Collaborative Causal Discovery in a Synthetic Environment: Empirical Evaluation of FLODO

No Author Given

No Institute Given

Abstract. Causal discovery aims to uncover the underlying cause-and-effect relationships from observational and interventional data, a task historically dominated by centralized statistical methods. However, traditional centralized approaches are often impractical in real-world applications, as the inherently distributed nature of modern data architectures imposes strict privacy constraints that prohibit the pooling required by single-agent systems. In this paper, we empirically test FLODO (Flock of Dodos), a collaborative multi-agent algorithm designed for decentralized causal structure learning. Inspired by distributed problem-solving, FLODO deploys a "flock" of autonomous agents, where each agent independently explores subsets of variables using the causal discovery algorithm DODO and proposes local structural priors. Through a consensus-driven negotiation protocol, agents debate the global causal structure, merging their localized findings into a globally consistent Directed Acyclic Graph (DAG). We evaluated FLODO in a rigorous synthetic setting, systematically varying node counts, edge densities, and noise distributions. Our experimental results highlight the difference in performance between various scenarios, and we speculate on the areas of improvement of the collaborative protocol.

Keywords: Causal Discovery · Collaborative Planning · Agentic AI.

1 Introduction

Modern causal discovery seeks to identify the unique directed acyclic graph that represents the generative process of a system of random variables [13, 16]. Although centralized statistical methods are well-established, they are often incompatible with the distributed nature of contemporary data architectures[10], where privacy constraints and communication costs prohibit global data pooling [11, 9, 8]. To bridge this gap, we present an empirical analysis of FLODO (Flock of Dodos), a decentralized causal discovery algorithm for multi-agent settings. In this framework, the global search space is partitioned among autonomous agents that perform local structural inference on variable subsets. These localized findings are integrated into a globally consistent directed acyclic graph through a consensus-driven negotiation protocol. Our study evaluates the efficacy of this collaborative approach across a variety of synthetic benchmarks, examining the impact of graph dimensionality, edge density, stochastic noise, and different agent

strategies. The results delineate the performance boundaries of the FLODO protocol and suggest specific refinements for decentralized structural alignment.

2 Brief Overview Causal Discovery

We define a Structural Causal Model (SCM) as a quadruple $M = \langle U, V, F, P(u) \rangle$, where U denotes a set of exogenous background variables governed by a probability distribution $P(u)$, and V denotes a set of endogenous observed variables. These variables are determined by a set of structural equations $F = \{f_i\}_{v_i \in V}$ such that each $v_i \leftarrow f_i(pa_i, u_i)$ is a deterministic function of its parents $pa_i \subseteq V$ and error terms $u_i \subseteq U$. This model induces a **Directed Acyclic Graph (DAG)** G, where directed edges represent direct functional dependencies [13]. Causal inference within this framework relies on the **do-operator**, denoted $do(X = x)$, which simulates a physical action by replacing the structural equation for X with the constant $X = x$. This induces a sub-model M_x where all arrows entering X are removed while the distribution $P(u)$ remains invariant, a property known as **modularity** or **invariance** [3]. Consequently, the interventional distribution $P(Y = y|do(X = x))$ describes the probability of $Y = y$ in M_x, which is generally distinct from the observational conditional probability $P(Y = y|X = x)$ due to confounding. To bridge this gap, the **Back-Door Criterion** identifies a set of variables Z that, when conditioned upon, blocks all "spurious" paths between X and Y, allowing the interventional distribution to be expressed in terms of purely observational data via the **adjustment formula** [12]. Extending this to the unit level, a counterfactual quantity $Y_x(u)$ is defined as the solution for Y in the sub-model M_x given a specific background context u. The probability of a counterfactual statement is computed by the three-step process of **abduction** (updating $P(u)$ to $P(u|e)$ given evidence e), **action** (enforcing $do(X = x)$), and **prediction** (computing Y in the modified model), representing the highest level in causal hierarchy [2].

