Curvature and Causal Inference in Network Data

Amirhossein Farzam
Duke University
a.farzam@duke.edu

Allen Tannenbaum
Stony Brook University
arobertan@cs.stonybrook.edu

Guillermo Sapiro
Duke University & Apple
guillermo.sapiro@duke.edu

Abstract

Learning causal mechanisms involving networked units of data is a notoriously challenging task with various applications. Graph Neural Networks (GNNs) have proven to be effective for learning representations that capture complex dependencies between data units. This effectiveness is largely due to the conduciveness of GNNs to tools that characterize the geometry of graphs. The potential of geometric deep learning for GNN-based causal representation learning, however, remains underexplored. This work makes three key contributions to bridge this gap. First, we establish a theoretical connection between graph curvature and causal inference, showing that negative curvatures pose challenges to learning the causal mechanisms underlying network data. Second, based on this theoretical insight, we present empirical results using the Ricci curvature to gauge the error in treatment effect estimates made from representations learned by GNNs. This empirically demonstrates that positive curvature regions yield more accurate results. Lastly, as an example of the potentials unleashed by this newfound connection between geometry and causal inference, we propose a method using Ricci flow to improve the treatment effect estimation on networked data. Our experiments confirm that this method reduces the error in treatment effect estimates by flattening the network, showcasing the utility of geometric methods for enhancing causal representation learning. Our findings open new avenues for leveraging discrete geometry in causal representation learning, offering insights and tools that enhance the performance of GNNs in learning robust structural relationships.

1 Introduction

Generalizability of parameters that characterize observed data to unobserved environments is a core objective of machine learning models. Parameters that are invariant across environments inherently facilitate this generalizability, leading to robust and domain-adaptable models. To this end, uncovering the inherent causal relationships between variables from data —causal representation learning— is a natural approach for developing models with such generalizability [35, 70, 71]. Meanwhile, deep learning has led to notable progress in causal inference [36, 47, 49, 56]. Inferring causal mechanisms from observational data is a fundamental task in various domains, from epidemiology and medicine to social sciences [39, 65, 80, 84]. Due to the endogeneity in the network structure, identifying causal parameters is particularly challenging on a network of units with non-trivial dependencies [79, 86]. Recently, GNN-based causal representation learning has been proposed to account for network-induced endogeneity in structured data [13, 22, 25, 30, 34, 51].

Despite these advances, the full potential of geometric deep learning [3, 5–7, 9, 53] remains underutilized for causal representation learning on networked data. GNNs enable the leveraging of inherent

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

geometry in graph-structured data, characterized, for instance, by discrete curvature on graphs, which has been used to improve GNN-based models [74, 78, 85]. Processes taking place on the graph and, hence, endogeneities rooted in the network, are significantly influenced by its geometry. However, the connection between the geometric properties of the graph, such as its curvature, and causal representation learning remains underexplored, making it an important subject to investigate.

This work aims to bridge geometry and geometric deep learning and causal representation learning by formally establishing the relationship between graph curvature and causal inference on networked data. We explore this connection both through theoretical results pointing to such a relationship, and through theoretically-informed experiments illustrating this connection using a GNN-based causal effect estimation method. Drawing from the theory of invariance and distributional robustness of causal models [8, 52, 61], the central premise of this paper is inspired by the proposition that curvature could serve as a practical measure for robustness in networks [14, 77]. Recent studies have established the connections between curvature, robustness, and entropy [62, 69, 77]. Meanwhile, the relationship between entropy and causal inference has been explored in the context of causal discovery [12]. Collectively, these works and the foundations here developed suggest that graph curvature could offer a powerful tool for enhancing GNN-based causal representation learning.

Main contributions. We present a theoretical layout of causal inference from a distributional robustness perspective, entropic causal inference, and curvature as a robustness indicator, which prepares the ground for establishing the connection between curvature and causal inference. This connection is formally implied from our Theorem 1, which suggests that identification of causal effects becomes more challenging where the curvature is negative. Applying this theoretical finding to causal inference on empirical networks using GNNs, our experiments show that treatment effect estimation error is lower in regions with non-negative curvature, firmly validating our theoretical foundations. Lastly, we propose an adjustment using the Ricci flow to flatten the network, which leads to a remarkable gain in estimating treatment effects on networked data.

2 Causality, invariance, and robustness

2.1 Preliminaries

Consider the causal mechanism involving features X and target Y. Suppose we are interested in evaluating the causal effect of a treatment T on Y for units with features X, which can be measured for each unit i by the individual treatment effect (ITE), or the expected effect conditioned on the features, known as the conditional average treatment effect (CATE). Given features x_i of an individual, the CATE is given by $\tau_i(x_i) := \mathbb{E}\left[Y_i|do(t_i = t, x_i = x) - Y_i|do(t_i = t', x_i = x)\right]$, where $Y_i|do(t_i, x_i)$ is the potential outcome of the unit with features x_i upon intervention by treatment t_i [58]. Following Shalit et al. [72] and Jiang & Sun [30], we adopt a conditional formulation of the ITE as the CATE for the features of an individual unit, and throughout our experiments, we refer to $\tau_i(x_i)$ as the ITE. Since the data is missing the *counterfactual outcome*, $\tau_i(x_i)$ is only a causal quantity and cannot be directly computed as a statistical quantity. Causal effect estimation is essentially estimating causal quantities from statistical quantities. Whether this estimation is possible —the *identification* problem— is the central question of causal inference [57]. Identification of the causal effect from the data is contingent on a set of assumptions. When estimating causal quantities on a network of units, relaxing two assumptions, *ignorability* and *stable unit treatment value assumption (SUTVA)* [28, 67], is likely essential due to peer effects on each unit from its neighbors' features and treatments [30]. Four common assumptions, including ignorability and SUTVA, are formally defined in Appendix A.

2.2 Invariance and distributional robustness of causal parameters

We follow Bühlmann [8] to formalize the derivation of causal quantities from statistical quantities as a worst-case risk minimization problem which leads to learning parameters that are invariant across environments. This results in the generalizability promised by causal representation learning. In this section, we briefly discuss problem formulations and results which explain that the minimization loss function upholding causal assumptions coincides with the one that leads to distributional robustness. A more detailed discussion is included in Appendix B. The discussion in this section formally presents the main claim regarding the distributional robustness of causal parameters. Linear anchor regression is used for the purpose of this formal discussion as an example, and concepts regarding distributional robustness are not limited to the specific settings of this example.

Adopting the notation in Bühlmann [8], let Y^e and X^e denote the random variable and the n_X-dimensional random vector corresponding to an observed environment $e \in \mathcal{E}$, and $\mathcal{F} \supseteq \mathcal{E}$ the union of observed and unobserved environments. Consider the causal mechanism between X and Y described by the structural equation $Y = f(X)$ for some function f. To infer f, we aim to learn a function g from observations $e \in \mathcal{E}$ such that $g(X^e)$ still provides accurate estimates for Y^e when $e \in \mathcal{F} \setminus \mathcal{E}$, under the assumption that e does not directly impact Y^e or change the mechanism between X^e and Y^e. Suppose $g \equiv g_\theta$ is a neural network parameterized by θ. This can be formulated as solving

$$\theta_{\text{causal}} = \underset{\theta}{\operatorname{argmin}} \max_{e \in \mathcal{F}} \mathcal{L}\left(Y^e, g_\theta(X^e)\right), \tag{1}$$

where $\mathcal{L}$ is the loss function [8]. If we can find a subset of covariate indices $S \subset \{1, \ldots, n_X\}$ such that $\mathcal{L}\left(Y^e, g_\theta(X_S^e)\right)$ is invariant with respect to $e \in \mathcal{F}$, to find θ_{causal} it suffices to minimize $\mathcal{L}\left(Y^e, g_\theta(X_S^e)\right)$ for any observable e. This requirement, although abstract, elucidates the invariance of causal estimation. Conditions which allow us to put this on a computational footing, detailed in Appendix B, are contingent on restrictive assumptions such as absence of hidden confounders.

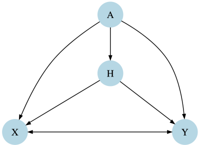

Figure 1: Causal graph for the anchor regression model. A, H, X, and Y denote the anchor, hidden confounders, covariates, and outcome.

