METRICS FOR THE EVALUATION OF LEARNED CAUSAL GRAPHS BASED ON GROUND TRUTH

Josephine Rehak[1], **Alexander Falkenstein**[1], **Frank Doehner**[1], **and Jürgen Beyerer**[1,2]
[1]Karlsruhe Institute of Technology, Kaiserstraße 12, Karlsruhe
[2]Fraunhofer IOSB, Fraunhoferstraße 1, Karlsruhe

ABSTRACT

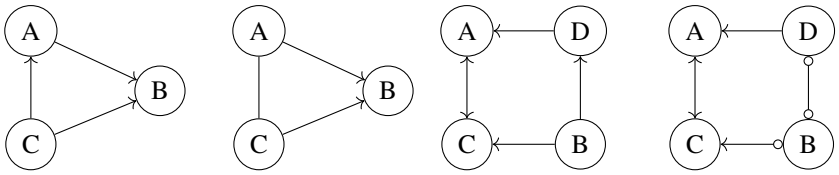
The self-guided learning of causal relations may contribute to the general maturity of artificial intelligence in the future. To develop such learning algorithms, powerful metrics are required to track advances. In contrast to learning algorithms, little has been done in regards to developing suitable metrics. In this work, we evaluate current state of the art metrics by inspecting their discovery properties and their considered graphs. We also introduce a new combination of graph notation and metric, which allows for benchmarking given a variety of learned causal graphs. It also allows the use of maximal ancestral graphs as ground truth.

Keywords causal graph · metric · causal discovery · ground truth · bayesian network structure learning · causal structure learning · ancestral graph · acyclic graph

1 Introduction

Causal Discovery (CD) is a domain of artificial intelligence, that targets the identification of cause-effect relationships in data [12]. For the evaluation of CD algorithms, it has become a standard approach to perform novel algorithms on popular benchmarking datasets, for which the ground truth is known [9, 18]. By choosing the metric with desired properties, the learned graph can be compared against the ground truth and the discovery capabilities of the algorithm can be assessed by the resulting score [4]. In this style of evaluation, the metrics and its properties play a vital role. This is why in this work, we will investigate current metrics of CD by assessing their applicability on different graph types learned by algorithms. We highlight desired criteria and inspect how different state of the art metrics facilitate these. Finally, we propose a new metric called the normed causal edit distance metric (nCED).

In Section 2, we give a basic introduction to CD. In Section 3, we first introduce a new graph notation to unify the several established causal graph types before proposing our new metric. In Section 4, we cover related work in the field of CD metrics and prior metric evaluations. In Section 5, we take a closer look into the general topic of metric evaluation. Finally, Section 6 concludes the paper.



(a) DAG (b) PDAG (c) MAG (d) PAG

Figure 1: Depicted are examples of common types of causal graphs in their established representation.

2 Fundamentals of Causal Discovery

2.1 Causal Graphs

In general, it is defined that a causal relation is present from variable a to b written as $a \to b$, if the value of b depends on the value of a. If there is no such causal influence present in either direction, they are independent $a \perp\!\!\!\perp b$.

Networks of such causal relations are represented as graphs $G = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V}$ and edges $\mathcal{E}$. Nodes are random variables, representing chosen events or states, while edges indicate the causal relations. In our definition, two variables a, b can be connected by a total of two edges $e_{a,b} \in \mathcal{E}$ and $e_{b,a} \in \mathcal{E}$.

In this work, we employ two different classes of causal graphs. Their main difference is the class of acyclic graphs and the corresponding learning algorithms assume causal sufficiency, the absence of latent variables. Ancestral graphs on the other hand do not. Each class has two representatives as depicted in Figure 1.

Starting with the class of acyclic graphs, Directed Acyclic Graphs (DAGs) [19] allow edges to be either present, indicated by an arrowhead pointed at the caused variable, or absent indicated by missing arrowheads or by the absence of any representation. Cycles in the directed edges are forbidden. This forbids minimal cycles like $a \leftrightarrow b$.

Complete Partial Directed Acyclic Graphs (CPDAGs) may contain besides present and absent edges, also undirected edge pairs like $a - b$. In this case, it is a causal relation exists between a and b but its direction, $e_{a,b}$ or $e_{b,a}$, is unknown. Under the constraint that no cycles are created, $a - b$ may be oriented in either direction to gain DAGs. These DAGs form a Markov equivalence class for given CPDAG [17].

