Value is in the Eye of the Beholder: A Framework for an Equitable Graph Data Evaluation

Francesco Paolo Nerini*
DIAG, Sapienza University of Rome
Rome, Italy
f.paolo.nerini@centai.eu

Paolo Bajardi
CENTAI Institute
Turin, Italy
paolo.bajardi@centai.eu

André Panisson
CENTAI Institute
Turin, Italy
panisson@centai.eu

ABSTRACT

Proprietary data is a valuable asset used to develop predictive algorithms that benefit a wide range of users, including customers, business owners, and decision-makers. Consequently, there is a growing interest in developing safe and robust techniques for sharing, learning models, and distributing predictions across a wide spectrum of potential stakeholders. However, a structured process to assess the value of data assets, and thus enabling collaborations among stakeholders, remains largely unexplored. This is particularly challenging when the data to be shared has a networked structure, where increasing the shared data samples potentially connects information observed by different data owners, providing new knowledge that is unavailable to any data owner individually. Here, we propose *E-GraDE*, a framework that assists organizations in assessing the value of their networked data to better address graph machine learning tasks. This framework includes a step-by-step analysis of the requirements of different stakeholders, such as the accuracy or fairness requisites of the models, ensuring a fair evaluation process and stronger alignment in the development of a data federation consortium. Additionally, we propose an approach to estimate the value of networked data to be shared while disclosing only a small fraction of the original information. We support our approach with extensive computational experiments, analysing each part of it through simulated use cases.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → **Database performance evaluation**; *Data exchange*; *Network data models.*

KEYWORDS

Dataset Evaluation, Graph Datasets, Graph Neural Networks, Shapley Values

ACM Reference Format:
Francesco Paolo Nerini, Paolo Bajardi, and André Panisson. 2024. Value is in the Eye of the Beholder: A Framework for an Equitable Graph Data Evaluation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3630106.3658919

*Also with CENTAI Institute.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658919

1 INTRODUCTION

In the modern digital age, proprietary data has emerged as a fundamental asset for organizations, providing the raw material for developing powerful predictive algorithms. These algorithms can serve a wide range of users, from end customers who benefit from personalized offerings to business owners who gain actionable insights for strategic decision-making, and policy-makers who rely on data-driven evidence for policy development. Given the broad spectrum of potential beneficiaries, there is an increasing interest in creating safe, robust, and efficient methodologies for sharing data, learning models and distributing predictions derived from them.

However, despite the wide recognition of the immense value that these data assets hold, there remains a conspicuous gap in the literature regarding the systematic assessment of this value. This is particularly true when the data to be shared has a networked structure. Networked data, under its interconnected nature, does not merely amplify the training size but rather enhances the depth and breadth of understanding of the problem at hand, by integrating data points from various data owners. Evaluating the value of such data becomes a complex task, further complicated by considerations of data privacy and the potential risk of information leakage.

In sectors like retail and telecommunications, where businesses possess large volumes of consumer behavior data, there are compelling use cases for the sharing of networked data. In such instances, sharing information could lead to better market predictions and customer service innovations. Similarly, sharing clinical data among healthcare providers, researchers, and policymakers, provides invaluable insights into disease patterns and treatment outcomes. In the financial sector, the large-scale effort by regulators and law-enforcement agencies to combat money laundering [32–34] is also a compelling use case. Institutions often access only a portion of the transaction network [2], limiting their ability to detect and prevent financial crimes. Therefore, alongside privacy-aware methodologies and regulatory frameworks, establishing a common standard for evaluating the importance of data assets could be crucial in unlocking the potential of novel public-private partnerships.

In response to these challenges, we introduce *E-GraDE* (Equitable Graph Data Evaluation), a new, practical framework, grounded in game theory to help organizations comprehensively assess the value of their proprietary graph data. Designed as a step-by-step guide to assist organizations navigate through the complex landscape of diverse stakeholder requirements, *E-GraDE* simplifies the task of

understanding and accommodating various stakeholder needs. It also promotes fairness in data valuation and strengthens coordination in a data sharing consortium. This approach represents a significant advancement in data value assessment, especially in the realm of collaborative data sharing.

In addition to the *E-GraDE* framework, we propose a pragmatic approach for estimating the value of networked data intended for sharing, while enhancing data privacy by revealing only a small portion of the original dataset. To demonstrate the feasibility and effectiveness of our approach, we corroborate it with extensive experiments. Our results highlight the capability of our methodology to address the critical need for data value assessment in today's data-rich world, paving the way for more equitable and efficient data sharing practices.

Notation. Let's consider a graph structured dataset as a simple, unweighted, undirected, attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ with a set of nodes $\mathcal{V}$, a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and its associated node attribute matrix $\mathcal{X} \in \mathbb{R}^{|\mathcal{V}| \times}$. That is, for each $\mathcal{v} \in \mathcal{V}$, there is an attribute vector $\mathcal{x}_{\mathcal{v}}$ associated with $\mathcal{v}$. Here we consider a node classification task, where each node $\mathcal{v} \in \mathcal{V}$ is associated to a class $\mathcal{y}_{\mathcal{v}} \in \mathcal{Y}$. Let $\mathcal{A} \mapsto \mathcal{Y}$ be a learned model, which takes as input a training graph dataset $\mathcal{T} \subseteq \mathcal{G}$ and assigns to each node a class. The predictive performance of $\mathcal{A}$ is evaluated on a transductive test set of nodes $\mathcal{E} \subset \mathcal{V}$, which implies that the features and the edges associated to the test nodes are available during training, but their labels are not.

We assume that the graph $\mathcal{G}$ and its associated features $\mathcal{X}$ have been formed as the union of different subgraphs, i.e. $\mathcal{G} = \bigcup_{\mathcal{v}=1}^{\mathcal{N}} \mathcal{G}_{\mathcal{v}}$. Each subgraph corresponds to data belonging to a different data owner. They may potentially overlap, but they mostly possess different information regarding the same data (e.g. one may have information regarding certain neighbours of a node, while another may have access to its label). Throughout the experiments that we run in this paper, stakeholders are artificially sampled from different common graph datasets.

Problem statement. Our most general problem can be stated as follows: given a finite set of data owners $\mathcal{N}$ and the union of their datasets $\mathcal{G} = \bigcup_{\mathcal{v}=1}^{\mathcal{N}} \mathcal{G}_{\mathcal{v}}$, each data owner with information regarding a subset of $\mathcal{G}$, we want to assign to each subset $\mathcal{G}_{\mathcal{v}}$ a value $\mathcal{q}_{\mathcal{v}}$. This value quantifies the contribution of the single subset $\mathcal{G}_{\mathcal{v}}$ to the performance of the learned model $\mathcal{A}$ concerning the test set $\mathcal{E}$.

The problem is strongly context-dependent by its nature. Depending on the learning algorithm used to train the model and the metrics used to measure performance, the value attributed to a certain data owner may change (i.e. the overall contribution it brings to improving the performance of the model with respect to the test set $\mathcal{E}$).

Contributions. We introduce the *E-GraDE* framework for evaluating context-specific graph data importance. It starts with the definition of the problem and the requirements specified by the problem owner. A key aspect of *E-GraDE* is its use of a game-theoretical approach, using Shapley values to assess the contribution of each participant. Additionally, *E-GraDE* aims to minimize data sharing within the consortium in the early evaluation phase. Moreover, it strives to conserve computational resources by limiting the amount of data on which the models are trained. This is achieved by (1) extracting Random Spanning Trees (RSTs) from the dataset and (2) calculating Shapley values considering the unshared portions of the datasets. In scenarios where RSTs may reveal excessive information, we propose to further limit the data sharing based on the concept of effective resistance. After ensuring that the model meets the specified requirements, *E-GraDE* provides each data owner with an evaluation of their dataset's contribution to the model's performance. This assessment is not meant to price dataset but to offer an insight into its contribution to the performance of the model, which could potentially be associated with an economic value. While Shapley values have been previously proposed to assess data importance [11], this work seeks to expand their application by evaluating datasets where data points are inherently interconnected, such as those in graph datasets.

