From Geometry to Causality- Ricci Curvature and the Reliability of Causal Inference on Networks

Amirhossein Farzam [1] **Allen Tannenbaum** [2] **Guillermo Sapiro** [1 3]

Abstract

Causal inference on networks faces challenges posed in part by violations of standard identification assumptions due to dependencies between treatment units. Although graph geometry fundamentally influences such dependencies, the potential of geometric tools for causal inference on networked treatment units is yet to be unlocked. Moreover, despite significant progress utilizing graph neural networks (GNNs) for causal inference on networks, methods for evaluating their achievable reliability without ground truth are lacking. In this work we establish for the first time a theoretical link between network geometry, the graph Ricci curvature in particular, and causal inference, formalizing the intrinsic challenges that negative curvature poses to estimating causal parameters. The Ricci curvature can then be used to assess the reliability of causal estimates in structured data, as we empirically demonstrate. Informed by this finding, we propose a method using the geometric Ricci flow to reduce causal effect estimation error in networked data, showcasing how this newfound connection between graph geometry and causal inference could improve GNN-based causal inference. Bridging graph geometry and causal inference, this paper opens the door to geometric techniques for improving causal estimation on networks.

1 Introduction

Inferring causal effects of interventions from observational data is a critical task across disciplines — spanning epidemiology, economics, and political science (Rothman & Greenland, 2005; Keele, 2015; Varian, 2016; Yao et al., 2021). Causal effect estimation methods aim to estimate causal quantities by statistical ones (Pearl, 2009a). Estimating treatment effects on networked treatment units poses considerable difficulties due to violations of standard identification assumptions (van der Laan, 2012; Zheleva & Arbour, 2021). Despite progress on leveraging graph neural networks (GNNs) for causal models for structured data (Guo et al., 2020; Wein et al., 2021; Ma & Tresp, 2021; Harada & Kashima, 2021; Kaddour et al., 2021; Jiang & Sun, 2022; Cristali & Veitch, 2022), the full potentials of GNNs and the graph structure they encode are yet to be unleashed for causal inference on networked data.

In the realm of geometric deep learning, GNNs enable the use of inherent geometry in graph-structured data (Bruna et al., 2014; Bronstein et al., 2017; Monti et al., 2017; Cao et al., 2020; Gong et al., 2020; Ye et al., 2020; Bronstein et al., 2021; Atz et al., 2021; Topping et al., 2021; Southern et al., 2023). While the geometry of the graph is a key driver of network-induced endogeneities and consequent challenges posed to causal inference, the potential of geometric tools for understanding and enhancing causal inference on networks remains largely unexplored, leaving a significant gap in GNN-based arsenal of tools for causal inference. Moreover, existing techniques cannot canonically assess reliability without ground truth, which significantly limits validation.

Our work sets out to address these gaps by formally connecting network geometry, in the form of discrete curvature, and causal inference on networked data. Guided by the proposition that curvature could serve as a practical measure for *system* robustness in networks (Demetrius & Manke, 2005; Tannenbaum et al., 2015), we take theoretical steps to enable exploiting this geometric signature of robustness to gauge the *confidence* of causal models on networked data. Validated by our empirical observations revealing a negative correlation between Ricci curvature and individual treatment effect (ITE) estimation error, our theoretical results show that negative Ricci curvature corresponds to greater *intrinsic* difficulty in identifying learned causal parameters. Drawing from the theory of invariance of causal models and

[1]Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States [2]Department of Computer Science, Stony Brook University, Stony Brook, NY, United States [3]Apple, Cupertino, CA, United States. Correspondence to: Amirhossein Farzam <a.farzam@duke.edu>, Guillermo Sapiro <guillermo.sapiro@duke.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

distributional robustness of causal parameters (Peters et al., 2016; Meinshausen, 2018; Bühlmann, 2020; Weichwald & Peters, 2021), we establish the theoretical connection between causal inference and curvature, bridging the gap using the connection between curvature and entropy (Tannenbaum et al., 2015; Sandhu et al., 2015; Pouryahya et al., 2017) and results from entropic causal inference (Kocaoglu et al., 2017; Compton et al., 2020; 2022). This contribution is summarized schematically in Figure 1, visually highlighting existing works in the literature, the gap between causal inference and Ricci curvature, and the path connecting multiple arms of machine learning that we formulate to bridge this gap.

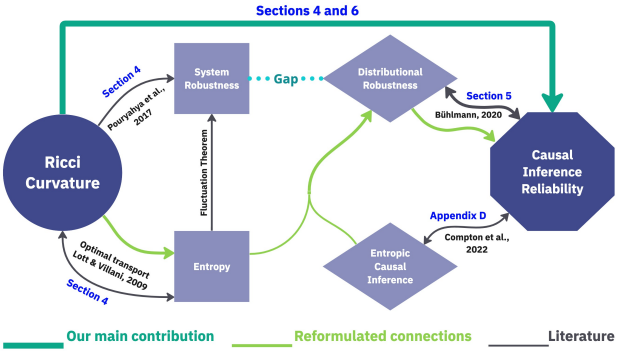

Figure 1: Visual summary of our theoretical contributions connecting the network geometry, via Ricci curvature, and causal inference. Our theory and methodological contributions are based in putting together novel results (e.g., filling the gap between different types of robustness), reformulating some of the existing links between relevant concepts in the literature, and connecting between different areas of machine learning. Arrow annotations mark the sections where the corresponding link is discussed and/or the most relevant related literature.

We present a theoretical layout of causal inference from a distributional robustness perspective, which prepares the ground for establishing the link to curvature as a robustness indicator. This connection is formally implied from our Theorem 6.1 and Corollary 6.3, which suggest 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 the ITE estimation error is lower in regions with non-negative curvature, firmly validating our theoretical foundations. Collectively, the works revisited in this paper and the foundations here developed suggest that graph curvature could offer a powerful tool for enhancing the performance of GNNs in causal inference tasks. Lastly, we propose an adjustment using the Ricci flow to flatten the network. Most empirical networks are sparse with locally tree-like structure, hence, this flattening results in increasing the Ricci curvature on a majority of edges. Thus, our proposed method leads to a remarkable gain in ITE estimation on observational networked data.

Main Contributions

The contributions of this work are threefold:

Theoretical Foundations. We establish a theoretical connection between Ricci curvature and causal inference on networks. Specifically, we show that the identification of causal parameters is more challenging in negatively curved regions of the network. This insight provides a foundational understanding of how the geometric properties of a network can influence causal analysis.

Application and Experimental Results. Guided by our theoretical findings, we report experiments showing that the estimation of the treatment effect tends to be most accurate in areas of the network with positive Ricci curvature. Demonstrating the practical applicability of our theory in real-world scenarios, our empirical results indicate that Ricci curvature effectively gauges the accuracy of causal estimates on networked treatment units *without ground-truth data*.

Methodological Contribution. We propose a novel method using the geometric Ricci flow to flatten the network, improving causal estimation on networked data. Our proposed method leads to superior performance in estimating ITE on empirical networks. Offering a new tool for enhancing the reliability of causal inference in complex networks, this method showcases how our newfound connection opens the door for utilizing graph geometry to improve GNN-based causal models.

2 Preliminaries

2.1 Causal Inference

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

$$\tau_i(x_i) \coloneqq \mathbb{E}\left[Y_i|do(t_i = t, x_i = x) - Y_i|do(t_i = t', x_i = x)\right], \quad (1)$$

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 (Pearl, 2009b). Consider an example where we are interested in the effect of a medication. Given an individual i with health and demographic features x_i, $\tau_i(x_i)$ quantifies how effective the

medication is, in average, on individuals with the same features. Following Shalit et al. (2017) and Jiang & Sun (2022), 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. This is referred to as *the fundamental problem of causal inference* (Holland, 1986). Hence, 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 (Pearl, 2003). 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)* (Imbens & Rubin, 2015; Rubin, 1980), is likely essential due to peer effects on each unit from its neighbors' features and treatments (Jiang & Sun, 2022). Four common assumptions, including ignorability and SUTVA, are formally defined in Appendix A.

2.2 Ricci Curvature

The Ricci curvature indicates how much the local geometry induced by a Riemannian metric deviates from that of a Euclidean space (Bauer et al., 2017). Extended to discrete structures such as graphs, the graph Ricci curvature captures the dispersion through an edge in its neighborhood. As visualized in Figure 2 for unweighted graphs, tree-like, grid-like, and dense neighborhoods are analogous to hyperbolic, Euclidean, and spherical spaces. Ricci curvature on graphs has been proven powerful for performing various computational tasks on GNNs (Topping et al., 2021; Southern et al., 2023; Di Giovanni et al., 2023; Liu et al., 2023). While our theoretical foundations are independent of the particular form of Ricci curvature, for the experiments in this paper we use the Ollivier-Ricci curvature (Ollivier, 2009) —a graph curvature notion rooted in optimal transport (Lott & Villani, 2009; Villani et al., 2009) (this brings yet another connection with a different arm of machine learning). A formal definition of Ollivier-Ricci curvature as well as an alternative Ricci-type curvature are included in Appendix B.