For the class of ancestral graphs, we consider Maximal Ancestral Graphs (MAGs). They allow one-directed and bidirected edges, but forbid any cycles formed by a combination of bidirected and directed edges. In ancestral graphs, bidirected edges $a \leftrightarrow b$ imply the presence of a latent variable l with $l \to a$ and $l \to b$. DAGs are a subset of MAGs which do not contain any bidirected edges [33]. Similar to CPDAGs, a Partial Ancestral Graph (PAG) [23, 24, 25] may additionally include unknown edges represented by circular arrowheads, implying a Markov equivalence class of several MAGs. These unknown edges may be present in single and opposing edges. As CD algorithms for MAG learning have not been developed yet, they will be of less importance in this work.

2.2 Established Metrics

Several metrics that use a ground truth for the evaluation of the discovered graph are currently established in literature.

To explain them, we make use of the following definitions. The count of True Positives edges (TP) is the number of discovered present edges that are also present in the ground truth. The count of True Negatives (TN) is the number of all edges that are neither present in the discovered graph and the ground truth. The count of False Positives (FP) is the number of discovered edges that are not present in the ground truth, and the False Negatives (FN) is the number of edges in the discovered graph that are not present but are missing in the ground truth [13]. We specify FN and FP as undirected, if only the presence of the edge is considered to be compared to the ground truth, but not its orientation.

The first metric, we introduce is the F_1 score [28]. It considers only the present edges and calculates the harmonic mean of precision and recall when comparing the inferred graph to the ground truth.

$$F_1 \text{ score} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \tag{1}$$

The Receiver Operator Curve Area Under Curve (ROC AUC) metric [10, 21] measures the True Positive Rate (TPR) on the False Positive Rate (FPR) several times to plot a curve.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{3}$$

The ROC AUC is measured once for an 'empty' graph with a TPR and FPR of zero, for the actual discovered graph, and also for a graph with maximized TPR and FPR to 100 percentage points. Then, the area under this curve is calculated. A ROC AUC of 1 indicates a discovered causal graph that is identical to the ground truth with a TPR of hundred percent and an FPR of zero percent. A ROC AUC of fifty percentage points indicates a discovery performance similar to a random guesser. Zero percentage points indicate the inverse of a successfully CD. The Precision Recall Curve Area Under Curve (PRC AUC) [7] operates similar. Instead of TPR and FPR, it uses precision and recall to calculate the area under curve. The Structural Hamming Distance (SHD) [31, 2] counts the overall number of changes in edges that are required to transform the discovered graph into the ground truth. In the established definition, the allowed changes include only the addition and the deletion of edges.

$$\text{SHD} = \text{FN} + \text{FP} \tag{4}$$

The adapted version for CPDAGs [31] converts the ground truth and the predicted graph to CPDAGs and then counts the operations required.

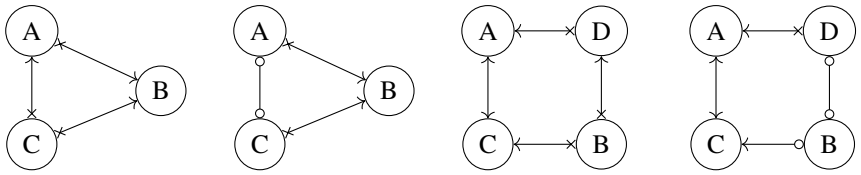
$$\text{SHD CPDAG} = \text{undirected FN} + \text{undirected FP} + \text{FN} + \text{FP} \tag{5}$$

No version of SHD is available for MAGs and PAGs.

3 Causal Edit Distance Metric

3.1 Universal Causal Graph Representation

To make all the various graph types comparable, we transfer them to a unified representation. Consider causal graphs $G = (\mathcal{V}, \mathcal{E})$ with variables V and edges $\mathcal{E}$. The edges $\mathcal{E}$ can be represented as an adjacency matrix $\mathbf{E}$. Each edge $e_{i,j} \in \mathbf{E}$ represents knowledge about a present causal relation from variable i to j and also has a counterpart



(a) DAG (b) PDAG (c) MAG (d) PAG

Figure 2: To compute the nCED, the example graphs of Figure 1 are adapted to match the new universal notation described in Section 3.1.