E-GraDE structure and processes are justified by a series of experiments performed by evaluating randomly extracted coalitions. First, we tested that the Shapley values are not trivially correlated to any of the graph metrics we measured over the datasets in isolation (see Appendix). This supports the use of Shapley values and a training phase as necessary to assess the correct contribution of the dataset. Then we show how the Shapley values are assigned depending on the particular context of the problem, i.e. the learning algorithm, the testing procedure and the evaluation metrics. Finally, we show that by using RSTs and approximation heuristics it is possible to obtain a reasonable approximation of the Shapley values without performing the training over the full datasets. The code to reproduce these experiments is available at https://github.com/FrappaN/graph-eqev.

2 PRELIMINARIES AND RELATED WORK

In the evolving landscape of data marketplaces, understanding the dynamics of data pricing and trading is becoming increasingly crucial [3, 13, 25]. In [45], the authors propose a formalization of existing technological solutions that might inform existing approaches to data governance in data collaborative initiatives. Rasouli and Jordan [27] provides an in-depth analysis of data markets for distributed machine learning, considering the unique aspects of data and presenting models for bilateral and unilateral exchanges.

Data pricing models aim to assign a fair and reasonable value to data assets, taking into account various factors such as data quality, granularity, and attribute relevance. Yang et al. [36] introduce a pricing model that assesses the quality of data using dimensional indicators. On the other hand, Shen et al. [28] propose a pricing model for personal data centred on tuple granularity. The model applies positive rating and reverse pricing methods by considering value-influencing data attributes like information entropy, weight value, data reference index, and cost. Agarwal et al. [1] introduce a marketplace model for the efficient exchange of training data for machine learning tasks.

Addressing the need to limit information sharing during data asset valuation, Xu et al. [35] propose a data appraisal stage that eliminates the need for full data sharing between data owners and buyers in data markets. Azcoitia et al. [4] propose a "Try Before You Buy" (TBYB) method. The TBYB method provides data buyers

in Data Marketplaces with a measure of machine learning performance on individual datasets, enabling them to make almost optimal purchase decisions without full information on all possible dataset subsets.

When it comes to sharing graph-structured data for graph machine learning tasks, most research focuses on federated learning methodologies to limit information exchange between partners. Existing studies, such as those summarized in [10], fall into various categories. Our work specifically addresses the category marked by cross-client missing information, a common situation in structured data. In these cases, each data owner owns a subgraph of the overall graph, and some nodes may have neighbours owned by other data owners. Examples of these methodologies include FedGraph [7], FedNI [26], PPSGCN [38], and FedSage+ [39], which primarily focus on healthcare scenarios. These scenarios range from using distributed healthcare data (e.g., patient interactions like co-staying in a hospital room or co-diagnosis of a disease) to building a powerful, generalizable model across multiple distributed subgraphs. As privacy regulations are increasing worldwide [29], these methodologies address more properly privacy risks. The survey of Zhang et al. [41] analyses in detail how different approaches can guarantee privacy, security, and robustness. The effectiveness of these approaches is often tested using popular benchmark datasets, such as Cora, Citeseer and PubMed.

2.1 Shapley values

Shapley values are a general solution concept to the problem of allocating gains in the context of cooperative games. We define a cooperative game by the tuple (S, v), where $S = \{1, \ldots, n\}$ is a (finite) set of n players and $v : 2^n \to \mathbb{R}$ is a coalition function: it assigns to each possible subset of the n players a real value. This value corresponds to the performance of a given subset of players when they are the only ones to play. The problem is to find an allocation $\{\phi_1, \ldots, \phi_n\}$ of the gains of the whole coalition, depending on the contribution of each player. Shapley values are considered to be an "equitable" allocation as it can be proven that they satisfy a set of desirable properties [19]. This problem is distinct from the problem of computing Shapley values for graph-restricted games (see, e.g., [30]), as coordination between players is not defined by a graph; instead, the players themselves are defined by graph datasets.

The Shapley value of player i can be written as $\phi_i(S, v) = \frac{1}{n!} \sum_{D \subseteq S \setminus \{i\}} |D|!\,(n - |D| - 1)!\,\big(v(D \cup \{i\}) - v(D)\big)$, where the sum is over all subsets D of players not containing player i. It is interpreted as a weighted sum over all possible "marginal contributions" of player i, with weights corresponding to the number of subsets of size $|D|$ not containing i. Each player obtains a "fair payment" for their work, which sums up with the others to the total win of the team, by checking the contribution the player gives to any possible sub-coalition. It can also be easily generalized to give value not only to single players but also to any sub-coalition of them ([12]).

The use of Shapley values is gaining traction as a novel approach to quantify the value of individual data points in a data economy [24]. Ghorbani and Zou [11] propose using data Shapley values in supervised machine learning to provide an equitable valuation for individual data points, even in contexts involving large datasets and complex learning algorithms. Both studies highlight the diverse utility of Shapley values in ensuring fair data valuation and underscore the complexity of achieving this in various contexts.

Moreover, computation of Shapley values in a federated learning scenario has also been proposed [20, 31, 42], in order to preserve users' privacy during data sharing.

2.2 Value of Data

Here, we tackle the problem of equitable data valuation [11, 17, 24] with the Shapley values paradigm, where each player corresponds to a dataset, and the coalition function v corresponds to an evaluation of a model after being trained over a subset of datasets. When considering datasets, each player includes also the test set (X_i, y_i). The coalition function v will depend on the trained model A, and on the metric κ used to evaluate the performance of the model. We can formally define the game as:

Game 1 (Data Evaluation)
Players: the set of datasets S (including their test sets)
Coalition function $: v_{\kappa, A}$, depending on the metric κ and on the training of a ML model A over the merged dataset of the coalition.

Existing literature shows how Shapley values assignment performs well in satisfying what we would expect from valuable data [5, 11], especially with respect to other metrics such as the size of a dataset or simpler solutions as the leave-one-out score. It can be easily applied both to single data points and to datasets.

Thanks to the flexibility of the definition of $v_{\kappa, A}$, different Shapley values can be assigned in different contexts to the same datasets. It must be always remembered that they carry meaning only inside a specific game, and changing the game (e.g. by considering a different set of players or a different scoring function) will also change the Shapley values.

Throughout the work, we will sometimes drop the subscript in $v_{\kappa, A}$, taking for granted the dependence of the coalition function on the chosen metric and trained model.

3 E-GRADE: A FRAMEWORK FOR EQUITABLE EVALUATION FOR NETWORKED DATA

The E-GraDE framework is defined to evaluate data within the context of a consortium or collaboration, focusing on three distinct types of users, each with unique interests in data evaluation:

(1) Problem owner: the user who has a given problem to solve. She believes that learning a model on a larger dataset could improve the performance of the corresponding algorithmic decision support system. She might have a set of constraints over the solutions (e.g. fairness or model transparency concerns) which must be evaluated.
(2) Consortium/Partnership: the collaboration itself which is formed to share individual datasets to train machine learning models. Its main concerns are related to the minimization of data sharing in the preliminary operations, the possibility of including a third party to perform independently the data evaluation, and satisfying the fairness and transparency requirements from the problem owner or a regulator for the final model.

(3) Data owner: the actual owner of the dataset. Her objective when evaluating the dataset is to maximize the value assigned to her data asset and minimize as much as possible the shared data.

These stakeholders, as identi ed, have objectives that may overlap, adjoin, or even con ict. E-GraDE provides a structured, step-by-step procedure to clearly identify and de ne the data evaluation process upfront.

As a rst step (see Figure 1) the problem owner de nes the requirements that the nal model trained on the consortium data needs to comply with. There are three main requirements that we show can impact directly the Shapley values (as shown from experiments in section 5.1): the model transparency, its generalizability and the evaluation metric, which can re ect fairness requirements. These aspects de ne di erent coalition functions, by directly a ecting the training and testing procedures. It is possible to consider the di erent alternatives as di erent games, where the same player can provide varying degrees of contribution. In this step, it is also possible to give full voice to the problem owner (and regulators) restraints over the model and its expected performance. It is therefore crucial for all parties to properly de ne the procedure in order to give value to the correct characteristic.