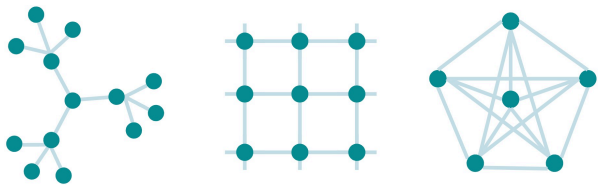

Figure 2: Graph Ricci curvatures of edges in tree-like (left), grid-like (middle), and dense (right) are analogous to hyperbolic (-), Euclidean (0), and spherical spaces (+).

3 Related Work

In this work we show 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 causal inference. While the related work is cited in each section as needed, and, as visualized in Figure 1, we connect between multiple arms of machine learning, in Appendix C we extend on the most relevant works in each aspect and point to other works in the literature. Although these works provide critical foundations and motivations for this work, none of them makes the explicit connections we developed in this paper and, in particular, the close connection between geometry/curvature, robustness, and causal inference.

4 Curvature, Robustness, and Causal Parameter Estimation Confidence

Here we present a preview of our main theoretical results and their implications. In order to do so, we first explain the link between Ricci curvature, robustness, and entropy, which critically enables the theoretical foundation leading to our results. The connection we establish between Ricci curvature and causal inference is formally stated in Section 6. The in-depth discussion in Section 6 requires additional background on our reinterpretation of causal parameters as distributionally robust estimators, which we expound in Section 5. Before delving deeper into such levels of detail, we informally describe our theoretical results in this section and discuss the core insights that follow.

Ricci Curvature and System Robustness. Ricci curvature is known to be an indicator of the robustness of systems (Pouryahya et al., 2017), due to its connection with entropy. The following result from optimal transport (Lott & Villani, 2009), offers bounds on entropy in terms of a lower bound on the Ricci curvature,

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

where $S(\cdot)$ denotes the Boltzmann entropy (Adkins, 1983), $\underline{k}$ is a lower bound on the Ricci curvature, $W_2\left(\mu_0, \mu_1\right)$ is the Wasserstein distance of order 2 between 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 (Pouryahya et al., 2017). This inequality indicates a positive correlation between Ricci curvature k_R and entropy (Pouryahya et al., 2017), i.e.,

$$\Delta S \times \Delta k_R \geq 0. \quad (3)$$

On the other hand, there is a correlation between *system robustness* and entropy. Characterized by the fluctuation decay rate (Demetrius & Manke, 2005), *system robustness* refers to the ability of the system to rapidly return to its stationary state after a perturbation. By the Fluctuation Theorem (Evans et al., 1993), there is a positive correlation between system robustness and entropy, which in turn implies that system robustness is positively correlated with curvature (Pouryahya et al., 2017), considering Equation 3.

Causality, Ricci Curvature, and the Tale of Two Robustness. Learning causal parameters could be formulated as minimizing a worst-case risk to find a *distributionally robust* estimator (Bühlmann, 2020; Rothenhäusler et al., 2021). Meanwhile, as explained above, Ricci curvature is known to be an indicator of *system robustness*. While these two notions of robustness are not equivalent, the correspondence of Ricci curvature with one and causal inference with the other suggests a possible connection (see Figure 1), which we show exists using the link between the two and entropy. This link is shown through Theorem 6.1 and Corollary 6.3. Informally, Theorem 6.1 states that positive Ricci curvature is more likely than negative Ricci curvature to further constrain exogenous variables that pose challenges to an accurate estimation of causal parameters. This implies that when the Ricci curvature in the neighborhood of a node is positive, the unobserved variables that impact the features and outcome of the node are more likely to be 'benign' — not causing additional difficulties for causal inference. This implication is proven in Corollary 6.3, which states that positive Ricci curvature corresponds to a higher probability of having a sufficiently large set of variables which are not observed in the data, for which the causal parameters learned by a model are identified. The practical consequences, summarized below, are confirmed through our experiments, and schematically illustrated in Appendix D.

The Consequences. Formally explained in Section 6, these theoretical results indicate that neighborhoods of the network with negative Ricci curvature tend to pose greater challenges to the identification of causal parameters, leading to less reliable causal estimates. Therefore, Ricci curvature could assess the confidence in causal estimation on networks based on data alone, as empirically confirmed by our experiments in Section 7. Next, we present a reinterpretation of previously established results to build up the steps that enable us to discuss our theoretical findings in depth.

5 Invariance and Distributional Robustness of Causal Parameters

In this section, we 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, revisiting the work by Bühlmann (2020). We briefly discuss the problem formulation and results which explain that the minimization loss function upholding causal assumptions coincides with the one that leads to *distributional robustness* (remember that Ricci curvature relates to *system robustness*; we connect between the two in this work). A more detailed discussion is included in Appendix E. 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 a pedagogic example, the concepts regarding distributional robustness are not limited to the specific settings of this example.

Adopting the notation in Bühlmann (2020), 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. Given a neural network $g \equiv g_\theta$ parameterized by θ, this can be formulated as solving

$$\theta_{\text{causal}} = \operatorname*{argmin}_{\theta} \max_{e \in \mathcal{F}} \mathcal{L}\left(Y^e, g_\theta(X^e)\right), \tag{4}$$

where $\mathcal{L}$ is the loss function (Bühlmann, 2020). If we can find a subset of covariate indices $S \subset \{1, \ldots, n_X\}$ such that $\mathcal{L}\left(Y^e, g_\theta(X^e_S)\right)$ is invariant with respect to $e \in \mathcal{F}$, to find θ_{causal} it suffices to minimize $\mathcal{L}\left(Y^e, g_\theta(X^e_S)\right)$ for any observable e. This requirement, although abstract, elucidates the invariance of causal estimation. When S is the set of indices corresponding to the causal parents of Y^e, the relationship between Y^e and X^e_S does not depend on the environment and hence the causal parameters minimize $\mathcal{L}\left(Y^e, g_\theta(X^e_S)\right)$. Conditions which allow us to put this on a computational footing, detailed in Appendix E, are contingent on restrictive assumptions such as absence of hidden confounders.

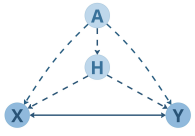

Figure 3: Causal graph for the anchor regression model. A, H, X, and Y denote the anchor, hidden confounders, covariates, and outcome.

To relax this assumption, consider the anchor regression model, involving an anchor variable A, covariates X, out-

come Y, and hidden confounders H, with the causal graph shown in Figure 3. The estimand for linear anchor regression, here used for illustration purposes, can be computed as the minimizer of the loss $\mathcal{L}_A$ given by,

$$\begin{aligned}\mathcal{L}_A(\gamma) =& \mathbb{E}\left[\left((I - P_A)\left(Y - X^T b\right)\right)^2\right] \\ &+ \gamma \mathbb{E}\left[\left(P_A(Y - X^T b)\right)^2\right], \end{aligned} \tag{5}$$

where P_A is the projection operator onto the column space of A. The anchor variable —the root in the causal graph— could be thought of as the determiner of the environment in Equation 4, thus the second term in Equation 5 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 E, that

$$\mathcal{L}_A(\gamma) = \sup_{v \in \mathcal{C}_\gamma} \mathbb{E}\left[\left(Y^v - (X^v)^T b\right)^2\right],$$

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 $\mathcal{O}(\gamma)$ (Bühlmann, 2020; Rothenhäusler et al., 2021). 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.

6 Curvature and Causal Inference

Equations 2 and 3 together with the Fluctuation Theorem (Evans et al., 1993; Pouryahya et al., 2017) provide the basis for the positive correlations pointed to in Section 4 between entropy, Ricci curvature, and *system robustness*. The discussion in Section 5 on the other hand, formulates the problem of learning causal parameters in terms of *distributionally robust* optimization. Having built intuition into these relationships in prior sections, we now formally present our core theoretical results elucidating the connection between Ricci curvature and causal inference, which was explained informally in Section 4. These results rigorously substantiate the theoretical grounds for using Ricci curvature as a practical proxy for the reliability of causal estimation on networked units. Moreover, this theoretical connection will, in turn, inform a methodological remedy utilizing the geometric Ricci flow to improve causal estimates on networks.

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 $Y_i = f_i(X_i, E_i)$, where $E \perp\!\!\!\perp X$ denotes the exogenous variables. Suppose an alternative model $X_i = g_i(Y_i, \tilde{E}_i)$, with alternative exogenous variables $\tilde{E}_i \perp\!\!\!\perp Y$ fits the data. Assume that the Ricci curvature corresponding to X_i is bounded from below by $\underline{k}_i$, for $i \in \{1, 2\}$. Then, under the assumptions stated in Appendix G, where we also provide the proof, the following holds:

Theorem 6.1. *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 entropy of $\tilde{E}_1$.*

The proof of this theorem, Appendix G, uses a result from entropic causal inference (Appendix F) (Kocaoglu et al., 2017; Compton et al., 2020; 2022). This provides a critical link between entropy and causal inference, allowing us to establish the connection with curvature using Equation 2.

