Complex Data: Learning Trustworthily, Automatically, and with Guarantees

Luca Oneto[1], Nicolò Navarin[2], Battista Biggio[3], Federico Errica[4], Alessio Micheli[4], Franco Scarselli[5], Monica Bianchini[5], Alessandro Sperduti[2]

[1]Università di Genova - Via Opera Pia 11a, 16145, Genova, Italy
[2]Università di Padova - Via Trieste 63, 35121, Padova, Italy
[3]Università di Cagliari - Piazza d'Armi, 09123, Cagliari, Italy
[4]Università di Pisa - Largo Bruno Pontecorvo 3, 56127, Pisa, Italy
[5]Università di Siena - Via Roma 56, 53100, Siena, Italy

Abstract. Machine Learning (ML) achievements enabled automatic extraction of actionable information from data in a wide range of decision-making scenarios. This demands for improving both ML technical aspects (e.g., design and automation) and human-related metrics (e.g., fairness, robustness, privacy, and explainability), with performance guarantees at both levels. The aforementioned scenario posed three main challenges: (i) Learning from Complex Data (i.e., sequence, tree, and graph data), (ii) Learning Trustworthily, and (iii) Learning Automatically with Guarantees. The focus of this special session is on addressing one or more of these challenges with the final goal of Learning Trustworthily, Automatically, and with Guarantees from Complex Data.

1 Context

The use of Machine Learning (ML) for extracting actionable information from data has recorded unprecedented success in a wide range of decision-making scenarios, ranging from healthcare to education and cybersecurity. However, the increasing digitalization and datification of all aspects of people's daily life, and the consequent growth in the use of personal data, are increasingly challenging the current development and adoption of ML algorithms. First, the increasing complexity and amount of data available in these applications strongly demands for ML models that can be trained directly on complex structures. Indeed, Graphs inherently capture information about entities, attributes, and relationships between them, rather than requiring domain experts and data scientists to face the challenging and time-consuming problem of designing a suitable vector-based data representation. Second, ML algorithms should not only be designed to achieve high technical and functional standards. As the automated decisions provided by these algorithms can have a relevant impact on the people's life, their behavior has to be aligned with the values and principles of individuals and society. This demands for designing automated algorithms that we, as humans, can trust, fulfilling the requirements of fairness, robustness, privacy, and explainability. Third, designing effective ML algorithms requires skills and expertise developed at different levels. This substantially hinders the democratization and widespread availability of such technology for society at large, which in turn demands for improving the level of automatization and systematization of their design process, while also providing guarantees on their performance. In summary, the next generation of ML algorithms should not only be able to learn from graph data but it should also be automatic, guaranteed, and trustworthy. To this end, there is a need to simultaneously tackle the aforementioned

challenges of (i) learning from graphs, (ii) learning trustworthily, and (iii) learning automatically with guarantees, within a cutting-edge unified theoretical yet practical framework.

2 State-Of-The-Art

The field of *learning from graphs* has a long history but yet fast evolving [1, 2]. Kernel-based ML methods constitute a reliable approach for many learning tasks obtaining state-of-the-art performance in multiple applications [3]. They are based on the idea of defining (possibly implicitly) substructures as features. In parallel, neural-based models were developed to efficiently and automatically learn these features, with approaches based on recursion [4] or on deep graph convolutions [5] that started the research field. The underlying idea is to compute a representation for each element that depends on its local neighborhood, iterating this process to let information flow across the structure. Reservoir computing [6], graph generative models [7], and Bayesian methods [8, 9] are just a few of the most effective further directions of the field. Still, learning from graphs poses many open research challenges like further improving efficiency by reservoir computing [6] or incremental approaches [5], analyzing in depth the properties related to the informative content of layers [10], studying long-term dependencies issues [11], and understanding causal factors encoded in the layers [12]. Cross contamination between different approaches is also needed by data-dependent representations kernel definition [13] via multiple kernel learning [14] or by incorporating priors provided by graph kernels in graph neural networks [15, 16]. Finally, enriching these methods with trustworthiness and automatization into a unified framework with guarantees on both technical and human-relevant metrics remains an open problem.