3 Collaborative Causal Discovery

The current State of the Art in Collaborative Causal Discovery addresses the problem of inferring a global Structural Causal Model $M^* = \langle U, V, F, P(u) \rangle$ from a set of partitioned datasets $\mathcal{D} = \{D_1, \ldots, D_k\}$ distributed across k decentralised agents, where direct data pooling is prohibited by privacy constraints or bandwidth limitations[6]. Contemporary methodologies typically formulate this as a constrained global optimisation problem, seeking a directed acyclic graph G that minimises a collective loss function $\mathcal{L}(G; \mathcal{D}) = \sum_{i=1}^{k} \mathcal{L}_i(G; D_i)$ subject to acyclicity constraints[18, 17], often employing consensus-based algorithms like the Alternating Direction Method of Multipliers (ADMM) to align the structural adjacency matrix $A \in \{0, 1\}^{|V| \times |V|}$ without exchanging raw observations [11]. A critical challenge in such decentralised settings is the violation of the Independent and Identically Distributed (I.I.D.) assumption, which requires the

adoption of invariant causal prediction frameworks where the structural equations $v_i \leftarrow f_i(pa_i, u_i)$ remain stable across diverse agent environments $e \in E$ even as marginal distributions $P(pa_i)$ vary [14]. This intersection of causal inference and distributed intelligence is central to the work of Franco Zambonelli and colleagues [4], who advance the paradigm of "causality-empowered agents" within Multi-Agent Systems (MAS); they posit that agents can transcend the identifiability limits of the standard Markov Equivalence Class (MEC) by leveraging their capability to perform local interventions $do(X = x)$. In this framework, the causal model functions not only as a statistical map but as a shared semantic artefact for coordination, allowing agents to prune the joint action space $\mathcal{A}$ by evaluating counterfactual independence relations $Y_{a,z} \perp W$ and thereby significantly reducing the sample complexity required to converge to optimal cooperative policies in stochastic environments [4].

4 FLODO approach to collaborative causal discovery

Let $\mathcal{A} = \{A_1, \ldots, A_k\}$ denote a set of k decentralized agents operating within a global environment governed by a latent Structural Causal Model (SCM) $M^* = \langle U, V, F, P(u) \rangle$. The collaborative FLODO algorithm orchestrates the distributed inference of the global causal graph G^* through DODO[7] (a sequential causal discovery protocol of alternating between distinct observational modes Φ_{obs} and interventional modes Φ_{int} with causal links identification and potential pruning), for each agent and lastly through a protocol of causal information pooling. In the initial DODO phase, each agent A_i collects an observational dataset $\mathcal{D}_i^{obs} \sim P(V_i)$ over its local variable set $V_i \subseteq V$, subsequently transitioning to an interventional phase where it performs a set of atomic interventions $do(X = x)$ to generate interventional data $\mathcal{D}_i^{int}$. Depending on the properties of the system, during the observation phase each agent can share the global timing of the observation phase (while still retaining observability for their own subgraph only), or have different observational timings; most importantly, the agents can either perform atomic interventions one after the other, each waiting for the previous agent to finish interacting with its own subsystem, or perform atomic interventions simultaneously on their own subgraph. Secondly, each agent constructs a local preliminary causal graph G_i via causal link detection and pruning as described in [7]; finally the protocol initiates a *Variable Alignment Phase* to resolve semantic heterogeneity, where agents are initially agnostic regarding the equivalence of their observed phenomena. This process establishes an ideally bijective mapping $\mu : \bigcup_{i=1}^{k} V_i \to \mathcal{V}_{global}$ that unifies latent, potentially overlapping variables across disjoint local observations into a global namespace. Upon establishing the unified variable set $\mathcal{V}_{global}$, the system enters the *Collaborative Completion Phase*, governed by a centralized global clock $\mathcal{T}$ that synchronizes agents sharing of information. In this stage, any agent A_i detecting a causal gap, defined as an undetermined edge or parameter in the combined graph $G_\cup$, issues a query $Q(X, Y)$ to the coalition. This query requests specific observational or interventional statistics from the peer agent A_j possessing the relevant access,

thereby effectively pooling the global causal knowledge without requiring the centralization of the raw dataset $\mathcal{D}_{\cup}$ and minimizing communication overhead.

