Semantic Variable Construction for Causal Discovery in Persuasive Dialogues

Anonymous ACL submission

Abstract

We investigate the problem of recovering causal structures from persuasive dialogues, i.e., how utterance-level strategies, persuadee responses, personality variables, and outcomes are causally organized throughout interaction. A central challenge is that persuasive dialogue annotations, such as persuasion strategies and response labels, are symbolic interaction states that cannot be directly used as variables for causal discovery. To address this issue, we propose a causal discovery framework that constructs continuous semantic variables from utterance-level symbolic annotations by estimating label-specific semantic values from dialogue context. We then combine these semantic variables with personality and outcome variables and apply causal discovery over persuasive interactions to recover interpretable causal structures. Since the recovered graph may contain edges that are difficult to interpret, we apply an edge filtering step based on semantic and temporal plausibility, and further evaluate whether the retained causal candidates remain stable across different sample sizes. Experiments on persuasive donation dialogues show that the recovered graphs reveal meaningful causal dynamics across six structurally plausible edge types. The results further suggest that semantic variable construction substantially affects the interpretability and stability of recovered causal structures.

1 Introduction

Existing studies on persuasive dialogue mainly focus on recognizing persuasion strategies, predicting persuasion outcomes, or generating persuasive responses (Wang et al., 2019; Samad et al., 2022; Song and Wang, 2024; Zeng et al., 2025b,a). While effective for prediction and generation, these approaches provide limited insight into the underlying causal structure of persuasion: how persuader strategies, persuadee responses, user traits, and final decisions are interdependently organized throughout the interaction. As a result, persuasive dialogue is often modeled as a sequence prediction problem rather than as a structured causal process. In this work, we address the problem of recovering interpretable causal mechanisms from persuasive dialogue.

A key challenge for causal discovery (Pearl, 2009; Spirtes et al., 2000) is variable construction, i.e., particularly when variables of interest are not directly measurable but instead represented through symbolic or latent interaction states. Persuasive conversations contain directly observable variables, such as agreement, donation amount, and personality traits, as well as symbolic dialogue annotations that include persuasion strategies and response labels. While these annotations describe semantic interaction states, they cannot be directly used as variables for causal discovery because they lack continuous and context-sensitive representations. Consequently, naïvely applying causal discovery to symbolic dialogue labels may produce noisy or weakly interpretable causal structures.

To address this issue, we construct continuous semantic variables from utterance-level dialogue annotations, where each label is treated as a variable and its value is estimated from the corresponding dialogue context. We then combine semantic variables with personality and outcome variables and apply the Greedy Relaxation of the Sparsest Permutation (GRaSP) algorithm for causal discovery (Lam et al., 2022). To improve interpretability, we then utilize an edge filtering step based on semantic and temporal plausibility, subsequently evaluating whether the retained causal candidates remain stable across different sample sizes.

Our experiments on persuasive donation dialogues show that the recovered causal graphs reveal interpretable persuasion structures, including persuader strategy organization, persuadee decision progression, and personality-conditioned outcome tendencies. The results further suggest that

graph quality is sensitive to semantic variable construction, while the edge filtering step and sample stability-awareness improve plausibility and stability.

The main contributions of our work are as follows. (***i***) We formulate persuasive dialogue analysis as a causal mechanism discovery problem, aiming to understand the causal logic behind persuasion. (***ii***) We construct continuous semantic variables from symbolic dialogue annotations and combine them with personality and outcome variables for personalized causal discovery. (***iii***) Lastly, we introduce persuasion-specific structural constraints and multi-scale sample evaluations to effectively eliminate process-inconsistent dependencies and isolate robust causal pathways.

2 Persuasion-aware Causal Discovery

We propose a persuasion-aware causal discovery framework for recovering the causal logic in persuasive dialogue. This consists of three stages: semantic variable construction, GRaSP-based causal discovery, and causal-edge refinement.

2.1 Problem Formulation

Given a persuasive dialogue from the PersuasionForGood (P4G) dataset (Wang et al., 2019)

$$\mathcal{D} = \{(u_t, s_t, r_t)\}_{t=1}^{T}, \tag{1}$$

where u_t is the utterance at turn t, $s_t \in \{\text{ER}, \text{EE}\}$ denotes speaker role (persuad<u>ER</u> or persuad<u>EE</u>), and r_t is the corresponding strategy or response label, our goal is to learn a causal graph over dialogue, personality, and outcome variables.