While the problem owner sets the requirements, data owners can share their data with the consortium, or at least a limited fraction of them. There are many sampling methodologies de ned over graphs in the literature of network science and computer science. In the proposed framework we show that sampling edges and nodes using the properties of random spanning trees produce the best results, as shown experimentally in section 5.2.

Ideally, the samples should be independent of the chosen model and evaluation metrics to ensure they accurately represent the original dataset, including its aws and strengths. Data sharing could either involve a trusted third party to perform data evaluation or be executed by the data owners in a distributed fashion. They will then assess each Shapley value on their own, providing a pooled estimate to the consortium. As outlined in 2 section, data sharing and the computation of Shapley values can also be performed in a secure and privacy-preserving setting within federated learning environments. Since E-GraDE does not make any assumption on the techniques adopted to compute the values or the learning framework, but rather focuses on training models with representative samples, it could be adopted and complement existing privacy-preserving methods.

If the data sharing involves a third party, estimates of the Shapley value can only rely on the shared samples. On the contrary, if each data owner trusts the other or the sharing procedure, it is possible to obtain better estimates, as illustrated in section 3.2 and in the experiment in section 5.2. Such estimations evaluate the contribution of Shapley values of full samples together with the sub-samples of the others. In this way, each data owner can obtain a more faithful estimation of the Shapley value of their own dataset.

After the estimation, the problem owner and data owner can raise legitimate concerns regarding both the model requirements and the data sharing. If the model does not satisfy the target requirements in terms of fairness, transparency or accuracy, the problem owner can propose to rede ne the requirements or to ask to share more data to obtain better performances. Similarly, the data owners could object that the chosen requirements are disadvantaging their dataset evaluation, or that by sharing more data they could obtain a better estimate as well. Hence, the three steps are re-iterated until all the stakeholders reach an agreement, or the collaboration is dismissed entirely.

If all the stakeholders reach an agreement, they can then proceed to full data sharing and a last ne-tuning of the model. It is then possible to evaluate fully and exactly the Shapley value of the whole dataset. Especially after changing the model, the nal value may di er from the original estimates.

3.1 Graph sampling strategies for partial data sharing

To accommodate data owners who want an initial estimate of their data's value before fully committing to a collaboration, E-GraDE includes an e ective sampling strategy. Here, we assume a fair and collaborative relationship between the data owners with no adversarial behaviour. We show experimentally how to select a smaller subgraph from the data to obtain the best possible estimate. We propose an enhanced approach that goes beyond standard methods of graph sampling, which often rely on random walks or uniformly sampling the edge list. This approach leverages the properties of uniform random spanning trees for more accurate and e ective results.

Uniform random spanning trees have some useful theoretical properties which justify their use in this context, especially their connection with commute time distance and the e ective resistance $' \, ^1\!4^{\circ}$ of an edge. In an unweighted graph, $^1\!4^{\circ}$ is equal to the probability of an edge being sampled in a uniform random spanning tree. We provide a brief overview of these concepts in the Appendix. For detailed mathematical explanations and proofs, please refer to [21].

Random spanning trees maintain the graph's full node set while revealing the minimum amount of information about its structure. However, when there is a need to control the amount of disclosed information, we suggest selecting edges and nodes based on the effective resistance. Speci cally, we prioritize edges with the highest e ective resistance, as they are more likely to appear in a random spanning tree. Our experiments, detailed in Section 5.3, demonstrate that both sampling random spanning trees or selecting the edges with the highest e ective resistance are e ective methods for choosing subgraphs. This nding supports the use of this sampling method in the E-GraDE framework.

Choosing nodes and edges based on other metrics (such as a centrality measure) may also be e ective. However, we found that Random Spanning Trees generally yield better results when compared to other sampling techniques we tested. If our assumption of homophily does not apply in the graph, alternative sampling strategies may be necessary.

3.2 Estimating Shapley values with partial data

As discussed in the previous section, sharing the entire data asset all at once to measure its value might be a barrier for data owners who are starting to collaborate. Therefore, we focus on the partial sharing of the dataset, where the aim is to obtain the best approximation of the Shapley values for the whole dataset. In what follows, we

Figure 1: General E-GraDE framework owchart. Di erent stakeholders (highlighted in di erent colours) are engaged in de ning di erent aspects of the data-sharing process that subsequently impacts the estimate of the Shapley values.

show how our proposed framework handles this task, maximizing the use of information depending on the sharing choices. In the following considerations, we assume that the partners have agreed on a modelling strategy, speci cally on the type of machine learning model to be trained. We keep it xed, along with the testing procedures (hence we omit the subscript for the coalition function $+_{<\bullet A}$ except in the game de nitions).

Let's consider the case where each data owner owns a dataset G_8, but they share only a fraction G_8^0 of their whole datasets. We refer to the data shared by each data owner as $S^0 = fG_8^0 \bullet 8 = 1\bullet \text{""""} \bullet g$. Each shared graph $G_8^0 = {}^1V_8^0 \bullet E_8^0 \bullet X_8^{0\flat}$ contains the shared nodes $V_8^0 \quad V_8$, the shared edges $E_8^0 \quad E_8$, and all the attributes corresponding to the shared nodes. The shared graphs are obtained by rst sampling a subset of nodes and edges, using one of the di erent techniques we considered, and then by adding a corresponding test set to the graph, such that it contains all the nodes of the test set $()_8 \quad V_8^0$) and the edges connecting them to the rest of the shared nodes (i.e. the set of edges $f4 = {}^1E_1 \bullet E_2{}^o \; j \; E_1 \; 2 \;)_8 \wedge \; E_2 \; 2 \; V_8^0 \; n)_8 g$ E_8^0).

We proceed by training the model using only the shared data. By doing so, we can estimate the Shapley values of the data shared within the collaboration, in a game similar to Game 1 but with a di erent set of players:

Game 2 (Centralised Estimation of q_8).
Players: the set of shared datasets S^0 (including their test sets).
Coalition function $: +_{<\bullet A}$ as de ned for the Data Evaluation game (Game 1).

The Shapley values of the players inside this game will then be:

$$q_8^{0\,1}+^o = \frac{1}{=} \tilde{O}_{D^0 \; S^0 n G_8^0} \frac{+{}^1D^0[fG_8^0 g^o \quad +{}^1D^{0\flat}}{\frac{=\;1}{jD^0j}}\text{"} \qquad (1)$$

In this approach, we assign Shapley values q_8^0 taking into account only the data that has been shared. We use these values as an approximation of the complete dataset's Shapley values q_8, assuming that q_8^0 is proportional to q_8. This assumption is based on the fact that the coalition function $+$ remains the same, and there is a direct match between each player in the Data Evaluation game (involving the whole dataset) and each player in the Centralised Estimation game (involving a shared fraction of the dataset). We refer to this as a centralised estimation because the dataset samples can be shared with a third party, who then calculates the estimated Shapley values. Alternatively, we explore a decentralized scenario in which data owners share their dataset samples directly with each other. In this setting, each owner can train on their full dataset as well as on the shared portions from others. This allows any owner, say data owner 8, to calculate the Shapley value of their complete dataset within a game that includes shared samples from the other data owners:

Game 3 (Decentralised Estimation of q_8).
Players: the full dataset G_8 and the set of shared datasets $S^0 = fG_8^0 \bullet 8 = 1\bullet \text{""""} \bullet 8 \; 1\bullet 8, \; 1\bullet \text{""""} \bullet g$ (including their test sets).
Coalition function $: +_{<\bullet A}$ as de ned for the Data Evaluation game (Game 1).

