

Explainable Graph Neural Networks for Fraud Detection in Financial Transaction Networks

Dominik Watson

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
dominik.watson212@missouri.edu

Paul D. Fields

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.
pdfields@uab.edu

Vishal R. Russell

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.
contactvishal@oregonstate.edu

Abstract

Financial transaction networks constitute large-scale, dynamic ecosystems in which illicit actors continuously evolve their strategies to circumvent detection. Graph neural networks have emerged as powerful instruments for modeling the relational structure inherent in these payment topologies, enabling the capture of subtle patterns of fraudulent behavior that elude traditional feature-based classifiers. However, the high-stakes nature of financial crime mitigation demands not only predictive accuracy but also transparent reasoning that can be audited, contested, and aligned with regulatory frameworks. This paper presents a comprehensive examination of explainable graph neural network architectures for fraud detection. We discuss the architectural trade-offs between local and global explanation paradigms, the integration of post-hoc interpreters with inherently interpretable message-passing schemes, and the implications for fairness, robustness, and adversarial resilience. System-level considerations including deployment infrastructure, latency constraints, data provenance, and long-term model sustainability are analyzed through the lens of real-world financial compliance environments. The discussion extends to governance and policy, examining how explainability serves as a bridge between automated detection pipelines and human-centered accountability under regulations such as anti-money laundering directives. We argue that purely post-hoc explanation modules, while convenient, introduce fragility and potential misalignment with the decision logic of the underlying model, and propose a layered architectural strategy that combines intrinsic model sparsity, attention-based signal tracing, and recourse-aware counterfactual generation. The paper concludes by outlining open challenges in standardizing explanation fidelity metrics, maintaining fairness across demographic and behavioral subgroups, and designing collaborative oversight mechanisms that involve regulators, financial institutions, and technology providers.

Keywords

graph neural networks, fraud detection, explainability, financial transaction networks, anti-money laundering, fairness, governance 1 Introduction The digitization of global financial

infrastructures has produced transactional environments of extraordinary scale and complexity. Payment processors, correspondent banking networks, and real-time settlement systems collectively process billions of transactions daily, each represented by parties connected through intricate webs of ownership, geography, and commercial relationship. Within these networks, fraudulent activities including money laundering, identity theft, and synthetic identity fraud operate as dynamic adversarial processes that continuously adapt to evade rule-based and traditional machine learning defenses. The relational nature of financial transactions, where the legitimacy of one entity's behavior cannot be determined in isolation but depends on its structural position and interaction patterns with others, has motivated the adoption of graph neural networks as a foundational analytical technology. Graph neural networks perform iterative information aggregation across neighborhoods defined by transaction edges, enabling the representation of nodes (accounts, merchants, institutions) to encode both their own attributes and the attributes of the wider network context in which they are embedded. This inductive capacity for relational reasoning has yielded substantial improvements in the detection of previously unseen fraud patterns, especially those involving coordinated groups of accounts whose individual transactional profiles appear innocuous when examined in isolation [1, 2]. Despite these performance gains, the operational deployment of graph neural networks in fraud detection confronts a critical tension: the very mechanisms that produce high-fidelity representations through multiple layers of non-linear neighborhood aggregation also render the models opaque to human understanding. In regulated financial environments, automated decisions to freeze accounts, decline transactions, or file suspicious activity reports are subject to stringent audit requirements, external challenge, and—increasingly—explicit mandates for explanation under frameworks such as the General Data Protection Regulation and evolving anti-money laundering directives. Explainability in this context is not merely a desirable property but a functional requirement for institutional trust, regulatory compliance, and the protection of individual rights. The challenge of providing faithful, understandable, and actionable explanations for graph-level predictions is compounded by the heterogeneous nature of financial graphs, which contain multiple node and edge types, temporal dynamics, and varying levels of data quality and completeness. Consequently, the design of explainable graph neural network architectures must navigate a complex landscape of technical, operational, and regulatory constraints that intersect across multiple organizational and societal levels. This paper engages with the system-level dimensions of explainable graph neural networks for financial fraud detection, moving beyond isolated model-level metrics to consider the integrated socio-technical infrastructure in which these models function. We examine the structural trade-offs between inherently interpretable computation graphs and post-hoc explanation surrogates, the implications of explanation granularity for human decision-making, and the downstream effects on fairness and robustness when explanations are used as instruments of institutional oversight. Our analysis draws on a broad corpus of research spanning graph representation learning, interpretable machine learning, financial crime analytics, and technology policy. Throughout, we emphasize the importance of multi-disciplinary alignment: technical architectures for explanation must co-evolve with audit practices, regulatory standards, and the procedural logics of financial compliance units if they are to achieve meaningful adoption.