$e_{j,i}$ pointing in the opposite direction. Edges may take on the values $e_{i,j} \in \{-1, 0, 1\}$. $e_{i,j} = 0$ indicates the absence of a causal relation from i to j. This is represented by $\times$ arrowheads in the universal representation. $e_{i,j} = -1$ indicates the edge is undirected. This is represented by the arrowhead $\circ$ in the universal graph representation. $e_{i,j} = 1$ in the adjacency matrix or $\rightarrow$ in the universal graph representation represent the presence of a causal relation. Hence, for $N = |\mathcal{V}|$, the adjacency matrix is $\mathbf{E} \in \{-1, 0, 1\}^{N \times N}$. The adjacency matrix representation allows cyclic causal relations if such graphs are not actively prohibited. Causal self-references of variables, the smallest possible cycles, are excluded $\forall i \in \mathcal{V}, e_{i,i} = 0$.

3.2 Calculating the Causal Edit Distance

The causal edit distance (CED) between two graphs G and G^* is using a comparison between the corresponding adjacency matrix elements $e_{i,j} \in \mathbf{E}$ and as $e^*_{i,j} \in \mathbf{E}^*$ follows.

$$\text{CED}(G, G^*) = \sum_{\substack{(i,j) \in V^2 \\ i \neq j}} f(e_{i,j}, e^*_{i,j}) \tag{6}$$

For its computation, we require the following function f using the parameter $0 < k < 0.5$.

$$f(e_{i,j}, e^*_{i,j}) = \begin{cases} 0 & \text{if } e_{i,j} = e^*_{i,j} \\ k & \text{if } e_{i,j} \neq e^*_{i,j} \wedge e_{i,j} = -1 \\ 1 & \text{else} \end{cases} \tag{7}$$

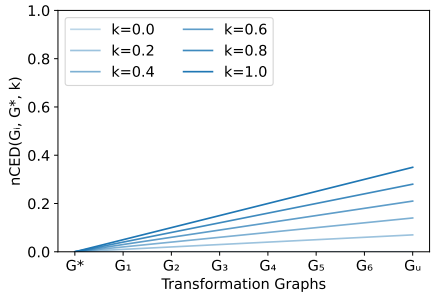
The normalized variant of the Causal Edit Distance metric is called in short nCED. It requires the division of the CED with the maximum possible editing costs $N(N-1)$ given $N = |\mathcal{V}|$ variables.

$$\text{nCED}(G, G^*) = \frac{\text{CED}(G, G^*)}{N(N-1)} \tag{8}$$

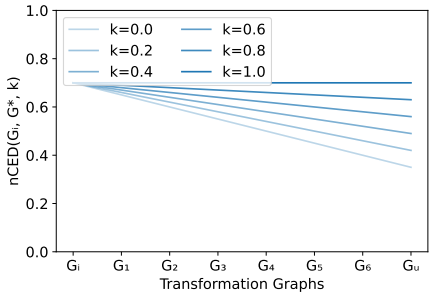
3.3 Analysis on Choosing the Parameter k

Due to the function f, the algorithm punishes each TP and TN with 0, each undirected FP and FN with k and each FP and FN with 1. Given $0 \leq k \leq 1$, k can be freely chosen. Consider, that $k \leq 1$ is required for the normalization step to be valid and $0 \leq k$ for edges to be scored independently.

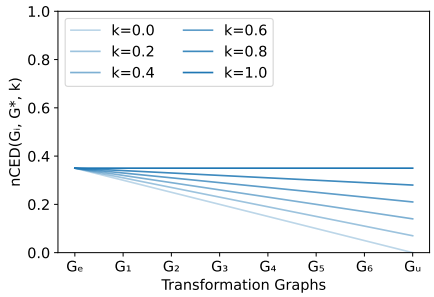
In the following, we propose our own preferences in the choice of k. As undirect TP edges correctly imply the presence of a TP edge, they should be evaluated better than



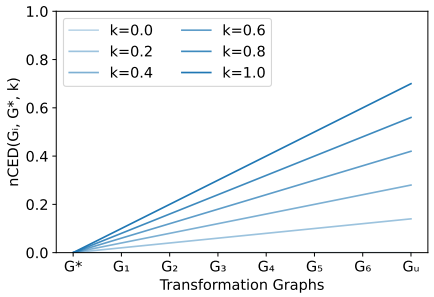
(a) Stepwise transformation from G^* to $G_{u,1}$



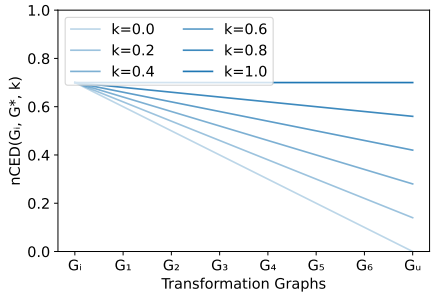
(b) Stepwise transformation from G_i to $G_{u,1}$