Theorem 6.1 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. The core insight from Theorem 6.1 is that a non-negative Ricci curvature corresponds to a wrong model that admits a smaller class of exogenous variables. When the Ricci curvature is non-negative, the exogenous variables corresponding to a fitting wrong model tend to require a larger entropy lower bound, hence, the set of exogenous variables which could make the wrong model fit the observed data is smaller. This is formally stated in Corollary 6.3 below, which specifically shows the implications of Ricci curvature for causal identification. This property is *intrinsic to the network* and independent of the exact algorithm used to estimate the causal relationships.

6.1 Ricci Curvature and Causal Identification

In order to explicitly show the implications of Theorem 6.1 for causal identification, we briefly formalize the definition of identified causal parameters over a family of unobserved variables. Consider the problem of learning a causal parameter characterizing the relationship between X and Y from data with population cumulative distribution function (CDF) $F_{X,Y,U} \in \mathcal{F}_{X,Y,U}$, where X and Y are observed and U unobserved variables, and $\mathcal{F}_{X,Y,U}$ is a family of

joint CDFs. Let $F_{X,Y} \in \mathcal{F}_{X,Y}$ denote the observed CDF of (X, Y). For the purpose of this formalization, we shall abstract away from the sampling process, hence consider $F_{X,Y}$ to be the population distribution of the observed variables, and let $h : \mathcal{F}_{X,Y,U} \to \mathcal{F}_{X,Y}$ denote the mapping that gives $F_{X,Y} = h(F_{X,Y,U})$. We are interested in the causal mechanism governing the population as described by a function $\vartheta : \mathcal{F}_{X,Y,U} \to \Theta$ that maps any distribution to its corresponding causal parameter $\theta \in \Theta$. Note that an estimation method we use to learn $\theta(F_{X,Y,U})$, such as a neural network in a deep learning model estimating a causal effect, has access only to $F_{X,Y}$. The following definition formally states what it means for the causal parameter to be identified over a set of possible unobserved variables $\mathcal{U}$.

Definition 6.2. We say the parameter is *point identified over* $\mathcal{U}$, if $\Theta^{\mathcal{U}}$, defined below, is a singleton,

$$\Theta^{\mathcal{U}} := \{\, \theta \in \Theta \;:\; \exists\, (U, F_{X,Y,U}) \in \mathcal{U} \times \mathcal{F}_{X,Y,U} \text{ s.t. } F_{X,Y} = h(F_{X,Y,U}) \text{ and } \theta = \vartheta(F_{X,Y,U}) \,\}\ . \quad (6)$$

$\Theta^{\mathcal{U}}$ is *the identified set over* $\mathcal{U}$. We further say that the parameter is *partially identified over* $\mathcal{U}$ if $\Theta^{\mathcal{U}} \subset \Theta$, and *completely not identified over* $\mathcal{U}$ if $\Theta^{\mathcal{U}} = \Theta$.

In other words, the identified set over $\mathcal{U}$ is the set of all parameters that describe the population distributions which are observationally equivalent.

Assume $\vartheta(\cdot)$ is injective, i.e., the causal parameter of interest are expressive of the population distribution. Consider X_i, Y_i, and $\underline{k}_i$ for $i \in \{1, 2\}$ in the setup for Theorem 6.1, and let θ_i denote the corresponding causal parameter. Suppose further that the variables not observed in the data include the exogenous variables E_i and $\tilde{E}_i$ as described for Theorem 6.1. Let $(\mathcal{U}, 2^{\mathcal{U}})$ be a measurable space equipped with the measure μ, where $\mathcal{U}$ denotes the set of all possible unobserved variables, i.e., $\left\{E_1, E_2, \tilde{E}_1, \tilde{E}_2\right\} \subseteq \mathcal{U}$. Under the conditions for Theorem 6.1, we have the following:

Corollary 6.3. *Let $\bar{\mathcal{U}}_i$ denote the maximal set of unobserved variables over which θ_i is point-identified. If $\underline{k}_1 < 0 \leq \underline{k}_2$, then $\mathbb{P}\left[\mu\left(\bar{\mathcal{U}}_2\right) > \upsilon\right] \geq \mathbb{P}\left[\mu\left(\bar{\mathcal{U}}_1\right) > \upsilon\right]$ for a constant υ.*

In words, the probability that the causal parameter is point identified over a set at least as large as υ is weakly higher for the case of non-negative curvature. The proof of the corollary is provided in Appendix G.

In light of the discussion in Section 5, Corollary 6.3 implies that positive Ricci curvature is more likely to allow the causal parameter to be point-identified over a larger class of unobserved variables, such as shift perturbations. Shedding light on a more direct implication of Ricci curvature for causal inference, this corollary suggests that learning identified causal parameters tends to be more challenging where the curvature is negative. Theorem 6.1 and its corollary ultimately suggest that a positive Ricci curvature is expected to correspond to a lower error in estimating causal parameters.

Intermezzo. To recap, Theorem 6.1 and Corollary 6.3 formally establish the connection between Ricci curvature and causal inference, demonstrating for the first time how the geometry of a network influences causal estimations. They show that positive curvature leads to causal parameters that are identified for a larger class of perturbations, which enhances distributional robustness, making the worst-case risk minimization in a system under perturbation a less challenging problem. 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.

6.2 Ricci Flow Adjustment for Improving Causal Effect Estimates

Informed by the theoretical connection between Ricci curvature and causal inference, we propose, if needed, to improve treatment effect estimates on network data using the discrete geometric Ricci flow (Jin et al., 2008; Ni et al., 2019). 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. This is, in fact, equivalent to the heat equation, considering the formulation of the Ricci flow as a scaling of the Laplacian of the metric tensor (Chow & Knopf, 2004). As a result, just as the temperature evolves towards a more uniform distribution under heat diffusion, the Ricci flow evolves to a uniform distribution of curvature (Hamilton, 1988; Jin et al., 2008). Similarly, its discrete analog iteratively updates the edge weights toward a flatter network, without changing unweighted connectivities in the graph.

Let w_{vu} be the weight on the edge $(v, u) \in E$, κ_{vu} its Ricci curvature, and $d(v, u)$ the geodesic distance. At iteration i, w_{vu} evolves under the discrete Ricci flow as

$$w_{vu}^{i+1} = (1 - \kappa_{vu})\, d(v, u)^i. \quad (7)$$

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 input 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 effects.

Since the anticipated error reduction results from eliminating negative curvature values, this Ricci flow adjustment is expected to be more effective in networks with larger proportions of neighborhoods with highly negative Ricci curvature.

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.

7.1 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 (Kaddour et al., 2021; Jiang & Sun, 2022; Chu et al., 2023). Jiang & Sun (2022) 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, the ITE formulation, the training loss, and implementation are included in Appendix H and Appendix I. While our experiments are primarily aimed at demonstrating our theoretical results in practice, and evaluating the performance of NetEst and our proposed enhancement of it in estimating ITE, the theory and methodology we develop in this paper concern the intrinsic graph structure of the treatment units; they neither depend on the estimation method nor are specific to the estimand it estimates. Additionally, in Section 7.3 we compare our results with several baseline causal effect estimation methods. These baselines include CFR (Shalit et al., 2017), TARNet (Shalit et al., 2017), NetDeconf (Guo et al., 2020), T-Learner and X-Learner (Künzel et al., 2019) with random forest (RF) regressors, and T-Learner and X-Learner implemented using a GNN encoder followed by a multilayer perceptron. 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) $\epsilon_{PEHE} := \sqrt{\frac{1}{N}\sum_{v\in V}(\tau_v - \hat{\tau}_v)^2}$, where τ_v and $\hat{\tau}_v$ denote the true and estimated ITEs for node v.

Consistent with standard practice in causal machine learning, we use semi-synthetic datasets, namely empirically observed network structures and features with simulated treatments and potential outcomes (Hill, 2011; Shalit et al., 2017; Veitch et al., 2019; Guo et al., 2020; Ma et al., 2021; Jiang & Sun, 2022). Following the original experiments on NetEst, we use the BlogCatalog (BC) and Flickr datasets (Guo et al., 2020; Ma et al., 2021). We supplement our experiments with numerous other empirical networks. All datasets are described in Appendix J.

7.2 Ricci Curvature as an Indicator of the Treatment Effect Estimation Error

We demonstrate the implications of Theorem 6.1 and Corollary 6.3 by inspecting the joint distribution of ε_{ITE} and the Ricci curvature, which in turn, provides empirical evidence for employing Ricci curvature to evaluate the reliability of causal estimates. Ollivier-Ricci curvature is inherently an edge-based measure (see Appendix B). To quantify the curvature of the region surrounding a node, we aggregate the curvature of its incident edges by taking their sum. We then compute the empirical joint distribution of the sum of edge curvatures and ε_{ITE} for each node. The joint distributions in Figure 4 show a negative correlation between Ricci curvature and ε_{ITE}, indicating that treatment effect estimations are more reliable in regions with non-negative curvature. Additional experiments in Appendix K show that these results are consistent not only across different datasets but also different notions of Ricci curvature on networks, supporting our theoretical results on the connection between this geometric feature and the reliability of causal estimation.

