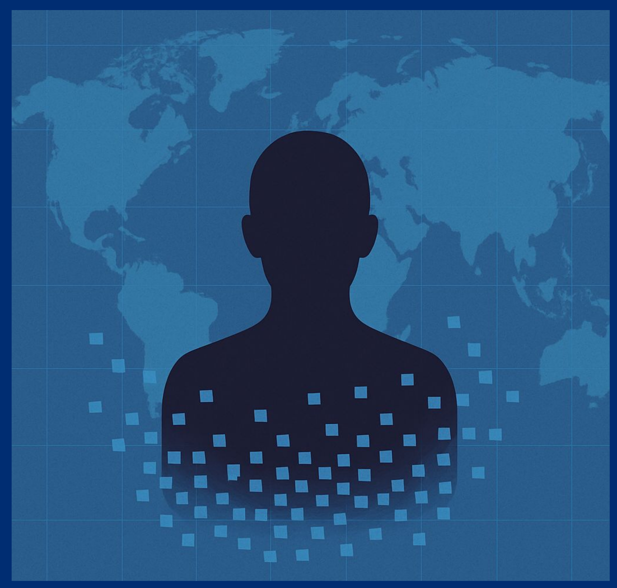
DRAWING PATTERNS IN HUMAN TRAFFICKING DATA THROUGH COVARIANCE ANALYSIS



Oren Wei, Marvin Larweh, Ryan Zhang, Hope Ugwuoke

PI: Dr. Joshua Vogelstein

Department of Biomedical Engineering, Department of Applied Mathematics and Statistics, Department of Computer Science

Johns Hopkins University, Baltimore, MD

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Background

- Data available on global patterns of human trafficking is often sparse and anonymized
- Through a variety of statistical methods, we can both fill in gaps in data and find similarities within it
- These similarities can help identify trends in trafficking which may be useful for policy making and law enforcement purposes

Objectives

- To use data simulation and imputation methods to fill gaps in the existing dataset
- To use graph covariance and geodesic distance to find similarities within the data
- To use large language models (LLM) to reference geopolitical context in order to interpret these similarities

The Datasets

Data Simulation

- 21 time steps, 7 variables with various relationships:
  - **X1** ~ Bernoulli(0.2) for all t, acting as the baseline.
  - **X2** ~ Bernoulli(0.1) for all t, which remains independent of X1 across all time points.
  - **X3** ~ Bernoulli(0.9) × X1 for all t, which is highly dependent on X1 throughout time.
  - **X4** ~ Bernoulli(0.1 + 0.1 t) × X1 for all t, which exhibits increasing dependence on X1 as time progresses.
  - **X5** ~ Bernoulli(0.9 − 0.1 t) × X1 for all t, which shows a decreasing dependence on X1 over time.
  - **X6** ~ Bernoulli(0.1 + 0.2 (t − 10)) × X1 for all t. This attribute has a shifting dependence on X1, with the strength of dependence decreasing from t = 1 to 10 and then increasing from t = 10 to 21.
  - **X7** ~ Bernoulli(0.1) for t < 21, and **X7** ~ Bernoulli(0.9) × X1 for t = 21. This attribute is independent of X1 until t = 21, where it experiences a sudden surge in dependence at that time point.

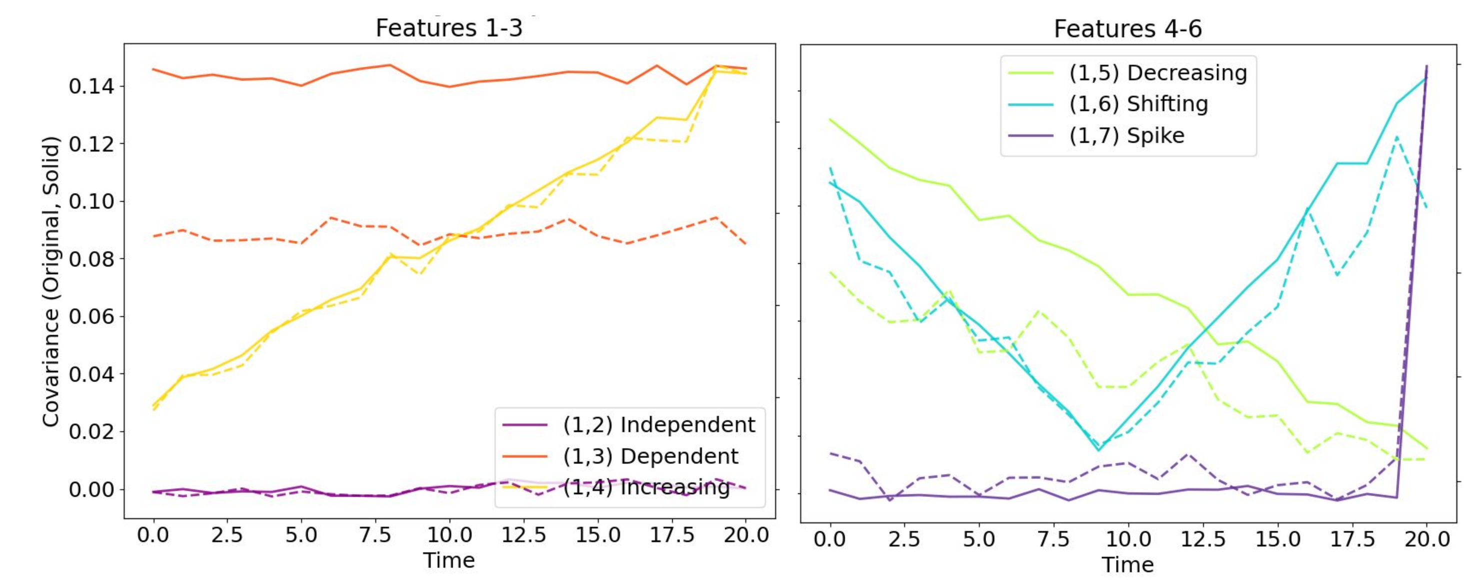
Counter Trafficking Data Collaborative (CTDC)[1]

