

Discovering Root Causes of Risks Using Counterfactual Knowledge Graphs (CKG)

Ying Zhao¹[0000–0001–8350–4033], Gabe E. Mata², Jesse Zhou³, and Charles Zhou³[0000–0001–9598–015X]

¹ Naval Postgraduate School, Monterey, USA
yzhao@nps.edu

² USMC, San Diego, CA, USA
gabino.mata@usmc.mil

³ Quantum Intelligence, Inc., Northern Monterey County, CA, USA
charles.zhou@quantumii.com

Abstract. Finding the root causes for risks regarding warfighting systems for the U.S. Marine Corps is critical for designing policies and taking actions in strategic, operational, and tactical levels to prevent future risks. These risks may be attributed to multiple factors and the data sources might reside in distributed environment and often are difficult to share. In this paper, we show collaborative learning agents (CLAs) to analyze data sources separately and then fuse the patterns and models using shared vocabularies in the patterns instead of fusing raw data. We also show causal learning, counterfactuals reasoning, and knowledge graphs to discover root causes from structured data. These counterfactuals knowledge graphs from unstructured data serve as context inputs to a large language model, which can generate more relevant and meaningful descriptive root cause analysis. We demonstrate the methodologies using the use case from the mishap and incident reports of a marine transportation equipment and related data resources from the I Marine Expeditionary Force (IMEF).

Keywords: counterfactual knowledge graph · CKG · causal learning · root causes · human factors · distributed systems · collaborative learning agents

Introduction

The U.S. Marine Corps (USMC) faces significant challenges in adapting its operational frameworks to embrace modern technologies, notably in the integration of advanced data sciences and AI/ML. For example, finding the root causes for risks such as mishaps or incidents regarding a warfighting system is critical for designing policies and taking actions in strategic, operational, and tactical levels to prevent future risks. These risks may be attributed to multiple factors, especially human factors such as who operates/drives certain equipment. The data sources detailing the incidents, systems, and operators might reside in distributed environment, sometime they are difficult to access without necessary

policies and rules to justify the needs. For example, operators' data may include personal identifiable information (PII), demographics, and personal history. In the context of WHI, different end users may not have the rights to share data directly because of sensitive nature of human elements. Furthermore, these data sets can be structured such as age, gender, education, military occupational specialty (MOS), etc., or unstructured such as descriptions of incidents.

It is critical to apply causal learning [2], counterfactual reasoning, and related theory to address the USMC challenges. The key factors for causal learning, according to [9], includes the three layers of a causal hierarchy - association, intervention, and counterfactuals.

0.1 Association

The common consensus is that data-driven analysis or data mining can discover initial statistically significant correlations and associations from big data. For example, an association can be usually represented as a conditional probability $P(E|C)$. A high conditional probability is a necessary condition for a causal relation. However, it is not a sufficient condition.

0.2 Intervention

In machine learning practices, the associations, correlations or probabilistic rules are typically cross-validated using separate or new data sets. In causal learning, this relates to the intervention layer. A typical question at this level would be: What will happen if one takes an action? For example, instead of examining $P(E|C)$, one might further ensure C is actionable. In other words, $P(E|do(C))$ [2] can be examined. The answers to the question are more than just mining the existing data. The action needs to generate new observable data as an effect of the intervention to determine if the underlying action causes the effect. When a factor X represents an observed characteristics of an entity such as age, gender, and health status, although the "do" requirement does not meet, $P(C|X)$ represents propensity of the characteristics with respect to the actionable cause.

0.3 Counterfactuals

A typical question asked is: "What if I had acted differently?" or counterfactual reasoning. If $P(E|C)$ is high-probability rule discovered from data and C is actionable, $P(E|Not C)$, $P(Not E|C)$, and $P(Not E|Not C)$ are the counterfactuals needed in the reasoning. Traditionally, the counterfactual is defined as the effect of an action for an entity and for the same entity without the action, i.e., $P(E|C)$ and $P(E|Not C)$. However, since different effect is impossible to directly observe for the same entity, this is commonly referred to as the fundamental problem of causal inference [2]. The potential outcome or counterfactual-based model of casual inference explores the idea of an entity-level treatment effect, although it is unobservable as well, it can be estimated in various ways.