Let $\mathcal{X}$ denote persuader-side strategy variables, $\mathcal{Y}$ denote persuadee-side response variables, $\mathbf{p}$ denote personality variables, and $\mathbf{o}$ denote outcome variables such as agreement and donation amount. We aim to infer a directed graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \quad \mathcal{V} = \mathcal{X} \cup \mathcal{Y} \cup \mathbf{p} \cup \mathbf{o}, \tag{2}$$

where each edge in $\mathcal{E}$ represents a potential causal dependency among variables involved in the persuasion process. In our formulation, personality variables are treated as pre-dialogue covariates, while semantic dialogue variables and outcomes are interpreted as evolving interaction states in the persuasion process.

2.2 Semantic Variable Construction

A central challenge in persuasive dialogue analysis is that persuasion strategies and response labels are symbolic annotations rather than directly usable numerical causal variables. We therefore transform each utterance-level label into a continuous semantic variable. Specifically, each label ℓ_k is treated as a semantic variable whose value is estimated from the corresponding utterance or dialogue context.

Given an utterance u_i with annotated label ℓ_i, we obtain its representation $\mathbf{z}_i$ using `all-mpnet-base-v2` (Reimers and Gurevych, 2019; Song et al., 2020). Let $\mathbf{q}_k$ denote the representation of candidate label ℓ_k, and K be the number of role-specific labels (27 for ER and 23 for EE). In the similarity-based setting, we use an annotation-aware semantic score:

$$x_{i,k}^{\text{sim}} = \lambda \mathbb{I}[\ell_i = \ell_k] + (1-\lambda)\frac{1+\cos(\mathbf{z}_i, \mathbf{q}_k)}{2}, \tag{3}$$

where $\lambda = 0.8$, and $\mathbb{I}[\ell_i = \ell_k]$ indicates whether u_i is annotated with ℓ_k. This keeps the annotated label dominant while allowing semantically related labels to receive weaker continuous values.

Alternatively, classifier-based semantic variables use the predicted probability of a fine-tuned RoBERTa (Liu et al., 2019) classifier:

$$x_{i,k}^{\text{clf}} = P(\ell_k \mid u_i) = \frac{\exp(h_{i,k})}{\sum_{j=1}^{K}\exp(h_{i,j})}, \tag{4}$$

where $h_{i,k}$ is the classifier logit for label ℓ_k. The classifiers are evaluated using 5-fold cross-validation and achieve approximately 85% top-3 accuracy for both persuader- and persuadee-side prediction. This is acceptable because dialogue utterances exhibit context-dependent semantic ambiguity.

The semantic variable vector for u_i is $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,K}]$. After aggregating utterance-level variables over each dialogue, we combine them with personality variables $\mathbf{p} \in \mathbb{R}^{32}$ and outcome variables $\mathbf{o} \in \mathbb{R}^{2}$:

$$\mathcal{V} = \{\mathbf{x}, \mathbf{p}, \mathbf{o}\}. \tag{5}$$

This formulation enables causal discovery over semantically grounded continuous variables rather than discrete dialogue symbolic labels alone.

2.3 Causal Discovery and Edge Filtering

We apply GRaSP to the constructed variable set $\mathcal{V}$ (from Eq. 5). GRaSP is suitable for

high-dimensional causal discovery and efficiently searches for a sparse directed causal structure over these variables by optimizing graph sparsity under conditional independence constraints (Lam et al., 2022). The learned structure is interpreted as a persuasion mechanism graph that captures causal dependencies among semantic dialogue states, personality variables, and outcomes.

The causal graph recovered by GRaSP may contain edges that are difficult to interpret in persuasive dialogue. We therefore apply an edge filtering step based on semantic and temporal plausibility. Since personality and demographic variables are treated as pre-dialogue covariates, edges from dialogue labels, agreement, or donation outcomes to these variables are excluded from the main interpretation.

We retain edges that are consistent with plausible persuasive-dialogue relations, including personality or demographic variables to dialogue labels or outcomes, persuader strategy to persuader strategy, persuader strategy to persuadee response, persuadee response to persuadee response, and dialogue label to outcome relations. This filtering step defines the set of interpretable causal candidates considered in our analysis.

The reliability of these retained causal candidates is not assumed from a single discovered graph. Instead, we evaluate whether they remain stable when the amount of data changes, refer to section 3.1. This connection between filtering and stability allows us to distinguish semantically plausible but sample-specific edges from semantically plausible and robust causal candidates.

3 Experiment

We evaluate the framework on 300 annotated dialogues from the P4G to assess its ability to recover interpretable and structurally plausible causal graphs. Since no ground-truth causal graph is available, we assess graph quality using two complementary criteria: edge plausibility and sampling. In particular, we examine whether different semantic variable construction strategies lead to more coherent and interpretable causal structures.

3.1 Evaluation Protocol

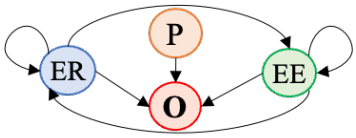

Figure 1: Plausible causal structure in persuasion.