To relax this assumption, consider the anchor regression model, involving an anchor variable A, covariates X, outcome Y, and hidden confounders H, with the causal graph shown in Figure 1. The estimand for linear anchor regression can be computed as the minimizer of the loss $\mathcal{L}_A$, given by,

$$\mathcal{L}_A(\gamma) = \mathbb{E}\left[\left(\left(I - P_A\right)\left(Y - X^T b\right)\right)^2\right] + \gamma \mathbb{E}\left[\left(P_A(Y - X^T b)\right)^2\right], \tag{2}$$

where P_A is the projection operator onto the column space of A. The anchor variable could be thought of as the determiner of the environment in Equation 1, thus the second term in Equation 2 encourages the invariance of $\mathcal{L}_A(\gamma)$ with respect to the environment, and can be considered a causal regularization term. Suppose now that we replace the effect of the anchor on Y, modeled as MA in the span of a constant matrix M, with a shift perturbation of the form $v \in \text{span}(M)$ generated by a vector independent of the noise on Y. It can be shown, under conditions described in Appendix B, that $\mathcal{L}_A(\gamma) = \sup_{v \in \mathcal{C}_\gamma} \mathbb{E}\left[\left(Y^v - \left(X^v\right)^T b\right)^2\right]$, where X^v and Y^v correspond to X and Y in the perturbed systems, and $\mathcal{C}_\gamma$ is the class of shift perturbations whose size is typically constrained by $\mathcal{O}(\gamma)$ [8, 64]. This implies that the anchor regression estimand corresponds to worst-case risk minimization in a perturbed system, and simultaneously promotes conditions conforming to the assumptions for causal identification. In other words, an estimand that satisfies the criteria for a causal parameter is also a distributionally robust optimizer, as we formally showcased through the above discussion of anchor regression. This concludes our discussion of the invariance and distributional robustness of causal parameters.

3 Curvature, robustness, and entropy

We now detail why we expect the Ricci curvature to be related to causal inference. Curvature controls how volume balls and geodesics on a Riemannian manifold behave in their local neighborhood [16]. The Ricci curvature indicates how much the local geometry induced by a Riemannian metric deviates from that of a Euclidean space [62]. Extended to discrete structures such as graphs, the graph Ricci curvature characterizes the deviation of the neighborhood of an edge from a grid, capturing the dispersion through the edge in its neighborhood. Ricci-type curvatures have been proven powerful for performing various computational tasks on graph neural networks [15, 45, 74, 78]. A formal definition of Ollivier-Ricci curvature [55], which we use for the experiments in this paper, as well as an alternative Ricci-type curvature, are included in Appendix C.

Curvature and entropy. The following result from optimal transport [46], offers bounds on the Boltzmann entropy in terms of a lower bound on the Ricci curvature,

$$S(\mu_\lambda) \geq (1 - \lambda)\, S(\mu_0) + \lambda S(\mu_1) + \underline{k} \frac{\lambda(1-\lambda)}{2} W_2\left(\mu_0, \mu_1\right)^2, \tag{3}$$

where $S(.)$ denotes the Boltzmann entropy [1], $\underline{k}$ is a lower bound on the Ricci curvature, $W_2\left(\mu_0, \mu_1\right)$ is the Wasserstein distance of order 2 between measures μ_0 and μ_1 in the metric space $(P(\mathcal{X}), W_2)$

of probability measures on $\mathcal{X}$, and μ_λ for $\lambda \in [0, 1]$ gives the geodesic between them [62]. This inequality indicates a positive correlation between Ricci curvature and entropy [62].

Curvature and robustness. Characterized by the fluctuation decay rate [14], *system robustness* refers to the ability of the system to rapidly return to its stationary state after a perturbation. By the Fluctuation Theorem [17], there is a positive correlation between system robustness and entropy, which in turn implies that system robustness is positively correlated with curvature [62]. Despite this correlation and the discussion on distributional robustness and causal inference in Section 2, claiming a connection between Ricci curvature and causal inference is premature at this stage, due to the difference between distributional robustness and system robustness. Thus, we utilize the correlation between entropy and Ricci curvature to formally establish this anticipated connection. To do so, we use the results from entropic causal inference [11], which we briefly review in the following section.

4 Entropic causal inference

Entropic causal inference is a framework that seeks the information-theoretically simplest structural explanation of the data to infer causality, following an Occam's razor-type principle [11, 41]. The central claim is that the true causal structural model is one that yields the minimum entropy in the system [12]. Under a set of assumptions, this principle is shown to facilitate the correct orientation of the edges in the causal graph, following the finding that fitting the wrong model to the data requires a higher entropy than the correct model [11, 12]. More precisely, let $Y = f(X, E)$ be the structural causal model, where $E \perp\!\!\!\perp X$ denotes the exogenous variables. Consider alternative exogenous variables $\tilde{E} \perp\!\!\!\perp Y$ for which the data fit an alternative model $X = g(Y, \tilde{E})$. With high probability,

$$H(X) + H(E) - H(Y) < H(\tilde{E}), \tag{4}$$

provided $H(E)$ is sufficiently small. Considering the correlation between curvature and entropy, this result on structural causal models and the entropies of the corresponding variables points to a connection between curvature and causal inference. We present this connection next.

5 Curvature and causal inference

The results discussed in Section 3 indicate a positive correlation between Boltzmann entropy and Ricci curvature, and in Section 4 we stated a bound on the Shannon entropy of the exogenous variables in an alternative structural causal model, different from the true model. We now show a connection between Ricci curvature and causal inference. This connection will, in turn, inform a methodological remedy utilizing the Ricci flow to improve treatment effect estimates.

Consider the problem of identifying the causal relationship between X_i and Y_i for $i \in \{1, 2\}$, corresponding to two sets of data with the true causal models given by $Y_i = f_i(X_i, E_i)$. Suppose an alternative model $X_i = g(Y_i, \tilde{E}_i)$, with alternative exogenous variables $\tilde{E}_i$ fits the data, and assume that the conditions described in Section 4 leading to Equation 4 are satisfied. Assume that the Ricci curvature corresponding to X_i is bounded below by $\underline{k}_i$, for $i \in \{1, 2\}$. Then, under the assumptions stated in Appendix D, where we provide the proof, the following holds:

Theorem 1. *If $\underline{k}_1 < 0 \leq \underline{k}_2$, there exists a value η, for which $\mathbb{P}\left[H(\tilde{E}_2) > \eta\right] \geq \mathbb{P}\left[H(\tilde{E}_1) > \eta\right]$, i.e., the probability that the Shannon entropy of $\tilde{E}_2$ is lower bounded by η is at least as high as the probability that η is a lower bound for the Shannon entropy of $\tilde{E}_1$.*

Theorem 1 states that if the lower bound on the Ricci curvature is negative for X_1 and non-negative for X_2, then the alternative exogenous variables with which the wrong model fits the data are more likely to have a larger entropy in the case of X_2 than X_1. In other words, for the wrong model to fit the data, we expect a higher entropy of the exogenous variables when the curvature is non-negative.

Theorem 1 formally establishes the connection between Ricci curvature and causal inference. The core insight from Theorem 1 is that a non-negative Ricci curvature corresponds to a wrong fitting model that admits a smaller class of exogenous variables. When the Ricci curvature is non-negative, the entropy of the exogenous variables corresponding to a fitting wrong model tends to have a larger lower bound, hence, a smaller class of exogenous variables could make the wrong model fit the observed data. This enhances distributional robustness, making the worst-case risk minimization in a system under perturbation a less challenging problem. As a result, the regression estimator is identified for a larger class of perturbations. This foundational result on the connection between geometric properties of networks and causal inference paves the way for improving causal effect estimation, as we show next and illustrate further with experiments in Section 7.

5.1 Ricci flow adjustment for improving causal effect estimates

Informed by the theoretical connection between Ricci curvature and causal inference, we propose to improve treatment effect estimates on network data using the discrete Ricci flow [31, 54]. Under the Ricci flow, at time t, the Riemannian metric g evolves as $\frac{\partial g_{ij}(t)}{\partial t} = -2R_{ij}$, where R_{ij} is the Ricci curvature tensor. The Ricci flow evolves to a uniform distribution of curvature [24, 31]. Similarly, its discrete analog leads to a flatter network. Let w_{vu} denote the weight on the edge $(v, u) \in E$, κ_{vu} its Ricci curvature, and $d(v, u)$ the geodesic distance. The discrete Ricci flow at iteration i is given by

$$w_{vu}^{i+1} = (1 - \kappa_{vu})\, d(v, u)^i. \tag{5}$$

In order to improve estimations of treatment effects, we propose modifying the edge weights via the discrete Ricci flow to obtain an adjusted shift operator for the graph convolution, which is the weighted adjacency matrix. This is, in essence, preprocessing the training data through computing a weight matrix by which we multiply the adjacency matrix, and hence, a cost-efficient one-time computation. Since real-world networks are predominantly sparse, this flattening increases the Ricci curvature of the majority of the edges in the network and, therefore, based on our theory, is expected to reduce the error in estimating causal treatment effects.

6 Related work

We establish a novel link between curvature and causal inference, bridging robustness in causal models with geometric deep learning. For a comprehensive review of related work, see Appendix E.

7 Experiments

Building upon these theoretical foundations, we now turn our attention to empirical validation. We employ numerical experiments on real-world network data to demonstrate the practical utility of Ricci curvature for causal effect estimation. Considering the success of neural networks in estimating causal effects, we use a GNN-based framework to estimate treatment effects on the nodes in networked data.