For example, the causal effect is typically measured using two randomized populations, one with the action (or with C) and another one without the action (Not C or control group). The two populations are randomized to ensure they are similar to each other (as if they were the same entity). This is the Randomized Control Treatment (RCT) theory [2], which is a standard practice found in social sciences, drug development, and clinic trials. Counterfactual knowledge graphs (CKG) are not considered formally and used as generic directed graphs in these previous studies.

Existing work [3] estimates propensities from interaction data when so called exposure data (i.e., whether an item has been recommended to a user or not), or the propensity scores (i.e., the probability of recommending/exposing an item to a user), are not observable at least during the training stage due to technical and privacy constraints. Existing work [4] applies invariant representations of causal inference, that is, the representation is not causally influenced by the environment. This formalization increases the resilience of anomaly detection models to different kinds of distribution shifts. Existing work [5] studies how to combine experiences for a learning agent who collects online experiences through direct interactions with the environment (interventional data), but has also access to a large collection of offline experiences, obtained by observing another agent interacting with the environment (observational data). The observed agent may interact with the environment based on hidden information, which is not observed by the learning agent. The method relies on learning a latent-based causal transition model that explains both the interventional and observational regimes. Existing work [8] develops an autoencoding variational Bayes algorithm using latent variables for causal disentanglement, with each intervention changing the mechanism of a latent variable. Existing work [6] learns causal structures via low-dimensional causal embedded spaces when the underlying graph involving the multivariate functions may have cycles. These existing works are mostly based the standard Bayesian framework, however, large-scale, data-driven counterfactual networks or knowledge graphs are rarely studied for real-life data [7].

The contribution of this paper is to demonstrate a systematic methodology using advanced data sciences and AI/ML to address these challenges in the following steps:

1. Provide distributed AI/ML agents infrastructure such as collaborative learning agents (CLAs) to analyze data sources separately and then fuse the patterns and models using shared vocabularies in the patterns instead of sharing raw data.
2. Design and apply CKGs for causal learning and reasoning to discover root causes of risks from structured and unstructured data. Leverage large language models (LLMs) to interpret the CKGs.

We demonstrate the methodologies using the use case from the mishap and incident reports of a marine transportation equipment and related data resources from the I Marine Expeditionary Force (IMEF).

1 Data Sets

The mishaps and incidents data from IMEF are described as follows:

- Data Set 1 (structured and unstructured): all incidents for IMEF private motor vehicles of two wheels (PMV2). The structured data include age, gender, training, and MOS; unstructured data include one-line descriptions of incidents. Two incident classes A and C are of interest of this paper. To simplify, we only focus on Incident Class A, which includes serious or fatal incidents, while incident Class C includes less serious incidents.
- Data Set 2 (structured and unstructured): all incidents of PMV2 for all USMC nationwide.
- Data Set 3 (structured): Training data for all USMC nationwide.

We use three CLAs for the three data sets and assume these data sets can not be shared and fused initially. However, these three CLAs can fuse the results by sharing vocabularies and patterns.

2 Methods

2.1 Traditional Statistical Graphs and Reports

One can usually start an analysis by generating traditional statistical graphs and reports from a data set. Fig. 1 shows an example of PMV2 incident breakdowns by organization and incident type within IMEF.

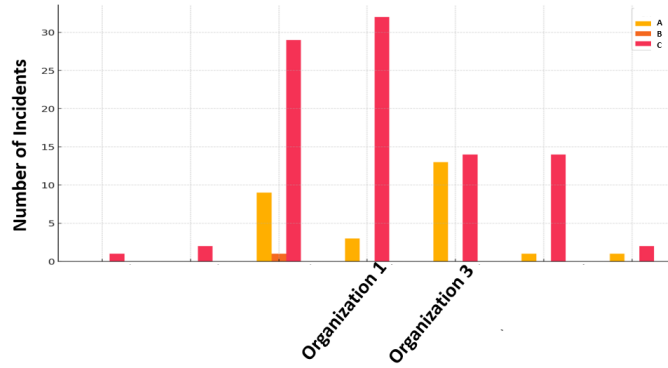


Fig. 1. Incident Breakdowns by Organization and Incident Type for IMEF Alone (Data Set 1)

2.2 Collaborative Learning Agents (CLA) [1]

A single CLA is capable of ingesting, indexing, and analyzing data; performing knowledge pattern discovery; and building AI/ML models from local data that the agent can access. As multiple CLAs work collaboratively in a peer-to-peer network, patterns are discovered by CLAs from different data sets. These patterns, which are not raw data sets, are fused through shared word features and vocabularies.

