Chapter 8

Conclusion and Perspectives

8.1 Introduction

Over the course of six chapters, we have explored the potential of federated learning to build collaborative intrusion detection systems. We termed this new approach Federated Intrusion Detection System (FIDS), and have shown how it can be used to address the limitations of traditional Collaborative Intrusion Detection Systems (CIDSs). We have observed the appealing properties of this approach, but also identified major limitations, some of which we have addressed in the second part of this manuscript. In this concluding chapter, we start with summarizing the contributions of this thesis (Section 8.2), before outlining how future work could address their limitations and further improve the state of the art of FIDSs in Section 8.2.2. We end this manuscript with a discussion on the research perspectives of FIDSs and CIDS in general, laying out a roadmap for future research in this area (Section 8.3).

8.2 Thesis Summary

In the Introduction chapter of this manuscript, after introducing the context and motivation of this work, we formalized our research objective as the following question: *Can Federated Learning (FL) serve as a trustable knowledge-sharing framework for collaboratively improving Intrusion Detection Systems (IDSs)?* We derived four research questions from this main objective (Questions RQ1 to RQ4) and structured the rest of this manuscript around these questions. The first part of this manuscript, Part I, was exploratory, and aimed at understanding the core characteristics of FIDSs. We started by providing the necessary background on intrusion detection systems and federated learning in Chapter 2, before reviewing the state of the art of CIDSs and FL in Chapter 3, addressing Question RQ1. This first part ended by illustrating the potential of FIDSs through a practical application in Chapter 4, and highlighting the challenges associated with Questions RQ2 and RQ3: data heterogeneity and malicious participants. The second part of this manuscript, Part II, focused on addressing some of the identified limitations of FIDSs to provide answers to Questions RQ2 to RQ4. We answer these research questions hereafter, with references to the corresponding chapters.

8.2.1 Answering the Research Questions

— Question RQ1: *What makes applying FL to IDSs specific?*

While introducing the preliminary concepts required for this thesis in Chapter 2, we enumerated challenges of CIDSs that motivated exploring of FL. That alone does not answer the question, so we performed a comprehensive review of the state of the art in Chapter 3, where we proposed a taxonomy that describes FIDSs along five axes: *data*, *local operation*, *federation settings*, *aggregation*, and *evaluation*. This highlights the first specificity of FIDSs, as two out of our five criteria cover most of the existing FL taxonomies: *federation settings* and *aggregation*.

We discuss in the same chapter other critical aspects of intrusion detection, emphasizing the importance of explainability, personalization, and adversarial robustness in this context. Further, the specific types of data distributions and partitioning encountered in IDSs are also unique to this domain, such as the attack classes overlap encountered in Chapter 5. Finally, the lack of datasets and the massive variety of IDS use cases make it difficult to generalize the results of FL research to this domain.

— Question RQ2: *Can FL be used to federate IDSs across heterogeneous data sources?*

Observed in both FL and FIDS literature (*c.f.* Chapter 3), data heterogeneity is a major challenge to real FL deployments. We illustrated this issue in Chapter 4, where we highlighted both the benefits of data heterogeneity (*e.g.*, improved generalization and knowledge-sharing) and the challenges it poses, particularly in terms of convergence and performance. Chapter 5 highlighted another issue with data heterogeneity: when participants are two different, identifying malicious contributions becomes more difficult.

We proposed in Chapter 6 a novel approach to address this issue, RADAR, which leverages three components: (i) a *cross-evaluation* scheme that modifies the FL workflow to collect evaluation feedbacks; (ii) a *clustering* algorithm that groups participants based on their feedback similarities, providing a more subjective view of the participants' differences; and (iii) a *reputation system* that analyses feedbacks over time to weight the aggregation process accordingly. However, because of the lack of appropriate distributed IDS datasets, we are limited in our evaluation. We propose in Chapter 7 an unconventional approach to address this issue, leveraging constraint-based topology composition to generate synthetic datasets that mimic the characteristics of independent organizations. Consequently, while we have evidences that FL can could be used to federate IDSs across heterogeneous data sources, this still represents a major research direction.