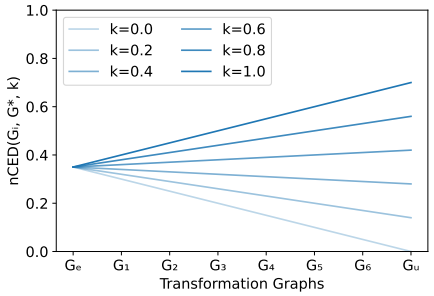
(c) Stepwise transformation from G_e to $G_{u,1}$



(d) Stepwise transformation from G^* to $G_{u,2}$



(e) Stepwise transformation from G_i to $G_{u,2}$



(f) Stepwise transformation from G_e to $G_{u,2}$

Figure 3: Overview of the nCEDs behavior for varying k on stepwise graph transformations of a five variable DAG.

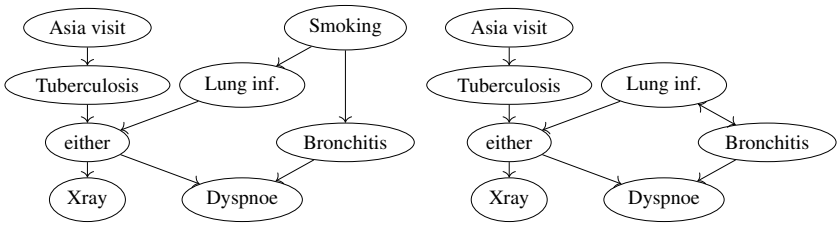
FN edges, but worse than TP edges as they do not imply it with certain orientation.

$$f(e_{i,j} = 1, e^*_{i,j} = 1) < f(e_{i,j} = -1, e^*_{i,j} = 1) < f(e_{i,j} = 0, e^*_{i,j} = 1) \tag{9}$$

Likewise, undirected FP edges should be scored worse than TN edges, as they imply faulty information, but they should be evaluated better than FP edges, since the certain discovery of a FP edge is more severe.

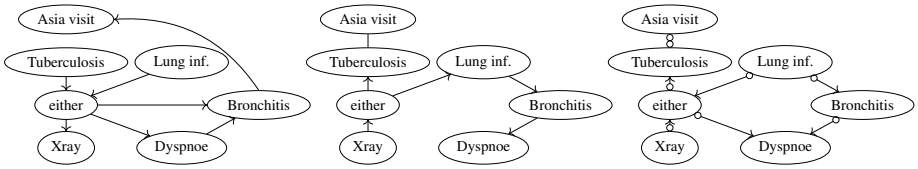
$$f(e_{i,j} = 0, e^*_{i,j} = 0) < f(e_{i,j} = -1, e^*_{i,j} = 0) < f(e_{i,j} = 1, e^*_{i,j} = 0) \tag{10}$$

Accordingly, the score of an undirected graph nCED(G_u, G^*) should be better than the score of an empty graph nCED(G_e, G^*) and better than the score of a graph with flipped



(a) Original DAG (b) Created MAG

Figure 4: The ground truth DAG and the created MAG for the Asia dataset. By eliminating the variable smoking, we create a hidden confounder between the lung infection and the bronchitis variable. The resulting graph is a ground truth MAG.



(a) Example DAG learned by HCS algorithm (b) Example CPDAG learned by PC algorithm (c) Example PAG learned by FCI algorithm

Figure 5: Example graphs as they were created by the different learning algorithms on application on the adapted Asia dataset.

edges nCED(G_i, G^*), but worse than nCED(G^*, G^*). This should hold for undirected graphs with individual undirected edges $G_{u,1}$ (as is possible in PAGs), but also where opposing edges are undirected $G_{u,2}$ (as is the case in CPDAGs).

We observed the behavior of the metric for different k by stepwise transforming G_e, G_i and G^* into $G_{u,2}$, which results in the plots shown in Figure 3. One observes nCED$(G^*, G^*) <$ nCED$(G_{u,2}, G^*)$ and nCED$(G_{u,2}, G^*) <$ nCED(G_i, G^*) to hold true for any k. But while nCED$(G_e, G^*) <$ nCED$(G_{u,1}, G^*)$ is valid for any k, we see in Figure 3f that nCED$(G_e, G^*) <$ nCED$(G_{u,2}, G^*)$ only holds if $k < 0.5$. This is why we propose to choose $0 < k < 0.5$.