5 Simulation Conditions

To systematically evaluate the proposed collaborative algorithms, we conducted a comprehensive set of simulations across a highly parameterized configuration space. Let $\mathcal{C}$ denote the complete set of experimental configurations. Each configuration $c \in \mathcal{C}$ is a tuple defined by the Cartesian product of environmental, topological, and algorithmic parameters. First, the underlying causal structure is modeled as a Directed Acyclic Graph (DAG), $\mathcal{G} = (V, E)$, pruned from an initialized Erdős-Rényi (ER) graph. The node set size is strictly defined as $|V| \in \mathcal{V} = \{10, 20\}$. The sparsity of the graph is controlled by an initial edge probability parameter ρ, which scales linearly to the final density of the pruned DAG. We evaluate five discrete tiers of density, $\rho \in \mathcal{P} = \{\rho_1, \rho_2, \ldots, \rho_5\}$. To ensure statistical robustness against topological artifacts, we sample graph structures using 20 distinct topology generation seeds, $S_{top} \in \{s_{t,1}, \ldots, s_{t,20}\}$. The partitioning of each agent's subgraph, $G_i = (V_i, E_i)$, is governed by an intra-subgraph topological distance parameter d. To ensure structural heterogeneity across conditions, we evaluate three discrete distance tiers, $d \in \mathcal{D} = \{d_1, d_2, d_3\}$, sampling nodes $v \in V_i$ at topological extremes (i.e., minimally or maximally distant). Stochasticity in the environment and agent behaviors is initialized via random number generator (RNG) seeds, $S_{RNG} \in \{s_{r,1}, s_{r,2}, s_{r,3}\}$. Each node in the environment is subjected to additive Gaussian noise, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where the noise variance scales across three severity tiers ranging from negligible to severe: $\sigma \in \Sigma = \{\sigma_1, \sigma_2, \sigma_3\}$. The collaborative sampling algorithm operates under a defined sampling budget B, representing the total allowable sampling instances. This budget is swept from 400 to 3400 with a step increase of 200, yielding the set $\mathcal{B} = \{b \mid b = 400{+}200k, k \in \mathbb{N}, b \leq 3200\}$. In the multi-agent context, the number of participating agents is denoted by $N_A \in \mathcal{A} = \{2, 3, 4\}$. Each agent is equipped with a policy drawn from the strategy space $\Omega = \{\omega_1, \omega_2\}$, corresponding to the two strategies of either sequential interventions across agents, or simultaneous agent interventions. Furthermore, the observational data gathering period is parameterized by a temporal synchronization variable $\tau \in \mathcal{T} = \{\tau_{sync}, \tau_{async}\}$. Under τ_{sync}, the observation time interval is identical for all agents, whereas under τ_{async}, observations occur at disjoint or varied moments; in both cases, agents are restricted to observing their respective localized subgraphs $\mathcal{G}_i \subseteq \mathcal{G}$. Accounting for all combinations of the aforementioned parameters, the total number of permutations evaluated in our simulations reaches $1,458,000$. We evaluate structural recovery using the F_1 score to address the inherent edge sparsity and class imbalance typical of Directed Acyclic Graphs (DAGs)[15, 1]. By balancing precision and recall, the F_1 score provides a robust measure of topological accuracy that penalizes both spurious discoveries and omitted links without being skewed by the vast majority of absent edges[5].

6 Results

In this section we evaluate the performance of the collaborative discovery algorithm; all of the results shown refer to the more complex case of 20 nodes graph. Fig. 1 showcases the complexity of the scenario at hand, as well as the wide range of performance across the 10 most performing simulations for the F1 score exclusively for graphs of 20 nodes, for each combination of the hyperparameters density, noise level, number of agents involved, whether they shared an observation period, they acted simultaneously or sequentially and lastly, if they could observe subgraphs with varying levels of geographical/topological consistency across nodes.

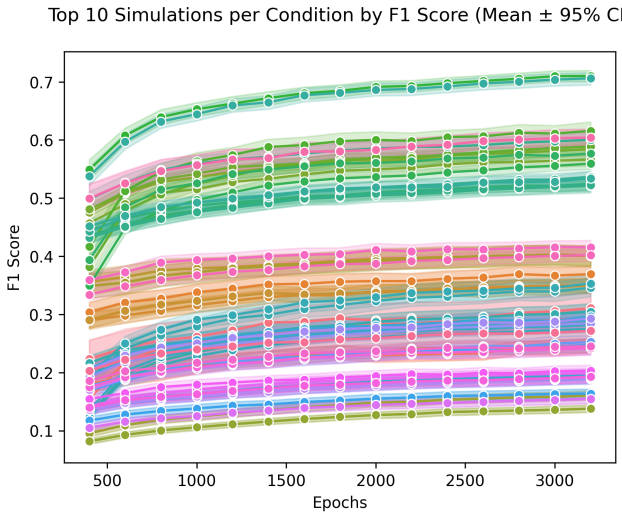

Fig. 1. F_1 score for the most performing 10 simulations, for each combination of agent operating condition, 20-node graphs; mean ± 95% confidence interval.