— Question RQ3: *How does FL handle malicious contributions in a federated IDS?*

Malicious participants represent a major threat to collaborative systems. FL is no exception, and we have illustrated in Chapter 4 how the lack of control over the participants' contributions is indeed a major concern. To better understand this issue, we performed a systematic analysis of the impact of label-flipping attacks against a FIDS in Chapter 5, and build an evaluation framework to facilitate this process. Label-flipping attacks are especially interesting for they are straightforward to implement and can be applied under any threat models. Our results indicate that the impact of such attacks can be quite significant, but also that FL's inherent construction also mitigates parts of their effect, until a certain threshold. With Chapter 6, we introduced a novel approach to detect malicious participants, including large groups of colluding attackers. Yet, our experiments call to be extended to generalize our findings.

— Question RQ4: *How can one assess and ensure the trustworthiness of the other participants' contributions?*

Even without malicious intent, uploaded model updates can have a negative impact on the global model. This can be explained by heterogeneity (see Question RQ2 and chapter 4), but also by the quality of the training data. RADAR (Chapter 6) is a first step towards addressing this issue, as it provides guarantees on the quality of the participants' contributions based on evaluation metrics. In fact, from the point of view of our aggregator, whether a participant is malicious or not is irrelevant: the collected feedbacks assess the model quality, not the participants' intentions. Most importantly, RADAR is, to the best of our knowledge, the only reputation-aware approach in FL that leverages actual participants' feedbacks. This represents a major shift in the way we assess the trustworthiness of the participants' contributions, by relaxing the assumption that participants cannot upload anything but model updates.

8.2.2 Future Work

The contributions of this thesis are a first step towards understanding the potential of FL for building collaborative IDSs. We identified above some limitations of our work, notably to extend our assessment study and our data-generation project. In this section, we discuss our future work to address these limitations, and provide steps to improve the state of the art by building on the contributions of this thesis.

Extending the assessment study. Chapter 5 provided the first systematic analysis of the impact of label-flipping attacks against a FIDS. While this chapter already extends our original publication at ARES 2024 [LBA24b], the results are still limited to two datasets, one model architecture and type of poisoning attack (*i.e.*, label-flipping).

Several steps can be taken to extend this work. First, the study can easily be extended to other types of poisoning attacks, such as backdoor attacks or model poisoning, and to other datasets. Likewise, implementing other aggregation baselines, including mitigation strategies, would provide a comprehensive toolbox to evaluate the robustness of FIDS and provide meaningful comparisons between approaches.

Improving RADAR's scalability. Recent works in FL have shown the potential of decentralized approaches to improve both the scalability and the robustness of such systems. This is in fact one of the major limitations of RADAR as it stands: the centralized nature of the FL approach, coupled with the need to collect evaluation feedbacks, makes it difficult to scale to large numbers of participants. Fortunately, RADAR's design makes it a good candidate for decentralization. In a Peer-to-Peer (P2P) network, for instance using a gossip-based protocol, participants could propagate model updates, but also evaluation feedbacks on other participants. Then, the clustering and reputation systems could be computed locally, from the point of view of each participant, each client could aggregate a personalized model based on the reputation of the other participants. As emphasized in the conclusion of Chapter 6, this would represent a key step towards a truly decentralized, trustworthy, and privacy-preserving CIDS.

Generating independent datasets. We mentioned several times throughout this manuscript the lack of appropriate truly distributed IT datasets for FIDS research. This is a major limitation to generalizing the community's findings to real-world applications. In Chapter 7, we proposed FedITN_gen, a novel approach to create synthetic network topologies enabling the generation of heterogeneous datasets. Our prototype is a first step towards this goal, but it does not support the execution of scenarios yet. Consequently, deploying scenarios and evaluating the generated datasets represents the natural next step. Moreover, the current implementation includes naive algorithms to filter out irrelevant sub-topologies after generation. We believe that the entire problem can be modeled as a Constraint Satisfaction Problem (CSP), and plan to investigate this direction. Finally, the recent advances in Large Language Models (LLMs) and Graph Neural Networks (GNNs) could be leveraged to bridge the gap between theoretical topologies generation and their deployment, by coupling these models with Infrastructure as Code (IAC) frameworks like Terraform.

8.3 Perspectives: Going beyond FIDSs

In the last section, we discussed the future work we envision to expand on the contributions of this thesis. However, the scope of this thesis is limited to specific aspects of FIDSs, and we believe that FL has opened new research directions in the field of CIDS

that are worth exploring. In this section, we discuss some of these perspectives, and in particular four research axes that we believe will shape the future of CIDS research.