Alternate desired choices of k might include $k > 0.5$, so that the algorithm treats undirected edges worse than FN and FP. It may be of use if the consideration of undirected edges is to be punished and only the learning of TP and TN is desired.

For $k = 0$, the nCED does not differ between TP and undirected TP edges. This might be of use if the orientations of an edge matters less than the knowledge of its presence.

3.4 Maximum Ancestral Graphs as Ground Truth

As the nCED may evaluate DAGs, CPDAGs, MAGs and PAGS because of the universal representation, and DAGs are a subclass of MAGs, we can also allow G^* to be a MAG instead of a DAG. As the edge representations of DAGs and MAGs are identical in the universal representation, this does not require an adaptation of nCED's definition. To our knowledge, nCED is the only metric that is capable of this. As a consequence, if $k > 0$ and a bidirected edge is present in the ground truth PAG, the nCED does not allow DAG-learning CD algorithms to achieve the optimal score anymore.

Table 1: Results of the nCED with a MAG as ground truth for graphs learned on the adapted Asia dataset. The lower the score, the better performed the structure learning algorithm.

	nCED (ours)					
	$k = 0$	$k = 0.1$	$k = 0.2$	$k = 0.3$	$k = 0.4$	$k = 0.5$
HCS	0.17	0.17	0.17	0.17	0.17	0.17
PC	0.12	0.13	0.14	0.15	0.17	0.18
FCI	0.12	0.14	0.14	0.15	0.17	0.18

In the following, we demonstrate how the novel nCED metric is able to evaluate results of structure learning algorithms using a PAG as ground truth. Regarding CD learning algorithms, we employed the Peter-Clark (PC) [30] algorithm, which delivers as result a CPDAG, the Fast Causal Inference (FCI) [34] algorithm delivers a PAG, while the hillclimb search (HCS) algorithm [15] delivers a DAG. For independence test the Fisher Z-test was used, and for score-based CD the Bayesian information criterion. Unfortunately, we know of no algorithm which is able to learn MAGs.

As a benchmark, we used the Asia dataset. It describes the causal relations of visits to Asia and smoking to the corresponding lung diseases and the medical options for identifying the disease. We created a ground truth MAG, by eliminating the smoking variable from the dataset as shown in Figure 4. In its place, it has a bidirected edge implying a latent variable between the bronchitis and the lung infection variable. After learning ten graphs using the introduced learning algorithms, including the example graphs shown in Figure 5, we calculated the averaged nCED scores for different k using the created MAG. The results are depicted in Table 1.

4 Related Work

4.1 Metric Types in Causal Research

In CD research, two different approaches exist to evaluate the learned causal graphs [5]. We focus on metrics, which inspect graph elements to compare the discovered graph against a provided ground truth.

As an alternative, scoring functions, which are commonly used as objective functions for score-based structure learning, can be used to evaluate the fit of a provided graph to given data. Established is for example the use of the Minimum Description Length [26], the Bayesian Information Criterion [29] and the Bayesian Dirichlet equivalence [14] for the score-based learning of causal graphs [3].

4.2 Evaluations of Metrics for Structure Learning

Several publication investigated the use of metrics using ground truth for the purpose of causal discovery.

De Jongh et al. [8] applied the Greedy Thick Thinning CD algorithm and the Greedy Equivalence Search algorithm on three datasets and inspected the resulting CPDAGs and DAGs. The examined metrics included the number of FN edges, the number of FP edges, the sum of undirected and directed TP edges, the number of TP edges, the number of TP edges with correct orientation, the number of edges with incorrect orientation, a

weighted sum of FN, FP and undirected and directed TP edges, a weighted combination of FN, FP, TP , and finally the Structural Hamming Distance (SHD) as the sum of TP and FN. They identified the SHD as a useful metric, but they also advise against the use of a single metric, as the choice on a single metric leads to information loss.

Constantinou et al. [5] did a general inspection of causal discovery metrics by inspecting and comparing the different types of metrics described above. The inspected metrics such as the TP, FP, TN, FN, Precision, Recall, the F1 measure, the SHD and the DAG Dissimilarity Metric (DDM) [6] and discovered imbalances in the metric scores. The use of metrics based on scoring functions may be biased due to the scoring function itself. Metrics using ground truth are reported to be biased by imbalances in the number of edges or independencies in the ground truth. TP, FP, TN, FN as well as Precision and Recall were reported to be biased as each does not integrate enough TP, FP, TN and FN values sufficiently to be independently meaningful.