2 Financial Transaction Networks as Relational Learning Domains

Financial transaction networks depart from canonical graph learning benchmarks in several fundamental respects that directly shape the requirements and design of explanation mechanisms. Transactions are timestamped events connecting payer and payee entities, often accompanied by narrative fields, amounts, currency codes, and intermediary routing information. The resulting graphs

are inherently heterogeneous, spanning multiple account types (retail, corporate, nostro/vostro), financial instruments, and jurisdictional boundaries. Temporal dynamics introduce strong non-stationarity, as both legitimate and fraudulent behavioral patterns evolve in response to seasonal commercial cycles, policy changes, and adversarial adaptation. Furthermore, class imbalance is extreme: the overwhelming majority of transactions are legitimate, with fraudulent instances representing a minute fraction that must nevertheless be identified with high recall to avoid regulatory penalties and reputational damage. These structural properties make graph neural networks particularly suitable, as message-passing architectures can propagate weak signals from known fraudulent nodes to neighboring unlabeled nodes, effectively leveraging the homophily and heterophily patterns characteristic of financial crime [3, 4]. Early adoptions of graph convolutional networks in anti-money laundering demonstrated that aggregating information from a target account’s immediate transaction neighborhood could substantially improve recall over logistic regression and random forest baselines that relied solely on account-level features [5]. Subsequent architectures introduced attention mechanisms to differentially weight the influence of neighboring transactions based on their relevance to the prediction task, enabling models to focus on suspicious patterns such as rapid fund layering through intermediary accounts [6]. Heterogeneous graph transformers extended this capability by learning type-specific projections for different node and edge categories, allowing a single model to jointly reason over retail customers, corporate entities, financial institutions, and the various transaction types that connect them [7]. The deployment of such architectures in production settings, however, has revealed interpretability gaps that limit their operational utility. A risk analyst presented with a high fraud probability score for a specific account needs to understand not only which features contributed to that score but which counterparties, transaction sequences, and structural motifs in the account’s neighborhood were most influential, and whether those influences reflect genuine criminal behavior or data artifacts.

3 Explainability Requirements in Financial Fraud Detection

Explainability in financial fraud detection operates at the intersection of technical model transparency, operational decision support, and legal accountability. Unlike many consumer-facing machine learning applications where explanations serve a persuasive or user-experience function, explanations in anti-fraud systems must withstand adversarial scrutiny: they may be examined by internal compliance officers, external auditors, law enforcement agencies, and defense counsel in legal proceedings. This environment imposes stringent requirements on explanation fidelity, which refers to the extent to which an explanation accurately reflects the true reasoning process of the model, as opposed to providing a plausible but post-hoc rationalization that may diverge from the model’s actual computation [8]. Fidelity is particularly difficult to guarantee in graph neural networks because predictions emerge from complex interactions between node features, graph topology, and the aggregation depth, creating a combinatorially large space of potential explanatory factors. The temporal dimension of transaction networks adds further complexity to explainability. Fraudulent schemes such as trade-based money laundering or structuring operate across extended time windows, and a prediction made at a current timestamp may integrate signals from transactions that occurred months earlier [9]. Explanations must therefore be capable of tracing influence backward through both the graph structure and the temporal sequence, highlighting specific historical events or dormant accounts whose reactivation contributed to the alert. Moreover, the operational tempo of financial crime units demands explanations that can be generated and consumed under time constraints. A system that requires minutes of computation to explain a single transaction alert is incompatible with real-time payment screening environments where decisions must be made within milliseconds.

This creates a direct trade-off between explanation depth and latency that must be addressed at the architectural level, for instance by pre-computing local explanation bases or maintaining incrementally updated explanation states for high-risk subgraphs. Finally, explainability must be evaluated not only by technical metrics such as fidelity and sparsity but also by its impact on human decision-making in the context of financial investigations. Studies in human-computer interaction and cognitive psychology have shown that explanations that are overly complex, highly technical, or poorly matched to the investigator's mental model can paradoxically degrade decision quality by inducing over-trust or confusion [10]. Designing explanation interfaces for fraud analysts requires a socio-technical perspective that integrates understanding of institutional workflows, regulatory reporting formats, and the collaborative nature of investigation teams that often include both data scientists and domain experts with decades of financial crime experience.

