Software Engineering and Cloud Computing Course

August 31, 2023

Sargam Gupta
Umeå University

Introduction to Research Topic

I am a PhD student in the WASP NEST CyberSec IT project. My major research area is mostly in privacy models and federated learning. The WASP NEST project clearly defines a few tasks that I must investigate in the due course of my Ph.D. The first task is related to decentralized privacy models. A lot of work has been done on using differential privacy as the model for the protection of data, but it comes with a trade-off between privacy and data utility. Also, what value of the parameter epsilon is appropriate for privacy guarantees is itself an open research question. So, I must investigate some other general privacy models like k-anonymity, integral privacy, plausible deniability, etc. in the distributed setup. The second task is mostly related to federated learning. Federated Learning is a decentralized machine learning technique in which different clients collaborate to form a model without the actual exchange of data. In federated learning, many different attacks are possible like poisoning attacks or backdoor attacks. This task requires me to investigate developing some secure aggregation methods based on federated learning that are resistant to these attacks and follow some privacy models as well. The third task is related to location privacy. Location data is a highly sensitive time series trajectory data that can reveal a lot of sensitive information if not handled with care. Nowadays, a lot of location-based services are available that need to process this location data. As of now, most techniques are based on centralized solutions. My task is to define some possible solutions to the problems related to smart cities and location services and IoT civilian infrastructures. I will also consider some human mobility prediction models. Currently, I am working on building some traffic flow prediction models in a federated environment with high accuracy.

Concepts from Robert's Lecture

0.0.1 Requirements Engineering

Requirements Engineering is the foundational step in the software development lifecycle. It mainly deals with the documentation and analysis of the user's needs for the project. Since my work is mainly in Federated Learning, requirements engineering can be used in my project to improve its working. Firstly, as in requirements engineering, we gather all the information related to the user, in a FL project we need to take into consideration the decentralized data owners' requirements like what kind of data can be used and what constraints and privacy requirements must be adhered to during the model training process. Another important point is that both concepts are concerned for data security and hence, during the requirements engineering process in any FL research project, we can address the privacy concerns of the project by decentralized design of the Federated mechanism. Also, since, the main basis of FL mechanism is the communication of the model parameters, requirements engineering can help in defining these coordination and communication requirements and make the system communicationally robust. Lastly, both Federated Learning and Requirements Engineering are iterative processes. With each communication round the server updates itself with the new updated model parameters and hence, makes the trained ML model more accurate which is very similar to how in requirements engineering, the requirements may evolve as the system is developed and tested to make a better system.

0.0.2 Software Testing

Software Testing is the process of evaluating and verifying any application or system by checking whether the system is doing what it is expected to do or not. In my opinion, software testing is a very crucial part of any research project as translating things from theory to application always needs rigorous testing so that whenever the algorithm or the proposed work is applied anywhere else it performs as expected. In my ongoing project on Location Privacy, software testing is a very helpful tool to ensure that the proposed system works correctly and securely. Firstly, I can do functional testing, in which I can test whether the data anonymization techniques are working as expected or not. Different test cases can be designed to simulate different location data scenarios

and validate the proposed privacy mechanisms. Since location data often deals with sensitive information, we need to ensure the security testing of the proposed work which includes data breaches, unauthorized access and data tampering. Another possible testing could be usability testing. In this type of testing, we can ensure that the anonymization technique applied to the location data leaves the data usable for research or analysis purposes yet ensures that the sensitive information is hidden. Lastly, there should be some compliance testing as there are many different legal and regulatory requirements, such as GDPR in Europe which have some regulations regarding the consent, data retention policies, and reporting capabilities.

Concepts from guest lectures

0.0.3 AutoML

AutoML stands for Automated Machine Learning which refers to the automation in different stages of a machine learning pipeline, like data preprocessing, feature selection, model selection, hyperparameter tuning, and model evaluation. Since, in most of the federated learning projects there is a big part of some automated ML tasks, we can use the AutoML tool for automating this aspect of my project. Data cleaning and Feature Engineering for distributed datasets collected from different data owners can be done using AutoML. This will help reduce the time and manual effort for doing this repeated task for each individual data owner. Another very important use case of AutoML in my project could using the AutoML tool for hyperparameter optimization for a selected ML model. This will help in achieving accurate results faster. Also, AutoML can help in privacy preservation as it can incorporate differential privacy, which is a de facto standard of privacy, by adding noise to the aggregated models in FL or in the individual client datasets. Some AutoML platforms include tools for detecting and mitigating bias and fairness issues in federated learning models which is another very interesting use case of AutoML in my research project. The only catch is that when implementing AutoML in federated learning projects, it's essential to select tools and platforms that align with the specific privacy and security requirements of the project.