Of the described publications, none investigated the use of current methods such as the Precision Recall Curve and the Receiver Operator Curve Area under Curve [22, 20]. Additionally, neither of them investigated metrics for ancestral graphs.

4.3 Metrics for Graph Comparison

In graph theory, there exists related literature regarding the comparison of graphs [11, 32]. Popular is for example, the graph edit distance (GED) metric [1] to compare the fit of two given graphs G_1 and G_2. It is the sum of costs over all required edge and node insertions, substitutions and deletions to transform G_1 into G_2. Since commonly, the nodes are fixed in CD, exact matching algorithms may be used as CD metrics.

As the GED itself is not in use in current CD, it was not part of our investigation. Instead, we investigate the SHD. It is a descendant of the GED, but other than the GED, it only computes the summed costs for the substitution, insertion and deletion of edges, since the nodes are expected to be fixed in CD. Additionally, its costs for each operation are set to one.

5 Metric Evaluation

5.1 Overview of Causal Discovery Metric Criteria

We pose a specific set of criteria on any metric m with a corresponding score s that uses solely the ground true graph G^* to evaluate the goodness of fit to any discovered causal graph G.

Lower and upper bound criterion We prefer s to be bound by maximum and minimum values. While the use of one bound is common to indicate a perfect match between G and G^*, the use of a second bound is helpful for orientation.

G^* identity criterion The criterion of G^* identity assumes if $m(G^*, G^*) = m(G^*, G)$ holds, then $G = G^*$ holds.

TP sensitivity criterion and FP / FN sensitivity criterion Since the true causal graph is assumed to contain only present and absent edges, a metric is required to be sensitive to changes regarding TP and FN edges given the present edges of the ground truth, and also to be sensitive to changes in TN or FP edges regarding the absent edges of the ground truth.

Scoring consistency criterion We try to assure consistent scoring for variations in TP / FN edges and for TN / FP edges by performing the required changes on G and checking for consistent linear changes in scores. For the TP and FP edges, this entails the continuous increase or decrease of absent edges in G and for the TP and FP edges the continuous increase or decrease of directed edges in G. This criterion holds if TP $= 0$. This allows comparability between different 'wrong' learned graphs.

Finally, we prefer metrics to handle any of the currently established graph types to make the causal discovery metrics comparable. This includes DAGs, PDAGs, MAGs and PAGs.

5.2 Investigation on Metric Applicability

To illustrate the applicability of the existing and the novel metric, we demonstrate their use on two example datasets. One is the established Asia dataset [16] with 10.000 data entries with 8 variables each. The second one is the popular Sachs datasets [27] containing 10.000 data entries with 11 variables each. For both, the ground truth is known. Three different CD algorithms were exercised on given datasets. MAGs were excluded as no respective discovery algorithm exists. Again the HCS, the PC Algorithm and the FCI algorithm were employed. All of the examined metrics were calculated for the resulting graphs. The scores were collected for 10 learned graphs each. In Table 2, we depict the resulting score average. We can observe that most metrics are primarily defined for DAGs. Several metrics such as the ROC AUC, PRC AUC, F_1 score, TPR and FPR do not consider undirected edges and only evaluate TP, FP, FN, and TN directed edges even if G^* contains an edge at this location. As PAGs and CPDAGs may contain directed edges as well, these metrics can be transferred without considering the undirected edges. Besides our own metric, only the CPDAG version of the SHD is defined for CPDAGs. As its definition deviates from the SHD for DAGs, we list it in a separate column. In our investigations, we did not find existing custom designed metrics for MAGs and PAGs.

5.3 Experiment on Scoring Consistency and Sensitivity

For this experiment, we considered a generic DAG with $N = 5$ and containing the maximum number of allowed directed edges resulting in seven directed edges and thus seven possible random structure changes. To inspect the metric behavior regarding TP - FN consistency, we stepwise changed the true DAG G^* into an empty DAG G_e randomly eliminating a TP and adding a FN one by one. Regarding consistency in TN / FP direction, we stepwise changed an empty graph G_e into an 'inverse' graph G_i by randomly adding FP directed edges directly opposing the edges of the ground truth. The score results are shown in Figure 6.

We observe in Figure 6a the PRC AUC to be sensitive, but to not perform consistent linear regarding constant changes in TP and FN edges. The zero gradient of the FPR proofs it to not be sensitive to said changes. Regarding sensitivity in TN and FP edges in Figure 6b, the TPR, F1 and PRC AUC show non-responsive by having a gradient of zero.The only metrics sensitive for changes in both individual trials show to be the ROC AUC, the SHD and the nCED.