Fig. 2 shows that three CLA agents that can represent three different data sets. Each agent can has a peer list of other CLAs that it can share output patterns or knowledge graphs.

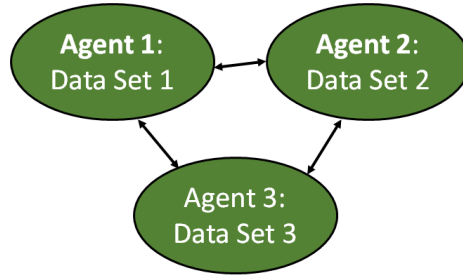


Fig. 2. Three CLA Agents Represent the Three Different Data Sets

2.3 Data Pre-processing

In each CLA, we first discretize structured data into word features: Each numeric variable and its value are represented in one of the intervals $(-\infty, m_k - d_k)$, $[m_k - d_k, m_k + d_k)$, $[m_k + d_k, \infty)$, where m_k, d_k represent the mean and standard deviation for the variable $k = 1, \dots$, respectively. Unstructured data are processed separately. In CLA, each agent learns in a list of features with specific vocabularies, lexicon terms, and/or values to describe the characteristics of the underlying domains the agents represent.

2.4 Patterns Fusing

Each CLA automatically pre-process structured or unstructured data into word features, clusters of word features, and displays them as word feature networks or knowledge graphs. Each CLA can include multiple indexes. Each index outputs a knowledge graph. Each node in the knowledge graph is word or word feature. The strength of the links between nodes are computed using causal learning and counterfactuals reasoning. The links between nodes are directional from a "cause" node to an "effect" node.

Each agent also has a peer list of other CLAs that it can share the output knowledge graph. Since the output knowledge graphs or patterns represented in counterfactuals knowledge graphs only include lexical terms found in the data, if two data sets share certain lexical terms, such as, age, and gender, the output patterns are fused from multiple CLAs.

2.5 Constructing CKGs

With recent data-driven approaches such as data mining and machine learning, if people can robustly estimate a local average effect using machine learning algorithms [10], causal reasoning is easier. These algorithms are mostly nonparametric models such as nearest neighbors and random forests [11], i.e., use the effect of the nearest neighbor of an entity as the surrogate for the unobservable effect of the same entity for the counterfactual action. In this paper, we show a CKG approach for causal reasoning, where the patterns from different CLAs are fused. Fig. 3 shows the relations of four counterfactuals.

We compute the counterfactual proportion difference in (1) as the strength of two associated two nodes E, C :

$$h_{EC} = \frac{[P(E|C) - P(E|Not\ C)]}{SE_0}, \quad (1)$$

where $P(E|C)$ is the probability of E occurs when C occurs and $P(E|Not\ C)$ is the probability of E occurs when C does not occur.

$$SE_0 = \sqrt{p(1-p) * (\frac{1}{N_C} + \frac{1}{N_{Not\ C}})}, \quad (2)$$

and

$$p = \frac{n_C + n_{Not\ C}}{N_C + N_{Not\ C}}, \quad (3)$$

where n_C , $n_{Not\ C}$, N_C , and $N_{Not\ C}$ are the number of times E occurs when C occurs, number of times E occurs when C does not occur, number of times C occurs in total, and number of times C does not occur in total, respectively.

Given that $P(E|C)$ and $P(E|not\ C)$ are derived from two independent samples since two counterfactual conditions are independent: $P(E|C)$ is estimated from samples when C occurs; $P(E|not\ C)$ is estimated when C does not occur, therefore, h_{EC} can be tested using the z-test of two independent proportions and it is a z-score [12]. If $h_{EC} > 1.96$, then the link between factors E and C has a statistically significant causal association in a directed graph ($C \rightarrow E$ and $p - value < 0.05$). If $h_{EC} < -1.96$, then the link between factors E and C has a statistically significant causal association in a directed graph ($Not\ C \rightarrow E$ and $p - value < 0.05$).