8.3.1 Federated Learning and Derivatives

FL is a core concept in this thesis, and more generally an essential component of FIDSs. However, FL is not the only collaborative learning framework available, as the term now encompasses a wide range of techniques that can be used to build collaborative systems. In this section, we discuss some of these techniques, and how they could be used to improve the state of the art of FIDSs.

Aggregation strategies Aggregation strategies are naturally at the core of FL, since they define how the participants' contributions are combined to build the global model. Furthermore, they can serve additional purposes: privacy preservation (*e.g.*, secure aggregation [?], differential privacy [?]), robustness (*e.g.*, model weighting [?], clipping [?], noise injection [?]), or incentivization. [?]A particularly interesting direction for FIDS is understanding and handling data heterogeneity. While a lot of research has been done on this topic in the context of FL [?], the performance of these strategies in the context of FIDSs is still an open question.

Horizontal architectures and trustworthiness A second import research direction in FL is relaxing its core architectural assumptions to overcome the central server dependency [Kai+21]. Multiple solutions have been proposed in the literature. However, this poses new challenges of trust that are particularly relevant in the context of CIDSs, as collaboration is typically done between trusted parties. Consequently, the question of how to build trust between participants is a major concern, where decentralized reputation systems [?]could play an important role. Since such systems usually rely on evaluation [?], can we trust participants to perform the evaluation correctly? Machine Learning (ML) training and evaluation in trusted enclaves [?]could become of great help to that regard. The lack of central authority also brings new interoperability challenges: Which model architectures is used? With which hyperparameters? With what Deep Learning (DL) framework? Standardization could help in that matter, whether it comes from entities like the Internet Engineering Task Force (IETF) or more ad-hoc consortiums. Recent efforts towards standardizing the exchange of Neural Networks (NNs), such as Open Neural Network Exchange (ONNX), could be envisioned to support interoperable and decentralized collaborative learning applications.

Leveraging the communication layers In FL, everything happens at the application layer, which makes it pretty agnostic of the architectural choices of the lower layers. However, some of these technologies present particularities that can be exploited to optimize

bandwidth consumption, computing resources, or both. A first example that connects with the decentralization objective mentioned above is Information-Centric Networking (ICN), which provide interesting properties, such as information caching and in-network computing. While few works have been published in this direction [**?**], we believe that ICN can provide FL with significant performance gains in large-scale settings. Yet, this obviously introduces new challenges, such as how to address content (such as model updates) without never accessing data. Another example is wireless protocols, were the information in inherently broadcasted on a common medium. Consequently, decentralized approaches (*e.g.*, gossip learning) could significantly benefit from sharing model updates with all available clients within range, instead of multiplicating one-to-one connections.

8.3.2 Modern Detection Techniques

Naturally, the second main component in FIDS the local algorithm. Yet, we have observed that most of the literature remained focused on scaling up the existing local intrusion detection approaches, mostly selecting *off-the-shelf* models to see how they work in federated settings. While these works have provided interesting first insight into FIDSs' abilities, we believe they only have scratched its surface.

Novel algorithms and representations The rise of DL algorithms has revolutionized the field of intrusion detection, especially with new classes able to capture temporal dependencies in data. Recurrent Neural Networks (RNNs), and more recently transformers [**?**], have been successfully applied to intrusion detection. Another breakthrough happened with the use of knowledge graphs to represent Network-based Intrusion Detection System (NIDS) data [Lei+20]. Yet, these techniques introduce more abstract knowledge extraction, and the question of aggregating such knowledge in FL and similar collaborative learning approaches remains to be addressed. A few works in the literature successfully federated RNNs, for instance for event prediction [Nas+22], but to the best of our knowledge, it is not the case with graph data. Finally, LLMs started to showcase applications in analyzing log data, and the recent proposal of federating LLMs might represent an opportunity to train such algorithms on actual data, going beyond the simple local fine-tuning.

Explainability and Semantic Explainable Artificial Intelligence (XAI) has received much attention over the last years, as the trustworthiness of ML algorithms started to be questioned. This becomes even more relevant with FL. Not only are the shared contributions black-boxes, but their averaging contributes to making the global model even more opaque. Consequently, the question of how to explain the global model's decisions is of great importance, but so is the ability to track and explain the contributions of each participant. This is particularly relevant in the context of FIDSs, where the global model's decisions can have significant impacts on the participants' security. A first step in this

direction is to apply semantic tagging to the shared contributions, to provide a more fine-grained understanding of the "knowledge" shared by each participant, and the resulting global model. Such tagging would also have applications in model personalization, as one could express to domains they want their model to be more sensitive to.