5.4 Elaboration on Metric Bounds and Normalization

To inspect the bounds of each metric, we inspected the definition of the metrics. The overview is presented in Table 3. The SHD metric only comes without an upper bound.

Table 2: Example on how the examined metrics perform in an actual application scenario. The higher the scores of the ROC AUC, PRC AUC, F_1, TPR and FPR metrics, the better performed the causal structure learning algorithm. For each SHD and nCED variant, lower scores are better.

(a) Asia dataset

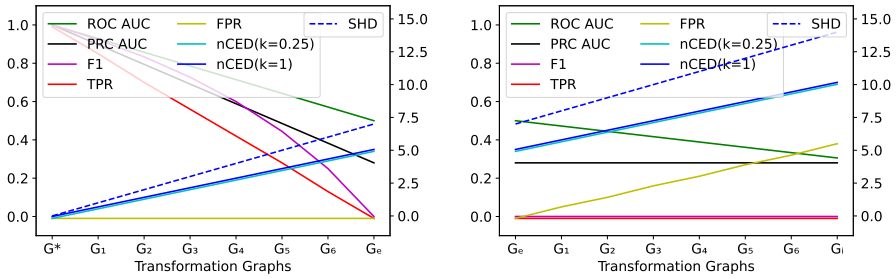
	ROC AUC	**PRC AUC**	**F_1 score**	**TPR**	**FPR**	**SHD** (DAG)	**SHD** (CPDAG)	**nCED** ($k = 0.2$)	**nCED** ($k = 0.4$)
HCS	0.62	0.19	0.33	0.38	0.12	12	-	0.20	0.20
PC	0.47 *	0.13 *	0.06 *	0.06 *	0.12 *	-	11	0.11	0.15
FCI	0.49 *	0.14 *	0.12 *	0.12 *	0.13 *	-	-	0.11	0.15

* The metric only considers present and absent directed edges, unaware of other edges types.

(b) Sachs dataset

	ROC AUC	**PRC AUC**	**F_1 score**	**TPR**	**FPR**	**SHD** (DAG)	**SHD** (CPDAG)	**nCED** ($k = 0.2$)	**nCED** ($k = 0.4$)
HCS	0.72	0.33	0.51	0.53	0.09	17	-	0.21	0.21
PC	0.47 *	0.13 *	0.06 *	0.06 *	0.12 *	-	10	0.22	0.26
FCI	0.57 *	0.16 *	0.25 *	0.29 *	0.16 *	-	-	0.23	0.26

* The metric only considers present and absent directed edges, unaware of other edges types.



(a) Stepwise transformation of TP edges of G^* to FN

(b) Stepwise transformation of TN edges in G_e to FP

Figure 6: Overview of the metrics behavior given the sequential use of random structure changes on a five variable DAG. The scale of the SHD is shown on the plots right side. We observe some metrics to have a gradient of zero and thereby to be not responsive for the induced changes. Of the responsive metrics, we observe most perform consistently indicated by a linear increase or decrease.

All other metrics come with an upper and lower bound and are therefore considered to be normalized. While most metrics use sums of FN, TP, FP or TN, but never all of them together, the nCED uses the number of graph variables for normalization. Mind, that $N = \text{TP} + \text{TN} + \text{FN} + \text{FP}$ holds and that is why the nCED considers each for normalization.

Table 3: Bounds and normalizability of the metrics described in Section 2.2

	Lower bound	**Upper bound**	**Normalization**	**Normalizing factor**
ROC AUC	✓	✓	✓	TP + FN and FP + TN
PRC AUC	✓	✓	✓	TP + TN and FP + TP
F_1 score	✓	✓	✓	$\frac{TP}{TP+TN}$ and $\frac{TP}{FP+TP}$
TPR	✓	✓	✓	TP + FN
FPR	✓	✓	✓	FP + TN
SHD (DAG)	✓	✗	✗	-
SHD (CPDAG)	✓	✗	✗	-
nCED (ours)	✓	✓	✓	$N(N-1)$

5.5 Elaboration on G^* identity

The criterion states that besides G^* no other graph G should exists for which the inspected metric evaluates $m(G, G^*) = m(G^*, G^*)$. For this investigation, we inspected if we were able to find evidences for each metric contradicting the criterion.