Edge plausibility. We evaluate whether each discovered edge can be interpreted under the semantic and temporal assumptions of persuasive dialogue. Let P, ER, EE, and O denote personality variables, persuader strategies, persuadee responses, and outcomes, respectively. Since P is measured before the dialogue and O represents final donation outcomes, edges from dialogue or outcome variables to P, as well as edges from O to non-outcome variables, are treated as invalid for interpretation. As shown in Figure 1, we consider six edge types as plausible: $P \rightarrow (ER, EE, O)$, $ER \rightarrow ER$, $ER \rightarrow EE$, $EE \rightarrow EE$, $EE \rightarrow ER$, and $(ER, EE, P) \rightarrow O$. We report PER as the proportion of valid edges that match these plausible types, and IDR as the proportion of edges with invalid directions.

Sampling stability. We evaluate whether core persuasion-related edges persist as the amount of data increases. Causal discovery is performed using a doubling schedule of sample sizes: $\mathcal{N} = \{500, 1000, 2000, 4000, 8000, N_{\text{full}}\}$, with $N_{\text{full}} = 10864$. For each edge e, we compute its persistence, where $\mathcal{E}_n$ is the edge set discovered using n ($\in N$) samples:

$$\text{Persistence}(e) = \frac{\sum_{n \in \mathcal{N}} \mathbb{I}[e \in \mathcal{E}_n]}{|\mathcal{N}|}. \quad (6)$$

3.2 Edge Plausibility Analysis

	Similarity-based				Classifier-based			
Setting	**PER**	**IDR**	**out**	**cross**	**PER**	**IDR**	**out**	**cross**
S	1.000	0.000	1	0	1.000	0.000	0	96
C	0.997	0.000	1	0	1.000	0.000	0	146
C&P	0.569	0.051	58	2	0.647	0.013	12	153
C&P(*)	0.621	0.009	25	1	0.683	0.002	3	65

Table 1: Ablation over semantic variable construction (**S**ingle, **C**ontext, and **P**ersonality), *–refined.

Table 1 shows that semantic variable construction strongly affects graph interpretability. The similarity-only settings, *sim*(S) and *sim*(C), produce highly plausible but sparse graphs, with almost no invalid directions and only one outcome-related edge. Adding personality variables in *sim*$(C\&P)$ greatly increases outcome-related edges, suggesting that user covariates help reveal donation-related dependencies, albeit the lower PER indicates a more complex graph. After refinement, *sim*$^*(C\&P)$ reduces invalid directions and

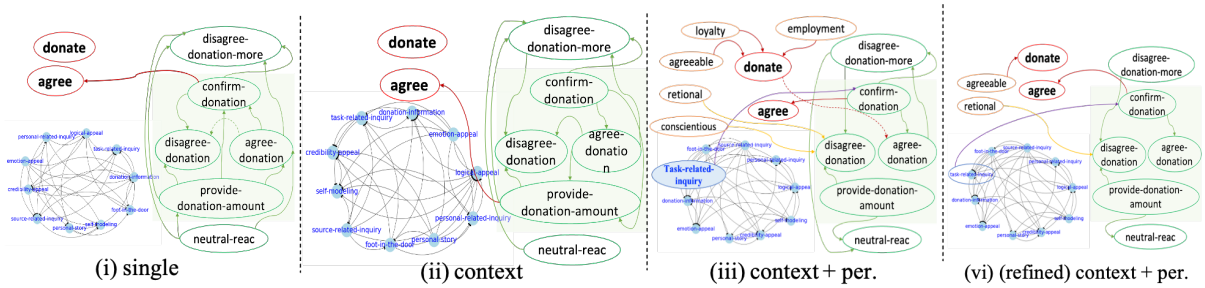

Figure 2: Comparison of recovered causal subgraphs across different variable configurations: single, +context, and +context&personality, using the cosine similarity method.

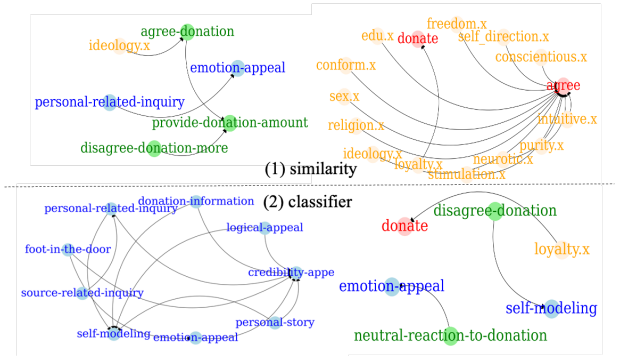

Figure 3: Causal subgraph showing persistent edges (Persistence=1.0) across six sample size steps.

preserves many outcome-related edges, indicating that filtering removes some process-inconsistent dependencies while keeping donation-relevant structure. In contrast, classifier-based settings produce many cross-role edges but few outcome-related edges. Although $cls(S)$ and $cls(C)$ have high PER, they do not explain donation outcomes. Even after adding personality and refinement, classifier-based graphs retain fewer outcome-related relations than similarity-based graphs. Overall, similarity-based variables better support interpretable, outcome-oriented causal analysis, while classifier-based variables produce denser cross-role structures.

