Causal Discovery on PHMSA Dataset

Shuai Feng[1], Yongming Liu[1], Hao Yan[1*]
[1] Arizona State University, [*] haoyan@asu.edu

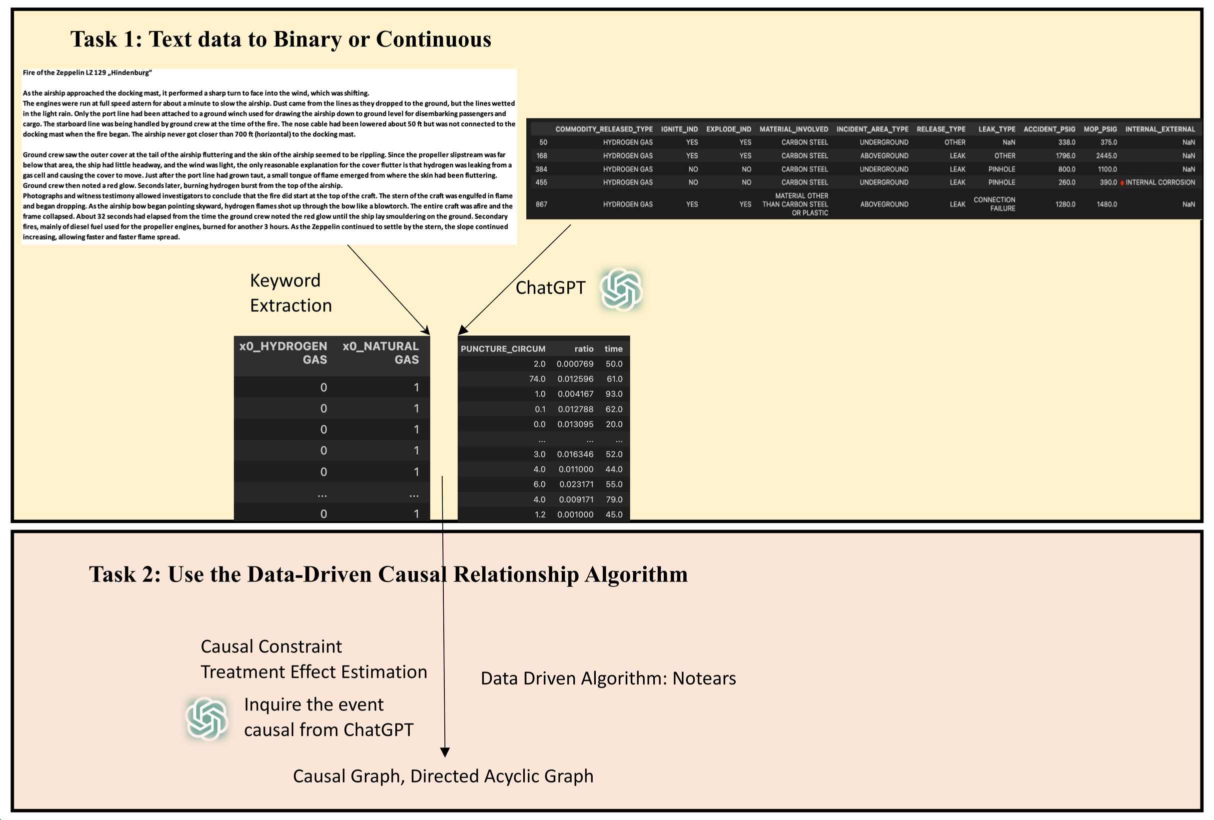
Ira A. Fulton Schools of Engineering, Arizona State University — U.S. Department of Transportation — Pipeline and Hazardous Materials Safety Administration

Objective & Challenge

Objective This project aims to develop a risk-based maintenance optimization framework based on **Natural Language Processing (NLP)** and **Bayesian causal network (BCN)** for hydrogen gas pipelines to provide decision support for pipeline integrity management.