Model and data. When treatments are applied to a network with non-trivial connections, traditional causal effect estimation methods fail due to violation of ignorability or SUTVA [10, 30, 34]. Jiang & Sun [30] proposed *NetEst*, a GNN-based model we use here, which yields identifiable estimates of the treatment effect on networked data in settings where SUTVA is violated due to peer exposure effect. Details of the NetEst model and the ITE formulation are included in Appendices F and G. Our experiments are primarily aimed at demonstrating our theoretical results in practice, and evaluating the performances of NetEst and our proposed enhancement of it. Additionally, we compare our results with multiple baseline methods. These baselines include TARNet [72], a T-Learner [42] with a random forest (RF) regressor, and a T-Learner with a GNN encoder followed by a multilayer perceptron. Further experiments with additional baselines are included in Appendix I.2. To evaluate the performance in estimating the treatment effects, we use the ITE error $\varepsilon_{ITE}(v) := |\tau_v - \hat{\tau}_v|$ and the Precision in Estimation of Heterogeneous Effect (PEHE) $\epsilon_{PEHE}^2 := \sum_{v \in V} (\tau_v - \hat{\tau}_v)^2 / N$, where τ_v and $\hat{\tau}_v$ denote the true and estimated ITEs for node v. Consistent with standard practice in causal representation learning, we use semi-synthetic datasets [22, 27, 30, 50, 72, 81]. Following the original experiments on NetEst, we use the BlogCatalog (BC) and Flickr datasets [22, 50]. We supplement our experiments with numerous other empirical networks, described in Appendix H.

Ricci curvature and treatment effect estimation error. We demonstrate the implications of Theorem 1 by inspecting the joint distribution of ε_{ITE} and the Ricci curvature. To quantify the curvature of the region surrounding a node, we aggregate the Ollivier-Ricci curvature of its incident edges by taking their sum. The joint distributions in Figure 2 show a negative correlation between Ricci curvature and ε_{ITE}, indicating that ITE estimations are less reliable in regions with negative curvature. Additional experiments in Appendix I show that these results are consistent not only across different datasets but also different notions of Ricci curvature.

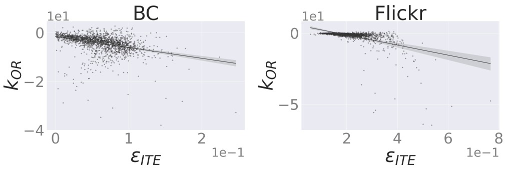

Figure 2: Joint distributions of the sum of Ollivier-Ricci curvatures in the neighborhood of each node and the ITE estimation error for the BC and Flickr networks. Regression lines with the corresponding 95% confidence intervals are marked on the plots.

Ricci flow adjustment for treatment effect estimation. The theory and experiments alike speak to the adverse effect of highly negative curvatures on estimating treatment effects. In line with this observation, in Section 5.1 we proposed a simple method to improve the estimation of treatment effects on networked data by flattening the network via the discrete Ricci flow. To evaluate this method, we apply this adjustment to the input graph of NetEst. We refer to the modified method as *f-NetEst*. The ϵ_{PEHE} values obtained from our experiments, Table 1, show that f-NetEst achieves the best performance on all datasets with relative gains of up to 52%. Comparing the distributions of the ITE estimation errors (Appendix I.3) further confirms that the Ricci flow adjustment leads to more accurate ITE estimates. Table 1 also reports the performances of multiple baseline models, and additional baselines are included in Appendix I.2. While our experiments primarily focus on NetEst, which outperforms all the baselines, we also explore the impact of the Ricci flow adjustment on the performance of T-Learner+GNN, which further confirms that our modification generally reduces estimation error in most GNN-based models.

Table 1: ϵ_{PEHE} for nine datasets, comparing the proposed f-NetEst against NetEst and baseline models. f-TLearner+GNN refers to the experiment with the proposed Ricci flow adjustment using T-Learner+GNN. Boldface and underline mark the best and second best performances. Green and yellow mark relative gains greater than 5% and less than -5% from the Ricci flow adjustment.

	BC	Flickr	Cornell	Texas	Wisconsin	chameleon	Cora	CiteSeer	Actor
T-Learner+RF	0.328	0.462	0.192	0.414	0.463	0.372	0.232	0.386	0.238
TARNet	0.969	1.024	0.705	1.028	0.711	1.212	0.679	0.638	0.796
T-Learner+GNN	4.178	9.630	5.125	4.437	0.559	16.715	0.285	0.529	7.912
f-TLearner+GNN	3.268	2.762	4.370	3.106	0.466	7.764	0.263	0.494	3.896
NetEst	0.069	0.213	0.165	0.330	0.147	0.247	0.082	0.176	0.094
f-NetEst (ours)	**0.033**	**0.208**	**0.127**	**0.308**	**0.142**	**0.230**	**0.078**	**0.165**	**0.088**

8 Conclusions, limitations, and ethical considerations

We delved into the unexplored territory of leveraging geometry for causal representation learning on networked data via GNNs. We established a theoretical connection between curvature and causal inference, uncovering the challenges posed by negative curvatures in identifying causal effects. We presented numerical results using graph Ricci curvature to predict the reliability of causal effect estimations on networked data, empirically validating that positive curvature regions lead to more accurate results. We then proposed using the Ricci flow to enhance treatment effect estimation on networked data, achieving superior performance through flattening the edges in the network. To the best of our knowledge, this work is the first to formally establish the connection between graph curvature and network causal inference; opening new avenues for applications of graph geometry in causal representation learning.

Limitations and future directions. Our proposed method cannot target specific neighborhoods of the network for improving causal effect estimation. Moreover, using Ricci flow to reduce treatment effect estimation error is a static adjustment on the graph that is not efficiently updated during training. Our proposed improvement effectively alters the graph by weighting the edges, requiring careful consideration regarding conceptual consistency of the edge weights with the context of the problem in hand. In future work, our aim is to incorporate these additional dimensions, enhancing the robustness and applicability of curvature-based techniques in causal representation learning.

Ethics statement. While all used data are standard in the community, they have the risk of being biased, this affecting the experimental results but not the theoretical work. Properly detecting causal factors and their uncertainty, as here introduced, can help with the development of fair ML.

Acknowledgments and Disclosure of Funding

This work was partially supported by ONR, NGA, and NSF.

References

[1] Clement John Adkins. Equilibrium Thermodynamics. Cambridge University Press, 1983.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

[3] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. Nature Machine Intelligence, 3(12):1023–1032, 2021.

[4] Frank Bauer, Bobo Hua, Jürgen Jost, Shiping Liu, and Guofang Wang. The geometric meaning of curvature: Local and nonlocal aspects of ricci curvature. Modern Approaches to Discrete Curvature, pp. 1–62, 2017.

[5] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4):18–42, 2017.

[6] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478, 2021.

[7] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. International Conference on Learning Representations, 2014.

[8] Peter Bühlmann. Invariance, causality and robustness. Statistical Science, 35(3):404–426, 2020.

[9] Wenming Cao, Zhiyue Yan, Zhiquan He, and Zhihai He. A comprehensive survey on geometric deep learning. IEEE Access, 8:35929–35949, 2020.

[10] Zhixuan Chu, Jianmin Huang, Ruopeng Li, Wei Chu, and Sheng Li. Causal effect estimation: Recent advances, challenges, and opportunities. arXiv preprint arXiv:2302.00848, 2023.

[11] Spencer Compton, Murat Kocaoglu, Kristjan Greenewald, and Dmitriy Katz. Entropic causal inference: Identifiability and finite sample results. In Advances in Neural Information Processing Systems, volume 33, pp. 14772–14782. Curran Associates, Inc., 2020.

[12] Spencer Compton, Kristjan Greenewald, Dmitriy A Katz, and Murat Kocaoglu. Entropic causal inference: Graph identifiability. In International Conference on Machine Learning, pp. 4311–4343. PMLR, 2022.

[13] Irina Cristali and Victor Veitch. Using embeddings for causal estimation of peer influence in social networks. In Advances in Neural Information Processing Systems, volume 35, pp. 15616–15628. Curran Associates, Inc., 2022.

[14] Lloyd Demetrius and Thomas Manke. Robustness and network evolution—an entropic principle. Physica A: Statistical Mechanics and its Applications, 346(3-4):682–696, 2005.

[15] Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio, and Michael M Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In International Conference on Machine Learning, pp. 7865–7885. PMLR, 2023.

[16] Manfredo Perdigao Do Carmo and J Flaherty Francis. Riemannian Geometry, volume 6. Springer, 1992.

[17] Denis J Evans, Ezechiel Godert David Cohen, and Gary P Morriss. Probability of second law violations in shearing steady states. Physical Review Letters, 71(15):2401, 1993.

[18] Laura Forastiere, Edoardo M Airoldi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. Journal of the American Statistical Association, 116(534):901–918, 2021.

[19] Forman. Bochner's method for cell complexes and combinatorial ricci curvature. Discrete & Computational Geometry, 29:323–374, 2003.

[20] Dennis Frauen and Stefan Feuerriegel. Estimating individual treatment effects under unobserved confounding using binary instruments. In International Conference on Learning Representations, 2022.