The Shapley value of player 8 in this game, $\hat{q}_8$, is then:

$$\hat{q}_8{}^{1}+^o = \frac{1}{=} \tilde{O}_{D^0 \; S^0 n G_8^0} \frac{+{}^1D^0[fG_8 g^o \quad +{}^1D^{0\flat}}{\frac{=\;1}{jD^0j}}\text{"} \qquad (2)$$

We assume that each marginal contribution of player 8 to the subcoalition D can be approximated as the marginal contribution it can give to the corresponding subcoalition of shared datasets $D^0 = fG_9^0 \; j \; G_9 \; 2 \; Dg$, i.e.:

$$+{}^1D\,[fG_8 g^o \quad +{}^1D^o{}' \quad +{}^1D^0[fG_8 g^o \quad +{}^1D^{0\flat}\text{"}$$

Thus, we can use the Shapley value of player 8, $\hat{q}_8$, in their corresponding Decentralised Estimation game as an estimate for its Shapley value q_8 in Game 1.

However, even if we use the full dataset G_8, we are evaluating its contributions in a game without the full datasets of the other data owners. Moreover, we are de ning a di erent game for each dataset, and comparing the Shapley values between di erent games. It is reasonable that the values will be comparable since the coalition function is always the same and the set of players mostly overlaps

between these games (the only di erence being the one player we are interested in computing the Shapley value of, i.e. the full dataset). However, we still lose some of the properties given by the Shapley values when performing their estimation through this strategy.

Therefore, we de ne a third strategy for estimating the Shapley values through a new game with a di erent coalition function, . We start by de ning, as a function of a coalition D and the corresponding shared coalition D', such that:

$$, \ ^1D \bullet D'^o = \frac{1}{jD'j} \tilde{O}_{G' 2 D'} + ^1 D' n fG'g \ [\ fGg^o \ " \qquad (3)$$

Consider a coalition of shared datasets D' (and the corresponding coalition of full datasets D). The function, takes the values given by + when substituting each dataset $G' 2 D'$ with its full version $G 2 D$, one at a time. Finally, it performs a mean between these values of + . In practice, it can be calculated if all data owners rst calculate the coalition function + of their own Game 3, where their own shared dataset is already substituted with their full dataset, and then share this function with the other data owners to compute the values of . Using this coalition function, we can then de ne the following game:

Game 4 (Mean Decentralised Estimation of q_8).
Players: the set of datasets S (including their test sets and their shared datasets S') .
Coalition function : the function, as de ned in equation 3.

In this game, each player is de ned by a triplet of elements: player 8 will be given by $^1G_8 \bullet G'_8 \bullet)_8{}^o$, where G_8 is the whole graph dataset, G'_8 G_8 is the shared counterpart of the graph, and $)_8$ is the test set of nodes.

The Shapley values of player 8 in this game will then be given by:

$$q_8{}^1, \ ^o = \frac{1}{\#} \tilde{O}_{D \ Sn \ 8} \frac{, \ ^1D \ [\ f \ 8g^o \ , \ ^1D^o}{= \ 1 \quad jDj} " \qquad (4)$$

When estimating the Shapley values of the game 1 through the values given by the Mean Decentralised Estimation game, we assume that, given a subcoalition of datasets D and the corresponding shared subcoalition D', , $^1D \bullet D'^o \ ' \ + ^1D^o$.

In all three estimation approaches, the estimate converges to the exact Shapley value when increasing the shared fraction of the datasets to 100%. Each approach accounts for more information in their estimate than the precedent but works on di erent assumptions. A step-by-step description of the estimation approaches is provided in the Appendix.

To test these di erent methods and compare them more quantitatively in a controlled environment, we applied them to a simple and transparent yet quite rich toy model game.

Toy model . Our toy model game is de ned by the following coalition function: given a coalition of three players $x = ^1G_1 \bullet G_2 \bullet G_3{}^o$, with each G_8 corresponding to the contribution given by player 8 (0 if it is absent from the coalition, 1 if it is contributing completely), we de ne the coalition function as:

$$+ ^1x^o = f \ \rangle F \ G_1 \ , \ F \ G_2 \ , \ ^1F \ , \ F \ ^oG_3 \ , \ F \ G_1G_2 \ , \ F \ G_1G_3 \frac{1}{2}$$

Table 1: Agreement between the rankings assigned by the di erent estimation strategies on the toy model game as percentages over all the games shown in Figure 4, obtained by varying F and F .

	CEN-DEC		DEC-MEAN		MEAN-CEN	
?	10%	50%	10%	50%	10%	50%
Both correct	0.9056	0.9658	0.9056	0.9725	0.9070	0.9753
First right	0.0014	0.0108	0.0000	0.0019	0.0065	0.0082
Second right	0.0000	0.0086	0.0079	0.0110	0.0000	0.0012
Both wrong	0.0930	0.0148	0.0865	0.0146	0.0865	0.0153

where the function $f \ ^1 \ ^o$ is a sigmoid function. This was selected as a nonlinear function that maps the coalition score to a range between 0 and 1, as we might expect in a classi cation task. We have three parameters which de ne a family of games with three players, where F corresponds to the maximum contribution that players 1 and 2 can obtain by themselves, F, in uences the contribution that player 3 can obtain alone with respect to the other two players and F modulates the cooperation between player 3 and the other two players, which do not interact between themselves.

By keeping F xed and modifying the other two parameters, we can model a variety of scenarios. In particular, the most important player can switch between player 1 (high F and low F) and player 3 (low F and high F), while player 2 is always the least contributing to the game. Varying these values allows to apply the approximation in di erent regimes, where the cooperation can be more or less important than the stand-alone value of each player.

In this toy model, we represent the partial sharing of the dataset by setting the player's contribution as $G_8 = ?, 0 \ddot{Y} ? \ddot{Y} 1$, instead of $G_8 = 1$. We set the three players' contributions to the same fraction $?$, we $x \ F \ = 0"4$ and vary the value of F and F . We analyze if the two di erent estimates produce di erences in the ranking of players in Table 1. In the toy example, rankings generally agree with one another, while the estimate with the lowest RMSE changes depending on the parameters of the game and the shared fraction $?$ (see Appendix). The mean decentralised estimation was shown to be slightly more accurate in predicting the correct rank, although it is not always the strategy with the lowest error. The best strategy depends on the parameters of the game: however, all three strategies can correctly identify the ranking of values in more than 90% of the cases even at low $?$.

4 EXPERIMENTAL SETTINGS

In our work, we emphasize the importance of using Shapley values for accurate graph dataset evaluation, which is a critical nal step in the E-GraDE framework. This is based on the understanding that dataset size isn't always a reliable indicator of its value, as highlighted in [5]. Similarly, since cooperative game theory provides a principled way to quantify contributions, we advise against relying on estimations based on other graph metrics. We provide experimental validation of this in the Appendix.

In this section, we present the experimental settings we adopted to test and examine the properties of the E-GraDE framework. We introduce the data, the machine learning models and the testing procedures we adopted.

Table 2: Datasets used in the work; a value with an asterisk refers to the largest connected component of the graph.

Dataset	#Nodes	#Edges	#Features	#Classes
Cora[37]	2485*	5069*	1433	7
PubMed[37]	19717	44324	500	3
Pokec-z[8]	67435*	617765*	276	2

4.1 Graph sources and arti cial coalitions

In order to conduct computational experiments on a simulated environment where di erent data owners share their (networked) data, we extract independently and randomly di erent graph datasets from larger graphs. We refer to these sets of datasets as arti cial coalitions . Each arti cial coalition is always extracted together from the same larger graph and the machine learning models are always trained for the task of node classi cation. We do not consider arti cial coalitions with datasets extracted from di erent graphs. We use three graph datasets as larger graphs from which to extract arti cial coalitions (see Table 2).

We aim at creating datasets inside the arti cial coalitions as di erent as possible, while also extracting them randomly. Each dataset inside a coalition is extracted independently from one another. The size (in terms of the number of edges) and the number of known labels of each dataset are uniformly distributed across all coalitions. The extraction technique for a dataset is chosen between four equiprobable methods: uniform edge sampling (in which we conserved only the nodes having at least one edge) and three different biased random walks sampling. The biased random walks which we use are the same as those of Node2Vec paper [14]. In particular, we modi ed the parameter of the walk in order to obtain a "Breadth-First Search-like" walk (= 0.001), a "Depth-First Search-like" walk (= 1000) and an unbiased random walk (= 1). The arti cial coalitions are formed by four datasets in all our experiments.