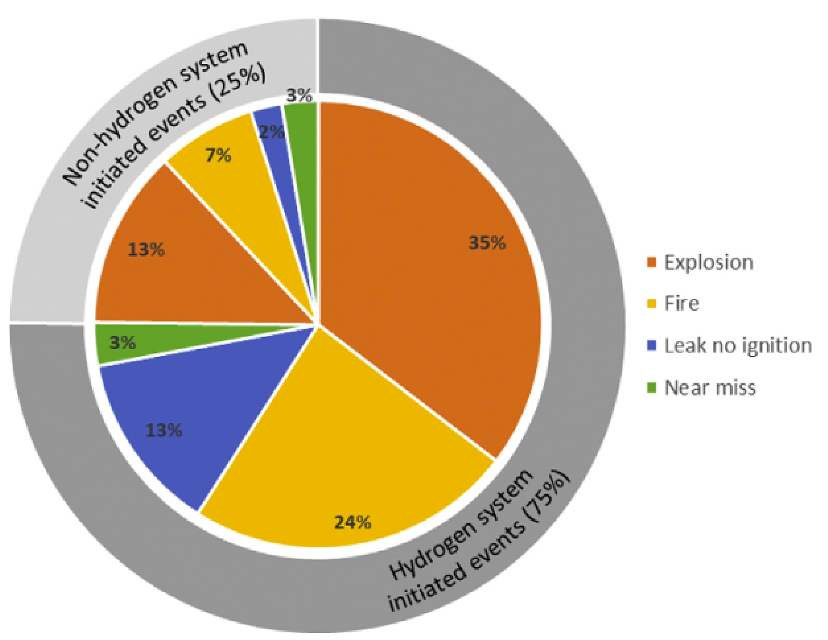
Challenge

- **Heterogeneous Text and Observation Data**: Unstructured Accident Report and continuous measurement data
- **Data Imbalance**: Few Hydrogen-Related Data



Dataset

- **PHMSA**: PHMSA's Office of Pipeline Safety (OPS) provides a variety of data about federally-regulated and state-regulated natural gas pipelines, hazardous liquid pipelines, and liquefied natural gas (LNG) plants. **Complete but only has 5 hydrogen-related accidents.**
- **HIAD 2.0**: With more than 700 hydrogen events – accidents, incidents, and near misses – collected in a database, HIAD was one of the main results of the NoE HySafe (2004-2009) and still represents one of the largest collections for hydrogen-specific data. **All Hydrogen related but incomplete.**



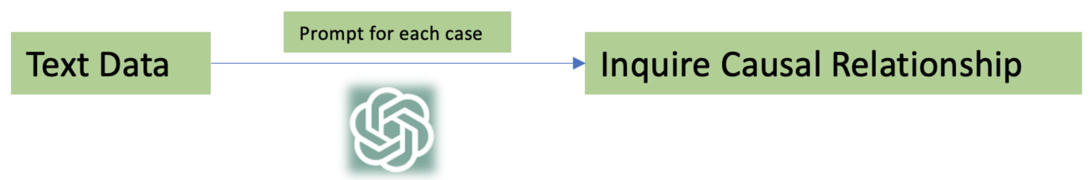
Methodology

Our goal is to estimate the Causal Graph based on the text data and observational data

- Use GPT model for feature extraction and analysis of text data
- Use Causal discovery algorithms NOTEARS for efficient causal graph estimation from observational data
- Use Causal Effect Estimation and Interpretable Machine Learning (Shapley value) to understand the effect of Hydrogen gas and other factors on safety issues.

Data Preprocessing

- **Event Extraction**: Use a combined Keyword Matching and GPT model to extract event from the accident Report
- **Data Completion**: Too many missing data, use GPT to extract and complete the missing entries if can be found in the accident report.



Understand Factors on Accident Severity

- Shapley value is a good way to get some important factors. This method considers every variable as an agent, and then quantifies their corresponding contribution to the outcome.
- Here we consider the injury or Fatal event as the accident severity.

- RELEASE_TYPE, IGNITE and EXPLODE increase severity significantly.

