# Knowledge-graph Embedding and Self Supervised Learning for Digital Forensics Analysis in Cyber-Physical Systems

## Summary

*It is highly required to collect the scattered data from different sources and correlate and consolidate them to arrive at critical conclusions when cyber incidents are investigated in cyber-physical systems. In this PhD research, the main aim is to induce knowledge graphs from the digital forensic artifacts obtained from various system components, adapt knowledge-graph embedding approaches for the data and apply machine learning for downstream tasks. This research would consider using self-supervising approaches to obtain the optimal knowledge from the unlabeled data. Additionally, time and user embedding techniques would be incorporated into the knowledge-graph representations.*

| | |
|---|---|
| Research field: | Information and communication technology |
| Supervisors: | Prof. Dr. Hayretdin Bahsi<br>Prof. Dr. Matthew James Sorell |
| Availability: | This position is available. |
| Offered by: | School of Information Technologies<br>Department of Software Science |
| Application deadline: | Applications are accepted between June 01, 2022 00:00 and June 30, 2022 23:59 (Europe/Zurich) |

## Description

Representation learning explores the unlabeled data and maps them to low-dimensional vectors in an embedding space [1]. Various downstream machine learning tasks (e.g., supervised, unsupervised) can be applied effectively in this new space. The advances in such embedding approaches have been complemented by self-supervising methods that have started a new era regarding optimal learning from unlabeled data. This is vital progress for areas where finding labeled data is very problematic due to enormous data, lack of expert resources, and privacy concerns.

On the other side, knowledge-graphs store the knowledge base in the form of entities and their semantic relations [2]. A similar embedding idea has been applied to these datasets to utilize them for machine learning tasks [3]. These tasks may enable us to predict the entities and relations which have not been identified in the original knowledge graphs so that new knowledge can be created. It is possible to embed time information to the knowledge graphs and, thus, their representation in the embedding space to handle dynamic knowledge bases that evolve in time [4]. If the source data could be mapped to users, then incorporating the user into this space is possible by using some user embedding approaches, which have been demonstrated especially in social media applications [5].

Cyber-physical systems take more role in connecting the physical and cyber spaces. The data obtained about a physical phenomenon is transferred to various other sub-systems, creating a huge number of interactions and information flows that are difficult to track. In case of an incident, let it be physical or cyber (e.g., a crash of a self-driving car or its compromise by a cyber attack), the data artifacts residing in end devices, edges/hubs/gateways, or central databases/clouds are valuable sources for digital forensic investigations. However, the data is highly distributed on heterogeneous system components with various hardware and OS types. It is important to collect the scattered data from different sources, correlate and consolidate them for arriving at critical conclusions when cyber incidents are investigated. The notion of knowledge-graph could be instrumental to do such consolidation in digital forensic investigations by identifying and enhancing the semantic relations between various artifacts. It is important to note that digital forensics is one of the problem domains in which labeled data is so scarce and very difficult to obtain. This is the main reason for the relatively low adaptation of machine learning to this problem domain.

In our research, we have already created ontologies and web semantic frameworks for digital forensic analysis. In one case study, we created a framework that correlates artifacts obtained from volatile and non-volatile memories of a device. In the second one, we focused on collecting artifacts from different devices in an IoT environment.

In this Ph.D. research, the main aim is to adapt knowledge-graph embedding approaches for mapping the digital forensic data to embedding space and apply machine learning for downstream tasks. As labeled data in the digital forensic domain is so scarce, this research would consider using self- supervising approaches to obtain the optimal knowledge from the unlabeled data.

Time and user information are critical in digital forensic investigations. Although it is possible to identify such information in some data sources, it may be missing or inaccurate in some situations. This research will investigate how time and user embedding techniques would be incorporated into the knowledge-graph representations and how the resulting vectors embedding space could be utilized for machine learning tasks requiring time- and user-based perspectives.

The Ph.D. candidate is expected to address relevant research questions that can be derived from real-world digital forensics problems.

**Support for Teaching and Supervising Activities**

The candidate will take part in developing courses regarding mobile and IoT forensics and supervising M.Sc./B.Sc. level theses in similar topics.

**Candidate's Background and Knowledge**

This position requires a solid background in machine learning or web semantic and familiarity with digital forensics concepts.

**References**


1. S. M. Kazemi *et al.*, "Representation Learning for Dynamic Graphs: A Survey.," *J. Mach. Learn. Res.*, vol. 21, no. 70, pp. 1–73, 2020.
2. S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Networks Learn. Syst.*, 2021.
3. Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, 2017.
4. J. Leblay and M. W. Chekol, "Deriving validity time in knowledge graph," in *Companion Proceedings of The Web Conference 2018*, 2018, pp. 1771–1776.
5. S. Pan and T. Ding, "Social media-based user embedding: A literature review," *arXiv Prepr. arXiv1907.00725*, 2019.


To get more information or to apply online, visit https://taltech.glowbase.com/positions/516 or scan the the code on the left with your smartphone.