- Largest publicly available individual-level data on human trafficking inc. over 220,000 victims across 197 countries
- Aggregated across several different data sources
- Anonymized via Microsoft Intelligence Toolkit[2]
- One-hot encoded: 21 time steps, 156 variables
- High degree of missingness (50+%)

Data Simulation

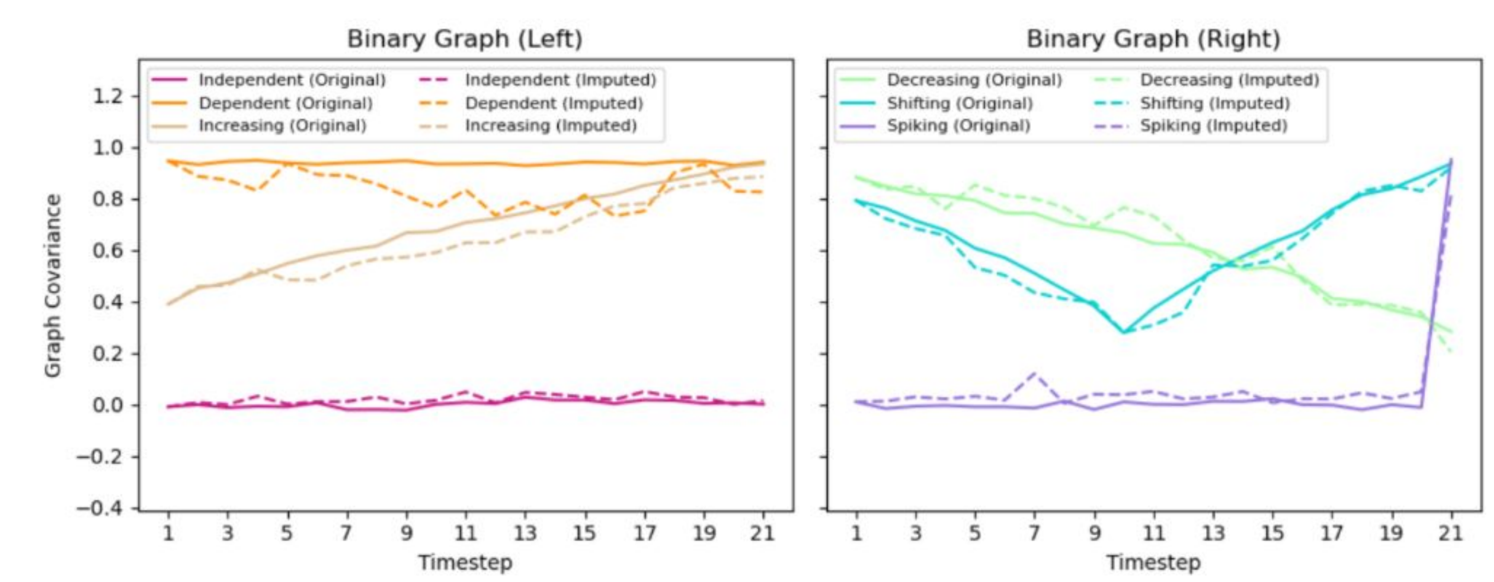
- Various data imputation methods were evaluated in their ability to preserve simulated trends when imputing missing data
- Metrics for assessing accuracy:
  - **Pearson Correlation** between original and imputed lines
  - **Normalized Mean Squared Error (NMSE)**, normalized by range
  - **Normalized Mean Absolute Error (NMAE)**, normalized by range