7.3 Geometric 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 6 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 geometric method, we apply this adjustment to the input graph of NetEst. We refer to the modified method as *f-NetEst* (for flow). 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 K.2) further confirms that the Ricci flow adjustment leads to more accurate ITE estimations.

Table 1 also reports the performances of several baseline models. 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 three baseline models featuring GNN encoders: T-Learner+GNN, X-Learner+GNN, and NetDeconf. These additional experiments confirm that our proposed modification results in a reduction in the treatment effect estimation error in most cases across other GNN-based models as well.

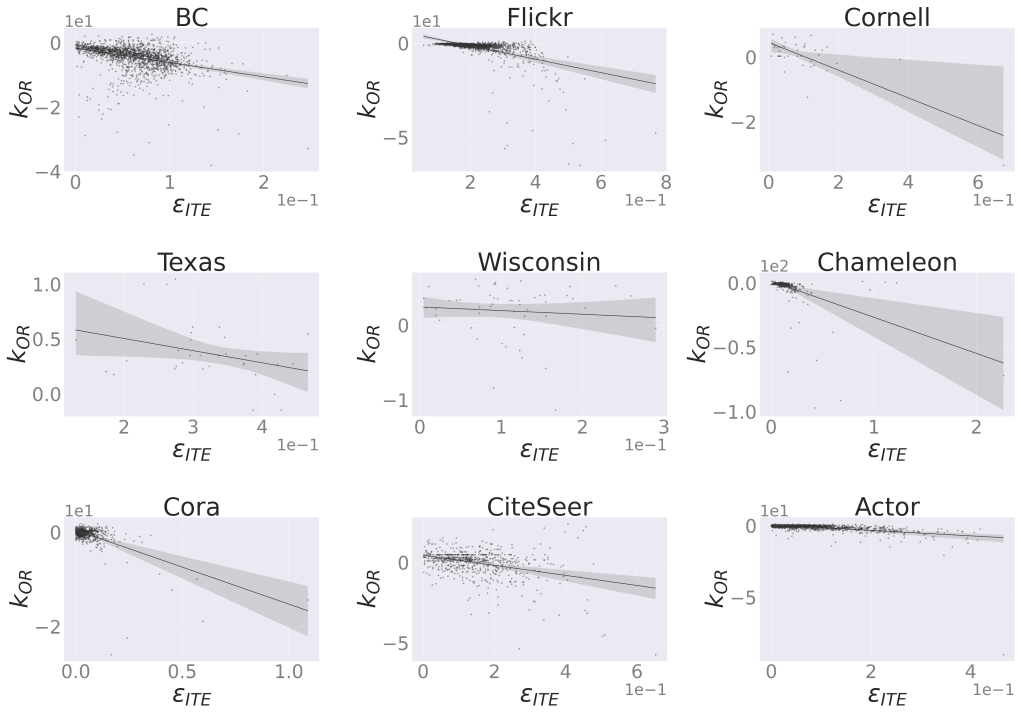

Figure 4: Joint distributions of the sum of Ollivier-Ricci curvatures in the neighborhood of each node and the estimation error of ITE for that node. The distributions for nine different networks are shown (all datasets are described in Appendix J). The regression lines with the corresponding 95% confidence intervals are marked on the plots. Additional experiments in Appendix K show that these results are consistent also for different notions of Ricci curvature, supporting our theory.

Table 1: ϵ_{PEHE} for nine datasets, comparing the proposed f-NetEst against NetEst and several 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, respectively, relative gains greater than 5% and less than -5% from the Ricci flow adjustment. Smaller ϵ_{PEHE} is better.

	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**

8 Conclusions and Limitations

We delved into the unexplored territory of leveraging geometry for causal inference 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 assess the reliability of causal effect estimations on networked data, empirically validating that positive curvature regions lead to more accurate results, and showing that curvature can serve as a practical measure of confidence in causal estimates with-

out any ground-truth data. This property is intrinsic to the network and independent of the method exploited for causal estimation. We then proposed using the geometric Ricci flow to enhance treatment effect estimation on networked data, achieving superior performance through flattening the edges.

To the best of our knowledge, this work is the first to formally establish the connection between graph curvature, as a proxy to geometry, and network causal inference. This opens new avenues for applications of graph geometry in causal inference, as well as neighboring tasks such as transfer learning, out-of-distribution generalization, and domain adaptation. The insights and tools presented in this paper lay the groundwork for future exploration, paving the way for innovative approaches to network causal analysis.

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 inference.

Acknowledgements

This work was partially supported by ONR, NGA, NSF, and Apple. The authors thank anonymous reviewers for their constructive comments.

Impact 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 as well as with the exploitation of ML for critical applications such as healthcare.

References

Adkins, C. J. Equilibrium Thermodynamics. Cambridge University Press, 1983.

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. Advances in Neural Information Processing Systems, 30, 2017.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

Atz, K., Grisoni, F., and Schneider, G. Geometric deep learning on molecular representations. Nature Machine Intelligence, 3(12):1023–1032, 2021.

Bauer, F., Hua, B., Jost, J., Liu, S., and Wang, G. The geometric meaning of curvature: Local and nonlocal aspects of ricci curvature. Modern Approaches to Discrete Curvature, pp. 1–62, 2017.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34 (4):18–42, 2017.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478, 2021.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. Spectral networks and locally connected networks on graphs. International Conference on Learning Representations, 2014.

Bühlmann, P. Invariance, causality and robustness. Statistical Science, 35(3):404–426, 2020.

Cao, W., Yan, Z., He, Z., and He, Z. A comprehensive survey on geometric deep learning. IEEE Access, 8: 35929–35949, 2020.

Chow, B. and Knopf, D. The Ricci Flow: An Introduction, volume 1. American Mathematical Society, 2004.

Chu, Z., Huang, J., Li, R., Chu, W., and Li, S. Causal effect estimation: Recent advances, challenges, and opportunities. arXiv preprint arXiv:2302.00848, 2023.

Compton, S., Kocaoglu, M., Greenewald, K., and Katz, D. Entropic causal inference: Identifiability and finite sample results. In Advances in Neural Information Processing Systems, volume 33, pp. 14772–14782. Curran Associates, Inc., 2020.

Compton, S., Greenewald, K., Katz, D. A., and Kocaoglu, M. Entropic causal inference: Graph identifiability. In International Conference on Machine Learning, pp. 4311–4343. PMLR, 2022.

Cristali, I. and Veitch, V. 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.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems, 26, 2013.

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

Di Giovanni, F., Giusti, L., Barbero, F., Luise, G., Lio, P., and Bronstein, M. M. 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.

Do Carmo, M. P. and Flaherty Francis, J. Riemannian Geometry, volume 6. Springer, 1992.

Evans, D. J., Cohen, E. G. D., and Morriss, G. P. Probability of second law violations in shearing steady states. Physical Review Letters, 71(15):2401, 1993.

Forastiere, L., Airoldi, E. M., and Mealli, F. Identification and estimation of treatment and interference effects in observational studies on networks. Journal of the American Statistical Association, 116(534):901–918, 2021.

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

Frauen, D. and Feuerriegel, S. Estimating individual treatment effects under unobserved confounding using binary instruments. In International Conference on Learning Representations, 2022.

Gong, S., Bahri, M., Bronstein, M. M., and Zafeiriou, S. Geometrically principled connections in graph neural networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11415–11424. IEEE, 2020.

Guo, R., Li, J., and Liu, H. 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.

Hägele, A., Rothfuss, J., Lorch, L., Somnath, V. R., Schölkopf, B., and Krause, A. Bacadi: Bayesian causal discovery with unknown interventions. In International Conference on Artificial Intelligence and Statistics, pp. 1411–1436. PMLR, 2023.

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

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

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. Journal of Causal Inference, 6(2):20170016, 2018.

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

Holland, P. W. Statistics and causal inference. Journal of the American Statistical Association, 81(396):945–960, 1986.

Imbens, G. W. and Rubin, D. B. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

Immer, A., Schultheiss, C., Vogt, J. E., Schölkopf, B., Bühlmann, P., and Marx, A. On the identifiability and estimation of causal location-scale noise models. In International Conference on Machine Learning, pp. 14316–14332. PMLR, 2023.

Jiang, S. and Sun, Y. 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.

Jin, M., Kim, J., Luo, F., and Gu, X. Discrete surface ricci flow. IEEE Transactions on Visualization and Computer Graphics, 14(5):1030–1043, 2008.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In International Conference on Machine Learning, pp. 3020–3029. PMLR, 2016.

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

Kaddour, J., Zhu, Y., Liu, Q., Kusner, M. J., and Silva, R. Causal effect inference for structured treatments. In Advances in Neural Information Processing Systems, volume 34, pp. 24841–24854. Curran Associates, Inc., 2021.