The field of *learning trustworthily* is no less fast evolving albeit much younger. It focuses on incorporating the human-relevant requirements of fairness, robustness, privacy, and interpretability into ML [17, 18]. Fairness concerns were raised by society when ML started to show human biases (e.g., gender or race biases) requiring the development of formal fairness metrics and mitigation strategies [19, 20]. Robustness requires ML models to be neither misled by carefully-crafted malicious input data nor induced into unexpected errors by poisoned training data [21]. Privacy-preserving ML addresses the self-contradictory problem of keeping private information about individual observations while learning useful information about a population. Current solutions [22–24] include corrupting data or models outcomes when data need to be centralized (e.g., via differential privacy) or keeping the learning procedure decentralized (e.g., via distributed protocols). Finally, explainability aims to provide model/outcome explanations for black-box ML (e.g., deep networks) engendering trust in their users. In fact, especially on graphs where glass-box ML tends to perform poorly [25], black-box ML often provides good performance. Recent research has unveiled that, not surprisingly, ML algorithms trained on graphs may suffer even more from untrustworthiness [26–29]. This highlights the need of addressing the problem of learning trustworthily on graphs by providing suitable definitions of robustness, fairness, privacy and interpretability and ways to optimize them. Moreover, most of the current works, with very few exceptions [30–32], are focused on a single aspect of trustworthiness, without considering the problem from an holistic perspective.

The third research field, *learning automatically with guarantees*, focuses on

the problem of tuning and assessing the ML performance trained on graphs by means of both technical and human-relevant metrics. Tuning the performance of a learning algorithm (e.g., optimal architecture and hyperparameter configuration) is a key problem for which effective and theoretically-grounded solutions have not been found yet [33]. The problem is even more challenging when the architectural hyperparameters are large in number, discrete, and non-linearly related. Gradient-based, gradient-free, and game-theoretic-based methods are the state-of-the-art solutions [34–36]. Nevertheless, when multiple metrics need to be optimized, multiple optimal solutions exist and finding the best compromise introduces an additional level of complexity. Guaranteeing the performance is closely related to theoretically characterizing the computational capabilities and the approximation abilities [37–40], the worst-case behavior [41, 42], and the asymptotic and finite sample behavior [33, 43] using different metrics of accuracy and trustworthiness. While some attempts to optimize and bound performance by specific metrics have been pursued [44–48], a unified theoretically-grounded approach for the graph domain is not yet available.

3 Future Perspectives

The problem of Learning Trustworthily, Automatically, and with Guarantees from Complex Data is to simultaneously face its three main challenges, (i) Learning from Graphs, (ii) Learning Trustworthily, and (iii) Learning Automatically with Guarantees, proposing a unifying framework to review, generalize, and advance the state of the art.

The first challenge needs to be tackled along four lines: Kernel-, Neural Network-, Incremental/Randomized-, and Hybrid-based approaches for graphs.

Kernel-based ML for graphs usually resorts to a convex learning problem, even though the kernel must be defined before seeing the data either implicitly (i.e., via kernel trick) or explicitly (i.e., via feature map). While implicit kernels can consider a combinatorial number of features, explicit kernels can be defined to be very efficient. Nevertheless, in both cases, the problem is to simultaneously define a rich and expressive feature space developing efficient and effective kernels. To this end, it is required to design kernel methods able to induce a data-dependent representation via incremental or multiple kernel learning. Kernels allow one to directly include constraints of fairness (fair kernels), robustness (including differentiability with respect to the original space), privacy (including noise injection in the feature design), explainability (leveraging convexity to evaluate the quality of the approximation performed by the interpretable models), and automatization (designing kernels for graphs that are differentiable with respect to the hyperparameters).

Graph Neural Networks - GNNs aim at learning a data-dependent representation from scratch using a layered approach in which each layer encodes a particular view of the structure. Some of the challenges of this field are related to the intrinsic difficulty in optimizing GNNs. A possible solution is the definition of GNNs of quasi-linear complexity in the size of the structure (e.g., via subsampling) and to exploit the minimum number of layers to effectively propagate contextual information. Other compelling challenges are the need for developing more expressive convolutional layers (e.g., via attention mechanisms), understanding and leveraging the effect of long-term dependencies in arbitrarily complex structures, fully exploiting hierarchy and compositionality of the feature spaces, and studying the dynamical properties of recurrent architectures.