4 Architectural Strategies for Explainable Graph Neural Networks

The provision of explanations for graph neural network predictions can be approached through two broad architectural philosophies: post-hoc interpretation applied to pre-trained black-box models, and inherently interpretable architectures that constrain the computation graph to produce explanations as a natural byproduct of inference. Post-hoc methods, such as perturbation-based explainers and gradient-based attribution techniques, offer the advantage of model-agnostic application, allowing financial institutions to retrofit explanation capabilities onto existing high-performance models without retraining [11]. In the graph domain, GNNExplainer learns a subgraph and a feature mask that maximize the mutual information between the extracted structure and the model's prediction, producing localized explanations that highlight the most influential neighborhood nodes and edges [12]. While such methods have been widely adopted in research settings, their fidelity in production financial networks remains an open concern: the optimization process underlying the explanation can converge to local minima that reflect spurious correlations, and small perturbations to the input graph can produce substantially different explanations for the same prediction, undermining the consistency that auditors require. Inherently interpretable architectures pursue a different design philosophy, embedding explanation generation into the forward pass of the network itself. Attention-based graph neural networks represent a prominent example, as the learned attention coefficients provide a natural mechanism for tracing the relative influence of neighboring nodes on a target node's representation [13]. When applied to transaction graphs, attention weights can be visualized as edge importance scores, allowing analysts to inspect which counterparties and transaction types the model deemed most relevant to its fraud assessment. However, the relationship between attention weights and feature importance is not straightforward: high attention may indicate either genuine predictive relevance or a form of redundancy where the model attends to uninformative nodes to suppress their influence, making raw attention distributions an unreliable explanatory signal without additional calibration [14]. To address this, recent work has proposed attention regularization schemes that promote sparsity and align attention patterns with semantically meaningful substructures derived from domain knowledge, such as known money laundering typologies [15]. A middle-ground architectural strategy gaining traction is the layered integration of local and global explanation modules. In such designs, the primary graph neural network is trained with a multi-objective loss that includes an auxiliary explanation fidelity term, encouraging the model to maintain representations from which faithful local explanations can be efficiently derived. Simultaneously, a global explanation component, often implemented as a concept bottleneck layer or a prototype-based reasoning module, provides case-level explanations that reference prototypical fraud patterns learned across the entire network [16]. This hybrid architecture

acknowledges that different stakeholders require explanations at different granularities: a transaction monitoring analyst needs an account-specific explanation to decide whether to escalate an alert, while a model risk manager needs a global understanding of the patterns the model has learned to assess its ongoing suitability and detect concept drift.

5 Trade-Offs Between Performance, Interpretability, and Robustness

The introduction of explainability mechanisms into graph neural network architectures inevitably introduces trade-offs with predictive performance and adversarial robustness. Architectures that achieve high interpretability through constrained computation graphs, such as those limiting message-passing depth or enforcing monotonic feature interactions, may sacrifice the representational capacity needed to capture rare and complex fraud patterns that span long transaction chains. This trade-off is particularly consequential in anti-money laundering, where detection of sophisticated layering schemes often requires aggregating information across five or more hops in the transaction graph, a depth that simpler interpretable models struggle to accommodate [17]. Financial institutions must therefore make context-dependent choices about where on this spectrum to operate, balancing regulatory explainability requirements against the substantial financial and reputational costs of missed detections. The robustness of explanations themselves has emerged as a critical system-level concern. Adversarial actors with knowledge of the explanation mechanism can theoretically craft transaction patterns that not only evade detection but also produce misleading explanations that direct investigator attention away from the true fraudulent activity. This form of explanation-based adversarial attack has been demonstrated in domains such as credit scoring and criminal recidivism prediction, and the interconnected nature of financial transaction graphs amplifies the attack surface: a single compromised account can be used to inject carefully structured transactions that perturb the local explanations for numerous legitimate accounts, potentially triggering wasteful investigations while masking genuine fraud elsewhere [18]. Defending against such attacks requires explanation mechanisms that are themselves subject to robustness verification, perhaps through randomized smoothing or certified radius techniques adapted from adversarial robustness literature to the specific topology of explanation faithfulness. Relatedly, fairness considerations intersect with explainability in subtle but important ways. Graph neural networks trained on historical transaction data may learn to associate certain geographic regions, nationalities, or business types with elevated fraud risk, reflecting historical biases in enforcement patterns rather than true criminal propensity. When explanations surface these biased associations—showing, for instance, that transactions routed through particular jurisdictions systematically receive higher attention weights—the institution faces not only a fairness problem but a transparency-mediated accountability challenge [19]. Explainability thus functions as a double-edged sword: it can expose discriminatory patterns that would otherwise remain hidden in the model’s numerical outputs, but also creates a documentary record that may be subject to legal discovery. Proactive fairness auditing of explanation patterns across demographic and behavioral subgroups is therefore essential, and methods for generating counterfactual explanations that are fair with respect to protected attributes are an active research frontier.

