterprise user who now uses product A will use product B should be analyzed and predicted;
3) From the public information, names of enterprise legal persons and senior executives can be obtained, but the name may refer to a wrong person who has the same name;
4) Finding ways to distinguish between different natural persons with the same name by using the public information should be studied, to establish the relationship between natural persons and enterprises without requiring sensitive information such as personal identities;
5) Enterprise relationship maps should be studied to enrich data dimensions in the maps to optimize enterprise marketing and risk models.

Back to Contents