EURECOM

Sophia Antipolis

Research topics	Trustworthy AI systems based on Knowledge Graphs and LLMs: Attacks and Defenses
Position (M/F)	PhD, thesis offer
Reference offer	SN/MÖ/trustworthyAI/032025
Research Department	Digital Security (SN)
Publication date	March 11, 2025
Start date	ASAP
Duration	Duration of the thesis

Description

The performance of AI technologies relies on access to large datasets of good quality and on the training of an accurate model. This dependence on large data makes AI systems vulnerable to *privacy attacks* that can leak privacy-sensitive information, to *adversarial attacks* that can manipulate inputs or model parameters in order to tamper with the training process, and to *fairness attacks* that aim to modify the existing behavior of the model to induce some bias. The literature features various proposals to mitigate each category of attacks. Nevertheless, these solutions usually work with neural networks, only, and sometimes require considerable and computationally heavy modification to existing algorithms. The aim of the PhD is to study and design new attacks under each category dedicated to AI systems making use of knowledge graphs (KGs), to evaluate them in centralized or distributed settings, and, finally, to propose dedicated defense strategies. More specifically:

- Privacy and confidentiality attacks [1-3] aim to learn information about the data set, the graph, and/or the model. Among potential privacy attacks against the datasets are membership inference, property inference, data leakage and prompt leakage attacks, etc. Some other attacks, like gradient leakage attacks, target the actual model and learn information about its gradients, hence its features. The goal is to investigate existing attacks against standalone or collaborative AI systems that use KGs and/or large language models (LLMs). The candidate will also study and implement new privacy-enhancing technologies (PETs) accordingly. These may be based on differential privacy (DP) or other empirical methods.
- Adversarial attacks [4,5] target the integrity and effectiveness of AI systems. The candidate will investigate existing model poisoning or backdoor attacks during training or adversarial examples at inference time (considering black box, grey box, white box adversary models) and further develop new ones customized to KGs or LLMs. Potential attacks include graph perturbation and word/sentence injection-based attacks for LLMs. The ultimate goal of the research is, once again, to develop defense strategies that can make use of DP mechanisms and/or synthetic data generation.
- Fairness attacks [6,7] aim at manipulating information to undermine AI systems' fairness. The candidate will study the transition from backdoor attacks to attacks against the model's fairness, assuming different types of actors (unintentional or malicious) and evaluate them in various settings, such as, for example when the data across clients are not independent or sampled from the same distribution (i.e., non-IID setting) or under distributed settings. Mitigation strategies will also be explored based on DP mechanisms or empirical solutions customized for knowledge graphs and/or LLMs exploiting the interrelations between fairness and differential privacy [8].

References

[1] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks against Machine Learning Models," in *IEEE Security and Privacy*, 2017.

[2] N. Carlini, F. Tramèr, E. Wallace, M. Hajielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea and C. Raffel, "Extracting Training Data from Large Language Models," in *Usenix Security Symposium*, 2021.

[3] O. Zari, C. Xu, J. Parra-Arnau, A. Ünsal, M. Önen, "Link inference attacks in vertical federated graph learning", in *ACSAC*, 2024.

[4] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin and V. Shmatikov, "How to backdoor federated learning," in *Proceedings of Machine Learning Research*, 2020.

[5] W. Zou, R. Geng, B. Wang, J. Jia, "PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models", in *Usenix Security Symposium,* 2024.

[6] N. Furth, A. Khreishah, G. Liu, N. Phan and Y. Jararweh, "Unfair Trojan: Targeted Backdoor Attacks Against Model Fairness. In H," *Handbook of Trustworthy Federated Learning. Cham: Springer International Publishing,* pp. 149-168, 2024.

[7] I. Gallegos, R. A. Rossi, Barrow, J., M. Tanjim, S. Kim, F. Demoncourt, T. Z. R. Yu and N. Ahmed, "Bias and Fairness in Large Language Models: A Survey," arXiv:2309.00770, 2024.

[8] F. Fioretto, C. Tran, P.V. Hentenryck, K. Zhu, "Differential Privacy and Fairness in Decisions in Learning Tasks: A Survey", September 2022, https://arxiv.org/abs/2202.08187

Requirements

- Education Level / Degree: MS.c
- Field / specialty: computer science with backgroung on privacy enhancing technologies and machine learning

Application

The application must include:

- Detailed curriculum,
- Transcript(s),
- Motivation letter,
- Name and address of a referee for recommendation.

Applications should be submitted by e-mail to secretariat@eurecom.fr with the reference: SN/MÖ/trustworthyAI/032025

About EURECOM

EURECOM is a major Engineering School and a Research Center in digital sciences founded in 1991 as a consortium in the international technology park of Sophia Antipolis. The IMT is a founding member of the GIE. Teaching and research activities are organized around 3 promising fields: digital security, communication systems and Data Science.

EURECOM has a staff of 150 (researchers and support teams) and welcomes 400 international students on the Campus Sophia Tech, the largest information science and technology campus of the region. EURECOM enjoys a privileged geographical environment on the French Riviera (Côte d'Azur), between sea and mountains, at the heart of a dynamic and multidisciplinary ecosystem that promotes high-level scientific and technological innovation.

Social advantages

- International and multicultural environment
- Attractive salary - Corporate saving plans
- Private retirement plan (executive, employer participation of 100%)
- Employee profit sharing policy
- Company health insurance (mutuelle) with high levels of guarantees for the whole family (employer participation of 60%)
- Restaurant vouchers (employer contribution of 60%)

EURECOM is one of Europe's leading engineering schools specializing in digital technologies. It is located in the heart of the Côte d'Azur, in Europe's Silicon Valley (Tech Park Sophia-Antipolis). EURECOM's research teams work in an international, multicultural environment.

EURECOM has a dynamic policy in terms of **inclusion and quality of life at work**. We are committed to diversity and give equal consideration to all applicants, without discrimination. Above all, we look for competence and team spirit.

All our positions are open to **people with disabilities**. EURECOM has set up a disability advisor to provide support and advice, organize accommodation and make positive commitments to personal integration.

As part of its **gender equality plan**, EURECOM encourages gender diversity within its teams. As part of our gender equality action plan, we encourage male applications for administrative positions, traditionally held by women, and female applications for IT and research positions, traditionally held by men.

EURECOM is taking positive action as part of its **CSR policy**. A CSR representative oversees EURECOM's CSR and energy transition policies (electric charging stations, solar panels, waste sorting, etc.).

Web site EURECOM: https://www.eurecom.fr/fr/eurecom/presentation
EURECOM in VIDEO: https://www.youtube.com/watch?v=uIlFcgNijnM
Employee experience:
https://www.youtube.com/watch?v=BHv9zIduzuQ
https://www.youtube.com/watch?v=hvbzzCBups8