[21] Shunwang Gong, Mehdi Bahri, Michael M Bronstein, and Stefanos Zafeiriou. Geometrically principled connections in graph neural networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11415–11424. IEEE, 2020.

[22] Ruocheng Guo, Jundong Li, and Huan Liu. Learning individual causal effects from networked observational data. In International Conference on Web Search and Data Mining, pp. 232–240. Association for Computing Machinery, 2020.

[23] Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. Bacadi: Bayesian causal discovery with unknown interventions. In International Conference on Artificial Intelligence and Statistics, pp. 1411–1436. PMLR, 2023.

[24] Richard S Hamilton. The ricci flow on surfaces, mathematics and general relativity. Contemporary Mathematics, 71:237–261, 1988.

[25] Shonosuke Harada and Hisashi Kashima. Graphite: Estimating individual effects of graph-structured treatments. In International Conference on Information & Knowledge Management, pp. 659–668. Association for Computing Machinery, 2021.

[26] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. Journal of Causal Inference, 6(2):20170016, 2018.

[27] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

[28] Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

[29] Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In International Conference on Machine Learning, pp. 14316–14332. PMLR, 2023.

[30] Song Jiang and Yizhou Sun. Estimating causal effects on networked observational data via representation learning. In International Conference on Information & Knowledge Management, pp. 852–861. Association for Computing Machinery, 2022.

[31] Miao Jin, Junho Kim, Feng Luo, and Xianfeng Gu. Discrete surface ricci flow. IEEE Transactions on Visualization and Computer Graphics, 14(5):1030–1043, 2008.

[32] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In International Conference on Machine Learning, pp. 3020–3029. PMLR, 2016.

[33] Jürgen Jost and Shiping Liu. Ollivier's ricci curvature, local clustering and curvature-dimension inequalities on graphs. Discrete & Computational Geometry, 51(2):300–322, 2014.

[34] Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. In Advances in Neural Information Processing Systems, volume 34, pp. 24841–24854. Curran Associates, Inc., 2021.

[35] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475, 2022.

[36] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In International Conference on Machine Learning, pp. 5067–5077. PMLR, 2020.

[37] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing, 20(1):359–392, 1998.

[38] Nan Rosemary Ke, Silvia Chiappa, Jane X Wang, Jorg Bornschein, Anirudh Goyal, Melanie Rey, Theophane Weber, Matthew Botvinick, Michael Curtis Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. In International Conference on Learning Representations, 2022.

[39] Luke Keele. The statistics of causal inference: A view from political methodology. Political Analysis, 23 (3):313–335, 2015.

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

[41] Murat Kocaoglu, Alexandros Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In AAAI Conference on Artificial Intelligence, volume 31:, 2017.

[42] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10): 4156–4165, 2019.

[43] Yong Lin, Linyuan Lu, and Shing-Tung Yau. Ricci curvature of graphs. Tohoku Mathematical Journal, 63 (4):605–627, 2011.

[44] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16021–16030. IEEE, 2022.

[45] Yang Liu, Chuan Zhou, Shirui Pan, Jia Wu, Zhao Li, Hongyang Chen, and Peng Zhang. Curvdrop: A ricci curvature based approach to prevent graph neural networks from over-smoothing and over-squashing. In Web Conference, pp. 221–230. Association for Computing Machinery, 2023.

[46] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. Annals of Mathematics, pp. 903–991, 2009.

[47] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[48] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In International Conference on Learning Representations, 2021.

[49] Yunan Luo, Jian Peng, and Jianzhu Ma. When causal inference meets deep learning. Nature Machine Intelligence, 2(8):426–427, 2020.

[50] Jing Ma, Ruocheng Guo, Chen Chen, Aidong Zhang, and Jundong Li. Deconfounding with networked observational data in a dynamic environment. In International Conference on Web Search and Data Mining, pp. 166–174. Association for Computing Machinery, 2021.

[51] Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In International Conference on Artificial Intelligence and Statistics, pp. 3700–3708. PMLR, 2021.

[52] Nicolai Meinshausen. Causality from a distributional robustness point of view. In IEEE Data Science Workshop, pp. 6–10. IEEE, 2018.

[53] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5115–5124. IEEE, 2017.

[54] Chien-Chun Ni, Yu-Yao Lin, Feng Luo, and Jie Gao. Community detection on networks with ricci flow. Scientific Reports, 9(1):1–12, 2019.

[55] Yann Ollivier. Ricci curvature of markov chains on metric spaces. Journal of Functional Analysis, 256(3): 810–864, 2009.

[56] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In Advances in Neural Information Processing Systems, volume 33, pp. 857–869. Curran Associates, Inc., 2020.

[57] Judea Pearl. Statistics and causal inference: A review. Test, 12:281–345, 2003.

[58] Judea Pearl. Causality: Models, Reasoning and Inference. Cambridge University Press, 2009.

[59] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In International Conference on Learning Representations, 2019.

[60] Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In Advances in Neural Information Processing Systems, volume 35, pp. 10904–10917. Curran Associates, Inc., 2022.

[61] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(5):947–1012, 2016.

[62] Maryam Pouryahya, James Mathews, and Allen Tannenbaum. Comparing three notions of discrete ricci curvature on biological networks. arXiv preprint arXiv:1712.02943, 2017.

[63] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

[64] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(2):215–246, 2021.

[65] Kenneth J Rothman and Sander Greenland. Causation and causal inference in epidemiology. American Journal of Public Health, 95(S1):S144–S150, 2005.

[66] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. Journal of Complex Networks, 9(2):cnab014, 2021.

[67] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American Statistical Association, 75(371):591–593, 1980.

[68] Areejit Samal, RP Sreejith, Jiao Gu, Shiping Liu, Emil Saucan, and Jürgen Jost. Comparative analysis of two discretizations of ricci curvature for complex networks. Scientific Reports, 8(1):8650, 2018.

[69] Romeil Sandhu, Tryphon Georgiou, Ed Reznik, Liangjia Zhu, Ivan Kolesov, Yasin Senbabaoglu, and Allen Tannenbaum. Graph curvature for differentiating cancer networks. Scientific Reports, 5(1):12323, 2015.

[70] Bernhard Schölkopf. Causality for machine learning. In Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 765–804. Association for Computing Machinery, 2022.

[71] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. Proceedings of the IEEE, 109(5): 612–634, 2021.

[72] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning, pp. 3076–3085. PMLR, 2017.

[73] Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In Uncertainty in Artificial Intelligence, pp. 1546–1555. PMLR, 2021.

[74] Joshua Southern, Jeremy Wayland, Michael Bronstein, and Bastian Rieck. Curvature filtrations for graph generative model evaluation. arXiv preprint arXiv:2301.12906, 2023.

[75] RP Sreejith, Karthikeyan Mohanraj, Jürgen Jost, Emil Saucan, and Areejit Samal. Forman curvature for complex networks. Journal of Statistical Mechanics: Theory and Experiment, 2016(6):063206, 2016.

[76] Suraj Srinivas, Kyle Matoba, Himabindu Lakkaraju, and François Fleuret. Efficient training of low-curvature neural networks. In Advances in Neural Information Processing Systems, volume 35, pp. 25951–25964. Curran Associates, Inc., 2022.

[77] Allen Tannenbaum, Chris Sander, Liangjia Zhu, Romeil Sandhu, Ivan Kolesov, Eduard Reznik, Yasin Senbabaoglu, and Tryphon Georgiou. Graph curvature and the robustness of cancer networks. arXiv preprint arXiv:1502.04512, 2015.

[78] Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In International Conference on Learning Representations, 2021.

[79] Mark J van der Laan. Causal inference for networks. Technical report, U.C. Berkeley Division of Biostatistics, 2012.

[80] Hal R Varian. Causal inference in economics and marketing. Proceedings of the National Academy of Sciences, 113(27):7310–7315, 2016.

[81] Victor Veitch, Yixin Wang, and David Blei. Using embeddings to correct for unobserved confounding in networks. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[82] Melanie Weber, Emil Saucan, and Jürgen Jost. Characterizing complex networks with forman-ricci curvature and associated geometric flows. Journal of Complex Networks, 5(4):527–550, 2017.

[83] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In International Conference on Machine Learning, pp. 40–48. PMLR, 2016.

[84] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. ACM Transactions on Knowledge Discovery from Data, 15(5):1–46, 2021.

[85] Ze Ye, Kin Sum Liu, Tengfei Ma, Jie Gao, and Chao Chen. Curvature graph network. In International Conference on Learning Representations, 2020.

[86] Elena Zheleva and David Arbour. Causal inference from network data. In SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 4096–4097. Association for Computing Machinery, 2021.