8.3.3 Evaluation

Throughout the manuscript, we discussed at length the challenges that come with evaluating FIDSs. Addressing some of them is even part of the contributions (*c.f.* Section 8.2.1) or future work (Section 8.2.2) of this thesis. Yet, we believe that the evaluation of FIDSs is a research axis in itself, and that it deserves more attention from the community.

Datasets and benchmarks The first striking issue is the lack of appropriate datasets, due to they way most of the public datasets are generated. Multiple strategies can be envisioned to address this issue, although we focused on generating heterogeneous topologies in this thesis. Other strategies could include generating synthetic datasets, using generative models or data augmentation techniques, as well as working on data transformation techniques to make variations of existing datasets, while preserving the validity of the generated data. Another direction lies in the creation of evaluation frameworks, that would allow researchers to evaluate their models in a more systematic way. We believe that this thesis provides a starting point with Eiffel to define a more robust and systematic evaluation methodology for FIDS, including the definition of relevant metrics and the development of benchmarks for this purpose.

Reproducibility For a long time, researchers have been publishing their results without providing the necessary tools to reproduce them. This is particularly true in the field of ML, where the choice of hyperparameters, the data preprocessing, or the model architecture can have a significant impact on the results. FL is particularly sensitive to these issues, as its distributed and partially asynchronous nature makes it difficult to reproduce the exact conditions of a given experiment.

Multiple initiatives have been launched to address this issue, such as the *Baselines* project of the Flower framework [Beu+20] which collects baselines developed by the community using the framework to facilitate comparisons. The research community in general has also started to pay more attention to reproducibility. In France, the Groupement de Recherche (GdR) Réseaux et Systèmes Distribués (RSD) has launched a working group on reproducibility [1], whose goal is to facilitate the community's discussions on this topic. At an international level, the Association for Computing Machinery (ACM) has launched

1. https://gdr-rsd.fr/gt-reproductibilite/

the *ACM Artifact Review and Badging* initiative, which aims at promoting the publication of artifacts associated with research papers, as well as a dedicated working group on reproducibility [2] and a novel venue for the topic, the *ACM Conference on Reproducibility and Replicability*. We followed these guidelines in this thesis, and specifically in our assessment study (Chapter 5), and we believe that this is a major step towards improving the reproducibility of our results.

TB!

8.3.4 Integration in the Regulatory Landscape

Sûr ? Ça me paraît fort pour le 1er.

Finally, the last research axis we believe is of great importance is the integration of FIDSs in the regulatory landscape. In Europe, the General Data Protection Regulation (GDPR) and the *Data Act* are two major regulations that have a significant impact on the way data is handled, making data sharing and processing more difficult for organizations. This is often used as a motivation to explore FL strategies in all kinds of applications, especially in the health sector. The more recent *AI Act* introduced new restrictions on the use of Artificial Intelligence (AI) and ML algorithms, and it is likely that FL will be impacted by these regulations as well. A first position paper on the topic has been recently shared on arXiv [Woi+24], although it does not address some of the key challenges introduced with this regulation, such as how to assess fairness without accessing the data.

In the specific context of cybersecurity, the French security agency, Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI), has published a series of guidelines on how to share information between organizations, and how to report incidents. Among the mission of the agency is also to share knowledge acquired while monitoring its beneficiaries, and we believe that FL could be a key technology to achieve this goal. More generally, *is FL compatible with the current and upcoming regulatory landscape for cybersecurity-related knowledge sharing?*

8.4 Closing Remarks

In this concluding chapter, we have summarized the contributions of this thesis, and outlined some of the future work that could be done to improve the state of the art of FIDSs. We have also discussed some of the research perspectives that we believe will shape the future of CIDS research, and in particular the role that FL and its derivatives will play in this context.

More than a distributed learning technique, FL has demonstrated how parametric models can be merged and modified using simple mathematical operations. This has opened a new dimension in the field of ML and all its applications. In the context of CIDS, this represents a major shift in the way we think about intrusion detection, and

2. ACM EIGREP: https://reproducibility.acm.org/

how we can leverage the knowledge of multiple organizations to build more robust and efficient systems.

Bravo! Ça sera une très belle thèse ! ;-)
Bon courage pour la dernière ligne droite !