In order to better ascertain the conditions most responsible for different performances, we conducted a parameter impact analysis, as shown in Fig. 2, calculating the correlation coefficient between a given hyperparameter and the F1 score at a specific epoch. By observing the asymptotic behavior of the correlation, we can partition the hyperparameter set into distinct subsets based on their impact. We identify a positive impact subset consisting of parameters that exhibit a strong positive correlation with the F1 score, namely the full graph nodes ID (signaling wether the agents have correctly identified all nodes as distinct) and density, proving that the final pooling of information is dependent on the proper distinction of each node in the causal graph. We also identify a negative impact subset containing parameters that exhibit a noticeable negative correlation, specifically noise percentage and multiple interventions, where increases

systematically degrade the score. In the case of multiple interventions, given its coding as a binary variable, this translates into degraded performance when the agents conduct multiple atomic interventions simultaneously in each subgraph, increasing cross-contamination of causal information. Finally, there is a negligible impact subset where the correlation approaches zero, implying statistical independence from the F1 score in marginal isolation, which includes shared observations (whether or not the agents could observe the system at the same or different time) and the separation parameter (defining the average topological distance between nodes for each subgraph) .

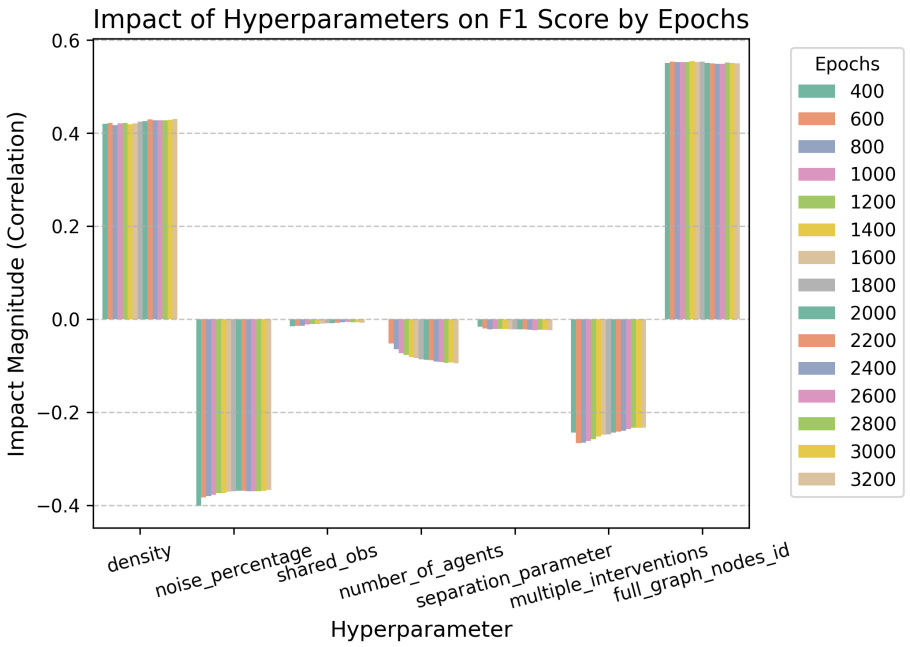

Fig. 2. Impact Analysis for the Correlation measure between selected hyperparameters and the F_1 score

We can evaluate the stability of the hyperparameters' impact over time by evaluating the rate of change of the correlation with respect to the epochs. For the vast majority of hyperparameters, excluding the number of agents, the correlation is highly stable across the evaluated interval. This suggests that the fundamental relationship between these parameters and the F1 score is established early and does not dynamically shift as the network continues to train. However, the number of agents exhibits a distinct, monotonic dynamic behavior. The magnitude of its negative correlation increases over time; specifically, the correlation drifts further negative, indicating that the negative impact of increasing the number of agents compounds as the model trains for more epochs.