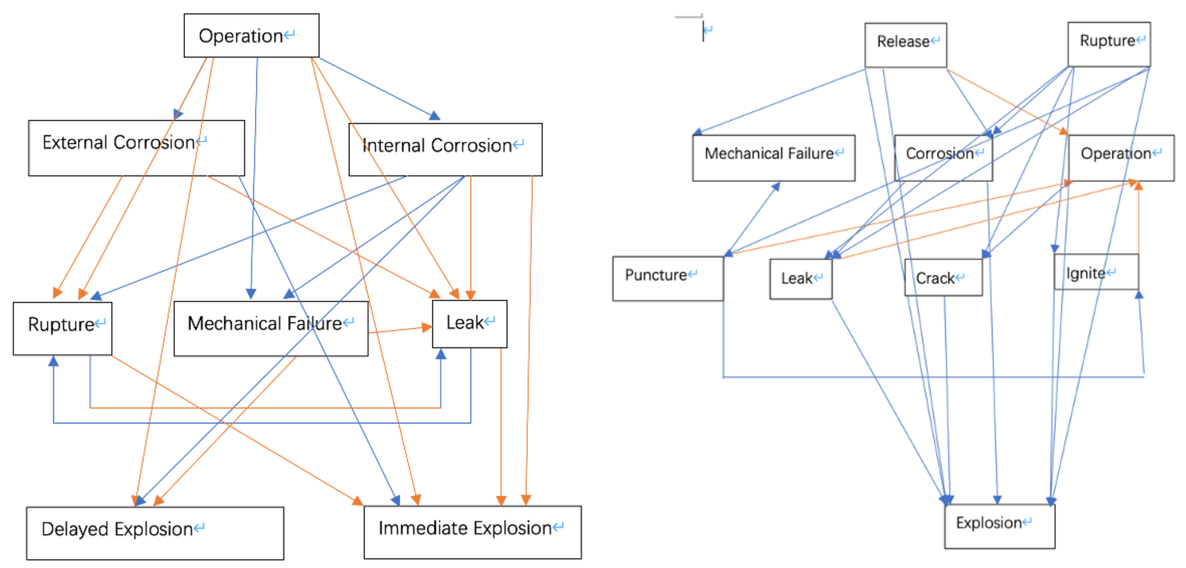
Causal Discovery

- **DAG Constraint**: Any DAG Graph satisfies $h(A) = tr(P(A)) - c_0 d = 0$, where $P(A) = c_0 I + c_1 A + \cdot + c_d A^d$ with $c_i \geq 0$
- **Knowledge Constraint**: Use knowledge queried from GPT as additional soft constraints for causal discovery
- **Algorithm**: The algorithm uses the dual ascent optimization method to solve the model such as Augmented Lagrangian Model

$$F(W) = \frac{1}{2}\|X - XW\|^2 + \lambda\|W\|_1 \qquad L^{\rho}(W,\alpha) = F(W) + \frac{\rho}{2}|h(W)|^2 + \alpha h(W)$$

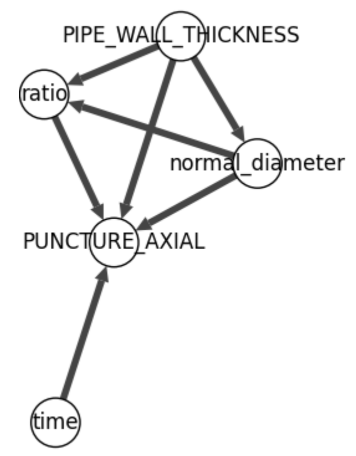
Causal Discovery on HIAD 2.0 with GPT

- Directly Ask GPT on Causal Graph Structures
- Combine Causal Discovery and GPT Results by constrained optimization on the adjacency matrix



Stress Corrosion Cracking (SCC) Causal Discovery

- Consider Stress Corrosion Cracking with variables, Report date, Diameter, Thickness, Specified minimum yield strength (SMYS), Manufactured year, Pressure at accident and extract **ratio features** $= \frac{\text{Pressure at accident}}{\text{SMYS}}$
- Treatment Effect, Y = THICKNESS, O = PUNCTURE_AXIAL, $E[O/Y = 1] - E[0/Y = 0] = -19.4111$, which means more thick, less puncture.