Further observations regarding the counterfactual knowledge graph can be summarized as follows:

- In a real-life application, effects are often observable data, one often wants to predict causes based on the data of effects, i.e., computing and validating $P(C|E)$. $P(C|E)$ is difficult to discover because causes are often hidden, anomalous, and capricious.
- Pattern Fusing: Each CLA first computes conditional probabilities $P(C|E)$, $P(E|C)$, $P(E)$, and $P(C)$ from the local data the agent can access directly. Each CLA has a peer list of other CLAs. Peers share the conditional probabilities $P(C|E)$, $P(E|C)$, $P(E)$, and $P(C)$. Each agent then updates the conditional probabilities $P(C|E)$, $P(E|C)$, $P(E)$, and $P(C)$ as if it has all the data in the peer network. Each agent also uses the relations among counterfactuals shown in (4), (5), (6), and (7) to update the counterfactuals.

$$P(E|C) - P(E|Not\ C) = P(Not\ E|C) - P(Not\ E|Not\ C) \quad (4)$$

$$= \frac{P(E)}{P(C)} \frac{P(C|E) - P(C)}{1 - P(C)} \quad (5)$$

$$P(C|E) - P(C|Not\ E) = P(Not\ C|E) - P(Not\ C|Not\ E) \quad (6)$$

$$= \frac{P(C)}{P(E)} \frac{P(E|C) - P(E)}{1 - P(E)} \quad (7)$$

Fig. 4 shows three causality relations, i.e., confounder, mediator, and collider as described in [3]. In Fig. 4(a), a confounder is a factor that causes both "Cause" and "Effect", therefore, usually is a hidden factor that might be discovered in the data to eliminate the relation between "Cause" and "Effect." In Fig. 4(b), a mediator is a factor that mediates a cause-and-effect relation so "Cause" and "Effect" relation might be realized and controlled. Fig. 4(c) shows a collider of "Cause" and "Effect" where their cause-and-effect relation might also be potentially true, however, not controlled by the collider.

3 Results and Discussion

Fig. 5 shows the part of the counterfactual graph generated from the three data sets. Solid arrows represent positive counterfactual strength computed from (1). Dotted lines represent negative counterfactual strength. We observe the following:

- From statistical analysis of Data Set 1, Organization 1 is associated with Incident Class C, Organization 3 is associated with Incident Class A. However, when fused with Data Set 2 and 3, the links change to -9 and -5 in the dotted lines in Fig. 5, respectively. The changes indicate there are no causal relations between these factors.
- Causal factors for Incident Class C are discovered from the fused data using three CLAs, listed as follows:
 - Month April
 - $26 < Age < 33$

- Female

Incident Class C is a collider for these causal factors, these causal factors can be potentially associated, however, associations are not statistically significant in the data sets.

- “Female” is a confounder factor for “Incident Class C,” “Organization 1,” and “ $1 < \textit{Training Counts} < 3$.”
- “Organization 3” is a mediator between “Rank Captain” and “ $\textit{Training Counts} \geq 5$,” therefore the link between “Rank Captain” and “ $\textit{Training Counts} \geq 5$ ” is a causal propensity. Similarly, “Organization 1” is a mediator between “Female” and “ $1 < \textit{Training Counts} < 3$,” therefore the link between “Female” and “ $1 < \textit{Training Counts} < 3$ ” is a causal propensity.
- “Training” is not a cause for “Incident Class A.” More data sets are needed to discover the causes.

Fig. 6 shows CLAs generate knowledge graphs as context from unstructured data (i.e., incident one-line descriptions from Data Set 1 and 2) and then feed them to a large language model (LLM) for summarizing and reasoning root causes for Incident Class A that link to the keywords “fatal” and “fatality.”

Prompt: Please analyze root causes based on this graph.