4.2 Models

In our analysis, we use two di erent models. The rst is a Label Propagation model, as described in [44]. This model predicts labels for all unlabeled nodes, including those in the test set, using only the edges and labels from a training set. Its key advantage is that it doesn't require any training, making it faster to apply across all potential coalitions of stakeholders compared to supervised training algorithms. Moreover, it ensures transparency and interpretability by relying solely on the graph structure and some initial labels for predictions. The second model is a multi-layer Graph Convolutional Network [18]. We tuned the hyperparameters by training over the complete graphs, excluding a validation set for each, and used them only for experiments on datasets extracted from the corresponding source graph.

4.3 Testing strategy

Although there are many metrics to choose from, in this work we consider only two di erent coalition scores for simplicity: accuracy and statistical parity (, 40). Both are measured over the test sets. We explore two distinct strategies for de ning these test sets and evaluate the models. In both cases, we work within a transductive setup.

The rst strategy, referred as shared test set , involves a common test set of nodes shared among all datasets in the coalition. This test set is extracted before the sampling of the arti cial coalition and then added to each of the graphs. This approach has the advantage of testing on a sample from the general, overarching distribution; however, the models may face a training set whose label distribution di ers from that of the general source population.

The second strategy, called individual test set , allows each stakeholder to have its own test set of nodes, which becomes part of its contribution when forming a coalition. These test sets are selected after the arti cial coalitions are extracted, ensuring that training and test sets are formed from the same population. Each model trained on a set is tested on its corresponding test set; a coalition of stakeholders is tested after merging the corresponding test sets (if a node in a test set belongs to the training of another, it is always removed from training when the datasets are merged).

5 E-GRADE APPLICATION IN SIMULATED USE CASES

We nally analyze the application of the framework in our simulated scenarios. By considering the di erent users of the framework we identi ed earlier, we examine how di erent steps impact and satisfy their objectives.

5.1 Problem owners' perspective: how model requirements impact the value of data

The rst step of E-GraDE asks the problem owner to set the requirements that the ML model should satisfy. This in turn allows for a de nition of a speci c cooperative game, as required from the Shapley values de nition. Di erent setting requirements can then lead to di erent value attributions. We identi ed three criteria for the game de nition which are closely related to the requirements of the nal model. The rst is the type of Machine Learning model used that could translate into more transparent or opaque decision systems (e.g. a simple deterministic Label propagation model or a GNN model); the second is the de nition of the test set which impacts the generalizability of the nal model and is tightly related to the context in which it will be deployed (see Section 4.3); lastly, the metric over which the model is evaluated, i.e. assessing the value of data on the accuracy gain or on fairness gain (such as the statistical parity).

We analyze how di erent requirements translate to di erent evaluations in our simulated environment by conducting multiple independent repetitions of arti cial coalition extraction and evaluation. Generally, we observe that the ranking based on the Shapley value of the dataset within a coalition varies signi cantly depending on the di erent modelling choices, as summarized in Table 3. This table includes data from 200 arti cial coalitions of 4 players each extracted from Pokec. Three scenarios are reported: changing the black box model for the evaluation (LP-GNN); evaluating the model through the (test) accuracy or the (test) statistical parity (ACC-SP); training the models on either a shared test set or on individual test sets (SHA-IND). In all three cases, the reference is to the ranking obtained by evaluating the accuracy of a GNN model

Table 3: Percentage of changes in ranking when modifying criteria of the Shapley values evaluation on datasets extracted from Pokec.

LP-GNN	
Same ranking	47.5%
One position swap	33.5%
More than one position swap(s)	19%
ACC-SP	
Same ranking	1.5%
One position swap	5.5%
More than one position swap(s)	93%
SHA-IND	
Same ranking	16%
One position swap	26%
More than one position swap(s)	58%

Table 4: RMSE of the di erent estimation techniques, when sharing subgraphs sampled through uniform (edge) random sampling. The lowest value for each fraction is in bold.

		Fraction (%)				
Dataset	Method	5	25	50	75	100
Cora	Centralised	0.077	0.024	0.013	0.010	0.0
	Decentralised	0.020	0.009	0.006	0.005	0.0
	Mean decentralised	0.016	0.005	0.004	0.003	0.0
PubMed	Centralised	0.031	0.011	0.006	0.005	0.0
	Decentralised	0.012	0.005	0.003	0.002	0.0
	Mean decentralised	0.004	0.002	0.002	0.002	0.0
Pokec	Centralised	0.022	0.010	0.010	0.007	0.0
	Decentralised	0.011	0.006	0.005	0.004	0.0
	Mean decentralised	0.004	0.003	0.003	0.003	0.0

on a shared test set. Rather than identifying an objective value of a dataset, Shapley values allow to give value to a dataset depending exactly on what it is considered important in the context given by the problem owner.

Even in our simpli ed scenario, the assigned values can still have some substantial change, and there is no guarantee a priori that the value will remain similar when changing the evaluation criteria. Thus, as the data does not have an intrinsic universal value and their importance strongly depends on the modelling choices shaped by the context and downstream application of the nal model, the framework is useful to unpack step-by-step the modelling choices that impact the nal data value. We show some of the cases where the changes were the most signi cant for the three conditions which we analyzed (see Figure 2).

5.2 Data owners' perspective: how to sub-sample data to share

In E-GraDE, we address data owners' hesitation to fully share their data without knowing the value they could gain. We recommend strategies for sharing only portions of their data and evaluate different sampling techniques to support this approach. As baseline approaches, we consider uniform edge sampling, where each edge has the same probability of being selected and nodes are included if any of their edges are chosen, and unbiased random walks, which sample the subgraph traversed by a random walker. However, our analysis shows that the best sampling strategy is based on the graph's e ective resistance. Here, we sample the 30% of edges ranked by e ective resistance and their corresponding nodes. We estimate the e ective resistance of each edge by assessing how frequently each appears when extracting random spanning trees from the graph. Lastly, we also compute the Shapley values of samples obtained by extracting one random spanning tree from the dataset, or one for each connected component in the case of a disconnected graph.

In these experiments, we sample 30 di erent arti cial coalitions of four datasets each. We compute their exact Shapley value. Then we proceed to sample increasing fractions of each dataset through di erent sampling techniques and to compute the Shapley value of each sample. For each dataset and method, we perform one sampling per edge fraction (except for random spanning trees, which have a xed percentage of edges). The Shapley values are computed by evaluating the accuracy of a GNN model on a shared test set. To compare the results across coalitions, we normalize each Shapley value, exact or approximate, such that the sum of the values of the corresponding coalition equals 1. We then compute the root mean square error (RMSE) of the estimation at each xed fraction, which is overall the average estimation error over 120 di erent dataset-sample pairs at xed sampling technique and edge fraction. We show the result of adopting di erent samplings, for coalitions extracted from PubMed, in Figure 3a. In the case of the last sampling technique, since each dataset has a random spanning tree of

Figure 2: Examples of how the Shapley values assigned to coalitions extracted from Pokec-z may change by changing the evaluating criteria. The Shapley values of each coalition were normalized such that their sum corresponded to 1.

(a) RMSE for di erent sampling methods.

(b) RMSE for di erent estimating methods.

Figure 3: Application of di erent sampling and estimating strategies for arti cial coalitions extracted from PubMed.

di erent size, we show a kernel density plot of the absolute errors of estimating the Shapley values against the sizes of the trees used (relative to the total number of edges of the corresponding datasets). The Shapley values are estimated through the mean decentralised approach.

As shown in Figure 3a, sampling random spanning trees (and thus sharing the totality of the nodes in the graph) generally allows us to obtain lower errors in estimating the Shapley value of the whole dataset. If we consider only a limited portion of the nodes of the graph, sampling the edges with the highest resistance shows the best result against the two naive sampling techniques.