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

Ke, N. R., Chiappa, S., Wang, J. X., Bornschein, J., Goyal, A., Rey, M., Weber, T., Botvinick, M., Mozer, M. C., and Rezende, D. J. Learning to induce causal structure. In

International Conference on Learning Representations, 2022.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

Kocaoglu, M., Dimakis, A., Vishwanath, S., and Hassibi, B. Entropic causal inference. In AAAI Conference on Artificial Intelligence, volume 31:, 2017.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10):4156–4165, 2019.

Lin, Y., Lu, L., and Yau, S.-T. Ricci curvature of graphs. Tohoku Mathematical Journal, 63(4):605–627, 2011.

Lin, Y., Dong, H., Wang, H., and Zhang, T. Bayesian invariant risk minimization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16021–16030. IEEE, 2022.

Liu, Y., Zhou, C., Pan, S., Wu, J., Li, Z., Chen, H., and Zhang, P. 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.

Lott, J. and Villani, C. Ricci curvature for metric-measure spaces via optimal transport. Annals of Mathematics, pp. 903–991, 2009.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Lu, C., Wu, Y., Hernández-Lobato, J. M., and Schölkopf, B. Invariant causal representation learning for out-of-distribution generalization. In International Conference on Learning Representations, 2021.

Luo, Y., Peng, J., and Ma, J. When causal inference meets deep learning. Nature Machine Intelligence, 2(8):426–427, 2020.

Ma, J., Guo, R., Chen, C., Zhang, A., and Li, J. 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.

Ma, Y. and Tresp, V. Causal inference under networked interference and intervention policy enhancement. In International Conference on Artificial Intelligence and Statistics, pp. 3700–3708. PMLR, 2021.

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

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. 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.

Ni, C.-C., Lin, Y.-Y., Luo, F., and Gao, J. Community detection on networks with ricci flow. Scientific Reports, 9(1):1–12, 2019.

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

Pawlowski, N., Coelho de Castro, D., and Glocker, B. Deep structural causal models for tractable counterfactual inference. In Advances in Neural Information Processing Systems, volume 33, pp. 857–869. Curran Associates, Inc., 2020.

Pearl, J. Statistics and causal inference: A review. Test, 12: 281–345, 2003.

Pearl, J. Causal inference in statistics: An overview. Statistics Surveys, 3:96–146, 2009a.

Pearl, J. Causality: Models, Reasoning and Inference. Cambridge University Press, 2009b.

Pei, H., Wei, B., Chang, K. C.-C., Lei, Y., and Yang, B. Geom-gcn: Geometric graph convolutional networks. In International Conference on Learning Representations, 2019.

Perry, R., Von Kügelgen, J., and Schölkopf, B. 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.

Peters, J., Bühlmann, P., and Meinshausen, N. 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.

Pouryahya, M., Mathews, J., and Tannenbaum, A. Comparing three notions of discrete ricci curvature on biological networks. arXiv preprint arXiv:1712.02943, 2017.

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

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(2):215–246, 2021.

Rothman, K. J. and Greenland, S. Causation and causal inference in epidemiology. American Journal of Public Health, 95(S1):S144–S150, 2005.

Rozemberczki, B., Allen, C., and Sarkar, R. Multi-scale attributed node embedding. Journal of Complex Networks, 9(2):cnab014, 2021.

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

Samal, A., Sreejith, R., Gu, J., Liu, S., Saucan, E., and Jost, J. Comparative analysis of two discretizations of ricci curvature for complex networks. Scientific Reports, 8(1): 8650, 2018.

Sandhu, R., Georgiou, T., Reznik, E., Zhu, L., Kolesov, I., Senbabaoglu, Y., and Tannenbaum, A. Graph curvature for differentiating cancer networks. Scientific Reports, 5 (1):12323, 2015.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning, pp. 3076–3085. PMLR, 2017.

Shi, C., Veitch, V., and Blei, D. M. Invariant representation learning for treatment effect estimation. In Uncertainty in Artificial Intelligence, pp. 1546–1555. PMLR, 2021.

Southern, J., Wayland, J., Bronstein, M. M., and Rieck, B. Curvature filtrations for graph generative model evaluation. In Advances on Neural Information Processing Systems, 2023.

Sreejith, R., Mohanraj, K., Jost, J., Saucan, E., and Samal, A. Forman curvature for complex networks. Journal of Statistical Mechanics: Theory and Experiment, 2016(6): 063206, 2016.

Srinivas, S., Matoba, K., Lakkaraju, H., and Fleuret, F. Efficient training of low-curvature neural networks. In Advances in Neural Information Processing Systems, volume 35, pp. 25951–25964. Curran Associates, Inc., 2022.

Tannenbaum, A., Sander, C., Zhu, L., Sandhu, R., Kolesov, I., Reznik, E., Senbabaoglu, Y., and Georgiou, T. Graph curvature and the robustness of cancer networks. arXiv preprint arXiv:1502.04512, 2015.

Topping, J., Di Giovanni, F., Chamberlain, B. P., Dong, X., and Bronstein, M. M. Understanding over-squashing and bottlenecks on graphs via curvature. In International Conference on Learning Representations, 2021.

van der Laan, M. J. Causal inference for networks. Technical report, U.C. Berkeley Division of Biostatistics, 2012.

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

Veitch, V., Wang, Y., and Blei, D. Using embeddings to correct for unobserved confounding in networks. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

Villani, C. et al. Optimal transport: old and new, volume 338. Springer, 2009.

Weber, M., Saucan, E., and Jost, J. Characterizing complex networks with forman-ricci curvature and associated geometric flows. Journal of Complex Networks, 5(4): 527–550, 2017.

Weichwald, S. and Peters, J. Causality in cognitive neuroscience: concepts, challenges, and distributional robustness. Journal of Cognitive Neuroscience, 33(2):226–247, 2021.

Wein, S., Malloni, W. M., Tomé, A. M., Frank, S. M., Henze, G.-I., Wüst, S., Greenlee, M. W., and Lang, E. W. A graph neural network framework for causal inference in brain networks. Scientific Reports, 11(1):8061, 2021.

Yang, Z., Cohen, W., and Salakhudinov, R. Revisiting semi-supervised learning with graph embeddings. In International Conference on Machine Learning, pp. 40–48. PMLR, 2016.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. ACM Transactions on Knowledge Discovery from Data, 15(5):1–46, 2021.

Ye, Z., Liu, K. S., Ma, T., Gao, J., and Chen, C. Curvature graph network. In International Conference on Learning Representations, 2020.

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

A Causal Identification Assumptions

A set of standard 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 their description as well as two other common assumptions (Forastiere et al., 2021; Imbens & Rubin, 2015; Rosenbaum & Rubin, 1983; Rubin, 1980),

- **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 effects (Jiang & Sun, 2022). However, Jiang & Sun (2022) 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 (Jiang & Sun, 2022),

- **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 Ricci Curvature Notions on Graphs

The Ricci curvature indicates deviation from the Euclidean space (Bauer et al., 2017; Do Carmo & Flaherty Francis, 1992; Pouryahya et al., 2017). On graphs, this translates to measuring how much the neighborhood of an edge differs from a grid. We used the Ollivier-Ricci curvature (Ollivier, 2009) for the experiments reported in the main text of the paper. The Forman-Ricci curvature (Forman, 2003) is an alternative notion of Ricci curvature on graphs, which we use, in addition to the Ollivier-Ricci curvature, in supplementary experiments included in Appendix K.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

$$\kappa_{OR}(v, u) := 1 - \frac{W_1(\mu_v, \mu_u)}{d_G(v, u)}, \tag{8}$$

where $d_G(.)$ is a distance metric on V and W_1 denotes the 1-Wasserstein distance (Jost & Liu, 2014; Lin et al., 2011). 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

$$\kappa_{FR}(v,u) \coloneqq w_{vu}\left[\frac{w_v}{w_{vu}} + \frac{w_u}{w_{vu}} - \sum_{(v',u')\in N_v\times N_u}\left(\frac{w_v}{\sqrt{w_{vu}w_{vv'}}} + \frac{w_u}{\sqrt{w_{vu}w_{uu'}}}\right)\right], \tag{9}$$

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 (Sreejith et al., 2016; Weber et al., 2017). 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.

In this paper we base our main experiments on the Ollivier-Ricci curvature for its proven effectiveness (see, e.g., Southern et al. (2023)) and link to optimal transport —an essential building block of our theoretical framework. However, with a combinatorial formulation, Forman-Ricci curvature is a more computationally efficient notion with linear computational complexity. Meanwhile, the complexity of computing the Ollivier-Ricci curvature of each edge (v, u) is $\mathcal{O}((d_v \times d_u)^c)$, where d_v and d_u are the degrees, and c is the exponent in the cost of matrix multiplication (2.37 by current fastest algorithms). In the most expensive case (fully-connected graph) this is $\mathcal{O}(n^{4.74})$, n is the number of nodes. Empirical networks are typically sparse with an average degree of constant order, d, the cost becomes $\mathcal{O}(d^{4.74})$ (Ye et al., 2020). Approximating the Wasserstein distance by Sinkhorn distances (Cuturi, 2013) reduces this complexity to a near-linear order in d (Altschuler et al., 2017; Ye et al., 2020).