Random Forest Imputation



Pearson	0.672
NMSE	0.064
NMAE	0.064

kNN Imputation (k=3, 5, 7)

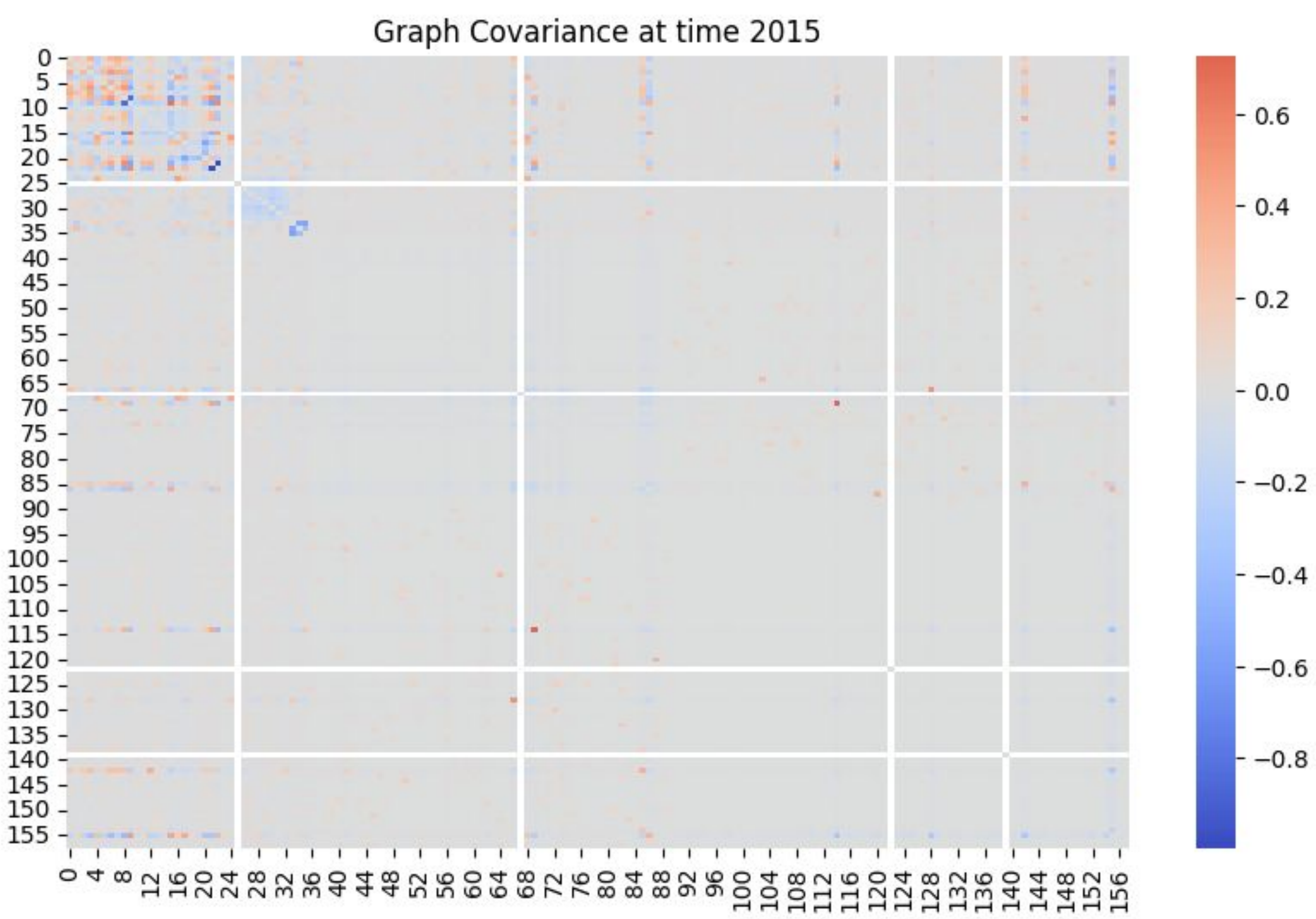


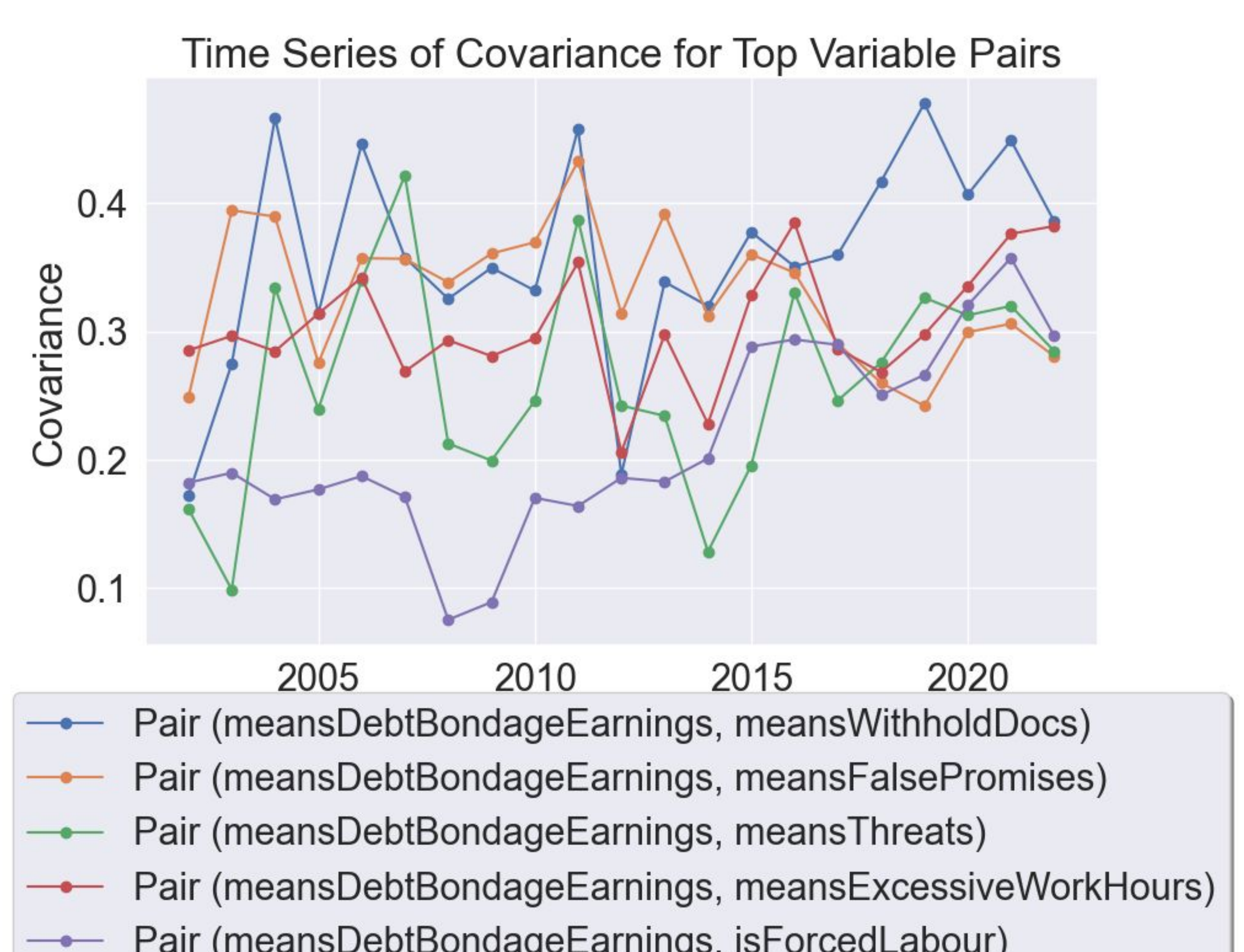
Pearson	0.815
NMSE	5.622
NMAE	0.922

Finding Variables of Interest

Objective: Find variables in the data that have high correlations with each other.

- Our approach uses graph covariance to determine which variables are highly correlated
- Our algorithm results in a **>5000x** speedup compared to the Microsoft Intelligence Toolkit
- Top variable pairs are extracted to be used for downstream prediction and root cause analysis





Large Language Model & Geopolitical Context

- LLMs can contextualize trafficking trends by linking data patterns to global events (e.g., recessions, COVID-19) and shifts in enforcement (e.g., TIP Reports, Palermo Protocol)
- We used **ChatGPT o4-mini-high** to conduct web searches and generate year-by-year interpretations of how debt-bondage earnings covary with other coercion tactics like document withholding, false promises, and forced labor
- The model helped identify key periods of change, such as:
  - **2002–2003:** Spikes in document withholding aligned with expanding transnational trafficking networks and weak oversight.
  - **2010–2012:** A boom-bust cycle driven by economic crisis and subsequent law-enforcement crackdowns.
  - **2016–2018:** Record highs in ID confiscation tied to stricter border control policies.
- This approach enables policymakers, NGOs, and researchers to not only track trends but also understand the why behind them, allowing for more data-informed interventions

Conclusions and Next Steps

- Data imputation methods are able to fill in otherwise missing data while preserving trends over time
- Our covariance algorithm provides an efficient means of assessing the similarity of relationships between variables
- Changes in covariance over time can be mapped to historical or geopolitical events which may help to explain those changes

References

1. https://www.ctdatacollaborative.org/page/global-synthetic-dataset
2. https://github.com/microsoft/intelligence-toolkit