5.3 Consortium's perspective: how to estimate the value of data

The estimation of Shapley values is the central and last technical step of E-GraDE. This process assumes that the fraction of data each owner shares is a representative sample of their whole dataset. Despite this and other assumptions, we can approximate the values in di erent ways, each requiring di erent levels of collaboration among consortium members. We test the methods proposed in the framework by comparing the quality of each approximation.

We consider the three di erent methods to compute the estimation of the Shapley value as described in Section 3.2. The rst is called centralised estimation and requires that each stakeholder shares a sample of their dataset to a third party, which computes independently the Shapley values of these samples. The second method is called decentralised estimation as each stakeholder receives a sample of the other datasets and can estimate the Shapley value of their own dataset by themselves by leveraging the full information of their own dataset. The third method is called mean decentralised estimation, where each coalition has a coalition score assigned by the function in equation 4. The function performs a mean over the coalition scores given by the decentralised estimation. We tested the methods by using the same strategy as before, measuring the RMSE for 120 samples at xed estimation method and edge fraction.

For a xed sampling technique, each method uses an increasing amount of information to estimate the Shapley value. This, in turn, results in obtaining better estimates, as mean decentralised estimation is shown to be the best method to evaluate the Shapley value in this arti cial setting. This can be seen in Figure 3b, where each sample is extracted through uniform (edge) random sampling. On average, the mean values method allows us to obtain a value closer to the actual Shapley value of the whole sample. This is further proven by looking at the di erent methods on arti cial coalitions extracted from the three graph sources as in Table 4. The mean decentralised estimation proves to greatly reduce the error, especially at low sampled fractions.

6 CONCLUSIONS

In this paper, we introduced E-GraDE, a framework designed to assess the value of context-speci c graph data. Through extensive computational experiments, we show how the framework facilitates a collaborative data sharing process, highlighting the critical requirements that must be addressed before assessing data value. Additionally, we show how E-GraDE provides accurate value estimates while minimizing the amount of shared data, a key aspect that could enhance collaborative agreements. A limitation of E-GraDE is its reliance on honest players with aligned goals. In adversarial settings where one or more players seek to maximize their individual payo , we might observe diverse sharing strategies aimed at maximising the player's Shapley value estimate, rather than minimizing the overall error.

While this paper presents E-GraDE solely in the context of node classi cation tasks, the framework is versatile enough to be adapted to other downstream tasks such as link prediction. Future work will focus on applying E-GraDE to these settings, documenting all the phases and iterations of the decision-making process. We aim to re ne the data valuation process to include the actual business value added through dataset sharing. Additionally, we may explore bidding dynamics that emerge when data value is assessed through increased data sample sharing. We are committed to promoting collaboration among multiple stakeholders to validate the usability of E-GraDE in real-world settings.

ACKNOWLEDGMENTS

This work was partially funded by the Horizon Europe project PRE-ACT, supported by the European Commission through the Horizon Europe Program (G.A. 101057746), the Swiss State Secretariat for Education, Research and Innovation (contract number 22 00058), and the UK government through Innovate UK (application number 10061955).

ETHICS STATEMENT

We do not foresee any evident malicious or unethical consequences that could arise from the utilization of the proposed framework. The only potential area of concern is related to data fusion in contexts where combining sources may pose inappropriate risks for users, such as in the case of personal information. However, we believe that the mitigation of these risks relies on the presence of robust legal regulations regarding data. Establishing and enforcing this kind of legislation is necessary to prevent any unintended consequences stemming from the amalgamation of datasets that are meant to be handled separately. E-GraDE has to be implemented within rigorous legal safeguards and in alignment with independently assessed ethical standards.

Finally, throughout our research, we used publicly available datasets. Any personal information contained within the Pokec dataset was already anonymized in the version we employed. The experiments conducted on these datasets were performed in an entirely arti cial setting, avoiding any ethical consequences from the dataset evaluation.

REFERENCES

[1] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In Proceedings of the 2019 ACM Conference on Economics and Computation (Phoenix, AZ, USA) (EC '19). Association for Computing Machinery, New York, NY, USA, 701 726. https://doi.org/10.1145/3328526.3329589

[2] Kartik Anand, Iman van Lelyveld, Ádám Banai, Soeren Friedrich, Rodney Garratt, Grzegorz Ha aj, Jose Fique, Ib Hansen, Serafín Martínez Jaramillo, Hwayun Lee, José Luis Molina-Borboa, Stefano Nobili, Sriram Rajan, Dilyara Salakhova, Thiago Christiano Silva, Laura Silvestri, and Sergio Rubens Stancato de Souza. 2018. The missing links: A global study on uncovering nancial network structures from partial data. Journal of Financial Stability 35 (2018), 107 119. https://doi.org/10.1016/j.jfs.2017.05.012 Network models, stress testing and other tools for nancial stability monitoring and macroprudential policy design and implementation.

[3] Santiago Andrés Azcoitia and Nikolaos Laoutaris. 2022. A Survey of Data Marketplaces and Their Business Models. SIGMOD Rec. 51, 3 (nov 2022), 18 29. https://doi.org/10.1145/3572751.3572755

[4] Santiago Andrés Azcoitia and Nikolaos Laoutaris. 2022. Try before you buy: a practical data purchasing algorithm for real-world data marketplaces. In Proceedings of the 1st International Workshop on Data Economy (Rome, Italy) (DE '22). Association for Computing Machinery, New York, NY, USA, 27 33. https://doi.org/10.1145/3565011.3569054

[5] Santiago Andrés Azcoitia, Marius Paraschiv, and Nikolaos Laoutaris. 2020. Computing the Relative Value of Spatio-Temporal Data in Wholesale and Retail Data Marketplaces. (2020). https://doi.org/10.48550/arXiv.2002.11193 arXiv:arXiv:2002.11193

[6] N Cesa Bianchi, Claudio Gentile, F Vitale, G Zappella, et al. 2010. Active learning on trees and graphs. In Proceedings of the 23rd Conference on Learning Theory (Haifa, Israel). Omnipress, 320 332.

[7] Fahao Chen, Peng Li, Toshiaki Miyazaki, and Celimuge Wu. 2022. FedGraph: Federated Graph Learning With Intelligent Sampling. IEEE Transactions on Parallel and Distributed Systems 33, 8 (2022), 1775 1786. https://doi.org/10.1109/TPDS.2021.3125565

[8] Enyan Dai and Suhang Wang. 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 (Virtual Event, Israel) (WSDM '21). Association for Computing Machinery, New York, NY, USA, 680 688. https://doi.org/10.1145/3437963.3441752

[9] Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. 2023. Fairness in Graph Mining: A Survey. IEEE Transactions on Knowledge and Data Engineering 35, 10 (2023), 10583 10602. https://doi.org/10.1109/TKDE.2023.3265598

[10] Xingbo Fu, Binchi Zhang, Yushun Dong, Chen Chen, and Jundong Li. 2022. Federated Graph Machine Learning: A Survey of Concepts, Techniques, and Applications. SIGKDD Explor. Newsl. 24, 2 (dec 2022), 32 47. https://doi.org/10.1145/3575637.3575644

[11] Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In Proceedings of the 36th International Conference on Machine Learning (Long Beach, California, USA) (Proceedings of Machine Learning Research, Vol. 97), Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2242 2251. https://proceedings.mlr.press/v97/ghorbani19c.html

[12] Michel Grabisch and Marc Roubens. 1999. An axiomatic approach to the concept of interaction among players in cooperative games. International Journal of Game Theory 28, 4 (01 Nov 1999), 547 565. https://doi.org/10.1007/s001820050125

[13] Ronen Gradwohl and Moshe Tennenholtz. 2022. Pareto-Improving Data-Sharing. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (Seoul, Republic of Korea) (FAccT '22). Association for Computing Machinery, New York, NY, USA, 197 198. https://doi.org/10.1145/3531146.3533085

[14] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16). Association for Computing Machinery, New York, NY, USA, 855 864. https://doi.org/10.1145/2939672.2939754

[15] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems (Long Beach, California, USA) (NIPS 2017, Vol. 30), I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc. https://proceedings.neurips.cc/paper_ les/paper/2017/ le/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

[16] William L Hamilton. 2020. Graph representation learning. Morgan & Claypool Publishers.