GNNs also allow for graph generation, which has the potential of being used in many applications or to extend small datasets. The generation of graphs using a hierarchical strategy, e.g., by generating a graph at different resolution levels, may potentially enable generation of larger graphs, overcoming the main limitation of current methods. These ideas can be instrumental towards the design of representation learning methods for graphs with fairness constraints, leveraging subsampling to enforce privacy, and exploiting attention in convolutional GNNs to facilitate and improve visual explanation. Finally, it is possible to define graph deterministic or probabilistic aggregators suited to federated learning (e.g., not requiring the synchronization of messages).

Incremental and randomized models for graphs address fundamental open issues concerning the design of efficient models. Indeed, graph ML models often incur a significant computational burden that prevents scalability to large datasets and structures in contemporary real-world applications. To this aim, it is possible to develop GNNs and Bayesian methods with randomized, incremental, and online approaches. Randomization via reservoir computing leads to efficient design of recurrent GNNs by keeping the recursive part untrained. Combined with the study of the properties of these systems (e.g., contractivity of the state dynamics), it is possible to further develop extremely efficient novel methodologies for graphs. Incremental approaches rely on the idea that the layers and their units can be incrementally trained and tuned by optimizing a metric of interest (including the human-relevant ones). Efficiency is obtained thanks to this layer-wise training strategy, which also enables an adaptive decomposition of the learning task in sub-tasks. Exploiting this idea in GNNs will also allow their self-design. Online learning enables one to learn a model on-the-fly using a possibly-infinite data stream (which can be easily enriched with adversarial streams to increase robustness) dealing also with concept drift (e.g., societal shift in fairness definitions). Strong time and memory constraints are present in this problem that can be faced via dynamic feature selection and model approximation techniques.

Hybrid approaches deal with the challenge of cross-contaminating the approaches just described. While in most applications there is a clear winner in the zoo of all the ML models (e.g., convolution on image recognition) when it comes to dealing with graphs it is unclear whether kernels or GNNs will provide the best solution. This motivates the effort of trying to take the best from the different approaches to derive new methods that can outperform the existing ones. It is then required to combine the flexibility and efficiency of GNNs with the probabilistic formulation of Bayesian networks. This will produce novel graph ML methods that can model the generation of latent causal factors in the structure by approximating intractable probability distributions via neural modules. Another possibility is to generate kernels from the representation learned by probabilistic models and GNNs, or vice-versa.

The second challenge, namely to learn trustworthily is exacerbated and demand for specific solutions when dealing with graphs.

First there is a need to focus on ensuring that ML models do not discriminate subgroups in the populations (e.g., based on gender, ethnicity, and political/sexual orientations). While many approaches have been proposed for vector-based ML, a unifying framework that summarizes and generalizes them is still missing. Moreover, their extension to the graph domain is largely unexplored. In particular, two fundamental questions must be addressed: (i) how to design notions of fairness that also account for the relations among entities and (ii) how to

impose them via pre-, in-, and post-processing methods. For (i) fairness metrics for graphs should be able to capture unfair behaviour which is not simply direct (e.g., discriminate someone based on its political orientation) but also indirect (e.g., favour someone based on connection with powerful people). These metrics can be then used for optimization and assessment of fairness requirements. For (ii) pre-processing methods require modification of data, which in the graph domain demands for the development of suitable methods to learn fair representations from discrete and complex structures, removing sensitive information via data unbiasing. In-processing methods require to further constrain the learning process with non-linear, non-convex, and non-differentiable constraints that need to be approximated or relaxed to be efficiently imposed during the learning phase (of both shallow or layered approaches learning fair representations) while maintaining their cognitive meaning. Finally, post-processing methods are less affected by the input structure as they simply require to trick the model outcome (e.g., via histogram matching), thus allowing one to easily exploit classical approaches even on graphs.