Appendix

A Causal identification assumptions

A set of assumptions, often referred to as *identification strategy*, are commonly considered for identifying the causal effect. In Section 2.1 we named two common assumptions. Here we include a description of these assumptions, as well as two other common assumptions [18, 28, 63, 67]:

- **Positivity:** For every unit i, $\mathbb{P}\left[t_i = 1 | x_i\right] \in (0, 1)$, i.e., each unit may or may not receive the treatment.
- **Consistency:** If the treatment and covariates of unit i are t_i and x_i, then $Y_i = Y_i | do(t_i, x_i)$. In other words, the potential outcome of the observed treatment and covariates is the same as the observed outcome.
- **Strong Ignorability:** Also referred to as *unconfoundedness*, this assumption is formally defined as $\{Y|do(T = 1), Y|do(T = 0)\} \perp\!\!\!\perp T|X$. In other words, conditional on all the measured covariates, the potential outcome does not depend on the treatment assignment.
- **Stable Unit Treatment Values Assumption (SUTVA):** The potential outcome of a unit is unaffected by treatment assignment of all other units.

These assumptions, although not always sufficient or necessary, could lead to identification of the treatment effect in various settings where there is no network effect, but fail to do so in the presence of network effect [30]. However, Jiang & Sun [30] show the identifiability of the treatment effect estimated by NetEst, under a set of modified assumptions that account for the covariates of neighbors and the peer effect. For a graph $G = (V, E)$ with treatments $\{t_v\}_{v \in V}$, features $\{x_v\}_{v \in V}$, peer exposures $\{z_v\}_{v \in V}$, and potential outcomes $\{Y_v\}_{v \in V}$, these assumptions are as follows [30]:

- **Positivity:** For every node $v \in V$, $\mathbb{P}\left[t_v = 1 | x_v, \{x_u\}_{u \in N_v}\right] \in (0, 1)$.
- **Consistency:** For every node $v \in V$, $Y_v = Y_v | do(t_v = t, z_v = z)$.
- **Strong Ignorability:** For every node $v \in V$, $Y_v | do(t_v, z_v) \perp\!\!\!\perp t_v, z_v | x_v, \{x_u\}_{u \in N_v}$.
- **Markov:** For any two sets of treatments $\{t_v\}_{v \in V}$ and $\{t'_v\}_{v \in V}$, given any node $w \in V$, if $t_w = t'_w$ and $Z(\{t_u\}_{u \in N_w}) = Z(\{t'_u\}_{u \in N_w})$, then $Y_w | do(\{t_v\}_{v \in V}) = Y_w | do(\{t'_v\}_{v \in V})$, where $Z(.)$ is the exposure function, and we use $do(\{t_v\}_{v \in V})$ to denote enforcing all treatments in $\{t_v\}_{v \in V}$. That is, the potential outcome of any node is only affected by its treatment and the treatments of its immediate neighbors.

B Causal inference, invariance, and distributional robustness

B.1 Causal inference as risk minimization

Following Bühlmann [8], we formalize the derivation of causal quantities from statistical quantities as a worst-case risk minimization problem. Adopting the notation in Bühlmann [8], let $\mathbf{X}$ and $\mathbf{Y}$ denote the covariates and the outcomes, let Y^e and X^e denote the random variable and the random vector corresponding to an observed environment $e \in \mathcal{E}$, and let $\mathcal{F} \supseteq \mathcal{E}$ denote the union of observed and unobserved environments encompassing the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. The causal relationship of $\mathbf{X}$ and $\mathbf{Y}$ is trivially revealed when $\mathcal{F} = \mathcal{E}$, hence, without loss of generality, we assume $\mathcal{E} \subset \mathcal{F}$.

Learning the relationship between $\mathbf{X}$ and $\mathbf{Y}$ can be described as predicting Y^e from X^e based on observations $e \in \mathcal{E}$, such that the prediction is robust under the choice of $e \in \mathcal{F}$. To this end, consider a linear model as an example; we can formulate a causal inference parameter, θ_{causal}, as the worst case regression estimand below, with the constraint that e does not directly impact the joint distribution of X^e and Y^e [8], hereafter referred to by $\mathbf{C}$,

$$\theta_{\text{causal}} = \underset{b}{\operatorname{argmin}} \max_{e \in \mathcal{F}} \mathbb{E}\left[\left(Y^e - (X^e)^T b\right)^2\right]. \tag{6}$$

B.2 Invariance of causal models

Invariance of this worst-case risk minimization is a core component behind inferring causality from data. Given a set of environments $\mathcal{G} \subseteq \mathcal{F}$, invariance can be formalized as the existence of a subset of covariate indices $S \subset \{1, \ldots, n_X\}$ satisfying $\mathbf{A}_S(\mathcal{G})$, defined below,

Definition 1 (Bühlmann [8]). *$\mathbf{A}_S(\mathcal{G})$ is defined as the property that $\{\mathcal{L}(Y^e|X_S^e)\,|e \in \mathcal{G}\}$ is a singleton, where X_S^e denotes the subset of covariates induced by indices in S, and $\mathcal{L}(Y^e|X_S^e)$ denotes the loss function* $\mathbb{E}\left[\left(Y^e - (X_S^e)^T b\right)^2\right]$.

If $\mathbf{A}_S(\mathcal{G})$ holds, the causal parameter in Equation 6 remains the same under variations in $e \in \mathcal{G}$. For causal inference, we are particularly interested in the invariance assumption $\mathbf{A}_S(\mathcal{G})$ when $\mathcal{G} = \mathcal{E}$ for estimating θ_{causal} from the data, or when $\mathcal{G} = \mathcal{F}$ for the more general case of determining causal parameters over the population. Assuming there exists an S for which $\mathbf{A}_S(\mathcal{F})$ holds, the problem of causal inference is then to find such $S = \text{pa}(Y) \subset \{1, \ldots, n_X\}$, where $\{X_i\}_{i \in \text{pa}(Y)}$ is the set of direct causal parents of Y. Taking a step towards computation, this problem can be formulated in terms of structural equation models (SEMs) between X and Y, as finding the set $\text{pa}(Y)$ such that $\mathbf{C}$ is satisfied [8]. This can be formalized as satisfying $\mathbf{B}(\mathcal{F})$, where $\mathbf{B}(\mathcal{G})$ is,

Definition 2 (Bühlmann [8]). *$\mathbf{B}(\mathcal{G})$ is defined as the property that $\left\{p_{\epsilon^e}|e \in \mathcal{G} \wedge Y^e = f\left(X_{pa(Y)}^e, \epsilon^e\right)\right\}$ is a singleton, where f determines the SEM, ϵ^e is independent of $X_{pa(Y)}^e$, and p_{ϵ^e} is the distribution of ϵ^e.*

The assumption $\mathbf{B}(\mathcal{F})$ in fact completes the formulation of causal inference problems from the perspective of invariance, with Proposition 1 in Bühlmann [8], which states that under $\mathbf{B}(\mathcal{F})$, $\text{pa}(Y)$ satisfies $\mathbf{A}_{\text{pa}(Y)}(\mathcal{F})$. It follows that an identification strategy, when computing θ_{causal} over the observed environments, is taking the intersection of all sets S satisfying $\mathbf{A}_S(\mathcal{E})$. The main issue, however, is that such an identification mechanism relies on assumption $\mathbf{B}(\mathcal{F})$ and condition $\mathbf{C}$. We next discuss the robustness of an estimator in a regression problem which allows for relaxing these constraints.

B.3 Distributional robustness and causal inference

One common situation where $\mathbf{C}$ fails is the presence of hidden confounders. We can use anchor regression [64] to relax $\mathbf{C}$ and allow for hidden confounders. In anchor regression, we consider an anchor variable A with $\text{pa}(A) = \emptyset$. We allow A to be a causal parent of the covariates X, outcome Y, and hidden confounders H, as described in Figure 1. The anchor variable could be considered as an environment that is not constrained by $\mathbf{C}$. The corresponding linear SEM is then

$$\begin{bmatrix} X \\ Y \\ H \end{bmatrix} = B \begin{bmatrix} X \\ Y \\ H \end{bmatrix} + \epsilon + MA, \tag{7}$$

where B and M are unknown constant real matrices and ϵ is the noise vector which satisfies $\epsilon \perp\!\!\!\perp A$. This yields the following anchor regression problem for regressing Y on X,

$$Y = X^T\beta + H^T\alpha + A^T\xi + \epsilon_Y. \tag{8}$$

Since the anchor variable is a source in the graphical model, the anchor regression estimator minimizes a risk in the column space of A. Let Π_A be the projection matrix onto the column space of A for the sample, and let P_A denote the corresponding projection operator for the population case. The anchor regression estimand $\beta_A(\gamma)$ and estimator $\hat{\beta}_A(\gamma)$ for regressing an $n \times 1$ outcome $\mathbf{Y}$ on an $n \times m$ matrix of covariates $\mathbf{X}$, corresponding to Y and X in Equation 8, are given by

$$\beta_A(\gamma) = \underset{b}{\text{argmin}} \left\{\mathbb{E}\left[\left((I - P_A)\left(Y - X^T b\right)\right)^2\right] + \gamma\mathbb{E}\left[\left(P_A(Y - X^T b)\right)^2\right]\right\}, \tag{9}$$

$$\hat{\beta}_A(\gamma) = \underset{b}{\text{argmin}} \left\{\frac{1}{n}\left\|(I - \Pi_A)\left(\mathbf{Y} - \mathbf{X}b\right)\right\|_2^2 + \frac{\gamma}{n}\left\|\Pi_A(\mathbf{Y} - \mathbf{X}b)\right\|_2^2\right\}, \tag{10}$$

where the second term in the objective functions encourages the residuals to be orthogonal to A [8]. We can compute $\hat{\beta}_A(\gamma)$ through the Ordinary Least Square estimator for regressing a transformed outcome variable $W_\gamma Y$ on the corresponding transformed covariate $W_\gamma X$, where $W_\gamma := I - \left(1 - \sqrt{\gamma}\right) \Pi_A$. Recall that A captures the influence of what we previously referred to as the environment, thus encouraging independence of residuals from the environment and leading to further invariance with respect to the environment.