[17] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019. Towards E cient Data Valuation Based on the Shapley Value. In Proceedings of the Twenty-Second International Conference on Arti cial Intelligence and Statistics (Valencia, Spain) (Proceedings of Machine Learning Research, Vol. 89), Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 1167 1176. https://proceedings.mlr.press/v89/jia19a.html

[18] Thomas N Kipf and Max Welling. 2016. Semi-supervised classi cation with graph convolutional networks. (2016). https://doi.org/10.48550/arXiv.1609.02907 arXiv:arXiv:1609.02907

[19] Harold William Kuhn and Albert William Tucker (Eds.). 1953. Contributions to the Theory of Games. Annals of Mathematics Studies, Vol. 2. Princeton University Press. 307 317 pages. https://books.google.it/books?id=EWCYDwAAQBAJ

[20] Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. 2022. GTG-Shapley: E cient and Accurate Participant Contribution Evaluation in Federated Learning. ACM Trans. Intell. Syst. Technol. 13, 4, Article 60 (may 2022), 21 pages. https://doi.org/10.1145/3501811

[21] Russell Lyons and Yuval Peres. 2017. Probability on trees and networks. Vol. 42. Cambridge University Press.

[22] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. Annual review of sociology 27, 1 (2001), 415 444.

[23] M. E. J. Newman. 2002. Assortative Mixing in Networks. Phys. Rev. Lett. 89 (Oct 2002), 208701. Issue 20. https://doi.org/10.1103/PhysRevLett.89.208701

[24] Marius Paraschiv and Nikolaos Laoutaris. 2019. Valuating User Data in a Human-Centric Data Economy. (2019). https://doi.org/10.48550/arXiv.1909.01137 arXiv:arXiv:1909.01137

[25] Jian Pei. 2022. A Survey on Data Pricing: From Economics to Data Science. IEEE Transactions on Knowledge and Data Engineering 34, 10 (2022), 4586 4608. https://doi.org/10.1109/TKDE.2020.3045927

[26] Liang Peng, Nan Wang, Nicha Dvornek, Xiaofeng Zhu, and Xiaoxiao Li. 2023. FedNI: Federated Graph Learning With Network Inpainting for Population-Based Disease Prediction. IEEE Transactions on Medical Imaging 42, 7 (2023), 2032 2043. https://doi.org/10.1109/TMI.2022.3188728

[27] Mohammad Rasouli and Michael I Jordan. 2021. Data sharing markets. (2021). https://doi.org/10.48550/arXiv.2107.08630 arXiv:arXiv:2107.08630

[28] Yuncheng Shen, Bing Guo, Yan Shen, Xuliang Duan, Xiangqian Dong, and Hong Zhang. 2016. A pricing model for Big Personal Data. Tsinghua Science and Technology 21, 5 (2016), 482 490. https://doi.org/10.1109/TST.2016.7590317

[29] Pranith Shetty. 2023. Data Privacy and Risk Management, Collaboration is Key on Tackling Privacy Risks/Issues. Journal of Arti cial Intelligence & Cloud Computing 224 (2023), 2 4. https://doi.org/10.47363/JAICC/2023

[30] Oskar Skibski, Tomasz P. Michalak, Talal Rahwan, and Michael Wooldridge. 2014. Algorithms for the shapley and myerson values in graph-restricted games. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (Paris, France) (AAMAS '14). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 197 204.

[31] Tianshu Song, Yongxin Tong, and Shuyue Wei. 2019. Pro t Allocation for Federated Learning. In2019 IEEE International Conference on Big Data (Big Data) (Los Angeles, CA, USA)(IEEE BigData 2019). 2577 2586. https://doi.org/10.1109/BigData47090.2019.9006327

[32] Michele Starnini, Charalampos E. Tsourakakis, Maryam Zamanipour, André Panisson, Walter Allasia, Marco Fornasiero, Laura Li Puma, Valeria Ricci, Silvia Ronchiadin, Angela Ugrinoska, Marco Varetto, and Dario Moncalvo. 2021. Smurf-Based Anti-money Laundering in Time-Evolving Transaction Networks. In Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track(Virtual Event), Yuxiao Dong, Nicolas Kourtellis, Barbara Hammer, and Jose A. Lozano (Eds.). Springer International Publishing, Cham, 171 186.

[33] Toyotaro Suzumura, Yi Zhou, Natahalie Baracaldo, Guangnan Ye, Keith Houck, Ryo Kawahara, Ali Anwar, Lucia Larise Stavarache, Yuji Watanabe, Pablo Loyola, Daniel Klyashtorny, Heiko Ludwig, and Kumar Bhaskaran. 2019. Towards Federated Graph Learning for Collaborative Financial Crimes Detection. (2019). https://doi.org/10.48550/arXiv.1909.12946 arXiv:arXiv:1909.12946

[34] Mark Weber, Jie Chen, Toyotaro Suzumura, Aldo Pareja, Tengfei Ma, Hiroki Kanezashi, Tim Kaler, Charles E. Leiserson, and Tao B. Schardl. 2018. Scalable Graph Learning for Anti-Money Laundering: A First Look. (2018). https://doi.org/10.48550/arXiv.1812.00076 arXiv:arXiv:1812.00076

[35] Xinlei Xu, Awni Hannun, and Laurens Van Der Maaten. 2022. Data Appraisal Without Data Sharing. InProceedings of The 25th International Conference on Arti cial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151), Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, 11422 11437. https://proceedings.mlr.press/v151/xu22e.html

[36] Jian Yang, Chongchong Zhao, and Chunxiao Xing. 2019. Big data market optimization pricing model based on data quality.Complexity2019 (2019), 1 10. https://doi.org/10.1155/2019/5964068

[37] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. 2016. Revisiting Semi-Supervised Learning with Graph Embeddings. InProceedings of The 33rd International Conference on Machine Learning(New York City, New York, USA)(Proceedings of Machine Learning Research, Vol. 48), Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, 40 48. https://proceedings.mlr.press/v48/yanga16.html

[38] Binchi Zhang, Minnan Luo, Shangbin Feng, Ziqi Liu, Jun Zhou, and Qinghua Zheng. 2021. PPSGCN: A Privacy-Preserving Subgraph Sampling Based Distributed GCN Training Method. (2021). https://doi.org/10.48550/arXiv.2110.12906 arXiv:arXiv:2110.12906 [cs.LG]

[39] Ke ZHANG, Carl Yang, Xiaoxiao Li, Lichao Sun, and Siu Ming Yiu. 2021. Subgraph Federated Learning with Missing Neighbor Generation. InAdvances in Neural Information Processing Systems(Virtual Event)(NeurIPS 2021, Vol. 34), M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.). Curran Associates, Inc., 6671 6682. https://proceedings.neurips.cc/paper_ les/paper/2021/ le/34adeb8e3242824038aa65460a47c29e-Paper.pdf

[40] Wenbin Zhang, Shimei Pan, Shuigeng Zhou, Toby Walsh, and Jeremy C. Weiss. 2023. Fairness Amidst Non-IID Graph Data: Current Achievements and Future Directions. (2023). https://doi.org/10.48550/arXiv.2202.07170 arXiv:arXiv:2202.07170

[41] Yifei Zhang, Dun Zeng, Jinglong Luo, Zenglin Xu, and Irwin King. 2023. A Survey of Trustworthy Federated Learning with Perspectives on Security, Robustness and Privacy. InCompanion Proceedings of the ACM Web Conference 2023(Austin, TX, USA,)(WWW '23 Companion). Association for Computing Machinery, New York, NY, USA, 1167 1176. https://doi.org/10.1145/3543873.3587681

[42] Shuyuan Zheng, Yang Cao, and Masatoshi Yoshikawa. 2023. Secure Shapley Value for Cross-Silo Federated Learning.Proc. VLDB Endow.16, 7 (mar 2023), 1657 1670. https://doi.org/10.14778/3587136.3587141

[43] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. 2020. Beyond Homophily in Graph Neural Networks: Current Limitations and E ective Designs. InAdvances in Neural Information Processing Systems(Virtual Event) (NeurIPS 2020, Vol. 33), H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.). Curran Associates, Inc., 7793 7804. https://proceedings.neurips.cc/paper_ les/paper/2020/ le/58ae23d878a47004366189884c2f8440-Paper.pdf

[44] Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from Labeled and Unlabeled Data with Label Propagation.Technical Report. Pittsburgh, PA, USA. CMU-CALD-02-107.