Understanding and improving adversarial robustness of ML algorithms trained on graphs poses three main issues: (i) a systematization of the threats that may affect these algorithms; (ii) a proper methodology and evaluation framework to assess their adversarial robustness under the envisioned threat models; and (iii) suitable countermeasures to mitigate the impact of adversarial attacks as well as the impact of natural concept drift that may occur during operation of such systems. For (i), there is a need for planning to extend existing threat models, including test-time (evasion) and training-time (poisoning) attacks, to graphs and discrete structures by enumerating feasible and practical data manipulation techniques (e.g., node injection or removal in graphs). For (ii), there is a need for using the identified perturbation models and threats to propose a systematic evaluation framework. This framework will provide a definition of adversarial robustness depending on the given perturbation model, and protocols and algorithms to evaluate it, either via empirical attack simulations or theoretical worst-case analyses. For (iii), there is a need for working on countermeasures to mitigate the impact of the envisioned attacks by leveraging two main research directions. The first countermeasure that can be developed is game-theoretical to model the interaction between the learning process and the considered attacks. The second can be to develop techniques to detect and reject samples that are out of the training distribution and cannot be thus classified with sufficient reliability from the ML algorithm.

In order to ensure that ML models preserve the privacy of the individuals while learning actionable information from graph data it is possible to picture two main scenarios: (i) when data must/need to be centralized and (ii) when data can/must be kept distributed. In scenario (i) data or outcomes of an ML model need to be corrupted with noise to keep the individual observations private while learning useful and actionable information. This process is more challenging in graphs where information is scattered in entities, attributes, and relations among entities. Hence, it is required to define novel privacy notions and noise injection methods able to preserve the privacy of the individuals both directly (e.g., differential attacks to the features of a single node in the graph) and indirectly (e.g., differential attacks to the structure of the graph). In scenario (ii) federating the learning process to guarantee privacy using also cooperative game-theoretic approaches where nodes are actors aiming to learn an optimal model can be a good solution. However, an honest-but-curious curator (server)

may still infer clients' information by examining their contributions to learning. To overcome this issue, it is possible to complement federated learning with secure aggregation. Moreover, by adding further synchronization/encryption steps, privacy can be guaranteed also against an active adversary. Finally, hybrid scenarios where data are partially centralized and partially distributed will also be investigated.

Finally there is a need for developing explainability techniques customized for graph domains. To this end, it is required to consider two main research directions: (i) to introduce notions/definitions of explainability suited to graphs, and (ii) to make black-box models more explainable by adapting and developing novel methods to graphs. For (i), the problem is how to define explainability to understand how attributes, entities and relationships contribute to the decisions of the model. For (ii), it is required to adopt different strategies, based on either local model/outcome explanation or global inspection methods. The first strategy is to define ML models whose inherent architectural properties ease the inspection of the model adapting current explainability methods (e.g., relying on an attention mechanism to automatically identify only the most relevant substructures). The second strategy is to leverage ML models which intrinsically exhibit model outcome explainability properties (e.g., exploit self-organizing maps, or their probabilistic counterparts, to identify relationships between different but functionally similar patterns, or using Bayesian methods to identify the latent causal factors in the data). The third strategy aims to design new explainable graph generation algorithms, using iterative algorithms whose decisions are controlled by explainable GNNs.

The third challenge, Learning Automatically and with Guarantees, requires the solution of fundamental theoretical and practical problems which become even more challenging when dealing with graphs.