This is due to compounding errors in the information sharing protocol between agents, and to the aforementioned cross contamination in a higher number of simultaneous atomic multiple interventions, as the number of agents increases. To optimize the causal discovery process, FLODO should prioritize the correct identification of all nodes in the nodes recognition phase of causal information sharing across agents, while actively minimizing noise percentage and multiple interventions. Furthermore, shared observations and the separation parameter can be ignored to reduce the dimensionality of the hyperparameter search space.

7 Conclusion

This study rigorously evaluates the FLODO algorithm within a distributed multi-agent framework designed to infer a global structural causal model from decentralized data. The empirical analysis, quantified through a balanced causal reconstruction accuracy metric, establishes clear partitions within the hyperparameter space based on their correlation with structural recovery performance. We define a subset of parameters exhibiting a strong positive impact, specifically the global node identification mechanism prior to the agents information pooling, which must be assured to optimize the structural fidelity of the inferred global directed acyclic graph. Conversely, variables such as environmental noise and the possibility of multiple simultaneous atomic interventions demonstrate a robust negative correlation, systematically degrading the accuracy of the model. Notably, while the influence of most parameters remains temporally invariant throughout the training epochs, the cardinality of the agent coalition exhibits a monotonically compounding negative impact over time. Consequently, to optimize the collaborative consensus protocol, the parameter search space can be effectively reduced by fixing stochastically independent variables, such as shared observation scenarios and the topological separation diversification of the subgraphs, which demonstrate negligible marginal impact. Ultimately, achieving optimal structural identifiability in such decentralized topologies hinges fundamentally on the precise resolution of the global namespace during the node recognition phase prior to iterative structural sharing of causal information.

References

1. Averin, P., Mellidou, I., Ganopoulou, M., Xanthopoulou, A., Moysiadis, T.: Evaluating directed acyclic graphs with dagmetrics: Insights from tuber and soil microbiome data. Agronomy **15**(4), 987 (2025). https://doi.org/10.3390/agronomy15040987
2. Bareinboim, E., Correa, J.D., Ibeling, D., Icard, T.: On pearl's hierarchy and the foundations of causal inference. In: Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 507–556. ACM (2022)
3. Bareinboim, E., Pearl, J.: Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences **113**(27), 7345–7352 (2016)

4. Briglia, G., Mariani, S., Zambonelli, F.: A roadmap towards improving multi-agent reinforcement learning with causal discovery and inference. arXiv preprint arXiv:2503.17803 (2025)
5. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). pp. 233–240 (2006). https://doi.org/10.1145/1143844.1143874
6. Gao, E., Chen, J., Shen, L., Liu, T., Gong, M., Bondell, H.: Federated causal discovery. arXiv preprint arXiv:2112.03555 (2021)
7. Gregorini, M., Boldrini, C., Valerio, L.: Dodo: Causal structure learning with budgeted interventions (2025), https://arxiv.org/abs/2510.08207
8. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Foundations and Trends® in Machine Learning **14**(1–2), 1–210 (2021)
9. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine **37**(3), 50–60 (2020)
10. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 1273–1282. PMLR (2017)
11. Ng, I., Zhang, K.: Towards federated bayesian network structure learning with continuous optimization. In: International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 2329–2352. PMLR (2022)
12. Pearl, J.: Causal diagrams for empirical research. Biometrika **82**(4), 669–688 (1995)
13. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, 2nd edn. (2009)
14. Peters, J., Bühlmann, P., 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)
15. Rehak, J., Falkenstein, A., Doehner, F., Beyerer, J.: Metrics for the evaluation of learned causal graphs based on ground truth. In: Machine Learning for Cyber Physical Systems (ML4CPS). pp. 1–13. Springer (2024). https://doi.org/10.24405/15305
16. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, Prediction, and Search. MIT Press, 2nd edn. (2000)
17. Yang, D., He, X., Wang, J., Yu, G., Domeniconi, C., Zhang, J.: Federated causality learning with explainable adaptive optimization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16308–16315 (2024)
18. Zheng, X., Aragam, B., Ravikumar, P.K., Xing, E.P.: Dags with no tears: Continuous optimization for structure learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)