C Related Work

In this work we formally 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) (Bühlmann, 2020; Heinze-Deml et al., 2018; Peters et al., 2016; Shi et al., 2021) and invariant risk minimization (IRM) (Arjovsky et al., 2019; Bühlmann, 2020; Lin et al., 2022; Shi et al., 2021). Bühlmann (2020) formally describes how IRM can lead to a *distributionally* robust estimator while imposing causal identification assumptions. Remember that curvature is related to *system* robustness, and connecting it with distributionally robustness as we did led to the foundations developed in this paper.

Geometric Deep Learning. Geometric tools have been instrumental to recent advances on GNNs (Bronstein et al., 2017; 2021; Bruna et al., 2014; Gong et al., 2020). Discrete Ricci curvatures on graphs, in particular, are well-established measures with roots in Riemannian geometry (Ollivier, 2009; Samal et al., 2018; Sandhu et al., 2015), with applications for GNNs (Southern et al., 2023; Topping et al., 2021). The connection between Ricci curvature and entropy is known from the optimal transport literature (Lott & Villani, 2009), based on which, (Pouryahya et al., 2017) uses Ricci curvature as a measure of *system* robustness. Moreover, curvature has been used by Srinivas et al. (2022) 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 (Compton et al., 2020; 2022; Kocaoglu et al., 2017).

Deep Learning for Causal Inference. Deep learning methods have had success in estimating treatment effect (Louizos et al., 2017; Shalit et al., 2017), counterfactual inference (Johansson et al., 2016; Pawlowski et al., 2020), and other problems in causal inference (Frauen & Feuerriegel, 2022; Hägele et al., 2023; Immer et al., 2023; Ke et al., 2022; Lu et al., 2021; Luo et al., 2020; Perry et al., 2022). Causal effect estimation on networked data on the other hand, is known to be notoriously challenging (van der Laan, 2012; Zheleva & Arbour, 2021). Various methods have been proposed for estimating the causal effect in structured data which violate traditional identification assumptions (Cristali & Veitch, 2022; Guo et al., 2020; Harada & Kashima, 2021; Jiang & Sun, 2022; Kaddour et al., 2021; Ma & Tresp, 2021). For instance, Guo et al. (2020) uses a *Network Deconfounder* to learn a representation of hidden confounders from the data, Kaddour et al. (2021) proposes an effect decomposition, and Veitch et al. (2019) and Cristali & Veitch (2022) use the embeddings to deal with unobserved confounders and the homophily effect. Another approach, taken in Jiang & Sun (2022); Ma & Tresp (2021); Harada & Kashima (2021), 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.

D The Consequences of Ricci Curvature for Causal Prameter Estimation on Networks

As we discuss in Sections 4 and 6, our theoretical results imply that negative Ricci curvature poses greater challenges to identifying causal parameters, hence is expected to lead to larger error in estimating them. In this appendix we visualize this consequence of our results, schematically illustrated it in Figure 5.

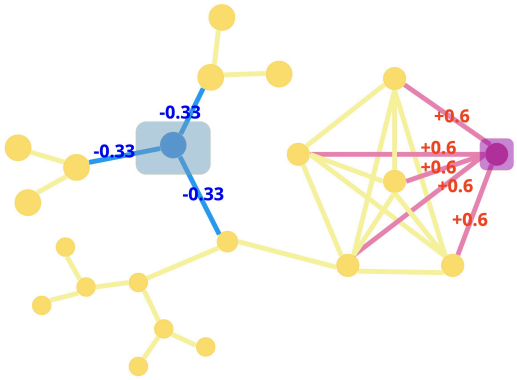

Figure 5: Our theoretical results suggest that positive/negative Ricci curvature correspond to smaller/larger error in estimating causal parameters. The Ollivier-Ricci curvature values of the edges in the neighborhoods of the blue and pink nodes are shown in this illustration, and the sizes of the shaded shapes around them mark our anticipated relative error of estimating causal parameters on them.

E Causal Inference, Invariance, and Distributional Robustness

E.1 Causal Inference as Risk Minimization

Following Bühlmann (2020), we formalize the derivation of causal quantities from statistical quantities as a worst-case risk minimization problem. Adopting the notation in Bühlmann (2020), 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 (Bühlmann, 2020), hereafter referred to by condition $\mathbf{C}$,

$$\theta_{\text{causal}} = \underset{b}{\operatorname{argmin}} \max_{e \in \mathcal{F}} \mathbb{E}\left[\left(Y^e - (X^e)^T b\right)^2\right]. \tag{10}$$

E.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 E.1 (Bühlmann (2020))**.** $\mathbf{A}_S(\mathcal{G})$ 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 10 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 condition $\mathbf{C}$ is satisfied (Bühlmann, 2020). This can be formalized as satisfying $\mathbf{B}(\mathcal{F})$, where $\mathbf{B}(\mathcal{G})$ is,

Definition E.2 (Bühlmann (2020))**.** $\mathbf{B}(\mathcal{G})$ is defined as the property that $\left\{p_{\epsilon^e} | e \in \mathcal{G} \wedge Y^e = f\left(X^e_{\text{pa}(Y)}, \epsilon^e\right)\right\}$ is a singleton, where f determines the SEM, ϵ^e is independent of $X^e_{\text{pa}(Y)}$, 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 (2020), 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 distributional robustness of an estimator in a regression problem which allows for relaxing these constraints.

E.3 Distributional Robustness and Causal Inference

One common situation where condition $\mathbf{C}$ fails is the presence of hidden confounders. We can use anchor regression (Rothenhäusler et al., 2021) to relax condition $\mathbf{C}$ 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 3 in the core manuscript. The anchor variable could be considered as an environment that is not constrained by condition $\mathbf{C}$. The corresponding linear SEM is then

$$\begin{bmatrix} X \\ Y \\ H \end{bmatrix} = B \begin{bmatrix} X \\ Y \\ H \end{bmatrix} + \epsilon + MA, \tag{11}$$

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,

$$Y = X^T\beta + H^T\alpha + A^T\xi + \epsilon_Y. \tag{12}$$

Since the anchor variable is a root 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 12, are given by

$$\beta_A(\gamma) = \underset{b}{\text{argmin}} \left\{ \mathbb{E}\left[\left((I - P_A)(Y - X^Tb)\right)^2\right] + \gamma \mathbb{E}\left[\left(P_A(Y - X^Tb)\right)^2\right] \right\}, \tag{13}$$

$$\hat{\beta}_A(\gamma) = \underset{b}{\text{argmin}} \left\{ \frac{1}{n} \left\|(I - \Pi_A)(\mathbf{Y} - \mathbf{X}b)\right\|_2^2 + \frac{\gamma}{n} \left\|\Pi_A(\mathbf{Y} - \mathbf{X}b)\right\|_2^2 \right\}, \tag{14}$$

where the second term in the objective functions encourages the residuals to be orthogonal to A (Bühlmann, 2020). 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 $W_\gamma \coloneqq I - \left(1 - \sqrt{\gamma}\right)\Pi_A$. 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 11. The SEM under perturbation can be written as

$$\begin{bmatrix} X^v \\ Y^v \\ H^v \end{bmatrix} = B \begin{bmatrix} X^v \\ Y^v \\ H^v \end{bmatrix} + \epsilon + v. \tag{15}$$

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 (Bühlmann, 2020; Rothenhäusler et al., 2021),

Proposition E.3. *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 E.3 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 13 (Bühlmann, 2020; Rothenhäusler et al., 2021),

Theorem E.4. *For any $b \in \mathbb{R}^m$*

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

Corollary E.5, which states that $\beta_A(\gamma)$ minimizes a worst case risk over the class of shift perturbations $\mathcal{C}_\gamma$, follows trivially considering Equation 13:

Corollary E.5. $\beta_A(\gamma) = \operatorname{argmin}_{b \in \mathbb{R}^m} \sup_{v \in \mathcal{C}_\gamma} \mathbb{E}\left[\left(Y^v - (X^v)^T b\right)^2\right]$.

Recall that the second term in the objective function of the anchor regression estimand in Equation 13 is essentially a causal regularization term that encourages the invariance of the residuals with respect to the environment. Theorem E.4 and Corollary E.5 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. As mentioned before, to connect causality with geometry via Ricci curvature, we developed a connection with system robustness. Entropic causal inference, as presented next, opened the door to that.