SHD for CPDAGs transforms the ground truth DAG to its CPDAG representation. This transformation is commonly surjective as unambiguous edge orientations are removed from the graph and thus several DAGs can share the same CPDAG representation. This is why the G^* identity criterion does not hold for the DAGs sharing the CPDAG representation with G^*.

Several metrics, such as the F1 score, PRC AUC, ROC AUC, TPR and FPR, do not consider undirected edges and they count them as absent edges. This is why the criterion does not hold for graphs identical to G^*, but with undirected edges instead of TN edges.

For $k = 0$, the nCED metric evaluates undirected edges like TP edges. Thus the criterion does not hold for graphs with arbitrary undirected TP and TP edges.

6 Conclusion

In this paper, we performed several inspections on established metrics for the use with ground truths. We proposed a new graph notation and a novel metric called the normalized Causal Edit Distance (nCED) that allows benchmarking of MAGs, PAGs, CPDAGs and DAGs. Additionally, we investigated if current causal discovery metrics fulfill a proposed set of criteria and found several peculiarities which should be taken into account before their application. Also, we found deficiencies in the applicability of current metrics as many perform on DAGs, but few perform on CPDAGs and even less on PAGs.

7 Acknowledgements

This work benefited from the Dagstuhl seminar 24031 "Fusing Causality, Reasoning and Learning for Fault Management and Diagnosis".

References

[1] Z. Abu-Aisheh, R. Raveaux, J.-Y. Ramel, and P. Martineau. An exact graph edit distance algorithm for solving pattern recognition problems. In *4th International Conference on Pattern Recognition Applications and Methods 2015*, 2015.

[2] S. Acid and L. M. de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, 2003.

[3] R. R. Bouckaert. *Bayesian belief networks: from construction to inference*. PhD thesis, 1995.

[4] L. Cheng, R. Guo, R. Moraffah, P. Sheth, K. S. Candan, and H. Liu. Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, 3(6):924–943, 2022.

[5] A. C. Constantinou. Evaluating structure learning algorithms with a balanced scoring function. *arXiv preprint arXiv:1905.12666*, 2019.

[6] A. C. Constantinou, N. Fenton, and M. Neil. How do some bayesian network machine learned graphs compare to causal knowledge? *arXiv preprint arXiv:2101.10461*, 2021.

[7] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[8] M. de Jongh and M. J. Druzdzel. A comparison of structural distance measures for causal bayesian network models. *Recent Advances in Intelligent Information Systems, Challenging Problems of Science, Computer Science series*, pages 443–456, 2009.

[9] C. C. Emezue, A. Drouin, T. Deleu, S. Bauer, and Y. Bengio. Benchmarking bayesian causal discovery methods for downstream treatment effect estimation. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

[10] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38, 2004.

[11] X. Gao, B. Xiao, D. Tao, and X. Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13:113–129, 2010.

[12] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[13] D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson. Further thoughts on precision. In *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011)*, pages 129–133. IET, 2011.

[14] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243, 1995.

[15] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[16] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

[17] C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of UAI 1995*, pages 403–410, 1995.

[18] G. Menegozzo, D. Dall'Alba, and P. Fiorini. Cipcad-bench: continuous industrial process datasets for benchmarking causal discovery methods. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 2124–2131. IEEE, 2022.

[19] J. Pearl. *Causality*. Cambridge university press, 2009.

[20] A. Pérez-Suay and G. Camps-Valls. Causal inference in geoscience and remote sensing from observational data. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1502–1513, 2018.

[21] W. Peterson, T. Birdsall, and W. Fox. The theory of signal detectability. *Transactions of the IRE professional group on information theory*, 4(4):171–212, 1954.

[22] S. R. Pfohl, T. Duan, D. Y. Ding, and N. H. Shah. Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference*, pages 325–358. PMLR, 2019.

[23] T. Richardson and P. Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[24] T. S. Richardson. *Models of feedback: interpretation and discovery*. PhD thesis, Carnegie-Mellon University, 1996.

[25] T. S. Richardson and P. Spirtes. Causal inference via ancestral graph models. *Oxford Statistical Science Series*, pages 83–105, 2003.

[26] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[27] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[28] Y. Sasaki et al. The truth of the f-measure. *Teach tutor mater*, 1(5):1–5, 2007.

[29] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[30] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[31] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[32] P. Wills and F. G. Meyer. Metrics for graph comparison: a practitioner's guide. *Plos one*, 15(2):e0228728, 2020.

[33] J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

[34] J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.