Figure 2 shows that context and personality variables change the recovered similarity-based graph. Adding context reveals earlier causes of the final agreement outcome, whereas incorporating personality features reveals novel interaction dynamics from the persuadee to the persuader ($EE \rightarrow ER$). After refinement, non-persistent edges are removed, producing a sparser graph that preserves the main donation-related structure.

3.3 Sampling Stability Analysis

We evaluate whether recovered edges persist as the sample size increases. Figure 3 shows the edges with Persistence $=$ 1.0, i.e., edges that appear across all six sample sizes. In the similarity-based graph, stable edges reveal three patterns: personalities consistently point to `agree`. `ee_agree-donation` reliably leads to `ee_provide-donation-amount`, and persuader strategies form stable internal structures such as `er_self-modeling`$\rightarrow$`er_personal-story` and `er_source-related-inquiry`$\rightarrow$`er_credibility-appeal`.

Several donation-decision related edges also appear from 1000 samples onward, including `ee_agree-donation`$\rightarrow$`agree`, `ee_agree-donation`$\rightarrow$`ee_confirm-donation`, `ee_provide-donation-amount`$\rightarrow$`ee_confirm-donation`, and `ee_provide-donation-amount`$\rightarrow$`agree`. These results suggest that donation agreement emerges through intermediate commitment states rather than a single direct transition. In contrast, classifier-based variables mainly preserve stable persuader-strategy relations, indicating that similarity-based variables better capture outcome-oriented decision progression.

4 Conclusion

We formulated persuasive dialogue as a causal discovery problem. We proposed a framework that transforms symbolic dialogue annotations into continuous semantic variables and combines them with personality and outcome variables for causal discovery. With this combination, the framework recovers interpretable persuasion structures from conversational interaction. Our results suggest that semantic variable construction substantially influences the plausibility, stability, and interpretability of the recovered causal graphs. In particular, context-aware and personality-aware semantic representations produce more coherent persuasion dynamics, while persuasion-aware graph refinement improves structural consistency by filtering weak or temporally inconsistent edges.

References

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. 2022. Greedy relaxations of the sparsest permutation algorithm. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1052–1062. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press, Cambridge, UK.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867.

Yuhan Song and Houfeng Wang. 2024. Would you like to make a donation? a dialogue system to persuade you to donate. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17707–17717, Torino, Italia. ELRA and ICCL.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2 edition. MIT Press, Cambridge, MA.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Donghuo Zeng, Roberto Legaspi, Yuewen Sun, Xinshuai Dong, Kazushi Ikeda, Peter Spirtes, and Kun Zhang. 2025a. Causal discovery and counterfactual reasoning to optimize persuasive dialogue policies. *BiT*, pages 1–15.

Donghuo Zeng, Roberto Legaspi, Yuewen Sun, Xinshuai Dong, Kazushi Ikeda, Peter Spirtes, and Kun Zhang. 2025b. Generative framework for personalized persuasion: Inferring causal, counterfactual, and latent knowledge. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 83–93.

A Limitations

The main limitation of this is that the recovered graph depends on the construction of semantic variables from symbolic dialogue labels. Consine-similarity-based variables depend on the quality of label definitions, prototypes, and embedding modes, and may capture general semantic relatedness rather than the pragmatic role of an utterance in persuasion. Classifier-based variables depend on classifier accuracy and calibration, and may be affected by annotation bias, class imbalance, or confusion among similar labels. In addition, this work is based on the Persuasion for Good (P4G) dataset, which focuses on charitable donation dialogues. Therefore, the recovered mechanisms may not directly generalize to the other persuasion domains. The discovered edges should be interpreted as causal candidates under specific representation and dataset assumptions, rather than definitive causal effects.

B Ethical considerations

Persuasive dialogue analysis can be misused to design systems that manipulate users. This work aims to improve the interpretability of persuasion mechanisms, not to optimize coercive or deceptive persuasion. Personality variables are used only as observed covariates for aggregate analysis, and we avoid drawing conclusions about individual users. Edges involving personality variables are interpreted cautiously, particularly when they violate temporal assumptions. Future applications of this research should ensure user consent, transparency, and safeguards that preserve user autonomy.