Consider the system under perturbation by a vector $v = M\delta$ for some δ replacing the anchor term in Equation 7. The SEM under perturbation can be written as

$$\begin{bmatrix} X^v \\ Y^v \\ H^v \end{bmatrix} = B \begin{bmatrix} X^v \\ Y^v \\ H^v \end{bmatrix} + \epsilon + v. \tag{11}$$

Let us impose $\delta \perp\!\!\!\perp \epsilon$ and constrain the norm of the expected perturbation by the order of a constant γ. That is, we consider a class of shift perturbations $\mathcal{C}_\gamma$ where the perturbation is generated in the column space of M by a vector δ independent of the noise, and where the typical size of the perturbation is $O(\gamma)$ as $\gamma \to \infty$. Also assume, without loss of generality, that X and Y are centered at 0. Under these conditions, if $\mathbb{E}\left[AA^T\right]$ is positive definite, the following proposition holds [8, 64].

Proposition 1. *Given any $b \in \mathbb{R}^m$, if A and $Y - X^T b$ are uncorrelated, $Y^v - (X^v)^T b$ in the perturbed system has the same distribution for all $v \in span(M)$.*

Proposition 1 points to what leads to the distributional robustness of the anchor regression estimand. This is due to an equality between a worst-case residual in the perturbed system and the objective function for the estimand in Equation 9 [8, 64],

Theorem 2. *For any $b \in \mathbb{R}^m$*

$$\sup_{v \in \mathcal{C}_\gamma} \mathbb{E}\left[\left(Y^v - (X^v)^T b\right)^2\right] = \mathbb{E}\left[\left(\left(I - P_A\right)\left(Y - X^T b\right)\right)^2\right] + \gamma \mathbb{E}\left[\left(P_A(Y - X^T b)\right)^2\right].$$

Corollary 1, which states that $\beta_A(\gamma)$ minimizes a worst case risk over the class of shift perturbations $\mathcal{C}_\gamma$, follows trivially considering Equation 9:

Corollary 1. $\beta_A(\gamma) = \operatorname{argmin}_{b \in \mathbb{R}^m} \sup_{v \in \mathcal{C}_\gamma} \mathbb{E}\left[\left(Y^v - (X^v)^T b\right)^2\right].$

Recall that the second term in the objective function of the anchor regression estimand in Equation 9 is essentially a causal regularization term that encourages the invariance of the residuals with respect to the environment. Theorem 2 and Corollary 1 establish that the anchor regression estimand corresponds to a worst-case risk minimization in a perturbed system, and simultaneously encourages conditions which bring us closer to a scenario where the assumptions for causal identification hold. In other words, an estimator that satisfies the criteria for a causal parameter is also a distributionally robust optimizer. This concludes our discussion of causal inference as a worst-case risk optimization, establishing the connection between causal inference and distributional robustness.

C Ricci curvature notions on graphs

The Ricci curvature indicates deviation from the Euclidean space [4, 16, 62]. On graphs, this translates to measuring how much the neighborhood of an edge differs from a grid. We used the Ollivier-Ricci curvature [55] for the experiments reported in the main text of the paper. Forman-Ricci curvature [19] is an alternative notion of Ricci curvature on graphs, which we use, in addition to Ollivier-Ricci curvature, in supplementary experiments included in Appendix I.1. In this section, we formally define these two Ricci-type graph curvatures.

The Ollivier-Ricci curvature is an optimal transport formulation of the Ricci curvature on graphs. Given a graph $G = (V, E)$, for an edge $(v, u) \in E$, with μ_v and μ_u probability measure on the nodes anchoring (v, u), the Ollivier-Ricci curvature is defined as

$$\kappa_{OR}(v, u) := 1 - \frac{W_1(\mu_v, \mu_u)}{d_G(v, u)}, \tag{12}$$

where $d_G(.)$ is a distance metric on V and W_1 denotes the 1-Wasserstein distance [33, 43]. Given the flexibility with respect to the choice of μ_v and μ_u, the Ollivier-Ricci curvature is a versatile tool for capturing the local geometry of edges in a graph.

The Forman-Ricci curvature is a combinatorial curvature notion. The Forman-Ricci curvature of an edge $(v, u) \in E$ in an undirected graph is given by

$$\kappa_{FR}(v,u) := w_{vu} \left[\frac{w_v}{w_{vu}} + \frac{w_u}{w_{vu}} - \sum_{(v',u') \in N_v \times N_u} \left(\frac{w_v}{\sqrt{w_{vu} w_{vv'}}} + \frac{w_u}{\sqrt{w_{vu} w_{uu'}}} \right) \right], \tag{13}$$

where w_v is the weight of the node v, w_{vu} is the weight of the edge (v, u), and N_v ist the set of neighbors of the node v [75, 82]. By convention, all weights are set to 1 in an unweighted graph, in which case the Forman curvature becomes $\kappa_{FR}(v, u) = 4 - d_v - d_u$, where d_v denotes the node degree.

D Theorem details

In this appendix we provide the proof of Theorem 1, which states the following under the assumptions listed in the appendix Section D.1: Given X_i and Y_i for $i \in \{1, 2\}$, corresponding to two sets of data with causal models $Y_i = f_i(X_i, E_i)$, if an alternative model $X_i = g(Y_i, \tilde{E}_i)$ fits the data, having non-negative and negative lower bounds on the Ricci curvatures corresponding to X_2 and X_1 implies that for some constant η, the probability that the Shannon entropy of $\tilde{E}_2$ is greater than η is greater than or equal to the probability that the entropy of $\tilde{E}_1$ is lower bounded by η.

D.1 Assumptions

Given the triplets (X_1, Y_1, E_1) and (X_2, Y_2, E_2), with structural causal models $Y_i = f_i(X_i, E_i)$ for $i = 1, 2$, we make the following assumptions:

(Ai) Considering probability measures μ_{X_1} and μ_{X_2} corresponding to X_1 and X_2, there exists a pair of measures μ_0 and μ_1 such that μ_{X_1} and μ_{X_2} are on the geodesics between μ_0 and μ_1 in a 2-Wassertein metric space.

(Aii) $H(Y_1) \approx H(Y_2)$ and $H(E_1) \approx H(E_2)$, where we use $\approx$ to denote sufficiently close, and $H(.)$ denotes the Shannon entropy.

(Aiii) The conditions for Conjecture 1 in Kocaoglu et al. [41] and Compton et al. [11]: $X \sim p(X)$ and $E \sim p(E)$, where $p(X)$ is a uniform random sample from the n-dimensional probability simplex, $p(E)$ is sampled uniformly from the points in the m-dimensional probability simplex satisfying $H(E) \leq \log(n) + \mathcal{O}(1)$, and f is sampled according to p_f satisfying $\left\| \frac{p_f}{p_U} \right\|_\infty \leq n^c$ for some constant c, where p_U is a uniform distribution [12].

In assumption (Aii) above, we use the term sufficiently close to refer to the existence of a sufficiently small upper bound on the distance between the two values.

Assumptions (Ai) and (Aiii) are primarily technical assumptions to ensure applicability of inequalities 3 and 4 used in the proof. Assumption (Aii) on the other hand, while facilitating steps of the proof, has a conceptual implication: (Aii) implies that the difference in the randomness of the two datasets is primarily due to X_1 and X_2.