F Entropic Causal Inference

Entropic causal inference is a framework that aims to learn the causal graph between variables from observational data, using an Occam's razor-type principle (Kocaoglu et al., 2017). This approach seeks the information-theoretically simplest structural explanation of the data to infer causality (Compton et al., 2020). The central claim is that the true causal structural model is one that yields the minimum entropy (Compton et al., 2022). Under a set of assumptions, this principle is shown to facilitate correct orientation of the edges in the causal graph in a two-variable setting (Compton et al., 2020); and is addressed in the more general case of multi-variable causal graphs in Compton et al. (2022), by finding the minimum entropy coupling between each pair of connected variables.

These results further point to the relationship between Shannon entropy and distributional robustness. Fitting the wrong model to the data requires a higher entropy than the correct model (Compton et al., 2020; 2022). More precisely, let $Y = f(X, E)$ be the structural causal model, where $E \perp\!\!\!\perp X$ denotes exogenous variables. If the entropy $H(E)$ is sufficiently small, for the data to fit an alternative structural model $X = g(Y, \tilde{E})$, with high probability, the Shannon entropy of the alternative exogenous variables $\tilde{E} \perp\!\!\!\perp Y$ is bounded from below,

$$H(X) + H(E) - H(Y) < H(\tilde{E}). \tag{16}$$

Considering the link between entropy and Ricci curvature as described in Section 4, Inequality 16 enables us to prove the connection we establish between Ricci curvature and causal inference through Theorem 6.1.

G Theorem Details

We now provide the proof of Theorem 6.1, which states the following under the assumptions listed in the appendix Section G.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.

Theorem 6.1 allows us to deduce Corollary 6.3, which sheds light on a direct connection between causal identification and Ricci curvature. In particular, with the setup and assumptions of Theorem 6.1, Corollary 6.3 states that the probability that the measure of the maximal set over which the causal parameter is point identified is at least as large as a constant υ is weakly higher in the case with non-negative Ricci curvature than when Ricci curvature is negative. The proof of this corollary is provided in this appendix following the proof of Theorem 6.1.

G.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. (2017) and Compton et al. (2020): $X \sim p(X)$ 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 (Compton et al., 2022).

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 2 and 16 used in the proof. Assumption (Aii) on the other hand, while facilitating steps of the proof, has a conceptual implication, it implies that the difference in the randomness of the two datasets is primarily due to X_1 and X_2.

G.2 Proofs

The proof of Theorem 6.1, under the assumptions above, relies on Inequality 2 from Pouryahya et al. (2017) and Lott & Villani (2009), and the results from Compton et al. (2020) and Compton et al. (2022) leading to Inequality 16. Given alternative models $X_i = g(Y_i, \tilde{E}_i)$ for $i = 1, 2$ with exogenous variables $\tilde{E}_i$, under (Aiii), Inequality 16 gives the following lower bound on the Shannon entropy of $\tilde{E}_i$,

$$H(X_i) - H(Y_i) + H(E_i) < H(\tilde{E}_i). \tag{17}$$

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 2, assuming (Ai), we have

$$\underline{s}_2 > \underline{s}_1, \tag{18}$$

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 18 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 $\Lambda_i := H(X_i) - H(Y_i) + H(E_i)$ 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 17, this implies

$$\mathbb{P}\left[H(\tilde{E}_2) > \eta\right] \geq \mathbb{P}\left[H(\tilde{E}_1) > \eta\right], \tag{19}$$

completing the proof of the theorem, establishing the link between causal inference and curvature.

Proof of Corollary 6.3. The result from Theorem 6.1 further allows us to prove Corollary 6.3, directly linking causal identification with Ricci curvature, in light of the formalization of the identified set over a class of unobserved variables introduced in Section 6. Consider the setup described above and assume the conditions for Theorem 6.1 hold. Let $\mathcal{U}$ be the set of all possible unobserved variables, and define $\mathcal{U}^\eta := \{U \in \mathcal{U} \ : \ H(U) \leq \eta\}$ to be the subset of $\mathcal{U}$ with the entropy at most η, where η is the constant in Inequality 19. Let θ_i denote the causal parameter that describes the causal mechanisms between X_i and Y_i. Following the notation described in Section 6 leading to Definition 6.2, let F_{X_i,Y_i,E_i} denote the true population CDF, while $F_{X_i,Y_i} = h(F_{X_i,Y_i,E_i})$ is the corresponding observed CDF, and let $\vartheta : \mathcal{F}_{X,Y,U} \to \Theta$ be the function that maps the population distribution to an expressive causal parameter. Note that an expressive $\vartheta(.)$ must be injective, reflecting differences between input distributions. Hence, given $Y_i = f_i(X_i, E_i)$ and $X_i = g_i(Y_i, \tilde{E}_i)$ described by causal parameters $\theta_i^f \equiv \vartheta(F_{X_i,Y_i,E_i})$ and $\theta_i^g \equiv \vartheta(F_{X_i,Y_i,E_i})$, as long as $F_{X_i,Y_i,E_i} \neq F_{X_i,Y_i,\tilde{E}_i}$, we have $\theta_i^f \neq \theta_i^g$, as reasonably desired. However, since $F_{X_i,f_i(X_i,E_i)} = F_{g_i(Y_i,\tilde{E}_i),Y_i}$, i.e., the distributions of (X_i, Y_i, E_i) and $(X_i, Y_i, \tilde{E}_i)$ are observationally equivalent, $\Theta^{\mathcal{D}}$ is not a singleton for any subset $\mathcal{D} \subseteq \mathcal{U}$ such that $\left\{E_i, \tilde{E}_i\right\} \subseteq \mathcal{D}$, as $\left\{\theta_i^f, \theta_i^g\right\} \subseteq \Theta^{\mathcal{D}}$. In other words, the causal parameter is not point-identified over any set of unobserved variables that contains both E_i and $\tilde{E}_i$. Let $\bar{\mathcal{U}}_i \subseteq \mathcal{U}$ denote the maximal set of unobserved variables over which θ_i is point-identified, which means $\tilde{E}_i \notin \bar{\mathcal{U}}_i$. Thus, we can write

$$\mathbb{P}\left[\mathcal{U}^\eta \subseteq \bar{\mathcal{U}}_i\right] = \mathbb{P}\left[\tilde{E}_i \notin \mathcal{U}^\eta\right] \equiv \mathbb{P}\left[H(\tilde{E}_i) > \eta\right]. \tag{20}$$

Let, $\upsilon := \mu(\mathcal{U}^\eta)$ for a measure μ over the measurable space $(\mathcal{U}, 2^{\mathcal{U}})$. Then Equation 20 implies that $\mathbb{P}\left[\mu\left(\bar{\mathcal{U}}_i\right) \geq \upsilon\right] = \mathbb{P}\left[H(\tilde{E}_i) > \eta\right]$. Hence, it follows immediately from Inequality 19 that $\mathbb{P}\left[\mu\left(\bar{\mathcal{U}}_2\right) > \upsilon\right] \geq \mathbb{P}\left[\mu\left(\bar{\mathcal{U}}_1\right) > \upsilon\right]$, as stated in Corollary 6.3. This completes the proof of this corollary.

H NetEst and f-NetEst

Given a graph $G = (V, E)$ with the adjacency matrix A, features X, observed outcome Y, and treatments T, the NetEst model (Jiang & Sun, 2022) 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

$$\tau(x_v) := \mathbb{E}\left[Y_v | do(t_v = t', z_v = z') - Y_v | do(t_v = t'', z_v = z'') \,\middle|\, x_v, \{x_u\}_{u \in N_v}\right], \tag{21}$$

which is identified under the assumptions described in Appendix A (Jiang & Sun, 2022).

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 $\mathcal{L} = \mathcal{L}_m + \alpha_t \mathcal{L}_t + \alpha_z \mathcal{L}_z$.

In Section 6.2, we propose a Ricci flow adjustment to improve the accuracy of causal parameter estimation on network data. Being a preprocessing of the input network structure, this adjustment lends itself to any causal inference method which takes graph-structured treatment units as its input. Considering the practical success of NetEst and the empirically-relevant identifiability results shown by Jiang & Sun (2022), we combined our proposed adjustment with NetEst. The pseudocode for the resulting method, which we refer to by *f-NetEst*, is included in Algorithm 1. We use f-NetEst as the primary method to validate our theoretical results and showcase its application in Section 7.

Algorithm 1 *f-NetEst*: Ricci Flow Adjustment for NetEst Training and Testing

Input: Graph of treatment units $G(V, E)$, node features X, labels Y for training, number of iterations T.
Output: Predicted labels $\hat{Y}$.