First it is required to focus on automated learning of architectures and hyperparameters. No-free-lunch theorems state that there is no way of building an ML algorithm able to perform better than others on a reasonably large range of applications. For this reason, tuning the performance of an ML algorithm, i.e., finding the optimal values of the hyperparameters or the right architecture for an ML model, is mandatory to reach satisfying accuracies. This problem, even if crucial, has very few theoretically-grounded solutions and researchers still largely rely on grid, random, or gradient-(free)-based search. In GNNs or multiple kernel learning, the number of hyperparameters or architecture configurations is so huge that gradient-based methods are becoming the only viable choice. However, the gradient may be available for some hyperparameters but not for others, requiring the design of efficient gradient-free and hybrid optimization methods. Another promising research direction is to exploit cooperative transferable-utility mathematical games where an objective (expressed, e.g., in terms of multiple metrics to optimize) and an utility (expressed, e.g., in terms of hyperparameters or architectural configurations) can be transferred from one player to another without any loss. Moreover, multiple objectives (e.g., accuracy, fairness, robustness, explainability) need to be simultaneously optimized to meet the previously mentioned trustworthy requirements. This makes the problem even more challenging as it requires constructing the corresponding Pareto Frontier to identify the best compromise solutions. It is thus crucial to consider that the architecture and hyperparameters of new ML models for graphs will need to be tuned with respect to both technical (e.g., accuracy and efficiency) and human-related metrics (e.g., trust-related metrics).

For what concerns the problem of learning with guarantees it is required to deliver theoretical analysis in terms of approximation, computational, statistical, and worst-case behaviour. For what concerns the computational and approximation capabilities, studying the topological characteristics of input–output functions is probably the most promising research direction. Another research direction focuses on networks for graphs extending two existing theories (i.e., Weisfeiler-Lehman test and unfolding equivalence), which separately provide only few suggestions about the relationship between the characteristics and the capabilities of the networks. In this way, it is possible to gain insights on how to overcome the limits of current architectures (e.g., many GNNs are not universal approximators) and how to design new ones. The choice of a suitable GNN architecture can be investigated by introducing an application-specific probability distribution on a set of classification functions and then estimating the expected correlation between functions sampled from this distribution and network input-output mappings. The effects of graph size (e.g., number of nodes or connections) on these correlations can be analysed by exploiting geometrical properties of high-dimensional spaces. ML model sparsity with respect to different measures can be estimated by introducing norms defined in the spaces of input-output mappings computable by a given ML model and expressed via convex hulls of sets of the ML sub-modules (e.g., graph convolutional layers). Statistical performance can be estimated via both asymptotic and finite sample bounds on the ML models generalization ability. Doing it for complex ML architectures is challenging as it requires rethinking generalization (e.g., overfitting does not imply poor generalization in deep networks). The problem is exacerbated when dealing with multiple technical and human-relevant metrics (e.g., accuracy, efficiency, privacy, fairness, explainability, and robustness) which suffer from poor statistical and mathematical properties (they are non-smooth, non-linear, and ill-defined). Moreover, some of these metrics have proven to be incompatible with each other and therefore, suitable approaches able to relax and combine these metrics while maintaining their compatibility for deriving asymptotic and finite sample bounds need to be designed. Studying the worst-case behavior is also a challenging task for which effective solutions exist only for small networks. The first problem is to find suitable abstractions for networks working on graphs. Abstraction-refinement has proven to be effective for software verification and, as the complexity of ML models increases, over approximation by computable abstraction is promising. Second problem is dealing with properties whose semantics are not given in terms of convex and compact data sets. Non-convex and non-compact data sets can be handled as unions of convex sets, but research is needed to make these representations practical. Finally, scalability is a crucial problem in worst-case methods and it is possible to leverage on algorithmic (e.g., abstraction) and computational (e.g., parallelization) approaches to make graph ML models amenable to verification.

The three challenges identified above are deeply interconnected and offer ideas for contamination and unification. However, these challenges are currently mostly faced independently because of their complexity. Consequently it is required to put them under a common framework in the effort of delivering a holistic approach to the problem of learning trustworthily, automatically, and with guarantees from graph data. Given a specific problem equipped with (i) a dataset, (ii) technical requirements (e.g., minimum satisfying accuracy, no centralized data collection, limited computational requirements), and (iii) human-relevant metrics (e.g., no discrimination against needy people or ensuring right of

explanation), it is required to automatically build an ML model able to address the problem by leveraging the information present in the database, automatically selecting the best technical solution (e.g., noise injection, encryption, or federation for the privacy requirements), the best architecture (e.g., Bayesian or convolutional networks or kernels), the best hyperparameters (e.g., number of units or kernel hyperparameters), and guaranteeing the final performance both in terms of technical and human-relevant metrics.