D.2 Proof

The proof of Theorem 1, under the assumptions above, relies on Inequality 3 from Pouryahya et al. [62] and Lott & Villani [46], and the results from Compton et al. [11] and Compton et al. [12] leading to Inequality 4. Given alternative models $X_i = g(Y_i, \tilde{E}_i)$ for $i = 1, 2$ with exogenous variables $\tilde{E}_i$, under (Aiii), Inequality 4 gives the following lower bound on the Shannon entropy of $\tilde{E}_i$,

$$H(X_i) - H(Y_i) + H(E_i) < H(\tilde{E}_i). \tag{14}$$

Suppose $\underline{k}_i < 0 \leq \underline{k}_2$ where $\underline{k}_i$ is a lower bound on the Ricci curvature corresponding to X_i. Then, by Inequality 3, assuming (Ai), we have

$$\underline{s}_2 > \underline{s}_1, \tag{15}$$

where $\underline{s}_i$ is a lower bound on the Boltzmann entropy corresponding to X_i. On the other hand, the Boltzmann entropy can be written as a constant scaling of the Shannon entropy. Thus, given lower bounds $\underline{h}_1$ and $\underline{h}_2$ on $H(X_1)$ and $H(X_2)$, Inequality 15 implies $\underline{h}_2 > \underline{h}_1$. Consider a constant $h \in (\underline{h}_1, \underline{h}_2)$. Since $\underline{h}_2$ is a lower bound for $H(X_2)$, it holds that $\mathbb{P}\left[H(X_2) \geq h\right] = 1 \geq \mathbb{P}\left[H(X_1) \geq h\right]$, where $\mathbb{P}(.)$ denotes the probability. Hence, under assumption (Aii), $\mathbb{P}\left[\Lambda_2 > \eta\right] = 1 \geq \mathbb{P}\left[\Lambda_1 > \eta\right]$, where $\Lambda_i \coloneqq H(X_i) - H(Y_i) + H(E_i)$ and $\eta \in (\underline{h}_1 - H(Y_1) + H(E_1), \underline{h}_2 - H(Y_2) + H(E_2))$ is a constant. Using the lower bounds in 14, this implies

$$\mathbb{P}\left[H(\tilde{E}_2) > \eta\right] \geq \mathbb{P}\left[H(\tilde{E}_1) > \eta\right],$$

completing the proof of the theorem connecting causal inference with curvature.

E Related work

We showed the connection between curvature and causal inference, bridging for the first time the works on invariance and robustness of causal models, geometric deep learning, and deep learning for causal inference. Although we have cited related works in each section as appropriate, we now mention the most relevant works in each aspect and reference other contributions in the literature. While these works provide essential foundations and motivations for the theory in this work, none of them establishes the explicit links we developed in the paper and, in particular, the close connection between geometry/curvature, robustness, and causal inference.

Invariance, robustness, and causal inference. Learning representations that are invariant across a set of environments is the primary goal of invariant causal prediction (ICP) [8, 26, 61, 73] and invariant risk minimization (IRM) [2, 8, 44, 73]. Bühlmann [8] formally describes how IRM can lead to a distributionally robust estimator while imposing causal identification assumptions.

Geometric deep learning. Geometric tools have been instrumental to recent advances on GNNs [5–7, 21]. Discrete Ricci curvatures on graphs, in particular, are well-established measures with roots in Riemannian geometry [55, 68, 69], with numerous applications for GNNs [74, 78]. The connection between Ricci curvature and entropy is known from the optimal transport literature [46], based on which, [62] uses Ricci curvature as a measure of system robustness. Moreover, curvature has been used by Srinivas et al. [76] to improve robustness in neural networks. However, the literature does not establish a connection with distributional robustness, a gap that we fill with the help of results from entropic causal inference [11, 12, 41].

Deep learning for causal inference. Deep learning methods have had success in estimating treatment effect [47, 72], counterfactual inference [32, 56], and other problems in causal inference [20, 23, 29, 38, 48, 49, 60]. Causal effect estimation on networked data on the other hand, is known to be notoriously challenging [79, 86]. Various methods have been proposed for estimating the causal effect in structured data which violate traditional identification assumptions [13, 22, 25, 30, 34, 51]. For instance, Guo et al. [22] uses a *Network Deconfounder* to learn a representation of hidden confounders from the data, Kaddour et al. [34] proposes an effect decomposition, and Veitch et al. [81] and Cristali & Veitch [13] use the embeddings to deal with unobserved confounders and the homophily effect. Another approach, taken in Jiang & Sun [30], Ma & Tresp [51], Harada & Kashima [25], is to account for the peer treatment effects in the network using GNN-based causal estimation methods, which allows the violation of SUTVA. However, the literature lacks a practical indicator of the local reliability of the estimates. We show that Ricci curvature can serve as such an indicator, and informed by this result, we propose a preprocessing using the Ricci flow to improve the causal effect estimates obtained from GNN-based methods.

F The NetEst model

Given a graph $G = (V, E)$ with the adjacency matrix A, features X, observed outcome Y, and treatments T, the NetEst model [30] uses a summary function $Z : 2^T \to [0, 1]$ to capture the peer effect on unit $v \in V$ through v's *peer exposure*, $z_v = Z(\{t_u\}_{u \in N_v})$, where N_v denotes the set of immediate neighbors of the node v. Assuming a Markov-type property that the peer effect can be learned from the signals received from immediate neighbors, the peer exposure function is set to be the average treatment of the neighbors, i.e., $z_v = \sum_{u \in N_v} t_u / |N_v|$. The ITE, $\tau(x_v)$, for two treatments t' and t'' is then defined as

$$\tau(x_v) \coloneqq \mathbb{E}\left[Y_v|do(t_v = t', z_v = z') - Y_v|do(t_v = t'', z_v = z'') \;\middle|\; x_v, \{x_u\}_{u \in N_v}\right], \tag{16}$$

which is identified under the assumptions described in Appendix A [30].

NetEst consists of four modules: an encoder, two regularizers, and an estimator. The *encoder* module learns a representation for the nodes using a graph convolutional network, producing an embedding $s_v = \phi(x_v, \{x_u\}_{u \in N_v}) \in S$ for every unit $v \in V$. The *estimator* module is trained to estimate the observed outcome from the embeddings $\{s_v\}_{v \in V}$ by minimizing a mean squared error (MSE) loss. This MSE loss $\mathcal{L}_m$ is the *potential outcome loss*, between $m(s_v, t_v, z_v)$ and the potential outcome $Y_v|do(t_v, z_v)$, where $m : S \times \{0, 1\} \times [0, 1] \to \mathcal{Y}$ denotes the estimator, assuming a binary treatment, and $\mathcal{Y}$ is the outcome space. The $p(t|x)$ and $p(z|x, t)$ *regularizer* modules are used in an adversarial training scheme to resemble randomized treatment assignment and uniform peer exposure, respectively, minimizing two MSE losses $\mathcal{L}_t$ and $\mathcal{L}_z$ on the embeddings and treatments. Hence, NetEst is trained by first training the discriminators in the regularizer modules, minimzing their respective loss values, then updating the estimator to minimze $\mathcal{L}_m$, and in the end, updating the encoder to optimize a total loss $\mathcal{L} = \mathcal{L}_m + \alpha_t \mathcal{L}_t + \alpha_z \mathcal{L}_z$.

G Implementation parameters and hardware specifications

Since the main purpose of our experiments was to inspect the joint distribution of estimation errors and evaluate the impact of our proposed data preprocessing, we followed the parameters and setup used by Jiang & Sun [30] for all implementation and training purposes of NetEst, TARNet, CFR, and NetDeconf. The encoder of NetEst contains 1 graph convolution layer, the estimator has 3 fully-connected hidden layers of size 32, and the two regularization terms in the total training loss of the encoder both have weight 0.5. The learning rate is 0.001 for 300 epochs of full batch training using an Adam optimizer [40]. The meta learner baselines with GNN encoders, T-learner+GNN and X-Learner+GNN, are implemented using a graph convolutional network followed by a three-layer multilayer perceptron. All meta learners were fine tuned with grid search. The system specifications for the experiments are reported in Table 2.

Table 2: System specifications for the experiments.

CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
GPU	Nvidia V100
OS	Ubuntu 22.04.2 LTS
Architecture	x86_64

H Data

Validating causal inference methods and theories through experiments often requires data that contain counterfactual outcomes. To this end, the standard practice in the literature is to use semi-synthetic data, where the features are empirically observed, while the treatments and potential outcomes are simulated [22, 27, 30, 50, 72, 81]. Jiang & Sun [30] use the BlogCatalog (BC) and Flickr datasets [22, 50] to evaluate the performance of NetEst. In addition to these two datasets, we supplement our experiments with additional network datasets used in the geometric deep learning and GNN literature: Cornell, Texas, and Wisconsin networks from the WebKB dataset [1]; Chameleon network from the Wikipedia networks dataset [66]; Cora and CiteSeer networks [83]; and Actor network [59]. Table 3 includes descriptive statistics on these networks. Note that we only use the largest connected component in each network. Following Jiang & Sun [30], we split each network data into training, validation and test sets using METIS [37]. The treatments and potential outcomes for all network data are synthesized following the formulation in Jiang & Sun [30].

[1] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/

Table 3: Descriptive statistics for the networks used in our experiments.

	BC	Flickr	Cornell	Texas	Wisconsin	Chameleon	Cora	CiteSeer	Actor
Nodes	5196	7600	183	183	251	2277	2708	3327	7600
Edges	171743	30019	298	325	515	36101	10556	9104	30019
Features	8189	932	1703	1703	1703	2325	1433	3703	932

I Further experiments

I.1 Ricci curvature and treatment effect estimation error

In this section we include additional plots showing the joint distributions of the ITE estimation error for each node $v \in V$, $\varepsilon_{ITE}(v)$, and the Ricci curvature in the neighborhood of the node, for both Forman and Ollivier Ricci curvatures (this partially repeated from Figure 2 for completeness and ease of visualization/comparison). The distributions for the nine networks are shown in Figure 3, with two plots (one per curvature) for each dataset. All ITE estimations in this figure have been obtained using NetEst [30]. These distributions and the regression lines marked on the plots further confirm our theoretical results, which imply that highly negative Ricci curvature makes causal effect estimation more challenging.