$G^0 \leftarrow G$.
for $t = 1$ to T **do**
    $K_G \leftarrow$ COMPUTERICCICURVATURE(G^{t-1})
    $G^t \leftarrow$ RICCIFLOW(G^{t-1}, K_G)
$M \leftarrow$ TRAINNETESTMODEL(G^T, X, Y)
$\hat{Y} \leftarrow$ TESTNETESTMODEL(G^T, X, M)

procedure COMPUTERICCICURVATURE(G)
    for each edge $(u, v) \in E$ **do**
        Compute Ricci curvature κ_{uv}
    return Ricci curvature values $K_G = \{\kappa_{uv} \ : \ (u, v) \in E\}$

procedure RICCIFLOW(G, K_G)
    for each edge $(u, v) \in E$ **do**
        Update edge weight $w_{uv} = (1 - \kappa_{uv})\, d(v, u)$
    return G with updated edge weights

procedure TRAINNETEST(G, X, Y)
    Compute peer exposure Z
    Initialize NetEst model M
    Train M using G, X, Z, and Y
    return M

procedure TESTNETESTMODEL(G, X, M)
    Predict labels $\hat{Y}$ using M, G, X, and Z
    return $\hat{Y}$

I Implementation Parameters and Hardware Specifications

Since the main purpose of our experiments is 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 (2022) 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 (Kingma & Ba, 2015). 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

J 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 (Guo et al., 2020; Hill, 2011; Jiang & Sun, 2022; Ma

et al., 2021; Shalit et al., 2017; Veitch et al., 2019). Jiang & Sun (2022) use the BlogCatalog (BC) and Flickr datasets (Guo et al., 2020; Ma et al., 2021) 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 (Rozemberczki et al., 2021); Cora and CiteSeer networks (Yang et al., 2016); and Actor network (Pei et al., 2019). Table 3 includes descriptive statistics on these networks. Note that we only use the largest connected component in each network. Following Jiang & Sun (2022), we split each network data into training, validation and test sets using METIS (Karypis & Kumar, 1998). The treatments and potential outcomes for all network data are synthesized following the formulation in Jiang & Sun (2022).

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

K Additional Experiments

K.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 4 for completeness and ease of visualization/comparison). The distributions for the nine networks are shown in Figure 6, with two plots (one per curvature) for each dataset. All ITE estimations in this figure have been obtained using NetEst (Jiang & Sun, 2022). 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.

K.2 ITE Error Distribution

In order to obtain a better understanding of how the geometric 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 (2022), as well as seven other networks described in Appendix J. As shown in Figure 7, 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.

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

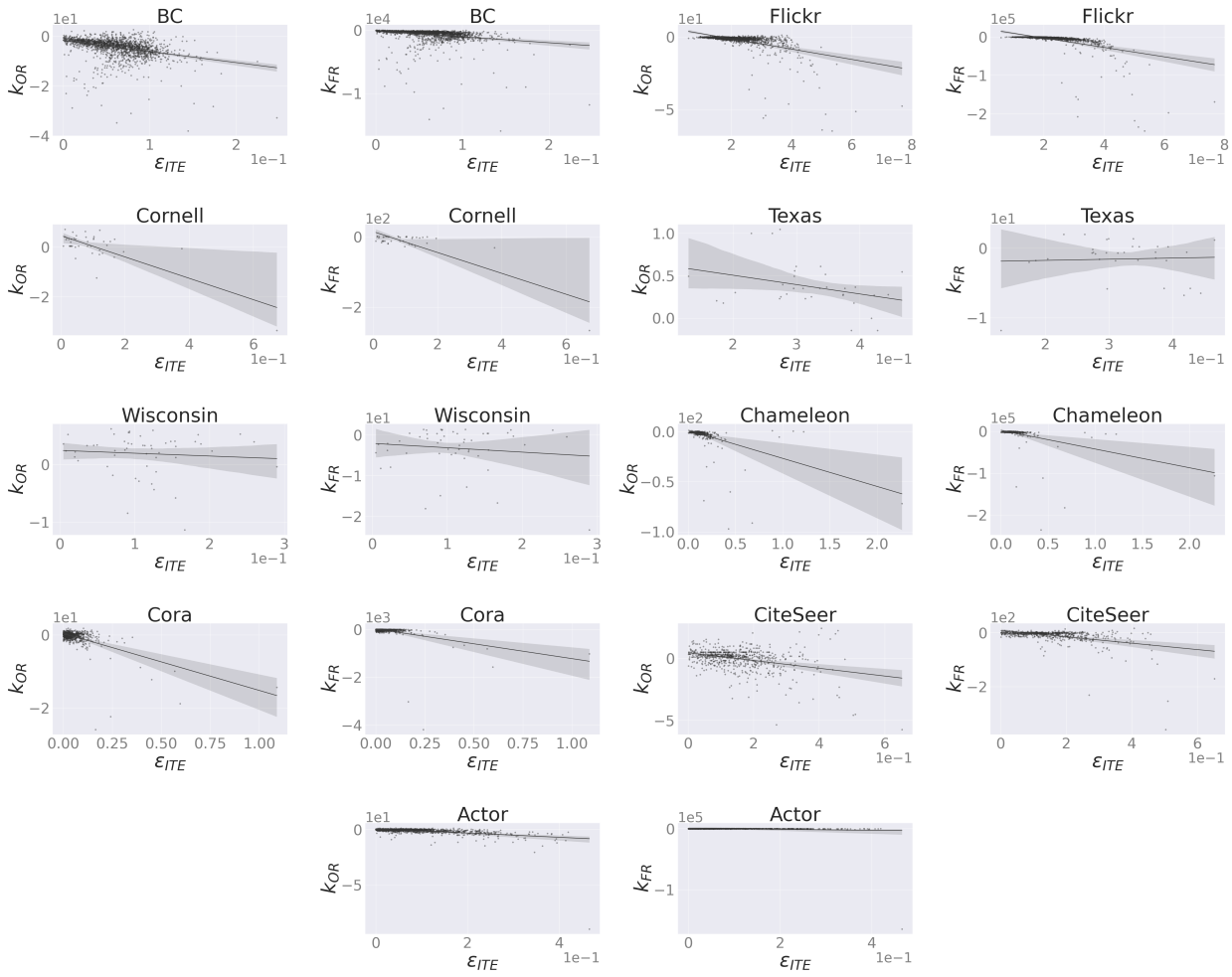

Figure 6: Joint distributions of the sum of Forman and Ollivier-Ricci curvatures in the neighborhood of each node and the estimation error of ITE for that node. The distributions for the nine networks are shown, with two plots (one per curvature) for each dataset. The regression lines with the corresponding 95% confidence intervals are marked on the plots.

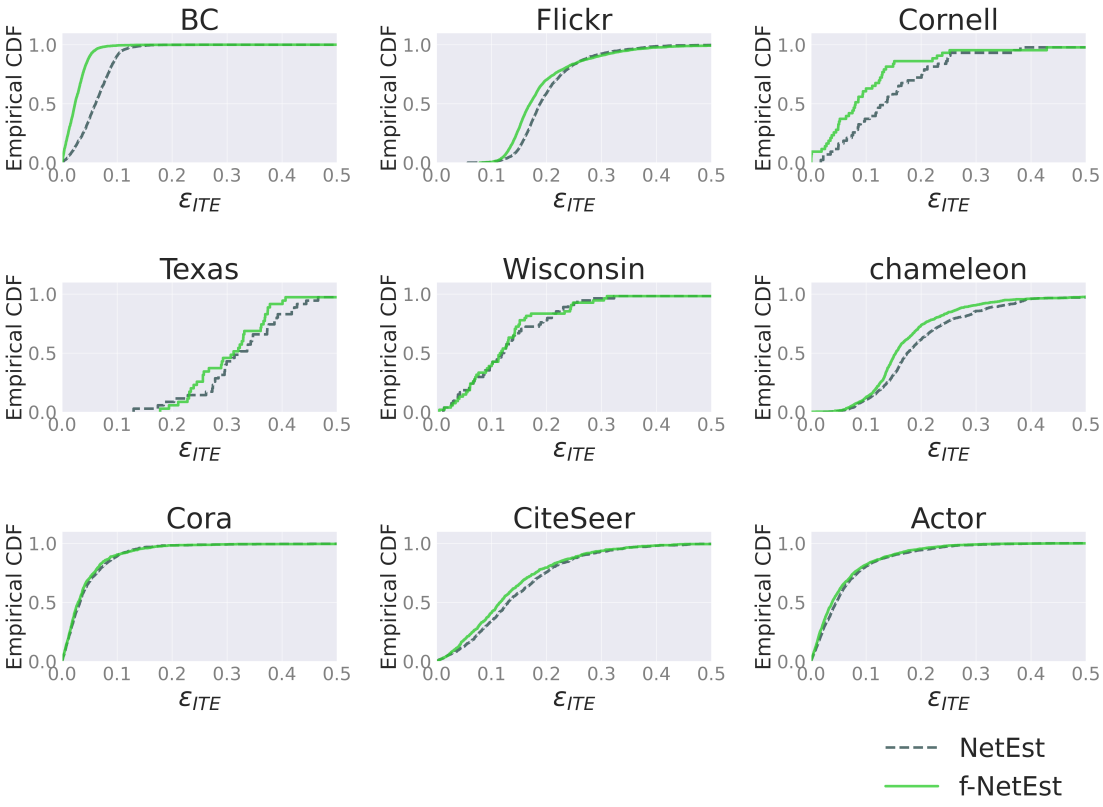

Figure 7: Empirical CDF of the ITE error, ε_{ITE}, obtained from NetEst (black) and f-NetEst (green).