6 Governance, Infrastructure, and Deployment Realities

Deploying explainable graph neural networks in live financial environments requires attention to a constellation of infrastructure and governance challenges that extend well beyond model architecture. Financial transaction data are distributed across multiple processing systems, often segregated by product line, jurisdiction, and data residency requirements. Building the unified graph representation needed for effective graph neural network inference demands substantial investment in data engineering pipelines, entity resolution systems that link accounts across disparate identifiers, and streaming architectures

capable of updating graph state with low latency [20]. The explanation subsystem must be integrated into this same data fabric, retrieving the contextual information—transaction narratives, customer due diligence profiles, historical alert dispositions—that give explanations their operational meaning for investigators. Governance frameworks for explainable fraud detection systems must define clear ownership and accountability for model behavior across its lifecycle. This includes specifying the conditions under which explanation-based overrides of model decisions are permitted, the processes for escalating disagreements between human analysts and automated recommendations, and the evidentiary standards that explanations must meet when included in suspicious activity reports submitted to financial intelligence units [21]. Many jurisdictions are moving toward requiring that regulated entities maintain comprehensive model documentation, including explanation methodology and validation results, as part of their model risk management programs aligned with supervisory guidance on artificial intelligence. The technical complexity of graph neural network explanations poses particular challenges for these documentation requirements, as the relevant concepts span multiple specialized domains and may be difficult to communicate effectively to supervisory examiners whose expertise is primarily in banking regulation rather than machine learning. From a sustainability perspective, the computational cost of generating explanations at the scale of global transaction networks is non-trivial. Running a perturbation-based explainer on every alert produced by a real-time fraud detection system can multiply the inference compute budget by an order of magnitude, challenging the cost efficiency and environmental footprint of the overall pipeline [22]. This has motivated research into amortized explanation generation, where a separate explanation network is trained to directly output explanations given the same input graph, effectively distilling the explanation process into a single forward pass. Such amortized approaches, while reducing latency and compute, introduce additional model complexity and the risk that the explanation network itself drifts relative to the primary predictor, requiring coordinated monitoring and retraining schedules. The long-term sustainability of explainable graph neural network deployments thus depends on institutional commitments to ongoing infrastructure investment, cross-training between engineering and compliance teams, and the development of shared evaluation benchmarks that allow financial institutions to compare explanation quality in a standardized manner.

7 Policy Implications and Regulatory Alignment

The regulatory landscape surrounding algorithmic decision-making in finance is in a period of rapid evolution, with explainability provisions becoming increasingly codified in binding instruments rather than merely aspirational guidance. The European Union’s proposed Artificial Intelligence Act classifies certain financial services applications as high-risk, triggering requirements for transparency, human oversight, and technical documentation that directly implicate the design of explanation systems for fraud detection models [23]. Banking supervisors in multiple jurisdictions have issued expectations that institutions using advanced analytics for anti-money laundering must be able to demonstrate to examiners how model outputs are derived and how they relate to the underlying risk indicators that form the legal basis for suspicious activity reporting. Explainable graph neural networks, if properly designed and validated, can serve as a technical foundation for meeting these obligations, but only if the explanation outputs are aligned with the specific evidentiary and procedural standards that regulators apply. One promising direction for regulatory alignment is the development of standardized explanation fidelity metrics that are accepted by both the technical community and supervisory authorities. Current explanation evaluation largely relies on proxy measures such as the drop in prediction confidence when explained features are removed, which may not correspond to the semantic correctness required in legal contexts. Collaborative efforts