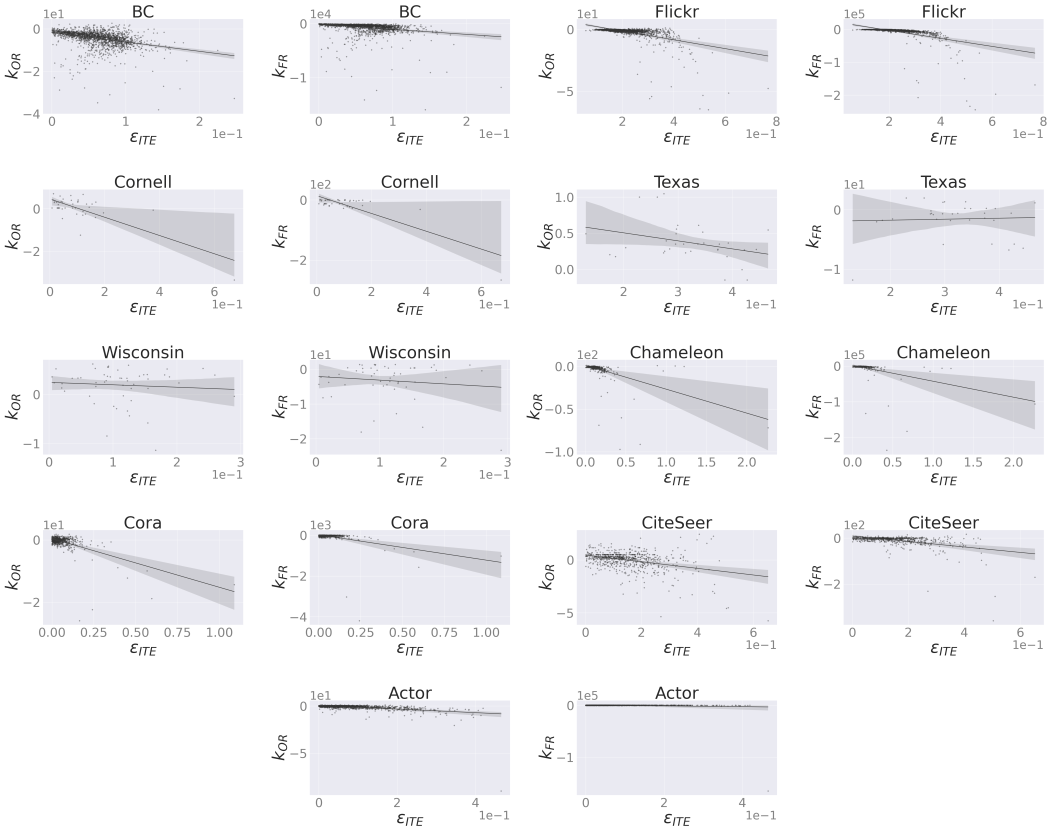

Figure 3: Joint distributions of the sum of Forman and Ollivier-Ricci curvatures in the neighborhood of each node and the estimation error of ITE for that node. The distributions for the nine networks are shown, with two plots (one per curvature) for each dataset. The regression lines with the corresponding 95% confidence intervals are marked on the plots.

I.2 Comparison with additional baselines for treatment effect estimation

We discussed our experiments on the improvement of treatment effect estimation on network data using the Ricci flow in Section 7. In this section, we supplement these experiments with additional baseline models, including CFR [72], NetDeconf [22], an X-Learner [42] with a random forest (RF) regressor, and an X-Learner implemented using a GNN encoder followed by a multilayer perceptron. Similar to the experiments reported in Section 7, we also report the impact of the Ricci flow adjustment on the performance of X-Learner+GNN and NetDeconf, which are the additional baseline models using GNN encoders. The performances of these baselines, as well as those reported in Section 7, are compared with our target method, f-NetEst, in Table 4. These results further confirm the superior performance of f-NetEst and the improvement resulting from the Ricci flow adjustment.

Table 4: ϵ_{PEHE} for nine datasets, comparing the proposed f-NetEst against NetEst and baseline models. The baselines include three models implemented with GNN encoders. The experiment with the Ricci flow adjustment for these models is marked with "f-". Boldface and underline mark the best and second best performances. Green and yellow mark relative gains greater than 5% and less than -5% from the Ricci flow adjustment.

	BC	Flickr	Cornell	Texas	Wisconsin	chameleon	Cora	CiteSeer	Actor
T-Learner+RF	0.328	0.462	0.192	0.414	0.463	0.372	0.232	0.386	0.238
X-Learner+RF	5.612	5.745	5.928	3.827	3.815	3.709	8.626	5.606	5.231
TARNet	0.969	1.024	0.705	1.028	0.711	1.212	0.679	0.638	0.796
CFR	0.895	0.960	0.806	1.038	0.849	0.926	0.570	0.620	0.735
T-Learner+GNN	4.178	9.630	5.125	4.437	0.559	16.715	0.285	0.529	7.912
X-Learner+GNN	4.627	3.933	20.461	1.995	16.244	329.959	3.165	4.428	4.296
NetDeconf	1.092	1.251	0.900	1.137	0.952	1.207	0.791	0.752	0.895
f-TLearner+GNN	3.268	2.762	4.370	3.106	0.466	7.764	0.263	0.494	3.896
f-XLearner+GNN	4.222	3.859	17.395	2.020	20.815	251.290	3.053	3.919	3.967
f-NetDeconf	1.088	1.245	0.900	1.143	0.954	1.200	0.810	0.767	0.898
NetEst	0.069	0.213	0.165	0.330	0.147	0.247	0.082	0.176	0.094
f-NetEst (ours)	**0.033**	**0.208**	**0.127**	**0.308**	**0.142**	**0.230**	**0.078**	**0.165**	**0.088**

I.3 ITE error distribution

In order to obtain a better understanding of how the Ricci flow adjustment impacts ITE estimation for each unit, we compare the empirical cumulative distribution functions (CDFs) of ε_{ITE} obtained from f-NetEst and NetEst, in the two datasets used by Jiang & Sun [30], as well as seven other networks described in Appendix H. As shown in Figure 4, the empirical CDF from f-NetEst is uniformly above that from NetEst for low ε_{ITE} values, which further confirms that flattening the edges leads to a larger proportion of units with low ITE estimation error.

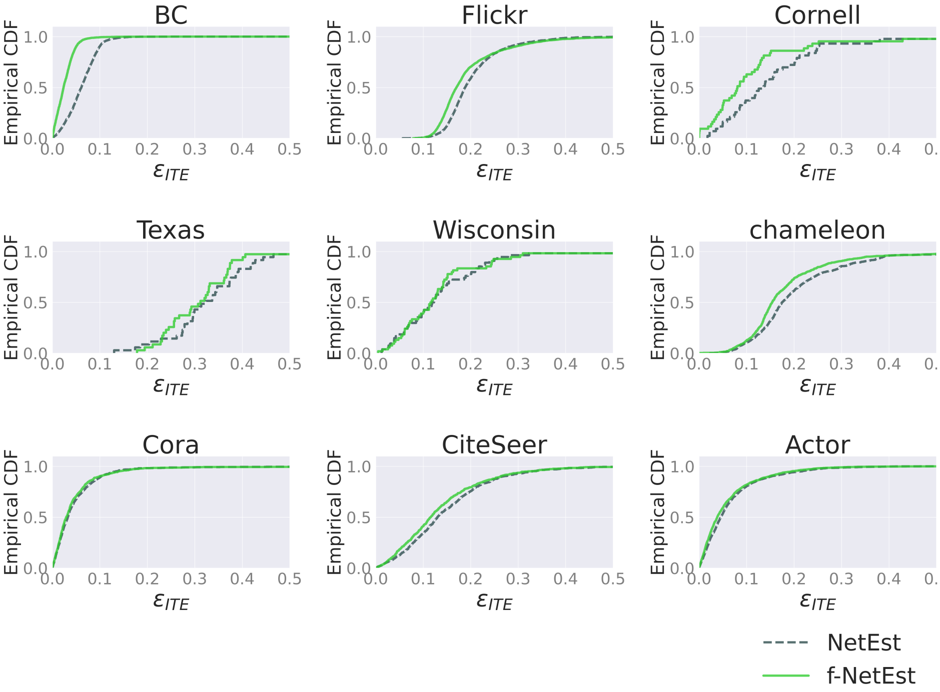

Figure 4: Empirical CDF of the ITE error, ε_{ITE}, obtained from NetEst (black) and f-NetEst (green).