Answer from the Microsoft Copilot:

Root Causes Analysis:

1. Motorcycle-Related Incidents:
 - Keywords: motorcycle, ejected, collision, operating, ride, pmv2
 - Analysis: Motorcycle accidents appear frequently linked to fatal outcomes, often resulting from ejections during collisions and issues related to operating or riding motorcycles.
2. Internal and Traumatic Injuries:
 - Keywords: internal, traumatic, injuries, suffered
 - Analysis: Fatalities are frequently caused by severe internal or traumatic injuries, which indicates the critical need for immediate medical attention and safety improvements.
3. Loss of Control:
 - Keywords: lost, control, crashed
 - Analysis: Loss of control over vehicles (including PMV—Personal Mobility Vehicles) leads to fatal crashes. This highlights the importance of control measures and driver training.
4. Struck and Collision:
 - Keywords: struck, oncoming, guardrail, multiple, pmv4
 - Analysis: Collisions with oncoming traffic, guardrails, or multiple vehicles are significant factors. This suggests the necessity for better roadway design and preventive measures.
5. Critical and Threatening Situations:
 - Keywords: critical, threatening, resulting

- Analysis: Certain scenarios lead to critical or life-threatening situations, implying that effective intervention strategies could reduce fatalities.
- 6. Environmental Factors:
 - Keywords: tree, guardrail
 - Analysis: Environmental obstacles like trees and guardrails contribute to fatal outcomes. This underscores the need for better road maintenance and hazard removal.

Summary: The network graph indicates that root causes of fatalities largely revolve around motorcycle incidents, severe injuries, loss of vehicle control, collisions, and environmental obstacles. Improvements in vehicle safety, road infrastructure, and emergency medical response could mitigate these fatal accidents.

4 Conclusions

We show human factors and data sources can reside in distributed environment for the USMC data enterprise. Difficulties exist to share personal identifiable information (PII), demographics, and personal history. We show that how to apply CLA and causal learning to analyze root causes of multi-modality data such as structured data of age, gender, education, military occupational specialty (MOS), and unstructured such as descriptions of incidents in the distributed data environment. The results can be used to design policies and take actions in strategic, operational, and tactical levels to prevent future risks. For example,

- Find causes of causes behind the causal propensity factors such as female, age group of 26 to 33, and Month April by linking more data sources and CLAs.
- Start strategic education and communication to focus on certain age and gender groups with respect to PMV2’s body type.

Disclaimer

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the U.S. Government.

References

1. Zhao, Y. & Zhou C. (2014) System and method for knowledge pattern search from networked agents. US Patent 8,903,756
2. Mackenzie, D. & Judea Pearl (2018). The Book of Why: The New Science of Cause and Effect.
3. Liu, Z., Fang, Y., & Wu, M. (2023). Estimating Propensity for Causality-based Recommendation without Exposure Data. In the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

4. Carvalho, J., Zhang, M., Geyer, R., Cotrini, C., & Buhmann, J. (2023). Invariant Anomaly Detection under Distribution Shifts: A Causal Perspective. In the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).
5. Gasse, M. (2021). Causal Reinforcement Learning using Observational and Interventional Data <https://arxiv.org/pdf/2106.14421.pdf>
6. Roy, S., Wong, R., Ni, Y. (2023). Directed Cyclic Graph for Causal Discovery from Multivariate Functional Data. <https://arxiv.org/abs/2310.20537>
7. Guo, Z., Xiao, T., Wu, Z., Aggarwal, C., Liu, H., & Wang, S. (2024) Counterfactual Learning on Graphs: A Survey. <https://arxiv.org/pdf/2304.01391v3>
8. Zhang, J., Squires, C., Greenewald, K., Srivastava, A., Shanmugam, K., & Uhler, C. (2023). Identifiability Guarantees for Causal Disentanglement from Soft Interventions.
9. Pearl, J. (2018). The Seven Pillars of Causal Reasoning with Reflections on Machine Learning. Retrieved from http://ftp.cs.ucla.edu/pub/stat_ser/r481.pdf
10. Gelman, A. (2018) Donald Rubin. Retrieved from <http://www.stat.columbia.edu/gelman/research/published/rubin.pdf>
11. Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.
12. Penn State University (PSU), (2021). Online Statistics: Normal Approximation Method Formulas. <https://online.stat.psu.edu/stat200/lesson/9/9.1/9.1.2/9.1.2.1>