Causal Discovery Algorithm

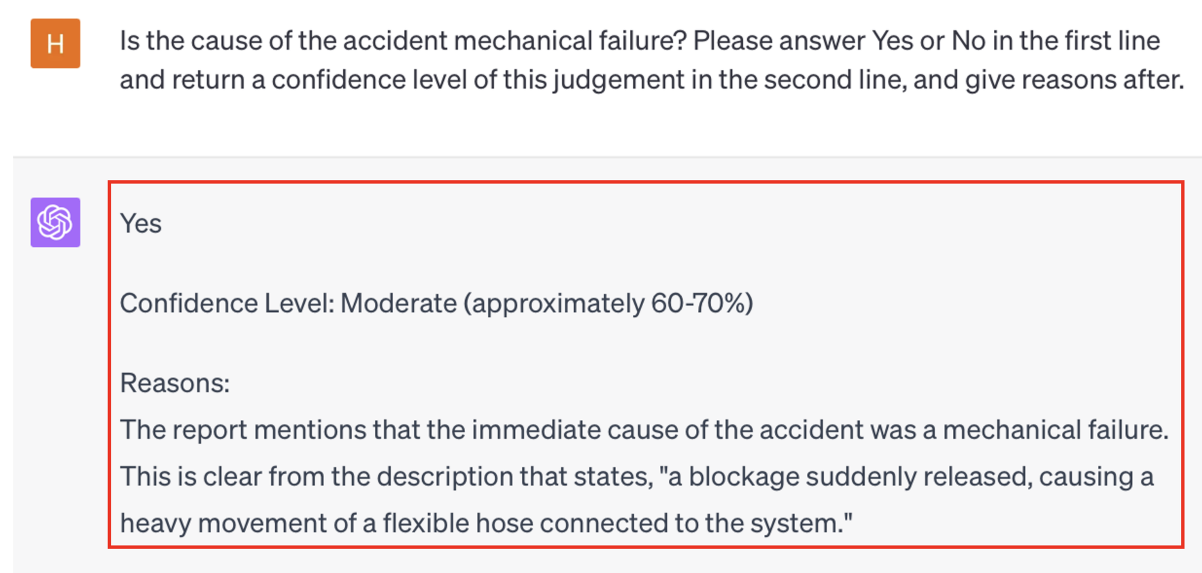
Goal: Understand the effect of Hydrogen on Total Cost, Internal Corrosion, and External Corrosion. The Treatment is defined as Hydrogen Gas.

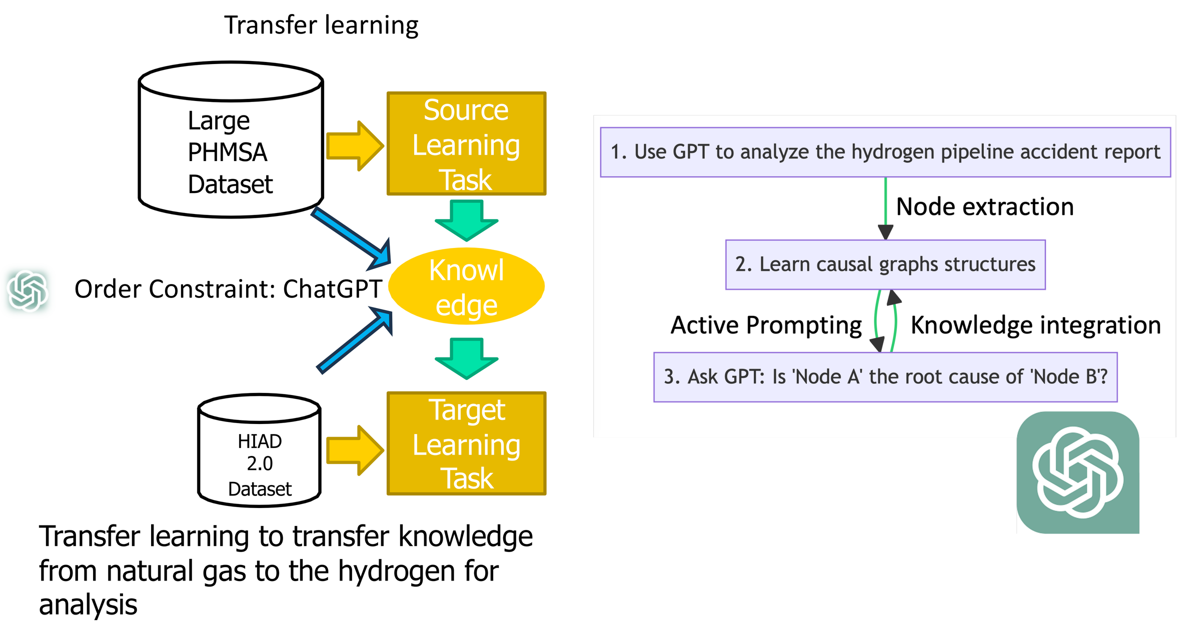
Outcome	Treatment Effect	CI
Total Cost	0.335	(0.314,0.533)
Internal Corrosion	0.27	(0.24,0.33)
Explode	0.6	(0.314, 0.886)

Future Work

- Combine the GPT Models and Causal Discovery Algorithm on Sequential Knowledge Inquiry
- Make the Constrained Causal Discovery Algorithm more stable

The prompt example for ChatGPT could be added for Causal Discovery.





Acknowledgement

We would like to thank the Department of Transportation for providing the funding support for the project and releasing the PHMSA dataset. We would also like to thank FCH 2 JU JRC for providing us with the complete HIAD 2.0 dataset. We like to thank Brian Enrhart from the HyRAM team for providing us with many useful guidelines throughout the projects.