0.0.4 Boundary Value Testing

Boundary value testing is a software testing technique that focuses on testing values at the boundaries of their input domains. In the context of federated learning and privacy, I can use boundary value testing to assess the robustness and security of federated learning systems with respect to privacy concerns. First and foremost, in privacy scenarios, the values of the privacy parameters like eps and delta in differential privacy are very crucial. Boundary Value Testing can help in evaluating how the proposed algorithm will behave when the values of these parameters approach their lower and upper limits. Another direct application of boundary value testing which I use quite often is the setting up of aggregation thresholds. In federated learning, data from multiple sources is aggregated to build a global model. Testing different aggregation thresholds helps ensure that the privacy of individual data sources is preserved while still achieving the desired model quality. Another way of incorporating boundary value testing in Federated Learning approaches is testing the frequency of model updates. Boundary value testing can help determine the optimal update frequency that balances privacy and model accuracy as frequent updates could result in models inadvertently memorizing the data. Lastly, Boundary value testing can also help in assessing model fairness by testing scenarios where model predictions are close to the boundary of fairness constraints. This ensures that the FL model doesn't make unfair decisions related to privacy-protected attributes.

Two topics from the list

0.0.5 Security and Privacy

Security and Privacy are two very close terms yet they have a distinct meaning in them. Both of them focus on protecting information, systems, and individuals from unauthorized access, misuse, and various threats.

Privacy usually encompasses the protection of personal data from unauthorized access, use, disclosure, or manipulation. It is a fundamental human right recognized by various laws and regulations worldwide. The main topics incorporated by privacy are

- Data Protection
- User Consent
- Privacy-Preserving Technologies
- Anonymity and Identity Protection
- Regulatory Compliance

Security is a broader concept that includes measures and practices designed to protect information, systems, assets, and resources from various threats, including unauthorized access, attacks, damage, theft, or disruption. Security aims to ensure the integrity, availability, and confidentiality of data and resources. The key aspects of security are

- Cybersecurity
- Access Control
- Cryptography
- Intrusion Detection and its prevention
- Network Security
- Security Standards and Certifications

My research topic is mainly focused on data privacy, especially in federated learning and location privacy. Research challenges in security and privacy are continually evolving as technology advances and new threats emerge. One of the major challenges is the privacy attacks happening on different ML models due to the growing usage of AI in various applications. These attacks include adversarial attacks, data poisoning attacks, and information retrieval attacks. Another challenge is in IoT system architectures and 5G communication systems. Since, these ecosystems are fast expanding, hence, securing the identity and sensitive information of a large number of users in high-speed, low latency networks is a task. There are a lot of issues and challenges in the security of cryptocurrency and blockchain technology as well. Also, since most of the authentication systems are turning to facial recognition and biometrics, there is need of developing secure methods of storing and usage of this sensitive data. Lastly, As technology becomes more ingrained in society, ethical considerations in security and privacy research are gaining importance. Researchers must consider the ethical implications of their work, including issues related to bias, fairness, and accountability.

0.0.6 Regulations and Compliance

Regulations and Compliances are rules, laws and specifications relevant to a particular process. Violations of regulatory compliance often result in legal punishment, including federal fines. Since I work in privacy I can relate to the privacy regulations. Privacy regulations and compliance requirements vary from one country to another, and they continue to evolve. Some of the privacy regulations around the world are stated below.

- General Data Protection Regulation (GDPR): This is the privacy guideline issued by the EU. It governs the collection, processing, and storage of personal data and imposes strict requirements on consent, data protection impact assessments, and data breach notifications. This applies to all EU member states.
- California Consumer Privacy Act (CCPA): It is a privacy law that applies in the California state of the US. It grants California residents specific rights over their personal data, including the right to access, delete, and opt-out of the sale of their data. Sometimes, it is also applied outside the California state to some business that match some specific criteria.
- Personal Information Protection and Electronic Documents Act (PIPEDA): PIPEDA governs the collection, use, and disclosure of personal information by private sector organizations in Canada. It requires obtaining consent for data processing, providing access to personal data, and notifying individuals of data breaches.

Some other countries and their privacy guidelines are the Privacy Act 1988 in Australia, Personal Data Protection Bill in India, Personal Information Protection Act (PIPA) in South Korea, Data Protection Act 2018 in the UK and Act on the Protection of Personal Information (APPI) in Japan.

When working with sensitive datasets in privacy, my work needs to follow the GDPR compliances. There are a few challenges that software systems may face as they have to adhere to complex and evolving legal and industry standards. Firstly, there is a high demand for developing software systems that have embedded privacy by design which inherently protects user data and minimizes the need for manual privacy controls. Another challenge is developing systems that allow the users to provide active and informed consent for data processing, as required by regulations like GDPR and CCPA. Next is cross-border compliances in which the systems developed should be of global regulations and adhere to multiple sets of regulations, each with its own nuances and requirements. Lastly, researching the legal and ethical implications of software design decisions, including the balance between compliance with regulations and ethical considerations, such as user privacy and fairness is also a big challenge.

Future Trends

The conjugation of privacy, especially, federated learning and software engineering is an emerging field that can revolutionize the way software is developed, deployed, and maintained. These trends include privacy-preserving software development, edge AI and usage of AutoML for Federated Learning. Also, software systems could be developed with user-centric privacy controls and strong consent management systems. All in all, these trends signify the growing importance of federated learning in software engineering, driven by the need for privacy-preserving AI applications and the evolving landscape of data privacy regulations.