4 The contributions of the ESANN special session

A total of four studies were accepted in the special session.

In *The Benefits of Adversarial Defense in Generalisation* [49], authors observed how recent researches have shown that models induced by ML, in particular by deep learning, can be easily fooled by an adversary who carefully crafts imperceptible, at least from the human perspective, or physically plausible modifications of the input data. This discovery gave birth to a new field of research, the adversarial ML, where new methods of attacks and defense are developed continuously, mimicking what is happening from a long time in cybersecurity. In this paper authors have shown that the drawbacks of inducing models from data less prone to be misled actually provides some benefits when it comes to assessing their generalisation abilities.

In *Boundary-Based Fairness Constraints in Decision Trees and Random Forests* [50], authors recall how popular Decision Trees and Random Forests are for practitioners in order to solve real-world problems. However, Decision Trees may sometimes learn rules that treat different groups of people unfairly, by paying attention to sensitive features like for example gender, age, income, language, etc. Even if several solutions have been proposed to reduce the unfairness for different ML algorithms, few of them apply to Decision Trees. This work aims to transpose a successful state-of-the-art method to reduce the unfairness in boundary based ML models [51] to Decision Trees.

In *Robust Malware Classification via Deep Graph Networks on Call Graph Topologies* [52], authors propose a malware classification system that is shown to be robust to some common intraprocedural obfuscation techniques. Indeed, by training the Contextual Graph Markov Model on the call graph representation of a program, authors classify it using only topological information, which is unaffected by such obfuscations. In particular, authors show that the structure of the call graph is sufficient to achieve good accuracy on a multi-class classification benchmark.

In *Slope: A First-order Approach for Measuring Gradient Obfuscation* [53], authors observe how evaluating adversarial robustness is a challenging problem. Many defences have been shown to provide a false sense of security by unintentionally obfuscating gradients, hindering the optimisation process of gradient-based attacks. Such defences have been subsequently shown to fail against adaptive attacks crafted to circumvent gradient obfuscation. In this work, authors present Slope, a metric that detects obfuscated gradients by comparing the expected and the actual increase of the attack loss after one iteration. Authors show that their metric can detect the presence of obfuscated gradients in many documented cases, providing a useful debugging tool towards improving adversarial robustness evaluations.

References

[1] G. Da San Martino and A. Sperduti. Mining structured data. *IEEE Computational Intelligence Magazine*, 5(1):42–49, 2010.

[2] D. Bacciu, F. Errica, A. Micheli, and M. Podda. A gentle introduction to deep learning for graphs. *Neural Networks*, 129:203–221, 2020.

[3] N. M. Kriege, F. D. Johansson, and C. Morris. A survey on graph kernels. *Applied Network Science*, 5(1):1–42, 2020.

[4] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2009.

[5] A. Micheli. Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, 20(3):498–511, 2009.

[6] C. Gallicchio and A. Micheli. Fast and deep graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2020.

[7] M. Podda, D. Bacciu, and A. Micheli. A deep generative model for fragment-based molecule generation. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[8] D. Bacciu, F. Errica, and A. Micheli. Contextual graph markov model: A deep and generative approach to graph processing. In *International Conference on Machine Learning*, 2018.

[9] D. Bacciu, F. Errica, and A. Micheli. Probabilistic learning on graphs via contextual architectures. *Journal of Machine Learning Research*, 21(134):1–39, 2020.

[10] C. Gallicchio, A. Micheli, and L. Pedrelli. Design of deep echo state networks. *Neural Networks*, 108:33–47, 2018.

[11] Q. Li, Z. Han, and X. M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018.

[12] M. Qu, Y. Bengio, and J. Tang. GMNN: Graph markov neural networks. In *International conference on machine learning*, 2019.

[13] D. Bacciu, A. Micheli, and A. Sperduti. Generative kernels for tree-structured data. *IEEE transactions on neural networks and learning systems*, 29(10):4932–4946, 2018.