involving central banks, financial intelligence units, and academic researchers to define explanation quality benchmarks that reflect investigative utility, rather than purely mathematical convenience, would represent a significant step toward harmonization [24]. Such benchmarks could be incorporated into regulatory sandboxes, allowing institutions to test explainable graph neural network systems under realistic conditions while providing feedback to both model developers and policymakers. The global nature of financial transaction networks necessitates consideration of cross-border regulatory compatibility. An explanation generated for a suspicious transaction report filed in one jurisdiction may be reviewed by authorities in another jurisdiction with different legal standards for evidence and privacy. Designing explanation architectures that can be configured to comply with multiple regulatory regimes—for instance by supporting configurable levels of detail and different formats for presenting explanation evidence—adds engineering complexity but is essential for institutions operating across borders. This international dimension also raises questions about the governance of shared explainability infrastructure, such as consortium-based graph databases and federated learning environments where multiple banks collaboratively train fraud detection models without centralizing sensitive customer data [25]. In such settings, the explanation mechanism itself must respect data-sharing restrictions, potentially requiring federated explanation protocols that can generate coherent cross-institutional accounts of model behavior without exposing individual transaction details.

8 Conclusion

Explainable graph neural networks represent a critical frontier in the application of relational machine learning to financial fraud detection, situated at the confluence of technical innovation, regulatory demand, and institutional accountability. The unique structural properties of financial transaction networks—their heterogeneity, temporal dynamics, extreme class imbalance, and adversarial operating environment—shape explanation requirements in ways that extend beyond the standard paradigms of interpretable machine learning. We have argued that effective explainability in this domain must be understood as a system-level property, emerging from the interaction of model architecture, explanation interface design, data infrastructure, institutional governance processes, and regulatory frameworks. The path forward requires progress on multiple fronts. From an architectural perspective, the development of inherently interpretable graph neural network designs that achieve competitive performance without sacrificing fidelity remains an open and important challenge. The trade-offs between local and global explanations, post-hoc and intrinsic methods, and human-centered versus model-centered evaluation must be navigated with careful attention to the specific decision contexts in which explanations will be used. Infrastructure investments in streaming graph processing, amortized explanation generation, and cross-institutional collaborative platforms will determine the feasibility of deploying these systems at the scale required by global financial networks. Governance structures must evolve to define clear accountabilities, establish validated benchmarks for explanation quality, and create feedback loops between investigative outcomes and model improvement. In parallel, policymakers and regulators must continue to refine the normative frameworks that give explainability its teeth, ensuring that technical progress serves the broader goals of financial integrity, consumer protection, and the rule of law.

References

1. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems* (pp. 1024–1034).

2. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.
3. Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020* (pp. 2704–2710). ACM.
4. Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., & Yu, P. S. (2019). Heterogeneous graph attention network. In *The World Wide Web Conference* (pp. 2022–2032). ACM.
5. Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. In *KDD Workshop on Anomaly Detection in Finance*.
6. Liu, Z., Chen, C., Li, L., Zhou, J., Li, X., Song, L., & Qi, Y. (2019). GeniePath: Graph neural networks with adaptive receptive paths. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 4424–4431.
7. Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (pp. 315–324).
8. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
9. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
10. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
11. Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems* (pp. 9240–9251).
12. Yuan, H., Tang, J., Hu, X., & Ji, S. (2020). XGNN: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 430–438).
13. Jain, A., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3543–3556).
14. Wiegrefe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 11–20).
15. Luo, D., Cheng, W., Yu, W., Zong, B., Ni, J., Chen, H., & Zhang, X. (2020). Learning to drop: Robust graph neural network via topological denoising. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 779–787).
16. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems* (pp. 8928–8939).

17. Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? In International Conference on Learning Representations.
18. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
19. Dai, E., & Wang, S. (2021). Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 680–688).
20. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). TensorFlow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (pp. 265–283).
21. Arner, D. W., Barberis, J., & Buckley, R. P. (2017). FinTech, RegTech, and the reconceptualization of financial regulation. *Northwestern Journal of International Law & Business*, 37(3), 371–413.
22. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.
23. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.
24. Financial Action Task Force. (2021). Opportunities and challenges of new technologies for AML/CFT. FATF, Paris.
25. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.