[45] Maciej Krzysztof Zuziak, Onntje Hinrichs, Aizhan Abdrassulova, and Salvatore Rinzivillo. 2023. Data Collaboratives with the Use of Decentralised Learning (FAccT '23). Association for Computing Machinery, New York, NY, USA, 615 625. https://doi.org/10.1145/3593013.3594029

APPENDIX

Uniform Random Spanning Trees

Consider a nite, weighted, undirected and connected graph $G = V_G \bullet E_G$. A spanning tree T is a subgraph of G such that T is connected and acyclic (i.e. it is a tree) and T spans over all the nodes of G, i.e. $V_T = V_G$.

For each graph, there are many possible alternative spanning trees. Throughout this work, we use Wilson's method to sample uniform random spanning trees. If we call by $2^{1}4^{o}$ the weight of an edge 4 and by $\%\rangle T \frac{1}{4}$ the probability of extracting the spanning tree T through Wilson's method, it is possible to show that $\%\rangle T \frac{1}{4} / \prod_{42T} 2^{1}4^{o}$. In other words, the probability of extracting the spanning tree T is proportional to the product of the weights of its edges. When dealing with unweighted graphs, this results in uniform sampling between all possible spanning trees at random.

By interpreting an undirected graph with positive weights as an electrical network, we identify the weight of an edge $2^{1}4^{o}$ as the conductance between two nodes, and thus it is possible to prove that, given an edge 4 and a uniform random spanning tree T:

$$\%\rangle 4 2 T \frac{1}{4} = 2^{1}4^{o}' {}^{1}4^{o}\bullet$$

where $' {}^{1}4^{o}$ is the e ective resistance of the edge. The e ective resistance depends not only on the conductance between the two nodes but on the conductance of all possible paths connecting the two. When the graph is unweighted and therefore $2^{1}4^{o} = 1 \bullet 84 2 E_G$, we have that $' {}^{1}4^{o} = 1$ when the edge 4 is the only connection between two subgraphs otherwise disconnected, while it is lower between two nodes belonging to a common dense subgraph.

The latest observation is particularly relevant in the context of learning on graphs. A general, common assumption which is taken for many applications, including many Graph Neural Networks architectures [15, 16, 43], is homophily [22]. Homophily, in the context of node classi cation, assumes that nodes belonging to the same class will be more likely to be connected. Nodes belonging to the same class will tend to form subgraphs which are more densely connected than the rest. It is therefore likely that edges with a high e ective resistance may divide two nodes belonging to di erent classes. Since a high e ective resistance corresponds to a high probability that the edge belongs to a uniform spanning tree T, it is possible to take advantage of this result for predicting the classes of the nodes. An example of such a method is [6], where spanning trees are integrated along Label Propagation in an active learning theoretical framework.

Step-by-step description of the estimation algorithms

We proceed to describe each of the estimation procedures we proposed in detail. As a general remark, none of these methods address the problem of approximating Shapley values due to computational constraints. While Shapley values are notoriously hard to compute exactly for large number of players, this problem is not relevant in our scenario, where we expect to involve a relatively small number of data owners, typically not exceeding 5-10 players. Since we explore whether the Shapley values found when players contribute only partially (i.e. with limited data sharing) to a game are similar to the Shapley values found when players contribute fully, improving

the computational cost of each "estimation" does not necessarily offer advantages over computing Shapley values exactly. Nonetheless, when required, Shapley value calculations can be approximated using established techniques.

Centralised estimation. The estimation of the Shapley value of a dataset is performed exclusively on the sample shared with the data owner:

(1) Each data owner shares a sample of her dataset to a third party;
(2) The third party computes the Shapley values with respect to a pre-de ned coalition function using exclusively these samples.

The method has the advantage that each data owner does not share her data with one another, assuring an higher level of safety. Howevever, the estimation could be strongly biased by the sample chosen by each data owners.

Decentralised estimation. The estimation of the Shapley value of a dataset is performed privately by each data owner, using the sample of the other data owners.

(1) Each data owner shares a sample of her dataset with the others;
(2) Each data owner computes the Shapley value with her own full dataset and the samples of the other;
(3) Each data owner shares the resulting Shapley value of her own dataset

While this method may compromise some of the data by sharing them with the other data owners, it has two advantages. First, it allows for a data owner to estimate more precisely the value of their full dataset. Second, each data owner could possibly identify more clearly which of the other dataset are important for her interest, by testing individually the di erent samples. Another drawback, however, is that the Shapley values obtained from this approach of own full dataset with another is formally impropriate, as each Shapley value was obtained from a game de ned by a slightly di erent set of playes.

Mean decentralised estimation. This method builds on the previous one, but by de ning a new coalition function, allows for a proper game de nition (see Game 4).

(1) Each data owner shares a sample of her dataset with the others;
(2) Each data owner computes completely the coalition function values with her own full dataset and the sample of the others;
(3) Each data owner shares the coalition function values they obtain;
(4) The new coalition function, is computed, according to Equation 3;
(5) The Shapley values according to the new coalition function are computed.

This latter method requires that the data owners share even more information regarding their datasets, in the form of the performances obtained with the samples of the others. However, it allows to properly de ne a game in which each Shapley value depends on the full dataset, without requiring to share it fully with the other data owners.

Toy Model for Estimating Shapley values with partial data

To test di erent Shapley value estimations, we de ned a toy model game in Section 3.2. Here, we compare the di erent estimates for each game two-by-two: for each couple of estimation methods we compute the ratio between the root mean square error (RMSE) over the three players in Figure 4.

Relation between Shapley values and graph metrics

To check whether the dataset Shapley value (that is potentially intensive to compute and requires some level of data disclosure) is necessary, we analyzed the relation between the Shapley value of graphs in an arti cial coalition and metrics de ned over the graph. Various properties of each dataset were quantitatively measured, in particular: diameter, average clustering coe cient, average degree, and label distance (measured as the sum of the absolute di erences between the frequency of a label in the graph and its frequency on the corresponding graph source), normalized cut size (the number of edges between two nodes belonging to di erent classes), degree-based assortative mixing [23].

In Figure 5, the metrics are presented for arti cial coalitions extracted from Cora. The Shapley values are calculated using the accuracy of the models as coalition score and were then normalized, such that the sum of their values in a coalition is always one. In each subplot, the colour corresponds to the relative size of each dataset in comparison to the full coalition (measured by the number of edges with respect to the number of edges of the entire graph). The extraction method used for each dataset is also represented through di erent symbols for the points. The analysis of the diameter metric does not reveal a clear correlation with the corresponding Shapley values. In the case of other metrics, it seems plausible that there is a relation between the metrics and the corresponding normalized Shapley value. However, there is also a huge variability and noise. Moreover, the properties corresponding to the larger Shapley value appear to be dependent on the graph source: when plotting similar plots for arti cial coalitions extracted from PubMed and Pokec we obtain di erent values corresponding to increasing Shapley values, although the qualitative behaviour is similar.

Based on these observations, we can conclude that predicting Shapley values based on these properties is a complex and context-dependent task, as variations in underlying graph sources, models, and testing procedures can lead to di erent results. Even if it is possible to compute these properties before training a model over the datasets, its contribution will not depend on these metrics always in the same manner.