[14] F. Aiolli, M. Donini, N. Navarin, and A. Sperduti. Multiple graph-kernel learning. In *IEEE Symposium Series on Computational Intelligence*, 2015.

[15] N. Navarin, D. Van Tran, and A. Sperduti. Learning kernel-based embeddings in graph neural networks. In *European Conference on Artificial Intelligence*, 2020.

[16] L. Oneto, N. Navarin, A. Sperduti, and D. Anguita. Multilayer graph node kernels: Stacking while maintaining convexity. *Neural Processing Letters*, 48(2):649–667, 2018.

[17] A. F. Winfield, K. Michael, J. Pitt, and V. Evers. Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]. *Proceedings of the IEEE*, 107(3):509–517, 2019.

[18] L. Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6):261–262, 2019.

[19] L. Oneto and S. Chiappa. Fairness in machine learning. In *Recent Trends in Learning From Data*, 2020.

[20] L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Neural Information Processing Systems*, 2020.

[21] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[22] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[23] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[24] T. Graepel, K. Lauter, and M. Naehrig. Ml confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*, 2012.

[25] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys*, 51(5):1–42, 2018.

[26] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Neural information processing systems*, 2019.

[27] A. Bose and W. Hamilton. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*, 2019.

[28] D. Zügner and S. Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[29] C. Meng, S. Rambhatla, and Y. Liu. Cross-node federated graph neural network for spatio-temporal data modeling. *arXiv preprint arXiv:2106.05223*, 2021.

[30] D. Franco, N. Navarin, M. Donini, D. Anguita, and L. Oneto. Deep fair models for complex data: Graphs labeling and explainable face recognition. *Neurocomputing*, In Press, 2021.

[31] L. Oneto, M. Donini, M. Pontil, and J. Shawe-Taylor. Randomized learning and generalization of fair and private classifiers: From pac-bayes to stability and differential privacy. *Neurocomputing*, 416:231–243, 2020.

[32] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. C. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[33] L. Oneto. *Model Selection and Error Estimation in a Nutshell*. Springer, 2020.

[34] X. He, K. Zhao, and X. Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

[35] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, 2015.

[36] J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[37] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[38] F. Scarselli, M. Gori, A. G. Tsoi, M. Hagenbuchner, and G. Monfardini. Computational capabilities of graph neural networks. *IEEE Transactions on Neural Networks*, 20(1):81–102, 2009.

[39] V. Garg, S. Jegelka, and T. Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, 2020.

[40] M. Bianchini and F. Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

[41] F. Leofante, N. Narodytska, L. Pulina, and A. Tacchella. Automated verification of neural networks: Advances, challenges and perspectives. *arXiv preprint arXiv:1805.09938*, 2018.

[42] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.

[43] L. Oneto, N. Navarin, M. Donini, S. Ridella, A. Sperduti, F. Aiolli, and D. Anguita. Learning with kernels: a local rademacher complexity-based analysis with application to graph kernels. *IEEE transactions on neural networks and learning systems*, 29(10):4660–4671, 2017.

[44] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*, 2018.

[45] D. Yin, R. Kannan, and P. Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, 2019.

[46] J. Li, V. Nagarajan, G. Plumb, and A. Talwalkar. A learning theoretic perspective on local explainability. *arXiv preprint arXiv:2011.01205*, 2020.

[47] L. Oneto, S. Ridella, and D. Anguita. Differential privacy and generalization: Sharper bounds with applications. *Pattern Recognition Letters*, 89:31–38, 2017.

[48] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[49] L. Oneto, S. Ridella, and D. Anguita. The benefits of adversarial defence in generalisation. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2021.

[50] G. Nanfack, V. Delchevalerie, and B. Frénay. Boundary-based fairness constraints in decision trees and random forests. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2021.

[51] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[52] F. Errica, G. Iadarola, F. Martinelli, F. Mercaldo, and A. Micheli. Robust malware classification via deep graph networks on call graph topologies. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2021.

[53] M. Pintor, L. Demetrio, G. Manca, B. Biggio, and F. Roli. Slope: A first-order approach for measuring gradient obfuscation. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2021.