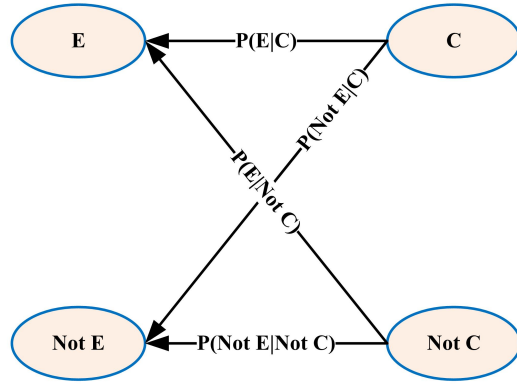


Fig. 3. Four Probabilistic Rules Related to Counterfactuals

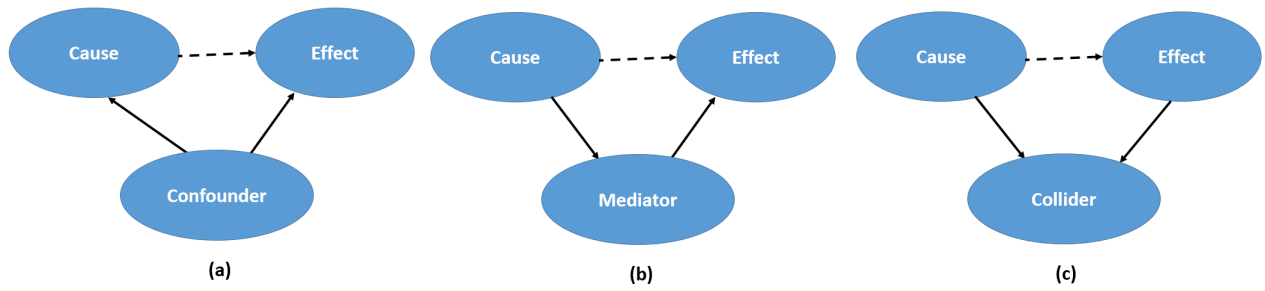


Fig. 4. Three Causality Networks: Confounder, Mediator, and Collider [3]

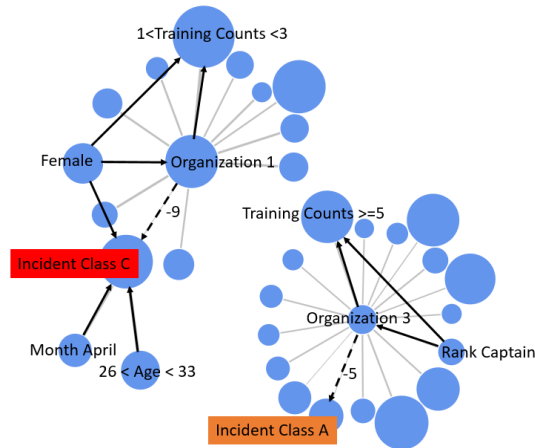


Fig. 5. Counterfactual Knowledge Graphs for the Fused Data Sets

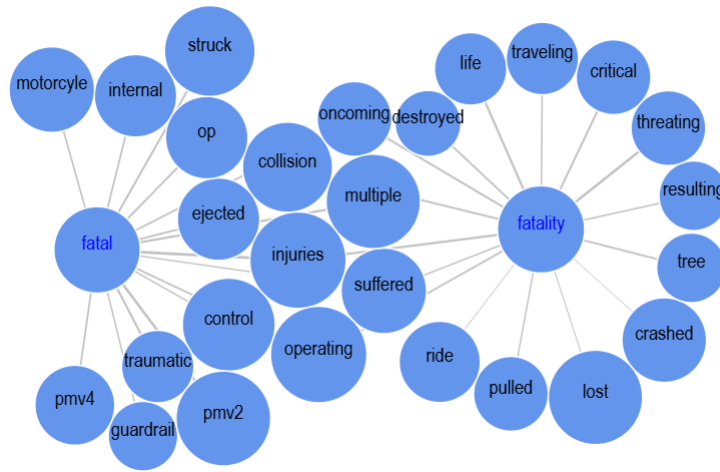


Fig. 6. CLAs Generate Knowledge Graphs from Unstructured Data and Feed Them to an LLM for Summarizing and Reasoning Root Causes